Thèse de doctorat en cotutelle internationale

Sorbonne Université,
Université de Sherbrooke

École doctorale 386 Sciences Mathématiques de Paris Centre
Équipe SERENA (Inria Paris),
Faculté des Sciences,
Laboratoire BISOUS (Université de Sherbrooke)

Une méthode de linéarisation robuste pour les problèmes de complémentarité

Un détour par les arrangements d'hyperplans

Soutenue par : Baptiste PLAQUEVENT-JOURDAIN

En vue de l'obtention du grade de Philosophiae Doctor (Ph.D.) en mathématiques

Encadrée par Jean-Pierre DUSSAULT et Jean Charles GILBERT

Présentée et soutenue publiquement le 16/07/2025 à Paris, France

Avec le jury présidé par Mounir Haddou et composé de :

Rapporteur	Mounir HADDOU, Professeur des Universités, INSA Rennes
Rapporteur	Miroslav RADA, Professeur assistant, Prague University of Economics and Business
Examinatrice	Ibtihel BEN GHARBIA, Dr, IFPEN, Rueil-Malmaison
Examinatrice	Virginie CHARETTE, Professeur titulaire, Université de Sherbrooke
Encadrant	Jean-Pierre DUSSAULT, Professeur titulaire, Université de Sherbrooke
Encadrant	Jean Charles GILBERT, DR, Inria Paris

Summary

Title : A Robust Linearization Method for Complementarity Problems – A Detour Through Hyperplane Arrangements.

Keywords : complementarity ; nonsmooth analysis ; hyperplane arrangements ; algorithms

Abstract : The initial goal of this thesis is the resolution of complementarity problems. These problems are reformulated here by the minimum C-function, which is piecewise linear, so nondifferentiable, and leads to nonsmooth systems of equations to solve. The globalization of pseudo-linearizing methods for such equations (semismooth Newton method for instance) may face the following difficulty : the computed directions are not necessarily descent directions for the associated merit function, used for linesearch methods (whereas in the smooth case, the opposite of the gradient is always suitable).

The piecewise nature of the merit function induced by the minimum C-function implies to choose one certain piece, and this thesis proposes, in its chapter 6, an approach on this question, via geometric observations allowing to describe the difficulty of the task.

In the case of the minimum C-function, a recent method replaces the direction of pseudo-linearization by finding a direction in a suitable convex polyhedron. However, to ensure all the stationary points of the generated sequence are solutions of the problem, they must verify a stringent regularity condition. This one then ensures the convex polyhedron is nonempty in the neighborhood of such points. The initial goal of this thesis was to avoid this regularity assumption, like for smooth systems, by using the Levenberg-Marquardt approach.

While searching to better understand this method and to analyze the B(ouligand)-differential of the minimum C-function, which plays a central role, it appeared that, in the simple case of linear (affine) problems, the inherent structure of this B-differential was the one of a hyperplane arrangement. This very classic problem of combinatorial geometry, that we discovered at this occasion, is in fact surprisingly rich and deep (which was fully acknowledged by specialists).

We propose, in chapters 3 and 5, an analysis related to the question of nonsmooth methods as well as improvements on a state-of-the-art algorithm to compute the chambers. In particular, "(primal-)dual" variants, which use explicitly a link between the chambers of an arrangement and the circuits of its underlying matroid, seem promising.

This long detour, which constitutes the major part of this thesis, ended up being insightful for the nonsmooth method and the choice of the piece – linearizing the functions results in a (B-)differential of the minimum of affine functions – but we believe that it brought to light interesting links between nondifferentiability and combinatorial geometry.

Titre : Une méthode de linéarisation robuste pour les problèmes de complémentarité – Un détour par les arrangements d'hyperplans

Mots-clés : complémentarité ; analyse non lisse ; arrangements d'hyperplans ; algorithmes

Résumé : Le but initial de cette thèse est la résolution de problèmes de complémentarité. Ces problèmes sont reformulés ici par la C-fonction minimum, qui est linéaire par morceaux, donc non différentiable, ce qui conduit à des systèmes d'équations non lisses à résoudre. La globalisation de méthodes pseudo-linéarisant de telles équations (Newton semi-lisse par exemple) se heurte généralement à la difficulté que les directions calculées ne sont pas nécessairement de descente pour la fonction de mérite associée, utilisée par les méthodes de recherche linéaire (alors que dans le cas d'équations différentiables, l'opposé du gradient de la fonction de mérite convient toujours).

Dans le cas de la C-fonction minimum, une méthode récente remplace la direction de pseudo-linéarisation par une direction trouvée dans un polyèdre convexe adapté. Cependant, pour s'assurer que tous les points stationnaires de la suite générée soient solutions du problème, ceux-ci doivent vérifier une condition de régularité contraignante. Celle-ci assure alors que polyèdre convexe n'est pas vide dans le voisinage de tels points. L'objectif initial de cette thèse était de se libérer de cette hypothèse de régularité, comme pour les systèmes lisses, en utilisant l'approche de Levenberg-Marquardt.

Le caractère différentiable par morceaux de la fonction de mérite induit par la C-fonction minimum implique de devoir choisir un certain morceau, et cette thèse propose, dans son chapitre 6, une approche sur cette question, via des observations géométriques permettant une description de la difficulté de la tâche.

En cherchant à mieux comprendre cette méthode et à analyser le B(ouligand)-différentiel de la fonction minimum, qui y joue un rôle central, il est apparu, dans les cas simples de problèmes linéaires (ou affines), que la structure inhérente à ce B-différentiel est celle d'un arrangement d'hyperplans. Ce problème très classique en géométrie combinatoire, que nous avons découvert à cette occasion, s'est révélé surprenamment riche et profond (ce que les spécialistes de ce domaine savaient parfaitement).

Nous proposons, aux chapitres 3 et 5, une analyse en lien avec la question des méthodes non lisses ainsi que des améliorations sur un algorithme de l'état de l'art identifiant les chambres. En particulier, des variantes "(primales-)duales", reliant explicitement les chambres d'un arrangement et les circuits du matroïde associé, semblent prometteuses.

Ce long détour, qui constitue la majeure partie du manuscrit, s'est révélé instructif pour l'algorithme non lisse et le choix du "morceau" – la linéarisation des fonctions faisant apparaître un (B-)différentiel du minimum de fonctions affines – mais nous pensons surtout que cela a permis de mettre en lumière des liens intéressants entre non-différentiabilité et géométrie combinatoire.

Remerciements

C'est fait. La longue aventure qu'est cette thèse touche à sa fin. Avant de plonger dans son sujet, je souhaite témoigner ma gratitude à toutes les personnes qui l'ont rendue possible.

Tout d'abord, à Jean-Pierre et Jean Charles, merci pour m'avoir permis de parcourir ce chemin durant ces quatre (et quelques) années. Grâce à vous, j'ai pu beaucoup réfléchir et travailler sur ce qui m'intéressait, de l'optimisation non différentiable à la géométrie computationnelle en passant par la combinatoire. Au-delà de me laisser explorer dans mon coin certains aspects un peu curieux, nous avons pu travailler ensemble, découvrir le fascinant domaine des arrangements et approfondir des questions d'analyse non lisse. Par ailleurs, la (toute) fin de la thèse relie les arrangements et les problèmes de complémentarité dans certaines preuves : bien que le travail ne soit pas encore fini, j'ai la conviction que l'on a réussi à éclaircir une partie des mystères du minimum et j'espère que l'on pourra continuer à y travailler ensemble. Votre dévouement à la recherche et la quête de la connaissance, votre passion, votre minutie, et bien d'autres, resteront toujours un objectif à atteindre pour moi. Encore une fois, merci.

L'organisation de cette thèse en cotutelle internationale a nécessité beaucoup d'aide et je souhaite remercier celles et ceux qui l'ont facilitée : Patricia Zizzo, Josée Lamoureux, Annie Carbonneau, Thomas Brüstle, Anne MacKay, Félix Camirand Lemyre, Derya Gök, et bien d'autres. J'aimerais aussi remercier les instituts canadiens MITACS et ISM pour leur soutien financier.

À ma famille, je ne saurais jamais vous remercier assez. Vous m'avez toujours aidé et soutenu, vous êtes toujours intéressé à ma progression et ce qui passait dans ma thèse, même si le sujet ne s'y prêtait pas nécessairement le mieux. Vous êtes merveilleux, je réalise la chance que j'ai de vous avoir, et j'espère pouvoir partager davantage avec vous à l'avenir.

Merci à Charles pour les encouragements et conseils systématiques. Aux membres étudiants du BISOUS : Luc, Tania, Vivien, Nicolas, Arthur, Josué, Christopher et les (nombreux !) stagiaires, merci pour ces étés passés ensemble. J'ai beaucoup apprécié le temps avec vous, partager de nombreux repas, surtout lors des gâteaudis, de passionnantes discussions et sorties. C'était assez nécessaire entre deux bizarreries de logement (poison pour cafards, démence, zoos…). Je vous souhaite le meilleur pour la suite. Merci aussi à Sylvain, et particulièrement à Guillaume pour les (*fort* nombreux) matins et les rapides tête-à-tête – dommage que j'ai été le seul autre matinal au BISOUS ! À SERENA, merci de nous avoir ac-

cueilli pour la seconde partie de ma thèse. Je reconnais ne pas avoir fait beaucoup d'efforts d'intégration, mais j'espère que vous aurez apprécié aussi les gâteaudis.

Je voudrais aussi remercier, bien que pour un rôle assez particulier et que l'on ne se soit jamais vus, les membres des communautés de "yellowhat" et "NumottheNummy", pour ces innombrables discussions, débats, Cavernes, trivia et ces infinis divertissements, le tout grâce à "Godfield".

Enfin, aux ami·es, merci pour m'avoir supporté, merci pour le soutien, et merci pour tout. Des excuses sincères aux personnes que je n'ai pas citées. Un merci tout particulier aux gens de la fusée d'Ariane : Aurore (et au passage à la légende d'Amogus), Basile, Servane, Ronan, Aliénor, Guillaume, Pauline. Juliette, Julieng, Léopold(ieu), Benoît B, Benoît S, pour le soutien moral, l'aide et les conseils. Elias, pour l'aide dans certains moments. Louis, Agnès, Yousra, Adrien, pour nos précieuses discussions. Étienne et Hugo, pour les (bien trop peu) nombreuses séances de jeux à la Maison Riquet. Pour le *gaming*, avec Quentin, Lucas et Florian. Également avec Olivier et Zoé, mais aussi pour de fantastiques sorties, malgré le soleil, l'air pur et le dehors ; Alice, pour tout ce que l'on a toujours partagé et que surtout on partage toujours – tu conviendras que suffisamment de mots ont été écrits… À Tom et Jérémy, merci pour toutes ces discussions et ces bons moments.

Table des matières

Summary i

Remerciements v

Liste des figures xviii

Liste des tableaux xxi

1 Introduction 1
- 1.1 Point de départ et motivations . . . 1
- 1.2 Problèmes reliés . . . 2
- 1.3 Formulations équivalentes et algorithmes . . . 3
- 1.4 Non-différentiabilité et géométrie combinatoire . . . 4
- 1.5 Plan et contributions . . . 6

2 Cadre général 9
- 2.1 Notations . . . 10
  - 2.1.1 Notations générales . . . 10
  - 2.1.2 Notations spécifiques . . . 12
- 2.2 Reformulations de problèmes de complémentarité . . . 12
  - 2.2.1 Quelques types de PCL . . . 12
  - 2.2.2 Quelques types de PCN(F) . . . 13
  - 2.2.3 Complexité générale . . . 14
  - 2.2.4 Équations généralisées et applications normales . . . 14
  - 2.2.5 Méthodes de points intérieurs . . . 15
  - 2.2.6 Équation de valeur absolue . . . 16
  - 2.2.7 Problèmes avec complémentarité dans les contraintes . . . 16
- 2.3 Cadre non lisse et algorithmes . . . 17
  - 2.3.1 Introduction aux C-fonctions . . . 17
  - 2.3.2 Quelques outils d'analyse non lisse . . . 19
  - 2.3.3 Premiers algorithmes non lisses . . . 25
  - 2.3.4 Traitement particulier du minimum . . . 32
  - 2.3.5 Autres méthodes non lisses . . . 35
  - 2.3.6 Techniques de lissage . . . 43
  - 2.3.7 Un commentaire sur la complexité . . . 46
- 2.4 Sur l'aspect combinatoire . . . 48
  - 2.4.1 Relation avec les sujets précédents . . . 48

2.4.2 Références classiques . . . 49
2.4.3 Quelques outils spécifiques . . . 49
2.4.4 Matroïdes orientés (et circuits) . . . 50
2.4.5 Logiciels d'algèbre . . . 52
2.4.6 Algorithmes spécifiques pour identifier les chambres . . . 52
2.4.7 Quelques exemples d'applications . . . 54

3 B-différentiel du minimum de deux fonctions vectorielles affines 57
3.1 Introduction . . . 58
3.2 Background . . . 61
3.3 Equivalent problems . . . 63
3.3.1 B-differential of the minimum of two affine functions . . . 64
3.3.2 Linear algebra problems . . . 64
3.3.3 Convex analysis problems . . . 70
3.3.4 Discrete geometry : hyperplane arrangements . . . 74
3.4 Description of the B-differential . . . 76
3.4.1 Some properties of the B-differential . . . 77
3.4.2 Cardinality of the B-differential . . . 79
3.4.3 Particular configurations . . . 84
3.4.4 A glance at the C-differential . . . 84
3.5 Computation of the B-differential . . . 85
3.5.1 Computation of a single Jacobian . . . 86
3.5.2 Computation of all the Jacobians . . . 87
3.6 Discussion . . . 105
Acknowledgments . . . 106
Statements & Declarations . . . 106

4 Éléments complémentaires sur le B-différentiel du minimum et les arrangements 107
4.1 Matériel complémentaire du chapitre précédent . . . 108
4.2 Notions de régularité et contre-exemples . . . 110
4.3 B-différentiel du minimum de F et G non linéaires . . . 111
4.3.1 Différentiels de H . . . 111
4.4 Différentiel de la fonction de mérite . . . 120
4.5 Détails sur les instances et algorithmes . . . 130
4.5.1 À propos des instances permutahedron . . . 130
4.5.2 À propos de l'arrangement de séparabilité du crosspolytope . . . 136
4.5.3 Instances parfaitement symétriques . . . 140

5 Approches primales et duales pour énumérer les chambres d'arrangements d'hyperplans 143
5.1 Introduction . . . 144
5.2 Background . . . 146
5.3 Hyperplane arrangements . . . 147
5.3.1 Presentation . . . 147
5.3.2 Properties . . . 150
5.3.3 Stem vectors . . . 155

5.3.4 Augmented matrix . . . 161
5.4 Chamber computation - Primal approaches . . . 170
5.4.1 Primal $\mathcal{S}$-tree algorithm . . . 171
5.4.2 Preventing some computations . . . 176
5.5 Chamber computation - Dual approaches . . . 178
5.5.1 Algorithms using all the stem vectors . . . 179
5.5.2 Algorithms using some stem vectors . . . 182
5.6 Compact version of the algorithms . . . 186
5.6.1 The compact $\boldsymbol{\mathcal{S}}$-tree . . . 187
5.6.2 Compact primal $\mathcal{S}$-tree algorithm . . . 188
5.6.3 Compact primal-dual $\mathcal{S}$-tree algorithm . . . 192
5.7 Numerical results . . . 195
5.7.1 Arrangement instances . . . 195
5.7.2 Assessed algorithms . . . 196
5.7.3 Numerical results . . . 198
5.8 Conclusion . . . 201
5.9 Appendix : tables with numerical results . . . 202

6 Globalisation de PNM par moindres-carrés et Levenberg-Marquardt 205
6.1 Modification de Newton-min polyédrique . . . 206
6.1.1 Présentation de la méthode . . . 206
6.1.2 Variante moindres-carrés et Levenberg-Marquardt . . . 208
6.1.3 Choix des poids et stationnarité . . . 218
6.1.4 Choix des poids et différentiels . . . 222
6.1.5 Régularité de solutions . . . 224
6.2 Un algorithme envisagé . . . 227
6.2.1 La méthode et ses propriétés . . . 228
6.2.2 Modifications et améliorations potentielles . . . 233

Conclusion 237

A Informations détaillées sur les instances affines et les algorithmes 239
A.1 Détails sur des propriétés du chapitre 5 . . . 239
A.2 Valeurs des instances . . . 242
A.3 Comportements algorithmiques . . . 244
A.3.1 Heuristiques primales . . . 244
A.3.2 Heuristiques duales . . . 247
A.3.3 Analyse de l'algorithme compact . . . 248
A.4 Instances linéaires et autres sujets . . . 254
A.4.1 Calcul des circuits . . . 255
A.4.2 Test de couverture récursive . . . 256

B Éléments de géométrie sur les polytopes 263
B.1 Polytopes et leurs faces . . . 263
B.2 Propriétés spécifiques des zonotopes . . . 267

C Inclusion de zonotopes 273

D Poids et élément du différentiel de Clarke **281**
D.1 Géométrie et ensembles de vecteurs de signes 281
D.2 Preuve principale 283
D.2.1 Contre-exemples détaillés (simplifiés) 290
D.2.2 Dégénérescences et corrections (théoriques) 300

Bibliographie **309**

Liste des figures

1.1 Illustration avec trois hyperplans et sept chambres, avec les hyperplans $H_1 = \{x \in \mathbb{R}^2 : x_1 = 0\}$, $H_2 = \{x \in \mathbb{R}^2 : x_2 = 0\}$ et $H_3 = \{x \in \mathbb{R}^2 : x_1 + x_2 = 1\}$. 5

3.1 The figure is related to the linear complementarity problem defined by example 3.3.2 : the v_i's are the columns of the matrix V (their third zero components are not represented). Each of the 6 sets of vectors plots the 3 vectors $\{s_i v_i : i \in [1:3]\}$, for each of the 6 sign vectors $s \in \mathcal{S}$ (given by the columns of the matrix S in (3.13)), as well as a direction d (given by the columns of D in (3.13), dashed lines) such that $s_i v_i^\mathsf{T} d > 0$ for all $i \in [1:3]$. Each conic hull of these vectors, namely $\mathrm{cone}\{s_i v_i : i \in [1:3]\}$, is pointed. The conic hulls of $\{v_1, v_2, v_3\}$ and $\{-v_1, -v_2, -v_3\}$ are both the space of dimension 2, hence there are not pointed, which confirms the fact that $(1,1,1)$ and $(-1,-1,-1)$ are not in $\mathcal{S}$. 71

3.2 Linearly separable bipartitions of a set of $p = 4$ points $\bar{v}_i$ in $\mathbb{R}^2$ (the dots in the figure). Possible separating hyperplanes are the drawn lines. We have not represented any separating line associated with the partition $(\varnothing, [1:p])$ or $([1:p], \varnothing)$, so that $|\mathcal{S}| = 2(n_s+1)$, where n_s is the number of represented separating lines. We have set $r := \dim(\mathrm{vect}\{\bar{v}_1, \ldots, \bar{v}_p\}) + 1$. 72

3.3 Illustration of problem 3.3.19 (arrangement of hyperplanes containing the origin) for the 3 vectors that are the columns on the matrix V in example 3.3.2 (since the last components of these v_i's vanish, only the first two ones are represented above). The hyperplanes $\mathcal{H}_i$ are defined by (3.28). The regions to determine are represented by the sign vectors here denoted $(s_1 s_2 s_3)$ with $s_i = \pm$: if $d \in \mathbb{R}^2$ belongs to the region $(s_1 s_2 s_3)$, then $s_i = +$ if $v_i^\mathsf{T} d > 0$ and $s_i = -$ if $v_i^\mathsf{T} d < 0$. We see that there are only $6 = 2p$ regions among the $8 = 2^p$ possible ones ; the regions $(+++)$ and $(---)$ are missing, which reflects the fact that $+ v_1 + v_2 + v_3 = 0$ and $-v_1 - v_2 - v_3 = 0$ (see problem 3.3.6). 75

3.4 Half of the $\mathcal{S}$-tree for example 3.3.2 (the other half is obtained by swapping the $+$'s and the $-$'s). Top-down arrows indicate descendence ; the sign sets $\mathcal{S}_k^1$ are defined by (3.42). 88

4.2 Illustration de l'idée du processus de récurrence : en violet les deux hyperplans ajoutés. Le point noir en haut représente un point avec $d_0 = 0$ et $d_{[1:n]}$ arbitraire qui a 4 descendants. Le point noir en bas représente un point avec $d_0 \neq 0$ et $d_{[1:n]} = 0$ qui n'a que 3 descendants (deux flèches plus lui-même). 139

5.1 Arrangements in $\mathbb{R}^2$ specified by the hyperplanes $H_1 := \{x \in \mathbb{R}^2 : x_1 = 0\}$, $H_2 := \{x \in \mathbb{R}^2 : x_2 = 0\}$, H_3(left) $:= \{x \in \mathbb{R}^2 : x_1 + x_2 = 0\}$, H_3(middle) $:= \{x \in \mathbb{R}^2 : x_1 + x_2 = 1\}$ and H_3(right) $:= \{x \in \mathbb{R}^2 : x_1 + x_2 = -1\}$. The origin is contained in all the hyperplanes but in H_3(middle) and H_3(right), so that the arrangement in the left-hand side is *linear* with 6 chambers and the other ones are *affine* with 7 chambers. 148

5.2 Symbolic representation of the sets $\mathfrak{S}(V,\tau)$, $\mathfrak{S}_s(V,\tau)$, $\mathfrak{S}_a(V,\tau)$, $\mathfrak{S}(V,0)$, $\mathfrak{S}_0(V,\tau)$ and $\mathfrak{S}([V;\tau^\mathsf{T}],0)$, respecting propositions 5.3.14, 5.3.21 and 5.3.23. The horizontal dashed line aims at representing the reflexion between a stem vector σ and its opposite $-\sigma$: $\mathfrak{S}_s(V,\tau)$, $\mathfrak{S}(V,0)$, $\mathfrak{S}_0(V,\tau)$ and $\mathfrak{S}([V;\tau^\mathsf{T}],0)$ are symmetric in the sense that $-\mathfrak{S}_s(V,\tau) = \mathfrak{S}_s(V,\tau)$, $-\mathfrak{S}(V,0) = \mathfrak{S}(V,0)$, $-\mathfrak{S}_0(V,\tau) = \mathfrak{S}_0(V,\tau)$ and $-\mathfrak{S}([V;\tau^\mathsf{T}],0) = \mathfrak{S}([V;\tau^\mathsf{T}],0)$. By propositions 5.3.15 and 5.3.21, the diagram simplifies when $\tau \in \mathcal{R}(V^\mathsf{T})$, since then $\mathfrak{S}_a(V,\tau) = \mathfrak{S}_a(V,-\tau) = \mathfrak{S}_0(V,\tau) = \varnothing$ and there is only one set left. . . . 157

5.3 Symbolic representation of the sets $\mathcal{S}(V,0)$, $\mathcal{S}(V,\tau)$, $\mathcal{S}_a(V,\tau)$ and $\mathcal{S}([V;\tau^\mathsf{T}],0)$, respecting (5.9), (5.10), (5.11) and propositions 5.3.6 and 5.3.18. The horizontal dashed line aims at representing the reflection between a sign vector s and its opposite $-s$: $\mathcal{S}(V,0)$, $\mathcal{S}([V;\tau^\mathsf{T}],0)$ and $\mathcal{S}([V;\tau^\mathsf{T}],0)^c$ are symmetric in the sense of definition 5.3.4. 162

5.4 $\mathcal{S}$-tree of the arrangement in the middle pane of figure 5.1. The gray node is actually absent from the tree, since there is no chamber associated with $s = (-1,-1,+1)$ (no x such that $s \bullet (V^\mathsf{T}x - \tau) > 0$). 171

5.5 Standard $\mathcal{S}$-trees (left) and compact $\mathcal{S}$-trees (right) of the arrangements in the middle pane (above, compare with figure 5.4) and the right-hand side pane (below) of figure 5.1. The sign vectors in the white boxes are in $\mathcal{T}(V,0)$, those in the blue/gray boxes are in $\mathcal{S}_a(V,\tau)$ and the one in the blue/gray box with bold edges is in $\mathcal{S}_a(V,-\tau)$; this last sign vector must be multiplied by -1 to get a sign vector in $-\mathcal{S}_a(V,-\tau) = \mathcal{S}_a(V,\tau) \subseteq \mathcal{S}(V,\tau)$. 188

5.6 Performance profiles of the RC, P, PD and D algorithms, for the computing time. 199

5.7 Performance profiles of the RC vs RC/C, P vs P/C, PD vs PD/C and D vs D/C algorithms, for the computing time. The dashed lines refer to the compact versions of the algorithms. 200

5.8 Performance profiles of the RC vs PD/C solvers, for the computing time. . 201

6.1 Gauche : courbes de niveau de φ_x avec $\gamma = (1/2, 1/2)$. Droite : courbes de niveau de θ ; les lignes pointillées sont les plis (θ n'est pas différentiable). Le point rouge correspond à un minimum local. Les lignes de niveau révèlent de trop grandes différences entre θ et φ_x, donc la direction donnée par φ_x *augmente* θ. 212

6.2 Gauche : courbes de niveau de φ_x avec $\gamma = (2/3, 1)$. Droite : courbes de niveau de θ ; les lignes pointillées sont les plis (θ n'est pas différentiable). Le point rouge est un minimum local. Les lignes de niveau de φ_x sont (du moins localement) assez proches de celles de θ pour qu'une direction de descente de φ_x *diminue* θ. 213

6.3 Illustration de l'exemple 6.1.4. Les courbes de niveau de θ ($\sqrt{\theta}$ pour la visibilité) sont en couleur, les lignes pointillées en bleu les plis de H (définis par $x_i = (Px+q)_i$ pour $i \in \{1,2\}$), le point rouge au-dessus est l'unique solution $\bar{x} = (0, 1/10)$ du problème et le point bleu est l'itéré courant. Les flèches en vert, bleu, rouge et noir correspondent à quatre $-g$ possibles pour des choix de γ extrémaux, celle en magenta à une direction de descente. . . 213

6.4 Illustration de la faible stationnarité. On a $F(x) = x$, $G(x) = 1 + (x-1)^2$, donc le problème a une solution en $x = 0$. À $x = 1$, ni H ni θ ne sont différentiables. Puisque $G'(1) = 0$, en prenant une suite $1 + t_k \to 1$, on a que $0 \in \partial\theta(1)$, mais $x = 1$ n'est clairement pas fortement stationnaire. De tels points sont parfois appelés "plis concaves", qui, comme on le verra, sont difficiles à traiter. 223

6.5 Les courbes au-dessus de θ sont les modèles quadratiques φ_{x_k}. Bien qu'il puisse y avoir convergence rapide vers $x = 1$, le point peut ne jamais être atteint ($x_k > 1$) donc $\forall\, k, \mathcal{E}(x_k) = \varnothing$. 233

6.6 Illustration de quelques itérés, avec $\tau > 0$ et $\mathcal{E}(x) := \{i \in [1:n] : |F_i(x) - G_i(x)| < \tau\}$, de l'algorithme 6.2.1. 235

A.1 Les lignes noires sont les hyperplans déjà considérés et x est un point de la région courante. Il est simple d'ajouter d'abord les deux hyperplans bleus, qui impliquent un seul descendant, puis d'ajouter les hyperplans en pointillés qui impliquent deux descendants. Bien que la figure soit montrée pour un arrangement centré, le principe est similaire pour le cas affine. 245

A.2 Illustration de (A.4) (et (A.3)) ("LOP" signifie POL). La correction, c'est-à-dire la différence entre (A.3) et (A.4), est notée $(*)$ et ajoutée aux images du bas ; pour les instances en haut à droite des graphiques de droite, cela place les points sur la ligne correspondant à la formule, donnée par $y = 1 + x/2$. Le (c) désigne le nombre de POL de la variante compacte. Pour les images de droite, quatre points, correspondant aux instances PERM, sont décalés vers la droite. Une explication possible est proposée ci-contre. 254

B.1 Illustration des lemmes B.1.1 (gauche) et B.1.4 (droite). À gauche, on voit que les intérieurs relatifs sont obtenus en retirant les parties de P où des égalités $A_{:,i}x = a_i$ sont vraies. Cependant, imaginons le même polytope en dimension 3 (donc avec un intérieur vide), avec les inégalités supplémentaires $e_3^\mathsf{T} x \leqslant 0$ et $-e_3^\mathsf{T} x \leqslant 0$, on ne peut prendre les inégalités strictes puisque l'on aurait un ensemble vide. C'est parce que ces inégalités forment en fait une égalité. À droite, on voit que l'intérieur relatif en bleu corresond à l'intérieur relatif de P (en tant que face de lui-même), alors que la frontière est composée est intérieurs relatifs des faces en magenta et des sommets en rouge. 265

B.2 Illustration des lemmes B.1.3 (gauche) et B.1.5 (droite). À gauche, on peut observer que la face verte correspond à l'ensemble $I = \{4\}$ de taille 1, alors que les sommet en rouge correspond à un ensemble $I = \{1, 2\}$ de taille 2. À droite, la même face verte a une unique (à une constante > 0 multiplicative près) normale c puisque la face est de dimension maximale $n-1$ alors que le sommet en rouge a de multiples normales non colinéaires possibles c (voir la remarque B.1.6). 266

B.3 Exemple d'un zonotope simple avec $V = [e_1\ e_2\ e_1+e_2\ e_3]$. La face supérieure en orange est une face de dimension deux, générée par $e_1, e_2, e_1 + e_2$, avec $I^* = \{4\}$. La face en violet en avant est générée par e_2 et e_3, avec $I^* = \{1, 3\}$. La face verte à droite, qui est une arête, est générée par e_3 avec $I^* = \{1, 2, 3\}$. tous les sommets sont aussi des faces sans générateurs. À droite, certains hyperplans orthogonaux aux normales sont ajoutés. Les faces de dimension maximale ont un seul hyperplan mais les arêtes en ont plusieurs (puisque la dimension est 3). 268

B.4 Gauche : "normal fan" d'un zonotope. Vert clair : sommets et leurs cônes normaux (frontières en pointillés puisque'elles correspondent aux normales des arêtes) ; violet : arêtes et leurs cônes normaux. Droite : Illustration de la proposition B.2.3. Ensembles et flèches en vert : faces et leurs normales. Points rouges : centres des faces (égaux aux faces pour les sommets). Gris : vecteurs générant le zonotope. Noir : les flèches pointillées sont les $V_{:,i}\kappa_i$ pour $i \in I^*$. On a translaté en le centre des faces pour voir plus facilement que les normales c en vert de la proposition B.1.7 ont un produit scalaire positif avec les flèches pointillées en noir. 269

B.5 Exemple d'un point tel que la direction point – projection ne vérifie pas la complémentarité stricte. Cela se produit à un point où la projection n'est pas différentiable. (La fonction distance elle-même est différentiable en-dehors des points sur la frontière du convexe. 272

C.1 Dans cet exemple particulier, la solution (unique) (Δ, β) du problème détaillé dans l'exemple C.0.5 donne $\lambda^* = 6$. Observons que lorsque l'on dilate Z_y par λ^*, on a $Z_x \subseteq \overline{y} + \lambda^* Y[-1, +1]^2$, mais en dilatant par tout $\lambda < \lambda^*$, l'inclusion n'est pas vérifiée (le point en haut de l'aire verte n'est pas contenu dans la l'aire violette). 275

D.1 Gauche : $[X\ Y][-1, +1]^4$, sommets en bleu et autres points en noir (chaque point est deux vecteurs de signes). Droite : $[X\ -Y][-1, +1]^4$, sommets en bleu et autres points en noir (chaque point est deux vecteurs de signes). Schématiquement, le bleu clair correspond aux zonotopes avec $[X\ Y]$ et $[X\ -Y]$ et le noir à $\partial\theta(x)$. 287

D.2 Illustration de l'aspect zonotope pour une situation présentant plusieurs difficultés. À gauche, le zonotope turquoise en haut à gauche est Z_x, tandis que Z_y est en magenta en bas à droite. Le violet clair représente la version dilatée (par λ^*) de Z_y. À droite, le zonotope bleu correspond à $\partial\theta(x)$ (voir (D.8)), les trois autres composantes nulles n'étant pas représentées. On observe que parmi les huit vecteurs de signes de $\partial_B H(x)$, seuls quatre forment l'enveloppe convexe du C-différentiel après multiplication par H. De plus, les "voisins" dans la figure ne correspondent pas à des vecteurs de signes adjacents. Enfin, comme décrit dans un exemple plus simple, la méthode de l'annexe C renvoie une valeur de $\mathcal{E}^{0+}(x)$ correspondant au point gauche de la zone turquoise (la dilatation du point inférieur de la frontière de Z_y, avec $\overline{\zeta}$), qui correspond à g (la projection est le point plus haut avec ζ^*) qui est le point rouge dans l'image de droite et est hors de $\partial\theta(x)$; cela provient du fait que les signes choisis ne sont pas dans $\mathcal{S}(V, 0)$. 290

D.3 Illustration du contre-exemple. À gauche : les zonotopes (turquoise pour Z_x, magenta pour Z_y), la flèche représente g pour $\eta = -1$, qui n'appartient pas à $\partial\theta(x)$. À droite : $\partial\theta(x)$ et les éléments de $\partial_B H(x)^\mathsf{T} H(x)$; g_1 est le centre du différentiel. 292

D.4 Illustration du contre-exemple. À gauche : les zonotopes correspondants (turquoise pour Z_x, magenta pour Z_y), les flèches représentent g pour $\eta = -1$ (hors de $\partial\theta(x)$) et $\eta = +1$ (dans $\partial\theta(x)$). À droite : illustration de $\partial\theta(x)$ (dans le plan $x_2 = 4$). Comme dans le contre-exemple D.2.3, tous les vecteurs de signes ne sont pas extrémaux ($(+,-,+,+)$ et $(-,+,+,-)$ correspondent au point bleu central) et selon η, g peut appartenir ou non à $\partial\theta(x)$. Les 6 autres vecteurs de signes correspondent à la face dans le plan $x_2 = 6$. 296

D.5 Illustration des dégénérescences, obtenues en ajoutant artificiellement une dimension (la troisième dimension est omise à droite, dans le plan $x_1 = -1$). À gauche : les zonotopes décrits par les équations précédentes. À droite : les gradients g obtenus ; le carré bleu représente le différentiel de θ, ses sommets correspondent à des vecteurs de signes non "adjacents" (dû ici à la colinéarité des vecteurs de X et Y). Le segment orange représente les g possibles selon $\mathcal{E}^{0+}(x)$, les points rouges ses sommets. Les sommets du différentiel (bleu) correspondent au zonotope exprimé dans (D.7). 298

D.6 Illustration des quantités dilatées, avec les données du contre-exemple D.2.3. 301

D.7 À gauche : Z_y en magenta, Z_x en bleu-vert, et le sous-ensemble de Z_x en vert correspondant aux $\gamma_{\mathcal{E}^{0+}(x)}(\eta)$ tels que les $\gamma_{\mathcal{E}^{-}(x)}(\zeta)$ obtenus après projection donnent $g \in \partial\theta(x)$. À droite : valeurs correspondantes dans $[-1,+1]^2$ (c'est-à-dire avec η). Noter que le point le plus haut de la figure à gauche correspond à $\eta = (1,-1)$ et le plus à gauche à $\eta = (-1,-1)$, ce qui explique le changement d'orientation (pour retrouver une forme similaire au graphique de gauche, effectuer une rotation de $+3\pi/4$ dans le sens contre-horaire puis une symétrie axiale verticale). 306

D.8 En magenta, Z_y et en bleu-vert, Z_x. À gauche : $\partial\theta(x)$ en bleu, l'intersection avec Z_x en vert foncé correspond aux η avec $g(\eta, \zeta = 0) \in \partial\theta(x)$. À droite : valeurs de $\theta'(x, -g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^{-}(x)}))$ pour les η extrémaux : tout g avec $\zeta = 0$ est une direction de descente. 307

D.9 En magenta, Z_y et en bleu-vert, Z_x. À gauche : $\partial\theta(x)$ en bleu, l'intersection avec Z_x en vert foncé correspond aux η avec $g(\eta, \zeta = 0) \in \partial\theta(x)$. À droite : valeurs de $\theta'(x, -g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^{-}(x)}))$ pour les η extrémaux : pour certains η spécifiques, $-g$ est une direction de montée. 308

Liste des tableaux

2.1 Résumé des propriétés de régularité pour le PCL. 31

3.1 Description of the test-problems and comparison of the "original RC algorithm in [208]", written in Python, and the "simulated RC algorithm 3.5.5", written in Matlab : "(n, p, r, ς)" are the features of the problem ($V \in \mathbb{R}^{n\times p}$ is of rank r and has ς circuits, this last number being known to be bounded by $\varsigma_{\max}$), "$|\partial_B H(x)|$" is the cardinality of the B-differential of H given by (3.3), "Schläfli's bound" is the right-hand side of (3.39), "Original RC" gives the number of linear optimization problems (LOPs) solved by the original piece of software in Python of Rada and Černý [208], "Simulated RC" gives the number of LOPs solved by the implementation in the Matlab code ISF of the Rada and Černý algorithm (see algorithm 3.5.5), "Difference" is the difference between the two previous columns. Note (1) : computer crash after several weeks of computation. 100

3.2 Evaluation of the efficiency of the solvers by the number of LOPs they solve : A (taking the rank of V into account), B (special handling of the case where $v_{k+1}^{\mathsf{T}} d \simeq 0$), C (changing the order of the vectors v_i's by taking i_{k+1} by (3.49)), D_1 (pre-computation of $2(p-r)$ stem vectors after the QR factorization), D_2 (D_1 and 2 additional stem vectors computed after solving a LOP, whose optimal value is nonnegative), D_3 (all the stem vectors are first computed and, for $(s, \pm 1) \in \mathcal{S}_{k+1}$, a LOP is solved to get a handle d), D_4 (all the stem vectors are first computed and no LOP is solved). The "Ratio" (acceleration ratio) columns give for each considered problem the ratio (*LOPs of the considered ISF version*)/(*LOPs of simulated RC*). Note (1) : interruption of the run after several days of computation. The Mean/Median rows give the mean and medan values of the ratios. 101

3.3 Evaluation of the efficiency of the solvers by their computing times. The "Ratio" (acceleration ratio) columns give for each considered problem the ratio (*Time of the considered ISF version*)/(*Time of simulated RC*). Note (1) : interruption of the run after several days of computation. The Mean/Median rows give the mean and medan values of the ratios. 104

5.1 Corresponding lines in algorithms 5.6.5 and 5.6.8. 193

5.2 Cardinality formulas for some instances, when $p > n$ and $\mathrm{rank}(V) = n$. . 196

5.3 Description of the 33 considered arrangements. The first column gives the problem names. The next two columns specify the dimensions of $V \in \mathbb{R}^{n\times p}$. The 4th column gives the upper bound on the number of circuits of V, recalled in remark 5.3.13(6); by remark 5.3.13(3), it is also an upper bound on $|\mathfrak{S}_s|/2 + |\mathfrak{S}_a|$, where $|\mathfrak{S}_s|$ (resp. $|\mathfrak{S}_a|$) is the number of symmetric (resp. asymmetric) stem vectors (definition 5.3.12) of the arrangement $\mathcal{A}(V,\tau)$; $|\mathfrak{S}_s|/2$ and $|\mathfrak{S}_a|$ are given in columns 5 and 6. Columns 7 and 8 give half the number of stem vectors of the arrangement $\mathcal{A}([V;\tau^{\mathsf{T}}],0)$ and its Schläfli upper bound, derived from (5.28). The last two columns give the number $|\mathcal{S}(V,\tau)|$ of chambers of the arrangement $\mathcal{A}(V,\tau)$ and its upper bound given by (5.30). 197

5.4 Computing times (in seconds) for the *standard* algorithms listed in section 5.7.2. For each algorithm A := P, PD or D, the second column gives the ratios time(RC)/time(A) 203

5.5 Computing times (in seconds) for the *compact* algorithms listed in section 5.7.2. For each algorithm A = RC, P, PD, or D, the first column gives the computing time of A/C in seconds, the second column gives the ratios time(A)/time(A/C) (upper bounded by 2, approximately) and the third column gives the ratios time(RC)/time(A/C). 204

A.1 Cardinalités connues pour certaines instances. Les valeurs des problèmes RAND-N-P et 2D-N-P sont obtenues par position générale affine, donc les propositions 5.3.31, 3.4.6 et la remarque 5.3.13 6). Rappelons que les vecteurs souches symétriques sont comptés par paires (d'où le facteur $1/2$) – le nombre de circuits *n'a pas* à considérer ce facteur. 242

A.2 Nombre de sous-problèmes résolus selon l'état du nœud courant. De plus, on note $\pm s \in \mathcal{S}_k \iff \boxed{\mathbb{S}} = 0$ et $(-)s \in \mathcal{S}_k \iff \boxed{\mathbb{S}} \neq 0$. 249

A.3 Proportions approximatives de chambres symétriques. Les instances 2D ont une proportion particulièrement faible de chambres symétriques. 252

A.4 Valeurs pertinentes pour les instances affines. Les colonnes 2 et 3 représentent les cardinalités des ensembles de vecteurs de signes, les colonnes 4-5-6 les nombres de problèmes réalisables résolus. La colonne 7 est la différence des deux précédentes. La dernière colonne représente le second terme du membre de droite de (A.4). Ce tableau est illustré ci-dessous dans la figure A.2. Les * représentent les irrégularités mentionnées précédemment (pas de position générale parfaite dans le sous-arrangement). 253

A.5 Temps de calcul en noir et ratios temps(RC) / temps(A) en bleu pour les instances linéaires et les différents algorithmes; les meilleurs sont en gras. Pour l'algorithme D, les vecteurs souches et les tests de couverture sont calculés de façon un peu plus efficace comme décrit en sections A.4.1 et A.4.2. 255

A.6 Temps de calcul des vecteurs souches dans les variantes régulières. Les deuxième et quatrième colonnes représentent les nombres de vecteurs souches, les troisième et cinquième colonnes le nombre de doublons. Les trois colonnes restantes indiquent le temps du calcul initial, le temps de calcul avec la forme échelonnée et leur ratio : si supérieur à 1, cela signifie que la forme échelonnée est plus rapide. 257

A.7 Temps de calcul des vecteurs souches dans les variantes compactes. Les deuxième et quatrième colonnes représentent les nombres de vecteurs souches, les troisième et cinquième colonnes les doublons. Les trois colonnes restantes indiquent le temps du calcul initial, le temps de calcul avec la forme échelonnée et leur ratio : si supérieur à 1, cela signifie que la forme échelonnée est plus rapide. 258
A.8 Illustration de l'implémentation récursive du test de couverture. Il y a $p = 5$ vecteurs dans $\mathbb{R}^n$, et la matrice des vecteurs souches est donnée à droite sous forme transposée dans la moitié droite. Par exemple, la première colonne signifie que $[v_1\ v_2\ -v_3]$ est de nullité un dans $\mathbb{R}_+^{\{1,2,3\}}$. À gauche, sur la ligne avec index = 1 et signe = +, le vecteur courant est $+M_{:,1}$ (la première ligne de la matrice transposée des vecteurs souches). Sur la ligne suivante, comme le signe est aussi $+$, la deuxième ligne des matrices de vecteurs souches est ajoutée. Sur la ligne avec index = 3 et signe = -, le vecteur courant est donc la première ligne de $\mathfrak{S}$ plus la deuxième moins la troisième. En particulier, la coordonnée 1 du vecteur courant est égale à 3, qui est la taille du premier vecteur souche (première colonne de $\mathfrak{S}$), donc le test de couverture arrête la récursion. 259
A.9 Temps d'exécution pour les instances linéaires avec l'option D3. Les colonnes 2-3 représentent les temps totaux et de couverture avec le produit matrice-vecteur complet dans le test. Les colonnes 4-5 représentent les temps totaux et de couverture effectués de manière récursive. La colonne 6 donne le nombre de vecteurs souches. Les colonnes 7-8 montrent les ratios pour les temps totaux et les temps de couverture. 260
A.10 Temps d'exécution pour les instances linéaires avec l'option D4. Les colonnes 2-3 représentent les temps totaux et de couverture avec le produit matrice-vecteur complet dans le test. Les colonnes 4-5 représentent les temps totaux et de couverture effectués de manière récursive. La colonne 6 donne le nombre de vecteurs souches. Les colonnes 7-8 montrent les ratios pour les temps totaux et les temps de couverture. 261

Chapitre 1

Introduction

1.1 Point de départ et motivations

Notre but principal est l'étude de *problèmes de complémentarité* (PC) sous le prisme algorithmique. Souvent étudiés en dimension finie, bien que des extensions existent en dimension infinie, ils consistent en un nombre d'(in)égalités et de relations de complémentarité qui peuvent s'écrire sous la forme suivante. Pour deux fonctions vectorielles $F, G : \mathbb{R}^n \to \mathbb{R}^n$, on cherche un $x \in \mathbb{R}^n$ tel que

$$F(x) \geqslant 0, \qquad G(x) \geqslant 0, \qquad F(x)^\mathsf{T} G(x) = 0, \tag{1.1}$$

où les inégalités se lisent composante par composante et $F(x)^\mathsf{T} G(x)$ est le produit scalaire euclidien entre les vecteurs $F(x)$ et $G(x)$. Une expression compacte est donnée par

$$0 \leqslant F(x) \perp G(x) \geqslant 0,$$

où $\perp$ signifie l'orthogonalité. En particulier, c'est équivalent à demander que pour chaque indice $i \in [1 : n] := \{1, \dots, n\}$, $F_i(x) \geqslant 0$, $G_i(x) \geqslant 0$ et $F_i(x)G_i(x) = 0$. Cette forme relativement générale fait des problèmes de complémentarité un outil pratique pour traiter un vaste panorama de situations, aussi bien les conditions d'optimalité de problèmes sous contraintes d'inégalité que des phénomènes physiques. Très souvent, la fonction G est l'identité, ce qui donne :

$$\mathrm{PCN}(F) \qquad 0 \leqslant x \perp F(x) \geqslant 0. \tag{1.2}$$

En outre, quand F est affine, disons $F(x) = Mx + q$, on a :

$$\mathrm{PCL}(M, q) \qquad 0 \leqslant x \perp Mx + q \geqslant 0, \tag{1.3}$$

qui est appelé problème de complémentarité linéaire [58]. Une part conséquente de la littérature se concentre sur ce cas linéaire, la part belle étant faite aux propriétés de la matrice M [181, 86, 48].

À partir du travail initial de Cottle dans sa thèse en 1964 [56, 57], la littérature sur les problèmes de complémentarité s'est beaucoup développée. Il existe moult méthodes pour

traiter les PC, auxquelles nombre d'auteur·rice·s, aussi bien des domaines théoriques que pratiques, ont contribué. Voici quelques domaines où la complémentarité apparaît, parfois après une discrétisation de l'espace pour se ramener en dimension finie, et des articles associés : problèmes de contact [2, 9, 62, 129, 128, 91, 261], simulations de fluides multiphases [18, 20, 36, 38, 163, 164, 249], EDP avec contraintes de complémentarité [21, 122], où des contraintes d'(in)égalité interviennent en plus de la complémentarité, imagerie [84], finance [100] ; voir aussi leurs références. En particulier, les états de l'art de Harker et Pang [120], Pang [193] puis Ferris et Pang [91] contiennent davantage de références, détails sur des applications de PC comme les équilibres de traffic routier et la théorie des jeux par exemple. Récemment, des PC en "algèbre tropicale" ont été étudiés [8].

1.2 Problèmes reliés

Avant d'évoquer des algorithmes résolvant les PC sous les reformulations discutées, nous mentionnons d'autres formes plus ou moins proches. Le PCL *vertical* [54, 91, 55, 58] consiste en la complémentarité entre plusieurs fonctions affines, contrairement à (1.3) qui a deux fonctions affines dont l'identité. Le PCL *horizontal* s'exprime comme

$$Ax + By = c, \qquad x \geqslant 0, \qquad y \geqslant 0, \qquad x^\mathsf{T} y = 0. \tag{1.4}$$

Dans [55], une version étendue avec $Ax + By \in K$ pour un convexe polyhédrique K est proposée. Le PC généralisé a la forme suivante

$$F(x) \in K, \qquad G(x) \in K^*, \qquad \langle F(x), G(x) \rangle = 0, \tag{1.5}$$

pour un cône convexe fermé K et son dual pour le produit scalaire euclidien $\langle \cdot, \cdot \rangle$ K^* [120, définition 2.3 p. 166]. Choisir $K = \mathbb{R}^n_+$ réduit le PC généralisé au PC usuel. C'est également le cas d'autres problèmes, telle que l'optimisation sous contraintes d'inégalité, lorsque l'on regarde les conditions d'optimalité. Considérons

$$\min \frac{1}{2} x^\mathsf{T} M x + c^\mathsf{T} x, \qquad \text{t.q.} \qquad Ax \leqslant b,$$

avec M symétrique. En effet, si λ est un multiplicateur positif, les conditions d'optimalité deviennent

$$Mx + c + A^\mathsf{T} \lambda = 0, \qquad 0 \leqslant \lambda \perp (b - Ax) \geqslant 0,$$

où la complémentarité s'applique pour une partie du système (le nombre de contraintes). De tels systèmes sont des problèmes de complémentarité *mixtes* (PCM), où une partie des équations sont des égalités et l'autre des conditions de complémentarité, s'écrivant par exemple [86, §9.4.2]

$$G(u, v) = 0, \qquad 0 \leqslant v \perp H(u, v) \geqslant 0, \tag{1.6}$$

où $G : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ et $H : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$.

Un autre cadre répandu est celui des inégalités variationnelles, le lien ayant été découvert par Karamardian [139]. Dans une forme générale, elles s'expriment comme

$$x \in C, \qquad \langle F(x), (y - x) \rangle \geqslant 0, \forall\, y \in C, \tag{1.7}$$

où $F : \mathbb{R}^n \to \mathbb{R}^n$ et C est un ensemble convexe fermé de $\mathbb{R}^n$. On les trouve également sous la forme équivalente suivante

$$0 \in F(x) + N_C(x), \tag{1.8}$$

où $N_C(x)$ est le cône normal à C en x [2, 47, 86, 98, 131]. En particulier, pour $C = \mathbb{R}^n_+$, on retrouve $\mathrm{PCN}(F)$. La situation spécifique de contraintes de bornes, i.e., $C = \{x \in \mathbb{R}^n : l \leqslant x \leqslant u\}$ pour des bornes inférieures $l \in (\{-\infty\} \cup \mathbb{R})^n$ et supérieures $u \in (\mathbb{R} \cup \{+\infty\})^n$, est traitée dans [47, 90, 179]. Pour le cas de la dimension infinie, voir [247].

1.3 Formulations équivalentes et algorithmes

Dans le cas le plus général, résoudre un PC est un problème NP-complet [49]. Même pour des types de matrices M précis, le problème peut rester difficile [60].

Nous démarrons par mentionner l'algorithme de Lemke [152] pour les PCL, qui rappelle l'algorithme du simplexe en optimisation linéaire. En supposant une hypothèse classique sur M, l'algorithme change la complémentarité des indices, i.e., choisit $x_i = 0$ ou $(Mx + q)_i = 0$. Sur certaines instances, tout comme la méthode du simplexe, un nombre exponentiel d'étapes peut être requis [180, 88].

Les méthodes de points intérieurs ont également été appliquées pour les PC, en relaxant la contrainte de complémentarité en $F_i(x)G_i(x) = \mu > 0$ pour tout $i \in [1 : n]$. Cela permet aux méthodes d'éviter la combinatoire inhérente de trouver, pour chaque indice, quelle quantité est nulle. Nous mentionnons quelques contributions : [145, 174, 144], ou [44] pour un point de vue un peu différent.

Les *équations généralisées*, introduites par Robinson, permettent également d'étudier les PC. Elles peuvent se voir comme des généralisations des (1.7) et (1.8), et ont par exemple la forme suivante :

$$0 \in F(x) + T(x)$$

où $F : \mathbb{R}^n \to \mathbb{R}^n$ est une fonction et $T : \mathbb{R}^n \multimap \mathbb{R}^n$ est une multifonction (une fonction de $\mathbb{R}^n$ vers l'ensemble des parties de $\mathbb{R}^n$). Avec $F(x) = Mx + q$ et $T(x) = N_{\mathbb{R}^n_+}(x)$ le cône normal de l'orthant positif, on retrouve le PCL [217, section 4]. Sur le sujet, voir aussi par exemple dans [220, 216].

Les méthodes de type activation de contraintes, en particulier la méthode "Primal-Dual Active Set" (PDAS) forment un autre aspect des algorithmes possibles (lorsque l'une des fonctions est l'identité). À un itéré, ces méthodes déterminent quels indices sont actifs ($x_i = 0$), puis traitent les différemment selon qu'ils indices soient associés à la partie x ou à la partie $F(x)$.

Cette méthode est employée dans [128] sur un cas provenant d'une application, dans [122] pour les inégalités variationnelles et dans [124] pour le cas de la dimension infinie, où cette méthode est en fait révélée être fonctionnellement identique à l'algorithme utilisant la C-fonction $\varphi_{\min}$ discutée ci-dessous.

Une technique assez répandue est la reformulation par les C-fonctions, qui est celle

principalement discutée dans cette thèse, voir la section 2.3.1. Brièvement, une C-fonction φ vérifie, pour a et b dans $\mathbb{R}$,

$$\varphi(a,b) = 0 \qquad \Longleftrightarrow \qquad a \geqslant 0, \qquad b \geqslant 0, \qquad ab = 0,$$

et permet de reformuler un PC en

$$\Phi(x) := \begin{pmatrix} \varphi(F_1(x), G_1(x)) \\ \vdots \\ \varphi(F_n(x), G_n(x)) \end{pmatrix} = 0, \tag{1.9}$$

ainsi qu'à la minimisation de la fonction dite de mérite $\Psi(x) := ||\Phi(x)||^2/2$, dont on cherche un zéro, un minimum local ou un point stationnaire selon la difficulté du problème.

Quelques articles importants sur les C-fonctions sont par exemple [159, 138, 94, 194, 92, 204, 99, 6]. Souvent, les systèmes résultants sont non différentiables, ce qui conduit à leur *lissage* :

$$\tilde{\Phi}(x,\mu) = \begin{pmatrix} \tilde{\varphi}(F_1(x), G_1(x), \mu) \\ \vdots \\ \tilde{\varphi}(F_n(x), G_n(x), \mu) \\ \mu \end{pmatrix} = 0,$$

où $\tilde{\varphi}(\cdot,\cdot,\mu)$ est différentiable quand $\mu > 0$ et $\tilde{\varphi}(a,b,0) = \varphi(a,b)$. Ce lissage du système est par exemple considéré dans [98, 45, 47, 86, 260, 157, 117, 190, 249].

Lorsque des C-fonctions sont utilisées, les PC sont reliés à la grande question de la résolution de systèmes (non lisses) d'équations. Sur ce sujet, voir par exemple [217, 216, 51, 195, 197, 205, 46, 127], parmi bien d'autres.

Les contributions citées ne représentent qu'un bref aperçu de la littérature ; leurs propres références contiennent des éléments supplémentaires. Un des buts de ce doctorat est l'étude de la C-fonction $\varphi_{\min}$. C'est à la fois une des plus simple, puisque linéaire par morceau en ses arguments (en F et G), mais est également la moins différentiable. Cet inconvénient semble avoir fait que $\varphi_{\min}$ a été moins étudiée (mais néanmoins très appréciée en pratique) que la fonction de Fischer φ_{FB} et sa progéniture.

1.4 Non-différentiabilité et géométrie combinatoire

Nous abordons maintenant brièvement le sujet des "arrangements d'hyperplans". Sa relation avec les sujets précédents provient d'un calcul d'un objet défini par la C-fonction $\varphi_{\min}$, comme discuté dans le chapitre 3. Bien que ce domaine puisse paraître inoffensif lorsque l'on vient de l'optimisation, c'est en réalité un domaine extrêmement varié et profond, un sujet tout à fait classique pour les spécialistes en algèbre, combinatoire et géométrie discrète.

Soit $n \geqslant 1$ un entier représentant la dimension, $v \in \mathbb{R}^n \backslash \{0\}$ et $t \in \mathbb{R}$. L'hyperplan $H_{v,t} := \{x \in \mathbb{R}^n : v^\mathsf{T} x = t\}$ scinde clairement $\mathbb{R}^n$ en lui-même et ses deux demi-espaces

ouverts associés :

$$H^+_{v,t} := \{x \in \mathbb{R}^n : v^\mathsf{T} x > t\} \qquad \text{et} \qquad H^-_{v,t} := \{x \in \mathbb{R}^n : v^\mathsf{T} x < t\}. \tag{1.10}$$

Maintenant, soit $V = [v_1 \ \cdots \ v_p] \in \mathbb{R}^{n\times p}$ et $\tau = (\tau_1, \ldots, \tau_p) \in \mathbb{R}^p$. On considère la collection d'hyperplans $\{H_i : i \in [1 : n]\}$ où $H_i := H_{v_i,\tau_i}$. Soit $\mathcal{A}(V, \tau)$ cette collection, appelée arrangement d'hyperplans. Le but est d'étudier la structure géométrique formée par les hyperplans.

Les arrangements peuvent être analysés sous de multiples points de vue ou pour des raisons variées, bien qu'une des questions qui revient souvent est le nombre de *chambres*, aussi appelées *régions* ou *cellules*. Plus précisément, soit $\{H_1, \ldots, H_p\}$ une collection de p hyperplans dans $\mathbb{R}^n$. De fait, $\mathbb{R}^n \backslash \cup_{i=1}^p H_i$ est décomposé en composantes connexes appelées chambres. Un example est donné en figure 1.1.

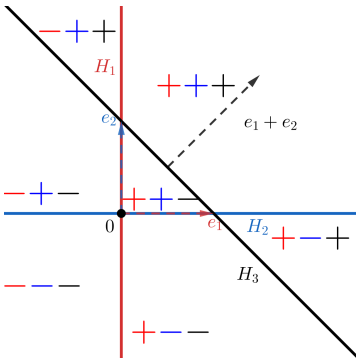


FIGURE 1.1 – Illustration avec trois hyperplans et sept chambres, avec les hyperplans $H_1 = \{x \in \mathbb{R}^2 : x_1 = 0\}$, $H_2 = \{x \in \mathbb{R}^2 : x_2 = 0\}$ et $H_3 = \{x \in \mathbb{R}^2 : x_1 + x_2 = 1\}$.

Chacune de ces régions est un sous-ensemble de H_i^+ ou H_i^-, pour chaque $i \in [1 : p]$, avec lequel on peut associer un signe $s_i \in \{\pm 1\}$. De fait, l'identification des chambres revient à trouver toutes les combinaisons dans $\{-1, +1\}^p$, appelées *vecteurs de signes*, qui correspondent à une intersection non vide. Compter les chambres est une question qui trouve sa source au moins au début du XIX^{ème} siècle, avec [239, 227, 215] ([227] est une œuvre posthume ; la partie sur les arrangements est entourée de moult autres contributions sur des sujets extrêmement variés). Notre but est d'identifier l'ensemble $\mathcal{S}$ des vecteurs de signes s décrits plus haut, et pas seulement leur nombre $|\mathcal{S}|$. Bien que ces deux questions puissent sembler proches – puisque identifier les chambres indique leur nombre et les compter peut être fait par énumération, notre compréhension est que ces deux questions sont en fait significativement différentes. De nombreuses contributions comptent le nombre de chambres sans les identifier, par l'usage d'outils combinatoires intrinsèques.

La découverte de ce domaine a été un aperçu de l'inimaginable – difficile de penser qu'un problème si simple donnait lieu à des choses si raffinées et encore plus de mentionner davantage que certaines contributions. Voici quelques articles et livres qui traitent de la question du comptage (puisque bien plus fréquent) et de l'énumération des chambres d'un arrangement.

De nombreux livres sur la combinatoire ou même les arrangements précisément vont bien au-delà du cadre et des notions intervenant dans cette thèse, par exemple les travaux

de Crapo et Rota [61], Orlik et Terao [187], Stanley [237, 238] et son cours du MIT [236]. Edelsbrunner a écrit un livre spécifiquement sur les algorithmes [81]. Plus récemment, voir la monographie d'Aguiar et Mahajan ainsi que l'état de l'art [118] de Halperin et Sharir.

Parmi les nombreuses contributions discutant du calcul du *nombre* de chambres (mais pas les chambres elles-mêmes), mentionnons le travail pionnier de Zaslavsky [257], via l'utilisation du "polynôme caractéristique". D'autres utilisations en sont faites dans [12, 238] par exemple.

Puisque l'on se concentre sur l'obtention des chambres d'arrangements, la plupart des logiciels du domaine de la combinatoire sont bien plus généraux, capables de traiter une foule de problèmes au-delà des arrangements. Nous en citons quelques-uns : Sagemath [68], Macaulay2 [111], OSCAR [67, 189], TOPCOM [214], polymake [140], et pour finir Counting_Chambers [35] (qui, malgré son nom, calcule en fait davantage que les chambres).

Pour terminer, nous citons certains algorithmes spécifiques. Bieri et Nef, dans [27], calculent l'arrangement complet, et pas uniquement les chambres de dimension n. De même, Edelsbrunner, O'Rourke et Seidel [83] conçoivent un algorithme avec une complexité théorique optimale. Pour se concentrer sur les chambres, notre intérêt principal, mentionnons Avis, Fukuda et Sleumer [14, 232]. Un meilleur algorithme est conçu par Rada et Černý dans [208], qui est celui que l'on améliore et développe dans cette thèse.

1.5 Plan et contributions

Cette thèse s'organise comme suit. Le chapitre 2 est une introduction détaillée, avec une présentation des notations, une discussion plus profonde de la littérature et une explication des buts de la thèse. D'abord, les problèmes de complémentarité (PC) sont présentés, principalement sous l'angle de reformulations et d'algorithmes non lisses. L'introduction termine par davantage d'éléments sur les arrangements.

La première partie, chapitres 3 et 4, parle du calcul du B-différentiel du minimum de deux fonctions affines et d'arrangements centrés. En effet, le calcul de ce B-différentiel, i.e., une question d'analyse non lisse provenant du contexte d'un algorithme particulier pour les PC, se résout en fait en identifiant les chambres d'un arrangement centré, i.e., dont tous les hyperplans ont un point en commun. Ces chapitres présentent de nombreuses propriétés des arrangements et d'autres problèmes équivalents, ainsi que des algorithmes identifiant les chambres. La seconde partie, chapitre 5, se concentre sur le cas où les hyperplans ne s'intersectent pas nécessairement tous simultanément, ce qui implique d'adapter les notions et algorithmes du chapitre 3.

Les deux parties présentent principalement des améliorations et variantes algorithmiques sur un algorithme de l'état de l'art, conçu par Rada et Černý [208], pour réduire les temps de calcul. Ces améliorations se basent sur des observations analytiques ou heuristiques afin d'élaguer l'arbre.

Une autre méthode, assez différente, dite "duale", est aussi proposée. En particulier, elle utilise les circuits du matroïde (orienté) sous-jacent [191]. Nous n'avons pas trouvé (du

moins algorithmiquement) de méthode reliant les circuits et les vecteurs de signes qui *ne* correspondent *pas* aux chambres. Au-delà de la découverte conceptuelle, cette apparition surprenante de la dualité conduit à des améliorations significatives de l'algorithme de Rada et Černý, surtout sur des instances avec une nature combinatoire provenant d'applications. Cette contribution [77] est publiée dans Mathematical Programming Computation ; le code Matlab correspondant et sa documentation sont accessibles en ligne [75, 76].

Le chapitre 4 présente des détails et compléments supplémentaires sur le sujet précédent : le cas de fonctions non linéaires et le rôle de leurs linéarisations, le B-différentiel de la fonction de mérite ainsi que des détails sur les instances testées.

Bien que le sujet précédent donne lieu à des arrangements centrés, il nous semblait important de savoir si les techniques introduites pour les arrangements centrés s'adaptaient au cas général d'arrangements affines (non centrés). C'est le sujet du chapitre 5, en préparation (initialement soumis à SIAM Journal on Discrete Mathematics [79]).

La dernière partie, le chapitre 6, discute de la globalisation d'une variante de l'algorithme de Newton-min, une méthode qui utilise la C-fonction minimum $\varphi_{\min}$ et résout un système linéaire à chaque itération, pour résoudre les PC [72]. Dans cette contribution, la direction de recherche est obtenue en trouvant un point dans un certain polyèdre, qui est non vide grâce à des hypothèses de régularité devant être valide partout autour des points considérés. Notre but dans ce manuscrit est de s'affranchir de telles conditions en utilisant la méthode de Levenberg-Marquardt, qui, en échange d'un vraisemblablement coût plus élevé, garantit l'existence d'une solution.

Cependant, sans hypothèse de ce type, la question suivante se pose : dans la structure par morceau, "Quelle morceau choisir ?". Dit autrement, en un itéré d'une méthode type Newton, "Quel élément du différentiel choisir ?". En particulier, on montre une relation entre la détection de la stationnarité "forte" (Dini) en un itéré (pas de meilleur point autour) et le choix du "morceau", ce qui caractérise que cette détection est un problème, en général, co-NP-complet. Nous discutons de plusieurs éléments autour de cette question, comme la globalisation par Levenberg- Marquardt d'une telle méthode. Ces contributions forment un projet d'article.

Chapitre 2

Cadre général

Ce chapitre a pour objectif de discuter des buts de cette thèse et de sa position par rapport à la littérature. Les deux problèmes en jeu sont les problèmes de complémentarité et les arrangements d'hyperplans. À vrai dire, les deux domaines dans lesquels ces problèmes apparaissent sont plutôt éloignés l'un de l'autre. Par conséquent, les deux parties principales de la thèse ne sont pas si étroitement liées.

Initialement, cette thèse visait à concevoir une méthode globalement convergente pour les problèmes de complémentarité basée sur la C-fonction minimum $\varphi_{\min}$. Notamment, cette reformulation est fortement non différentiable. Ainsi, à moins de faire certaines hypothèses de régularité fortes, la non-différentiabilité entrave considérablement la globalisation d'algorithmes locaux tels que l'algorithme de Newton-min. Dans [72], les auteurs modifient cette méthode locale standard en considérant des systèmes polyédraux au lieu de systèmes linéaires à chaque itération. La motivation initiale était d'utiliser une technique de Levenberg-Marquardt sur ces systèmes pour éviter des hypothèses de régularité malgré des itérations plus coûteuses.

Ce sujet est présenté dans le chapitre 6. Cependant, l'approche Levenberg-Marquardt était en quelque sorte liée à la question de choisir un élément *approprié* dans un différentiel afin d'obtenir une propriété de descente sur la fonction de mérite. Cette question nous a conduits à essayer de mieux comprendre le différentiel impliqué, c'est-à-dire le différentiel du minimum (composante par composante) de deux fonctions. Il s'est avéré que lorsque les fonctions impliquées sont affines, calculer ce différentiel équivaut à identifier les chambres d'un certain arrangement d'hyperplans, où tous les plans s'intersectent.

C'est l'objet du chapitre 3, où d'autres problèmes équivalents, propriétés et algorithmes pour résoudre cette question sont discutés. En particulier, une nouvelle méthode basée sur la dualité et les circuits de matroïdes est conçue. Des compléments et détails sont proposés dans le chapitre 4. Le chapitre 5 complète la discussion précédente en étudiant le cas des arrangements avec des hyperplans affines, et étend le travail des chapitres précédents. L'annexe A discute de divers compléments sur les arrangements.

Enfin, le chapitre 6 aborde la motivation initiale des problèmes de complémentarité (PC) et de la globalisation. En particulier, nous verrons une certaine forme de "symbiose" avec le

sujet des autres chapitres : les arrangements ont été considérés car ils représentent la face cachée d'un problème de différentiabilité issu des PC, et ont été utiles pour développer de nouvelles perspectives sur la résolution algorithmique des PC. Les annexes B, C, D discutent de précisions techniques complémentant le chapitre 6.

Après avoir précisé les notations utilisées dans les parties suivantes, ce chapitre présente diverses notions introductives et discute plus en détail certains des choix qui ont conduit à ce travail.

2.1 Notations

La première partie, section 2.1.1, présente des notations relativement communes, tandis que d'autres moins fréquentes sont introduites à la section 2.1.2.

2.1.1 Notations générales

- $\mathbb{N}$ et $\mathbb{N}^*$ représentent les entiers positifs et strictement positifs.
- $[1:n] := \{1, \dots, n\}$ sont les n premiers entiers strictement positifs ; $[n_1 : n_2] := \{n_1, \dots, n_2\}$ pour $n_1 \leqslant n_2$ dans $\mathbb{N}$.
- $\mathbb{R}$ et $\mathbb{R}_+$ sont les réels et réels positifs.
- $\mathbb{R}^n$, $\mathbb{R}^n_+$, et $\mathbb{R}^n_{++}$ sont respectivement l'espace réel de dimension n, les orthants positifs et strictement positifs dans $\mathbb{R}^n$; $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x \geqslant 0\}$ où les inégalités se lisent composante par composante ; $\mathbb{R}^n_{++} = \{x \in \mathbb{R}^n : x > 0\}$.
- $t_+ := \max(t, 0)$ pour $t \in \mathbb{R}$ (aussi nommé "ReLU") ; t^2_+ est une notation compacte pour $(t_+)^2$ (puisque $(t^2)_+ = t^2$) ; pour des vecteurs $v \in \mathbb{R}^n$, $v = v_+ - v_-$ avec $(v_+)_i = \max(v_i, 0)$ et $(v_-)_i = \max(-v_i, 0) = -\min(v_i, 0)$.
- sgn est la fonction $\mathbb{R} \to \mathbb{R}$ définie par $\mathrm{sgn}(t) = -1$ si $t < 0$, $\mathrm{sgn}(t) = +1$ si $t > 0$ et $\mathrm{sgn}(0) = 0$; $\mathrm{sgn}(v) := (\mathrm{sgn}(v_i))_{i \in [1:n]}$ pour un $v \in \mathbb{R}^n$.
- $u \cdot v$ est le produit de Hadamard de deux vecteurs u et v de $\mathbb{R}^n$ défini par $(u \cdot v)_i = u_i v_i$ pour $i \in [1 : n]$.
- $|v| := (|v_i|)_{i \in [1:n]} = \mathrm{sgn}(v) \cdot v$ est la valeur absolue composante par composante d'un vecteur $v \in \mathbb{R}^n$.
- e est le vecteur de composantes égales à 1, dont la taille se déduit du contexte ; en particulier, $e \cdot u = u$ pour $u \in \mathbb{R}^n$; $\{e_1, \dots, e_n\}$ est la base canonique de $\mathbb{R}^n$.
- $||\cdot|| := ||\cdot||_2$ est la norme 2 de $\mathbb{R}^n$: $||x||_2^2 = \sum_{i=1}^n x_i^2$; $||\cdot||_1$ est la norme 1 : $||x||_1 := \sum_{i=1}^n |x_i|$; $||\cdot||_\infty$ est la norme infinie : $||x||_\infty := \max_i\{|x_i|\}$.
- $\mathrm{supp}(v) := \{i \in [1 : n] : v_i \neq 0\}$ pour un vecteur $v \in \mathbb{R}^n$; le support peut aussi être défini pour un ensemble d'indices (pas nécessairement identifié avec $[1 : n]$ pour un $n \in \mathbb{N}^*$).
- I est la matrice identité dont la taille est claire par le contexte. Éventuellement, on pourra l'appeler Id.

- $A_{I,J}$ est, pour une matrice $A \in \mathbb{R}^{m\times n}$, des sous-ensembles $I \subseteq [1:m]$ et $J \subseteq [1:n]$, la sous-matrice $(A_{ij})_{i\in I, j\in J}$; ":" signifie aucune sélection ($I = [1:m]$ ou $J = [1:n]$), de fait la i-ème ligne de A est $A_{i,:}$ et sa j-ème colonne est $A_{:,j}$.
- $\mathcal{N}(A)$, $\mathcal{R}(A)$ sont le noyau et l'image de $A \in \mathbb{R}^{m\times n}$; $\mathrm{rank}(A) := \dim \mathcal{R}(A)$ est son rang, $\mathrm{null}(A) := \dim \mathcal{N}(A)$ sa nullité ; par le théorème du rang, $\mathrm{rank}(A) + \mathrm{null}(A) = n$.
- $\cdot^\mathsf{T}$ sert à indiquer la transposée d'un vecteur ou d'une matrice.
- $[A; B]$ est la concaténation verticale de matrices $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{m'\times n}$ et appartient à $\mathbb{R}^{(m+m')\times n}$; $[A\ B] := [A^\mathsf{T}; B^\mathsf{T}]^\mathsf{T}$ est la concaténation horizontale de matrices $A \in \mathbb{R}^{m\times n}$ et $B \in \mathbb{R}^{m\times n'}$, appartenent à $\mathbb{R}^{m\times(n+n')}$; cela s'applique aussi aux vecteurs.
- $\det(A)$ est le déterminant de A quand A est carrée.
- $\mathrm{sp}(A)$ est le spectre de A, l'ensemble des valeurs propres de A quand A est carrée.
- $A \in \mathbb{R}^{n\times n}$ est dite (semi)définie positive si pour tout $x \in \mathbb{R}^n$, $x^\mathsf{T}Ax > 0$ ($\geqslant 0$).
- $\mathrm{Diag}(v)$ pour $v \in \mathbb{R}^n$ est la matrice diagonale de diagonale égale à v.
- $\mathcal{S}^n, \mathcal{S}^n_+, \mathcal{S}^n_{++}$: ensembles des matrices symétriques, symétrique semi-définies positives et symétriques définies positives.
- $|S|$ est la cardinalité d'un ensemble S.
- $\cup$ est le symbole de l'union disjointe.
- S^c est le complémentaire de l'ensemble S dans un ensemble (plus grand) clair dans le contexte.
- S^J est l'ensemble des vecteurs dont les éléments sont dans S et indexés les éléments de J ; dit autrement, l'ensemble des fonctions de J vers S.
- 2^S est l'ensemble des parties/sous-ensembles de S (dont S et $\varnothing$) ; dit autrement, c'est l'ensemble des fonctions de S vers $\{0, 1\}$.
- $\mathrm{vect}(S)$ est l'espace vectoriel engendré par un sous-ensemble S d'un espace vectoriel.
- $V^\perp := \{x \in \mathbb{R}^n : x^\mathsf{T}v = 0, \forall\, v \in V\}$ est l'orthogonal du sous-espace $V \subseteq \mathbb{R}^n$.
- $\mathrm{conv}(S)$ représente l'enveloppe convexe du sous-ensemble S d'un espace vectoriel.
- $\mathrm{P}_C(x)$ est la projection orthogonal de x sur le convexe fermé C, définie par $\mathrm{P}_C(x) = \mathrm{argmin}_{y\in C}||y - x||^2/2$.
- $C^* := \{v \in \mathbb{R}^n : \langle v, c\rangle \geqslant 0, \forall\, c \in C\}$ pour un sous-ensemble $C \subseteq \mathbb{R}^n$ représente le cône dual de C pour le produit scalaire euclidien.
- $N_C(x) := \{v : \forall\, y \in C, \langle v, y - x\rangle \leqslant 0\}$ pour un convexe fermé C est le cône normal en x à C. Le cône tangent est le cône dual du cône normal, défini par $T_C(x) := N_C(x)^*$.
- $F'(x)$ est la dérivée d'une fonction régulière $F : \mathbb{R}^n \to \mathbb{R}^m$ en x.
- $F'_I(x)$ pour une fonction différentiable $F : \mathbb{R}^n \to \mathbb{R}^m$ est à comprendre comme $(F_I)'(x) \in \mathbb{R}^{I\times n}$ pour $I \subseteq [1:m]$.
- $\nabla f(x) \in \mathbb{R}^n$ est le gradient d'une fonction différentiable $f : \mathbb{R}^n \to \mathbb{R}$ pour le produit scalaire euclidien.
- $q(t) = o(t)$ désigne une quantité telle que $\lim_{t\neq 0, t\to 0} q(t)/t = 0$; $q(t) = O(t)$ désigne une quantité telle que $q(t)/t$ est bornée pour $t \neq 0$.

2.1.2 Notations spécifiques

Ici, on présente quelques notations un peu plus restreintes.

- $\mathfrak{B}(S)$ est l'ensemble des bipartitions de S, i.e., toutes les paires différentes de sous-ensembles I et J tels que $I \cap J = \varnothing$, $I \cup J = S$, en considérant (I, J) et (J, I) comme différentes.
- $H_{v,t} := \{x \in \mathbb{R}^n : v^\mathsf{T} x = t\}$ est l'hyperplan orthogonal à v contenant $tv/||v||^2$.
- $\mathcal{D}_H$ est utilisé pour décrire le domaine différentiable d'une fonction (vectorielle) H, les points où la fonction H est différentiable.
- $\partial_B H(x) := \{J \in \mathcal{L}(\mathbb{E}, \mathbb{F}) : H'(x_k) \to J, \mathcal{D}_H \ni x_k \to x\}$ (B pour Bouligand).
- $\partial_C H(x) := \operatorname{conv}(\partial_B H(x))$, aussi écrit $\partial H(x)$.
- $\partial_\times H(x) := \partial_C H_1(x) \times \cdots \times \partial_C H_n(x)$. Parfois, la notation ∂_C est utilisée dans la littérature, mais "C" peut être compris comme une référence à Clarke.
- φ est une C-function, $\Phi : \mathbb{R}^n \mapsto \mathbb{R}^n$ la fonction vectorielle utilisant la C-fonction φ composante par composante et $\Psi := ||\Phi||^2/2$ la fonction de mérite associée.

2.2 Reformulations de problèmes de complémentarité

Rappelons de la section 1.1 que les problèmes considérés sont, dans l'ordre, le PCG pour deux fonctions F et G, le PCN pour une fonction F et le PCL pour une matrice M et un vecteur q :

$$\operatorname{PCG}(G, F) \qquad 0 \leqslant F(x) \perp G(x) \geqslant 0, \tag{2.1a}$$

$$\operatorname{PCN}(F) \qquad 0 \leqslant x \perp F(x) \geqslant 0, \tag{2.1b}$$

$$\operatorname{PCL}(M, q) \qquad 0 \leqslant x \perp Mx + q \geqslant 0. \tag{2.1c}$$

2.2.1 Quelques types de PCL

Cette section vise à présenter quelques-unes des nombreuses classes de matrices impliquées dans les PCL, qui distinguent les types et variantes de PCL. Les définitions sont principalement tirées de [58, 86, 181]. Aucune de ces matrices n'est nécessairement symétrique. Premièrement, la classe des $\mathbf{P}$-matrices garantit qu'il existe une unique solution à $\operatorname{PCL}(M, q)$ pour tout $q \in \mathbb{R}^n$ (voir aussi [225]).

Définition 2.2.1 ($\mathbf{P}$-matrices). M est une $\mathbf{P}$-matrice, notée $M \in \mathbf{P}$, si elle vérifie l'une des conditions équivalentes suivantes :

(i) $\forall\, q \in \mathbb{R}^n$, $\operatorname{PCL}(M, q)$ a exactement une solution,
(ii) $\forall\, I \subseteq [1 : n]$, $\det(M_{I,I}) > 0$,
(iii) $\forall\, I \subseteq [1 : n]$, $\operatorname{sp}(M_{I,I}) \cap \mathbb{R} \subseteq \mathbb{R}_{++}$,
(iv) tout $x \in \mathbb{R}^n$ vérifiant $x \bullet (Mx) \leqslant 0$ s'annule. □

Comme le suggèrent les points (ii) et (iii), vérifier si M appartient à $\mathbf{P}$ n'est pas simple. Coxson [60] a montré que c'est un problème co-NP-complet, en utilisant un résultat sur la NP-complétude d'un problème traitant de la non-singularité d'un intervalle de matrices contenant une matrice non singulière. De nombreuses autres caractérisations des $\mathbf{P}$-matrices existent. Ben Gharbia et Gilbert, dans [23, 24], donnent une définition équivalente basée sur l'absence de cyclage entre deux points dans un algorithme spécifique (voir section 2.3.3). Rump dans [222] propose une autre caractérisation basée sur les spectres matriciels et la transformée de Cayley. L'article montre également qu'aucun sous-ensemble ne peut être ignoré dans les points (ii) et (iii) de la définition 2.2.1.

Les propriétés discutées sont également observées, bien qu'énoncées dans d'autres termes, dans l'article de Samelson, Thrall et Wesler [225], qui donne une propriété sur la manière dont la paire $(I, -M)$ peut décomposer l'espace. Leur théorème est énoncé avec 2 matrices mais l'une peut être l'identité.

La classe suivante, pour laquelle le PCL a toujours (au moins) une solution, n'a pas de caractérisation algébrique équivalente connue.

Définition 2.2.2 (Matrices $\mathbf{Q}$)**.** M est une $\mathbf{Q}$-matrice, notée $M \in \mathbf{Q}$, si pour tout $q \in \mathbb{R}^n$, $\mathrm{LCP}(M, q)$ a (au moins) une solution. □

La classe suivante discute d'une sous-classe particulière de matrices inversibles.

Définition 2.2.3 (Matrices $\mathbf{ND}$)**.** M est une $\mathbf{ND}$-matrice, notée $M \in \mathbf{ND}$ si elle vérifie l'une des conditions équivalentes suivantes :

(i) $\forall\, I \subseteq [1:n]$, $\det(M_{I,I}) \neq 0$,
(ii) tout $x \in \mathbb{R}^n$ vérifiant $x \cdot (Mx) = 0$ s'annule. □

Un aperçu des relations entre les classes présentées et bien d'autres peut être trouvé dans [20, p. 37, fig. 2.2.1, en français]. Un aperçu de nombreux problèmes de complexité sur les classes de matrices peut être trouvé dans un article de Tseng [245] : déterminer la $\mathbf{P}$-matricité, la monotonie stricte, la suffisance colonne (corollaire 1 p. 187), la $\mathbf{P}_0$-matricité, la semi-monotonie (corollaire 2 p. 190), le caractère $\mathbf{R}_0$ et la non-dégénérescence (corollaire 3 p. 191) sont tous des problèmes co-NP-complets.

2.2.2 Quelques types de PCN(F)

À l'instar de la section précédente, celle-ci vise à présenter quelques classes de fonctions impliquées dans les PCN(F), les définitions étant tirées de [243] par exemple. Certaines d'entre elles sont analogues à certaines classes de matrices rencontrées pour le PCL, où $F'(x)$ (pour tout $x \in \mathbb{R}^n$) joue le rôle de la matrice M.

Définition 2.2.4 (types de fonction $\boldsymbol{F}$)**.** Soit $x, y \in \mathbb{R}^n$. On dit que F est une fonction…

$\mathbf{P}_0$	si $\exists\, i$	$(x_i - y_i)(F_i(x) - F_i(y)) \geqslant 0$	$\Leftrightarrow F'(x) \in \mathbf{P}_0$,
$\mathbf{P}$	si $\exists\, i$	$(x_i - y_i)(F_i(x) - F_i(y)) > 0$,	
$\mathbf{P}$ − *uniforme*	si $\exists\, i$	$(x_i - y_i)(F_i(x) - F_i(y)) \geqslant \mu\|\|x - y\|\|^2$	$\Leftrightarrow F'(x) \in \mathbf{P}$,
monotone	si	$(x - y)^\mathsf{T}(F(x) - F(y)) \geqslant 0$,	
strictement monotone	si	$(x - y)^\mathsf{T}(F(x) - F(y)) > 0$,	
fortement monotone	si	$(x - y)^\mathsf{T}(F(x) - F(y)) \geqslant \mu\|\|x - y\|\|^2$	

□

Au lieu de $\exists\, i$, un maximum sur les i est parfois utilisé. Dans les trois premières, $x_i \neq y_i$, et $x \neq y$ dans la cinquième. On a clairement

$$\text{monotone} \Rightarrow \mathbf{P}_0, \text{ strictement monotone} \Rightarrow \mathbf{P}, \text{ fortement monotone} \Rightarrow \text{uniformément } \mathbf{P}$$

en utilisant que $(x - y)^\mathsf{T}(F(x) - F(y)) = \sum_i (x_i - y_i)(F_i(x) - F_i(y))$. La monotonicité est par exemple utilisée par Subramanian [240]. Megiddo [168] a trouvé un exemple simple en dimension 2 d'un $\mathrm{PCN}(F)$ sans solution bien que F soit monotone.

2.2.3 Complexité générale

Avant d'aborder les reformulations, évoquons la complexité des problèmes de complémentarité. Comme les PCL sont un cas particulier des PCN, les résultats de complexité sur les PCL montrent déjà la difficulté des problèmes considérés. Comme discuté dans la section 2.2.1, même déterminer le type de la matrice impliquée dans un PCL peut être difficile.

Chung [49] a montré que le PCL est, en général, NP-complet, et même fortement NP-complet.

Lorsque $M \in \mathbf{P}$, Megiddo a montré que le problème n'était probablement pas "difficile" [169] : un théorème montre que si $\mathrm{PCL}(M, q)$ est NP-difficile, alors NP = co-NP. De plus, lorsque M est symétrique et (semi-)définie positive, la méthode des ellipsoïdes peut être utilisée pour résoudre le PCL en temps polynomial [181]. Lorsque $M \in \mathbf{P}_0$, la complexité devient NP-complète comme décrit par Kojima, Megiddo, Noma et Yoshisa [144].

2.2.4 Équations généralisées et applications normales

Les problèmes de complémentarité peuvent être exprimés et formulés de nombreuses manières différentes, ce qui souligne l'importance et la pertinence de leur étude. Les PCL peuvent être énoncés, par exemple, comme une équation généralisée :

$$0 \in Mx + q + N_{\mathbb{R}^n_+}(x). \tag{2.2}$$

En utilisant une formule de produit cartésien pour le cône normal [221], [105, proposition 2.30 2) p. 49], on obtient :

$$N_{\mathbb{R}^n_+}(x) = \prod_{i=1}^{n} N_{\mathbb{R}_+}(x_i) = \begin{cases} 0 & x_i > 0 \\ \mathbb{R}_- & x_i = 0 \\ \varnothing & x_i < 0 \end{cases},$$

ce qui indique que les solutions du PCL et de (2.2) sont les mêmes. Étudiées en détail par Robinson [217, 220, 216], elles représentent un cadre plus large avec de nombreuses applications telles que les conditions d'optimalité ou d'autres problèmes d'équilibre.

Ralph [211] présente une reformulation similaire pour le $\mathrm{PCN}(F)$, une *application normale*, qui prend la forme (rappelons que $z = z_+ + z_-$) :

$$F(z_+) + (z - z_+) = 0 = F(z_+) + z_-. \tag{2.3}$$

Les solutions des deux problèmes sont liées par $z = x - F(x)$ pour x une solution de (2.1b) et $x = z_+$ pour z une solution de (2.3). En effet, si x résout (2.1b), pour certains $i \in [1 : n]$, $z_i = x_i - F_i(x)$. Par complémentarité de x et $F(x)$, on a $z_+ = x$ et l'équation de l'application normale devient $F(x) + (x - F(x) - x)$ qui s'annule clairement. Inversement, si z résout (2.3), $x = z_+ \geqslant 0$, $F(x) = z_+ - z = -z_- \geqslant 0$. Alors, $x_i F_i(x) = (z_+)_i (z_-)_i = 0$, ce qui signifie que la complémentarité est respectée.

En plus de la méthode de Ralph, nous mentionnons la contribution de Sun et Qi [243] où ils modifient l'application normale par une technique de lissage (voir section 2.3.6).

Les applications normales ont été étudiées par Robinson dans [219], où une transformation linéaire est considérée mais avec un ensemble convexe polyédrique C au lieu de l'orthant non négatif. Après lui, Josephy [134] propose un schéma qui linéarise la fonction, transformant $\mathrm{PCN}(F)$ en une séquence de PCL à résoudre, de même dimension.

Dans (2.3), on pourrait aussi utiliser $F(z_+) + z_- = 0$. Cette forme est employée par exemple par Harker et Xiao [121], où elle est appelée application de Minty d'après [173]. Leur algorithme résout une équation avec la B-dérivée pour obtenir une direction avant de lancer une recherche linéaire. Par conséquent, il traite également un PCL mixte à chaque itération. Cependant, il est différent pendant la phase de recherche linéaire, où ils suggèrent qu'il peut se comporter mieux que la reformulation usuelle avec le minimum.

2.2.5 Méthodes de points intérieurs

Ici, nous mentionnons brièvement certains algorithmes traitant des problèmes de complémentarité par l'utilisation de méthodes de type intérieur. Pour l'optimisation sous contraintes d'inégalités de la forme

$$\min f(x), \qquad \text{t.q.} \qquad g(x) \leqslant 0,$$

les méthodes de points intérieurs traitent le système de conditions d'optimalité et remplacent les conditions de complémentarité $0 \leqslant \lambda \perp (-g(x)) \geqslant 0$ par $\lambda_i(-g(x))_i = \mu$ pour

un certain $\mu > 0$. Pour les propriétés générales sur les méthodes de points intérieurs, voir par exemple le livre de Wright [254].

Clairement, les méthodes de points intérieurs peuvent être adaptées pour les problèmes de complémentarité. Le livre de Kojima, Megiddo, Noma et Yoshise [144] discute le cas des PCL. Kojima et Yoshise, avec Mizuno [145], toujours pour les PCL, ont amélioré la complexité de $O(Ln^{7/2})$ ($O(Ln^{1/2})$ itérations, puisque chacune divise la fonction de mérite par $(1-\eta/\sqrt{n})$, et chacune résout un système linéaire donc $O(n^3)$) à $O(Ln^3)$ avec L la taille de l'entrée et n la dimension. La réduction à n^3 est basée sur des manipulations astucieuses des systèmes et sous-problèmes à résoudre. Une approche similaire est considérée dans [44].

Une application issue de l'analyse numérique peut être trouvée dans [21], où la discrétisation conduit à un système avec une contrainte de complémentarité, c'est-à-dire un système "mixte". La méthode de points intérieurs est comparée à d'autres approches populaires discutées ci-dessous. Des informations supplémentaires peuvent être trouvées dans les références citées.

2.2.6 Équation de valeur absolue

Les PCL sont également équivalents, après un changement de variables, à un type de problèmes appelés "équations de valeur absolue" (AVE). Elles prennent la forme

$$Ax - |x| = b \qquad \text{ou} \qquad x - A|x| = b \tag{2.4}$$

où $|x| = (|x_i|)_{i\in[1:n]}$ est la valeur absolue composante par composante de x. La calcul reliant PCL et équations de valeur absolue peut être trouvée par exemple dans un article de Mangasarian et Meyer [160], avec une transformation affine (mais non triviale) entre la variable du PCL et celle de l'AVE. Ils discutent des propriétés concernant l'existence de solutions en étudiant les spectres de certaines matrices. Radons et Tonelli-Cueto [210] discutent également des propriétés de l'AVE et de l'application associée $x \mapsto x - A|x|$ en étudiant un spectre adapté de A. Leur travail utilise la théorie du degré, qui intervient également dans l'étude des PCL (en particulier les propriétés d'existence), voir [58, chapitre 6] et [86, section 2.1, pp. 126-145]. Le cas non linéaire $F(x) - |x| = 0$, relié au PCN, a récemment été considéré dans [63].

2.2.7 Problèmes avec complémentarité dans les contraintes

Maintenant, nous mentionnons des contributions traitant de ce qu'on peut appeler des Programmes Mathématiques avec Contraintes d'Équilibre / de Complémentarité (MPECs / MPCCs). Par exemple, Scheel et Scholtes [226] traitent des problèmes de la forme

$$\min f(z), \qquad \text{t.q.} \qquad g(z) \leqslant 0, \qquad h(z) = 0, \qquad \min(F^1(x), \dots, F^l(x)) = 0$$

avec $F^1, \dots, F^l$ des fonctions lisses de $\mathbb{R}^n$ vers $\mathbb{R}^m$. Ils discutent des moyens de décomposer les contraintes et diverses formes de stationnarité pour de tels problèmes. Les MPECs ont également été considérés comme une application dans [197].

Hintermüller et Kopacka [125] discutent le cas en dimension infinie, où les notions de stationnarité doivent être adaptées aux difficultés rencontrées. Leur algorithme résout des sous-problèmes régularisés, qui sont non lisses, et utilise une approche de recherche linéaire pour globaliser la convergence, bien qu'"il n'y ait pas de descente garantie le long d'un tel chemin" (voir leur remarque 5.4 p. 888).

2.3 Cadre non lisse et algorithmes

Cette section est consacrée au cadre non lisse découlant de (la plupart) des C-fonctions présentées dans la section 2.3.1. Nous rappelons quelques notions importantes liées dans la section 2.3.2. Ensuite, les sections 2.3.3 et 2.3.5 discutent *quelques* méthodes existantes : hypothèses, conditions sur la solution, sous-problèmes résolus, etc. Entre elles, la section 2.3.4 discute quelques contributions sélectionnées qui ont motivé cette thèse. Suite à cela, la notion de *lissage*, c'est-à-dire modifier légèrement le système pour obtenir un système lisse, est présentée dans la section 2.3.6 ainsi que certaines contributions liées. Enfin, la section 2.3.7 évoque quelques questions de complexité, pour mieux situer les différents résultats.

2.3.1 Introduction aux C-fonctions

Une C-fonction (C pour complémentarité) est une fonction scalaire à deux variables définie comme suit.

Définition 2.3.1 (C-fonctions)**.** Une fonction $\varphi : \mathbb{R}^2 \to \mathbb{R}$ est dite C-fonction si elle vérifie la condition suivante pour tous a, b dans $\mathbb{R}^n$:

$$\varphi(a, b) = 0 \qquad \Longleftrightarrow \qquad a \geqslant 0, \qquad b \geqslant 0, \qquad ab = 0. \tag{2.5}$$

□

L'utilisation principale des C-fonctions est la propriété suivante :

$$0 \leqslant F(x) \perp G(x) \geqslant 0 \qquad \Longleftrightarrow \qquad \Phi(x) := \begin{pmatrix} \varphi(F_1(x), G_1(x)) \\ \vdots \\ \varphi(F_n(x), G_n(x)) \end{pmatrix} = 0, \tag{2.6}$$

où la C-fonction est appliquée composante par composante. Équivalemment, minimiser la fonction de mérite $\Psi(x) := ||\Phi(x)||^2/2$ (avec une valeur optimale de 0 si le PC initial a une solution) est une autre reformulation. Bien qu'il existe de nombreuses C-fonctions, deux semblent avoir un rôle particulier comme fonctions de base qui ont inspiré la plupart des autres.

$$\begin{aligned} \varphi_{FB}(a, b) &:= \sqrt{a^2 + b^2} - (a + b) \\ \varphi_{\min}(a, b) &:= \min(a, b) = a - (a - b)_+ = b - (b - a)_+ \end{aligned} \tag{2.7}$$

Dans la première, FB signifie Fischer-Burmeister, bien que parfois seul Fischer soit cité [92]. Le minimum a été utilisé pour la première fois par Kostreva [147] en 1976, et les deux autres expressions ont été observées pour la première fois par Wierzbicki dans [252]. Voir ci-dessous pour plus de commentaires.

Plusieurs C-fonctions sont non lisses, c'est-à-dire que le PC devient l'équation $H(x) = 0$ avec $H : \mathbb{R}^n \to \mathbb{R}^n$ non lisse. Malgré l'existence de reformulations lisses, elles peuvent ne pas être aussi appropriées, comme détaillé dans [86, prop. 9.1.1, pp. 794-795]. En bref, les algorithmes résultant d'une reformulation lisse (C-fonction) ne peuvent pas bénéficier d'une convergence locale rapide autour de solutions *dégénérées* (voir la fin de cette section).

Mangasarian [159] propose un cadre plus général pour obtenir des C-fonctions.

Définition 2.3.2 (Reformulation de Mangasarian)**.** Soit $\rho : \mathbb{R} \to \mathbb{R}$ une fonction (strictement) croissante avec $\rho(0) = 0$. Alors x résout le problème de complémentarité 2.1a (et de même 2.1b ou 2.1c) si et seulement si

$$\rho(|F_i(x) - G_i(x)|) - \rho(F_i(x)) - \rho(G_i(x)) = 0, \forall\, i \in [1 : n]. \tag{2.8}$$

□

Autrement dit, le cadre de Mangasarian énonce que

$$\tilde{\rho}(a, b) := \rho(|a - b|) - \rho(a) - \rho(b)$$

pour ρ vérifiant les propriétés de la définition est une C-fonction. En particulier, le choix $\rho(t) = t$ donne la C-fonction $-2\varphi_{\min}$. Mangasarian discute également de la non-singularité du Jacobien à une solution pour $\mathrm{PCN}(F)$ de (2.1b). En particulier (p. 91), si $\rho'(0) = 0$, le système devient différentiable.

Un exemple de contribution basée sur la reformulation lisse de Mangasarian [159] est celle de Subramanian [241], qui nécessite des hypothèses fortes sur la solution x^* pour obtenir des résultats de convergence : non-dégénérescence (ou complémentarité stricte, $x_i^* + F_i(x^*) > 0$) de la solution, $\nabla F(x^*)$ est une matrice non dégénérée. Cette hypothèse de non-dégénérescence de la solution est particulièrement forte puisqu'elle dépend purement du problème lui-même et n'est pas nécessairement souvent vérifiée.

Un autre cadre de construction est donné par Luo et Tseng dans [156], également utilisé par Kanzow, Yamashita et Fukushima dans [138]. Il consiste à combiner d'autres fonctions sous la forme suivante (ils utilisent une condition opposée sur le signe des variables dans les C-fonctions)

$$\Psi(x) = \sum \varphi_i(x_i, F_i(x)), \quad \varphi_i(a, b) = \psi_0(ab) + \psi_i(-a, -b),$$

où les fonctions ψ sont continues et égales à 0 sur l'orthant négatif. Par exemple, $\psi_0(t) = (t_+)^p$,

$$\psi_i(a, b) \in \begin{cases} (a_+ + b_+)^p \\ (a_+^2 + b_+^2)^{p/2} \\ ((\sqrt{a^2 + b^2} + a + b)_+)^p \\ \max(0, a, b)^p \end{cases}$$

avec $p \geqslant 1$ un entier positif.

Beaucoup des C-fonctions ne sont pas lisses (partout), mais elles sont dites *semi-lisses*, une propriété entre la lipschitzianité et la lissité, voir la section 2.3.2. Nous nous concentrons essentiellement sur les C-fonctions "symétriques" (en a et b), bien qu'il existe de nombreuses fonctions asymétriques, qui peuvent être intéressantes pour traiter différemment x et $F(x)$ (resp. $Mx + q$) dans $\mathrm{PCN}(F)$ (resp. $\mathrm{PCL}(M, q)$) par exemple.

Des informations supplémentaires sur les C-fonctions peuvent être trouvées dans l'état de l'art de Fischer et Jiang [94] discutant des principales propriétés des C-fonctions, telles que la lissité, les formes des ensembles de sous-niveaux, l'efficacité des directions... voir aussi [138] et les références citées. Malgré des décennies d'innovation sur les C-fonctions, il existe encore de nouvelles fonctions conçues : Galántai [99] discute comment les C-fonctions peuvent être construites ou *non* construites. Une manière de décomposer les C-fonctions est également évoquée. Il donne un bestiaire contenant environ 30 C-fonctions, dont beaucoup sont basées sur la fonction FB d'une manière ou d'une autre. Dans [6], les auteurs discutent de constructions plus élaborées de C-fonctions basées sur une généralisation de la fonction Fischer, utilisant la p-norme au lieu de la 2-norme de $(a; b)$. Les considérations algorithmiques sont détaillées dans la section 2.3.3.

Le terme "dégénérescence" est souvent utilisé dans divers contextes mathématiques, allant de l'optimisation linéaire à la géométrie combinatoire (voir plus loin) en passant par les problèmes de complémentarité, pour mettre en évidence certaines difficultés techniques pertinentes. Dans les problèmes de complémentarité, une solution x^* est dite dégénérée s'il existe certains indices i pour lesquels $x_i^* = 0 = F_i(x^*)$ (avec des définitions similaires pour le PCL et d'autres variantes de PC). Ces indices, qui dépendent uniquement du problème lui-même, rendent la plupart des C-fonctions non différentiables, ce qui explique la difficulté : il faut avoir des conditions de régularité pour que *toutes* les matrices jacobiennes à la solution soient non singulière.

2.3.2 Quelques outils d'analyse non lisse

Comme la reformulation par C-fonction (2.6) conduit souvent à un système d'équations non lisse, l'analyse non lisse est donc un outil important pour étudier ces systèmes. La référence principale avec laquelle nous commençons est le livre de Clarke [51], qui suppose que les fonctions sont lipschitziennes, une propriété qui est uniformément vérifiée dans cette thèse, bien qu'il existe des extensions sans lipschitzianité (faites par Rockafellar par exemple). Pour simplifier, nous supposons que l'espace courant est $\mathbb{R}^n$, donc un espace de dimension finie qui simplifie certaines notations.

Définition 2.3.3 (Fonction lipschitzienne)**.** Une fonction $f : \mathbb{R}^n \to \mathbb{R}$ est localement L-lipschitzienne en $x \in \mathbb{R}^n$ si pour tout y et z dans un voisinage de x, $|f(z) - f(y)| \leqslant L||z - y||$. □

Elle est dite localement lipschitzienne si elle est lipschitzienne sur chaque sous-ensemble borné de son domaine, et (globalement) lipschitzienne si l'inégalité est vérifiée pour tout x

et y. Habituellement, pour les fonctions lisses, on a le développement de Taylor

$$f(x+d) = f(x) + \nabla f(x)^{\mathsf{T}} d + o(||d||) = f(x) + f'(x)d + o(||d||). \tag{2.9}$$

Cependant, lorsque f est non lisse en x, on ne peut pas écrire un tel développement aussi facilement. Nous verrons quelques généralisations de ce développement. Néanmoins, grâce au théorème de Rademacher [209] (voir aussi [123]), les fonctions lipschitziennes (en dimension finie) bénéficient de la propriété suivante. Rappelons que $\mathcal{D}_H$ est le domaine différentiable de H, les points où H est différentiable.

Théorème 2.3.4 (Théorème de Rademacher)**.** *Les fonctions lipschitziennes sont différentiables presque partout. Autrement dit, si H est une fonction lipschitzienne,* $\mathrm{mes}(\mathcal{D}_H^c) = 0$ *où* mes *est la mesure de Lebesgue.* □

Lorsque f est non lisse en x, on peut généraliser le terme $f'(x)d = \nabla f(x)^{\mathsf{T}} d$ dans (2.9).

Définition 2.3.5 (Dérivée directionnelle)**.** Soit f lipschitzienne au voisinage de $x \in \mathbb{R}^n$, la dérivée directionnelle en x dans la direction $d \in \mathbb{R}^n$ est notée et définie par

$$f'(x;d) := \lim_{t \searrow 0} \frac{f(x+td) - f(x)}{t}. \tag{2.10}$$

□

Comme cette limite peut ne pas être définie partout, parfois une autre définition, avec une limsup, est utilisée, la dérivée directionnelle supérieure de Dini.

Définition 2.3.6 (Dérivée directionnelle supérieure de Dini)**.** Soit f lipschitzienne au voisinage de $x \in \mathbb{R}^n$, la dérivée directionnelle supérieure de Dini en x dans la direction $d \in \mathbb{R}^n$ est notée et définie par

$$f^D(x;d) := \limsup_{t \searrow 0} \frac{f(x+td) - f(x)}{t} \tag{2.11}$$

□

En particulier, $f^D(x;d) = f'(x;d)$ lorsque ce dernier existe. Clarke [51, section 2.1, p. 25] utilise plutôt la définition ci-dessous, la $\limsup$ s'appliquant également à la quantité dans $\mathbb{R}^n$ (et pas seulement le scalaire).

Définition 2.3.7 (Dérivée directionnelle généralisée)**.** Soit f lipschitzienne au voisinage de $x \in \mathbb{R}^n$, la dérivée directionnelle généralisée en x dans la direction $d \in \mathbb{R}^n$ est notée et définie par

$$f^\circ(x;d) := \limsup_{y \to x, t \searrow 0} \frac{f(y+td) - f(y)}{t}. \tag{2.12}$$

□

Puisque dans (2.12) le y peut être pris égal à x, on a clairement $f^\circ(x;d) \geqslant f^D(x;d)$. Comme le montre l'exemple suivant, le minimum est une fonction qui expose la différence entre ces dérivées directionnelles.

Exemple 2.3.8 (Cas du minimum). *Soit $f(x) = -|x| = \min(x, -x)$, qui est clairement lipschitzienne. Alors $f'(0; d) = -|d|$ et $f^\circ(0; d) = +|d|$.*

En effet, (2.10) *donne* $\lim_t -|td|/t = \lim_t -|d| = -|d|$, *alors que dans la* $\limsup$, *prendre* $y = -td \to 0 = x$ *donne* $+|d|$. □

Clarke utilise la définition 2.3.7 pour définir un certain différentiel :

Définition 2.3.9 (Gradient généralisé, [51, p. 27])**.**

$$\partial f(x) := \{\zeta \in \mathbb{R}^n : f^\circ(x; d) \geqslant (\zeta, d) \; \forall \, d \in \mathbb{R}^n\} \tag{2.13}$$

En particulier : [51, proposition 2.1.2, p. 27] pour tout $d \in \mathbb{R}^n$, $f^\circ(x; d) = \max\{\zeta^\mathsf{T} d; \zeta \in \partial f(x)\}$; [51, proposition 2.1.5, p. 29] $\zeta \in \partial f(x) \Leftrightarrow f^\circ(x; d) \geqslant \zeta^\mathsf{T} d$ pour tout $d \in \mathbb{R}^n$, $\partial f(x)$ est fermé (fermé en topologie faible* sans l'hypothèse de dimension finie). □

Certaines notions de stationnarité liées à ces dérivées directionnelles sont discutées ci-dessous dans la section 2.3.7.

Par exemple, on a que $\partial(-|\cdot|)(0) = \partial(|\cdot|)(0) = [-1, +1]$. Dans ce qui suit, les différentiels "usuels", les différentiels au sens de Clarke, sont notés ∂ ou ∂_C (le différentiel construit avec le produit cartésien composante par composante des différentiels de f_i avec $f = (f_i)_i$ est noté avec $\times$ pour éviter la confusion possible). Un autre type de différentiels sont les B-différentiels, où le "B" signifie Bouligand.

Définition 2.3.10 (B-différentiel scalaire, [51, p. 63])**.** Le B-différentiel de f en x est noté et défini par

$$\partial_B f(x) := \{v : \exists \, \{x_k\} \subseteq \mathcal{D}_f, x_k \to x, \nabla f(x_k) \to v\}. \tag{2.14}$$

En particulier, f est différentiable aux points x_k de la suite. □

Par exemple, on a $\partial_B(|\cdot|)(0) = \{-1, +1\} = \partial_B(-|\cdot|)(0)$. Comme détaillé dans [51, théorème 2.5.1 p. 63], les points x_k peuvent aussi être pris hors d'un ensemble S de mesure nulle (puisque nous considérons la dimension finie), et une relation clé entre ∂_B et $\partial := \partial_C$ est la suivante.

Proposition 2.3.11 (C-différentiel scalaire)**.** *On a l'égalité suivante*

$$\partial_C f(x) = \operatorname{conv} \partial_B f(x). \tag{2.15}$$

□

En particulier [51, proposition 2.2.7], pour les fonctions convexes univoques lipschitziennes f, $\partial f(x)$ est le sous-différentiel convexe usuel, et $f^\circ = f'$. Pour les fonctions non convexes, une fonction f vérifiant $f' = f^\circ$ est dite régulière par Clarke.

Définition 2.3.12 (Régularité [51, p. 39])**.** La fonction f est dite régulière en $x \in \mathbb{R}^n$ si $f'(x; d)$ existe pour tout $d \in \mathbb{R}^n$ et $f'(x; d) = f^\circ(x; d)$. □

Dans de nombreux articles, cette notion est appelée "régularité sous-différentielle", puisque "régularité" est déjà utilisé pour des conditions qui assurent que les points d'accumulation des algorithmes sont des solutions du problème considéré (par exemple dans [119]).

En particulier, $-|\cdot|$ n'est pas régulière en 0. Les difficultés de la fonction $-|\cdot|$, liées au minimum par $-|x| = \min(x, -x)$, ont été observées par exemple par Pang, Han et Rangaraj dans [197, p. 60]. Certaines des définitions peuvent être adaptées aux fonctions vectorielles [51, sections 2.2 p. 30 et 2.6 p. 70].

Définition 2.3.13 (Dérivée directionnelle vectorielle [51, p. 30])**.** Soit $F : \mathbb{R}^n \to \mathbb{R}^m$, la dérivée directionnelle de F en x dans la direction d est notée et définie par

$$F'(x; d) := \lim_{t \searrow 0} \frac{F(x + td) - F(x)}{t} \tag{2.16}$$

On dit que F a une dérivée au sens de Gâteaux si cette limite est égale à $DF(x)d$ pour un élément $DF(x) \in \mathbb{R}^{m \times n}$ et tout $d \in \mathbb{R}^n$. □

Cela peut être utilisé pour définir la notion de BD-régularité par Qi [204, p. 232]. Les conditions de régularité sont souvent utilisées à/autour d'une solution pour assurer de bonnes propriétés locales des algorithmes.

Définition 2.3.14 (BD-régularité)**.** Soit F différentiable directionnellement en x, F est dite BD-régulière en x si la condition suivante est vérifiée :

$$\forall\, h \in \mathbb{R}^n \backslash \{0\}, \quad F'(x; h) \neq 0. \tag{2.17}$$

En particulier, cela implique $||h|| \leqslant c||F'(x; h)||$ pour tout h et une constante $c > 0$. □

Définition 2.3.15 (B-différentiel vectoriel)**.** Soit $F : \mathbb{R}^n \to \mathbb{R}^m$ une fonction lipschitzienne, le B-différentiel de F en x est noté et défini par

$$\partial_B F(x) := \{J \in \mathbb{R}^{m \times n} : \exists\, \{x_k\} \in \mathcal{D}_F, x_k \to x, F'(x_k) \to J\} \tag{2.18}$$

□

Encore une fois, prendre l'enveloppe convexe du B-différentiel donne le C-différentiel.

Définition 2.3.16 (C-différentiel vectoriel)**.** Soit $F : \mathbb{R}^n \to \mathbb{R}^m$ une fonction lipschitzienne, le C-différentiel de F en x est noté et défini par

$$\partial_C F(x) := \mathrm{conv}(\partial_B F(x)) \tag{2.19}$$

Le C-différentiel est non vide fermé convexe compact [51, proposition 2.6.2a-b]. □

Le B-différentiel de la définition 2.3.15, bien que moins populaire que sa version convexifiée de la définition 2.3.16, est utilisée par Qi pour définir une autre forme de régularité. Elle est souvent appelée "BD-régularité", reconnaissant que Qi a initialement utilisé ce terme pour la définition 2.3.14 (et même par Qi lui-même !).

Définition 2.3.17 (forte BD-régularité [204, p. 233])**.** Une fonction F est dite fortement BD-régulière en x si pour toute matrice $V \in \partial_B F(x)$, V est inversible.

En particulier [204, lemma 2.6], la forte BD-régularité est une propriété diffusante, dans le sens où toutes les matrices jacobiennes aux points voisins de x sont également inversibles, et toutes les inverses en ces points voisins peuvent être majorées par une même constante. □

Les inclusions suivantes (le plus souvent strictes) seront pertinentes ultérieurement.

Proposition 2.3.18 (inclusion dans le différentiel produit, [51, proposition 2.6.2e])**.**

$$\begin{aligned} \partial_B F(x) \subseteq \partial_B^\times F(x) &:= \partial_B F_1(x) \times \cdots \times \partial_B F_m(x), \\ \partial_C F(x) \subseteq \partial_\times F(x) &:= \partial_C F_1(x) \times \cdots \times \partial_C F_m(x). \end{aligned} \tag{2.20}$$

De plus, par définition, $\partial_B F(x) \subseteq \partial_C F(x)$ avec égalité si F est continûment différentiable. □

Parmi les propriétés les plus utiles, on trouve les règles de la chaîne. Nous citons seulement l'une des trois versions de Clarke (les autres s'appliquant à certains cas différents mais dans un esprit similaire).

Proposition 2.3.19 (règle de la chaîne, [51, proposition 2.6.6, pp. 72-73])**.** *Soit $f = g \circ F$, où $F : \mathbb{R}^n \to \mathbb{R}^m$ est lipschitzienne en x et où $g : \mathbb{R}^m \to \mathbb{R}$ est lipschitzienne en $F(x)$ pour un $x \in \mathbb{R}^n$. Alors f est lipschitzienne en x et on a*

$$\partial f(x) \subseteq \operatorname{conv}\{\partial g(F(x))\partial F(x)\}. \tag{2.21}$$

Si en plus g est strictement différentiable en $F(x)$, alors il y a égalité (et conv *est superflu).* □

En particulier, continûment différentiable implique strictement différentiable, qui est moins restrictif. Si l'ouvrage de Clarke [51] reste un travail primordial, pour le cas spécifique des équations non lisses issues des C-fonctions, certaines notions supplémentaires ont été introduites, basées sur les travaux de Clarke. La première est la B-dérivée, utilisée par Robinson dans [218, appendice, p. 62 et suivantes], dans le contexte de l'étude de la structure locale de l'espace autour des solutions des problèmes d'optimisation. En fait, grâce à un résultat de Shapiro [230], en dimension finie, la dérivabilité directionnelle et la B-différentiabilité sont équivalentes.

Définition 2.3.20 (B-dérivée et B-différentiabilité)**.** Soit $F : \mathbb{R}^n \to \mathbb{R}^m$, elle est dite B-différentiable en $x \in \mathbb{R}^n$ s'il existe une fonction $BH(x) : \mathbb{R}^n \to \mathbb{R}^m$, appelée la B-dérivée de F en x, vérifiant $BH(x)(\lambda v) = \lambda BH(x)(v)$ pour $\lambda \in \mathbb{R}_+$ telle que

$$F(x+v) = F(x) + BF(x)(v) + o(||v||). \tag{2.22}$$

Elle est B-différentiable si cette propriété est vraie pour tout $x \in \mathbb{R}^n$. Grâce au résultat de Shapiro, (2.22) peut aussi s'écrire (la dimension finie est supposée dans cette thèse)

$$F(x+v) = F(x) + F'(x;v) + o(||v||).$$ □

Une autre notion cruciale est la semi-lissité. Ces fonctions sont souvent considérées comme intermédiaires entre les fonctions Lipschitz et C^1. Mifflin [171] les a introduites pour la minimisation de f sous la contrainte $h(x) \leqslant 0$ mais avec des fonctions non lisses et non convexes.

Définition 2.3.21 (fonction scalaire semi-lisse)**.** La fonction $f : \mathbb{R}^n \to \mathbb{R}$ est semi-lisse en x si

- f est Lipschitz au voisinage de x,
- pour tout $d \in \mathbb{R}^n$, pour toutes suites $\mathbb{R}_+ \supseteq \{t_k\} \searrow 0$, $\{y_k\} \subseteq \mathbb{R}^n$, $\{g_k\} \subseteq \mathbb{R}^n$ avec $y_k/t_k \to 0$, $g_k \in \partial f(x + t_k d + y_k)$, alors $\{g_k^{\mathsf{T}} d\}$ admet exactement un point d'accumulation. □

Pour être appliquée aux systèmes d'équations non lisses, une version vectorielle de la non-lissité est nécessaire. Cela a été réalisé par Qi et Sun [206, p. 354 et suivantes].

Définition 2.3.22 (fonction vectorielle semi-lisse)**.** L'application $F : \mathbb{R}^n \to \mathbb{R}^m$ est semi-lisse en x si la limite suivante existe pour tout $d \in \mathbb{R}^n$

$$\lim_{V \in \partial F(x+td'), d' \to d, t \searrow 0} V d'. \tag{2.23}$$

Des définitions équivalentes sont données dans [206, théorème 2.3 p. 356]. □

Qu'il s'agisse de fonctions scalaires ou vectorielles, la classe des fonctions semi-lisses inclut les fonctions lisses, lisses par morceaux, convexes, et est stable par addition, multiplication et composition, comme l'a montré Mifflin. Qi et Sun démontrent également une propriété simple [206, lemme 2.2, p. 356].

Lemme 2.3.23 (*F'* et *∂F*). *Si F est Lipschitz et $F'(x; d)$ existe pour des points x et d, alors on a $F'(x; d) = Vd$ pour une certaine matrice $V \in \partial F(x)$.* □

Une version forte de la semi-lissité existe également : ces fonctions sont plus "lisses" et bénéficient donc de propriétés de convergence plus fortes [204, lemme 2.3].

Définition 2.3.24 (semi-lissité forte)**.** L'application F est dite fortement semi-lisse en $x \in \mathbb{R}^n$ si la condition suivante est vérifiée :

$$\forall\, V \in \partial F(x + h), h \to 0, Vh - F'(x; h) = O(||h||^2). \tag{2.24}$$

□

En particulier, la plupart des C-fonctions de la section 2.3.5 sont fortement semi-lisses. Enfin, considérons une classe de fonctions qui "se situent entre les fonctions C^1 et C^2" ([65, p. 412]).

Définition 2.3.25 (fonctions SC^1)**.** L'application F est dite SC^1 si elle est C^1 et que son gradient est semi-lisse. □

D'autres points de vue sur l'analyse non lisse, utilisant des notions différentes, existent, comme la théorie développée par Mordukhovich [176, 177, 175].

2.3.3 Premiers algorithmes non lisses

Dans le cadre des C-fonctions que l'on développera dans la section 2.3.5, les problèmes de complémentarité deviennent des équations de la forme $H(x) = 0$. Lorsque H est lisse, une méthode classique parmi beaucoup d'autres [188] pour résoudre de tels systèmes d'équations est la méthode de Newton. Lorsqu'elle est démarrée suffisamment près de la solution, on sait qu'elle converge à un taux quadratique si la dérivée de H à la solution est inversible. L'équation de Newton, qui donne l'itéré $k+1$ à partir de l'itéré k, s'écrit :

$$x^{k+1} = x^k - H'(x_k)^{-1}H(x_k).$$

Sans l'hypothèse de lissité, deux questions se posent : comment remplacer la dérivée de H en x_k, et qu'est-ce qui assure une convergence locale rapide puisque la dérivée à la solution peut ne pas être définie ? De nombreuses variantes de la méthode de Newton ont été conçues pour les équations non lisses en jeu, issues de problèmes de complémentarité ou de problèmes équivalents associés. Dans cette section, nous discutons de certaines contributions concernant ces équations, en utilisant les outils définis dans la section précédente.

Idée générale : adapter la méthode de Newton

Commençons par mentionner l'observation de Kummer dans [150], où une certaine fonction affine par morceaux non lisse est construite et la méthode de Newton échoue même si elle est démarrée infiniment proche de la solution. Cet exemple justifie la nécessité de prendre en compte l'absence de lissité.

Le cadre général est le suivant [86, section 7.2, p. 638]. D'abord, considérons l'équation de Newton (dans le cas lisse) sous la forme

$$H(x^k) + H'(x^k)(x^{k+1} - x^k) = 0$$

et posons $d^k = x^{k+1} - x^k$. Puisque le second terme peut ne pas être correctement défini, considérons une *approximation*

$$H(x^k) + A(x^k, d^k) = 0,$$

où la quantité A reste vague pour l'instant. Autour d'un point $\overline{x}$ fixé, si l'approximation vérifie $A(x, 0) = 0$ et

$$\frac{G(x) + A(x, \overline{x} - x) - G(\overline{x})}{||x - \overline{x}||} = o(||x - \overline{x}||),$$

c'est une approximation de Newton. Si $O(||x-\overline{x}||^2)$ remplace le terme $o(||x-\overline{x}||)$, l'approximation est forte (elle donne une convergence plus rapide). Après avoir résolu l'équation de Newton approchée et obtenu une solution d^k, on met à jour $x^{k+1} = x^k + d^k$. De nombreuses possibilités pour A sont discutées dans [86], comme la résolution inexacte de l'équation approchée, le cas où H est une sélection par morceaux parmi plusieurs fonctions (lisses) ou est une composition…

Un choix particulier de l'approximation $A(x^k, d^k)$ est donné par le choix suivant :

$$A(x^k, d^k) := Jd^k, J \in \partial H(x^k) \tag{2.25}$$

où le différentiel de H en x_k est introduit dans les définitions 2.3.15 et 2.3.16. Ce choix plutôt naturel conduit cependant à une question importante, discutée plus en détail dans le chapitre 6 : comment choisir l'élément J du différentiel ? Souvent, des hypothèses sont faites pour que tout J soit approprié (voir ci-dessous).

Quelques améliorations générales

Dans cette section, nous mentionnons brièvement quelques techniques souvent utilisées dans les contributions évoquées ci-dessous. La première est la recherche linéaire, une méthode extrêmement classique, présentée par exemple sous la forme d'Armijo, nommée d'après l'article fondateur d'Armijo [11]. Dans un algorithme itératif minimisant une fonction Ψ, en supposant qu'une direction d^k est obtenue à l'itéré x^k, au lieu de mettre à jour $x^{k+1} = x^k + d^k$, on calcule le plus petit entier i tel que

$$\Psi(x^k + 2^{-i}d^k) \leqslant \Psi(x^k) + \beta 2^{-i}\nabla\Psi(x^k)^\mathsf{T}d^k, \quad \text{puis } x^{k+1} = x^k + 2^{-i}d^k, \tag{2.26}$$

où $\beta \in (0, 1/2)$ est une constante fixée. Beaucoup des articles ci-dessous utilisent une recherche linéaire modifiée, une variante "non monotone" introduite par Grippo, Lampariello et Lucidi [113]. Bien qu'initialement conçue pour des problèmes génériques, elle est néanmoins applicable. Son concept est plutôt simple : au lieu d'utiliser $\Psi(x^k)$ dans le membre de droite de (2.26), on utilise $\max\{\Psi(x^k), \Psi(x^{k-1}, \ldots, \Psi(x^{k-m(k)}))\}$ où $m(k)$ est une fonction vérifiant, selon l'article original, $m(0) = 0$ et $0 \leqslant m(k) \leqslant \min(M, 1 + m(k-1))$; en bref, les valeurs des derniers itérés sont considérées au lieu de seulement l'itéré courant x^k. Cela empêche l'algorithme d'avoir une propriété de descente trop forte, ce qui peut conduire à des pas trop petits, c'est-à-dire des i grands dans (2.26) (voir [65, p. 431, dernier paragraphe]).

L'autre technique principale que nous évoquons est la méthode classique de Levenberg-Marquardt, pour laquelle on présente un cas simple. En supposant qu'on veut minimiser $f = ||H||^2/2$ avec H C^1, on utilise les notations suivantes

$$J(x) := H'(x), \quad g(x) := \nabla f(x) = J(x)^\mathsf{T}H(x).$$

Une direction de descente est obtenue à partir de x en résolvant

$$(J(x)^\mathsf{T}J(x) + \lambda S)d = -J(x)^\mathsf{T}H(x)$$

où $\lambda \geqslant 0$ est le paramètre à adapter et $S \succ 0$ est une matrice symétrique définie positive (souvent $S = I$ l'identité). Cette méthode de globalisation a pour différence principale avec la recherche linéaire que le paramètre λ définit une *courbe* de solutions, contrairement à une recherche linéaire le long d'une demi-droite. Une itération de Levenberg-Marquardt se présente par exemple de la façon suivante [105, algorithme 19.10].

Algorithme 2.3.26 (itération de Levenberg-Marquardt)**.** L'algorithme utilise les constantes suivantes : $0 < \tau_1 < 1 < \tau_2$ pour la mise à jour de λ et $0 < \kappa_1 < \kappa_2 < 1$ comme seuils de satisfaction de la décroissance de f.

1. *Critère d'arrêt.* Si $g(x) \simeq 0$, arrêt de l'algorithme.
2. *Déplacement.* Prendre $\lambda_0 = \lambda$ et répéter les opérations suivantes pour $i \in \mathbb{N}$.

   2.1. Calculer la solution d_i du système linéaire

   $$(J(x)^\mathsf{T} J(x) + \lambda_i S) d_i = -J(x)^\mathsf{T} H(x).$$

   2.2 Si

   $$f(x + d_i) \leqslant f(x) + \kappa_1 g(x)^\mathsf{T} d_i,$$

   sortir de la boucle courante avec $d = d_i$, sinon $\lambda_{i+1} = \tau_2 \lambda_i$.
3. *Nouveau facteur de pénalisation.* Si $f(x + d) \leqslant f(x) + \kappa_2 g(x)^\mathsf{T} d$, $\lambda_+ = \tau_1 \lambda_i$, sinon $\lambda_+ = \lambda_i$.
4. *Nouvel itéré.* $x_+ = x + d$.
5. *Nouvelle matrice.* Choisir $S_+ \succ 0$.

Une règle de mise à jour possible est de la forme $\lambda_k = \kappa ||H(x^k)||^\delta$ pour une constante $\kappa > 0$ et $\delta \in [0, 2]$. La méthode a été initialement présentée par Levenberg [153] et plus tard redécouverte par Marquardt [165]. D'autres options comme la pénalisation sont considérées par exemple dans [183].

Contributions sans C-fonctions

Ici, nous discutons de certains articles qui considèrent des équations non lisses $H(x) = 0$ ne provenant pas nécessairement de PC et C-fonctions. Kojima et Shindo [146] discutent le cas des applications PC^1, pour par morceaux C^1, c'est-à-dire que H est sélectionnée parmi un nombre fini de fonctions C^1 H_i, qui peuvent inclure des PC et des applications normales. Leur algorithme choisit essentiellement un morceau correspondant à x^k et applique un pas de Newton avec ce morceau. L'algorithme converge si, à une solution z, pour tous les morceaux contenant z, les jacobiennes en z $H_i'(z)$ sont inversibles et les dérivées H_i' sont lipschitziennes autour de z. Ils considèrent également une variante quasi-Newton avec la mise à jour de Broyden, en stockant une approximation sur chaque morceau.

Discutons plus en détail le cas de (2.25). Qi et Sun [206] utilisent une méthode de Newton non lisse de la forme

$$x^{k+1} = x^k - V_k^{-1} H(x^k), \quad V_k \in \partial H(x^k). \tag{2.27}$$

Puisque H est localement lipschitzienne, ∂H est bien défini. Une propriété clé assurant la convergence locale est le fait que la non-singularité est une propriété "diffusante".

Proposition 2.3.27 ([206, proposition 3.1], [204, lemma 2.6])**.** *Soit x tel que chaque $V \in \partial H(x)$ est inversible. Alors il existe $C > 0$ et un voisinage $N(x)$ de x tels que pour tout $y \in N(x)$ et $W \in \partial H(y)$, W est inversible et $||W^{-1}|| \leqslant C$.* □

Cette propriété est essentielle puisque, si x^* est une solution vérifiant les hypothèses de la proposition 2.3.27, alors dans un voisinage de x^* l'équation de type Newton (2.25)

peut être résolue, donc l'algorithme est bien défini. Si en plus la fonction est semi-lisse, l'algorithme converge localement. Une propriété de convergence globale est montrée sous les hypothèses que la proposition 2.3.27 et plusieurs inégalités de type Lipschitz sont toutes vérifiées globalement.

Dans [204], Qi remplace ∂ par ∂_B (toujours sous l'hypothèse que H est lipschitzienne et semi-lisse). L'intérêt principal de ce changement est de réduire la "taille" du différentiel considéré, ce qui rend donc l'hypothèse de BD-régularité forte ("toutes les jacobiennes à la solution sont inversibles", définition 2.3.17) plus faible et plus facilement vérifiable. C'est une des motivations principales de cette thèse, en particulier puisque l'article traite de (2.1a) comme une application.

Avec Pang [198], ils conçoivent une méthode pour montrer que la convergence superlinéaire est valable pour une itération de la forme (2.27), ce qui s'applique également pour des itérations plus générales n'utilisant pas de différentiel. Elle nécessite également la semi-lissité et la BD-régularité forte (voir la définition 2.3.17). Ils discutent également des techniques de globalisation dans des idées similaires à celles discutées ci-dessous dans la section 2.3.4.

Maintenant, nous mentionnons une approche développée par Śmietański [233], qui combine (2.25) et une technique dans le cas unidimensionnel

$$x_{k+1} = x_k \exp\left(-\frac{f(x_k)}{x_k f'(x_k)}\right).$$

Observons que cette équation se réduit à celle de Newton en prenant $e^t \simeq 1 + t$. Cela devient, dans le cadre n-dimensionnel,

$$H(x^k) + V_k d = 0, \qquad x_i^{k+1} = \exp(d_i/x_i^k)x_i^k, i \in [1:n]$$

où $V_k \in \partial_B H(x^k)$. Elle converge sous les mêmes hypothèses que la méthode classique sans la mise à jour exponentielle, la semi-lissité de l'application et la BD-régularité forte (définition 2.3.14) à la solution.

Enfin, mentionnons un algorithme en trois sous-étapes par Solodov et Svaiter [235], centré sur $\mathrm{PCN}(F)$ avec F monotone. Chaque itération est composée de : une résolution inexacte de la linéarisation régularisée, puis une recherche linéaire sur la direction obtenue, et enfin une projection. Leur algorithme n'utilise pas de fonction de mérite particulière mais plutôt la quantité $(F(x + \cdot), \cdot - x)$ issue de la formulation variationnelle des PC. Le coût principal de chaque itération est un PCL, bien qu'aucune solution exacte ne soit requise. L'algorithme est globalement convergent sous les hypothèses que $\nabla F(x^*) \succ 0$, et que ∇F vérifie une régularité supplémentaire autour de x^*. Ce travail est basé sur un article similaire des mêmes auteurs [234], fournissant en particulier une illustration (p. 767) de l'intuition sous-jacente à l'algorithme.

Reformulation avec le minimum

Définition 2.3.28 (C-fonction min)**.** La C-fonction min est notée et définie par

$$\varphi_{\min}(a, b) := \min(a, b). \tag{2.28}$$

Elle est parfois écrite comme $\min(a,b) = a - (a-b)_+ = (a+b-|a-b|)/2$. □

Puisque le minimum est une C-fonction, pour résoudre des PC on peut considérer

$$H(x) = \min(x, Mx+q) \quad \text{ou} \quad H(x) = \min(x, F(x)) \quad \text{ou} \quad H(x) = \min(F(x), G(x)).$$

Le minimum a été utilisé par de nombreux auteurs [3, 195, 192, 204, 207, 95, 66, 20, 73]. C'est la C-fonction la moins différentiable mais elle est linéaire par morceaux dans ses arguments, et est aussi souvent appelée "résidu naturel" ([243, p. 202], [138, p. 116], [5, p. 1]). Ce nom semble venir du fait que le minimum sert souvent de critère d'erreur ou d'arrêt même pour d'autres C-fonctions.

Il nous semble que la fonction de Fischer (et d'autres C-fonctions inspirées par elle) ont été étudiées plus en profondeur que le minimum, en particulier puisque son carré (donc la fonction de mérite associée) est différentiable. Ceci est particulièrement pertinent pour la globalisation des algorithmes locaux. Néanmoins, certaines de ses propriétés comme la terminaison finie pour les PCL (voir ci-dessous) ou la présence du minimum dans les critères d'erreur sont des motivations particulières pour étudier le minimum ci-dessous (voir aussi la section 2.3.4) et dans les chapitres suivants.

Commençons par [194], où Pang énonce d'abord un critère d'erreur

$$||x - x^*|| \leqslant \kappa ||\min(x, Mx+q)|| \tag{2.29}$$

autour de x^* lorsque c'est une solution régulière, ce qui est énoncé en termes de PCL perturbés autour de x^*. Il est aussi prouvé (lemme 3 p. 59) que, pour $\mathrm{PCN}(F)$, l'application $\min(x, F(x))$ est lipschitzienne avec une constante $\sqrt{n}\max(1,\alpha)$ où α est la constante de Lipschitz de F. Un schéma approximatif, utilisant une linéarisation de F, est justifié convergent sous hypothèse de régularité de la solution x^*. En 1990, Pang [195] utilise la notion de B-dérivée (ou dérivée directionnelle, grâce à [230]) pour considérer l'équation $H(x) = 0$ et sa fonction de mérite $||H||^2/2$, résolue par l'itération

$$H(x^k) + H'(x^k; d^k) = 0, \quad x^{k+1} = x^k + d^k. \tag{2.30}$$

En supposant que l'application H est différentiable à la solution et que la dérivée est inversible et "forte" (c'est-à-dire plus régulière), une convergence locale est obtenue, et une convergence globale en appliquant une recherche linéaire. Cependant, en utilisant (2.30), soit la fonction H est différentiable en x^k et cela se réduit à un pas de Newton traditionnel, soit l'itération devient un PCL mixte puisque l'équation n'est pas linéaire en d. Notons que, lorsqu'elle est appliquée aux PCN, les sous-problèmes sont des PCL avec moins d'indices de complémentarité (comparer par exemple avec l'approche de Josephy [134] évoquée dans la section 2.2.4). Ainsi, soit des hypothèses supplémentaires sont nécessaires pour assurer une résolution simple des itérations, soit on doit accepter de résoudre des PCL mixtes.

Le travail de Pang conduit à l'algorithme appelé "Newton-min" pour résoudre l'équation non lisse $H(x) = \min(x, F(x)) = 0$. À chaque itération, il consiste à résoudre une version simplifiée de (2.30) en divisant les indices pour lesquels $x_i = F_i(x)$, qui (peuvent) causer que H soit non différentiable, dans la partie x ou la partie F du système, pour obtenir un système linéaire à résoudre. Un aperçu est donné ci-dessous.

Algorithme 2.3.29 (NEWTON-MIN)**.** Considérer un point de départ $x^0 \in \mathbb{R}^n$.

1. *Critère d'arrêt.* Si $H(x^k) = 0$, arrêter.
2. *Décomposition des indices.* Définir les ensembles d'indices suivants :

$$\begin{aligned}\alpha(x^k) &:= \{i \in \mathbb{R}^n : F_i(x^k) < x_i^k\}\\ \beta(x^k) &:= \{i \in \mathbb{R}^n : F_i(x^k) = x_i^k\}\\ \gamma(x^k) &:= \{i \in \mathbb{R}^n : F_i(x^k) > x_i^k\}.\end{aligned}$$

   Soit $(I^k, J^k) \in \mathfrak{B}(\beta(x^k))$ une partition de β^k, poser $\overline{\alpha}^k := \alpha(x^k) \cup I$ et $\overline{\gamma}^k := \gamma(x^k) \cup J$. Résoudre le système linéaire avec variable d.

$$\begin{cases} x^k_{\overline{\gamma}^k} + d_{\overline{\gamma}^k} = 0 \\ F(x^k)_{\overline{\alpha}^k} + F'(x^k)_{\overline{\alpha}^k} d = 0 \end{cases} \tag{2.31}$$

3. *Mise à jour.* Poser $x^{k+1} = x^k + d^k$ avec d^k la solution de (2.31).

Clairement, (2.31) peut être réduit à un système avec $|\overline{\gamma}^k|$ variables par substitution directe de $d_{\overline{\gamma}^k}$. Sa convergence locale est liée à des questions de régularité discutées ci-dessous, et la classe de matrices M pour lesquelles l'algorithme Newton-min converge a été étudiée par Ben Gharbia et Gilbert dans [23, 24]. Bien que ce soit discuté plus en détail dans le chapitre 6, mentionnons que le choix des indices dans $\beta(x^k)$ n'est pas anodin : des choix "incorrects" peuvent conduire à $||H(x^k + d^k)|| > ||H(x^k)||$, voir [20, exemple 5.8].

Une propriété particulière de l'algorithme 2.3.29 basé sur la reformulation avec le minimum a été observée pour la première fois par Fischer et Kanzow [95] : pour les PCL avec une hypothèse de régularité appropriée, on obtient une terminaison finie. Selon [86, p. 853], une telle propriété ne peut pas être obtenue pour la C-fonction de Fischer. Plus précisément, soit x^* une solution de (2.1c), et définissons

$$\begin{aligned}\alpha^* &:= \{i \in [1:n] : x_i^* > 0 = (Mx^* + q)_i\},\\ \beta^* &:= \{i \in [1:n] : x_i^* = 0 = (Mx^* + q)_i\},\\ \gamma^* &:= \{i \in [1:n] : x_i^* = 0 < (Mx^* + q)_i\}.\end{aligned}$$

Si pour toute bipartition $(I, J) \in \mathfrak{B}(\beta^*)$, la matrice

$$\begin{bmatrix} M_{:,\alpha^* \cup I} \\ I_{:,\gamma^* \cup J} \end{bmatrix}$$

est inversible, le nombre d'itérations est fini (voir aussi [86, théorème 7.2.18 p. 660]). Ils considèrent deux régularités pour le PCL.

Définition 2.3.30 (point régulier pour $\mathbf{PCL}(\boldsymbol{M}, \boldsymbol{q})$ [95, définition 2 p. 281])**.** *Un point x^* est dit b-régulier si $M_{\delta,\delta}$ est inversible pour $\alpha^* \subseteq \delta \subseteq \alpha^* \cup \beta^*$. Il est dit R-régulier (pour Robinson [220]) si*

$$\det(M_{\alpha^*,\alpha^*}) \neq 0 \text{ et } M_{\beta^*,\beta^*} - M_{\beta^*,\alpha^*} M^{-1}_{\alpha^*,\alpha^*} M_{\alpha^*,\beta^*} \in \mathbf{P}. \qquad \square$$

condition	propriété	référence (dans [95])
$M \in \mathbf{P}$ (pour tout x)	chaque $V \in \partial_C H(x)$ est inversible	théorème 5, p. 284
$M \in \mathbf{ND}$ (pour tout x)	chaque $V \in \partial_B H(x)$ est inversible	théorème 6, p. 285
solution x^* R-régulière	chaque $V \in \partial_C H(x^*)$ est inversible	théorème 9, p. 286
solution x^* b-régulière	chaque $V \in \partial_B H(x^*)$ est inversible	théorème 10, p. 287

TABLE 2.1 – Résumé des propriétés de régularité pour le PCL.

Bien que ces propriétés soient pertinentes à une solution x^*, certaines propriétés plus générales sont valables en tout point ; elles sont résumées dans le tableau ci-contre.

Ceci est lié au travail de Qi [204] sur la méthode de Newton semi-lisse avec le B-différentiel, puisqu'elle nécessite des hypothèses plus faibles sur le différentiel de l'application H à la solution, car il y a moins d'éléments dedans, que son homologue avec le différentiel de Clarke. De même, on mentionne que le B-différentiel de la fonction de Fischer contient plus d'éléments que celui du minimum [132, p. 151].

La notion de point régulier (pas celle de la définition 2.3.12) pour $x \mapsto \min(x, F(x))$ est définie comme suit [195, définition 2 p. 327], [87, définition 2.1 p. 230].

Définition 2.3.31 (point régulier pour $\mathbf{PCN}(\boldsymbol{F})$)**.** Soit $x \in \mathbb{R}^n$, définissons $\alpha := \{i \in [1:n] : F_i(x) < x_i\}$ et $\beta := \{i \in [1:n] : x_i = F_i(x)\}$. Un point x est appelé vecteur b-régulier de $\min(x, F(x))$ si, pour $\delta \subseteq \beta$, la matrice $\nabla F(x)_{\alpha\cup\delta,\alpha\cup\delta}$ est inversible. Il est dit un vecteur R-régulier si $[\nabla F(x)]_{\alpha,\alpha}$ est inversible et son complément de Schur dans $\alpha \cup \beta$, $\nabla F(x)_{\beta,\beta} - \nabla F(x)_{\beta,\alpha}[\nabla F(x)_{\alpha,\alpha}]^{-1}\nabla F_{\alpha,\beta}$ est une $\mathbf{P}$-matrice (voir définition 2.2.1). □

D'autres types de régularité, appelés semistabilité et hémirégularité, existent aussi [30]. Numériquement, Pang [195] suggère, pour réduire le coût de calcul, d'utiliser une méthode inexacte ou une approche par moindres-carrés. Pang adapte aussi ce schéma aux inégalités variationnelles dans [192], en utilisant le minimum composante par composante sur les "conditions d'optimalité" associées. Voir aussi la section 2.3.4 pour une discussion plus spécifique sur certaines contributions. Ces résultats ont été améliorés dans l'article précédemment mentionné de Qi et Sun [206, p. 362, avant-dernier paragraphe].

L'approximation basée sur le différentiel fut approfondie dans [204], où le B-différentiel de la définition 2.3.15 remplace le différentiel de Clarke. Sous la BD-régularité forte de la définition 2.3.17, *n'importe quel* élément du B-différentiel suffit pour l'algorithme, c'est pourquoi dans les "Final Remarks" ([204, p. 243]), un seul élément est calculé. Dans [255], un élément spécifique de ∂H est utilisé. Cette question est développée plus en détail dans le chapitre 6. Qi dérive aussi un résultat de convergence sans semi-lissité et BD-régularité forte en remplaçant ces hypothèses par des conditions de descente suffisante (théorème 5.1). Un algorithme mélangeant la méthode basée sur le différentiel et la méthode basée sur le B-différentiel, donc nommé "hybride", est montré comme globalement convergent.

2.3.4 Traitement particulier du minimum

Cette section vise à discuter d'une méthode particulière ou d'un type de méthodes apparaissant avec la reformulation avec le minimum. Nous l'avons observée dans des articles de Pang [192], puis Han, Pang et Rangaraj [119, 197] ou Pang et Gabriel [196], Qi et Sun [207], ainsi que le livre de Facchinei et Pang [86]. Elle est aussi discutée en détail dans [72], qui est un des points de départ de cette thèse. Pour la présentation, nous considérons le problème $\mathrm{PCN}(F)$: rappelons que lorsqu'on utilise la C-fonction minimum (définition 2.3.28), souvent trois sous-ensembles de $[1:n]$ (nous omettons "$i \in [1:n]$" ci-dessous) d'indices apparaissent en un point donné x :

$$\alpha(x) := \{i : F_i(x) < x_i\}, \quad \beta(x) := \{i : F_i(x) = x_i\}, \quad \gamma(x) := \{i : F_i(x) > x_i\}.$$

En particulier $H(x) = \min(x, F(x))$ est différentiable en x lorsque $F'(x)_{\beta(x),:} = I_{\beta(x),:}$ ou $\beta(x) = \varnothing$. Cependant, observons que $F_i(x) = x_i > 0$ viole la complémentarité mais $F_i(x) = x_i < 0$ viole la complémentarité *et* la non-négativité, ce qui peut expliquer un traitement différent des deux ensembles. Une remarque similaire peut être faite pour les indices dans $\alpha(x)$ et $\gamma(x)$. Bien que Pang considère des inégalités variationnelles, nous énonçons son formalisme adapté aux PC, où des problèmes étroitement liés sont évoqués dans [192, section 4, p. 109]. D'abord, définissons

$$\alpha_-(x) := \{i : F_i(x) < x_i < 0\} \quad \text{et} \quad \gamma_-(x) := \{i : x_i < F_i(x) < 0\},$$

puis regroupons ces sous-ensembles de $\alpha(x)$ et $\gamma(x)$ avec $\beta(x)$:

$$\overline{\alpha}(x) := \alpha(x) \backslash \alpha_-(x), \quad \overline{\beta}(x) := \beta(x) \cup \alpha_-(x) \cup \gamma_-(x), \quad \overline{\gamma}(x) := \gamma(x) \backslash \gamma_-(x).$$

Dans [72], les auteurs utilisent une idée similaire, sauf que $\alpha_-(x)$ et $\gamma_-(x)$ prennent seulement les indices pour lesquels $F_i(x)$ et x_i sont proches (avec les mêmes inégalités). Dans l'article de Pang [192], cela transforme certaines égalités du sous-problème en inégalités, pour éviter de se "reduce to just one single system of linear equations" (p. 110 ligne 4). Pang définit ensuite des conditions suffisantes pour la régularité, qui s'énoncent, dans ce cadre, $\nabla F(x)_{\alpha_+,\alpha_+}$ inversible et

$$\nabla F(x)_{\tilde{\beta},\tilde{\beta}} - \nabla F(x)_{\tilde{\beta},\alpha_+}[\nabla F(x)_{\alpha_+,\alpha_+}]^{-1}\nabla F(x)_{\alpha_+,\tilde{\beta}} \ \in \mathbf{P}$$

où $\alpha_+ := \{i \in [1:n] : x_i > F_i(x), x_i > 0\}$ et $\tilde{\beta}(x) := \{i \in [1:n] : x_i \leqslant 0, F_i(x) \leqslant 0\}$. Lorsqu'on compare à la définition 2.3.31, on voit que $\alpha(x)$ est devenu $\alpha_+(x)$, et $\beta(x)$ a augmenté avec une partie de $\alpha(x)$ et une partie de $\gamma(x)$, cette dernière étant absente de la définition initiale de la R-régularité.

Pang, Han et Rangaraj [197] considèrent un problème général de minimisation non lisse $\min f(x)$, avec des sous-problèmes de la forme

$$\min_d \ \psi(x^k, d) + \frac{1}{2} d^\intercal B_k d$$

où ψ remplace le terme du premier ordre (puisque le gradient peut ne pas être défini) et B_k approxime le terme du second ordre. Ils obtiennent une convergence globale vers un point

stationnaire au sens de Dini sous des hypothèses techniques sur ψ (et $\alpha I \preceq B_k \preceq \beta I$ pour $0 < \alpha \leqslant \beta$). Détaillons maintenant leur choix de ψ lors de la minimisation de $||H||^2/2$. C'est assez similaire à ce qui a été détaillé pour [192], qui utilise des ensembles d'indices très similaires (en enlevant la dépendance en x).

$$\alpha_{0+} := \left\{ i : \begin{array}{l} x_i > F_i(x) \\ x_i \geqslant 0 \end{array} \right\}, \quad \gamma_{0+} := \left\{ i : \begin{array}{l} F_i(x) > x_i \\ F_i(x) \geqslant 0 \end{array} \right\}, \quad \hat{\beta} := [1:n] \backslash (\alpha_{0+} \cup \gamma_{0+}).$$

Ensuite, ils définissent

$$\psi(x,d) = H(x)^\mathsf{T} G(x), \quad G_i(x) = \left\{ \begin{array}{ll} d_i & \text{si } i \in \gamma_{0+}, \\ \nabla F_i(x)^\mathsf{T} d & \text{si } i \in \alpha_{0+}, \\ \min(d_i, \nabla F_i(x)^\mathsf{T} d) & \text{si } i \in \hat{\beta}, \end{array} \right.$$

ce qui est une modification de $(||H||^2/2)'(x,d)$ en prenant la formule du minimum sur un ensemble plus grand d'indices. Ils obtiennent une convergence globale vers une solution si le point limite est régulier au sens de la définition 2.3.31. Peu après, les mêmes auteurs [119] discutent de ce cadre pour des équations non lisses, et proposent aussi un schéma où le terme du second ordre peut être défini comme souhaité.

L'article de Pang et Gabriel [196] présente une méthode où la contrainte $x \geqslant 0$ est ajoutée, ce qui simplifie les différents ensembles d'indices intervenants. En particulier, pour les indices i où $x_i = F_i(x)$, le sous-problème minimise la quantité $(x_i + d_i)^2/2$. Bien que la contrainte supplémentaire rende les notions de régularité légèrement plus compliquées, elle conduit à des sous-problèmes quadratiques convexes. La méthode bénéficie de fortes propriétés de convergence, observées sur de multiples exemples numériques, sous deux types de régularité.

Le livre de Facchinei et Pang [86] évoque aussi cette technique dans un contexte plus simple, modifiant seulement les termes avec indices dans $\beta(x)$ [86, section 8.3, p. 767]. Ils proposent d'utiliser, comme majorant du terme du premier ordre,

$$\begin{aligned} &\sum_{i \in \alpha(x)} F_i(x) F_i'(x) d + \sum_{i \in \gamma(x)} x_i d_i + \sum_{i \in \beta(x)} \max(x_i d_i, F_i(x) F_i'(x) d) \\ = &\sum_{i \in \alpha(x)} F_i(x) F_i'(x) d + \sum_{i \in \gamma(x)} x_i d_i + \sum_{x_i = F_i(x) \geqslant 0} H_i(x) \max(d_i, F_i'(x) d) \\ &\qquad + \sum_{x_i = F_i(x) < 0} H_i(x) \min(d_i, F_i'(x) d). \end{aligned}$$

où le second $\max$ est devenu un $\min$ puisque $H_i(x) = x_i = F_i(x) < 0$. Cette modification des termes d'"égalité positive" rend en fait le problème convexe (une observation similaire est faite dans [197, p. 72, ligne 23], où il est dit que la convexité est vérifiée s'il n'y a pas de tels indices). Nous discuterons plus en détail ce point dans le chapitre 6.

Qi et Sun [207] considèrent un problème plus général de minimisation d'une fonction non lisse par des sous-problèmes de région de confiance approchés. Pour $\mathrm{PCN}(F)$, ils utilisent la même approximation du premier ordre que décrite ci-dessus. En particulier, l'algorithme résout des sous-problèmes quadratiques par morceaux convexes bornés. Proposons une explication de leur résultat lorsqu'il est appliqué à $\mathrm{PCN}(F)$. En effet, pour que leur

algorithme trouve un point Dini-stationnaire (voir section 2.3.7 et chapitre 6), l'approximation ψ doit vérifier $\liminf \psi(x, td)/t \leqslant f^D(x; d)$ ce qui n'est clairement pas vérifié pour ce qui précède, puisque ψ est positivement homogène en d et a été construite comme un majorant de f^D, avec $f = ||H||^2/2$. Par conséquent, ils ne peuvent pas montrer directement que les points d'accumulation de leur algorithme sont Dini-stationnaires (ce qui est le cas sous l'hypothèse sur ψ). Ils montrent que le résultat est valable sous l'hypothèse que $\nabla F(x^*)_{\alpha,\alpha}$ est inversible et qu'une certaine matrice (plus compliquée que le complément de Schur habituel) est une **Q**-matrice.

Enfin, dans [72], le travail précurseur au chapitre 6, les auteurs discutent des améliorations des directions de type Newton qui peuvent être utilisées dans la méthode Newton-min, algorithme 2.3.29. Rappelons que celle-ci résout le système linéaire

$$\begin{cases} x^k_{\overline{\gamma}^k} + d_{\overline{\gamma}^k} = 0 \\ F(x^k)_{\overline{\alpha}^k} + F'(x^k)_{\overline{\alpha}^k} d = 0 \end{cases}$$

où $\alpha \subseteq \overline{\alpha}$ et $\gamma \subseteq \overline{\gamma}$ avec $(\overline{\alpha}, \overline{\gamma})$ une partition de $[1 : n]$. Les auteurs mentionnent une observation de [20] où, pour une répartition inappropriée des indices, la direction obtenue n'est pas une direction de descente pour $||H||^2/2$, ce qui vient de sa non-lissité. La question de choisir les bons indices (qui est liée à la question de choisir la bonne jacobienne dans un algorithme basé sur le différentiel) est discutée plus en détail dans le chapitre 6. Ils proposent des modifications du système ci-dessus, en remplaçant certaines égalités par des paires d'inégalités : soit $\mathcal{E}_\mathcal{F} \cup \mathcal{E}_\mathcal{G}$ une partition de $\mathcal{E}(x) := \{i \in [1 : n] : F_i(x) = G_i(x)\}$ et $\mathcal{E}^-(x) := \{i \in [1 : n] : F_i(x) = G_i(x) < 0\}$, on cherche d tel que

$$\begin{cases} F_i(x) + F_i'(x)d = 0 & \text{si } i \in \mathcal{F} \cup \mathcal{E}_\mathcal{F}, \\ G_i(x) + G_i'(x)d = 0 & \text{si } i \in \mathcal{G} \cup \mathcal{E}_\mathcal{G}, \\ F_i(x) + F_i'(x)d \geqslant 0 & \text{si } i \in \mathcal{E}^-(x), \\ G_i(x) + G_i'(x)d \geqslant 0 & \text{si } i \in \mathcal{E}^-(x). \end{cases} \tag{2.32}$$

Ce changement particulier est fait seulement pour les indices correspondant aux "plis négatifs", c'est-à-dire tels que $x_i = F_i(x) < 0$. Cette méthode assure, lorsque le polyèdre (puisque les équations viennent de linéarisations, on a des contraintes affines à vérifier) est non vide, que la direction obtenue est une direction de descente. Il est cependant envisageable que les systèmes puissent ne pas avoir de solutions, par exemple en raison du nombre d'(in)équations qui devient plus grand que la dimension. Cette transformation d'"une égalité en deux inégalités" est aussi observée dans [196]. Une version du système utilisant deux inégalités pour les indices tels que $|F_i(x) - G_i(x)| < \tau$ pour un $\tau \in \mathbb{R}^*_+$ est aussi utilisée, ce qui évite d'employer des égalités, par exemple pour l'aspect numérique.

Pour assurer que les systèmes polyédraux obtenus ont des solutions, certaines hypothèses de régularité doivent être faites autour de la solution. En particulier, une qualification de contrainte de type Mangasarian-Fromovitz doit être vérifiée pour toute partition possible qui peut exister autour du point concerné, ce qui est susceptible d'être difficile à vérifier sans une hypothèse globale sur les fonctions. Celle-ci considère toutes les partitions possibles $\mathcal{E}_\mathcal{F} \cup \mathcal{E}_\mathcal{G}$ de $\mathcal{E}(x) = \{i \in [1 : n] : F_i(x) = G_i(x)\}$ où x est proche d'un point régulier $\overline{x}$,

pour lesquelles il faut

$$\sum_{i\in\mathcal{F}\cup\mathcal{E}_\mathcal{F}} \alpha_i \nabla F_i(\overline{x}) + \sum_{i\in\mathcal{G}\cup\mathcal{E}_\mathcal{G}} \beta_i \nabla G_i(\overline{x}) + \sum_{i\in\mathcal{E}^-(x)} [\alpha_i \nabla F_i(\overline{x}) + \beta_i \nabla G(\overline{x})] = 0$$
$$\text{et } (\alpha_{\mathcal{E}^-(x)}, \beta_{\mathcal{E}^-(x)}) \geqslant 0 \text{ implique que } (\alpha, \beta) = 0.$$

En utilisant une recherche linéaire, cette méthode bénéficie d'une convergence globale, et en combinant la procédure polyédrique avec la direction classique de Newton-min (lorsque celle-ci appropriée), on s'assure d'obtenir les bonnes propriétés de convergence locale. Numériquement, l'algorithme montre de très bonnes performances, y compris sur des problèmes aléatoires ou provenant d'applications.

Ce qui est discuté dans le chapitre 6 est une tentative de contourner ces hypothèses difficiles, en remplaçant le système polyédrique par un problème d'optimisation de moindres-carrés. Ceux-ci ont toujours une solution (pas besoin d'hypothèses), mais ont naturellement un coût plus élevé.

2.3.5 Autres méthodes non lisses

Aperçu des contributions pour la fonction de Fischer

Fischer a initialement introduit sa fonction pour l'optimisation sous contraintes d'inégalités dans [92]. Cela résulte en une légère différence de convention de signe puisque les multiplicateurs et les contraintes ont des signes opposés alors que dans les PC les deux quantités ont le même signe. Burmeister est mentionné pour avoir défini la fonction comme une "distance à l'[orthant non négatif]". Nous utiliserons les deux noms ou seulement Fischer dans le reste de la thèse.

Définition 2.3.32 (C-fonction de Fischer [92])**.** La C-fonction de Fischer (Fischer-Burmeister) est notée et définie par

$$\varphi_{FB}(a, b) := \sqrt{a^2 + b^2} - (a + b). \tag{2.33}$$

□

Elle peut s'exprimer différemment, par exemple pour éviter la perte de précision numérique lors de la soustraction de quantités positives [179, p. 306]. L'intérêt principal de la fonction de Fischer est que, bien que non différentiable partout, elle concentre la non-différentiabilité à l'origine, dans le sens où elle est différentiable en dehors de l'origine, mais son différentiel à l'origine contient plus d'éléments que le minimum par exemple ([132, p. 151]). Néanmoins, son carré est différentiable partout, ce qui signifie qu'une fonction de mérite associée à φ_{FB} est différentiable, ce qui conduit à des algorithmes plus simples. La contribution initiale de Fischer [92] détaille les préoccupations algorithmiques. Cette fonction est "équivalente" à la fonction minimum dans le sens suivant [86, lemme 9.1.3 p. 798]

$$\frac{2}{2+\sqrt{2}} |\min(a, b)| \leqslant |\varphi_{FB}(a, b)| \leqslant (2 + \sqrt{2}) |\min(a, b)|.$$

Le différentiel de la fonction de Fischer, lorsqu'elle est appliquée à la complémentarité, a été identifié dans un article de Facchinei et Soares [87, section 3, pp. 232-236] (bien que l'article de Fischer discute aussi en partie de celui-ci). L'application $H(x) = 0$ avec la C-fonction φ_{FB} de la définition 2.3.32 est semi-lisse et la fonction de mérite est lisse partout. Elle est aussi SC^1 si F (chaque F_i) est SC^1. De plus, quand x est R(resp. b)-régulier (définition 2.3.31), toutes les $J \in \partial_C H(x)$ (resp. $\partial_B H(x)$) sont inversibles et sont de la forme (à une transposition près)

$$[\mathrm{Diag}(a(x)) - I] + \nabla F(x)[\mathrm{Diag}(b(x)) - I]$$

où les vecteurs $a(x)$ et $b(x)$ sont définis par

$$a_i(x) = \frac{x_i}{\sqrt{x_i^2 + F_i(x)^2}}, \qquad b_i(x) = \frac{F_i(x)}{\sqrt{x_i^2 + F_i(x)^2}}$$

tant que $(x_i, F_i(x)) \neq 0$; lorsque $x_i = 0 = F_i(x)$, $(a_i(x), b_i(x)) = (\xi_i, \rho_i)$ pour toute paire (ξ_i, ρ_i) avec $\sqrt{\xi_i^2 + \rho_i^2} \leqslant 1$.

De plus (section 4, pp. 236-237), si F est une $\mathbf{P}_0$- fonction (déf 2.2.4), les points stationnaires de $\Psi := ||H_{FB}||^2/2$ sont des solutions et si c'est une P-fonction uniforme (déf 2.2.4), les ensembles de niveaux de Ψ sont bornés. Utilisant ces propriétés, ils conçoivent un algorithme hybride qui calcule une direction soit par une approche Newton-min soit en utilisant $-\nabla\Psi$, qui est bien définie et une direction de descente puisque la fonction est lisse. Ensuite, une procédure de recherche linéaire peut être utilisée.

Fischer considère aussi le $\mathrm{PCN}(F)$ avec F Lipschitz semi-lisse mais monotone dans [93], et développe deux algorithmes. Le premier est une méthode basée sur la descente qui évite de calculer la dérivée de F et converge sous les hypothèses habituelles liées à la fonction de Fischer. Le second est une méthode de type Newton (inspirée des travaux de Qi et Sun dans les sections précédentes), et converge vers une solution x^* sous une condition de régularité sur F en x^* (légèrement différente des précédentes).

Le cas fortement monotone a par exemple été analysé par Geiger et Kanzow dans [103], où une méthode basée sur la descente avec recherche linéaire est présentée ; en particulier, elle ne calcule que les dérivées de la fonction de Fischer mais pas les dérivées de F elle-même, bien que F doit avoir une jacobienne Lipschitz. Quand F est seulement monotone, ils utilisent une méthode de recherche linéaire avec BFGS pour manipulation du terme d'ordre 2.

Facchinei, De Luca et Kanzow [65] proposent un algorithme non lisse de "type Qi" qui calcule une direction en résolvant une équation de type Newton avec un élément de $\partial_B H(x^k)$ ou, s'il n'est pas satisfaisant, utilise $-\nabla\Psi$. Cet algorithme génère des points stationnaires de Ψ (puisqu'il est lisse) et si un point d'accumulation est fortement BD-régulier et F est SC^1, la suite complète converge vers lui, en utilisant la direction de type Newton avec pas unitaire, superlinéairement (quadratiquement si $\nabla^2 F$ est Lipschitz). Notons que le critère d'arrêt utilisé est $||\min(x, F(x))||$ et non une quantité directement basée sur φ_{FB}. Les expériences numériques rapportées indiquent que la direction alternative $-\nabla\Psi(x^k)$ se produit rarement. Dans [85], Facchinei et Kanzow considèrent une version inexacte de Levenberg-Marquardt (utilisant aussi l'opposé du gradient si nécessaire) avec des propriétés de convergence plutôt similaires.

En 2000, De Luca, Facchinei et Kanzow [66] comparent plusieurs algorithmes. Chacun calcule d'abord une direction à partir d'une équation de type Newton ; si aucune solution ne peut être trouvée ou si une propriété de descente n'est pas vérifiée, on revient à $-\nabla\Psi(x^k)$. Ils comparent, en donnant un théorème de convergence pour chacun,

- $H_{FB}(x^k) + J_k d = 0, J_k \in \partial H_{FB}(x^k)$, une variante exacte et une variante inexacte de Levenberg-Marquardt (qui a toujours une solution) ;
- $H_{\min}(x^k) + J_k d = 0, J_k \in \partial H_{\min}(x^k)$, une variante exacte et une variante inexacte de Levenberg-Marquardt (qui a toujours une solution).

Avant de lancer les algorithmes, ils utilisent aussi une minimisation grossière de Ψ_{FB} par projections du gradient sur l'orthant non négatif, ce qui augmente la robustesse de l'algorithme ([66, p. 195]) surtout pour les algorithmes exacts, qui peuvent sous-performer autrement. Ils observent que l'algorithme exact basé sur le minimum nécessite moins de temps d'exécution que la version exacte basée sur FB ([66, p. 199]), ce qui est dû aux systèmes linéaires plus simples produits par le minimum. Cependant, cela peut s'inverser si les évaluations de fonction étaient coûteuses ([66, section 5.3, p. 202]) car les versions basées sur le minimum nécessitent plus d'évaluations de fonction pendant la recherche linéaire : comme elle utilise Ψ_{FB}, la reformulation de Fischer est mieux adaptée que le minimum.

Les C-fonctions $\varphi_{\min}$ et φ_{FB} ont aussi été comparées par Pieraccini, Gasparo et Pasquali dans [200]. Ils rapportent des résultats sur plusieurs algorithmes utilisant $\varphi_{\min}$ et/ou φ_{FB}. Leurs observations sont les suivantes : quand tous les algorithmes convergent, ceux utilisant le minimum tendent à nécessiter moins d'itérations (p. 379), sauf pour les problèmes hautement singuliers, où le manque de lissité du minimum était coûteux (p. 380). Ils mentionnent aussi que la fonction de mérite FB peut manquer d'efficacité (p. 381).

Liao, Qi et Qi [154] développent un cadre minimisant l'"énergie" représentée par Ψ_{FB} à travers des EDOs et des réseaux de neurones. Un travail similaire utilisant la fonction minimum peut être trouvé dans [5].

Dans les parties suivantes, nous discutons des contributions et résultats sur certaines C-fonctions existantes, bien que la fonction de Fischer reviendra quand nous évoquerons certaines techniques de lissage dans la section 2.3.6.

La C-fonction de Kanzow-Kleinmichel

Dans [137], Kanzow et Kleinmichel proposent une famille de C-fonctions.

Définition 2.3.33 (C-fonction Kanzow-Kleinmichel [137])**.** La C-fonction Kanzow-Kleinmichel est notée et définie par

$$\varphi_{KK}(a,b) := \sqrt{(a-b)^2 + \lambda ab} - (a+b) \quad \text{ou} \quad \tilde{\varphi}_{KK}(a,b) = \sqrt{a^2+b^2+\delta ab} - (a+b). \tag{2.34}$$

avec $\lambda \in (0,4)$ ou $\delta \in (-2,+2)$. □

Pour $\lambda = 2$ ($\delta = 0$), on obtient la fonction FB, et pour $\lambda \to 0$ ($\delta \to -2$), on retrouve la fonction minimum à un facteur -2 près. Elle bénéficie clairement de propriétés

de différentiabilité similaires à φ_{FB}, en particulier elle est fortement semi-lisse quand F est différentiable avec dérivée localement Lipschitz. Une surestimation de son différentiel de Clarke est naturellement dérivée de celui de la fonction FB, ainsi que des propriétés de régularité. On a le critère d'erreur suivant [137, lemme 3.6 p. 241]

$$(1 - \lambda/4)|\min(a, b)| \leqslant \varphi_{KK}(a, b) \leqslant (2 + \sqrt{2})|\min(a, b)|,$$

ce qui, combiné avec (2.29), signifie que la fonction de mérite de φ_{KK} majore $||x - x^*||^2$ quand F est une P-fonction uniforme. Une méthode de type Newton non lisse est proposée, avec des propriétés plutôt similaires à celles de la fonction FB. Ils rapportent que $\lambda = 2$ (Fischer) donne une meilleure convergence globale et $\lambda \simeq 0$ une meilleure convergence locale, et proposent une version avec un λ mis à jour pour combiner les bénéfices des deux principales C-fonctions.

La C-fonction de Chen-Chen-Kanzow

La C-fonction suivante a été introduite par Chen, Chen et Kanzow [43].

Définition 2.3.34 (C-fonction C-C-K [43])**.** La C-fonction Chen-Chen-Kanzow est notée et définie par

$$\varphi_{CCK}(a, b) := \lambda\varphi_{FB}(a, b) - (1 - \lambda)a_+b_+, \qquad \lambda \in (0, 1). \tag{2.35}$$

□

Conçue pour éviter la platitude de φ_{FB} dans l'orthant positif (voir ci-dessous pour un commentaire sur la fonction de Mangasarian-Solodov), la formule pour son gradient généralisé sont facilement obtenue à partir de celle de φ_{FB}, bien que non différentiable sur l'ensemble $\{(a, b) \in \mathbb{R}^2_+ : ab = 0\}$. Ses propriétés de régularité sont similaires aux reformulations de Fischer-Burmeister ou Kanzow-Kleinmichel.

Les auteurs utilisent un schéma de recherche linéaire non monotone pour la globalisation. Les points d'accumulation de l'algorithme sont stationnaires puisque la fonction de mérite est différentiable, et avec des hypothèses supplémentaires classiques, l'algorithme converge vers des solutions. Malgré le choix de $\lambda = 0.95$, donc une fonction assez proche de la fonction Fischer-Burmeister, ils observent des résultats particulièrement bons, probablement dus à la pertinence d'améliorer le paysage et les ensembles de niveaux de la fonction [p. 25 du rapport détaillé de l'article] : "In fact, we believe that our new NCP function is close to be an *optimal* C-function.".

Les C-fonctions de Sun-Qi

Dans une veine similaire, Sun et Qi [243] ont proposé les C-fonctions suivantes.

Définition 2.3.35 (Variantes de Sun-Qi de la C-fonction FB [243])**.** Les C-fonctions Sun-Qi

sont notées et définies par ($\alpha > 0$)

$$\begin{array}{ll} \varphi_{SQ1}(a,b) := \sqrt{[\varphi_{FB}(a,b)]_+^2 + \alpha(ab)_+^2} & \varphi_{SQ2}(a,b) := \sqrt{[\varphi_{FB}(a,b)]^2 + \alpha(a_+b_+)^2} \\ \varphi_{SQ3}(a,b) := \sqrt{[\varphi_{FB}(a,b)]^2 + \alpha(ab)_+^4} & \varphi_{SQ4}(a,b) := \sqrt{[\varphi_{FB}(a,b)]^2 + \alpha(ab)_+^2}. \end{array} \tag{2.36}$$

□

Ces quatre variantes de la fonction FB partagent la semi-lissité, les fonctions de mérite différentiables et les conditions pour les ensembles de sous-niveaux bornés ou les solutions aux points stationnaires. Sun et Qi comparent les quatre fonctions avec un algorithme utilisant la recherche linéaire : les variantes ne performent pas significativement différemment.

La C-fonction de Fukushima

Cette C-fonction, initialement conçue (seule la fonction de mérite associée) pour les inégalités variationnelles, est aussi lisse. Grâce à cette propriété, Fukushima [98] développe une méthode de descente globalement convergente.

Définition 2.3.36 (C-fonction écart de Fukushima)**.** La C-fonction Fukushima est notée et définie, sur l'ensemble $\{(a,b) : a \geqslant 0\}$, par

$$\varphi_F(a,b) = ab + \frac{\alpha}{2}[(a - b/\alpha)_+^2 - a^2], \qquad \alpha > 0, \tag{2.37}$$

□

Le lagrangien implicite (C-fonction Mangasarian-Solodov)

Bien qu'initialement pas énoncée comme une C-fonction, le lagrangien implicite de Mangasarian et Solodov [162, 161] est défini comme suit, généralisant le travail de Fukushima.

Définition 2.3.37 (C-fonction M-S [162])**.** La C-fonction Mangasarian-Solodov est notée et définie, pour $\alpha > 1$, par (notons que contrairement à la C-fonction précédente (2.37), aucune hypothèse sur a n'est nécessaire)

$$\begin{aligned} M(x,\alpha) &:= x^\mathsf{T} F(x) + \frac{1}{2\alpha}\left(||(x - \alpha F(x))_+||^2 - ||x||^2 + ||(F(x) - \alpha x)_+||^2 - ||F(x)||^2\right) \\ \varphi_{MS}(a,b)^2 &:= ab + \frac{1}{2\alpha}[(a - \alpha b)_+^2 - a^2 + (b - \alpha a)_+^2 - b^2]. \end{aligned} \tag{2.38}$$

En particulier, elle est différentiable quand F l'est, a un gradient nul aux solutions, et, sous une condition de type qualification à une solution non dégénérée, son minimiseur est un minimum global localement unique, en plus d'être plus net autour d'eux (voir figure 9.1 p. 797 dans [86]). Cela conduit en particulier à la possibilité d'une méthode de Newton standard sur $\nabla M = 0$. Yamashita et Fukushima [256] montrent que ∇F définie positive implique que les points stationnaires de $M(\cdot, \alpha)$ sont des solutions. Quand F est μ-fortement

monotone et F a un gradient Lipschitz, ils proposent une méthode de type descente. En particulier, les dérivées de F ne sont pas utilisées dans la méthode, ce qui la rend sans dérivée, et est montrée comme convergeant linéairement quand l'hypothèse de Lipschitz est valide autour de l'itéré initial et son ensemble de sous-niveaux.

De plus, le lagrangien implicite bénéficie des critères d'erreur suivants, d'abord exprimés par Mangasarian, Solodov, Luo et Ren [155] pour des x proches d'une solution x^* et une constante κ

$$\begin{aligned} \frac{\alpha-1}{\alpha}||\min(x,F(x))||^2 \leqslant M(x,\alpha) \leqslant (\alpha-1)||\min(x,F(x))||^2 \\ ||x-x^*||^2 \leqslant \frac{2\kappa^2\alpha}{\alpha-1}M(x,\alpha) \\ ||x-x^*||^2 \leqslant \frac{2\kappa^2\alpha}{\alpha-1}\max(1,||x||)M(x,\alpha) \end{aligned} \tag{2.39}$$

où le second critère est obtenu via le critère de Pang (2.29), et le troisième est une version globale pour tout x. Quand F est μ-fortement monotone et L-localement Lipschitz, on peut prendre $\kappa = (L+1)/\mu$; ceci est aussi exprimé par Yamashita et Fukushima.

Subséquemment, dans [161], les auteurs conçoivent une méthode qui ne dérive que la C-fonction mais pas la fonction du problème de complémentarité, et garantit la convergence quand la fonction F dans (2.1b) est fortement monotone.

Généralisation des fonctions de mérite

Mentionnons une contribution particulière de Kanzow [135]. Dans cet article, plusieurs C-fonctions (carrées) sont regroupées dans un cadre unique, incluant $\varphi_{\min}$, φ_{FB} et φ_{MS} (voir déf. 2.3.28, 2.3.32, 2.3.37). Un algorithme simple de type Newton basé sur la résolution de $\nabla\Psi(x) = 0$ pour une fonction de mérite Ψ est proposé. Il est bien défini, en particulier la hessienne $\nabla^2\Psi$ est montrée comme définie positive près des solutions. Cependant, ce cadre vient avec des hypothèses plutôt fortes : la complémentarité stricte / non-dégénérescence est vérifiée à la solution ainsi qu'une condition d'indépendance linéaire sur les gradients des composantes actives de x et $F(x)$ en plus de F ayant une hessienne Lipschitz. De telles hypothèses sont quelque peu attendues puisque le minimum (une C-fonction très non lisse) est aussi considéré.

Considérons maintenant une contribution de Tseng, Yamashita et Fukushima [246], traitant du CP général de (1.5). Ils reformulent un problème approché comme suit, où $\alpha < 1$

$$\begin{aligned} \tilde{m}_\alpha(x) := F(x)^\mathsf{T}G(x) - G(x)^\mathsf{T}\mathrm{P}_K(F(x)+\alpha G(x)) - F(x)^\mathsf{T}\mathrm{P}_{K^*}(G(x)+\alpha F(x)) \\ + \frac{1}{2\alpha}\left(||F(x)-\mathrm{P}_K(F(x)+\alpha G(x))||^2 + ||G(x)-\mathrm{P}_{K^*}(G(x)+\alpha F(x))||^2\right). \end{aligned}$$

Le cadre qu'ils développent généralise l'écart régularisé de Fukuhsima (voir ci-dessous) et le lagrangien implicite de Mangasarian et Solodov. Cette fonction bénéficie de multiples propriétés intéressantes comme la différentiabilité.

Méthodes spécifiques pour les inégalités variationnelles

Mentionnons quelques contributions centrées sur les inégalités variationnelles, problèmes avec la forme suivante : trouver un x^* tel que

$$x^* \in S, \qquad (F(x^*), x - x^*) \geqslant 0, \qquad \forall\, x \in S.$$

Fukushima [98] étudie la fonction suivante, reliée à la fonction de mérite appelée "écart régularisé" plus bas.

$$f_\alpha(x) = \max_{y \in S} \{(F(x), x - y) - \frac{\alpha}{2}||x - y||^2\}. \tag{2.40}$$

À partir de cette formulation, un algorithme de descente avec recherche linéaire est proposé, sous l'hypothèse que l'application F est monotone. Plus tard, avec Majig et Fukushima [158], sans l'hypothèse que F est monotone, proposent une méthode de type Josephy-Newton, résolvant l'inégalité variationnelle avec F linéarisée autour de l'itéré x^k. Même avec la linéarisation, le sous-problème reste difficile, donc ils restreignent le point variable dans l'inégalité variationnelle à une boule autour de x^k. De plus, une technique de recherche linéaire est employée ainsi qu'une projection sur l'ensemble S si la partie Josephy-Newton ne fournit pas une direction satisfaisante. La convergence repose sur la définie positivité de la jacobienne de F à la solution. Les auteurs proposent un autre algorithme, basé sur une technique évolutive, signifiant qu'une population de points candidats est étudiée et modifiée au fil des itérations.

À partir de l'écart régularisé de (2.40), on peut considérer l'"écart D(ifférence)" $f_\alpha(x) - f_\beta(x)$ pour $\alpha < \beta$. Il a par exemple été analysé par Sun, Fukushima et Qi [242], où ils proposent une approximation du second ordre pour cette fonction lisse qui n'est pas deux fois différentiable. Observons que l'écart régularisé de (2.40) nécessite une projection sur l'ensemble S, qui doit être un brin traitable. Quand $S = \{x : h(x) \leqslant 0\}$, une qualification de contrainte de rang constant (sur les gradients des contraintes actives) peut être utilisée pour analyser la projection. Ensuite, une méthode de type Newton généralisée est présentée, qui est basée sur les hypothèses que la qualification de contrainte est valide à la solution et que tous les éléments dans la hessienne généralisée sont inversibles. Elle est ensuite globalisée par des régions de confiance, qui converge vers une solution quand F est fortement monotone et soit Lipschitz soit S compact. De plus, quand les éléments de la hessienne généralisée sont définis positifs, la vitesse de convergence est superlinéaire/quadratique. L'écart D est utilisé par Fukushima et Kanzow [136] pour concevoir un algorithme hybride qui calcule une direction par une équation de type Newton ou revient à l'opposé du gradient pour vérifier une propriété de descente. Le cas particulier avec $\beta = 1/\alpha > 1$ a été étudié par Peng [199], qui dérive des bornes supérieures et inférieures sur la fonction de mérite.

Dans une approche similaire, Polak et Qi [202] introduisent la notion d'opérateur newtonien, qui vise à généraliser une dérivée seconde pour la fonction de mérite d'une inégalité variationnelle (ou une équation non lisse) quand celles-ci ne sont pas proprement deux fois différentiables. À partir de là, ils adaptent la méthode de Newton à leurs opérateurs newtoniens ; la convergence est obtenue en supposant que tous les éléments dans l'opérateur à la solution sont inversibles.

L'article d'Ito et Kunisch [131] discute aussi d'équations non lisses issues d'inégalités variationnelles, surtout avec des contraintes de borne :

$$l \leqslant x^* \leqslant u, \qquad (F(x^*), x - x^*) \geqslant 0 \qquad \forall\, x \in [l, u].$$

Observons que l'hypothèse A.4 p. 351 est que la fonction de mérite est "régulière sous-différentiellement" ou régulière au sens de Clarke (définition 2.3.12) ; rappelons que le minimum (composante par composante) ne vérifie pas cette propriété, ce qui est un problème par rapport aux illustrations qu'ils donnent (équations (1.3) et (1.5) p. 348). Ensuite, ils discutent d'une méthode basée sur des ensembles d'indices rappelant l'algorithme PDAS ou Newton-min 2.3.29. En effet, comme observé dans un article avec Hintermüller [124], les deux méthodes peuvent conduire aux mêmes itérés (section 2). Ils discutent aussi de la situation en dimension infinie, venant d'autres domaines comme les applications de contrôle optimal issues de problèmes d'obstacles.

Ce point de vue est aussi considéré par He et Yang [122], où F est T-monotone, i.e., $(F(v) - F(w), (v - w)_+) \geqslant 0$. Les problèmes qu'ils étudient, ayant des contraintes de borne, conduisent à un cadre similaire à ceux mentionnés ci-dessous dans la section 2.3.6, où ils utilisent des techniques de lissage.

Munson, Facchinei, Ferris, Fischer et Kanzow [179] discutent aussi d'inégalités variationnelles avec contraintes de borne, et proposent une reformulation avec deux C-fonctions, composante par composante, dépendant des valeurs de l_i et u_i :

$$\Phi_i(x) := \begin{cases} \phi_1(x_i - l_i, F_i(x)), & \text{si } l_i \in \mathbb{R}, u_i = +\infty, \\ -\phi_1(u_i - x_i, -F_i(x)), & \text{si } l_i = -\infty, u_i \in \mathbb{R}, \\ \phi_2(x_i - l_i, \phi_1(u_i - x_i, -F_i(x))), & \text{si } l_i \in \mathbb{R}, u_i \in \mathbb{R}, \\ -F_i(x), & \text{si } l_i = -\infty, u_i = +\infty. \end{cases}$$

où ϕ_1 et ϕ_2 sont des C-fonctions. Ils suggèrent deux options : $\phi_1 = \varphi_{FB} = \phi_2$ et $\phi_1 = \varphi_{CCK}$, $\phi_2 = \varphi_{FB}$. Ils discutent de nombreux aspects numériques comme les redémarrages, les perturbations régularisantes, les moindres carrés, le préconditionnement, les erreurs de précision, la recherche linéaire non monotone…et obtiennent de bonnes performances numériques.

D'autres notions en analyse non lisse

Pour conclure, mentionnons rapidement quelques articles utilisant d'autres notions d'analyse non lisse, comme celles introduites par Mordukhovich. Avec Hoheisel, Kanzow et Phan, dans [127], une adaptation de la méthode de Newton est analysée, où le sous-problème est de la forme $-H(x^k) \in DH(x^k)(d^k)$ où DH est une dérivée généralisée et la mise à jour est $x^{k+1} = x^k + d^k$.

Gfrerer et Outrata [104] définissent une version adaptée de la semi-lissité pour les ensembles et applications multivoques, afin de traiter des linéarisations des deux termes dans une équation généralisée $0 \in F(x) + T(x)$. Cela leur permet de mettre en place une méthode de type Newton.

2.3.6 Techniques de lissage

Lissage général

Comme discuté dans les sections précédentes, il existe de nombreuses méthodes pour traiter les formulations non lisses. Ici, nous mentionnons quelques références utilisant le lissage, qui, dans sa forme générale, introduit un paramètre positif supplémentaire ε (aussi appelé μ, τ,...), modifie l'équation non lisse (parfois directement la C-fonction utilisée) pour obtenir un système

$$\begin{pmatrix} \tilde{H}(x,\varepsilon) \\ h(\varepsilon) \end{pmatrix} = 0,$$

où $\tilde{H}(x,\varepsilon)$ est différentiable si $\varepsilon > 0$ et $h(\varepsilon) = 0 \Leftrightarrow \varepsilon = 0$ (le plus souvent, $h(\varepsilon) = \varepsilon$ et ε intervient comme ε^2 dans $\tilde{H}$). Observons que cela ne résout pas les problèmes d'inversibilité aux solutions, ce qui, comme mentionné avant, n'est pas souhaitable ([86, prop. 9.1.1, pp. 794-795]), mais évite la non-différentiabilité dans le reste de l'espace.

Par exemple, Chen, Qi et Sun [47] traitent de l'inégalité variationnelle suivante, où X est convexe fermé :

$$q(x) \in X, \quad (y - q(x))^{\mathsf{T}} p(x) \geqslant 0, \forall\, y \in X.$$

Ils se concentrent sur le cas $X := \{x : l \leqslant x \leqslant u\}$, ce qui conduit à la formulation non lisse

$$\begin{aligned} F(x) &:= q(x) - \operatorname{mid}(l, u, q(x) - p(x)) = 0, \\ \operatorname{mid}(a,b,c) &:= (\operatorname{mid}(a_i, b_i, c_i))_i = \min(b_i, \max(a_i, c_i)) = \mathrm{P}_{[a_i,b_i]}(c_i) \qquad a \leqslant b. \end{aligned}$$

Ils approchent directement F par $f(x,\varepsilon)$ tel que $||f(x,\varepsilon) - F(x)|| \leqslant \mu\varepsilon$, f vérifie la "propriété de consistance jacobienne" :

$$\lim_{\varepsilon} \operatorname{dist}(\nabla_x f(x,\varepsilon)^{\mathsf{T}}, \partial_\times F(x)) = 0,$$

ce qui signifie que l'approximation est aussi adaptée pour la dérivée, où $\partial_\times$ est le produit composant par composant défini dans la proposition 2.3.18. Le lissage est fait par une formule intégrale qui lisse l'opérateur mid. Ensuite, ils discutent d'une technique de recherche linéaire basée sur la globalisation, convergente quand $\nabla_x f$ est inversible et convergeant vers des solutions quand F est fortement régulière, i.e., toutes ses jacobiennes à la solution sont inversibles (définition 2.3.17).

Cette "séparation" de F en f lisse et $F - f$ non lisse petit a été utilisée par Chen dans [45], où le lissage vérifie la "propriété de consistance de la dérivée directionnelle"

$$\lim_{h} \frac{F(x+h) - F(x) - f^0(x+h)h}{||h||} = 0, \quad f^0(x) := \lim_{\varepsilon \searrow 0} \nabla_x f(x,\varepsilon).$$

En particulier ([45, lemme 2.3 p. 110]), pour les applications semi-lisses, la "consistance jacobienne" implique la "propriété de consistance de la dérivée directionnelle". Ce cadre est

utilisé pour développer une méthode quasi-Newton avec la règle de mise à jour de Broyden. L'autrice, avec Nashed et Qi [46] discute de problèmes similaires mais dans des espaces de Banach, où la "différentiabilité oblique" est utilisée pour adapter la semi-lissité, permettant de traiter la dimension infinie issue d'EDP non lisses par exemple.

Dans [243], Sun et Qi proposent un lissage de l'application normale de la forme $F(z_+) + z - z_+ + \alpha(z_+) \cdot [F(z_+)]_+$ pour $\alpha > 0$, qui possède de meilleurs ensembles de niveaux (théorème 5) que l'application usuelle $F(z_+) + z - z_+$. Leur fonction de lissage est donnée par

$$\psi(\mu, w) := \frac{w + \sqrt{w^2 + 4\mu^2}}{2}.$$

Ensuite, pour un paramètre de lissage $u \in \mathbb{R}^n$ (n variables au lieu d'une), étendre ψ à $\mathbb{R}^n$ par $\psi(u, z)_i = \psi(u_i, z_i)$. La fonction suivante est étudiée :

$$\Psi(u, z) := \begin{pmatrix} F(\psi(u, z)) + z - \psi(u, z) + \alpha\psi(u, z) \cdot \psi(u, F(\psi(u, z))) \\ u \end{pmatrix} = 0.$$

Cette application est (fortement) semi-lisse quand F' est localement Lipschitz et différentiable quand $u > 0$. Comme pour le cas d'une variable de lissage, les notions ou régularité s'appliquent aux matrices dans $\partial\Psi(0, z^*)$, bien que leur structure par blocs rende les conditions plus simples à vérifier.

Un autre type de cadre de lissage a été développé par Haddou et coauteurs. Il a été appliqué aux PCL [190], PCN [117], MPECs [116, 249, 248] (voir section 2.2.7) et équations de valeur absolue [1, 63] (voir section 2.2.6). Considérons ω une fonction scalaire croissante négative sur $\mathbb{R}_-$, égale à 0 en 0 et tendant vers 1 à $+\infty$. Des exemples incluent $\omega(t) = t/(t+1)$ ou $\omega(t) = 1 - e^{-t}$. Définissons $\omega_r(t) = \omega(t/r)$. La complémentarité (exprimée pour $\mathrm{PCN}(F)$ par exemple) est remplacée par l'équation lisse $\omega_r(x_i) + \omega_r(F_i(x)) = 1$. Des conditions supplémentaires sont requises pour s'assurer que cette reformulation (et les équivalentes) aient des propriétés adaptées, en particulier pour assurer la convergence de la méthode, qui peut employer par exemple des techniques similaires aux points intérieurs [190].

Lissage avec Levenberg-Marquardt

Maintenant, nous discutons de quelques références qui emploient des techniques de Levenberg-Marquardt pour la globalisation en plus des méthodes de lissage. Qi [205] décompose aussi H en "lisse + non lisse petit", séparant $H(x) = \min(F(x), G(x))$ en $H = p + q$, où, pour $w > 0$ fixé,

$$p_i(x) = \begin{cases} H_i(x) & |F_i(x) - G_i(x)| \geqslant w \\ \hat{p}_i(x) & |F_i(x) - G_i(x)| < w \end{cases} \quad \text{et} \quad q_i(x) = \begin{cases} 0 & |F_i(x) - G_i(x)| \geqslant w \\ \hat{q}_i(x) & |F_i(x) - G_i(x)| < w \end{cases}$$

où les expressions de $\hat{p}$ et $\hat{q}$ sont données par

$$\hat{p}_i(x) = \frac{G_i(x) + H_i(x)}{2} - \frac{1}{4w}((F_i(x) - G_i(x))^2 + w^2), \quad \hat{q}_i(x) = \frac{1}{4w}(|F_i(x) - G_i(x)| - w)^2.$$

En particulier, $|q_i(x)| \leqslant w/4$ et p est lisse. La dérivée de p est utilisée dans une méthode LM-région de confiance avec convergence globale sous l'hypothèse que q ne varie pas "trop vite".

Zhang et Chen [259] utilisent un lissage de la fonction Fischer-Burmeister de la forme $\varphi_\varepsilon(a,b) = a + b - \sqrt{a^2 + b^2 + 2\varepsilon^2}$ et résolvent la reformulation du PCL

$$\begin{pmatrix} Mx + q - y \\ \varphi_\varepsilon(x,y) \\ \varepsilon \end{pmatrix} = 0.$$

Puisque la fonction de mérite est lisse, les points d'accumulation sont stationnaires, et deviennent des solutions quand M est une $\mathbf{P}_0$-matrice. En supposant que la suite calculée est bornée et que l'algorithme converge vers une solution strictement complémentaire, la vitesse est quadratique (ou superlinéaire, selon la mise à jour LM). La seconde hypothèse peut être remplacée en supposant que l'inverse de l'application est uniformément borné. Dans une veine similaire, Zhang avec Zhang [260] traitent le cas de $\mathrm{PCN}(F)$.

Une variante du PCL est traitée par Tian, Yu et Yuan dans [244] : ils considèrent

$$x \geqslant 0, \qquad s \geqslant 0, \qquad x \cdot s = w, \qquad F(x,s,y) = 0$$

avec $w \geqslant 0$ donné et $F(x,s,y) = Px + Qs + Ry - a$. Ils utilisent une modification appropriée d'une C-fonction pour introduire un poids (pour w) : $\varphi^c(a,b) = 0 \Leftrightarrow a \geqslant 0, b \geqslant 0, ab = c$. Par exemple, on peut utiliser une des suivantes, où $c \geqslant 0$ et $\mathbb{Q} := \{3, 5, \ldots\}$.

$$\begin{aligned}
&\varphi^c(a,b) = (1+\varepsilon)(a+b) - \sqrt{(a+\varepsilon b)^2 + (\varepsilon a + b)^2 + 2c + 2\varepsilon^2}, && \\
&\varphi^c(a,b) = \sqrt{a^2 + b^2 - 2\theta ab + 2(1+\theta)c + 2\varepsilon} - a - b, && -1 < \theta \leqslant 1, \\
&\varphi^c(a,b) = (a+b)^q - \sqrt{a^2 + b^2 + (\tau - 2)ab + (4-\tau)c}^q, && 0 \leqslant \tau < 4, q \in \mathbb{Q}, \\
&\varphi^c(a,b) = (a+b)^q - \sqrt{\tau(a-b)^2 + (1-\tau)(a^2+b^2) + 2(1+c)\tau}^q, && 0 \leqslant \tau \leqslant 1, q \in \mathbb{Q}.
\end{aligned}$$

Leur algorithme, employant le quatrième choix, met à jour le paramètre LM comme un multiple de $||F(x^k, s^k, y^k)||^2$ et finit par faire une recherche linéaire.

Ma, Tang et Chen [157] discutent de $\mathrm{PCN}(F)$ en abordant la minimisation de

$$\frac{1}{2}\left[\sum_{i=1}^{n} (x_i F_i(x))^2 + (x_i)_-^2 + (F_i(x))_-^2\right].$$

Bien que $t \mapsto (t)_-^2$ soit lisse, elle n'est pas deux fois différentiable, donc un autre lissage est utilisé, qui est deux fois différentiable et a un gradient fortement semi-lisse :

$$\varphi(\varepsilon, t) = \begin{cases} 0 & t \geqslant \varepsilon \\ \frac{(\varepsilon - t)^3}{12\varepsilon} & |t| < \varepsilon \\ \frac{t^2}{2} + \frac{\varepsilon^2}{6} & t \leqslant -\varepsilon \end{cases}, \ \min\left[\sum_{i=1}^{n} \frac{(1+\varepsilon)[x_i F_i(x)]^2}{2} + \varphi(\varepsilon, x_i) + \varphi(\varepsilon, F_i(x))\right].$$

Au prix de quelques calculs, ils obtiennent un algorithme de Levenberg-Marquardt qui possède une convergence locale et globale en supposant que les ensembles de sous-niveaux sont compacts. Leurs règles de mise à jour utilisent le ratio fonction-modèle similaire aux méthodes de région de confiance et le paramètre LM est mis à jour comme la norme du gradient.

2.3.7 Un commentaire sur la complexité

Comme discuté dans les parties précédentes, les problèmes de complémentarité sont souvent formulés comme des systèmes d'équations (non lisses), qui peuvent être résolus par des méthodes de type Newton. Puisque les problèmes lisses peuvent déjà être difficiles, les problèmes non lisses sont en général également ardus. Dans l'optimisation non lisse, où la fonction coût f ou les contraintes sont non lisses, trouver des solutions est souvent trop exigeant. Cela conduit à la recherche de points stationnaires, c'est-à-dire des points vérifiant une certaine "condition d'optimalité", souvent de la forme $0 \in \partial f(x^*)$, où un certain différentiel est utilisé (principalement celui de la définition 2.3.9 dans la section 2.3.2). Beck et Hallak [19, p. 57] appellent cela "criticité", alors qu'ils nomment stationnarité "l'absence de directions réalisables". Énonçons ces propriétés.

Définition 2.3.38 (stationnarité)**.** Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction lipschitzienne, un point x est dit stationnaire si $0 \in \partial f(x) = \partial_C f(x)$. □

Dans la définition, le différentiel utilisé est celui de Clarke de la définition 2.3.9.

Définition 2.3.39 (stationnarité "forte")**.** Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction lipschitzienne, un point x est dit fortement stationnaire si pour tout $d \in \mathbb{R}^n$, $f'(x; d) \geqslant 0$. □

Elle est parfois appelée "stationnarité de Dini" (voir par exemple [197, p. 60]), puisque cette définition a un sens lorsque f' est définie, donc $f' = f^D$ ce qui explique la référence à Dini.

En combinant l'observation que $f^\circ \geqslant f'$, que $(\mathbb{R} \ni)0 = 0^\mathsf{T} d$, et la définition 2.3.9, la stationnarité forte implique la stationnarité. Il est assez clair que dans le cadre général de l'optimisation non lisse non convexe, trouver un point stationnaire n'est pas facile. Dans le chapitre 6, nous détaillons le fait que, dans le cas spécifique de la reformulation par le minimum (qui conduit à une équation non lisse et à de l'optimisation non lisse), même vérifier la stationnarité forte en un point peut être difficile. Cela peut être comparé au travail de Murty et Kabadi [182], où de nombreux problèmes quadratiques sont montrés NP-complets, et en particulier vérifier l'optimalité locale en optimisation lisse non convexe est également NP-complet. De plus, dans le cas non lisse également, les points stationnaires peuvent ne pas être des optima locaux.

Maintenant, évoquons quelques contributions qui discutent de questions de complexité dans un cadre d'optimisation lisse. Bien qu'elles concernent le cas lisse, ces travaux récents exposent des techniques et résultats intéressants. La première paire d'articles, par Carmon, Duchi, Hinder et Sidford [39, 40], discute de la difficulté d'obtenir un point ε-stationnaire, c'est-à-dire un x tel que $||\nabla f(x)|| \leqslant \varepsilon$. Ils obtiennent une borne inférieure sur la complexité dans le pire cas de tout algorithme de la forme $c_p \Delta L_p^{1/p} \varepsilon^{-\frac{p+1}{p}}$ où f est lipschitzienne jusqu'à l'ordre p avec constante L_p, Δ est une mesure de proximité de l'itéré initial, et c_p est une constante dépendant de p. En particulier, ils retrouvent pour la méthode du gradient $p = 1$ le taux en ε^{-2}, en $\varepsilon^{-3/2}$ pour la méthode de Newton $p = 2$, et la formule générale $\varepsilon^{-\frac{p+1}{p}}$ pour les méthodes d'ordre p. Essentiellement, l'idée est de partir de la fonction compliquée

de Nesterov [184],

$$f_{\text{Nesterov}}(x) = \frac{1}{2}(x_1 - 1)^2 + \sum_{i=1}^{n-1} \frac{1}{2}(x_i - x_{i+1})^2,$$

puis de modifier progressivement la fonction par des compositions successives pour rendre son paysage suffisamment tortueux. En particulier, la construction est telle que les algorithmes "découvrent" la valeur d'exactement une coordonnée de la solution à chaque itération. Par conséquent, en choisissant judicieusement la dimension de l'espace, la borne inférieure sur le nombre d'itérations peut être obtenue.

Ils présentent également le cas des méthodes du premier ordre [40], pour des fonctions pouvant avoir (non utilisées dans l'algorithme pour des raisons pratiques) des dérivées d'ordre supérieur lipschitziennes. Par exemple, lorsque la fonction a un hessien lipschitzien (d'ordre deux), l'ordre de convergence en précision ε est entre $\varepsilon^{-12/7}$ et $\varepsilon^{-7/4}$. Cet intervalle devient $[\varepsilon^{-8/5}, \varepsilon^{-5/3}]$ pour des dérivées lipschitziennes au-delà du second ordre. Plus de détails peuvent être trouvés dans les articles et leurs références.

Plus récemment, Carmon et Duchi ainsi que Arjevani, Foster, Srebro et Woodworth présentent des questions similaires pour l'optimisation stochastique [10].

Évoquons maintenant quelques contributions pour le cas non lisse. Jordan, Lin et Zampetakis [133] affirment que même obtenir un point stationnaire au sens de Clarke, c'est-à-dire $0 \in \partial f(x)$, n'est pas très réaliste. Cela conduit à un problème plus réalisable $\min\{||g|| : g \in \partial f(x)\} \leqslant \varepsilon$, qui est un analogue du cas lisse. Puisque cet objectif reste difficile, le différentiel de Clarke est remplacé par le différentiel δ-Goldstein, noté et défini par

$$\partial_\delta f(x) := \text{conv}(\cup_{y:||y-x||\leqslant\delta} \partial f(y)).$$

Introduit par Goldstein dans [107], il possède des propriétés similaires au différentiel de Clarke, et conduit également à une méthode de descente. Ils rapportent (voir les références citées) pour obtenir un point ε-approché dans le sous-différentiel δ-Goldstein, des vitesses de convergence de l'ordre : $O(\delta^{-1}\varepsilon^{-3})$ pour f différentiable directionnellement en utilisant la randomisation, $O(d^{3/2}\delta^{-1}\varepsilon^{-4})$ sans utiliser de gradients. Ils prouvent également que sans randomisation, les algorithmes ont toujours besoin d'un nombre d'itérations proportionnel à la dimension.

Dans le cas non lisse non convexe lipschitzien, leur tableau 1 regroupe une comparaison entre le cas déterministe et le cas randomisé, ainsi que des commentaires sur les algorithmes utilisant uniquement l'oracle du premier ordre. En particulier, les algorithmes déterministes utilisant uniquement le gradient peuvent ne pas converger. Les fonctions compliquées impliquées sont différentes de celles utilisées dans [39], impliquant des minima et/ou maxima. Ils obtiennent également une borne de complexité pour une méthode de lissage, utilisant un remplacement lisse du maximum :

$$\text{softmax}_a(z_1, z_2) = \frac{1}{a}\ln(\exp(az_1) + \exp(az_2)), \qquad a > 0,$$

qui est 1-lipschitzienne et a un gradient $a/2$-lipschitzien.

Le différentiel de Goldstein est également utilisé par Gebken dans [102], où les méthodes de descente sont analysées et la vitesse de convergence suivante est obtenue : si l'algorithme converge vers un minimum x^* vérifiant $f(x) \geqslant f(x^*) + c||x - x^*||^p$, alors

$$||x^k - x^*|| \leqslant C \max(\varepsilon_k^{1/p}, \delta_k^{1/(p-1)})$$

où ε_k et δ_k sont les constantes du différentiel de Goldstein à l'étape k. Des exemples simples et des expériences numériques sont présentés.

2.4 Sur l'aspect combinatoire

Cette section est organisée comme suit : d'abord, nous expliquons le lien principal avec le sujet de la complémentarité. Ensuite, nous mentionnons quelques références classiques et quelques notions importantes de ce domaine. Nous terminons par une brève mention des matroïdes orientés et certaines de leurs propriétés les plus basiques, suivie par des logiciels en géométrie discrète et combinatoire et enfin des algorithmes les plus proches de notre sujet, l'identification des chambres.

2.4.1 Relation avec les sujets précédents

Dans les sections précédentes, nous avons discuté des problèmes de complémentarité, avec un accent sur les méthodes non lisses, en particulier celles issues des reformulations utilisant des C-fonctions. Détaillons comment cela peut être lié, sous un certain angle, à un problème spécifique de géométrie discrète computationnelle. Nous résumons les éléments principaux, discutés plus en détail dans le chapitre 3.

- Le B-différentiel de l'application minimum composante par composante peut intervenir dans les algorithmes résolvant les problèmes de complémentarité.
- Le calcul du B-différentiel du minimum de deux fonctions affines est équivalent à identifier les chambres d'un arrangement centré d'hyperplans.
- Nous proposons une nouvelle approche du problème d'identification des chambres d'un arrangement, que nous appelons l'"approche duale", via le théorème d'alternative de Gordan.
- Cette méthode duale est basée sur les circuits du matroïde (orienté) sous-jacent à l'arrangement.

Comme nous le verrons, calculer entièrement le B-différentiel est difficile, puisqu'un nombre potentiellement exponentiel d'éléments doit être calculé. Rappelons que la méthode de type Newton non lisse ne nécessite qu'un élément du B-différentiel et pas tous, bien que l'algorithme correspondant doive supposer des hypothèses appropriées pour pouvoir utiliser n'importe quel élément et fonctionner néanmoins.

Cependant, les arrangements étant un domaine extrêmement développé en soi, des algorithmes efficaces sur les arrangements sont intéressants.

2.4.2 Références classiques

Dans leur énoncé plus général, lorsque les hyperplans ne s'intersectent pas tous en un point (commun), les arrangements représentent la manière dont les lignes divisent le plan ou les plans divisent l'espace (dimension 2 ou 3). Cette question très générale a été considérée dès le début du XIX^{ème} siècle, avec Steiner [239] et Roberts [215] ; par exemple, l'article d'Alexanderson et Wetzel [7] discute le cas de dimension 3. Voir aussi Schläfli [227] (posthume, parmi des contributions dans bien d'autres domaines), qui a contribué à l'hypothèse parfois supposée de *position générale*, qui s'énonce

"En dimension n, k hyperplans s'intersectent en un sous-espace de dimension $n-k$."

où une dimension négative signifie l'ensemble vide. Parfois, une configuration pas en position générale est dite dégénérée. En particulier, Schläfli a donné la borne de position générale pour une dimension arbitraire, qui énonce que, pour p hyperplans en dimension n, le nombre de chambres est majoré par

$$\sum_{i=0}^{i=n} \binom{p}{i}.$$

La position générale a également été utilisée par Buck dans [37] pour dériver le nombre d'objets de toute dimension (intersections d'hyperplans et/ou demi-espaces) dans un arrangement.

Depuis lors, le domaine s'est développé et a prospéré bien au-delà de ce qui peut être évoqué ici. Mentionnons quelques livres classiques, qui couvrent bien plus que ce qui peut être utile dans cette thèse : Crapo et Rota [61], où ils discutent également du cas des hyperplans dans un corps fini (voir [186] pour les nombres complexes), un travail plus récent de Stanley [237, 238] et sa partie introductive (basée sur un cours au MIT) [236]. Un autre classique est le livre d'Orlik et Terao [187]. Edelsbrunner a consacré un livre aux aspects algorithmiques [81], voir aussi le livre de De Loera, Rambau et Santos [64]. Aguiar et Mahajan [4] proposent un traitement plus récent un peu plus orienté vers la recherche actuelle. Voir aussi une synthèse sur le sujet par Halperin et Sharir [118].

2.4.3 Quelques outils spécifiques

Certaines contributions emploient une notion très utile et puissante mais complexe appelée polynôme caractéristique. Il est défini ainsi : pour un arrangement $\mathcal{A}(H_1, \ldots, H_p)$, soit $L(\mathcal{A})$ l'ensemble de toutes les intersections d'un nombre arbitraire d'hyperplans, on a

$$\chi_{\mathcal{A}}(t) = \sum_{E \in L(A)} \mu(E) t^{\dim(E)}, \tag{2.41}$$

où μ est la fonction de Möbius de l'arrangement, une fonction plutôt complexe définie récursivement sur l'ensemble $L(\mathcal{A})$. Cette définition récursive signifie qu'un calcul direct peut être difficile. Une des formules principales, due à Zaslavsky [257], énonce que le nombre

de chambres est égal à $(-1)^n \chi_{\mathcal{A}}(-1)$. Également discutée dans les livres mentionnés précédemment, puisqu'elle encode beaucoup d'informations sur l'arrangement, elle est par exemple utilisée par Athanasiadis dans [12] pour obtenir le nombre de chambres de plusieurs arrangements "classiques", en utilisant des raisonnements combinatoires astucieux. Le code de [35] calcule en fait le polynôme caractéristique, en utilisant des symétries sous-jacentes pour simplifier la charge computationnelle totale. Comme les auteurs expliquent (page 1357, troisième paragraphe), si on veut spécifiquement identifier les chambres, les options sont limitées.

Néanmoins, le polynôme caractéristique reste extrêmement utile pour analyser certains arrangements particuliers, lorsque les vecteurs normaux aux hyperplans sont de la forme $e_i - e_j$ pour $1 \leqslant i < j \leqslant p$ [12, 203].

Une technique dans le domaine de la combinatoire est le principe de suppression-restriction. Essentiellement utilisé dans des raisonnements par induction, il décompose la structure considérée en deux plus petites, l'une avec une dimension réduite. Il peut rappeller la formule de Pascal, puisqu'il s'écrit, sous une forme très brute,

$$\text{Problème}(p, n) \leftrightarrow \text{Problème}(p-1, n) + \text{Problème}(p-1, n-1).$$

Winder a appliqué ce principe au nombre de chambres dans [253], obtenant une formule en termes des dégénérescences des sous-ensembles d'hyperplans. Brysiewicz, Eble et Kühne [35] utilisent également ce principe dans leur algorithme.

Les techniques introduisant de l'aléatoire peuvent améliorer l'efficacité d'algorithmes [53, 52].

2.4.4 Matroïdes orientés (et circuits)

Les matroïdes et matroïdes orientés sont des notions abstraites qui généralisent l'(in)dépendance linéaire des vecteurs. Pour un ensemble donné d'objets, cela consiste à donner les sous-ensembles qui sont indépendants (ou autres, voir ci-dessous). Alors que l'(in)dépendance linéaire est claire avec des vecteurs dans $\mathbb{R}^n$, elle peut en fait être généralisée à des objets plus abstraits. Bien que certains des livres mentionnés précédemment discutent des matroïdes, puisqu'ils sont un concept très central en combinatoire, nous mentionnons deux livres plus "orientés" vers eux : celui d'Oxley [191] et le classique de Björner, Las Vergnas, Sturmfels, White et Ziegler [28]. Une revue récemment mise à jour sur le domaine peut être trouvée dans [264]. Toutes les définitions suivantes considèrent les notations du livre de Ziegler [263], qui est plus orienté vers la géométrie. Dans ce qui suit, un ensemble de vecteurs $V = \{v_1, \ldots, v_p\} \subseteq \mathbb{R}^n$ est identifié à la matrice $V = [v_1 \ \ldots \ v_p]$; $e_p \in \mathbb{R}^p$ est le vecteur de taille p composé de 1. Rappelons que le support supp désigne les indices des composantes non nulles.

Définition 2.4.1 ("vecteurs")**.** L'ensemble des vecteurs $\mathcal{V}(V)$ est composé des signes (composante par composante) des vecteurs de dépendances affines, c'est-à-dire,

$$\mathcal{V}(V) := \{\operatorname{sgn}(z) : z \in \mathbb{R}^p, Vz = 0, e_p^{\mathsf{T}} z = 0\}. \tag{2.42}$$

□

Définition 2.4.2 ("circuits"). *L'ensemble des circuits $\mathcal{C}(V)$ est composé des signes (composante par composante) des vecteurs de dépendances minimales, c'est-à-dire,*

$$\mathcal{C}(V) := \{\mathrm{sgn}(z) : z \in \mathbb{R}^p, Vz = 0, e_p^\mathsf{T} z = 0, [V; e^\mathsf{T}]_J \text{ injectif } \forall J \subsetneq \mathrm{supp}(z)\}. \tag{2.43}$$

□

Les deux définitions suivantes sont en quelque sorte duales des deux précédentes.

Définition 2.4.3 ("covecteurs"). L'ensemble des (signes des) covecteurs $\mathcal{V}^*(V)$ est composé des signes (composante par composante) des fonctions affines des vecteurs, c'est-à-dire,

$$\mathcal{V}^*(V) := \{\mathrm{sgn}(c^\mathsf{T} V - c_0 e_p^\mathsf{T}) : c \in \mathbb{R}^n, c_0 \in \mathbb{R}\}. \tag{2.44}$$

□

Définition 2.4.4 ("cocircuits"). L'ensemble des (signes des) cocircuits $\mathcal{C}^*(V)$ est composé des signes (composante par composante) des fonctions affines de support minimal des vecteurs, c'est-à-dire,

$$\mathcal{C}^*(V) := \{\mathrm{sgn}(c^\mathsf{T}[V; e_p^\mathsf{T}]) : c \in \mathbb{R}^{n+1}, \forall\, d \in \mathbb{R}^{n+1}, \mathrm{supp}(c^\mathsf{T}[V; e_p^\mathsf{T}]) \subseteq \mathrm{supp}(d^\mathsf{T}[V; e_p^\mathsf{T}])\}. \tag{2.45}$$

□

Définition 2.4.5 (matroïde orienté). Soit $V = \{v_1, \ldots, v_p\} \subseteq \mathbb{R}^n$ un ensemble de vecteurs avec $\mathrm{span}(V) = \mathbb{R}^n$. Le matroïde orienté de V est une structure donnée par l'ensemble des circuits, l'ensemble des vecteurs, l'ensemble des cocircuits et l'ensemble des covecteurs. □

En fait, chacune de ces quatre quantités peut déterminer les trois autres, ce qui signifie qu'une seule d'entre elles est nécessaire. Il est également possible d'utiliser les bases, qui sont les plus grands sous-ensembles indépendants, peuvent également être utilisées.

Définition 2.4.6 (bases). L'ensemble des bases $\mathcal{B}(V)$ est composé des plus grands sous-ensembles indépendants de V, c'est-à-dire,

$$\mathcal{B}(V) := \{B \subseteq [1:p] : V_{:,B} \text{ injectif}, \mathrm{null}(V_{:,B\cup\{i\}}) = 1\ \forall\, i \in [1:p]\backslash B\}. \tag{2.46}$$

□

Qu'ils soient orientés ou non, les matroïdes ont été étudiés sous de nombreux angles. Mentionnons quelques contributions liées à nos préoccupations. Minieka propose des algorithmes pour passer des bases aux circuits [172]. Dósa, Szalkai et Laflamme indiquent, sous des hypothèses très légères, les nombres maximaux et minimaux de circuits et de bases [70] (voir aussi les chapitres 3, 4, 5 et l'annexe A).

En général, les opérations sur les matroïdes tendent à avoir une complexité exponentielle. Néanmoins, des algorithmes en "temps polynomial incrémental" existent, c'est-à-dire que le coût pour passer d'un élément au suivant est polynomial (et souvent court). Un exemple peut être trouvé dans [229] et dans les algorithmes définis dans la section 2.4.6. Les contributions discutant de la complexité de divers problèmes incluent [141, 143, 166]. Les

matroïdes, avec les hyperplans, peuvent servir à des applications diverses comme l'analyse d'efficacité d'un ensemble de caméras [213].

Concluons cet aperçu des matroïdes orientés en énonçant qu'ils interviennent dans les chapitres 3 et 5 (et leurs compléments chapitres 4 et A). L'approche duale découle de la nature duale de deux théorèmes classiques d'analyse convexe, l'alternative de Gordan [108] et l'alternative de Motzkin [178] (bien que d'autres noms ou formulations puissent être plus appropriés).

2.4.5 Logiciels d'algèbre

Il existe de nombreuses options de logiciels dédiés à l'algèbre qui peuvent, parmi bien d'autres fonctionnalités, traiter les matroïdes et les arrangements. Le plus souvent, puisque les logiciels traitent bien plus que simplement obtenir les chambres, nous n'avons pas comparé avec des résultats numériques de ces logiciels ; de plus, comparer un code de recherche et un logiciel pleinement implémenté n'est pas toujours significatif. Voici une liste de quelques logiciels de ce domaine.

D'abord, SageMath [68] est un logiciel à large spectre qui peut traiter des matroïdes ou des arrangements, parmi bien d'autres fonctionnalités. La page principale peut être trouvée ici, et la documentation ici.

Ensuite, créé en 1992, Macaulay2 est un logiciel plus spécialisé mettant l'accent sur la recherche en géométrie et algèbre (page principale) [111]. Il possède des packages qui traitent les arrangements ou les matroïdes.

Le logiciel polymake [101] est plus orienté vers la géométrie, comme son nom le suggère. Il est capable de faire des calculs sur les matroïdes ; un package dédié aux hyperplans a été conçu par Kastner et Panizzut, documenté dans [140].

Pour se concentrer un peu plus sur les matroïdes, mentionnons le projet dédié de Kingan et Kingan, Oid [142], et un autre logiciel, par Rambau, TOPCOM (Triangulations Of Points Configurations and Oriented Matroids). Outre la page principale [214], il a récemment servi à traiter des problèmes spécifiques [212].

Enfin, le projet OSCAR est codé en Julia. Il agrège de nombreux packages et outils précédents (comme polymake par exemple). Il a été utilisé dans [35], ainsi que d'autres packages liés (voir les références citées), pour obtenir de nouveaux résultats combinatoires.

2.4.6 Algorithmes spécifiques pour identifier les chambres

Globalement, calculer les chambres d'un arrangement est #P-difficile [250, chapitre 6] [1]. Ce n'est pas très surprenant puisque la plupart des problèmes en combinatoire nécessitent

1. La référence discute d'un type plutôt spécifique d'arrangements, et peut être généralisé à tout arrangement ; sur un forum, Timothy Chow a confirmé cette complexité.

l'énumération d'une liste de taille (potentiellement) exponentielle. Néanmoins, de nombreux algorithmes ont été conçus pour la myriade de problèmes existants.

Par exemple, dans [83], Edelsbrunner, O'Rourke et Seidel présentent un algorithme avec une complexité théorique optimale pour construire un arrangement. Il est plutôt complexe, mais construit complètement l'arrangement, en ajoutant les hyperplans un par un de manière incrémentale. Il consiste en un balayage topologique de l'espace, une technique répandue qui scanne l'espace pour trouver des intersections (facilement puisqu'elles sont faites par algèbre linéaire) entre le plan de balayage et la construction courante. Leur algorithme élaboré est (page 361, deuxième paragraphe de la section 5) meilleur qu'un algorithme similaire précédent, celui de Bieri et Nef [27], qui utilise également un plan de balayage mais de manière inductive sur la dimension. Dans le plan, c'est-à-dire en dimension 2, au lieu d'une ligne (un hyperplan), on peut utiliser une ligne courbe [82] ; les auteurs donnent des applications et adaptent également la méthode pour être utilisée en dimension supérieure.

Un autre algorithme très intéressant utilise la propriété sous-jacente énonçant que le graphe des chambres est connexe (voir chapitre 3). Par conséquent, on peut se demander s'il existe un moyen d'explorer le graphe des chambres. Cependant, puisque les nœuds sont inconnus (nous voulons les identifier !), on a besoin d'un type spécial d'exploration. Une réponse à cette question est la recherche inverse, un algorithme introduit par Avis et Fukuda dans [13, 14]. L'idée est la suivante : considérer la chambre contenant l'origine, puis tester ses voisins potentiels en essayant d'aller de l'autre côté d'un hyperplan, puis appeler récursivement la procédure lorsqu'une chambre est trouvée. Parmi d'autres propriétés, cet algorithme peut être fait de manière à ne pas stocker toutes les chambres, donc nécessite une mémoire limitée (il est "compact"), est "à temps polynomial incrémental", ce qui signifie trouver une chambre supplémentaire se fait en temps polynomial, et sensible à la sortie, ce qui signifie que la complexité computationnelle est bornée supérieurement par un polynôme en le nombre de chambres. Sa nature le rend également facilement implémentable en parallèle. En améliorant l'une des composantes, Sleumer [232, 231] a diminué la complexité requise.

La recherche inverse apparaît également dans d'autres problèmes combinatoires, voir par exemple les applications aux triangulations, circuits et cocircuits dans [212]. Elle est liée à la méthode du simplexe dans le sens où l'algorithme passe d'un sommet à un autre (les sommets d'un polytope forment un graphe connexe) mais sans les calculer explicitement à l'avance. Voir [34] pour des exemples, ou [97] pour une application au calcul de la somme de polytopes. Une nouvelle approche, peut-être pas pour l'énumération des chambres mais proposant de trouver des chemins hamiltoniens, est proposée dans [170].

Le dernier algorithme que nous discutons a été conçu par Rada et Černý [208]. Il est plutôt simple : avec un hyperplan, l'espace est divisé en ses deux demi-espaces. Maintenant, considérons le deuxième hyperplan, vérifions si les deux demi-espaces du premier hyperplan sont eux-mêmes divisés en deux par le nouveau plan ajouté, et ainsi de suite. Les auteurs ont montré que, après une certaine reformulation, leur algorithme est meilleur que la recherche inverse (même avec l'accélération de Sleumer). Il bénéficie des mêmes propriétés, bien qu'ayant un champ d'application plus étroit. Il est discuté en détail, ainsi que certains réglages de celui-ci, dans les chapitres 3 et 5.

2.4.7 Quelques exemples d'applications

Terminons cette section sur les arrangements en évoquant quelques situations où des algorithmes calculant les chambres d'un arrangement peuvent être utiles. Quelques éléments supplémentaires sont discutés dans le chapitre 3 section 3. D'abord, il y a un usage "dual" : calculer les sommets de zonotopes, des polytopes particuliers qui interviendront dans le chapitre 6.

Définition 2.4.7 (zonotopes)**.** Soit $V = \{v_1, \ldots, v_p\} \subseteq \mathbb{R}^n$ une collection de vecteurs et soit $c \in \mathbb{R}^n$ un vecteur additionnel. Le zonotope défini par V et c est le polytope noté et défini par :

$$Z(V, c) := c + \sum_{i=1}^{p} v_i[-1, +1] = c + V[-1, +1]^p. \tag{2.47}$$

Dans (2.47), $v_i[-1, +1]$ est le segment $[-v_i, +v_i]$, V est la matrice $[v_1 \ \ldots \ v_p]$, $V[-1, +1]^p$ est l'expression compacte de la somme. Le zonotope $Z(V, c)$ est symétrique centralement autour de son centre c, ce qui s'écrit :

$$z \in Z(V, c) \qquad \Longleftrightarrow \qquad 2c - z \in Z(V, c). \qquad \square$$

Les zonotopes sont des objets étonnamment utiles dans certains domaines de recherche. Ils peuvent être utilisés en contrôle, où les états possibles du système peuvent former un zonotope, mis à jour d'un pas de temps à un autre avec les variations possibles du système. Puisqu'ils sont des images affines (souvent linéaires avec $c = 0$) d'hypercubes unitaires, ils possèdent des propriétés combinatoires plutôt intéressantes. Quelques éléments sont donnés en annexe B, sinon voir par exemple le livre de Grünbaum [114], celui de Ziegler [263] ou des articles dédiés, comme celui de McMullen [167]. Mentionnons brièvement la relation entre arrangements et zonotopes (voir corollaire 7.17, section 7.3, page 205 du livre de Ziegler avec des notations légèrement différentes ; nous nommons "faces" les faces de toutes dimensions).

Proposition 2.4.8 (lien entre zonotopes et arrangements)**.** *Soit $V \in \mathbb{R}^{n \times p}$ une matrice et $Z(0, V)$ le zonotope associé. Les vecteurs de signes des faces de $Z(0, V)$ sont en bijection avec les vecteurs de signes des objets dans l'arrangement formé par les vecteurs de V.* $\square$

Un usage surprenant des zonotopes peut par exemple être trouvé dans la maximisation convexe d'une quadratique convexe sur l'hypercube (en utilisant soit l'hypercube $\{0, 1\}$ ou $\{-1, +1\}$, qui ne diffèrent que par un changement affine de variables) sous une hypothèse de rang fixé, où la nature combinatoire peut être résolue par des zonotopes [89]. À la lumière de la proposition 2.4.8 ci-dessus, les applications des zonotopes sont ainsi liées aux arrangements.

Une autre application est l'estimation du rang en régression robuste, où la question d'ordonnancement des résidus conduit à une forme spécifique d'arrangements [42].

Les arrangements sont également liés aux fonctions de seuil, une notion entre informatique théorique et mathématiques : $f(x) = \operatorname{sgn}(a_0 + a^{\mathsf{T}} x)$, avec $x \in \{-1, +1\}^n$ pour un

certain $a_0 \in \mathbb{R}$ et $a \in \mathbb{R}^n$. Voir par exemple [253, 130]. Des considérations similaires sont traitées dans [251].

Dans [16] par exemple, les auteurs voient l'expression ci-dessus de f comme l'évaluation d'un neurone sur l'entrée x, ce qui relie notre sujet aux réseaux de neurones et à l'apprentissage profond, lorsque nous restreignons la fonction d'évaluation à être linéaire (pour une vue d'ensemble générale sur l'apprentissage profond lui-même, voir la célèbre synthèse de Schmidhuber [228]).

Certains types spécifiques d'arrangements ont également eu des articles dédiés, comme les arrangements avec des hyperplans de la forme $H_{ij} := \{x : x_i \pm x_j = \ldots\}$, voir par exemple [12, 203]. Les arrangements de *résonance*, avec des hyperplans de la forme $H_v := \{x : c^{\mathsf{T}} x = 0\}$ pour $c \in \{0, 1\}^n \backslash \{0\}$, sont liés à la théorie quantique [148, 35]. Pour plus de détails, voir les articles mentionnés et leurs références.

Chapitre 3

B-différentiel du minimum de deux fonctions vectorielles affines

Ce chapitre est constitué d'un article publié dans Mathematical Programming Computation [77]. Il décrit une question spécifique d'analyse non lisse : l'obtention du B-différentiel (voir définition 2.3.15) du minimum, composante par composante, de deux fonctions affines. Cette question provient de l'étude de l'algorithme de Newton-min 2.3.29, décrit à la section 2.3.3.

On y montre que cette question d'analyse non lisse en apparence assez restreinte est équivalente à moult autres problèmes, donc les arrangements (centrés) d'hyperplans. Ce lien avec les domaines de la combinatoire et de la géométrie computationnelle est utile pour mettre au point des algorithmes apportant une réponse numérique. Plusieurs améliorations sur un algorithme de l'état de l'art sont proposées et évaluées numériquement.

Ce chapitre va de pair avec le chapitre 4, qui contient des détails comme des preuves ou des commentaires supplémentaires, ainsi que des compléments sur des questions proches.

Par ailleurs, pour un souci d'harmonisation avec le reste, les références sont groupées avec celles de la thèse, et ne sont donc pas ajoutées à la fin de l'article (certaines dates des références comme les classiques de la littérature peuvent être différences de la version publiée). De même, la mise en page, les polices et tailles d'écriture sont différentes.

Note : l'Université de Sherbrooke demande que, pour les articles insérés, la contribution du doctorant soit précisée. Ce sujet était mentionné dans une liste de questions à considérer, que j'ai brièvement étudié avant que l'on ne travaille ensemble dessus. La majeure partie du travail a été réalisée de façon commune, en discutant fréquemment pour coordonner les contributions et points de vue, le code et la partie rédaction majoritairement faits par mes encadrants. Cet article publié a été rédigé de façon conjointe via un dépôt Git, de façon à laisser chacun contribuer.

On the B-differential of the componentwise minimum of two affine vector functions

Jean-Pierre Dussault [1], Jean Charles Gilbert [2] and Baptiste Plaquevent-Jourdain [3]

This paper focuses on the description and computation of the B-differential of the componentwise minimum of two affine vector functions. This issue arises in the reformulation of the linear complementarity problem with the Min C-function. The question has many equivalent formulations and we identify some of them in linear algebra, convex analysis and discrete geometry. These formulations are used to state some properties of the B-differential, like its symmetry, condition for its completeness, its connectivity, bounds on its cardinality, *etc.* The set to specify has a finite number of elements, which may grow exponentially with the range space dimension of the functions, so that its description is most often algorithmic. We first present an incremental-recursive approach avoiding to solve any optimization subproblem, unlike several previous approaches. It is based on the notion of matroid circuit and the related introduced concept of stem vector. Next, we propose modifications, adapted to the problem at stake, of an algorithm introduced by Rada and Černý in 2018 to determine the cells of an arrangement in the space of hyperplanes having a point in common. Measured in CPU time on the considered test-problems, the mean acceleration ratios of the proposed algorithms, with respect to the one of Rada and Černý, are in the range 15..31, and this speed-up can exceed 100, depending on the problem, the approach and the chosen linear optimization and matroid solvers.

Keywords : B-differential • Bipartition of a finite set • C-differential • Complementarity problem • Complexity • Componentwise minimum of functions • Connectivity • Dual approach • Gordan's alternative • Hyperplane arrangement • Matroid circuit • Pointed cone • Schläfli's bound • Stem vector • Strict linear inequalities • Symmetry • Winder's formula.

AMS Subject classification : 05A18, 05C40, 26A24, 26A27, 46N10, 47A50, 47A63, 49J52, 49N15, 52C35, 65Y20, 65K15, 90C33, 90C46.

3.1 Introduction

Let $\mathbb{E}$ and $\mathbb{F}$ be two real vector spaces of finite dimensions $n := \dim \mathbb{E}$ and $m := \dim \mathbb{F}$. The *B-differential* (B for Bouligand [218]) *at* $x \in \mathbb{E}$ of a function $H : \mathbb{E} \to \mathbb{F}$ is the set denoted and defined by

$$\partial_B H(x) := \{J \in \mathcal{L}(\mathbb{E}, \mathbb{F}) : H'(x_k) \to J \text{ for } \{x_k\} \subseteq \mathcal{D}_H \text{ converging to } x\}, \tag{3.1}$$

1. J.-P. Dussault, Département d'Informatique, Fac. des Sciences, Univ. de Sherbrooke, Québec, Canada
Jean-Pierre.Dussault@Usherbrooke.ca, ORCID 0000-0001-7253-7462

2. J.Ch. Gilbert, Inria Paris (Serena team), 48 rue Barrault, CS 61534, 75647 Paris Cedex, France
Département de Mathématiques, Fac. des Sciences, Univ. de Sherbrooke, Québec, Canada
Jean-Charles.Gilbert@inria.fr, ORCID 0000-0002-0375-4663

3. B. Plaquevent-Jourdain, Département de Mathématiques, Fac. des Sciences, Univ. de Sherbrooke, Québec, Canada, Inria Paris (Serena team), 48 rue Barrault, CS 61534, 75647 Paris Cedex, France
Baptiste.Plaquevent-Jourdain@USherbrooke.ca, ORCID 0000-0001-7055-4568

where $\mathcal{L}(\mathbb{E}, \mathbb{F})$ is the set of linear (continuous) maps from $\mathbb{E}$ to $\mathbb{F}$, $\{x_k\}$ denotes a sequence and $\mathcal{D}_H$ is the set of points at which H is (Fréchet) differentiable (its derivative at x is denoted by $H'(x)$). Recall that a locally Lipschitz continuous function is differentiable almost everywhere in the sense of the Lebesgue measure (Rademacher's theorem [209]) and this property has the consequence that the B-differential of a locally Lipschitz function is nonempty and bounded everywhere [51]. The B-differential is an intermediate set used to define the C-differential (C for Clarke [51]) of H at x, which is denoted and defined by

$$\partial_C H(x) := \operatorname{co} \partial_B H(x), \tag{3.2}$$

where $\operatorname{co} S$ denotes the convex hull of a set S [221, 126, 32]. Both intervene in the specification of conditions ensuring the local convergence of the semismooth Newton algorithm [206, 204, 233], which can be a motivation for being interested in that concept.

In this paper, we focus on the description of the B-differential of H at x when $H : \mathbb{R}^n \to \mathbb{R}^m$ is the componentwise minimum of two affine functions $x \mapsto Ax + a$ and $x \mapsto Bx + b$, where A, $B \in \mathbb{R}^{m\times n}$ and a, $b \in \mathbb{R}^m$. Hence, H is defined at x by

$$H(x) = \min(Ax + a, Bx + b), \tag{3.3}$$

where the minimum operator "min" acts componentwise (for two vectors $u, v \in \mathbb{R}^m$ and $i \in [1 : m] := \{1, \ldots, m\}$: $[\min(u, v)]_i := \min(u_i, v_i)$). This function is usually nonsmooth. A motivation to look at the B-differential of that function H comes from the fact that, when $m = n$ and H is given by (3.3), as explained below, the equation

$$H(x) = 0 \tag{3.4}$$

is a reformulation of the *balanced* [73] *Linear Complementarity Problem* (LCP)

$$0 \leqslant (Ax + a) \perp (Bx + b) \geqslant 0. \tag{3.5}$$

This system expresses the fact that a point $x \in \mathbb{R}^n$ is sought such that $Ax + a \geqslant 0$, $Bx+b \geqslant 0$ and $(Ax+a)^{\mathsf{T}}(Bx+b) = 0$ (the superscript "$^{\mathsf{T}}$" is used here and below to denote vector or matrix transposition). Problem (3.5) is a special case of the so-called (*extended*) *vertical LCP*, which uses more than two matrices and vectors in its formulation [54, 110, 258]. In the *standard LCP*, A is the identity matrix and $a = 0$ [181, 58].

The reformulation (3.4) of (3.5) is based on the fact that, for two real numbers α and β, $\min(\alpha, \beta) = 0$ if and only if $\alpha \geqslant 0$, $\beta \geqslant 0$ and $\alpha\beta = 0$ [3, 192]. This reformulation serves as the basis for a number of solving methods and investigations [3, 146, 195, 192, 193, 86, 22, 23, 132, 24, 71, 72, 73]. If (3.5) stands alone, it is appropriate to have $m = n$, but (3.5) may be part of a system with other constraints to satisfy [163, 164, 25], in which case $m \leqslant n$. In the computation of the B-differential of the Min function (3.3), m and n may be unrelated. Note that there are many other ways of reformulating problem (3.5) as a nonsmooth system of equations. It is frequent to use the *Fischer function*, whose B-differential is computed in [87]. The function H in (3.3) has been less studied and used than the Fischer function, although it has various advantages : it is piecewise affine (but has more nondifferentiability kinks), the local convergence of a semi-smooth Newton algorithm using it can be established under weaker assumptions and may be finitely locally convergent for linear complementarity problems [86, § 9.2].

Occasionally, we shall refer to the nonlinear version of the above problem, in which a function $\tilde{H} : \mathbb{E} \to \mathbb{R}^m$ is defined at $x \in \mathbb{E}$ by

$$\tilde{H}(x) := \min(F(x), G(x)), \tag{3.6}$$

where F and $G : \mathbb{E} \to \mathbb{R}^m$ are two functions and the "min" operator still acts componentwise. The equation $\tilde{H}(x) = 0$ is then a reformulation of the complementarity problem "$0 \leqslant F(x) \perp G(x) \geqslant 0$".

As a first general remark, let us quote the fact that the B-differential of H cannot be deduced from the knowledge of the B-differential of its scalar components $H_i : x \in \mathbb{E} \to H_i(x) \in \mathbb{R}$, for $i \in [1 : m]$, which is trivial in the present context. Indeed, it is known that [51, proposition 2.6.2(e)]

$$\partial_B H(x) \subseteq \partial_B^\times H(x) := \partial_B H_1(x) \times \cdots \times \partial_B H_m(x), \tag{3.7}$$

but equality in this inclusion may not hold (see [86, § 7.1.15], counter-example 3.2.3 and almost all the examples and test-cases below). Therefore, all the components of H must be taken into account simultaneously.

The B-differential of H at x is a finite set, made of Jacobians whose ith row is $A_{i,:}$ or $B_{i,:}$ (proposition 3.2.2). Consequently, its cardinality can be exponential in m and it occurs that its full mathematical description is a tricky task, essentially when there are many indices i for which $(Ax + a)_i = (Bx + b)_i$ and $A_{i,:} \neq B_{i,:}$, a situation that makes H nondifferentiable (lemma 3.2.1). Then, a rich panorama of configurations appears, which is barely glimpsed in this contribution. Note that the proposed computation methods do not require any assumptions on A or B.

The paper starts with a background section (section 3.2), which recalls a basic property of the minimum of two functions (lemma 3.2.1) and gives us a first perception of the structure of the B-differential of the function H, in particular its finite nature (proposition 3.2.2). A useful technical lemma is also presented (lemma 3.2.6).

In section 3.3, it is shown that the problem of computing $\partial_B H(x)$ has a rich panel of equivalent formulations, related to various areas of mathematics. We have quoted two forms of the problem in *linear algebra*, which are dual to each other (section 3.3.2), two equivalent problems in *convex analysis* (section 3.3.3) and a last equivalent problem, which arises in *computational discrete geometry* and deals with the arrangement of hyperplanes having the origin in common (section 3.3.4).

Section 3.4 gives some properties of the B-differential of H, recalls Winder's formula of its cardinality, provides some lower and upper bounds on this one, proves necessary and sufficient conditions so that two extreme configurations occur and highlights two links between the B-differential and C-differential.

Section 3.5 presents algorithms for computing one (section 3.5.1) or all (section 3.5.2) the Jacobians of $\partial_B H(x)$. In the latter case, the algorithms construct a tree incrementally and recursively (section 3.5.2), as proposed by Rada and Černý [208]. On the one hand (section 3.5.2), an algorithm based on the notion of matroid circuit of the matrix V expressing the "derivative gap" is proposed; it has the nice feature of requiring no linear optimization

problem (LOP) to solve. On the other hand (section 3.5.2), various modifications of the algorithm of Rada and Černý [208] are proposed with the goal of decreasing the number of LOPs to solve. Numerical experiments are reported (section 3.5.2), showing that the proposed algorithms significantly improve the performance of the Rada and Černý method, with mean (resp. median) acceleration ratios in the range 15..31 (resp. 5..20), measured by the computing time. This speed-up exceeds 100, for some algorithms and test-problems.

This paper is an abridged version of the more detailed report [78].

NOTATION. We denote by $|S|$ the number of elements of a set S (i.e., its *cardinality*). The *power set* of a set S is denoted by $\mathfrak{P}(S)$. The set of *bipartitions* (I, J) of a set K is denoted by $\mathfrak{B}(K) : I \cup J = K$ and $I \cap J = \varnothing$. The sets of nonzero natural and real numbers are denoted by $\mathbb{N}^*$ and $\mathbb{R}^*$, respectively. The *sign of a real number* is the multifunction $\mathrm{sgn} : \mathbb{R} \multimap \mathbb{R}$ defined by $\mathrm{sgn}(t) = \{1\}$ if $t > 0$, $\mathrm{sgn}(t) = \{-1\}$ if $t < 0$ and $\mathrm{sgn}(0) = [-1, 1]$. We note $\mathbb{R}^n_+ := \{x \in \mathbb{R}^n : x \geqslant 0\}$ and $\mathbb{R}^n_{++} := \{x \in \mathbb{R}^n : x > 0\}$ (strict inequalities must also be understood componentwise ; hence $x > 0$ means $x_i > 0$ for all indices i). For a subset S of a vector space, we denote by $\mathrm{vect}(S)$ the subspace spanned by S. The vector of all one's, in a real space whose dimension is given by the context, is denoted by e. The *Hadamard product* of u and $v \in \mathbb{R}^n$ is the vector $u \bullet v \in \mathbb{R}^n$ whose ith component is $u_i v_i$. The *range space* of an $m \times n$ matrix A is denoted by $\mathcal{R}(A)$, its *null space* by $\mathcal{N}(A)$, its *rank* is $\mathrm{rank}(A) := \dim \mathcal{R}(A)$ and its *nullity* is $\mathrm{null}(A) := \dim \mathcal{N}(A) = n - \mathrm{rank}(A)$ by the rank-nullity theorem. The ith row (resp. column) of A is denoted by $A_{i,:}$ (resp. $A_{:,i}$). Transposition operates after a row/column selection : $A^\mathsf{T}_{i,:}$ is a short notation for the column vector $(A_{i,:})^\mathsf{T}$ and $A^\mathsf{T}_{:,i}$ is a short notation for the row vector $(A_{:,i})^\mathsf{T}$. For a vector α, $\mathrm{Diag}(\alpha)$ is the square diagonal matrix with the α_i's on its diagonal.

3.2 Background

Recall that $F : \mathbb{E} \to \mathbb{F}$ is said to be (*Fréchet*) *differentiable* at x if $F(x + d) = F(x) + Ld + o(\|d\|)$ for some $L \in \mathcal{L}(\mathbb{E}, \mathbb{F})$, in which case one denotes by $F'(x) = L$ the *derivative* of F at x. We say below that F is *continuously differentiable at* x if it is differentiable near x (like in [51], "near" means here and below "in a neighborhood of" in the topological sense) and if its derivative is continuous at x.

The next famous lemma recalls a necessary and sufficient condition guaranteeing the differentiability of the minimum of two scalar functions (see [204, 1993, final remarks (1)], [255, 2011, theorem 2.1] and [78]).

Lemma 3.2.1 (differentiability of the Min function)**.** *Let f and $g : \mathbb{E} \to \mathbb{R}$ be two functions and $h : \mathbb{E} \to \mathbb{R}$ be defined by $h(\cdot) := \min(f(\cdot), g(\cdot))$. Suppose that f and g are differentiable at a point $x \in \mathbb{E}$.*

1) *If $f(x) < g(x)$, then h is differentiable at x and $h'(x) = f'(x)$.*
2) *If $f(x) > g(x)$, then h is differentiable at x and $h'(x) = g'(x)$.*

3) *If $f(x) = g(x)$, then h is differentiable at x if and only if $f'(x) = g'(x)$. In this case, $h'(x) = f'(x) = g'(x)$.*

The previous lemma shows the relevance of the following index sets :

$$\mathcal{A}(x) := \{i \in [1:m] : (Ax+a)_i < (Bx+b)_i\}, \tag{3.8a}$$
$$\mathcal{B}(x) := \{i \in [1:m] : (Ax+a)_i > (Bx+b)_i\}, \tag{3.8b}$$
$$\mathcal{E}(x) := \{i \in [1:m] : (Ax+a)_i = (Bx+b)_i\}, \tag{3.8c}$$
$$\mathcal{E}^{=}(x) := \{i \in \mathcal{E}(x) : A_{i,:} = B_{i,:}\}, \tag{3.8d}$$
$$\mathcal{E}^{\neq}(x) := \{i \in \mathcal{E}(x) : A_{i,:} \neq B_{i,:}\}. \tag{3.8e}$$

To simplify the presentation, we assume in the sequel that

$$\mathcal{E}^{\neq}(x) = [1:p], \tag{3.9}$$

for some $p \in [0:m]$ ($p = 0$ if and only if $\mathcal{E}^{\neq}(x) = \varnothing$).

The next proposition describes the superset $\partial_B^{\times} H(x)$ of $\partial_B H(x)$ given in the right-hand side of (3.7) (see [136, 1998, § 2] in a somehow different context, [66, 2000, before (8)] and [78] for a meticulous proof). This Cartesian product actually reads

$$\partial_B^{\times} H(x) := \{J \in \mathcal{L}(\mathbb{E}, \mathbb{R}^m) : \begin{array}{l} J_{i,:} = A_{i,:}, \text{ if } i \in \mathcal{A}(x), \\ J_{i,:} = B_{i,:}, \text{ if } i \in \mathcal{B}(x), \\ J_{i,:} = A_{i,:} = B_{i,:}, \text{ if } i \in \mathcal{E}^{=}(x), \\ J_{i,:} \in \{A_{i,:}, B_{i,:}\}, \text{ if } i \in \mathcal{E}^{\neq}(x)\}. \end{array} \tag{3.10}$$

Proposition 3.2.2 (superset of $\boldsymbol{\partial_B H(x)}$)**.** *One has $\partial_B H(x) \subseteq \partial_B^{\times} H(x) = \partial_B H_1(x) \times \cdots \times \partial_B H_m(x)$. In particular, $|\partial_B H(x)| \leqslant 2^p$.*

The following counter-example shows that one can have $\partial_B H(x) \neq \partial_B^{\times} H(x)$ and highlights the interest of the B-differential for the convergence of the semismooth Newton algorithm on (3.4).

Counter-example 3.2.3. Let $n = 2$, $m = 2$, $A = \left(\begin{smallmatrix} -1 & 1 \\ -1 & -1 \end{smallmatrix}\right)$, $B = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$ and $a = b = \binom{0}{0}$. One has $\mathcal{A}(0) = \mathcal{B}(0) = \varnothing$, $\mathcal{E}(0) = \mathcal{E}^{\neq}(0) = \{1, 2\}$, $\partial_B H(0) = \{A, B\}$, while $\partial_B^{\times} H(0) = \{A, B, \left(\begin{smallmatrix} -1 & 1 \\ 1 & -1 \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 & 1 \\ -1 & -1 \end{smallmatrix}\right)\}$. This example also shows that all the Jacobians of $\partial_B H(0)$ can be nonsingular, while the Jacobian $\left(\begin{smallmatrix} 1 & 1 \\ -1 & -1 \end{smallmatrix}\right)$ of $\partial_B^{\times} H(0)$ is singular and the central Jacobian (3.41), namely $\frac{1}{2}(A+B) = \left(\begin{smallmatrix} 0 & 1 \\ 0 & -1 \end{smallmatrix}\right) \in \partial_C H(0)$, is also singular. Therefore, in this case, H is BD-regular at 0 in the sense of [198, 132] (this notion is named *strong* BD-regularity in [204, p. 233]) and the conditions ensuring the local convergence of the semismooth Newton algorithm are satisfied [204, theorem 3.1]. □

The previous proposition shows that $\partial_B H(x)$ is a finite set. It also naturally leads to the next definition.

Definition 3.2.4 (complete B-differential)**.** We say that the B-differential of H at $x \in \mathbb{R}^n$ is *complete* if $\partial_B H(x) = \partial_B^{\times} H(x)$ or, equivalently, if $|\partial_B H(x)| = 2^p$. □

Definitions 3.2.5 (symmetry in $\boldsymbol{\partial_B H(x)}$)**.** For $x \in \mathbb{E}$, we say that the Jacobian $\tilde{J} \in \partial_B^\times H(x)$ is *symmetric* to the Jacobian $J \in \partial_B^\times H(x)$ if

$$\tilde{J}_{i,:} = \begin{cases} A_{i,:} & \text{if } i \in \mathcal{E}^{\neq}(x) \text{ and } J_{i,:} = B_{i,:}, \\ B_{i,:} & \text{if } i \in \mathcal{E}^{\neq}(x) \text{ and } J_{i,:} = A_{i,:}. \end{cases}$$

The B-differential $\partial_B H(x)$ itself is said to be *symmetric* if each Jacobian $J \in \partial_B H(x)$ has its symmetric Jacobian $\tilde{J}$ in $\partial_B H(x)$. □

We shall use several times the following lemma, which, for the sake of generality, is written in a slightly more abstract formalism than the one we need below (one could take for $\mathbb{E}$ a subspace of $\mathbb{R}^q$, for some $q \in \mathbb{N}^*$, and the Euclidean scalar product for $\langle \cdot, \cdot \rangle$). It is a refinement of [255, lemma 2.1].

Lemma 3.2.6 (discriminating covectors)**.** *Suppose that $(\mathbb{E}, \langle \cdot, \cdot \rangle)$ is a Euclidean vector space, $p \in \mathbb{N}^*$ and v_1, …, v_p are p distinct vectors of $\mathbb{E}$. Then, the set of vectors $\xi \in \mathbb{E}$ such that $|\{\langle \xi, v_i \rangle : i \in [1:p]\}| = p$ is dense in $\mathbb{E}$.*

Proof. Denote by Ξ the set of vectors $\xi \in \mathbb{E}$ such that $|\{\langle \xi, v_i \rangle : i \in [1:p]\}| = p$ (i.e., $\{\langle \xi, v_i \rangle : i \in [1:p]\}$ has p distinct values in $\mathbb{R}$). We have to show that Ξ is dense in $\mathbb{E}$.

Take $\xi_0 \notin \Xi$, so that $\langle \xi_0, v_i \rangle = \langle \xi_0, v_j \rangle$ for some $i \neq j$ in $[1:p]$. By continuity of the scalar product, for any $\varepsilon_0 > 0$ sufficiently small, the vector $\xi_1 := \xi_0 - \varepsilon_0(v_i - v_j)$ guarantees

$$\langle \xi_1, v_{i_1} \rangle < \langle \xi_1, v_{i_2} \rangle$$

for all i_1 and $i_2 \in [1:p]$ such that $\langle \xi_0, v_{i_1} \rangle < \langle \xi_0, v_{i_2} \rangle$ (in other words, ξ_1 maintains strict the inequalities that are strict with ξ_0). In addition

$$\langle \xi_1, v_i \rangle - \langle \xi_1, v_j \rangle = \underbrace{\langle \xi_0, v_i - v_j \rangle}_{=0} - \underbrace{\varepsilon_0 \|v_i - v_j\|^2}_{>0} < 0.$$

Therefore, one gets one more strict inequality with ξ_1 than with ξ_0. Pursuing like this, one can finally obtain a vector ξ in Ξ. This vector is arbitrarily close to ξ_0 by taking the ε_i's positive and sufficienty small. □

3.3 Equivalent problems

The problem of determining the B-differential of the piecewise affine function, that is the minimum (3.3) of two *affine* functions, appears in various contexts, sometimes with non straightforward connections with it (this one is recalled in section 3.3.1). We review some equivalent formulations in this section (see also [253, 14, 16] and the references therein) and give a few properties of the B-differential in this piecewise affine case. As suggested by proposition 3.2.2, these problems have an enumeration nature, since a finite list of mathematical objects has to be determined. This list may have a number of elements exponential

in p, which makes its content difficult to specify (in this respect, the particular case where the B-differential is complete is a trivial exception). Some formulations, such as the one related to the arrangement of hyperplanes containing the origin (section 3.3.4), have been extensively explored, others much less. Each formulation sheds a particular light on the problem and is therefore interesting to mention and keep in mind. They also offer the possibility of introducing new algorithmic approaches to describe the B-differential.

3.3.1 B-differential of the minimum of two affine functions

The problem of this section was already presented in the introduction and is sometimes referred to, in this paper, as the *original problem*.

Problem 3.3.1 (B-differential of the minimum of two affine functions)**.** Let be given two positive integers n and $m \in \mathbb{N}^*$, two matrices A, $B \in \mathbb{R}^{m\times n}$ and two vectors a, $b \in \mathbb{R}^m$. It is requested to compute the B-differential at some $x \in \mathbb{R}^n$ of the function $H : \mathbb{R}^n \to \mathbb{R}^m$ defined by (3.3). □

When $\mathcal{E}^{\neq}(x) \neq \varnothing$, the rows of $B - A$ with indices in $\mathcal{E}^{\neq}(x)$ will play a key role below. We denote its transpose by

$$V := (B - A)^{\mathsf{T}}_{\mathcal{E}^{\neq}(x),:} \in \mathbb{R}^{n\times p}. \tag{3.11}$$

Note that, due to their indices in $\mathcal{E}^{\neq}(x) = [1 : p]$ and the definition of this index set, the columns of V are nonzero. This matrix may not always have full rank, however.

The following example will accompany us throughout this section.

Example 3.3.2 (a simple example)**.** Consider the trivial linear complementarity problem $0 \leqslant x \perp (Mx + q) \geqslant 0$ defined by

$$M = \begin{pmatrix} 2 & 0 & 0 \\ -\alpha & 1+\beta & 0 \\ -\alpha & -\beta & 1 \end{pmatrix} \qquad \text{and} \qquad q = 0,$$

where $\alpha := -\cos(2\pi/3) = 1/2 > 0$ and $\beta := \sin(2\pi/3) \in (\alpha, 2\alpha)$. Note that, at the unique solution $x = 0$ to the problem, one has $\mathcal{A}(x) = \mathcal{B}(x) = \mathcal{E}^{=}(x) = \varnothing$ and $\mathcal{E}(x) = \mathcal{E}^{\neq}(x) = [1\,{:}\,3]$, so that $p = 3$ and

$$V = \begin{pmatrix} 1 & -\alpha & -\alpha \\ 0 & \beta & -\beta \\ 0 & 0 & 0 \end{pmatrix}.$$

□

3.3.2 Linear algebra problems

Signed feasibility of strict inequality systems

We call *sign vector* a vector whose components are $+1$ or -1. Many proofs below leverage the equivalence between the original problem 3.3.1 and the following one. The reason

is that working on problem 3.3.3 often allows us to propose shorter proofs. In addition, the algorithms of section 3.5 all focus on the generation of the sign vectors s forming the set $\mathcal{S}$ in (3.12) below. Recall the definition of the Hadamard product : $(u \cdot v)_i = u_i v_i$.

Problem 3.3.3 (signed feasibility of strict inequality systems)**.** Let be given two positive integers n and $p \in \mathbb{N}^*$ and a matrix V in $\mathbb{R}^{n\times p}$ with nonzero columns. It is requested to determine the set

$$\mathcal{S} := \{s \in \{\pm 1\}^p : \ s \cdot (V^\mathsf{T} d) > 0 \text{ holds for some } d \in \mathbb{R}^n\}. \tag{3.12}$$

□

By routine verification, one can see that the sign vectors s in $\mathcal{S}$ for example 3.3.2 are given by the columns of the matrix S below and possible associated directions d such that $s \cdot (V^\mathsf{T} d) > 0$ are given by the corresponding columns of the matrix D :

$$S = \begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 & -1 \end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix} 2 & 2 & 2 & -2 & -2 & -2 \\ 2 & 1 & -2 & -2 & -1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{3.13}$$

The sign vectors $\pm e := \pm(1,1,1)$ are not in $\mathcal{S}$ since $Ve = 0$ (there is not $d_\pm$ such that $(\pm e) \cdot (V^\mathsf{T} d_\pm) > 0$, since this would imply that $0 < \pm e^\mathsf{T} V^\mathsf{T} d_\pm = 0$, a contradiction). Therefore, there are only 6 sign vectors in $\mathcal{S}$ instead of the 8 sign vectors in $\{\pm 1\}^3$.

The link between problems 3.3.1 and 3.3.3 is established by the following map :

$$\sigma : J \in \partial_B^\times H(x) \mapsto s \in \{\pm 1\}^p, \text{ where } s_i = \begin{cases} +1 & \text{if } i \in \mathcal{E}^{\neq}(x),\ J_{i,:} = A_{i,:}, \\ -1 & \text{if } i \in \mathcal{E}^{\neq}(x),\ J_{i,:} = B_{i,:}, \end{cases} \tag{3.14a}$$

where we have used the definition (3.9) of p. The map is well defined since $A_{i,:} \neq B_{i,:}$ when $i \in \mathcal{E}^{\neq}(x)$. Furthermore, σ is bijective since two Jacobians in $\partial_B^\times H(x)$ only differ by their rows with index in $\mathcal{E}^{\neq}(x)$ and that these rows can take any of the values $A_{i,:}$ or $B_{i,:}$. Actually, its reverse map is

$$\sigma^{-1} : s \in \{\pm 1\}^p \mapsto J \in \partial_B^\times H(x), \text{ where } J_{i,:} = \begin{cases} A_{i,:} & \text{if } i \in \mathcal{E}^{\neq}(x),\ s_i = +1, \\ B_{i,:} & \text{if } i \in \mathcal{E}^{\neq}(x),\ s_i = -1. \end{cases} \tag{3.14b}$$

The question that arises is whether σ is also a bijection between $\partial_B H(x)$ and $\mathcal{S}$.

Proposition 3.3.4 (bijection $\boldsymbol{\partial_B H(x) \leftrightarrow \mathcal{S}}$)**.** *Let $H : \mathbb{R}^n \to \mathbb{R}^m$ be given by* (3.3)*, x be a point in $\mathbb{R}^n$ such that $p \neq 0$ and V be given by* (3.11)*. Then, the map σ is a bijection from $\partial_B H(x)$ onto $\mathcal{S}$. In particular, the following properties hold.*

1) *If $J \in \partial_B H(x)$, then $\exists\, d \in \mathbb{R}^n$ such that $\sigma(J) \cdot (V^\mathsf{T} d) > 0$.*
2) *If $s \in \{\pm 1\}^p$ and $\exists\, d \in \mathbb{R}^n$ such that $s \cdot (V^\mathsf{T} d) > 0$, then $\sigma^{-1}(s) \in \partial_B H(x)$.*
3) *Let $J \in \partial_B^\times H(x)$. Then, $J \in \partial_B H(x) \iff \sigma(J) \cdot (V^\mathsf{T} d) > 0$ holds for some $d \in \mathbb{R}^n$.*

Proof. The properties 1, 2 and 3 in the statement of the proposition are straightforward consequences of the bijectivity of $\sigma : \partial_B H(x) \to \mathcal{S}$. Now, the discussion before the proposition has shown that $\sigma : \partial_B^\times H(x) \mapsto \{\pm 1\}^p$ is a bijection. Therefore, $\sigma : \partial_B H(x) \mapsto \{\pm 1\}^p$

is injective and it suffices to prove that

$$\sigma(\partial_B H(x)) = \mathcal{S}. \tag{3.15a}$$

[$\subseteq$ or point 1] Let $J \in \partial_B H(x)$. We have to show that $s := \sigma(J) \in \mathcal{S}$, which means that one can find a $d \in \mathbb{R}^n$ such that $s \bullet (V^\mathsf{T} d) > 0$. By $J \in \partial_B H(x)$, there exists a sequence $\{x_k\} \subseteq \mathcal{D}_H$ converging to x such that

$$H'(x_k) \to J. \tag{3.15b}$$

For $i \in \mathcal{E}^{\neq}(x)$, one cannot have $(Ax_k + a)_i = (Bx_k + b)_i$, since $A_{i,:} \neq B_{i,:}$ would imply that $x_k \notin \mathcal{D}_H$ (lemma 3.2.1). Therefore, one can find a subsequence $\mathcal{K}$ of indices k and a partition $(\mathcal{A}_0, \mathcal{B}_0)$ of $\mathcal{E}^{\neq}(x)$ such that for all $k \in \mathcal{K}$:

$$(Ax_k + a)_{\mathcal{A}_0} < (Bx_k + b)_{\mathcal{A}_0} \qquad \text{and} \qquad (Ax_k + a)_{\mathcal{B}_0} > (Bx_k + b)_{\mathcal{B}_0}. \tag{3.15c}$$

Now, fix $k \in \mathcal{K}$ and set $d := x_k - x$. Since $(Ax + a)_i = (Bx + b)_i$ for $i \in \mathcal{E}^{\neq}(x)$, one deduces from (3.15c) that

$$(B - A)_{\mathcal{A}_0,:} d > 0 \qquad \text{and} \qquad (B - A)_{\mathcal{B}_0,:} d < 0.$$

Recalling the definitions of V in (3.11) and $\mathcal{S}$ in (3.12), we see that, to conclude the proof of the membership $\sigma(J) \in \mathcal{S}$, it suffices to show that $[\sigma(J)]_{\mathcal{A}_0} = +1$ and $[\sigma(J)]_{\mathcal{B}_0} = -1$ or, equivalently, by the definition of σ, ($J_{i,:} = A_{i,:}$ for $i \in \mathcal{A}_0$) and ($J_{i,:} = B_{i,:}$ for $i \in \mathcal{B}_0$). This is indeed the case, since by (3.15c), for all $k \in \mathcal{K}$, one has ($H'_i(x_k) = A_{i,:}$ for $i \in \mathcal{A}_0$) and ($H'_i(x_k) = B_{i,:}$ for $i \in \mathcal{B}_0$); now, use the convergence (3.15b) to conclude.

[$\supseteq$ or point 2] Let $s \in \mathcal{S}$. We have to find a $J \in \partial_B H(x)$ such that $\sigma(J) = s$, that is, which satisfies for $i \in [1 : p]$:

$$(J_{i,:} = A_{i,:} \text{ if } s_i = +1) \quad \text{and} \quad (J_{i,:} = B_{i,:} \text{ if } s_i = -1). \tag{3.15d}$$

Since $s \in \mathcal{S}$, there is a $d \in \mathbb{R}^n$ such that

$$s \bullet (V^\mathsf{T} d) > 0. \tag{3.15e}$$

Take a real sequence $\{t_k\} \downarrow 0$ and define the sequence $\{x_k\} \subseteq \mathbb{R}^n$ by

$$x_k := x + t_k d.$$

Then, $x_k \to x$. We claim that, for k sufficiently large, $x_k \in \mathcal{D}_H$ and $H'(x_k)$ is a constant matrix J satisfying (3.15d), which will conclude the proof. Let $i \in [1 : m]$.

- If $i \in \mathcal{A}(x)$, $(Ax_k + a)_i < (Bx_k + b)_i$ for k large, so that $x_k \in \mathcal{D}_H$ and $H'_i(x_k) = A_{i,:}$.
- If $i \in \mathcal{B}(x)$, $(Ax_k + a)_i > (Bx_k + b)_i$ for k large, so that $x_k \in \mathcal{D}_H$ and $H'_i(x_k) = B_{i,:}$.
- If $i \in \mathcal{E}^{=}(x)$, then $A_{i,:} = B_{i,:}$, so that $x_k \in \mathcal{D}_H$ and $H'_i(x_k) = A_{i,:} = B_{i,:}$.
- If $i \in \mathcal{E}^{\neq}(x)$, subtract side by side $(Ax_k + a)_i = (Ax + a)_i + t_k A_{i,:} d$ and $(Bx_k + b)_i = (Bx + b)_i + t_k B_{i,:} d$, use $(Ax + a)_i = (Bx + b)_i$ and next (3.15e) to get

$$(Bx_k + b)_i - (Ax_k + b)_i = t_k (B_{i,:} - A_{i,:}) d = t_k V^\mathsf{T}_{i,:} d \begin{cases} > 0 & \text{if } s_i = +1, \\ < 0 & \text{if } s_i = -1. \end{cases}$$

Hence, $x_k \in \mathcal{D}_H$, ($H'_i(x_k) = A_{i,:}$ if $s_i = +1$) and ($H'_i(x_k) = B_{i,:}$ if $s_i = -1$). □

Equivalence 3.3.5. (B-differential ↔ signed feasibility of strict inequality systems) The equivalence between the original problem 3.3.1 and the signed feasibility of strict inequality system problem 3.3.3 is a consequence of the previous proposition with V given by (3.11), which shows the bijectivity of the map $\sigma : \partial_B H(x) \to \mathcal{S}$ defined by (3.14a). Therefore, knowing σ by its definition (3.14), determining $\partial_B H(x)$ or $\mathcal{S}$ are equivalent problems. □

Orthants encountered by the null space of a matrix

Recall the definition of $\mathcal{S}$ in (3.12), which is associated with some matrix $V \in \mathbb{R}^{n\times p}$ with nonzero columns, which may or not come from (3.11). The equivalent form of problem 3.3.3 (hence of problem 3.3.1 when V is defined by (3.11)) introduced in this section is based on a bijection between the *complementary set* of $\mathcal{S}$ in $\{\pm 1\}^p$, denoted $\mathcal{S}^c := \{\pm 1\}^p \setminus \mathcal{S}$, and a collection $\mathcal{I}$ of subsets of $[1 : p]$ (i.e., $\mathcal{I} \subseteq \mathfrak{P}([1 : p])$), which refers to a collection of orthants of $\mathbb{R}^p$, those encountered by the null space of V. This equivalence will play a major part in the conception of the algorithms in section 3.5.2, in particular, but not only, in an algorithm describing the *complementary set* of $\partial_B H(x)$, which is interesting when $|\partial_B^\times H(x) \setminus \partial_B H(x)|$ is small. The concept of *stem vector*, defined in the second part of this section, has proven useful in this regard. The equivalence rests on a duality concept through Gordan's alternative.

Problem 3.3.6 (orthants encountered by the null space of a matrix)**.** Let be given two positive integers n and $p \in \mathbb{N}^*$ and a matrix V in $\mathbb{R}^{n\times p}$ with nonzero columns. Associate with $I \subseteq [1 : p]$ the following orthant of $\mathbb{R}^p$:

$$\mathcal{O}_I^p := \{y \in \mathbb{R}^p : y_I \geqslant 0,\ y_{I^c} \leqslant 0\},$$

where $I^c := [1 : p] \setminus I$. It is requested to determine the set

$$\mathcal{I} := \{I \subseteq [1 : p] : \mathcal{N}(V) \cap \mathcal{O}_I^p \neq \{0\}\}. \qquad \square$$

Note that, if $I \in \mathcal{I}$, then $I^c \in \mathcal{I}$ (because $y \in (\mathcal{N}(V) \cap \mathcal{O}_I^p) \setminus \{0\}$ implies that $-y \in (\mathcal{N}(V) \cap \mathcal{O}_{I^c}^p) \setminus \{0\}$), so that $|\mathcal{I}|$ is even (just like $|\mathcal{S}|$ and $|\mathcal{S}^c|$, see proposition 3.4.1).

The equivalence between problems 3.3.3 and 3.3.6 is obtained thanks to the following bijection

$$\imath : s \in \{\pm 1\}^p \to \imath(s) := \{i \in [1 : p] : s_i = +1\} \in \mathfrak{P}([1 : p]), \tag{3.16}$$

whose reverse map is $\imath^{-1} : I \in \mathfrak{P}([1 : p]) \to s \in \{\pm 1\}^p$, where $s_i = +1$ if $i \in I$ and $s_i = -1$ if $i \notin I$. As announced above, this equivalence relies on Gordan's theorem of the alternative [108, p. 1873] : for a matrix $A \in \mathbb{R}^{m\times n}$,

$$\exists\, x \in \mathbb{R}^n : Ax > 0 \qquad \Longleftrightarrow \qquad \nexists\, \alpha \in \mathbb{R}_+^m \setminus \{0\} : A^\mathsf{T}\alpha = 0. \tag{3.17}$$

Proposition 3.3.7 (bijection $\mathcal{S}^c \leftrightarrow \mathcal{I}$)**.** *The map $\imath$ defined by* (3.16) *is a bijection from $\mathcal{S}^c$ onto $\mathcal{I}$.*

Proof. Let $s \in \{\pm 1\}^p$ and set $I := \imath(s) = \{i \in [1:p] : s_i = +1\}$. Define $A := \mathrm{Diag}(s)V^\mathsf{T}$ to make the link with Gordan's alternative (3.17). One has the equivalences

$$\begin{aligned} s \in \mathcal{S}^c &\iff \nexists\, x \in \mathbb{R}^n : Ax > 0 \qquad \text{[definition of } \mathcal{S} \text{ in (3.12)]} \\ &\iff \exists\, \alpha \in \mathbb{R}^m_+ \setminus \{0\} : A^\mathsf{T}\alpha = 0 \qquad \text{[Gordan's alternative (3.17)]} \\ &\iff \exists\, \alpha \in \mathbb{R}^m_+ \setminus \{0\} : s \bullet \alpha \in \mathcal{N}(V) \\ &\iff \mathcal{N}(V) \cap \mathcal{O}^p_I \neq \{0\} \qquad \text{[see below]} \\ &\iff I \in \mathcal{I} \qquad \text{[definition of } \mathcal{I}\text{]}. \end{aligned} \tag{3.18}$$

The implication "$\Rightarrow$" in (3.18) is due to the fact that $s \bullet \alpha$ is nonzero and belongs to both $\mathcal{N}(V)$ and $\mathcal{O}^p_I$. The reverse implication "$\Leftarrow$" in (3.18) is due to the fact that there is a nonzero $y \in \mathcal{N}(V) \cap \mathcal{O}^p_I$, implying that $\alpha := s \bullet y$ is nonzero and $\geqslant 0$ and is such that $s \bullet \alpha = y \in \mathcal{N}(V)$.

Since $\imath : \{\pm 1\}^p \to \mathfrak{P}([1:p])$ is a bijection, the above equivalences show that $\imath$ is also a bijection from $\mathcal{S}^c$ onto $\mathcal{I}$. □

Equivalence 3.3.8 ($\mathcal{S}^c \leftrightarrow \mathcal{I}$). The equivalence between problems 3.3.3 and 3.3.6 is a consequence of the bijectivity of $\imath : \mathcal{S}^c \to \mathcal{I}$, established in proposition 3.3.7 : to determine $\mathcal{S}$, it suffices to determine $\mathcal{S}^c = \imath^{-1}(\mathcal{I})$, hence to determine $\mathcal{I}$, and vice versa. □

In example 3.3.2, one has $\mathcal{N}(V) = \mathbb{R}e$, which only encounters the orthants $\mathcal{O}^3_\varnothing$ and $\mathcal{O}^3_{[1:3]}$ outside the origin; hence $\mathcal{I} = \{\varnothing, [1:3]\}$. We have seen that $\mathcal{S}^c = \{\pm(1,1,1)\}$ for this problem. Clearly, $\imath$ maps $\mathcal{S}^c$ onto $\mathcal{I}$ bijectively, as claimed in proposition 3.3.7.

Recall that the *nullity* of a matrix A, denoted by $\mathrm{null}(A)$, is the dimension of its null space. Let us introduce the following collection of index sets (from now on, J usually denotes a set of indices rather than a Jacobian matrix) :

$$\mathcal{C} := \{J \subseteq [1:p] : J \neq \varnothing,\ \mathrm{null}(V_{:,J}) = 1, V_{:,J_0} \text{ is injective if } J_0 \subsetneq J\}, \tag{3.19}$$

where "$\subsetneq$" is used to denote strict inclusion. In the terminology of the *vector matroid* formed by the columns of V and its subsets made of linearly independent columns [191, proposition 1.1.1], the elements of $\mathcal{C}$ are called the *circuits* of the matroid [191, proposition 1.3.5(iii)]. The particular expression (3.19) of the circuit set is interesting in the present context, since it readily yields the following implication :

$$J \in \mathcal{C} \quad \Longrightarrow \quad \text{any nonzero } \alpha \in \mathcal{N}(V_{:,J}) \text{ has none zero component.} \tag{3.20}$$

From (3.19) and (3.20), one can associate with $J \in \mathcal{C}$ a pair of sign vectors $\pm\tilde{s} \in \{\pm 1\}^J$ by $\tilde{s} := \mathrm{sgn}(\alpha)$ for some nonzero $\alpha \in \mathcal{N}(V_{:,J})$; the sign vectors $\pm\tilde{s}$ do not depend on the chosen $\alpha \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ since $\mathrm{null}(V_{:,J}) = 1$. We call such a sign vector a *stem vector*, because of proposition 3.3.10 below, which shows that any $s \in \mathcal{S}^c$ can be generated from such a stem vector.

Definition 3.3.9 (stem vector)**.** A *stem vector* is a sign vector $\tilde{s} = \mathrm{sgn}(\alpha)$, where $\alpha \in \mathcal{N}(V_{:,J})$ for some $J \in \mathcal{C}$. □

Note that there are twice as many stem vectors as circuits and that the stem vectors do not have all the same size.

The matrix V in example 3.3.2 has $J = [1:3]$ as single circuit. Since $Ve = 0$, the associated stem vectors are $\pm e = \pm(1,1,1)$. The next proposition now confirms that $\pm(1,1,1)$ are the only elements of $\mathcal{S}^c$.

Proposition 3.3.10 (generating $\mathcal{S}^c$ from the stem vectors)**.** *For $s \in \{\pm 1\}^p$,*

$$s \in \mathcal{S}^c \quad \Longleftrightarrow \quad s_J = \tilde{s} \text{ for some } J \subseteq [1:p] \text{ and some stem vector } \tilde{s}. \tag{3.21}$$

Proof. [$\Rightarrow$] The index set $J \subseteq [1:p]$ in the right-hand side of (3.21) can be determined as one satisfying the following two properties :

$$\{d \in \mathbb{R}^n : s_j v_j^\mathsf{T} d > 0 \text{ for all } j \in J\} = \varnothing, \tag{3.22a}$$

$$\forall\, J_0 \subsetneq J,\ \{d \in \mathbb{R}^n : s_j v_j^\mathsf{T} d > 0 \text{ for all } j \in J_0\} \neq \varnothing. \tag{3.22b}$$

To determine such a J, start with $J = [1:p]$, which verifies (3.22a), since $s \in \mathcal{S}^c$. Next, remove an index j from $[1:p]$ if (3.22a) holds for $J = [1:p] \setminus \{j\}$. Pursuing the elimination of indices j in this way, one arrives to an index set J satisfying (3.22a) and $\{d \in \mathbb{R}^n : s_j v_j^\mathsf{T} d > 0 \text{ for all } j \in J \setminus \{j_0\}\} \neq \varnothing$ for all $j_0 \in J$. Then, (3.22b) clearly holds. We claim that, for a J satisfying (3.22a) and (3.22b), s_J is a stem vector, which will conclude the proof of the implication.

To stick to definition 3.19, we start by showing that J is a matroid circuit. By (3.22a), $J \neq \varnothing$. By Gordan's alternative (3.17), (3.22a) and (3.22b) read

$$\exists\, \alpha \in \mathbb{R}^J_+ \setminus \{0\} \text{ such that } \textstyle\sum_{j\in J} s_j v_j \alpha_j = 0, \tag{3.22c}$$

$$\forall\, J_0 \subsetneq J, \nexists\, \alpha' \in \mathbb{R}^{J_0}_+ \setminus \{0\} \text{ such that } \textstyle\sum_{j\in J_0} s_j v_j \alpha'_j = 0. \tag{3.22d}$$

From these properties, one deduces that $\alpha > 0$ and that $\mathrm{null}(V_{:,J}) \geqslant 1$. To show that $\mathrm{null}(V_{:,J}) = 1$, we proceed by contradiction. Suppose that there is a nonzero $\alpha'' \in \mathbb{R}^J$ that is not colinear with α and that verifies $\sum_{j\in J} s_j v_j \alpha''_j = 0$. One can assume that $t := \max\{\alpha''_j/\alpha_j : j \in J\}$ is > 0 (take $-\alpha''$ otherwise). Set $J_0 := \{j \in J : \alpha''_j/\alpha_j < t\}$. By the non-colinearity of α and α'', on the one hand, and the definition of t, on the other hand, one has $\varnothing \subsetneq J_0 \subsetneq J$. Furthermore, $\alpha' := \alpha - \alpha''/t \geqslant 0$, $\alpha'_j > 0$ for $j \in J_0$ and $\alpha'_j = 0$ for $j \in J \setminus J_0$. Therefore, $\sum_{j\in J_0} s_j v_j \alpha'_j = \sum_{j\in J} s_j v_j \alpha'_j = 0$, yielding a contradiction with (3.22d).

To show that $J \in \mathcal{C}$, we still have to prove that $V_{:,J_0}$ is injective when $J_0 \subsetneq J$. Equivalently, it suffices to show that any $\beta \in \mathcal{N}(V_{:,J})$ with some zero component vanishes. We proceed by contradiction. If there is a $\beta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ with a zero component, $s_J \cdot \alpha$ and β would be two linearly independent vectors in $\mathcal{N}(V_{:,J})$ (since $s_J \cdot \alpha$ has no zero component), contradicting $\mathrm{null}(V_{:,J}) = 1$.

Now, since $s_J = \mathrm{sgn}(s_J \cdot \alpha)$, since $s_J \cdot \alpha \in \mathcal{N}(V_{:,J})$ by (3.22c) and since J is a matroid circuit of V, s_J is a stem vector.

[$\Leftarrow$] Since s_J is a stem vector, it follows that $s_J := \mathrm{sgn}(\alpha)$ for some $\alpha \in \mathbb{R}^J$ with nonzero components that satisfies $V_{:,J}\alpha = 0$. Then, there is no $d \in \mathbb{R}^n$ such that $s_J \cdot$

$(V_{:,J}^{\mathsf{T}} d) > 0$ (otherwise, $(s_J \cdot \alpha \cdot s_J) \cdot (V_{:,J}^{\mathsf{T}} d) > 0$, because $s_J \cdot \alpha > 0$, or $\alpha \cdot (V_{:,J}^{\mathsf{T}} d) > 0$, implying that $0 = \alpha^{\mathsf{T}}(V_{:,J}^{\mathsf{T}} d) > 0$, a contradiction). Hence, there exists certainly no $d \in \mathbb{R}^n$ such that $s \cdot (V^{\mathsf{T}} d) > 0$. This implies that $s \in \mathcal{S}^c$. □

To determine the stem vectors, which are based on the matroid circuits of V defined by (3.19), one has to select subsets of columns of V forming a rank one matrix, whose strict subsets form injective matrices. Actually, this last condition can be simplified by the following property.

Proposition 3.3.11 (matroid circuit detection)**.** *Suppose that $I \subseteq [1 : p]$ is such that $\mathrm{null}(V_{:,I}) = 1$ and that $\alpha \in \mathcal{N}(V_{:,I}) \setminus \{0\}$. Then, $J := \{i \in I : \alpha_i \neq 0\}$ is a matroid circuit of V and the unique one included in I.*

Proof. 1) Let us show that J is a matroid circuit.

Since $\alpha \neq 0$, one has $J \neq \varnothing$.

Let us show that $\mathrm{null}(V_{:,J}) = 1$. Since $J \subseteq I$, one has $\mathrm{null}(V_{:,J}) \leqslant \mathrm{null}(V_{:,I}) = 1$. Furthermore, $\alpha_J \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ implies that $\mathrm{null}(V_{:,J}) \geqslant 1$.

Now, let $J_0 \subsetneq J$ and suppose that $V_{:,J_0}\beta = 0$. We have to show that $\beta = 0$. Since $V_{:,J}(\beta, 0_{J\setminus J_0}) = 0$, it follows that $(\beta, 0_{J\setminus J_0}) \in \mathcal{N}(V_{:,J})$, which is of dimension 1, so that $(\beta, 0_{J\setminus J_0})$ is colinear to α. Since the components of α are $\neq 0$, we get that $\beta = 0$.

2) Let us now show that J is the unique matroid circuit of V included in I.

Let J' be a matroid circuit of V included in I. Then $\mathrm{null}(V_{:,J'}) = 1$ and there is a nonzero $\alpha' \in \mathcal{N}(V_{:,J'})$. By (3.20), α' has nonzero components. Furthermore, $(\alpha', 0_{I\setminus J'}) \in \mathcal{N}(V_{:,I})$, which has unit dimension and contains α. Therefore, α and $(\alpha', 0_{I\setminus J'})$ are colinear. Since the components of α are $\neq 0$, we get that $J' = J$. □

3.3.3 Convex analysis problems

The formulation of the original problem 3.3.1 in the form of the convex analysis problems 3.3.12 and 3.3.15 below may be useful to highlight some properties of $\partial_B H(x)$, thanks to the tools of that discipline.

Pointed cones by vector inversions

Recall that a *convex cone* K of $\mathbb{R}^n$ is a convex set verifying $\mathbb{R}_{++}K \subseteq K$ (or, more explicitly, $tx \in K$ when $t > 0$ and $x \in K$). A *closed* convex cone K is said to be *pointed* if $K \cap (-K) = \{0\}$ [32, p. 54], which amounts to saying that K does not contain a line (i.e., an affine subspace of dimension one) or that K has no nonzero direction z such that $-z \in K$. For $P \subseteq \mathbb{R}^n$, we also denote by "$\mathrm{cone}\, P$" the smallest *convex* cone containing P.

Problem 3.3.12 (pointed cones by vector inversions)**.** Let be given two positive integers n and $p \in \mathbb{N}^*$ and p vectors $v_1, \ldots, v_p \in \mathbb{R}^n \setminus \{0\}$. It is requested to determine all the sign vectors $s \in \{\pm 1\}^p$ such that $\mathrm{cone}\{s_i v_i : i \in [1:p]\}$ is pointed. □

The solution to problem 3.3.12 for the vectors that are the columns of the matrix V in example 3.3.2 is illustrated in figure 3.1.

The equivalence between the original problem 3.3.1 and problem 3.3.12 is obtained thanks to the next proposition, which gives another property ("cone pointedness") that is equivalent to those in (3.17) and that is adapted to the present concern. For a proof, see [112, theorem 2.3.29] or [78].

Proposition 3.3.13 (pointed polyhedral cone)**.** *For a finite collection of nonzero vectors* $\{w_i : i \in [1:p]\} \subseteq \mathbb{R}^n$*, the following properties are equivalent :*

(i) $\mathrm{cone}\{w_i : i \in [1:p]\}$ *is pointed,*
(ii) $\nexists\, \alpha \in \mathbb{R}^p_+ \setminus \{0\} : \sum_{i\in[1:p]} \alpha_i w_i = 0$,
(iii) $\exists\, d \in \mathbb{R}^n,\ \forall\, i \in [1:p] : w_i^\mathsf{T} d > 0$.

Equivalence 3.3.14 (signed linear system feasibility $\leftrightarrow$ pointed cone by vector inversion)**.** The equivalence $(i) \Leftrightarrow (iii)$ of the previous proposition shows that the set $\mathcal{S}$ defined by (3.12) is also given by

$$\mathcal{S} = \{s \in \{\pm 1\}^p :\ \mathrm{cone}\{s_i v_i : i \in [1:p]\} \text{ is pointed}\}. \tag{3.23}$$

To put it in words, denoting by $v_1, \ldots, v_p$ the columns of the matrix V defined by (3.11), the signed feasibility problem 3.3.3 is equivalent to problem 3.3.12. □

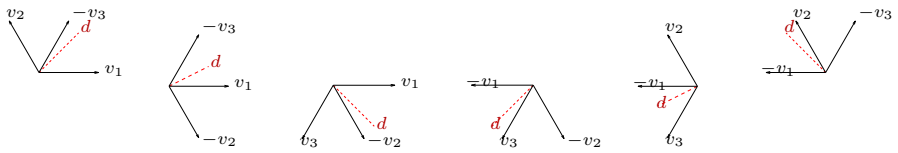


FIGURE 3.1 – The figure is related to the linear complementarity problem defined by example 3.3.2 : the v_i's are the columns of the matrix V (their third zero components are not represented). Each of the 6 sets of vectors plots the 3 vectors $\{s_i v_i : i \in [1:3]\}$, for each of the 6 sign vectors $s \in \mathcal{S}$ (given by the columns of the matrix S in (3.13)), as well as a direction d (given by the columns of D in (3.13), dashed lines) such that $s_i v_i^\mathsf{T} d > 0$ for all $i \in [1:3]$. Each conic hull of these vectors, namely $\mathrm{cone}\{s_i v_i : i \in [1:3]\}$, is pointed. The conic hulls of $\{v_1, v_2, v_3\}$ and $\{-v_1, -v_2, -v_3\}$ are both the space of dimension 2, hence there are not pointed, which confirms the fact that $(1,1,1)$ and $(-1,-1,-1)$ are not in $\mathcal{S}$.

Linearly separable bipartitions of a finite set

This section extends section 3.3.3 and adopts its concepts and notation. The point of view presented in this section was also shortly considered by Zaslavsky [257, 1975, § 6A].

This enumeration problem appears in the study of neural networks [251]. Baldi and Vershynin [16] make the connection with *homogeneous linear threshold functions* and highlight its impact in deep learning [228, 15].

Problem 3.3.15 (linearly separable bipartitioning)**.** Let be given an affine space $\mathbb{A}$ and $p \in \mathbb{N}^*$ vectors $\bar{v}_1, \ldots, \bar{v}_p \in \mathbb{A}$. Let $\mathbb{A}_0 := \mathbb{A} - \mathbb{A}$ be the vector space parallel to $\mathbb{A}$, endowed with a scalar product $\langle \cdot, \cdot \rangle$. It is requested to find all the ordered bipartitions (i.e., the partitions made of two subsets) (I, J) of $[1 : p]$ for which there exists a vector $\xi \in \mathbb{A}_0$ (also called *separating covector* below) such that

$$\forall\, i \in I,\ \forall\, j \in J : \quad \langle \xi, \bar{v}_i \rangle < \langle \xi, \bar{v}_j \rangle. \qquad \square$$

Of course, if (I, J) is an appropriate ordered bipartition to which a separating covector ξ corresponds, then (J, I) is also an appropriate ordered bipartition with separating covector $-\xi$. Therefore, only half of the appropriate ordered bipartitions (I, J) must be identified, a fact that is related to the symmetry of $\partial_B H(x)$ (proposition 3.4.1). Figure 3.2 shows the

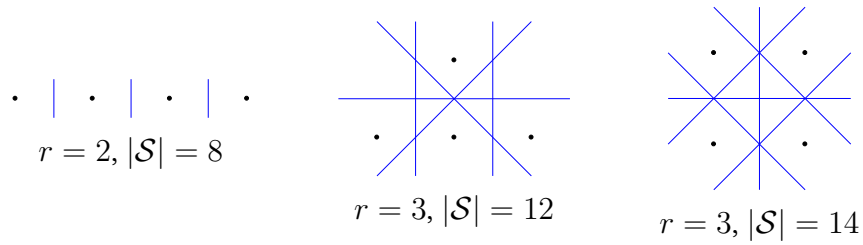


FIGURE 3.2 – Linearly separable bipartitions of a set of $p = 4$ points $\bar{v}_i$ in $\mathbb{R}^2$ (the dots in the figure). Possible separating hyperplanes are the drawn lines. We have not represented any separating line associated with the partition $(\varnothing, [1 : p])$ or $([1 : p], \varnothing)$, so that $|\mathcal{S}| = 2(n_s + 1)$, where n_s is the number of represented separating lines. We have set $r := \dim(\mathrm{vect}\{\bar{v}_1, \ldots, \bar{v}_p\}) + 1$.

solution to this problem by drawing the separating hyperplanes $\{\bar{v} \in \mathbb{A} : \xi^{\mathsf{T}} \bar{v} = t\}$ corresponding to some separating covector ξ and some $t \in \mathbb{R}$, for three examples with $p = 4$. Since it will be shown that $|\mathcal{S}|$ is the number of these searched linearly separable bipartitions, this one is denoted that way in the figure. Obviously, $|\mathcal{S}|$ not only depends on p and $r := \dim(\mathrm{vect}\{\bar{v}_1, \ldots, \bar{v}_p\}) + 1$, but it also depends on the arrangement of the $\bar{v}_i$'s in the affine space $\mathbb{A}$. We also see that $|\mathcal{S}|$ cannot take all the even values (proposition 3.4.1) between its lower bound $2p = 8$ and its upper bounds 8 (if $r = 2$) and 14 (if $r = 3$) given by propositions 3.4.7 and 3.4.10.

The equivalence between the linearly separable bipartitioning problem 3.3.15 of this section and the vector inversion problem 3.3.12 (hence, with the original problem 3.3.1) is grounded on the following construction and proposition.

Construction 3.3.16. 1) Let be given two integers n and $p \in \mathbb{N}^*$ and p nonzero vectors $v_1, \ldots, v_p \in \mathbb{R}^n$ such that $K := \mathrm{cone}\{v_k : k \in [1 : p]\}$ is a pointed cone. From proposition 3.3.13, there is a direction $d \in \mathbb{R}^n$ such that

$$\|d\| = 1 \qquad \text{and} \qquad \big(\forall\, k \in [1 : p] : \quad v_k^{\mathsf{T}} d > 0\big).$$

Define

$$\mathbb{A} := \{\bar{v} \in \mathbb{R}^n : d^\mathsf{T}\bar{v} = 1\}, \qquad \mathbb{A}_0 := \mathbb{A} - \mathbb{A} = \{v \in \mathbb{R}^n : d^\mathsf{T}v = 0\},$$
$$\forall\, k \in [1:p]: \quad \bar{v}_k := v_k/(v_k^\mathsf{T}d) \in \mathbb{A}.$$

2) For a given bipartition (I, J) of $[1:p]$, define

$$K_I := \mathrm{cone}\{v_i : i \in I\} \quad \text{and} \quad K_J := \mathrm{cone}\{v_j : j \in J\}, \tag{3.24a}$$
$$C_I := K_I \cap \mathbb{A} \quad \text{and} \quad C_J := K_J \cap \mathbb{A}, \tag{3.24b}$$

with the convention $K_\varnothing = \{0\}$ and $C_\varnothing = \varnothing$. □

Proposition 3.3.17 (pointed cone after vector inversions)**.** *Adopt the construction* 3.3.16 *and take a partition* (I, J) *of* $[1:p]$*. Then, the following properties are equivalent :*

(i) $\mathrm{cone}((-K_I) \cup K_J)$ *is pointed,*
(ii) $K_I \cap K_J = \{0\}$,
(iii) $C_I \cap C_J = \varnothing$,
(iv) *there exists a vector* $\xi \in \mathbb{A}_0$ *such that* $\max_{i\in I}\, \xi^\mathsf{T}\bar{v}_i < \min_{j\in J}\, \xi^\mathsf{T}\bar{v}_j$.

Proof. $[(i) \Rightarrow (ii)]$ We show the contrapositive. If there is $v \in (K_I \cap K_J) \setminus \{0\}$, then $-v \in (-K_I) \subseteq \mathrm{cone}((-K_I) \cup K_J)$ and $v \in K_J \subseteq \mathrm{cone}((-K_I) \cup K_J)$. Therefore, $\mathrm{cone}((-K_I) \cup K_J)$ is not pointed.

$[(ii) \Rightarrow (iii)]$ $\varnothing = \mathbb{A} \cap \{0\} = \mathbb{A} \cap K_I \cap K_J$ $[(ii)] = (\mathbb{A} \cap K_I) \cap (\mathbb{A} \cap K_J) = C_I \cap C_J$.

$[(iii) \Rightarrow (iv)]$ We claim that

$$C_I \text{ is nonempty, convex and compact.}$$

Indeed, since C_I is nonempty (it contains the vectors $\bar{v}_i$ for $i \in I \neq \varnothing$), convex (because K_I and $\mathbb{A}$ are convex) and closed (because K_I and $\mathbb{A}$ are closed), it suffices to show that C_I is bounded or that its asymptotic cone (or recession cone in [221, p. 61]), namely $C_I^\infty = K_I \cap \mathbb{A}_0$, is reduced to $\{0\}$ [221, theorem 8.4]. This is indeed the case since $v^\mathsf{T}d > 0$ for all $v \in K_I \setminus \{0\}$. For the same reason,

$$C_J \text{ is nonempty, convex and compact.}$$

Now, since $C_I \cap C_J = \varnothing$ by (iii), one can strictly separate the convex sets C_I and C_J in $\mathbb{A}$ [221, corollary 11.4.2] : there exists $\xi \in \mathbb{A}_0$ such that $\xi^\mathsf{T}v < \xi^\mathsf{T}w$, for all $v \in C_I$ and all $w \in C_J$. This shows that (iv) holds.

$[(iv) \Rightarrow (i)]$ Since $\mathrm{cone}((-K_I) \cup K_J) = \mathrm{cone}(\{-v_i : i \in I\} \cup \{v_j : j \in J\})$, by proposition 3.3.13, it suffices to find $d_{(I,J)} \in \mathbb{R}^n$ such that

$$\Big(-v_i^\mathsf{T}d_{(I,J)} > 0, \quad \forall\, i \in I\Big) \qquad \text{and} \qquad \Big(v_j^\mathsf{T}d_{(I,J)} > 0, \quad \forall\, j \in J\Big). \tag{3.25}$$

By (iv) and the fact that $\theta \in (0,\pi) \to \cot\theta \in \mathbb{R}$ is surjective, one can determine $\theta \in (0,\pi)$ such that

$$\max_{i\in I} \frac{\xi^{\mathsf{T}} v_i}{v_i^{\mathsf{T}} d} < -\cot\theta < \min_{j\in J} \frac{\xi^{\mathsf{T}} v_j}{v_j^{\mathsf{T}} d}. \tag{3.26}$$

Since $\sin\theta > 0$ for $\theta \in (0,\pi)$ and since $v_k^{\mathsf{T}} d > 0$ for all $k \in [1:p]$, this is equivalent to

$$\max_{i\in I} v_i^{\mathsf{T}}[(\cos\theta)d + (\sin\theta)\xi] < 0 < \min_{j\in J} v_j^{\mathsf{T}}[(\cos\theta)d + (\sin\theta)\xi].$$

Therefore, (3.25) is satisfied with $d_{(I,J)} := (\cos\theta)d + (\sin\theta)\xi$. □

One can now establish the link between the pointed cone problem of section 3.3.3 (problem 3.3.12) and the linearly separable bipartitioning problem (problem 3.3.15).

Equivalence 3.3.18 (pointed cone $\leftrightarrow$ linearly separable bipartitioning)**.** Let be given a matrix $V \in \mathbb{R}^{n\times p}$ with nonzero columns denoted by v_1, …, v_p and take $s \in \mathcal{S}$, which is nonempty. By (3.23), $\mathrm{cone}\{s_i v_i : i \in [1:p]\}$ is pointed. Use the construction 3.3.16(1) with $v_i \curvearrowright s_i v_i$.

For $\tilde{s} \in \{\pm 1\}^p$, define a partition (I,J) of $[1:p]$ by

$$I := \{i \in [1:p] : \tilde{s}_i s_i = -1\} \quad \text{and} \quad J := \{i \in [1:p] : \tilde{s}_i s_i = +1\}.$$

Define also K_I and K_J by (3.24a) with $v_i \curvearrowright s_i v_i$. We claim that

$$\mathrm{cone}\{\tilde{s}_i v_i : i \in [1:p]\} \text{ is pointed} \iff \exists \xi \in \mathbb{A}_0 : \max_{i\in I} \xi^{\mathsf{T}} \bar{v}_i < \min_{j\in J} \xi^{\mathsf{T}} \bar{v}_j. \tag{3.27}$$

Indeed, one has

$$\begin{aligned} &\mathrm{cone}\{\tilde{s}_i v_i : i \in [1:p]\} \text{ is pointed} \\ &\quad\iff \mathrm{cone}\{\tilde{s}_i s_i (s_i v_i) : i \in [1:p]\} \text{ is pointed} \\ &\quad\iff \mathrm{cone}((-K_I) \cup K_J) \text{ is pointed} \\ &\quad\iff \exists \xi \in \mathbb{A}_0 : \max_{i\in I} \xi^{\mathsf{T}} \bar{v}_i < \min_{j\in J} \xi^{\mathsf{T}} \bar{v}_j, \end{aligned}$$

where we have used the equivalence $(i) \Leftrightarrow (iv)$ of proposition 3.3.17 ($v_i \curvearrowright s_i v_i$).

The equivalence (3.27) establishes the expected equivalence between the pointed cone problem 3.3.12 (in which one looks for all the $\tilde{s} \in \{\pm 1\}^p$ such that $\mathrm{cone}\{\tilde{s}_i v_i : i \in [1:p]\}$ is pointed) and the linearly separable bipartitioning problem 3.3.15 of the vectors $\bar{v}_i = s_i v_i/(s_i v_i^{\mathsf{T}} d) = v_i/(v_i^{\mathsf{T}} d)$, $i \in [1:p]$, where d is associated with the pointed cone $\mathrm{cone}\{s_i v_i : i \in [1:p]\}$ by the equivalence $(i) \Leftrightarrow (iii)$ of proposition 3.3.13. □

3.3.4 Discrete geometry : hyperplane arrangements

The equivalent problem examined in this section has a long history, going back at least to the XIXth century [239, 215]. More recently, it appears in *Computational Discrete Geometry* (the discipline has many other names), under the name of *hyperplane arrangements*.

Contributions to this problem, or a more general version of it, with a discrete mathematics point of view, have been reviewed in [114, 81, 236, 4, 118]. It has many applications [83, 231, 42]. From an algorithmic point of view, the algorithms developed in this domain can immediately be used to compute $\mathcal{S}$ defined by (3.12) or $\partial_B H(x)$ defined by (3.1) and (3.3).

Problem 3.3.19 (arrangement of hyperplanes containing the origin)**.** Let be given two positive integers n and $p \in \mathbb{N}^*$ and p nonzero vectors $v_1, \ldots, v_p \in \mathbb{R}^n$. Consider the hyperplanes containing the origin :

$$\mathcal{H}_i := \{d \in \mathbb{R}^n : v_i^{\mathsf{T}} d = 0\}. \tag{3.28}$$

Figure 3.3 illustrates problem 3.3.19 for the linear complementarity problem 3.3.2. It is requested to list the regions of $\mathbb{R}^n$ that are separated by these hyperplanes, which are the connected components of $\mathbb{R}^n \setminus (\bigcup_{i\in[1:p]} \mathcal{H}_i)$. Such a region is called a *cell* or a *chamber*, depending on the authors [14, 232, 4]. More specifically, let us define the half-spaces

$$\mathcal{H}_i^+ := \{d \in \mathbb{R}^n : v_i^{\mathsf{T}} d > 0\} \qquad \text{and} \qquad \mathcal{H}_i^- := \{d \in \mathbb{R}^n : v_i^{\mathsf{T}} d < 0\}.$$

The problem is to determine the following set of open sectors or cells of $\mathbb{R}^n$, indexed by the bipartitions (I_+, I_-) of $[1 : p]$:

$$\mathfrak{C} := \left\{(I_+, I_-) \in \mathfrak{B}([1 : p]) : (\cap_{i\in I_+} \mathcal{H}_i^+) \cap (\cap_{i\in I_-} \mathcal{H}_i^-) \neq \varnothing\right\}, \tag{3.29}$$

where $\mathfrak{B}([1 : p])$ denotes the set of bipartitions of $[1 : p]$. □

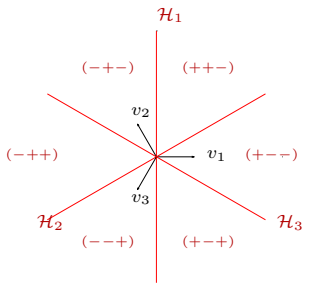


FIGURE 3.3 – Illustration of problem 3.3.19 (arrangement of hyperplanes containing the origin) for the 3 vectors that are the columns on the matrix V in example 3.3.2 (since the last components of these v_i's vanish, only the first two ones are represented above). The hyperplanes $\mathcal{H}_i$ are defined by (3.28). The regions to determine are represented by the sign vectors here denoted $(s_1 s_2 s_3)$ with $s_i = \pm$: if $d \in \mathbb{R}^2$ belongs to the region $(s_1 s_2 s_3)$, then $s_i = +$ if $v_i^{\mathsf{T}} d > 0$ and $s_i = -$ if $v_i^{\mathsf{T}} d < 0$. We see that there are only $6 = 2p$ regions among the $8 = 2^p$ possible ones; the regions $(+++)$ and $(---)$ are missing, which reflects the fact that $+v_1 + v_2 + v_3 = 0$ and $-v_1 - v_2 - v_3 = 0$ (see problem 3.3.6).

The link between problem 3.3.19 and the signed feasibility of strict linear inequality systems of section 3.3.2 is obtained from the bijection

$$\eta : (I_+, I_-) \in \mathfrak{B}([1 : p]) \mapsto s \in \{\pm 1\}^p, \text{ where } s_i = \begin{cases} +1 & \text{if } i \in I_+, \\ -1 & \text{if } i \in I_- \end{cases} \tag{3.30}$$

and the setting $V = \begin{pmatrix} v_1 & \cdots & v_p \end{pmatrix}$, whose columns are nonzero by assumption, here and in section 3.3.2. Recall the definition (3.12) of the set of sign vectors $\mathcal{S}$.

Proposition 3.3.20 (bijection $\mathfrak{C} \leftrightarrow \mathcal{S}$)**.** *For the matrix $V \in \mathbb{R}^{n\times p}$, with nonzero columns v_i's, the map η given by* (3.30) *is a bijection from $\mathfrak{C}$ onto $\mathcal{S}$.*

Proof. Let $(I_+, I_-) \in \mathfrak{B}([1:p])$ and $s := \eta((I_+, I_-))$. Then,

$$\begin{aligned}(I_+, I_-) \in \mathfrak{C} &\iff \exists\, d \in (\cap_{i\in I_+} \mathcal{H}_i^+) \cap (\cap_{i\in I_-} \mathcal{H}_i^-)\\ &\iff \exists\, d \in \mathbb{R}^n : (v_i^\mathsf{T} d > 0 \text{ for } i \in I_+) \text{ and } (v_i^\mathsf{T} d < 0 \text{ for } i \in I_-)\\ &\iff \exists\, d \in \mathbb{R}^n : s \bullet (V^\mathsf{T} d) > 0\\ &\iff s \in \mathcal{S}.\end{aligned}$$

These equivalences show the bijectivity of η from $\mathfrak{C}$ onto $\mathcal{S}$. □

Equivalence 3.3.21 (signed linear system feasibility $\leftrightarrow$ hyperplane arrangement)**.** The equivalence between problems 3.3.3 and 3.3.19 follows from the bijection of the map $\eta : \mathfrak{C} \to \mathcal{S}$ claimed in proposition 3.3.20. □

3.4 Description of the B-differential

This section gives some elements of description of the B-differential $\partial_B H(x)$, when H is the piecewise affine function given by (3.3) and $x \in \mathbb{R}^n$. This description is often carried out in terms of the matrix V defined by (3.11), whose p columns are denoted by $v_1, \ldots, v_p \in \mathbb{R}^n$ and are nonzero by construction. When the properties are given for $\mathcal{S}$, one may have $p \geqslant n$ and the referenced matrix $V \in \mathbb{R}^{n\times p}$ is *assumed* to have nonzero columns, which implies that $\mathcal{S} \neq \varnothing$. Some properties of $\partial_B H(x)$ are given in section 3.4.1, including those that are useful in [74]. Section 3.4.2 deals with the cardinality $|\partial_B H(x)|$ of the B-differential. Section 3.4.3 analyzes more precisely two particular configurations. Section 3.4.4 highlights two links between the B-differential and the C-differential of H.

Besides their theoretical relevance, the properties of the B-differential of H given in this section will also be useful to design the algorithms presented in section 3.5 and to check the correctness of their implementation.

As a preliminary remark, let us mention a way of proceeding that seems to us to be a dead end when one focuses on the B-differential $\partial_B H(x)$ and that does not make possible the description of a hyperplane arrangement governed by a matrix $V \in \mathbb{R}^{n\times p}$ with $p > n$. Therefore, this approach is not followed below. If $\partial_B H(x)$ is the main concern, one can write $H(x) = Ax + a - K(x)$, where $K(x) := \mathrm{P}_{\mathbb{R}^n_+}[Mx + q]$, $\mathrm{P}_{\mathbb{R}^n_+}$ is the orthogonal projector on the positive orthant, $M = A - B$ and $q = a - b$, so that $\partial_B H(x) = A - \partial_B K(x)$. To take advantage of the explicit formula of $\partial_B \mathrm{P}_{\mathbb{R}^n_+}$, one can look for conditions ensuring that the chain rule applies for the composition defining the map $K : \mathbb{R}^n \to \mathbb{R}^n$. It can be shown, however, that, when the chain rule applies, the B-differential $\partial_B H(x)$ is complete in the sense of definition 3.2.4, which is a very particular case ; see [78] for more details. Therefore, this approach is of too limited an interest.

3.4.1 Some properties of the B-differential

Let us start with a basic property of $\partial_B H(x)$, which is its symmetry in the sense of definitions 3.2.5. This property has been observed by many in other contexts [4, § 1.1.4], so that we leave its short proof, based on the equivalence 3.3.5, to [78]. It is useful for the algorithms since it implies that only half of the B-differential has to be computed.

Proposition 3.4.1 (symmetry of $\boldsymbol{\partial_B H(x)}$)**.** *Suppose that $p > 0$. Then, the B-differential $\partial_B H(x)$ is symmetric and $|\partial_B H(x)|$ is even.*

We now give a necessary and sufficient condition ensuring the completeness of $\partial_B H(x)$ in the sense of definition 3.2.4. The condition was shown to be sufficient in [255, corollary 2.1(i)] for the nonlinear case (3.6), using a different proof, but we shall see in [74] that it is an easy consequence of that property in the affine case (3.3). Thanks to the equivalence 3.3.5, the present proof is short. This property is also useful in the development of algorithms, as a test that these must pass : $|\partial_B H(x)| = 2^p$ if and only if $V \in \mathbb{R}^{n\times p}$ is injective.

Proposition 3.4.2 (completeness of the B-differential)**.** *The B-differential $\partial_B H(x)$ of H at x is complete if and only if the matrix $V \in \mathbb{R}^{n\times p}$ in* (3.11) *is injective. Hence, this property can hold only if $p \leqslant n$.*

Proof. [$\Rightarrow$] We show the contrapositive. Assume that V is not injective, so that $V\alpha = 0$ for some nonzero $\alpha \in \mathbb{R}^p$. With $s \in \operatorname{sgn}(\alpha)$, one can write

$$\sum_{i\in[1:p]} |\alpha_i| s_i v_i = 0.$$

By Gordan's alternative (3.17), it follows that there is no $d \in \mathbb{R}^n$ such that $s \bullet (V^\mathsf{T} d) > 0$. By (3.12), this implies that $s \notin \mathcal{S}$. According to the equivalence 3.3.5, $\sigma^{-1}(s) \notin \partial_B H(x)$, showing that the B-differential is not complete.

[$\Leftarrow$] Assume the injectivity of V. Let $s \in \{\pm 1\}^p$. Since V^T is surjective, the system $V^\mathsf{T} d = s$ holds for some $d \in \mathbb{R}^n$. For this d, $s \bullet (V^\mathsf{T} d) = e$, so that $s \bullet (V^\mathsf{T} d) > 0$ holds for some $d \in \mathbb{R}^n$, which implies that the selected s is in $\mathcal{S}$. We have shown that $\mathcal{S} = \{\pm 1\}^p$ or that $\partial_B H(x) = \sigma^{-1}(\{\pm 1\}^p)$ (σ^{-1} is defined by (3.14b)) is complete. □

We focus now on the connectivity of $\partial_B H(x)$, a notion that is more easily presented in terms of $\mathcal{S} \subseteq \{\pm 1\}^p$ but that can be transferred straightforwardly to $\partial_B H(x)$ by the bijection σ defined in (3.14). This property was implicitly used, for instance, in the algorithms proposed by Avis, Fukuda and Sleumer [14, 232] for hyperplane arrangements.

Definition 3.4.3 (adjacency in $\boldsymbol{\{\pm 1\}^p}$)**.** Two sign vectors s^1 and $s^2 \in \{\pm 1\}^p$ are said to be *adjacent* if they differ by a single component (i.e., the vertices s^1 and s^2 of the cube $\operatorname{co}\{\pm 1\}^p$ can be joined by a single edge). □

Definitions 3.4.4 (connectivity in $\{\pm 1\}^p$)**.** A *path of length l in a subset S* of $\{\pm 1\}^p$ is a finite set of sign vectors $s^0, \ldots, s^l \in S$ such that s^i and s^{i+1} are adjacent for all $i \in [0\,{:}\,l-1]$; in which case the path is said to be *joining* s^0 to s^l. One says that a subset S of $\{\pm 1\}^p$ is *connected* if any pair of points of S can be joined by a path in S. □

Proposition 3.4.5 (connectivity of the B-differential)**.** *The set $\mathcal{S}$ defined by* (3.12) *is connected if and only if V has no colinear columns. In this case, any points s and $\tilde{s}$ of $\mathcal{S}$ can be joined by a path of length $l := \sum_{i\in[1:p]} |\tilde{s}_i - s_i|/2 \leqslant p$ in $\mathcal{S}$.*

Proof. [⇒] We prove the contrapositive. Suppose that the columns v_i and v_j of V are colinear : $v_j = \alpha v_i$, for some $\alpha \in \mathbb{R}^*$. Assume that $\alpha > 0$ (resp. $\alpha < 0$). By (3.12), for any $s \in \mathcal{S} \neq \varnothing$, one can find $d \in \mathbb{R}^n$ such that $s \bullet (V^{\mathsf{T}} d) > 0$, implying that $s_i = s_j$ (resp. $s_i = -s_j$). Therefore, one cannot find a path in $\mathcal{S}$ joining $s \in \mathcal{S}$ and $-s \in \mathcal{S}$ (proposition 3.4.1), since one would have to change the two components with index in $\{i, j\}$ and that these components must be changed simultaneously for the sign vectors in $\mathcal{S}$, while the adjacency property along a path prevents from changing more than one sign at a time.

[⇐] We leave to [78] the proof of this implication and of the last claim of the proposition, since the conclusion of the implication is given in [4, section 1.10.4] as a simple observation with a very different point of view, related to graph theory. □

For $k \in [1 : p]$, we introduce

$$\mathcal{S}_k := \{s \in \{\pm 1\}^k : \ \exists\, d \in \mathbb{R}^n \text{ such that } s_i v_i^{\mathsf{T}} d > 0 \text{ for } i \in [1 : k]\}. \tag{3.33}$$

We also note $\mathcal{S}_k^c := \{\pm 1\}^k \setminus \mathcal{S}_k$. Hence $\mathcal{S} = \mathcal{S}_p$ and $\mathcal{S}^c = \mathcal{S}_p^c$. Point 1 of the next proposition will be used to motivate an improvement of algorithm 3.5.5 in section 3.5.2 and its points 2 and 3 will be used to get the equivalence in proposition 3.4.13, related to a fan arrangement.

Proposition 3.4.6 (incrementation)**.** 1) *If $s \in \mathcal{S}_k^c$, then $(s, \pm 1) \in \mathcal{S}_{k+1}^c$. In particular, $|\mathcal{S}_{k+1}^c| \geqslant 2|\mathcal{S}_k^c|$.*

2) *If $v_{k+1} \notin \operatorname{vect}\{v_1, \ldots, v_k\}$, then, $(s, \pm 1) \in \mathcal{S}_{k+1}$ for all $s \in \mathcal{S}_k$. In particular, $|\mathcal{S}_{k+1}| = 2|\mathcal{S}_k|$ and $|\mathcal{S}_{k+1}^c| = 2|\mathcal{S}_k^c|$.*

3) *If v_{k+1} is not colinear to any of the vectors v_1, …, v_k, then, $[(s, \pm 1)$ and $(-s, \pm 1) \in \mathcal{S}_{k+1}$ for one $s \in \mathcal{S}_k]$ and $[(s', +1)$ or $(s', -1) \in \mathcal{S}_{k+1}$ for any $s' \in \mathcal{S}_k]$. In particular, $|\mathcal{S}_{k+1}| \geqslant |\mathcal{S}_k| + 2$.*

Proof. 1) If $s \in \mathcal{S}_k^c$, there is no $d \in \mathbb{R}^n$ such that $s_i v_i^{\mathsf{T}} d > 0$ for $i \in [1 : k]$. Therefore, there is no $d \in \mathbb{R}^n$ such that ($s_i v_i^{\mathsf{T}} d > 0$ for $i \in [1 : k]$) and $\pm v_{k+1}^{\mathsf{T}} d > 0$. Hence, $(s, \pm 1) \in \mathcal{S}_{k+1}^c$. This implies that $|\mathcal{S}_{k+1}^c| \geqslant 2|\mathcal{S}_k^c|$.

2) Let P be the orthogonal projector on $\operatorname{vect}\{v_1, \ldots, v_k\}^\perp$ for the Euclidean scalar product. By assumption, $\mathrm{P}\, v_{k+1} \neq 0$. Let $s \in \mathcal{S}_k$, so that there is a direction $d \in \mathbb{R}^n$ such that $s_i v_i^{\mathsf{T}} d > 0$ for $i \in [1 : k]$. For any $t \in \mathbb{R}$ and $i \in [1 : k]$, the directions $d_\pm := d \pm t\, \mathrm{P}\, v_{k+1}$ verify $s_i v_i^{\mathsf{T}} d_\pm = s_i v_i^{\mathsf{T}} d > 0$ (because $v_i^{\mathsf{T}}\, \mathrm{P}\, v_{k+1} = 0$). In addition, for $t > 0$ sufficiently large, one has $\pm v_{k+1}^{\mathsf{T}} d_\pm = \pm v_{k+1}^{\mathsf{T}} d + t \| \mathrm{P}\, v_{k+1}\|^2 > 0$ (because $\mathrm{P}^2 = \mathrm{P}$ and $\mathrm{P}^{\mathsf{T}} = \mathrm{P}$). We have shown that both $(s, +1)$ and $(s, -1)$ are in $\mathcal{S}_{k+1}$. Therefore, $|\mathcal{S}_{k+1}| \geqslant 2|\mathcal{S}_k|$.

Now, $|\mathcal{S}_k| + |\mathcal{S}_k^c| = 2^k$, $|\mathcal{S}_{k+1}| + |\mathcal{S}_{k+1}^c| = 2^{k+1}$ and $|\mathcal{S}_{k+1}^c| \geqslant 2|\mathcal{S}_k^c|$ by point 1. Therefore, one must have $|\mathcal{S}_{k+1}| = 2|\mathcal{S}_k|$ and $|\mathcal{S}_{k+1}^c| = 2|\mathcal{S}_k^c|$.

3) We claim that one can find a direction $d \in \mathbb{R}^n$ such that

$$\Big(\forall\, i \in [1:k] :\ v_i^\mathsf{T} d \neq 0\Big) \qquad \text{and} \qquad v_{k+1}^\mathsf{T} d = 0. \tag{3.34}$$

Indeed, let $\mathbb{E} := \{d \in \mathbb{R}^n : v_{k+1}^\mathsf{T} d = 0\}$ and P be the orthogonal projector on $\mathbb{E}$ for the Euclidean scalar product. By lemma 3.2.6, one can find a direction $d \in \mathbb{E}$ (hence $v_{k+1}^\mathsf{T} d = 0$) such that $|\{(\mathrm{P}\, v_i)^\mathsf{T} d : i \in [1:k+1]\}| = |\{\mathrm{P}\, v_i : i \in [1:k+1]\}|$. Since $\mathrm{P}\, v_{k+1} = 0$ and $\mathrm{P}\, v_i \neq 0$ for $i \in [1:k]$ (because the v_i's are not colinear with v_{k+1}), one has $(\mathrm{P}\, v_i)^\mathsf{T} d \neq 0$ for $i \in [1:k]$. Since, $0 \neq (\mathrm{P}\, v_i)^\mathsf{T} d = v_i^\mathsf{T}\, \mathrm{P}\, d = v_i^\mathsf{T} d$, (3.34) follows.

Taking $s_i := \mathrm{sgn}(v_i^\mathsf{T} d)$ for $i \in [1:k]$, one deduces from (3.34) that there is a direction $d \in \mathbb{R}^n$ such that

$$\Big(\forall\, i \in [1:k] :\ s_i v_i^\mathsf{T} d > 0\Big) \qquad \text{and} \qquad v_{k+1}^\mathsf{T} d = 0.$$

It follows that, for $\varepsilon > 0$ sufficiently small, the directions $d_\pm := d \pm \varepsilon v_{k+1}$ satisfy

$$\Big(\forall\, i \in [1:k] :\ s_i v_i^\mathsf{T} d_\pm > 0\Big) \qquad \text{and} \qquad \pm\, v_{k+1}^\mathsf{T} d_\pm > 0.$$

This means that $(s, \pm 1) \in \mathcal{S}_{k+1}$. By symmetry (proposition 3.4.1), one also has $(-s, \pm 1) \in \mathcal{S}_{k+1}$, so that we have found 4 vectors in $\mathcal{S}_{k+1}$. Now, since, for any $s' \in \mathcal{S}_k \setminus \{\pm s\}$ (in number $|\mathcal{S}_k| - 2$), either $(s', +1) \in \mathcal{S}_{k+1}$ or $(s', -1) \in \mathcal{S}_{k+1}$, it follows that $|\mathcal{S}_{k+1}| \geqslant 4 + (|\mathcal{S}_k| - 2) = |\mathcal{S}_k| + 2$. □

3.4.2 Cardinality of the B-differential

Information on the cardinality of $\partial_B H(x)$ can be useful to check the correctness of the number of elements computed by the algorithms presented in section 3.5.2.

Winder's formula

Giving the exact number of elements in $\partial_B H(x)$, that is $|\partial_B H(x)| = |\mathcal{S}| = |\mathfrak{C}| = 2^p - |\mathcal{S}^c| = 2^p - |\mathcal{I}|$, with the notation (3.12), (3.29) and (3.16), is a tricky task, even in the present affine case, since it subtly depends on the arrangement of the vectors v_i's in the space (see figure 3.2). Many contributions have been done on this subject; the earliest we cite dates from 1826 [239, 215, 114, 257, 151, 7, 81, 52, 236, 4]. The formula (3.35) for $|\partial_B H(x)|$ is due to Winder [253, p. 1966] and reads for the matrix V with nonzero columns given by (3.11)

$$|\partial_B H(x)| = \sum_{I \subseteq [1:p]} (-1)^{\mathrm{null}(V_{:,I})}, \tag{3.35}$$

where $\mathrm{null}(V_{:,I})$ is the nullity of $V_{:,I}$ and the term in the right-hand side corresponding to $I = \varnothing$ is 1 (one takes the convention that $\mathrm{null}(V_{:,\varnothing}) = 0$). Note that, in this formula,

the columns of V can be colinear with each other. This amazing expression, with its only algebraic nature, potentially made of positive and negative terms, is explicit but, to our knowledge, has not been at the origin of a method to list the elements of $\partial_B H(x)$. We give in [78] a proof of (3.35) that follows the same line of reasoning as the one of Winder [253], but that is more analytic in that it uses the sign vectors introduced in section 3.3.2 rather than geometric arguments (i.e., the hyperplane arrangements of section 3.3.4).

Bounds

When p is large, computing the cardinality $|\partial_B H(x)|$ from (3.35) by evaluating the 2^p ranks $\mathrm{rank}(V_{:,I})$ for $I \subseteq [1:p]$ could be excessively expensive. Therefore, having simple-to-compute lower and upper bounds on $|\partial_B H(x)|$ may be useful in some circumstances, including theoretical ones. Proposition 3.4.7 gives elementary lower and upper bounds, while proposition 3.4.10 reinforces the upper bound, thanks to a lower semicontinuity argument (proposition 3.4.8). Necessary and sufficient conditions ensuring equality in the left-hand side or right-hand side inequalities in the next proposition are given in section 3.4.3.

Proposition 3.4.7. (lower and upper bounds on $|\partial_B H(x)|$) *Let $V \in \mathbb{R}^{n\times p}$ given by* (3.11) *and $r := \mathrm{rank}(V)$. Then,*

$$2^r \leqslant |\partial_B H(x)| \leqslant 2^p. \tag{3.36a}$$

If V has no colinear columns, then,

$$\max(2p, 2^r) \leqslant 2^r + 2(p-r) \leqslant |\partial_B H(x)|. \tag{3.36b}$$

Proof. [(3.36a)] One can assume that the first r columns of V are linearly independent, so that $|\mathcal{S}_r| = 2^r$ (notation (3.33) and proposition 3.4.6(2)). Since a sign vector has at least one descendant, $|\mathcal{S}_{k+1}| \geqslant |\mathcal{S}_k|$, for $k \in [r:p]$, which proves the lower bound. The upper bound was already mentioned in proposition 3.2.2.

[(3.36b)] The first inequality is clear since $p \geqslant r \geqslant 1$ and $2r \leqslant 2^r$. Consider now the second inequality. Like above, one can assume that the first r columns of V are linearly independent, so that $|\mathcal{S}_r| = 2^r$. Next, by proposition 3.4.6(3) and the non-colinearity of the columns of V, $|\mathcal{S}_{r+1}| \geqslant 2^r + 2$. By induction, The inequality follows. □

Proposition 3.4.10 below provides a refinement of the upper bound given by (3.36a). The next proposition will be useful for this purpose. Recall that a function $\varphi : x \in \mathbb{T} \to \varphi(x) \in \mathbb{R}$, defined on a topological space $\mathbb{T}$, is said to be *lower semicontinuous* if, for any $x \in \mathbb{T}$ and any $\varepsilon > 0$, there is a neighborhood $\mathcal{V}$ of x such that, for all $\tilde{x} \in \mathcal{V}$, one has $\varphi(\tilde{x}) \leqslant \varphi(x)+\varepsilon$. It is known that the rank of a matrix can only increase in the neighborhood of a given matrix, which implies its lower semicontinuity. The next lemma shows that the same property holds for $|\mathcal{S}| \in \mathbb{N}^*$, viewed as a function of V. Recall that the bijection σ is defined by (3.14).

Proposition 3.4.8 (lower semicontinuity of $|\partial_B H(x)|$)**.** *Suppose that the set $\mathcal{S}$, defined by* (3.12)*, is viewed as a function of $V \in \mathbb{R}^{n\times p}$. Then, $\mathcal{S}(V) \subseteq \mathcal{S}(\tilde{V})$ for $\tilde{V}$ near V in $\mathbb{R}^{n\times p}$. In particular, $V \in \mathbb{R}^{n\times p} \mapsto |\mathcal{S}(V)| \in \mathbb{N}^*$ is lower semicontinuous.*

Proof. By the definition (3.12) of $\mathcal{S}(V)$, for all $s \in \mathcal{S}(V)$, there is a $d_s \in \mathbb{R}^n$ such that $s \bullet (V^\mathsf{T} d_s) > 0$. Clearly, one still has $s \bullet (\tilde{V}^\mathsf{T} d_s) > 0$, for $\tilde{V}$ near V. Since $\mathcal{S}(V)$ is finite, there is a neighborhood $\mathcal{V}$ of V, such that, for $\tilde{V} \in \mathcal{V}$ and $s \in \mathcal{S}(V)$, there is a $d \in \mathbb{R}^n$ such that $s \bullet (\tilde{V}^\mathsf{T} d) > 0$ or $s \in \mathcal{S}(\tilde{V})$. We have shown that $\mathcal{S}(V) \subseteq \mathcal{S}(\tilde{V})$ for $\tilde{V}$ near V.

As a direct consequence of this inclusion, we have that $|\mathcal{S}(V)| \leqslant |\mathcal{S}(\tilde{V})|$ for $\tilde{V}$ near V. The lower semicontinuity of $V \mapsto |\mathcal{S}(V)|$ follows. □

Proposition 3.4.2 establishes a necessary and sufficient condition to have completeness of $\partial_B H(x)$. Here follows a less restrictive assumption, called *general position*, which is equivalent to have equality in (3.39) below. In connection with this assumption, it is worth noting that, for a matrix $V \in \mathbb{R}^{n\times p}$ of rank r, one has

$$\forall\, I \subseteq [1:p]: \quad \operatorname{rank}(V_{:,I}) \leqslant \min(|I|, r). \tag{3.37}$$

Definition 3.4.9 (general position)**.** The vectors $v_1, \ldots, v_p \in \mathbb{R}^n$ are said to be in *general position*, if the matrix $V := \begin{pmatrix} v_1 & \cdots & v_p \end{pmatrix}$ verifies

$$\forall\, I \subseteq [1:p]: \quad \operatorname{rank}(V_{:,I}) = \min(|I|, r), \tag{3.38}$$

where $r := \operatorname{rank}(V)$. □

In the matroid terminology, the vector matroid formed by the columns of V in general position is said to be uniform [191, example 1.2.7]. The general position notion is used by Winder [253] when $r = n$. Example of vectors in general position are those in the left-hand side and right-hand side panes in figure 3.2 (the points are the normalized vectors $\bar{v}_i$'s so that the v_i's are actually in $\mathbb{R}^3$); note that in the first case $2 = r < n = 3$. Those in the middle pane are not in general position. This is due to the fact that $r := \operatorname{rank}(V) = 3$ while for the 3 bottom vectors, with indices in I say, one has $\min(|I|, r) - \operatorname{rank}(V_{:,I}) = 3 - 2 \neq 0$.

Equality in the upper estimate (3.39) of the next proposition was shown by Winder [253, 1966, corollary] when the columns of V are in general position and $r = n$, thanks to the identity (3.35). Long before him, the Swiss mathematician Ludwig Schläfli [227, p. 211] established the identity under the same assumptions, before 1852 [227, p. 174], without reference to (3.35), which was probably not known at that time. Note that equality does not hold in (3.39) for the middle configuration in figure 3.2 since $|\partial_B H(x)| = 12$, while the right-hand side of (3.39) reads $2[\binom{3}{0} + \binom{3}{1} + \binom{3}{2}] = 14$ (we have seen that the vectors in this pane are not in general position). The bound (3.39) is also useful to check the behavior of the algorithms for test-cases in which the columns of V are in general position. This is likely to be so for randomly generated V, and it was verified by all our random test-cases in section 3.5.2(B.1).

Proposition 3.4.10 (upper bound on $|\partial_B \boldsymbol{H}(\boldsymbol{x})|$)**.** *For V given by* (3.11) *and $r := \operatorname{rank}(V)$, one has*

$$|\partial_B H(x)| \leqslant 2 \sum_{i\in[0:r-1]} \binom{p-1}{i}, \tag{3.39}$$

with equality if and only if (3.38) *holds.*

Proof. 1) The proof of the implication "(3.38) ⇒ (3.39) with equality" is established in [253, corollary], using the identity (3.35). See also [78].

2) Let us now show that (3.39) holds. Below, we systematically identify $\partial_B H(x)$ and $\mathcal{S}$, thanks to the equivalence 3.3.5. We also note $\mathcal{S} \equiv \mathcal{S}(V)$ to stress the dependence of $\mathcal{S}$ on V. Let β be the right-hand side of (3.39). We proceed by contradiction, assuming that there is a matrix $V \in \mathbb{R}^{n\times p}$ of rank r such that

$$|\mathcal{S}(V)| > \beta. \tag{3.40a}$$

It certainly suffices to show that one can find a matrice $\tilde{V} \subseteq \mathbb{R}^{n\times p}$ of rank r arbitrarily close to V that satisfies

$$|\mathcal{S}(\tilde{V})| = \beta, \tag{3.40b}$$

since then one would have the expected contradiction with the lower semicontinuity of $V \mapsto |\mathcal{S}(V)|$ ensured by proposition 3.4.8 :

$$|\mathcal{S}(\tilde{V})| = \beta < |\mathcal{S}(V)|.$$

To find $\tilde{V}$ of rank r arbitrarily close to V verifying (3.40b), we proceed as follows. Since (3.40a) holds, the first part of the proof implies that V does not satify (3.38). Our goal is to construct from V a matrix $\tilde{V}$ of rank r arbitrarily close to V with columns in general position. Then, $\tilde{V}$ satisfies (3.40b) by the first part of the proof.

In view of (3.37) and since V does not satisfy (3.38), there is some $I \subseteq [1 : p]$ such that $\mathrm{rank}(V_{:,I}) < \min(|I|, r)$. By linear algebra arguments (see [78] for more details), one can get an arbitrarily small perturbation $\tilde{V}_{:,I}$ of $V_{:,I}$, such that $\mathrm{rank}(\tilde{V}_{:,I}) = \min(|I|, r)$ and $\mathcal{R}(\tilde{V}_{:,I}) \subseteq \mathcal{R}(V)$. Next, one forms $\tilde{V} \in \mathbb{R}^{n\times p}$ by setting $\tilde{V}_{:,I^c} = V_{:,I^c}$, so that $\tilde{V}$ is as close to V as desired and verifies $\mathcal{R}(\tilde{V}) \subseteq \mathcal{R}(V)$. The perturbation $\tilde{V}_{:,I}$ of $V_{:,I}$ can also perturb $V_{:,I'}$ for other index sets $I' \subseteq [1 : p]$. However, one has $\mathrm{rank}(\tilde{V}_{:,I'}) \leqslant \min(|I'|, r)$ by (3.37). Now, by the property of the rank, which can only increase in a neighborhood of a given matrix, if the perturbation taken above is sufficiently small, one has $\mathrm{rank}(V_{:,I'}) \leqslant \mathrm{rank}(\tilde{V}_{:,I'}) \leqslant \min(|I'|, r)$ for any $I' \subseteq [1 : p]$. Therefore, $\mathrm{rank}(V_{:,I'}) = \min(|I'|, r)$ implies that $\mathrm{rank}(\tilde{V}_{:,I'}) = \min(|I'|, r)$. As a result, the modification of V into $\tilde{V}$ described above increases by at least one the number of intervals $I' \subseteq [1 : p]$ such that $\mathrm{rank}(\tilde{V}_{:,I'}) = \min(|I'|, r)$. Since the number of such intervals is finite, proceeding similarly with all the nonempty index sets $I'' \subseteq [1 : p]$ such that $\mathrm{rank}(\tilde{V}_{:,I''}) < \min(|I''|, r)$, one finally obtains a matrix $\tilde{V}$, arbitrarily close to V, such that (3.38) holds : $\mathrm{rank}(\tilde{V}_{:,I}) = \min(|I|, r)$ for all $I \subseteq [1 : p]$.

3) One still has to show that "(3.39) with equality ⇒ (3.38)". We proceed by contradiction, assuming that (3.39) holds with equality for $\partial_B H(x) \equiv \mathcal{S}(V)$, but that (3.38) does not hold. By (3.37), there exists $I \subseteq [1 : p]$ such that

$$\mathrm{rank}(V_{:,I}) < \min(|I|, r). \tag{3.40c}$$

Let $\beta = |\mathcal{S}(V)|$ be the right-hand side of (3.39). It certainly suffices to show that, thanks to (3.40c), one can find a matrix $\tilde{V} \in \mathbb{R}^{n\times p}$ such that $\operatorname{rank}(\tilde{V}) \leqslant r$ and $|\mathcal{S}(\tilde{V})| > \beta$, since this would be in contradiction with what has been shown in part 2 of the proof. This matrix $\tilde{V}$ is obtained by perturbing V. By proposition 3.4.8, if the perturbation is sufficiently small, one has $\mathcal{S}(V) \subseteq \mathcal{S}(\tilde{V})$, so that it suffices to show that $\mathcal{S}(\tilde{V})$ contains a sign vector s that is not in $\mathcal{S}(V)$.

We claim that (3.40c) implies that one can find an index set $J \subseteq I$ such that

$$V_{:,J} \text{ is not injective} \qquad \text{and} \qquad |J| \leqslant r. \tag{3.40d}$$

Indeed, if $|I| \leqslant r$, one can take $J = I$ to satisfy (3.40d), since $\operatorname{rank}(V_{:,I}) < |I|$ by (3.40c), so that $V_{:,I}$ is not injective. If $|I| > r$, then $\operatorname{rank}(V_{:,I}) < r$ by (3.40c), which implies that any $J \subseteq I$ such that $|J| = r$ satisfies (3.40d).

Since $V_{:,J}$ is not injective, one can find $\alpha_J \in \mathbb{R}^J \setminus \{0\}$ such that

$$0 = \sum_{j\in J} \alpha_j v_j = \sum_{j\in J} \tilde{s}_j |\alpha_j| v_j,$$

for some $\tilde{s}_J \in \{\pm 1\}^J$ satisfying $\tilde{s}_j \in \operatorname{sgn}(\alpha_j)$ for all $j \in J$. Then, by Gordan's alternative (3.17),

$$\nexists\, d \in \mathbb{R}^n : \quad \tilde{s}_j v_j d > 0, \ \text{ for all } j \in J.$$

This implies that there is no $s \in \mathcal{S}(V)$ such that $s_J = \tilde{s}_J$. To conclude the proof, it suffices now to show that one can construct an arbitrarily small perturbation $\tilde{V}$ of V, such that $\mathcal{R}(\tilde{V}) \subseteq \mathcal{R}(V)$ and with an $s \in \mathcal{S}(\tilde{V})$ satisfying $s_J = \tilde{s}_J$.

Let $J^c := [1:p]\backslash J$. By (3.40d), $|J| \leqslant r \leqslant n$ so that one can find vectors $\{\tilde{v}_j : j \in [1:p]\}$, such that $\tilde{v}_j = v_j$ for $j \in J^c$, the vectors $\{\tilde{v}_j : j \in J\}$ are linearly independent, $\tilde{v}_j - v_j$ is arbitrarily small and $\{\tilde{v}_j : j \in [1:p]\} \subseteq \mathcal{R}(V)$. Since the vectors $\{\tilde{v}_j : j \in J\}$ are linearly independent, one can find a direction $d_0 \in \mathbb{R}^n$ such that $\tilde{v}_j^\mathsf{T} d_0 = \tilde{s}_j$ for $j \in J$, hence

$$\tilde{s}_j \tilde{v}_j^\mathsf{T} d_0 > 0, \quad \forall\, j \in J. \tag{3.40e}$$

Set $\tilde{s}_j = 1$ for $j \in J^c$. Let d be a discriminating covector given by lemma 3.2.6 (there denoted ξ) for the vectors $\{0\} \cup \{\tilde{s}_i v_i : i \in [1:p]\}$ sufficiently close to d_0. It results that $\tilde{s}_j \tilde{v}_j^\mathsf{T} d > 0$ for $j \in J$ (by (3.40e)) and that $\tilde{s}_j \tilde{v}_j^\mathsf{T} d \neq 0$ for $j \in J^c$. Finally, we see that the sign vector $s \in \{\pm 1\}^p$ defined by $s_i = \operatorname{sgn}(\tilde{v}_i^\mathsf{T} d)$ for all $i \in [1:p]$ is in $\mathcal{S}(\tilde{V})$ and satisfies $s_J = \tilde{s}_J$, as desired. □

Corollary 3.4.11 (stability of the sign vector set). *The sign vector set $\mathcal{S} \subseteq \{\pm 1\}^p$ defined by* (3.12) *is unchanged by small variations of the matrix $V \in \mathbb{R}^{n\times p}$ preserving its rank, provided the columns $v_1, \ldots, v_p \in \mathbb{R}^n$ of V are in general position in the sense of definition* 3.4.9.

Proof. If $\tilde{V}$ is near V, $\mathcal{S}(V) \subseteq \mathcal{S}(\tilde{V})$ by proposition 3.4.8. If the columns of V are in general position, proposition 3.4.10 tells us that $|\mathcal{S}(V)| = \beta$, where β is the right-hand side of Schläfli's bound (3.39) with $r = \operatorname{rank}(V)$. Now, by the fact that $\operatorname{rank}(\tilde{V}) = r$, proposition 3.4.10 ensures that $|\mathcal{S}(\tilde{V})| \leqslant \beta$. Therefore, one must have $\mathcal{S}(\tilde{V}) = \mathcal{S}(V)$. □

3.4.3 Particular configurations

We consider in this section some particular matrices $V \in \mathbb{R}^{n\times p}$ given by (3.11), which may be useful to get familiar with the B-differential of H. For these V's, $|\partial_B H(x)|$ can be computed easily. We consider two matrices V with the property that $r := \operatorname{rank}(V)$ takes the value 2 or p; they yield the lower and upper bounds on $|\partial_B H(x)|$ given by proposition 3.4.7. The lower bound $2p$ applies to the left-hand side pane of figure 3.2. As shown by the intermediate pane in figure 3.2, however, $|\partial_B H(x)|$ does not only depend on r.

Proposition 3.4.12 (injective matrix)**.** *The matrix $V \in \mathbb{R}^{n\times p}$ given by* (3.11) *is injective if and only if $|\partial_B H(x)| = 2^p$.*

Proof. Indeed, by proposition 3.4.2, the B-differential $\partial_B H(x)$ is complete (meaning that it is equal to $\partial_B^\times H(x)$, given by (3.10)) if and only if V is injective. Clearly, the completeness of $\partial_B H(x)$ is equivalent to $|\partial_B H(x)| = 2^p$. □

Proposition 3.4.13 (fan arrangement)**.** *If $p \geqslant 2$ and the vectors v_i's are not two by two colinear, one has $\operatorname{rank}(V) = 2$ if and only if $|\partial_B H(x)| = 2p$.*

Proof. [⇒] A short proof leverages Schläfli's bound (3.39) with equality. Since the v_i's are not two by two colinear, one has for any $I \subseteq [1:p]$:

$$\operatorname{rank}(V_{:,I}) = \begin{cases} |I| & \text{if } |I| \leqslant 2 \\ 2 & \text{if } |I| > 2. \end{cases}$$

Therefore (3.38) holds. By proposition 3.4.10, this implies that equality holds in (3.39), that is, with $r := \operatorname{rank}(V) = 2 : |\partial_B H(x)| = 2\sum_{i\in[0:1]} \binom{p-1}{i} = 2p$.

[⇐] If $|\partial_B H(x)| = 2p$, (3.36b) yields $2p \leqslant \max(2p, 2^r) \leqslant 2^r + 2(p-r) \leqslant 2p$, so that equality holds in these inequalities. By the last one, $2^r = 2r$, which only occurs for $r \in \{1, 2\}$. Since $p \geqslant 2$ and the vectors are not colinear, one has $r = 2$. □

3.4.4 A glance at the C-differential

The section presents two links between the B-differential and the C-differential of the function H given by (3.3). The first proposition tells us that, whilst $\partial_C H(x)$ can be obtained

from $\partial_B H(x)$ by taking its convex hull (it is its definition (3.2)), the latter can be obtained from the former by taking its extreme points. For a proof, see [78].

Proposition 3.4.14 (a link with the C-differential)**.** $\partial_B H(x) = \operatorname{ext} \partial_C H(x)$.

The second proposition restates theorem 2.2 of Xiang and Chen [255, p. 2011], which applies to the more general nonlinear function (3.6). The interest of this restatement comes from its proof that is short, thanks to the use of the symmetry of the B-differential (proposition 3.4.1), and from the fact that proposition 3.4.15 can be used, straightforwardly, to recover Xiang and Chen's central C-Jacobian of $\tilde{H}$, given by (3.6); see [74]. Recall the notation (3.8) of the index sets.

Proposition 3.4.15 (the central C-Jacobian)**.** *One has $J \in \partial_C H(x)$ for the Jacobian whose ith row, $i \in [1:m]$, is defined by*

$$J_{i,:} = \begin{cases} A_{i,:} & \text{if } i \in \mathcal{A}(x), \\ \frac{1}{2}[A_{i,:} + B_{i,:}] & \text{if } i \in \mathcal{E}(x), \\ B_{i,:} & \text{if } i \in \mathcal{B}(x). \end{cases} \tag{3.41}$$

Proof. Let $M \in \partial_B H(x)$, which is known to be nonempty. By proposition 3.2.2, $M_{i,:} = A_{i,:}$ for $i \in \mathcal{A}(x)$, $M_{i,:} = B_{i,:}$ for $i \in \mathcal{B}(x)$ and $M_{i,:} = A_{i,:} = B_{i,:}$ for $i \in \mathcal{E}^{=}(x)$. By the symmetry of $\partial_B H(x)$ (proposition 3.4.1), M' defined by $M'_{:,i} = M_{:,i}$ if $i \in \mathcal{A}(x) \cup \mathcal{E}^{=}(x) \cup \mathcal{B}(x)$ and by

$$M'_{i,:} = \begin{cases} B_{i,:} & \text{if } i \in \mathcal{E}^{\neq}(x) \text{ and } M_{i,:} = A_{i,:} \\ A_{i,:} & \text{if } i \in \mathcal{E}^{\neq}(x) \text{ and } M_{i,:} = B_{i,:} \end{cases}$$

is also in $\partial_B H(x)$. Therefore, $J = (M + M')/2$ is in $\operatorname{co} \partial_B H(x) = \partial_C H(x)$, by (3.2). This is the formula of J given in the statement of the proposition. □

Instead of taking $J_{1/2} := \frac{1}{2}(M + M')$ in the preceeding proof, one could also have taken $J_t := (1-t)M + tM'$, which is also in $\operatorname{co} \partial_B H(x) = \partial_C H(x)$ for any $t \in [0,1]$. The inconvenient of this latter choice, when $t \neq 1/2$, is that M is usually not known. In particular, it is not necessarily known whether $M_{i,:}$ may be $A_{i,:}$ or $B_{i,:}$, for a particular $i \in \mathcal{E}^{\neq}(x)$, while J_t depends on this value when $t \neq 1/2$. In contrast, $J_{1/2}$ has an explicit formula that does not require the knowledge of the value of $M_{i,:}$ for $i \in \mathcal{E}^{\neq}(x)$.

3.5 Computation of the B-differential

This section describes techniques for computing a single Jacobian (section 3.5.1) or all the Jacobians (section 3.5.2) of the B-differential $\partial_B H(x)$, in exact arithmetic, when H is the piecewise affine function given by (3.3). The algorithms are presented as tools for computing the sign vector set $\mathcal{S} \equiv \mathcal{S}(V)$, defined by (3.12) from a matrix $V \in \mathbb{R}^{n \times p}$, which makes

them appropriate, even when $p > n$. When V is defined by (3.11), one has $p \leqslant n$ and the equivalence 3.3.5 tells us that $\mathcal{S}$ is then in bijection with $\partial_B H(x)$, so that the algorithms actually compute Jacobians of the B-differential $\partial_B H(x)$. The piece of software ISF has been written to test the algorithms [76, 75].

3.5.1 Computation of a single Jacobian

An interest of the problem equivalence highlighted in proposition 3.3.4(3) is to provide a method to find rapidly an element of $\partial_B H(x)$, which complements Qi's [204, 1993, final remarks (1)]. It is shown in [74], that this method extends to the computation of an element of the B-differential in the nonlinear case, i.e., when H is the function $\tilde{H}$ given by (3.6). The method is based on the following algorithm, which associates with p nonzero vectors v_1, …, v_p, which may be identical or colinear, a direction d such that $v_i^\mathsf{T} d \neq 0$ for all $i \in [1:p]$; it is a variant of the technique used in the proof of [255, lemma 2.1]. When the v_i's are also distinct, the direction d can also be derived from lemma 3.2.6, by adding the vector $v_0 = 0$.

Algorithm 3.5.1 (computes $\boldsymbol{d \in \mathbb{R}^n}$ such that $\boldsymbol{v_i^\mathsf{T} d \neq 0}$ for all $\boldsymbol{i}$).

Let be given p nonzero vectors v_1, …, v_p in $\mathbb{R}^n$ and take $d \in \mathbb{R}^n \setminus \{0\}$.

Repeat :

1. If $I := \{i \in [1:p] : v_i^\mathsf{T} d = 0\} = \varnothing$, exit.
2. Let $i \in I$.
3. Take $t > 0$ sufficiently small such that, for all $j \notin I$, $(v_j^\mathsf{T} d)(v_j^\mathsf{T}[d + t v_i]) > 0$.
4. Update $d := d + t v_i$.

Explanation. In step 3, any sufficiently small $t > 0$ is appropriate (the proof of [255, lemma 2.1] computes bounds explicitely), since $(v_j^\mathsf{T} d)(v_j^\mathsf{T}[d + t v_i])$ is positive for $t = 0$. The new direction d set in step 4 is such that $v_i^\mathsf{T}(d + t v_i) = t\|v_i\|^2 > 0$, so that this direction makes at least one more $v_j^\mathsf{T} d$ nonzero than the previous one. This implies that the algorithm finds an appropriate direction in at most p loops. □

The next procedure uses a direction d computed by algorithm 3.5.1 to obtain a single element of $\partial_B H(x)$. Recall that the map σ is defined by (3.14a) and is a bijection from $\partial_B H(x)$ onto $\mathcal{S}$, defined by (3.12) (proposition 3.3.4).

Algorithm 3.5.2 (computes a single Jacobian in $\boldsymbol{\partial_B H(x)}$).

Let H be given by (3.3), $x \in \mathbb{R}^n$ and suppose that $p \neq 0$.

1. Compute $V \in \mathbb{R}^{n\times p}$ by (3.11) and denote its columns by $v_1, \ldots, v_p \in \mathbb{R}^n$.
2. By algorithm 3.5.1, compute $d \in \mathbb{R}^n$ such that $v_i^\mathsf{T} d \neq 0$ for all $i \in [1:p]$.
3. Define $s \in \mathcal{S}$ by $s_i := \operatorname{sgn}(v_i^\mathsf{T} d)$, for $i \in [1:p]$.
4. Then, $\sigma^{-1}(s) \in \partial_B H(x)$.

Explanation. When $p = 0$, $\partial_B H(x) = \partial_B^\times H(x)$ contains a single Jacobian that is given by (3.10), which explains why algorithm 3.5.2 focuses on the case when $p > 0$. The sign vector s computed in step 3 is such that $s_i v_i^\mathsf{T} d > 0$ for all $i \in [1:p]$, so that it is indeed in $\mathcal{S}$ and, by proposition 3.3.4, $\sigma^{-1}(s)$ is a Jacobian in $\partial_B H(x)$. □

3.5.2 Computation of all the Jacobians

This section presents two basic algorithms, and some more efficient variants, for computing all the B-differential of H. They use the notion of $\mathcal{S}$-tree presented in section 3.5.2(A). The first algorithm is grounded on the notion of stem vector (section 3.3.2) and is described in section 3.5.2. The second algorithm is the outcome of a series of improvements brought to an algorithm by Rada and Černý [208, p. 2018] (section 3.5.2(B)) for computing the cells of a hyperplane arrangement, which is known to be an equivalent problem to the one of computing the B-differential of H when the hyperplanes contain zero (see section 3.3.4). The improvements are detailed in section 3.5.2 and the resulting algorithm is described in section 3.5.2. Finally, numerical experiments are presented in section 3.5.2 to compare the efficiency of the algorithms.

Algorithms for listing the elements of the finite set $\partial_B H(x)$ can be designed by looking at one of the various forms of the problem, those described in section 3.3 and others [14]; this is what we shall do. Most algorithms we have found in the scientific literature take the point of view of hyperplane arrangements of section 3.3.4 and can be used for more general arrangements than those needed to describe $\partial_B H(x)$ (i.e., in which case the hyperplanes pass through zero). One can quote the contributions by Bieri and Nef [27, p. 1982], Edelsbrunner, O'Rourke and Seidel [83, p. 1986], Avis and Fukuda [14, p. 1996], improved by Sleumer [232, p. 1998], and, more recently, Rada and Černý [208, p. 2018], which is described in section 3.5.2(B). See also [79].

Incremental-recursive algorithms

The algorithms described in this section are incremental in the sense that the considered sign vectors have their length increased by one at each step. Furthermore, the algorithms explore the $\mathcal{S}$-tree described in subsection A below by recursive procedures, whose names are recognizable by their suffix "-REC". All the procedures end by returning to their calling program.

A. THE $\mathcal{S}$-TREE. A common feature of the algorithms considered in this paper is the construction of the $\mathcal{S}$-tree described below, incrementally and recursively. This idea was probably introduced by Rada and Černý [208, p. 2018]. See figure 3.4 for an illustration.

The level k of the $\mathcal{S}$-tree is formed of a set of sign vectors denoted by

$$\mathcal{S}_k^1 := \{s \in \mathcal{S}_k : s_1 = +1\}, \tag{3.42}$$

where $\mathcal{S}_k$ is the subset of $\{\pm 1\}^k$ defined by (3.33). In particular, the level 1 or root of the $\mathcal{S}$-tree contains the unique sign vector $+1 \in \{\pm 1\}^1$. The $\mathcal{S}$-tree has p levels, where p is the number of vectors v_i, or columns of the given matrix $V \in \mathbb{R}^{n\times p}$. Note that there is no reason to compute $\{s \in \mathcal{S} : s_1 = -1\}$ since this part of $\mathcal{S}$ is equal to $-\{s \in \mathcal{S} : s_1 = 1\}$ by the symmetry property of $\mathcal{S}$ (proposition 3.4.1). In order to avoid the memorization of the elements of $\mathcal{S}_k^1$, the $\mathcal{S}$-tree is constructed by a *depth-first search*, which can be schematized as follows.

Algorithm 3.5.3 (STREE (V))**.** Let be given $V \in \mathbb{R}^{n\times p}$, with n and $p \in \mathbb{N}^*$, having nonzero columns.

1. Execute the recursive procedure STREE-REC$(V, +1)$.

Algorithm 3.5.4 (STREE-REC (V, s))**.** Let be given $V \in \mathbb{R}^{n\times p}$, with n and $p \in \mathbb{N}^*$, having nonzero columns, and a sign vector $s \in \mathcal{S}_k^1$ for some $k \in [1 : p]$.

1. If $k = p$, print s and return.
2. If $(s, +1) \in \mathcal{S}_{k+1}^1$, execute STREE-REC$(V, (s, +1))$.
3. If $(s, -1) \in \mathcal{S}_{k+1}^1$, execute STREE-REC$(V, (s, -1))$.

The method used to determine whether $(s, \pm 1)$ is in $\mathcal{S}_{k+1}^1$ depends on the specific algorithm and may or may not use a direction d intervening in (3.33). Note that, as emphasized in proposition 3.4.6(3), at least one of the sign vectors $(s, +1)$ and $(s, -1)$ belongs to $\mathcal{S}_{k+1}^1$ (maybe both). It is justified not to explore the $\mathcal{S}$-tree below an $(s, \pm 1)$ that is not in $\mathcal{S}_{k+1}^1$, since then $(s, \pm 1, s') \notin \mathcal{S}$ for any $s' \in \{\pm 1\}^{p-k-1}$. By construction, the algorithm STREE prints all the elements of $\mathcal{S}_p^1 \equiv \mathcal{S}^1 := \{s \in \mathcal{S} : s_1 = +1\}$ in step 1 of the STREE-REC procedure.

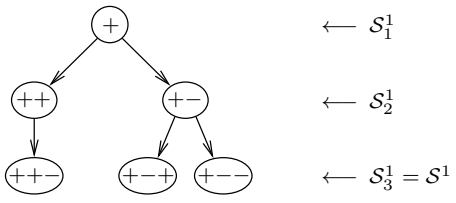


FIGURE 3.4 – Half of the $\mathcal{S}$-tree for example 3.3.2 (the other half is obtained by swapping the $+$'s and the $-$'s). Top-down arrows indicate descendence ; the sign sets $\mathcal{S}_k^1$ are defined by (3.42).

B. Rada and Černý's algorithm. The algorithm proposed by Rada and Černý [208, p. 2018], which is referenced below as the RC algorithm, deals with the determination of the cells associated with a general hyperplane arrangement. We describe it below for an arrangement of hyperplanes containing all zero (see section 3.3.4), which is the case when V results from (3.11) in the computation of the B-differential $\partial_B H(x)$. We also use the linear algebra language of section 3.3.2, viewing the problem as the one of determining the set $\mathcal{S}$ defined by (3.11)); in contrast, the language used in [208] is more geometric. The algorithm builds the $\mathcal{S}$-tree of the previous section A and, for each $s \in \mathcal{S}^1_k$, it solves a single problem (LOP) to determine whether $(s,+1)$ or $(s,-1)$ is in $\mathcal{S}^1_{k+1}$.

The RC algorithm succeeds in solving only one LOP to determine whether $(s,+1)$ and $(s,-1)$ are in $\mathcal{S}^1_{k+1}$, at the node $s \in \mathcal{S}^1_k$, thanks to the memorization of a direction d such that $s \bullet (V_k^\mathsf{T} d) > 0$ (we note $V_k := V_{:,[1:k]}$). Indeed, one has

$$\begin{array}{lll} v_{k+1}^\mathsf{T} d < 0 & \Longrightarrow & (s,-1) \in \mathcal{S}^1_{k+1}, \\ v_{k+1}^\mathsf{T} d > 0 & \Longrightarrow & (s,+1) \in \mathcal{S}^1_{k+1}, \end{array}$$

and one of these two cases takes place if we exclude the case where $v_{k+1}^\mathsf{T} d = 0$. In [208, Algorithm 1], the case where $v_{k+1}^\mathsf{T} d = 0$ is not dealt with completely since $(s,+1)$ is declared to belong to $\mathcal{S}^1_{k+1}$ in that case, while it is clear that $(s,-1)$ is also in $\mathcal{S}^1_{k+1}$. Indeed, in our implementation of the RC algorithm, we modify slightly d by adding a small positive or negative multiple of v_{k+1} to d when $v_{k+1}^\mathsf{T} d \simeq 0$, so that both $(s,\pm 1)$ are accepted in $\mathcal{S}^1_{k+1}$ in that case. This choice may be at the origin of the differences that one observes in table 3.1 below between the statistics of the original RC algorithm in [208] and those of our implementation.

Next, when $(s, s_{k+1}) \in \{\pm 1\}^{k+1}$ is observed to belong to $\mathcal{S}^1_{k+1}$, the question of whether $(s, -s_{k+1})$ also belongs to $\mathcal{S}^1_{k+1}$ arises. In the RC algorithm, the answer to this question is obtained by solving a LOP similar to

$$\left\{ \begin{array}{l} \min_{(d,t)\in\mathbb{R}^n\times\mathbb{R}} \ t \\ s_i v_i^\mathsf{T} d \geqslant 1, \quad \forall\, i \in [1:k] \\ -s_{k+1} v_{k+1}^\mathsf{T} d \geqslant -t \\ t \geqslant -1. \end{array} \right. \tag{3.43}$$

When $s \in \mathcal{S}^1_k$, this problem is feasible (take d satisfying $s_i v_i^\mathsf{T} d \geqslant 1$, for all $i \in [1:k]$, and t sufficiently large) and bounded (its optimal value is $\geqslant -1$), so that it has a solution [50, 31, 29, 106]. Solving these LOPs is a time consuming part of the algorithms and in the numerical experiments of section 3.5.2, in particular in table 3.2, following [208], we measure the efficiency of the algorithms by the number of LOPs they solve.

One can now formally describe our version of the RC algorithm (the change is in step 2 of the RC-REC algorithm, which is not considered in the original RC algorithm).

Algorithm 3.5.5 (RC (V))**.** Let be given $V \in \mathbb{R}^{n\times p}$, with n and $p \in \mathbb{N}^*$, having nonzero columns.

1. Execute the recursive procedure RC-REC$(V, v_1, +1)$.

Algorithm 3.5.6 (RC-REC (V, d, s))**.** Let be given $V \in \mathbb{R}^{n\times p}$, with n and $p \in \mathbb{N}^*$, having nonzero columns, a direction $d \in \mathbb{R}^n$ and a sign vector $s \in \{\pm 1\}^k$ for some $k \in [1 : p]$, such that $s_i v_i^\mathsf{T} d > 0$ for all $i \in [1 : k]$.

1. If $k = p$, print s and return.
2. If $v_{k+1}^\mathsf{T} d \simeq 0$, then
   2.1. Execute RC-REC$(V, d_+, (s, +1))$, where $d_+ := d + t_+ v_{k+1}$ with $t_+ > 0$ chosen in the nonempty open interval
   $$\left(0, \min_{\substack{i\in[1:k]\\ s_i v_i^\mathsf{T} v_{k+1}<0}} \frac{-v_i^\mathsf{T} d}{v_i^\mathsf{T} v_{k+1}}\right).$$
   2.2. Execute RC-REC$(V, d_-, (s, -1))$, where $d_- := d + t_- v_{k+1}$ with $t_- < 0$ chosen in the nonempty open interval
   $$\left(\max_{\substack{i\in[1:k]\\ s_i v_i^\mathsf{T} v_{k+1}>0}} \frac{-v_i^\mathsf{T} d}{v_i^\mathsf{T} v_{k+1}},\ 0\right).$$
3. Else $s_{k+1} := \operatorname{sgn}(v_{k+1}^\mathsf{T} d)$.
   3.1. Execute RC-REC$(V, d, (s, s_{k+1}))$.
   3.2. Solve the LOP (3.43) and denote by (d, t) a solution.
   If $t = -1$, execute RC-REC$(V, d, (s, -s_{k+1}))$.

In steps 2.1 and 2.2, the minimum and maximum are supposed to be infinite if their feasible set is empty. One can check that the directions $d_\pm$ computed in steps 2.1 and 2.2 are such that $s_i v_i^\mathsf{T} d_\pm > 0$ for $i \in [1 : k + 1]$ and $s_{k+1} = \pm 1$, provided $|v_{k+1}^\mathsf{T} d|$ is sufficiently small, which justifies the recursive call to RC-REC with the given arguments. The test $v_{k+1}^\mathsf{T} d \simeq 0$ done at the beginning of step 2 is supposed to take into account floating point arithmetic ; admittedly it is not very rigorous, but the algorithm is designed to be as close as possible to the original RC algorithm in [208] ; a more careful treatment of this situation is presented in section 3.5.2(B). The most time-consuming part of the RC algorithm comes from the possible numerous LOPs to solve in step 3.2 of RC-REC.

An algorithm using stem vectors

When $s \in \mathcal{S}_k$, it is conceptually easy to check whether $(s, \pm 1)$ is in $\mathcal{S}_{k+1}$, provided a list of all the stem vectors associated with V is known. Indeed, by proposition 3.3.10, if no subvector of $(s, +1)$ (resp. $(s, -1)$) is a stem vector, then $(s, +1)$ (resp. $(s, -1)$) belongs

to $\mathcal{S}_{k+1}$. Note also that, because any $s \in \mathcal{S}_k$ has at least one descendant in the $\mathcal{S}$-tree (proposition 3.4.6(3)), if it is observed that $(s,+1) \notin \mathcal{S}_{k+1}$, then, necessarily, $(s,-1) \in \mathcal{S}_{k+1}$. This observation prevents the algorithm from checking whether $(s,-1)$ contains a stem vector, which is a time consuming operation when the list of stem vectors is large. For future reference, we formalize this algorithm below.

Algorithm 3.5.7 (STEM (V))**.** Let be given $V \in \mathbb{R}^{n\times p}$, with n and $p \in \mathbb{N}^*$, having nonzero columns.

1. Compute all the stem vectors associated with V.
2. Execute the recursive procedure STEM-REC$(V,+1)$.

Algorithm 3.5.8 (STEM-REC (V, s))**.** Let be given $V \in \mathbb{R}^{n\times p}$, with n and $p \in \mathbb{N}^*$, having nonzero columns and a sign vector $s \in \{\pm 1\}^k$ for some $k \in [1:p]$.

1. If $k = p$, print s and return.
2. If no subvector of $(s,+1)$ is a stem vector, execute STEM-REC$(V,(s,+1))$.
3. If $(s,+1) \notin \mathcal{S}_{k+1}$ or no subvector of $(s,-1)$ is a stem vector, execute STEM-REC$(V,(s,-1))$.

This algorithm is improved below, as the option AD_4 of the ISF algorithm (see paragraphs A and D of section 3.5.2).

Note that, this algorithm need not generate directions d satisfying $s\bullet(V_k^{\mathsf{T}}d) > 0$, like the RC algorithm and need not solve any linear optimization problem. Nevertheless, regarding the computation time, the algorithm has two bottlenecks that we now describe.

The first bottleneck comes from the fact that the algorithm must compute all the stem vectors (or the set $\mathcal{C}$ of matroid circuits in (3.19)) associated with V. This is usually an expensive operation [141, 166, 212]. For example, if V is randomly generated and of rank r, like in the test-cases DATA_RAND_* in the experiments of section 3.5.2, any selection of r columns of V is likely to form an independent set of vectors, so that $\mathcal{C}$ is likely to be the sets of column indices of size $r+1$. In this case, the number of circuits is likely to be the combination $\binom{p}{r+1}$ (and it is actually that number, see section 3.5.2(B.1)), which can be exponential in p (this number is bounded below by $2^{p/2}/(p+1)$ if p is even and $r+1 = p/2$ [59, (11.52)]). In the implemented ISF code, numerically tested in section 3.5.2, only the sets of columns whose cardinality is in $[3 : r+1]$ are examined (since any group of two columns of V is supposed to be linearly independent and a group of $r+2$ columns or more is of nullity $\geqslant 2$, hence such group cannot form a matroid circuit ; see (3.19)).

The second bottleneck is linked to the detection of a stem vector in the current sign vectors $(s,\pm 1)$. This operation requires to examine the long list of stem vectors, which is a time consuming operation.

We shall see in the numerical experiments of section 3.5.2 that algorithm 3.5.7 is generally the fastest, provided the number of stem vectors is not too large.

Linear optimization problem and stem vector

The property described in this section will be useful for the improvement D_2 of the ISF algorithm, described in section 3.5.2(D). It shows that a stem vector can be obtained easily from the dual solution of the linear optimization (LOP) (3.43), when $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}$. Consider indeed the LOP (3.43) and denote by (d, t) one of its solutions (these have been shown to exist). Then, either $t \geqslant 0$ (equivalently, $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}$) or $t = -1$ (equivalently, $(s, -s_{k+1}) \in \mathcal{S}_{k+1}$).

Let σ_i, $i \in [1:k+1]$, be the multipliers associated with the first $k+1$ constraints of (3.43) and τ be the multiplier associated with its last constraint. Then, the Lagrangian dual of (3.43) reads [31, 26, 29, 105]

$$\left\{\begin{array}{l} \max_{(\sigma,\tau)\in\mathbb{R}^{k+1}\times\mathbb{R}} \ \sum_{i\in[1:k]} \sigma_i - \tau \\ \sigma \geqslant 0 \\ \tau \geqslant 0 \\ \sigma_{k+1} + \tau = 1 \\ \sigma_{k+1} s_{k+1} v_{k+1} = \sum_{i\in[1:k]} \sigma_i s_i v_i. \end{array}\right. \equiv \left\{\begin{array}{l} \max_{\sigma\in\mathbb{R}^{k+1}} \ \sum_{i\in[1:k+1]} \sigma_i - 1 \\ \sigma \geqslant 0 \\ \sigma_{k+1} \leqslant 1 \\ \sigma_{k+1} s_{k+1} v_{k+1} = \sum_{i\in[1:k]} \sigma_i s_i v_i, \end{array}\right. \tag{3.44}$$

where the second form of the dual is obtained by eliminating τ from the first form. By strong duality in linear optimization, the dual problems in (3.44) are feasible, have a solution and have the same optimal value as the primal problem. Let $(\sigma, \tau) \in \mathbb{R}^{k+1} \times \mathbb{R}$ be a dual solution. Then, $(s, -s_{k+1}) \in \mathcal{S}_{k+1}$ if and only if $t = -1$ if and only if $\sum_{i\in[1:k]} \sigma_i = 0$ and $\sigma_{k+1} = 0$. We have shown that

$$(s, -s_{k+1}) \in \mathcal{S}_{k+1} \quad \Longleftrightarrow \quad \sigma = 0.$$

Therefore, $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}$ if and only if $\sigma \neq 0$ if and only if $\sigma_{k+1} = 1$ (if $\sigma_{k+1} = 0$, one can make the dual objective value as large as desired by multiplying σ by a factor going to $+\infty$; if $\sigma_{k+1} \in (0, 1)$, the dual objective would be increased by replacing σ by σ/σ_{k+1}; in both cases the optimality of σ would be contradicted) if and only if $\tau = 0$. We have shown that

$$(s, -s_{k+1}) \notin \mathcal{S}_{k+1} \quad \Longleftrightarrow \quad s_{k+1} v_{k+1} \in \operatorname{cone}\{s_i v_i : i \in [1:k]\}.$$

The next proposition shows how a matroid circuit can be detected from the dual solution σ when $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}$.

Proposition 3.5.9. (matroid circuit detection) *Suppose that $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}$ and that (σ, τ) is a solution to the dual problem in the left-hand side of* (3.44) *located at an extreme point of its feasible set. Then, $\{i \in [1:k+1] : \sigma_i > 0\}$ is a matroid circuit of V.*

Proof. We have seen that $\sigma_{k+1} = 1$ and $\tau = 0$ when $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}$. The fact that $(\sigma, 0)$ is an extreme point of the feasible set of the problem in the left-hand side of (3.44) implies that the vectors [50, 105]

$$\left\{ \begin{pmatrix} 0 \\ s_i v_i \end{pmatrix}_{i\in[1:k],\ \sigma_i>0}, \begin{pmatrix} 1 \\ -s_{k+1} v_{k+1} \end{pmatrix} \right\} \text{ are linearly independent.}$$

In particular, the vectors

$$\{s_i v_i : i \in [1:k],\ \sigma_i > 0\} \text{ are linearly independent.}$$

Since $s_{k+1}v_{k+1} = \sum_{i\in[1:k]} \sigma_i s_i v_i$, it follows that

$$\{s_i v_i : i \in [1:k+1],\ \sigma_i > 0\} \text{ has nullity one.}$$

The conclusion of the proposition follows from proposition 3.3.11. □

Recall that the dual-simplex algorithm finds a dual solution at an extreme point of the dual feasible set. For this reason, we use this approach in the ISF algorithm with option D_2 (see section 3.5.2(D)).

Improvements of the RC and STEM algorithms

This section presents several modifications of the RC algorithm and one modification of the STEM algorithm that significantly improve their performance. The modifications are indicated by the letters A, B, C and D, with reference to the sections where they are introduced. Additional numeric indices specify variants of the D option. The version AD_4 (modifications A and D_4) can be considered as an improvement of the new algorithm 3.5.7.

A. TAKING THE RANK OF V INTO ACCOUNT. Instead of starting with the vector $s = +1$, one can take into account the rank $r := \mathrm{rank}(V)$ to determine 2^r initial vectors s, hence avoiding to solve linear optimization problems (LOPs) to determine these initial s's. This is especially useful when $p - r$ is small. In particular, when $p = r$, $\mathcal{S}$ is straightforwardly determined.

The algorithm selects $r := \mathrm{rank}(V)$ linearly independent vectors v_i, among the columns of $V \in \mathbb{R}^{n\times p}$. These vectors can be obtained by a QR factorization of

$$VP = QR,$$

where $P \in \{0,1\}^{p\times p}$ is a permutation matrix, $Q \in \mathbb{R}^{n\times n}$ is orthogonal (i.e., $Q^{\mathsf{T}}Q = I_n$) and $R \in \mathbb{R}^{n\times p}$ is upper triangular with $R_{[r+1:n],:} = 0$. To simplify the presentation, one can assume, without loss of generality, that $P = I$, in which case the vectors v_1, …, v_r are linearly independent (in practice, the vectors are symbolically reordered by using the permutation matrix P). By proposition 3.4.2 and with the notation (3.33) :

$$\mathcal{S}_r = \{\pm 1\}^r. \tag{3.45}$$

Furthermore, for each $s \in \mathcal{S}_r$, we have, using $S := \mathrm{Diag}(s)$, $Q_r := Q_{:,[1:r]}$ and $R_r := R_{[1:r],[1:r]}$, that the vector

$$d_s = Q_r R_r^{-\mathsf{T}} s \tag{3.46}$$

is such that $s \bullet (V_{:,[1:r]}^{\mathsf{T}} d_s) = e > 0$, as desired.

For each $s \in \mathcal{S}_r$ and the associated d_s given by (3.46), the modified algorithm 3.5.5 runs the recursive function RC-REC(V, d_s, s) (see algorithm 3.5.11 below).

B. SPECIAL HANDLING OF THE CASE WHERE $v_{k+1}^{\mathsf{T}} d \simeq 0$. Directions $d_\pm := d + t_\pm v_{k+1}$ ensuring that $(s, \pm 1) \bullet (V_{k+1}^{\mathsf{T}} d_\pm) > 0$ can be computed not only when $v_{k+1}^{\mathsf{T}} d \simeq 0$ like in step 2 of the RC-REC algorithm 3.5.6, but also when $v_{k+1}^{\mathsf{T}} d$ is in the interval specified by (3.47) below. Note that the left-hand side in (3.47) is negative and the right-hand side is positive (this can be seen by multiplying numerators and denominators by s_i and by using $s_i v_i^{\mathsf{T}} d > 0$ for all $i \in [1:k]$), so that these inequalities are verified when $v_{k+1}^{\mathsf{T}} d = 0$. With the additional flexibility that (3.47) offers, the ISF algorithm can sometimes avoid solving a significant number of LOPs of the form (3.43). For a proof of the next proposition, see [78].

Proposition 3.5.10 (two descendants without optimization)**.** *Suppose that $s \in \{\pm 1\}^k$ verifies $s \bullet (V_k^{\mathsf{T}} d) > 0$, that $v_{k+1} \neq 0$ and that*

$$\max_{\substack{i \in [1:k] \\ s_i v_i^{\mathsf{T}} v_{k+1} > 0}} \frac{-v_i^{\mathsf{T}} d}{v_i^{\mathsf{T}} v_{k+1}} < \frac{-v_{k+1}^{\mathsf{T}} d}{\|v_{k+1}\|^2} < \min_{\substack{i \in [1:k] \\ s_i v_i^{\mathsf{T}} v_{k+1} < 0}} \frac{-v_i^{\mathsf{T}} d}{v_i^{\mathsf{T}} v_{k+1}}. \tag{3.47}$$

1) *The direction $d_+ := d + t_+ v_{k+1}$ verifies $s \bullet (V_k^{\mathsf{T}} d_+) > 0$ and $v_{k+1}^{\mathsf{T}} d_+ > 0$ if and only if t_+ is in the nonempty open interval*

$$\left(\frac{-v_{k+1}^{\mathsf{T}} d}{\|v_{k+1}\|^2}, \min_{\substack{i \in [1:k] \\ s_i v_i^{\mathsf{T}} v_{k+1} < 0}} \frac{-v_i^{\mathsf{T}} d}{v_i^{\mathsf{T}} v_{k+1}} \right). \tag{3.48a}$$

2) *The direction $d_- := d + t_- v_{k+1}$ verifies $s \bullet (V_k^{\mathsf{T}} d_-) > 0$ and $-v_{k+1}^{\mathsf{T}} d_- > 0$ if and only if t_- is in the nonempty open interval*

$$\left(\max_{\substack{i \in [1:k] \\ s_i v_i^{\mathsf{T}} v_{k+1} > 0}} \frac{-v_i^{\mathsf{T}} d}{v_i^{\mathsf{T}} v_{k+1}}, \frac{-v_{k+1}^{\mathsf{T}} d}{\|v_{k+1}\|^2} \right). \tag{3.48b}$$

C. CHANGING THE ORDER OF THE VECTORS v_i'S. Each node s of the $\mathcal{S}$-tree described in section 3.5.2(A) has one or two descendants : $(s, +1)$ and/or $(s, -1)$. Since there is at most one LOP solved per node of the $\mathcal{S}$-tree, decreasing the number of nodes should decrease the number of LOPs to solve, which significantly count in the computing time. To reach that goal, one can try to get as much as possible at the top of the tree the nodes having a single descendant. As shown below, this can be achieved by changing the order in which the vectors v_i's, the columns of V, are considered in the *depth-first search* of the tree ; previously, the order was imposed by the modification A, taking into account the rank of V. As we shall

see, a new order is not fixed once and for all, but is determined during the construction of the $\mathcal{S}$-tree, is reconsidered at each node and depends on the path going from the root of the $\mathcal{S}$-tree to its leaves.

To implement this strategy, one associates with each node $s \in \mathcal{S}_k^1$ of the $\mathcal{S}$-tree, $k \in [1 : p-1]$, the list of vectors considered so far at that node, denoted by $T_s := \{i_1, \ldots, i_k\} \subseteq [1:p]$. Hence, we have to choose the next vector $v_{i_{k+1}}$ be selecting an index i_{k+1} in $T_s^c := [1:p] \setminus T_s$. Now, a natural idea is to restrict the set of possible indices to T_s^b, the set of indices j of T_s^c for which one of the intervals (3.48a) or (3.48b), with $v_{k+1} \equiv v_j$, is empty (implying that the technique used in the modification B will not give two descendants), if there is such an index, or T_s^c otherwise. To determine the index in T_s^b, we take

$$i_{k+1} = \text{argmax}_{i \in T_s^b} \frac{|v_i^\mathsf{T} d|}{\|v_i\|}, \tag{3.49}$$

which favors the vectors v_i for which $|v_i^\mathsf{T} d|/\|v_i\|$ is away from zero.

As table 3.2 indicates (section 3.5.2(C.3)), this modification has a significant impact on the decrease of the number of LOPs to solve.

D. Using stem vectors. We present in this section various modifications that use the concept of *stem vector*, introduced in the second part of section 3.3.2. These stem vectors are used to detect infeasible sign vectors, i.e., elements of $\mathcal{S}^c$, thanks to proposition 3.3.10. If $s \in \mathcal{S}_k^1$ and $(s, s_{k+1}) \in \mathcal{S}^c$ for $s_{k+1} \in \{\pm 1\}$, s has no descendant in $\mathcal{S}$ along (s, s_{k+1}), so that this part of the $\mathcal{S}$-tree does not need to be explored. From this point of view, computing all the stem vectors looks attractive, but, to our knowledge, this is a time consuming process, so that this option is not necessarily the most efficient one. The modifications presented below use more and more stem vectors, whose computation requires more and more time.

- D$_1$) Natural candidates as stem vectors are those obtained from the matroid circuits I made of $r+1$ columns of V ($r = \text{rank}(V)$) formed of the r linear independent columns selected by the QR factorization of section 3.5.2(A) and one of the remaining $p-r$ columns of V. By proposition 3.3.11, such I contains exactly one circuit. Therefore, one detects in this way $p-r$ circuits and $2(p-r)$ stem vectors. This is not much compared to the total number of stem vectors, which may depend exponentially on p, so that the number of infeasible sign vectors detected by these stem vectors is usually relatively small (see table 3.2).
- D$_2$) With this option, when a LOP (3.43) is solved at a certain node $s \in \mathcal{S}_k^1$ to see whether (s, s_{k+1}) belongs to $\mathcal{S}_{k+1}^1$, for $s_{k+1} \in \{\pm 1\}$, the dual solution is used to determine a matroid circuit, as shown by proposition 3.5.9. For this purpose, the ISF code solves the LOP with the dual-simplex algorithm, so that the computed dual solution is at a vertex of the dual feasible set.
- D$_3$) With this option, all the stem vectors are computed, before running the recursive process that builds the $\mathcal{S}$-tree. At each note $s \in \mathcal{S}_k^1$, the algorithm still computes

a direction $d \in \mathbb{R}^n$ such that $s_i v_i^\mathsf{T} d > 0$ for all $i \in T_s$ (the set of vector indices considered so far at s). The advantage of this direction is to allow the algorithm to use the beneficial modifications B and C and to easily determine one or two signs $s_{k+1} \in \{\pm 1\}$ such that $(s, s_{k+1}) \in \mathcal{S}^1_{k+1}$. If a single sign $s_{k+1} \in \{\pm 1\}$ is selected, the stem vectors can decide whether $(s, -s_{k+1}) \in \mathcal{S}^1_{k+1}$. If this is the case, this option D_3 has the inconvenient of still requiring to solve a LOP to get a direction associated with $(s, -s_{k+1})$. These LOPs (3.43) have an optimal value -1 and should not be solved exactly. Indeed, as soon as a feasible direction d for (3.43) gives a negative value to the objective of the problem, one could stop solving it, since this d verifies $s_i v_i^\mathsf{T} d > 0$ for all $i \in T_{(s,-s_{k+1})}$. We have not implemented that inexact solve of the LOPs, by lack of flexibility of the solver LINPROG in Matlab.

D_4) Like with the option D_3, all the stem vectors are computed, before running the recursive process that builds the $\mathcal{S}$-tree. But now, unlike with option D_3, the algorithm computes no direction $d \in \mathbb{R}^n$. When option A is also activated, the resulting approach can be viewed as an improvement of the algorithm 3.5.7 (STEM) presented in section 3.5.2.

Note that, knowing all the stem vectors, one could compute the complementary set $\mathcal{S}^c$ rather easily by completing with ± 1 the unspecified components of the stem vectors. Next, $\mathcal{S}$ could be obtained from $\mathcal{S}^c$ by taking its complementary set in $\{\pm 1\}^p$, but a straightforward implementation of this last operation looks rather expensive, so that we have not experimented it numerically.

ISF algorithm

We have named ISF (for Incremental Signed Feasibility) the algorithm that improves the RC algorithm 3.5.5 or the STEM algorithm 3.5.7 with the enhancements described in section 3.5.2. For the purpose of precision and reference, we formally state it in this section. It would be cumbersome and confusing, hence inappropriate, to mention all the options in its description, in particular because all of them have been specified separately in the previous section. As an example of algorithm, we provide a description with the options $ABCD_2$. It starts with a hat procedure ISF, similar to that of the RC algorithm but with the additional easy determination of $\mathcal{S}_r$ (modification A) and the computation of some stem vectors (modification D_1). Then, the hat procedure calls the recursive procedure ISF-REC.

Algorithm 3.5.11 (ISF (V), with options $ABCD_2$)**.** Let be given $V \in \mathbb{R}^{n\times p}$, with n and $p \in \mathbb{N}^*$, having nonzero columns.

1. Compute the QR factorization of V. Let $r = \mathrm{rank}(V)$ and $T_r := \{i_1, \ldots, i_r\}$ be the indices of r selected linear independent columns of V.
2. Compute the $p - r$ matroid circuits (see option D_1).
3. For each $s \in \mathcal{S}_r$, given by (3.45), and its associated d_s, given by (3.46), call the recursive

procedure ISF-REC(V, T_r, d_s, s).

Algorithm 3.5.12 (ISF-REC (V, T, d, s), with options BCD$_2$)**.** Let be given $V \in \mathbb{R}^{n\times p}$, with n and $p \in \mathbb{N}^*$ of rank r, having nonzero columns v_i, T a selection of k columns of V (with $k \in [r:p]$), a direction $d \in \mathbb{R}^n$ and a sign vector $s \in \{\pm 1\}^k$ for some $k \in [r:p]$. It is assumed that $s_i v_i^\mathsf{T} d > 0$ for all $i \in T$.

1. If $k = p$, print s and return.
2. Determine the index $i_{k+1} \in [1:p] \setminus T$ of the next vector to consider by option C and set $T_+ := T \cup \{i_{k+1}\}$.
3. If (3.47) holds (with $[1:k]$ changed into T and $k+1$ into i_{k+1}), then
   3.1. Execute ISF-REC$(V, T_+, d_+, (s, +1))$, where $d_+ := d + t_+ v_{i_{k+1}}$ and t_+ is chosen in the nonempty open interval (3.48a).
   3.2. Execute ISF-REC$(V, T_+, d_-, (s, -1))$, where $d_- := d + t_- v_{i_{k+1}}$ and t_- is chosen in the nonempty open interval (3.48b).
4. Else $s_{k+1} := \operatorname{sgn}(v_{i_{k+1}}^\mathsf{T} d)$.
   4.1. Execute ISF-REC$(V, T_+, d, (s, s_{k+1}))$.
   4.2. If $(s, -s_{k+1})$ contains a stem vector, return.
   4.3. Solve the LOP (3.43) (with $[1:k]$ changed into T and $k+1$ into i_{k+1}) by the dual-simplex algorithm and denote by (d, t) a solution.
      4.3.1. If $t = -1$, execute ISF-REC$(V, T_+, d, (s, -s_{k+1}))$.
      4.3.2. Else, use the dual solution to store two more stem vectors by option D$_2$.

Numerical experiments

We present in tables 3.1, 3.2 and 3.3 the results obtained by running the algorithms 3.5.7 and 3.5.11 (with several variants) on a small number of problems and compare it with our implementation of the RC algorithm 3.5.5, simulating algorithm 1 (IE) in [208].

A. COMPUTER AND PROBLEM PRESENTATION. The implementations have been done in Matlab (version "9.11.0.1837725 (R2021B) UPDATE 2") on a MACBOOKPRO18,2/10CORES (parallelism is not implemented however) with the system MACOS MONTEREY, version 12.6.1. The linear optimization problem solver is LINPROG.

Computation in ISF is done in floating point numbers, so that numerical roundoff errors may occur. To deal with this difficulty, the code uses various tolerances, for instance, to detect almost identical normalized vectors (columns of V), to identify nonzero components of circuits, *etc.* The Julia code described in [79], which deals with more general hyperplane arrangements, offers the user the possibility of requiring a computation in rational numbers, so as to have a computation in exact arithmetic.

We have assessed the codes on randomly generated problems (function RAND in Matlab, names prefixed by RAND and SRAND) and problems adapted/taken from [208] (names prefixed by RC) and [35] (names prefixed by BEK). Their relevant features are given in table 3.1 and their specifications are now given.

- The RAND-N-P-R problems have their data formed of a randomly generated matrix $V \in \mathbb{R}^{N\times P}$ with prescribed rank R.
- For the problems SRAND-N-P-Q, the first N columns of $V \in \mathbb{R}^{N\times P}$ form the identity matrix and the last $P-N > 0$ columns have Q nonzero random integer elements ($0 < Q \leqslant P-N$), randomly positioned.
- The matrix $V \in \mathbb{R}^{N\times P}$ of problem RC-2D-N-P is formed of 4 blocs : $V_{1:2,1:N-2} = 0$, $V_{3:N,N-1:P} = 0$, and the remaining blocks have random integer data.
- The problems RC-PERM-N refer to the hyperplane arrangements that are called *permutahedron* in [208] : the matrix $V \in \mathbb{R}^{N\times P}$ is such that $V_{:,[1:N]}$ is the identity matrix and $V_{:,[N+1:P]}$ is a Coxeter matrix [203] (each column is of the form $e_i - e_j$ for some $i \neq j$ in $[1 : N]$, where e_k is the kth basis vector of $\mathbb{R}^N$); while $P = N(N+1)/2$.
- The problems RC-RATIO-N-P-R refer to the problems that are controlled by a degeneracy ratio ρ in [208] : the first N columns of the matrix $V \in \mathbb{R}^{N\times P}$ are randomly generated, while the other $P-N > 0$ columns can either (with a probability ρ) be linear combination of the previously generated columns or randomly generated.
- The problems BEK-THRESHOLD-N refer to the *threshold arrangements* in [35, § 6.2] : for $N \geqslant 2$, each column of $V \in \mathbb{R}^{N\times p}$ is formed of the components of $(1, w)$ where $w \in \mathbb{R}^{N-1}$ are all the vectors of $\{0, 1\}^{N-1}$ (hence $p = 2^{N-1}$). This arrangement appears in the study of neural networks [251].
- The problems BEK-RESONANCE-N refer to the *resonance arrangements* in [35, § 6.3] : the columns of $V \in \mathbb{R}^{N\times p}$ are all the nonzero vectors with components in $\{0, 1\}$ (hence $p = 2^N - 1$). Note that, for this arrangement, the number of chambers (i.e., $|\mathcal{S}|$ in our notation) is only known for $N \leqslant 9$. Our approach, which does not use the particular structure of this arrangement, can get $|\mathcal{S}|$ in a reasonable time on a laptop for $N \leqslant 6$, which is to be compared to $N \leqslant 9$ in [35]. See [148] for applications.
- The problems BEK-CROSSPOLYTOPE-N refer to the *cross-polytope arrangements* in [35, § 6.4] : for $N \geqslant 2$, each column of $V \in \mathbb{R}^{N\times p}$ is formed of the components of $(1, w)$ where $w \in \mathbb{R}^{N-1}$ are all the $\pm e_i$ for $i \in [1 : N-1]$; hence $p = 2(N-1)$. For these problems, one numerically observes that $|\mathcal{S}| = 2^1 3^{N-1} - 2^{N-1}$ for $N \leqslant 12$ (this observation is made for $N \leqslant 21$ in [35]).
- The problems BEK-DEMICUBE-N refer to the *demicube arrangements* in [35, § 6.6] : the columns of $V \in \mathbb{R}^{N\times p}$ are the components of $(1, w)$ where $w \in \{w' \in \{0, 1\}^{N-1} : \sum_i w'_i \text{ is odd}\}$.

We have retained 3 problems per family, the most difficult that ISF can solve in a reasonable time for the RC-PERM and BEK families. These test-problems are available on Github and Software Heritage [75].

B. Observations on table 3.1. The dimensions n, p and r of the problems are given in columns 2-4 of table 3.1. Column 5 gives the number ς of matroid circuits of V, which is known to be bounded by $\varsigma_{\max} := \binom{p}{r+1}$ ($= 0$ if $r = p$) [70, 2006, theorem 2.1], whose value is given in column 6. In columns 7 and 8, one finds the cardinality $|\partial_B H(x)| = |\mathcal{S}|$ of the B-differential $\partial_B H(x)$ and the Schläfli upper bound (the right-hand side of (3.39)). The codes will be compared on the number of linear optimization problems (LOPs) they solve, which is a good image of their computation effort, measured independently of the computer used to run the codes and the features of the LOP solver. A first example of comparison is given in columns 9–11 of table 3.1, where one finds the number of LOPs solved by the original RC algorithm and the simulated RC algorithm implemented in the ISF code, as well as the difference between these two numbers. The latter code will be used next, in the comparison with its improved versions, both regarding the LOP counters (table 3.2) and the CPU times (table 3.3).

1) The randomly generated problems RAND are likely to provide vectors v_i's (the columns of V) in general position, in the sense of definition 3.4.9. This can be seen indirectly on the numbers in table 3.1.
   - It is known from proposition 3.4.10 that (3.38) implies equality in (3.39). This equality indeed holds, as we can observe by comparing columns 7 and 8.
   - The same phenomenon occurs with the bound $\varsigma_{\max}$, which is reached by ς if and only if the vectors are in general position [70, 2006, theorem 2.1].
2) The number of matroid circuits, given in the column labeled by ς, depends on the determination of the nonzero elements of the normalized vector $\alpha \in \mathcal{N}(V_{:,I}) \setminus \{0\}$ for the selected index set I (proposition 3.3.11). This operation is sensitive to a threshold value that is set to $10^5\varepsilon$, where $\varepsilon > 0$ is the machine epsilon; smaller values for this threshold have occasionally given larger numbers of matroid circuits. In other words, due to the floating point calculation, there is no certainty that the given number of circuits is the one that would be obtained in exact arithmetic. With a computation in rational numbers, this difficulty is avoided [79].

Problem	n	p	r	ς	$\varsigma_{\max}$	$\lvert\partial_B H(x)\rvert$	Schläfli's bound	LOPs solved in: Original RC	LOPs solved in: Simulated RC	LOPs solved in: Difference
RAND-8-15-7	8	15	7	6435	6435	12952	12952	9908	9907	1
RAND-9-16-8	9	16	8	11440	11440	32768	32768	22821	22818	3
RAND-10-17-9	10	17	9	19448	19448	78406	78406	50643	50642	1
SRAND-8-20-2	8	20	8	540	167960	24544	188368	28748	28620	128
SRAND-8-20-4	8	20	8	84390	167960	157192	188368	136133	135566	567
SRAND-8-20-6	8	20	8	159702	167960	186430	188368	167545	167262	283
RC-2D-20-6	6	20	6	560	77520	512	33328	1936	1927	9
RC-2D-20-7	7	20	7	455	125970	960	87592	3392	3343	49
RC-2D-20-8	8	20	8	364	167960	1792	188368	5888	5855	33
RC-PERM-6	6	21	6	1172	116280	5040	43400	10417	9346	1071
RC-PERM-7	7	28	7	8018	4292145	40320	795188	99155	90169	8986
RC-PERM-8	8	36	8	62814	94143280	362880	17463696	1036897	953009	83888
RC-RATIO-20-5-7	5	20	5	34556	38760	8470	10072	13798	13785	13
RC-RATIO-20-6-7	6	20	6	56184	77520	26194	33328	32993	32980	13
RC-RATIO-20-7-7	7	20	7	112576	125970	76790	87592	82751	82738	13
BEK-THRESHOLD-4	4	8	5	20	28	104	128	88	87	1
BEK-THRESHOLD-5	5	16	5	1348	8008	1882	3882	2758	2757	1
BEK-THRESHOLD-6	6	32	6	353616	3365856	94572	412736	248522	248521	1
BEK-RESONANCE-4	4	15	4	638	3003	370	940	705	635	70
BEK-RESONANCE-5	5	31	5	100091	736281	11292	63862	37766	36311	1455
BEK-RESONANCE-6	6	63	6	(1)	553270671	1066044	14137242	6272462	6164040	108422
BEK-CROSSPOLYTOPE-11	11	20	11	45	125970	117074	709044	111442	86526	24916
BEK-CROSSPOLYTOPE-12	12	22	12	55	497420	352246	2802584	339958	260601	79357
BEK-CROSSPOLYTOPE-13	13	24	13	66	1961256	1058786	11092764	1032162	788970	243192
BEK-DEMICUBE-5	5	8	5	6	28	146	198	106	99	7
BEK-DEMICUBE-6	6	16	6	460	11440	3756	9888	4752	4719	33
BEK-DEMICUBE-7	7	32	7	324640	10518300	291558	1885298	678453	674663	3790

Table 3.1 – Description of the test-problems and comparison of the "original RC algorithm in [208]", written in Python, and the "simulated RC algorithm 3.5.5", written in Matlab : "(n, p, r, ς)" are the features of the problem ($V \in \mathbb{R}^{n\times p}$ is of rank r and has ς circuits, this last number being known to be bounded by $\varsigma_{\max}$), "$|\partial_B H(x)|$" is the cardinality of the B-differential of H given by (3.3), "Schläfli's bound" is the right-hand side of (3.39), "Original RC" gives the number of linear optimization problems (LOPs) solved by the original piece of software in Python of Rada and Černý [208], "Simulated RC" gives the number of LOPs solved by the implementation in the Matlab code ISF of the Rada and Černý algorithm (see algorithm 3.5.5), "Difference" is the difference between the two previous columns. Note (1) : computer crash after several weeks of computation.

	Number of linear optimization problems (LOPs) solved and acceleration ratio (Ratio) for various options														
	Simulated	ISF (A)		ISF (AB)		ISF (ABC)		ISF ($ABCD_1$)		ISF ($ABCD_2$)		ISF ($ABCD_3$)		ISF (AD_4)	
Problem	RC	LOP	Ratio	LOP	Ratio	LOP	Ratio	LOP	Ratio	LOP	Ratio	LOP	Ratio	LOP	Ratio
RAND-8-15-7	9907	9844	1.01	7641	1.30	5210	1.90	5199	1.91	4355	2.27	3638	2.72	0	—
RAND-9-16-8	22818	22691	1.01	17586	1.30	13046	1.75	13023	1.75	11185	2.04	9943	2.29	0	—
RAND-10-17-9	50642	50387	1.01	38167	1.33	28849	1.76	28839	1.76	25370	2.00	23266	2.18	0	—
SRAND-8-20-2	28620	28620	1.00	20207	1.42	6668	4.29	5535	5.17	2881	9.93	2851	10.04	0	—
SRAND-8-20-4	135566	136027	1.00	113493	1.19	60066	2.26	59267	2.29	45569	2.97	42445	3.19	0	—
SRAND-8-20-6	167262	167351	1.00	137450	1.22	77800	2.15	77752	2.15	62694	2.67	54980	3.04	0	—
RC-2D-20-6	1927	1904	1.01	1680	1.15	912	2.11	688	2.80	40	48.17	0	—	0	—
RC-2D-20-7	3343	3296	1.01	2912	1.15	2208	1.51	1792	1.87	52	64.29	0	—	0	—
RC-2D-20-8	5855	5760	1.02	4992	1.17	2752	2.13	1984	2.95	28	209.11	0	—	0	—
RC-PERM-6	9346	9280	1.01	7898	1.18	2076	4.50	1836	5.09	92	101.59	61	153.21	0	—
RC-PERM-7	90169	90094	1.00	79049	1.14	17230	5.23	16558	5.45	960	93.93	855	105.46	0	—
RC-PERM-8	953009	952597	1.00	856597	1.11	160781	5.93	158989	5.99	9766	97.58	9393	101.46	0	—
RC-RATIO-20-5-7	13669	15341	0.89	14028	0.97	7108	1.92	7064	1.94	3644	3.75	2467	5.54	0	—
RC-RATIO-20-6-7	32883	35882	0.92	31992	1.03	17797	1.85	17505	1.88	10669	3.08	8765	3.75	0	—
RC-RATIO-20-7-7	82447	81428	1.01	72272	1.14	47798	1.72	47748	1.73	30442	2.71	25841	3.19	0	—
BEK-THRESHOLD-4	87	79	1.10	54	1.61	46	1.89	37	2.35	26	3.35	16	5.54	0	—
BEK-THRESHOLD-5	2757	2884	0.96	2399	1.15	1270	2.17	1180	2.34	502	5.49	370	3.75	0	—
BEK-THRESHOLD-6	248521	261728	0.95	236027	1.05	71963	3.45	70410	3.53	21339	11.65	19184	3.19	0	—
BEK-RESONANCE-4	635	672	0.94	546	1.16	171	3.71	138	4.60	31	20.48	0	—	0	—
BEK-RESONANCE-5	36311	37607	0.97	34056	1.07	6700	5.42	6569	5.53	1141	31.82	810	44.83	0	—
BEK-RESONANCE-6	6164040	6269410	0.98	5956586	1.03	760930	8.10	760457	8.11	155555	39.63	(1)	—	0	—
BEK-CROSSPOLYTOPE-11	86526	110418	0.78	58954	1.47	17569	4.92	15265	5.67	6085	14.22	6049	14.30	0	—
BEK-CROSSPOLYTOPE-12	260601	337910	0.77	182575	1.43	46900	5.56	41780	6.24	18785	13.87	18740	13.91	0	—
BEK-CROSSPOLYTOPE-13	788970	1028066	0.77	560013	1.41	124828	6.32	113564	6.95	57299	13.77	57244	13.78	0	—
BEK-DEMICUBE-5	99	90	1.10	33	3.00	24	4.12	12	8.25	3	33.00	0	—	0	—
BEK-DEMICUBE-6	4719	4761	0.99	3659	1.29	1882	2.51	1741	2.71	665	7.10	588	8.03	0	—
BEK-DEMICUBE-7	674663	704553	0.96	623160	1.08	175870	3.84	175595	3.84	60876	11.08	58333	11.57	0	—
Mean			0.97		1.28		3.45		3.88		31.54		24.52		—
Median			1.00		1.17		2.51		2.95		11.65		5.54		—

TABLE 3.2 – Evaluation of the efficiency of the solvers by the number of LOPs they solve : A (taking the rank of V into account), B (special handling of the case where $v_{k+1}^{\mathsf{T}} d \simeq 0$), C (changing the order of the vectors v_i's by taking i_{k+1} by (3.49)), D_1 (pre-computation of $2(p-r)$ stem vectors after the QR factorization), D_2 (D_1 and 2 additional stem vectors computed after solving a LOP, whose optimal value is nonnegative), D_3 (all the stem vectors are first computed and, for $(s, \pm 1) \in \mathcal{S}_{k+1}$, a LOP is solved to get a handle d), D_4 (all the stem vectors are first computed and no LOP is solved). The "Ratio" (acceleration ratio) columns give for each considered problem the ratio (*LOPs of the considered* ISF *version*)/(*LOPs of simulated* RC). Note (1) : interruption of the run after several days of computation. The Mean/Median rows give the mean and medan values of the ratios.

3) A comparison between the "Original rc code" in Python and its "Simulated rc code" in Matlab shows that the latter is slightly more effective in terms of the number of LOPs solved. This is probably due to the special treatment in step 2 of the case where $v_{k+1}^{\mathsf{T}} d \simeq 0$ in algorithm 3.5.6, which is not considered in the original code.

C. Observations on table 3.2. Table 3.2 shows the effect of the modifications discussed in section 3.5.2 on the number of LOPs solved, which significantly counts in the computing time. This will lead us to select three algorithms, those which bring the best profit on the LOP counter. The columns labeled "Ratio" show the acceleration ratio with respect to the simulated rc code in terms of LOPs, that is the ratio of the LOP counter of the considered algorithm divided by the LOP counter of the simulated rc algorithm. On the last two lines of the table, one finds the mean and median values of these acceleration ratios, which may be viewed as a summary of the effect of the considered modification. These mean/median values must be taken with caution when a solver fails to solve a problem as is the case with isf($ABCD_3$) and isf(AD_4) on problem bek-resonance-6.

1) The modification A, proposed in section 3.5.2(A), which uses the QR factorization to get r linearly independent columns of V, does not bring a large benefit ("Ratio" is close to 1) and sometimes increases the number of LOPs to solve. The benefit is not important since it "only" prevents $\sum_{i\in[0\,:\,r-1]} 2^i = 2^r - 1$ nodes from running the LOP solver, which is usually a small fraction of the total number of nodes of the $\mathcal{S}$-tree. One also observes that the number of solved LOPs may increase (acceleration ratio < 1), which is sometimes due to the fact that the number 2^{r-1} of nodes at level r with modification A is larger than the one without modification A, which contributes to increase the total number of nodes of the constructed $\mathcal{S}$-tree and, therefore, tends to increase the number of LOPs to solve. Furthermore, the order in which the vectors are considered without/with modification A is not identical, which has also an impact on the number of solved LOPs (see section 3.5.2(C)).
2) The modification B, proposed in section 3.5.2(B), which is able to detect two descendants of an $\mathcal{S}$-tree node, without solving any LOP, has a significant impact on the total number of these problems. We see, indeed, that the (mean, median) acceleration ratio is raised to (1.28, 1.17).
3) Consider now the modification C, described in section 3.5.2(C), which changes the order in which the vectors v_i's are considered. We use the test-problem rand-7-13-5 to show its effect in the next table.

	Number of nodes per level													Total
With modifications AB	1	2	4	8	16	31	57	99	163	256	386	562	794	2379
With modifications ABC	1	2	4	8	16	26	43	69	107	168	270	443	794	1951
$\mathcal{S}$-tree levels	1	2	3	4	5	6	7	8	9	10	11	12	13	

The table gives the number of nodes for each level in the $\mathcal{S}$-tree, with the modifications

AB and with the modifications ABC. Since $\mathrm{rank}(V) = 5$ for this problem and since the modification A is used in both cases, the number of nodes per level, only starts to differ from level 6 (before that it is equal to 2^{l-1}, where l is the $\mathcal{S}$-tree level). The final level is 13 (since there are $p = 13$ vectors) and its number of leaves is $|\mathcal{S}|/2 = 794$ (an observation from the table above), necessary identical in both cases. The effect of the modification C can be seen on the smaller number of nodes per level and in all the $\mathcal{S}$-tree (rightmost column). This contributes to the decrease of the number of LOPs to solve : the (mean, median) acceleration ratio is raised to (3.45, 2.51).

4) The modifications D, described in section 3.5.2(D), deal with the contribution of the computed stem vectors, whose number increases from modification D_1 ($2(p-r)$ stem vectors after the QR factorization of V), D_2 (more stem vectors from the dual solution of the LOP (3.43) when this one has a nonnegative optimal value), D_3 and D_4 (all the stem vectors).
   - We see that the option D_1 yields already some improvement (less LOPs to solve), but not much, raising the (mean, median) acceleration ratio from (3.45, 2.51) to (3.88, 2.95).
   - The use of the option D_2 is more beneficial since the (mean, median) acceleration ratio now goes up to (31.54, 11.65). We understand this fact to have its origin in the increase in the number of stem vectors detected from the dual solutions of some solved LOP. Note that this last operation does not require much computation time.
   - With option D_3, only the LOPs (3.43) with the optimal value -1 are solved, while, with option D_4, no LOP is solved. The efficiency of these modifications largely depends on the total number 2ς of stem vectors. If this one is not too large, the modifications have an important benefit. Otherwise, it can lead to execution failure, as for problem BEK-RESONANCE-6, which requires days of computation.

In conclusion of these observations, one could retain the following three solvers for a comparison on their computing time.

- ISF(ABCD_2) is the most efficient solver that does not compute all the stem vectors.
- The solvers ISF(ABCD_3) and ISF(AD_4) cannot be compared with the other solvers on the results of table 3.2 since both use all the stem vectors, so that the time to compute and use these must be taken into account, and ISF(AD_4) does not solve any LOP, which is the measure of efficiency in table 3.2.

D. OBSERVATIONS ON TABLE 3.3. Measuring the efficiency of the algorithms by the number of LOPs solved during execution, like in table 3.2, is sometimes misleading. If this is the main cost item for some algorithms, it is no longer the case when a large amount of stem vectors is computed. For two reasons. First, the time spent in the computation of these stem vectors is not negligible, far from it, at least in our implementation, in which each of them requires the computation of the nullity of a matrix and a null space vector. Next, verifying that a

	CPU times (in sec)						
	Simulated	ISF ($ABCD_2$)		ISF ($ABCD_3$)		ISF (AD_4)	
Problem	RC	Time	Ratio	Time	Ratio	Time	Ratio
RAND-8-15-7	71.77	33.27	2.16	32.91	2.18	5.62	12.77
RAND-9-16-8	151.39	75.45	2.01	82.30	1.84	14.43	10.49
RAND-10-17-9	347.32	185.05	1.88	198.18	1.75	55.96	6.21
SRAND-8-20-2	174.44	16.91	10.32	19.64	8.88	3.66	47.68
SRAND-8-20-4	832.74	309.15	2.69	450.35	1.85	349.83	2.38
SRAND-8-20-6	1011.30	483.97	2.09	732.82	1.38	746.49	1.35
RC-2D-20-6	11.01	0.32	34.71	0.25	43.53	0.22	50.95
RC-2D-20-7	19.88	0.50	39.95	0.50	39.68	0.38	52.97
RC-2D-20-8	35.87	0.41	87.97	0.74	48.56	0.63	56.78
RC-PERM-6	53.29	0.76	70.05	2.10	25.41	1.90	28.00
RC-PERM-7	549.04	7.44	73.78	45.62	12.04	67.10	8.18
RC-PERM-8	6171.22	74.93	82.36	1233.80	5.00	3355.22	1.84
RC-RATIO-20-5-7	83.34	22.71	3.67	28.36	2.94	18.58	4.49
RC-RATIO-20-6-7	202.09	72.04	2.81	101.51	1.99	112.12	1.80
RC-RATIO-20-7-7	504.52	247.99	2.03	351.08	1.44	353.15	1.43
BEK-THRESHOLD-4	0.61	0.18	3.44	0.11	5.46	0.01	74.64
BEK-THRESHOLD-5	17.43	3.56	4.89	2.83	6.16	0.35	50.40
BEK-THRESHOLD-6	1758.16	194.75	9.03	4577.26	0.38	6532.56	0.27
BEK-RESONANCE-4	3.97	0.22	17.71	0.09	46.12	0.08	48.99
BEK-RESONANCE-5	228.41	7.90	28.90	44.78	5.10	183.84	1.24
BEK-RESONANCE-6	38296.20	1988.60	19.26	(1)	—	(1)	—
BEK-CROSSPOLYTOPE-11	480.07	34.35	13.97	39.27	12.22	7.63	62.95
BEK-CROSSPOLYTOPE-12	1579.19	108.76	14.52	124.66	12.67	25.22	62.62
BEK-CROSSPOLYTOPE-13	5017.73	322.43	15.56	404.80	12.40	104.22	48.15
BEK-DEMICUBE-5	0.55	0.02	25.24	0.01	85.73	0.01	108.69
BEK-DEMICUBE-6	27.38	4.15	6.59	4.09	6.69	0.43	63.82
BEK-DEMICUBE-7	4310.35	510.25	8.45	2405.08	1.79	6396.66	0.67
Mean			21.71		15.12		31.14
Median			10.32		5.81		20.39

Table 3.3 – Evaluation of the efficiency of the solvers by their computing times. The "Ratio" (acceleration ratio) columns give for each considered problem the ratio (*Time of the considered* ISF *version*)/(*Time of simulated* RC). Note (1) : interruption of the run after several days of computation. The Mean/Median rows give the mean and medan values of the ratios.

sign vector contains a stem vector (proposition 3.3.10) is also time consuming when there are many stem vectors. Therefore a comparison of the CPU time of the runs is welcome. This is done for a selection of versions of the ISF codes in table 3.3, those selected at the end of section 3.5.2(C). Here are some observations on the statistics of this table.

1) A first observation is that the good behavior of the selected versions of the ISF codes is confirmed, even though the acceleration ratios are not as large as the one based on the number of LOPs solved. This can be explained by the fact that the time spent in solving LOPs is counterbalanced by the handling of stem vectors for the versions $ABCD_3$ and AD_4. Anyway, one observes that the CPU time acceleration ratios have (mean, median) values in the ranges (15..31, 5..20), which is significant.
2) The most effective combination of code options depends actually on the considered problems. It is difficult to state a rule that would predict which code behaves best because some solvers are better on some phases of the run, but worse on others (the three main phases are the detection of the stem vectors, the execution of LOPs and the search for stem vectors covered by a given sign vector). However, an inductive rule manifests itself : the purely dual method AD_4 is ahead for problems with a reasonable number of stem vectors (or matroid circuits), but can require a too large number of computing time if this number becomes large (this is the case of problems BEK-THRESHOLD-6, BEK-RESONANCE-6 and BEK-DEMICUBE-7). This conclusion could be invalidated if better techniques are used to enumerate and use the stem vectors.

3.6 Discussion

This paper deals with the description and computation of the B-differential of the componentwise minimum of two affine vector functions. The fact that this problem has many equivalent formulations, some of them being highlighted in section 3.3, implies that the present contribution has an impact on several domains, including on the description of the arrangement of hyperplanes in the space. To this respect, a singular aspect of this contribution is to propose a dual approach to solve the problem, using some or all the stem vectors, a concept made useful thanks to the convex analysis tool that is Gordan's alternative. Besides this contribution, the paper also brings various improvements of an algorithm of Rada and Černý [208], which was designed to determine the cells of an arrangement of hyperplanes in the space.

Even in the spirit of the methods proposed in this article, there is still room for improvement, in relation to three identified bottlenecks : (i) we have mentioned that with the option D_3, the LOP (3.43) can be solved inexactly, since, in that case, the optimal value is -1, while any negative objective value for a feasible unknown would suffice, but this requires a better tuning of the linear optimization solver, (ii) computing more efficiently all the stem vectors (or matroid circuits) of the matrix V is certainly a source of improvement,

(iii) a better algorithm to decide more rapidly that a sign vector contains a stem vector is also welcome. Some of these possible improvements are also linked to a better choice of programming language, probably one using a compilation phase.

This contribution has also various possible extensions. A first one would be to develop a dual approach to the problem of the arrangement in the space of hyperplanes *having no point in common* [79]. Another natural extension would be to see the implications of this work for computing the B-differential of the componentwise minimum of *nonlinear* vector functions [74]. Finally, the possibility to take profit of the computation of the full B-differential of the function H in (3.3) in a Newton-like approach to solve (3.4) is a subject that deserves reflection.

Acknowledgments

We thank Michal Černý and Miroslav Rada for providing their code and test problems, those used in [208] ; part of these were used in the numerical experiments. We also thank the referees for their remarks and recommendations, which have helped us make the paper more readable.

Statements & Declarations

Financial interests. The authors have no relevant financial or non-financial interests to disclose.

Conflict of interest. All authors declare that they have no conflicts of interest.

Code and data availability. The code isf described in this paper and the data on which it has been assessed are publicly available on Github and Software Heritage. URLs are included in the reference [75].

Chapitre 4

Éléments complémentaires sur le B-différentiel du minimum et les arrangements

Ce chapitre plus court vise à fournir des éléments complémentaires au précédent. Ces éléments consistent en des propriétés connexes mais ne se concentrent pas spécifiquement sur le calcul de $\partial_B H(x)$ avec $H : x \mapsto \min(Ax + a, Bx + b)$. Après quelques preuves, nous commençons ce chapitre en donnant quelques exemples de situations pouvant survenir concernant les propriétés de régularité introduites dans la section 2.3.2.

Ensuite, le cas des fonctions non linéaires lisses F et G (au lieu de $F(x) \equiv Ax + a$ et $G(x) \equiv Bx + b$) est considéré : les ensembles $\{x \in \mathbb{R}^n : F_i(x) = G_i(x)\}$ pour chaque $i \in [1 : n]$ ne sont plus des hyperplans, ce qui complique nettement tout calcul direct. Nous considérons les linéarisations de F et G en x, et proposons des conditions nécessaires et suffisantes pour garantir que le B-différentiel du minimum de ces linéarisations est égal au vrai B-différentiel ($\subseteq$ est toujours vrai mais la réciproque est généralement fausse).

La section suivante détaille une propriété de règle de chaîne sur $\partial_B \theta(x)$, où $\theta = ||H||^2/2$. Essentiellement, nous montrons que $\partial_B \theta(x) = \partial_B H(x)^\intercal H(x)$. Bien que cette formule ne soit pas très surprenante, sa preuve repose fortement sur la structure inhérente du minimum composante par composante.

Enfin, nous donnons des précisions supplémentaires concernant certaines instances testées dans le chapitre 3, comme les structures et cardinaux de $\mathcal{S}$ et $\mathfrak{S}$. Lorsque possible, nous proposons également des explications des comportements des heuristiques/algorithmes testés sur certains types d'instances.

Rappelons que nous considérons le problème général de complémentarité non linéaire suivant

$$0 \leqslant F(x) \perp G(x) \geqslant 0,$$

où F et G sont des fonctions lisses de $\mathbb{R}^n$ dans $\mathbb{R}^n$, en utilisant la reformulation via la fonction C-minimum, conduisant au système non lisse et à la minimisation de sa fonction de mérite associée

$$H(x) = \min(F(x), G(x)) = 0, \qquad \min \theta(x) := \frac{1}{2}||H(x)||^2.$$

Éventuellement, on pourra s'intéresser au cas linéaire/affine (voir chapitre 3) avec $F(x) = Ax + a$ et $G(x) = Bx + b$,

$$0 \leqslant Ax + a \perp Bx + b \geqslant 0,$$

Dans les deux cas, on considère les ensembles d'indices :

$$\begin{array}{rcl} \mathcal{E}(x) & := & \{i \in [1:n] : F_i(x) = G_i(x)\}, \\ \mathcal{F}(x) & := & \{i \in [1:n] : F_i(x) < G_i(x)\}, \\ \mathcal{G}(x) & := & \{i \in [1:n] : F_i(x) > G_i(x)\}. \end{array} \tag{4.1}$$

4.1 Matériel complémentaire du chapitre précédent

Donnons les preuves de quelques propriétés présentées dans le chapitre précédent, mais qui n'ont pas été jugées assez importantes pour figurer dans l'article.

Proposition 4.1.1 (sur-ensemble de $\partial_B H(x)$)**.** *On a*

$$\partial_B H(x) \subseteq \partial_B H_1(x) \times \cdots \times \partial_B H_m(x) = \partial_B^\times H(x). \tag{4.2}$$

En particulier, $|\partial_B H(x)| \leqslant 2^p$.

Preuve. L'inclusion dans (4.2) est claire puisque, lorsque $H'(x_k)$ converge vers un certain J, $H_i'(x_k) \to J_{i,:}$, pour tout $i \in [1:m]$. L'égalité est également claire en conséquence du lemme 3.2.1 ([78, lemme 2.1.4)]).

La dernière affirmation découle directement du fait que $J_{i,:}$ ne peut prendre que deux valeurs différentes, $A_{i,:}$ ou $B_{i,:}$, uniquement pour les indices $i \in \mathcal{E}^{\neq}(x)$ (rappelons que $|\mathcal{E}^{\neq}(x)| = p$). □

Proposition 4.1.2 (un lien avec le C-différentiel)**.** $\partial_B H(x) = \operatorname{ext} \partial_C H(x)$.

Preuve. Remarquons d'abord que, puisque $\mathcal{S}$ donné par (3.12) est inclus dans $\{\pm 1\}^p$, on a $\mathcal{S} = \operatorname{ext}(\operatorname{conv}\mathcal{S})$. Pour obtenir le résultat, il suffit maintenant de transporter cette identité dans $\mathbb{R}^{p\times n}$ via l'application affine $\tau : \mathbb{R}^p \to \mathbb{R}^{p\times n}$ définie en $s \in \mathbb{R}^p$ par

$$\tau(s) = \frac{1}{2}\left[(I - \operatorname{Diag}(s))B_{\mathcal{E}^{\neq}(x),:} + (I + \operatorname{Diag}(s))A_{\mathcal{E}^{\neq}(x),:}\right].$$

La restriction de τ à $\mathcal{S}$ est $\tau|_{\mathcal{S}} = \sigma^{-1}$, définie par (3.14b). De plus, τ est injective, car $A_{i,:} \neq B_{i,:}$ pour $i \in \mathcal{E}^{\neq}(x)$. Par conséquent, en appliquant τ aux deux membres de l'identité $\mathcal{S} = \mathrm{ext}(\mathrm{conv}\mathcal{S})$, on obtient

$$\begin{aligned} \tau(\mathcal{S}) &= \mathrm{ext}(\tau(\mathrm{conv}\mathcal{S})) \qquad [\text{injectivité de } \tau \text{ [105, prop. 2.12(2)]}] \\ &= \mathrm{ext}(\mathrm{conv}(\tau(\mathcal{S}))) \qquad [\text{affinité de } \tau \text{ [105, prop. 2.5(1)]}]. \end{aligned}$$

Le résultat découle maintenant du fait que $\tau(\mathcal{S}) = \sigma^{-1}(\mathcal{S}) = \partial_B H(x)$ (proposition 3.3.4) et $\partial_C H(x) = \mathrm{conv}\partial_B H(x)$. □

Enfin, mentionnons une approche, suggérée par un rapporteur de [77] avant acceptation pour publication, qui nous semble trop restrictive lorsqu'on se concentre sur le B-différentiel $\partial_B H(x)$ et qui ne permet pas de décrire un arrangement d'hyperplans gouverné par une matrice $V \in \mathbb{R}^{n\times p}$ avec $p > n$. Si $\partial_B H(x)$ est l'objet principal, on peut écrire $H(x) = Ax + a - K(x)$, où $K(x) := \mathrm{P}_{\mathbb{R}^n_+}[Mx + q]$, $\mathrm{P}_{\mathbb{R}^n_+}$ est le projecteur orthogonal sur l'orthant positif, $M = A - B$ et $q = a - b$, de sorte que

$$\partial_B H(x) = A - \partial_B K(x). \tag{4.4}$$

Pour exploiter la formule explicite de $\partial_B \mathrm{P}_{\mathbb{R}^n_+}$, on peut chercher des conditions garantissant que la règle de chaîne s'applique pour la composition définissant l'application $K : \mathbb{R}^n \to \mathbb{R}^n$. Nous affirmons que, lorsque la règle de chaîne s'applique, le B-différentiel $\partial_B H(x)$ est complet au sens de la définition 3.2.4, ce qui est un cas très particulier. Cette approche présente donc un intérêt trop limité.

Supposons en effet que la règle de chaîne s'applique pour le calcul de $\partial_B K(x)$. On aurait alors

$$\partial_B K(x) = [\partial_B \mathrm{P}_{\mathbb{R}^n_+}(Mx + q)]M.$$

Or, $\partial_B \mathrm{P}_{\mathbb{R}^n_+}(z)$ est connu explicitement comme le produit cartésien de

$$[\partial_B \mathrm{P}_{\mathbb{R}^n_+}(z)]_i = \begin{cases} \{0\} & \text{si } z_i < 0, \\ \{0,1\} & \text{si } z_i = 0, \\ \{1\} & \text{si } z_i > 0, \end{cases}$$

pour $i \in [1:n]$. Par conséquent, $\partial_B K(x)$ est le produit cartésien de

$$[\partial_B K(z)]_i = \begin{cases} \{0\} & \text{si } (Ax+a)_i < (Bx+b)_i, \\ \{0, A_{i,:} - B_{i,:}\} & \text{si } (Ax+a)_i = (Bx+b)_i, \\ \{A_{i,:} - B_{i,:}\} & \text{si } (Ax+a)_i > (Bx+b)_i, \end{cases}$$

pour $i \in [1:n]$. Avec (4.4), $\partial_B H(x)$ s'écrit comme le produit cartésien de

$$[\partial_B H(x)]_i = \begin{cases} \{A_{i,:}\} & \text{si } (Ax+a)_i < (Bx+b)_i, \\ \{A_{i,:}, B_{i,:}\} & \text{si } (Ax+a)_i = (Bx+b)_i, \\ \{B_{i,:}\} & \text{si } (Ax+a)_i > (Bx+b)_i, \end{cases}$$

pour $i \in [1:n]$. C'est la formule du B-différentiel complet.

4.2 Notions de régularité et contre-exemples

Cette section vise à décrire quelques (contre-)exemples pour montrer comment les notions de BD-régularité 2.3.14 et de BD-régularité forte interagissent avec le B-différentiel. Ils montrent en particulier que : la BD-régularité n'implique pas la BD-régularité forte, que la BD-régularité forte n'implique pas que le B-différentiel soit complet et que le B-différentiel peut être complet malgré le fait que tous ses éléments soient singuliers.

Définissons $V := B^{\mathsf{T}}_{\mathcal{E}(x),:} - A^{\mathsf{T}}_{\mathcal{E}(x),:}$, les matrices jacobiennes de $\partial_B H(x)$, le B-différentiel de H, sont les matrices $J(s)$ définies par

$$J(s)_{\mathcal{F}(x)\cup\mathcal{G}(x),:} = \begin{bmatrix} F'_{\mathcal{F}(x)}(x) \\ G'_{\mathcal{G}(x)}(x) \end{bmatrix}, \quad J(s)_{\mathcal{E}(x),:} = \frac{s+e}{2} \cdot F'_{\mathcal{E}(x)}(x) + \frac{e-s}{2} \cdot G'_{\mathcal{E}(x)}(x).$$

pour des vecteurs s définis par

$$s \in \{\{\pm 1\}^{\mathcal{E}(x)} : \exists\, d, s \cdot V^{\mathsf{T}} d > 0\}.$$

Ces exemples montrent la particularité de $\partial_B H(x)$, où ses éléments sont déterminés par la matrice $F'(x) - G'(x)$ mais leur (non-)singularité est déterminée par $F'(x)$ et $G'(x)$.

Exemple 4.2.1 (BD régulier mais pas fortement BD régulier)**.** Soit $\mathcal{F}(x) = \varnothing = \mathcal{G}(x)$ et $\mathcal{E}(x) = \{1, 2\}$. Soient A et B définis par

$$A = I_2, \qquad B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Puisque $V = B^{\mathsf{T}} - A^{\mathsf{T}} = [0\ 1; 1\ 0]$ est injective, on a $\mathcal{S} = \{\pm 1\}^2$. Cependant, pour $s = (-1, -1)$, on a $J(s) = B$ qui est singulière, donc il n'y a pas de BD-régularité forte. Vérifions que la BD-régularité est satisfaite. Cela donne

$$\begin{pmatrix} \min(d_1, d_1 + d_2) \\ \min(d_2, d_1 + d_2) \end{pmatrix},$$

qui peut être nul si et seulement si $(d_1, d_2) = 0$. En effet, la BD-régularité s'écrit (voir définition 2.3.14) $d \neq 0 \Rightarrow H'(x; d) \neq 0$. □

Exemple 4.2.2 (toutes les matrices sont inversibles mais pas toutes dans le B-différentiel)**.** Considérons l'exemple suivant :

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad a = 0, \qquad B = \begin{bmatrix} 2 & 1 \\ 2 & 3 \end{bmatrix}, \qquad b = 0, \qquad x = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

On obtient clairement que $Ax + a = Bx + b = x$. De plus, les 4 matrices jacobiennes potentielles sont

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 0 \\ 2 & 3 \end{bmatrix}, \qquad \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}, \qquad \begin{bmatrix} 2 & 1 \\ 2 & 3 \end{bmatrix},$$

qui sont toutes non singulières. Cependant, en x, la matrice impliquée dans la détermination de $\partial_B H(x)$ est

$$V = (B - A)^\mathsf{T} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix},$$

donc seulement deux des quatre matrices jacobiennes appartiennent à $\partial_B H(x)$. □

Exemple 4.2.3 (B-différentiel complet & jacobiennes singulières)**.** Considérons l'exemple suivant :

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 2 & 2 \end{bmatrix}, \qquad a = 0, \qquad B = \begin{bmatrix} -1 & 3 & 2 \\ 1 & 0 & 1 \\ -1 & 0 & -1 \end{bmatrix}, \qquad b = 0, \qquad x = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}.$$

Au point x, on obtient

$$Ax + a = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad Bx + b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

ce qui signifie $\mathcal{E}(x) = \{2, 3\}$. Maintenant, la matrice impliquée dans le B-différentiel est

$$V = (B - A)^\mathsf{T}_{\mathcal{E}(x)} = \begin{bmatrix} 1 & -1 \\ -1 & -2 \\ 0 & -3 \end{bmatrix}.$$

Puisque ses deux colonnes sont indépendantes, le B-différentiel est complet et les quatre jacobiennes

$$\begin{bmatrix} -1 & 3 & 2 \\ 0 & 1 & 1 \\ 0 & 2 & 2 \end{bmatrix}, \qquad \begin{bmatrix} -1 & 3 & 2 \\ 0 & 1 & 1 \\ -1 & 0 & -1 \end{bmatrix}, \qquad \begin{bmatrix} -1 & 3 & 2 \\ 1 & 0 & 1 \\ 0 & 2 & 2 \end{bmatrix}, \qquad \begin{bmatrix} -1 & 3 & 2 \\ 1 & 0 & 1 \\ -1 & 0 & -1 \end{bmatrix}$$

sont toutes singulières. □

4.3 B-différentiel du minimum de F et G non linéaires

4.3.1 Différentiels de H

Cette première section discute de quelques propriétés de $\partial_B H(x)$ et $\partial_C H(x)$. L'élément principal que nous discutons est le lien entre $\partial_B H(x)$ et $\partial_B(\mathcal{L}_x H)(x)$, le B-différentiel du minimum des linéarisations de F et G. Plus précisément, définissons

$$(\mathcal{L}_x H)(y) := \min\left(F(x) + F'(x)(y - x), G(x) + G'(x)(y - x)\right). \tag{4.5}$$

L'utilisation principale de la linéarisation est la suivante. Comme il s'agit d'un minimum de fonctions affines, il est régi par les propriétés décrites dans le chapitre 3. Ensuite, nous verrons que les différentiels de $\mathcal{L}_x H$ sont des sous-ensembles du "vrai" B-différentiel de H.

Naturellement, si les fonctions F et G sont affines, elles sont égales à leur linéarisation et on a égalité dans (4.6a) et (4.6b). Sinon, on peut facilement ne pas avoir l'égalité, et même avoir un B-différentiel avec un cardinal impair (contre-exemple 4.3.2 plus bas).

Proposition 4.3.1 (sous-ensemble de $\boldsymbol{\partial_B H(x)}$)**.** *Supposons que F et G sont continûment différentiables en x. Soit $\mathcal{L}_x$ l'opérateur défini par* (4.5)*. Alors,*

$$\partial_B(\mathcal{L}_x H)(x) \subseteq \partial_B H(x), \tag{4.6a}$$
$$\partial_C(\mathcal{L}_x H)(x) \subseteq \partial_C H(x). \tag{4.6b}$$

Preuve. L'inclusion (4.6b) peut être déduite de (4.6a) en prenant les enveloppes convexes de ses deux côtés, donc nous devons seulement nous concentrer sur (4.6a).

Premièrement, observons que les ensembles d'indices $\mathcal{F}(x)$, $\mathcal{E}^{\neq}(x)$, $\mathcal{E}^{=}(x)$ et $\mathcal{G}(x)$ sont identiques pour H et $\mathcal{L}_x H$. Soit $\tilde{J} \in \partial_B(\mathcal{L}_x H)(x)$. Nous voulons montrer que $\tilde{J} \in \partial_B H(x)$. Par l'équation (3.10), on obtient

$$\tilde{J}_{i,:} = \begin{cases} F_i'(x) & \text{si } i \in \mathcal{F}(x), \\ F_i'(x) = G_i'(x) & \text{si } i \in \mathcal{E}^{=}(x), \\ F_i'(x) \text{ ou } G_i'(x) & \text{si } i \in \mathcal{E}^{\neq}(x), \\ G_i'(x) & \text{si } i \in \mathcal{G}(x). \end{cases} \tag{4.7a}$$

De plus, on peut trouver une direction $d \in \mathbb{R}^n$ telle que, pour tout $i \in \mathcal{E}^{\neq}(x)$, on a

$$F_i'(x)d - G_i'(x)d < 0, \quad \text{si } \tilde{J}_{i,:} = F_i'(x), \tag{4.7b}$$
$$F_i'(x)d - G_i'(x)d > 0, \quad \text{si } \tilde{J}_{i,:} = G_i'(x). \tag{4.7c}$$

Construisons maintenant une jacobienne $J \in \partial_B H(x)$ et montrons que $J = \tilde{J}$, ce qui conclura la preuve de la proposition. Considérons la suite $\{x_k\}$ définie par

$$x_k := x + t_k d + \sigma(t_k), \tag{4.7d}$$

où $\{t_k\} \downarrow 0$ et $\sigma : \mathbb{R}_+ \to \mathbb{R}^n$ est une fonction (petite perturbation) qui n'a pas besoin d'être continue mais est choisie telle que $\sigma(0) = 0$, $\sigma(t_k) = o(t_k)$ et $x_k \in \mathcal{D}_H$ (ceci est possible par le théorème de Rademacher [209]). En extrayant une sous-suite $\mathcal{K}$ de $\mathbb{N}$ si nécessaire, on peut supposer que, pour chaque indice $i \in [1 : n]$, l'une des trois propriétés suivantes est vérifiée

$$\forall\, k \in \mathcal{K} : \quad F_i(x_k) < G_i(x_k), \tag{4.7e}$$
$$\forall\, k \in \mathcal{K} : \quad F_i(x_k) = G_i(x_k), \tag{4.7f}$$
$$\forall\, k \in \mathcal{K} : \quad F_i(x_k) > G_i(x_k). \tag{4.7g}$$

Par construction de $\{x_k\}$, $x_k \in \mathcal{D}_H$, donc $H'(x_k)$ existe. Montrons que, pour $i \in [1:n]$ et quand $k \to \infty$ dans $\mathcal{K}$, on a

$$\begin{aligned} &\text{(4.7e) est vérifiée} \quad \Longrightarrow \quad H'_i(x_k) \to F'_i(x), && \text{(4.7h)}\\ &\text{(4.7f) est vérifiée} \quad \Longrightarrow \quad H'_i(x_k) \to F'_i(x) = G'_i(x), && \text{(4.7i)}\\ &\text{(4.7g) est vérifiée} \quad \Longrightarrow \quad H'_i(x_k) \to G'_i(x). && \text{(4.7j)} \end{aligned}$$

L'implication (4.7h) (resp. (4.7j)) est claire, puisque $H'_i(x_k) = F'_i(x_k)$ (resp. $H'_i(x_k) = G'_i(x_k)$), $x_k \to x$ et F'_i (resp. G'_i) est continue en x. L'implication (4.7i) vient du fait que, lorsque (4.7f) est vérifié, $H'_i(x_k) = F'_i(x_k) = G'(x_k)$ par la différentiabilité de H_i en x_k (utiliser le lemme 3.2.1), impliquant à nouveau que $H'_i(x_k) \to F'_i(x) = G'_i(x)$. Par définition, la limite J de $H'(x_k)$ est dans $\partial_B H(x)$. Il reste à montrer que $J = \tilde{J}$ pour conclure la preuve de la proposition.

Regardons une ligne arbitraire $i \in [1:n]$ de $J \in \partial_B H(x)$ et $\tilde{J} \in \partial_B(\mathcal{L}_x H)(x)$ et montrons que $J_{i,:} = \tilde{J}_{i,:}$. Par la propriété de différentiabilité de F et G en x et par (4.7d) avec $\sigma(t_k) = o(t_k)$, on a pour $k \in \mathcal{K}$:

$$\begin{cases} F_i(x_k) = F_i(x) + t_k F'_i(x)d + o(t_k), \\ G_i(x_k) = G_i(x) + t_k G'_i(x)d + o(t_k). \end{cases} \tag{4.7k}$$

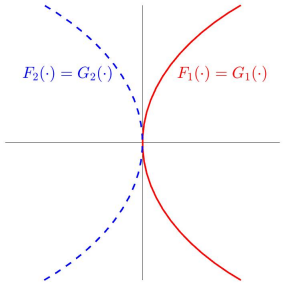
- Si $i \in \mathcal{F}(x)$ (resp. $i \in \mathcal{G}(x)$), on a $F_i(x) < G_i(x)$ (resp. $F_i(x) > G_i(x)$), donc, par continuité de F et G en x, on a aussi $F_i(x_k) < G_i(x_k)$ (resp. $F_i(x_k) > G_i(x_k)$). Par conséquent, le cas (4.7e)+(4.7h) (resp. cas (4.7g)+(4.7j)) s'applique et nous avons $J_{i,:} = F'_i(x)$ (resp. $J_{i,:} = G'_i(x)$). Donc $J_{i,:} = \tilde{J}_{i,:}$ par (4.7a).
- Si $i \in \mathcal{E}^{\neq}(x)$ et $\tilde{J}_{i,:} = F'_i(x)$ (resp. $\tilde{J}_{i,:} = G'_i(x)$), on a $F_i(x) = G_i(x)$, donc (4.7b) (resp. (4.7c)) et (4.7k) donnent $F(x_k) < G(x_k)$ (resp. $F(x_k) > G(x_k)$) pour k suffisamment grand. Par conséquent, le cas (4.7e)+(4.7h) (resp. cas (4.7g)+(4.7j)) s'applique et nous avons $J_{i,:} = F'_i(x)$ (resp. $J_{i,:} = G'_i(x)$). Donc $J_{i,:} = \tilde{J}_{i,:}$ dans ce cas également.
- Enfin, si $i \in \mathcal{E}^{=}(x)$, $\tilde{J}_{i,:} = F'_i(x) = G'_i(x)$, par (4.7a). Regardons maintenant la valeur de $J_{i,:}$ en considérant les trois cas possibles (4.7e)-(4.7g).
  - Si (4.7e) (resp. (4.7g)) se produit, on a $H'_i(x_k) = F'_i(x_k)$ (resp. $H'_i(x_k) = G'_i(x_k)$) et on obtient à la limite $J_{i,:} = F'_i(x)$ (resp. $J_{i,:} = G'_i(x)$). Donc $J_{i,:} = \tilde{J}_{i,:}$ dans ces cas.
  - Si (4.7f) se produit, on a $F'_i(x_k) = G'_i(x_k)$ d'après le lemme 3.2.1, car $F_i(x_k) = G_i(x_k)$ et $x_k \in \mathcal{D}_H$. À la limite, on obtient $J_{i,:} = F'_i(x) = G'_i(x)$. Donc $J_{i,:} = \tilde{J}_{i,:}$ dans ce cas également.

□

Contre-exemple 4.3.2 (pas d'égalité dans (4.6))**.** Considérons le cas où $n = 2$, $F(x) \equiv x$

et $G(x) \equiv (x_1^2 + x_2^2 - x_1, x_1^2 + x_2^2 + 2x_1 + x_2)$:

$$\begin{aligned} H(x) &= \min\left(x, \begin{pmatrix} x_1^2 + x_2^2 - x_1 \\ x_1^2 + x_2^2 + 2x_1 + x_2 \end{pmatrix}\right) \\ (\mathcal{L}_0 H)(x) &= \min\left(x, \begin{pmatrix} -1 & 0 \\ 2 & 1 \end{pmatrix} x\right). \end{aligned}$$



On a $\mathcal{F}(0) = \mathcal{G}(0) = \varnothing$ et $\mathcal{E}(0) = \{1, 2\}$, et les deux matrices impliquées A et B sont

$$A = I \text{ et } B = \begin{pmatrix} -1 & 0 \\ 2 & 1 \end{pmatrix}, \qquad \text{ainsi } V := (B - A)^\mathsf{T} = \begin{pmatrix} -2 & 2 \\ 0 & 0 \end{pmatrix}.$$

Puisque $\mathrm{rank}(V) = 1 < 2$, il y a moins de 2^2 éléments dans $\partial_B(\mathcal{L}_0 H)(0)$. En fait, il n'y a que deux vecteurs de signes $s \in \{\pm 1\}^2$ tels que $s \cdot V^\mathsf{T} d > 0$ est réalisable pour d, car on doit avoir $s_1 = -s_2$, à savoir $s = (1, -1)$ et $s = (-1, 1)$. De cette observation, on déduit que

$$\partial_B(\mathcal{L}_0 H)(0) = \left\{ \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \right\}.$$

Les jacobiennes dans $\partial_B(\mathcal{L}_0 H)(0)$ sont aussi dans $\partial_B H(0)$ (proposition 4.3.1), mais ce dernier contient également une autre jacobienne. En effet, prenons des points de la forme $x^t = (0, t)$, avec $t \downarrow 0$ ou $t \uparrow 0$. Alors $F_1(x^t) < G_1(x^t)$ et $F_2(x^t) < G_2(x^t)$ pour tout $t \neq 0$. Par conséquent, $H'(x^t) = I$ et on a la jacobienne $I \in \partial_B H(0)$, qui n'est pas dans $\partial_B(\mathcal{L}_0 H)(0)$. Ce contre-exemple montre aussi que, bien que $|\partial_B(\mathcal{L}_x H)_{\mathcal{E}^{\neq}(x)}(x)|$ soit pair, $|\partial_B H_{\mathcal{E}^{\neq}(x)}(x)|$ peut être impair. □

Utilisons la notation : $(G'(x) - F'(x))_{\mathcal{E}^{\neq}(x)} = V^\mathsf{T}$, $v_i = \nabla G_i(x) - \nabla F_i(x)$ et $\mathcal{V}_i := \{y \in \mathbb{R}^n : F_i(x) = G_i(x)\}$ pour $i \in \mathcal{E}^{\neq}(x)$. Lorsque V est surjective, le B-différentiel de la linéarisation est complet, ce qui signifie que l'égalité est vérifiée dans (4.6). Il y a en fait une équivalence entre ces propriétés. Pour avoir l'égalité mais pas nécessairement la complétude, on doit affaiblir l'hypothèse de surjectivité.

Proposition 4.3.3 (CNS pour avoir l'égalité dans (4.6)). *Supposons que F et G sont $\mathcal{C}^1$ en x. Alors, on a les relations suivantes équivalentes : $(i) \Leftrightarrow (ii) \Leftrightarrow (iii) \Rightarrow (iv) \Leftrightarrow (v)$.*

- (i) *pour tout $\alpha \in \mathbb{R}^{|\mathcal{E}^{\neq}(x)|} \setminus \{0\}$ et $s \in \{\pm 1\}^{|\mathcal{E}^{\neq}(x)|}$ tels que $V\alpha = 0$ et $s \cdot \alpha \geqslant 0$, il existe un voisinage $\mathcal{U}$ de x tel que*

$$\{x' \in \mathcal{U} : s \cdot [G(x') - F(x')]_{\mathcal{E}^{\neq}(x)} > 0\} = \varnothing, \tag{4.8}$$

- (ii) *même condition que (i) sauf qu'on utilise*

$$\{x' \in \mathcal{U} : s_I \cdot [G(x') - F(x')]_I > 0\} = \varnothing, \tag{4.9}$$

où $I := \{i \in \mathcal{E}^{\neq}(x) : \alpha_i \neq 0\}$.

(iii) *l'égalité est vérifiée dans* (4.6)*, c'est-à-dire* $\partial_B(\mathcal{L}_x H)(x) = \partial_B H(x)$.

(iv) *pour tout* $i_0 \in \mathcal{E}^{\neq}(x)$*, tel que*

$$v_{i_0} = \sum_{i \in I_0} \alpha_i v_i, \quad \textit{pour certains } I_0 \subseteq \mathcal{E}^{\neq}(x) \setminus \{i_0\} \textit{ et } \alpha_i \in \mathbb{R}^*, \tag{4.10a}$$

il existe un voisinage $\mathcal{U}$ *de* x *tel que*

$$\bigcap_{i \in I_0} (\mathcal{V}_i \cap \mathcal{U}) \quad \subseteq \quad \mathcal{V}_{i_0}, \tag{4.10b}$$

(v) *pour tout* $i_0 \in \mathcal{E}^{\neq}(x)$*, tel que* (4.10a) *est vérifié avec des vecteurs linéairement indépendants* $\{v_i : i \in I_0\}$*, il existe un voisinage* $\mathcal{U}$ *de* x *tel que* (4.10b) *est vérifié.*

Preuve. $[(i) \Rightarrow (ii)]$ Supposons que (i) soit vérifié et que $\alpha_i = 0$ pour un certain $i \in \mathcal{E}^{\neq}(x)$. Alors, (4.8) est satisfait pour $s_i = \pm 1$, de sorte que pour $\mathcal{U}$ suffisamment petit

$$\{x' \in \mathcal{U} : s_j[G_j(x') - F_j(x')] > 0 \text{ pour tout } j \neq i \text{ et } G_i(x') - F_i(x') > 0\} = \varnothing, \tag{4.11a}$$

$$\{x' \in \mathcal{U} : s_j[G_j(x') - F_j(x')] > 0 \text{ pour tout } j \neq i \text{ et } G_i(x') - F_i(x') < 0\} = \varnothing. \tag{4.11b}$$

Nous affirmons que, pour $\mathcal{U}$ suffisamment petit, on a aussi

$$\{x' \in \mathcal{U} : s_j[G_j(x') - F_j(x')] > 0 \text{ pour tout } j \neq i \text{ et } G_i(x') - F_i(x') = 0\} = \varnothing. \tag{4.11c}$$

En effet, si ce n'était pas le cas, on pourrait trouver un x' dans l'ensemble du membre de gauche de (4.11c), avec $\mathcal{U}$ suffisamment petit pour avoir $F_i'(x') \neq G_i'(x')$ (rappelons que $i \in \mathcal{E}^{\neq}(x)$). Mais alors, $x_t := x' + t[G_i'(x') - F_i'(x')]$ avec $t > 0$ suffisamment petit serait tel que $s_j[G_j(x_t) - F_j(x_t)] > 0$ pour tout $j \neq i$ et

$$\begin{aligned} G_i(x_t) - F_i(x_t) &= G_i(x') - F_i(x') + t[G_i'(x') - F_i'(x')]^2 + o(t) \\ &= t[G_i'(x') - F_i'(x')]^2 + o(t) \qquad [G_i(x') = F_i(x')] \\ &> 0 \qquad [G_i'(x') \neq F_i'(x')]. \end{aligned}$$

Alors, x_t serait dans l'ensemble du membre de gauche de (4.11a), ce qui contredirait sa vacuité. En combinant (4.11a)-(4.11c), on obtient

$$\{x' \in \mathcal{U} : s_j[G_j(x') - F_j(x')] > 0 \text{ pour tout } j \neq i\} = \varnothing.$$

On peut poursuivre de la même manière avec les autres indices donnant une composante nulle à α. Ceci montre que $(i) \Rightarrow (ii)$.

$[(ii) \Rightarrow (i)]$ Cette implication est claire puisque l'ensemble dans (4.8) est plus grand que celui dans (4.9).

$[(i) \Rightarrow (iii)]$ Nous procédons par contraposition, en supposant que (iii) n'est pas vérifié. Alors, par l'inclusion (4.6), il existe $J \in \partial_B H(x)$ qui n'est pas dans $\partial_B(\mathcal{L}_x H)(x)$. En utilisant $J \in \partial_B H(x)$, il existe une suite $\{x_k\} \subseteq \mathcal{D}_H$ convergeant vers x telle que $H'(x_k) \to J$.

L'examen de la suite $\{x_k\}$ nous permet de déterminer J. Puisque $\{x_k\} \subseteq \mathcal{D}_H$ et $F_i'(x) \neq G_i'(x)$ pour $i \in \mathcal{E}^{\neq}(x)$, on doit avoir $F_i(x_k) \neq G_i(x_k)$ pour $i \in \mathcal{E}^{\neq}(x)$ et k suffisamment grand (lemme 3.2.1). Alors, on peut trouver une partition (I_-, I_+) de $\mathcal{E}^{\neq}(x)$ et une sous-suite d'indices k telle que

$$\begin{cases} G_i(x_k) < F_i(x_k) & \text{pour } i \in I_-, \\ G_i(x_k) > F_i(x_k) & \text{pour } i \in I_+. \end{cases} \tag{4.12a}$$

Cela implique que pour k suffisamment grand dans la sous-suite sélectionnée (pour les indices dans $\mathcal{E}^{=}(x)$, on utilise à nouveau le lemme 3.2.1 [1]) :

$$H_i'(x_k) = \begin{cases} F_i'(x_k) & \text{si } i \in \mathcal{F}(x), \\ F_i'(x_k) \text{ ou } G_i'(x_k) & \text{si } i \in \mathcal{E}^{=}(x), \\ G_i'(x_k) & \text{si } i \in I_- \subseteq \mathcal{E}^{\neq}(x), \\ F_i'(x_k) & \text{si } i \in I_+ \subseteq \mathcal{E}^{\neq}(x), \\ G_i'(x_k) & \text{si } i \in \mathcal{G}(x) \end{cases}$$

et donc, la jacobienne $J \in \partial_B H(x)$ mentionnée ci-dessus a sa i-ème ligne donnée par

$$J_{i,:} = \begin{cases} F_i'(x) & \text{si } i \in \mathcal{F}(x), \\ F_i'(x) = G_i'(x) & \text{si } i \in \mathcal{E}^{=}(x), \\ G_i'(x) & \text{si } i \in I_- \subseteq \mathcal{E}^{\neq}(x), \\ F_i'(x) & \text{si } i \in I_+ \subseteq \mathcal{E}^{\neq}(x), \\ G_i'(x) & \text{si } i \in \mathcal{G}(x). \end{cases} \tag{4.12b}$$

Maintenant, définissons $s \in \{\pm 1\}^{|\mathcal{E}^{\neq}(x)|}$ par

$$s_i = \begin{cases} -1 & \text{si } i \in I_-, \\ +1 & \text{si } i \in I_+. \end{cases} \tag{4.12c}$$

Puisque la jacobienne J donnée par (4.12b) n'est pas dans $\partial_B(\mathcal{L}_x H)(x)$, on sait que

$$\nexists\, d \in \mathbb{R}^n : \quad s \bullet V^{\mathsf{T}} d > 0.$$

Maintenant, l'alternative de Gordan implique qu'on peut trouver $\alpha \in \mathbb{R}^{|\mathcal{E}^{\neq}(x)|} \setminus \{0\}$ tel que

$$V\alpha = 0 \qquad \text{et} \qquad s \bullet \alpha \geqslant 0.$$

On voit que cette paire (α, s) satisfait les propriétés dans la prémisse de (i), mais, par (4.12a) et (4.12c), il existe une suite $\{x_k\} \to x$ telle que

$$s \bullet [G(x_k) - F(x_k)]_{\mathcal{E}^{\neq}(x)} > 0,$$

ce qui contredit (4.8). Par conséquent, (i) n'est pas vérifié, comme attendu.

1. Les valeurs possibles de $H'(x_k)$ sont claires si $F_i(x_k) \neq G_i(x_k)$. Si $F_i(x_k) = G_i(x_k)$, on doit avoir $F_i'(x_k) = G_i'(x_k)$ et $H_i'(x_k) = F_i'(x_k) = G_i'(x_k)$ par la différentiabilité de H_i en x_k (lemme 3.2.1).

$[(iii) \Rightarrow (i)]$ Nous procédons également par contraposition, en supposant que (i) n'est pas vérifié. Alors, il existe une paire $(\alpha, s) \in (\mathbb{R}^{|\mathcal{E}^{\neq}(x)|} \setminus \{0\}) \times \{\pm 1\}^{|\mathcal{E}^{\neq}(x)|}$ telle que

$$V\alpha = 0 \qquad \text{et} \qquad s \cdot \alpha \geqslant 0, \tag{4.12d}$$

mais aucun voisinage $\mathcal{U}$ de x tel que (4.8) soit vérifié. On déduit de ce dernier fait qu'il existe une suite $\{x_k\} \to x$ telle que

$$s \cdot [G(x_k) - F(x_k)]_{\mathcal{E}^{\neq}(x)} > 0, \tag{4.12e}$$

Par (4.12d) et l'alternative de Gordan,

$$\nexists\, d \in \mathbb{R}^n : \quad s \cdot V^{\mathsf{T}} d > 0.$$

Alors, cela implique qu'il n'existe pas de $J \in \partial_B(\mathcal{L}_x H)(x)$ satisfaisant [2]

$$J_{i,:} := \begin{cases} F_i'(x) & \text{si } s_i = +1, \\ G_i'(x) & \text{si } s_i = -1. \end{cases} \tag{4.12f}$$

Montrons maintenant qu'il existe un J dans $\partial_B H(x)$ qui satisfait (4.12f), ce qui montrera que (i) n'est pas vérifié, comme attendu. Par (4.12e), la continuité de F et G en x et le fait que $x_k \to x$, on a (notons que rien n'est dit sur les indices dans $\mathcal{E}^{=}(x)$ ou la différentiabilité de H en x_k)

$$\begin{cases} F_i(x_k) < G_i(x_k) & \text{si } i \in \mathcal{F}(x), \\ F_i(x_k) < G_i(x_k) & \text{si } i \in \mathcal{E}^{\neq}(x) \text{ et } s_i = +1, \\ F_i(x_k) > G_i(x_k) & \text{si } i \in \mathcal{E}^{\neq}(x) \text{ et } s_i = -1, \\ F_i(x_k) > G_i(x_k) & \text{si } i \in \mathcal{G}(x). \end{cases} \tag{4.12g}$$

Nous affirmons qu'on peut légèrement perturber la suite $\{x_k\}$ pour obtenir $\{x_k'\} \subseteq \mathcal{D}_H$ convergeant vers x et satisfaisant (donc, la différentiabilité de H en x_k' est maintenant garantie)

$$\begin{cases} F_i(x_k') < G_i(x_k') & \text{si } i \in \mathcal{F}(x), \\ F_i(x_k') < G_i(x_k') & \text{si } i \in \mathcal{E}^{\neq}(x) \text{ et } s_i = +1, \\ F_i(x_k') > G_i(x_k') & \text{si } i \in \mathcal{E}^{\neq}(x) \text{ et } s_i = -1, \\ F_i(x_k') > G_i(x_k') & \text{si } i \in \mathcal{G}(x). \end{cases}$$

En effet, pour prouver cela, il suffit de préciser ce qui est fait pour les indices dans $\mathcal{E}^{=}(x)$ pour s'assurer que $\{x_k'\}$ est dans $\mathcal{D}_H$ et converge vers x, puisque les inégalités dans (4.12g) sont préservées par une petite perturbation grâce à la continuité de F et G en x_k et garantissent que la composante correspondante de H est différentiable en x_k'. Prenons un

2. Montrons cette affirmation par contraposition. Si un tel $J \in \partial_B(\mathcal{L}_x H)(x)$ existait, on aurait par l'affirmation précédente

$$\nexists\, d \in \mathbb{R}^n : \quad \begin{cases} [G_i'(x) - F_i'(x)]d > 0 & \text{si } J_{i,:} = F_i'(x), \\ [G_i'(x) - F_i'(x)]d < 0 & \text{si } J_{i,:} = G_i'(x). \end{cases}$$

Ceci serait en contradiction avec $J \in \partial_B(\mathcal{L}_x H)(x)$.

premier indice $i \in \mathcal{E}^{=}(x)$. Si $F_i(x') = G_i(x')$ pour x' près de x_k, alors H_i est différentiable en x_k et aucune perturbation de x_k n'est nécessaire. Sinon, il existe un x' arbitrairement proche de x_k avec $F_i(x') \neq G_i(x')$; H_i est différentiable en ce x'. En procédant ainsi pour les autres indices possibles dans $\mathcal{E}^{=}(x)$, tout en préservant les inégalités strictes obtenues jusqu'à présent, nous obtenons une suite $\{x'_k\} \subseteq \mathcal{D}_H$. Dans cette procédure, x'_k peut être pris arbitrairement proche de x_k afin de garantir que $\{x'_k\} \to x$. Puisque

$$\begin{cases} H'_i(x'_k) = F'_i(x'_k) & \text{si } i \in \mathcal{E}^{\neq}(x) \text{ et } s_i = +1, \\ H'_i(x'_k) = G'_i(x'_k) & \text{si } i \in \mathcal{E}^{\neq}(x) \text{ et } s_i = -1, \end{cases}$$

nous voyons que $H'(x'_k)$ converge vers une jacobienne $J \in \partial_B H(x)$ satisfaisant (4.12f), comme attendu.

$[(iv) \Rightarrow (v)]$ C'est clair, puisque (4.9) doit être vérifié pour moins de triplets (i_0, I_0, α) satisfaisant (4.8).

$[(iii) \Rightarrow (v)]$ Nous procédons par contraposition, en supposant que (v) n'est pas vérifié. Alors, il existe un indice $i_0 \in \mathcal{E}^{\neq}(x)$, un ensemble d'indices $I_0 \subseteq \mathcal{E}^{\neq}(x) \setminus \{i_0\}$ et $\alpha_i \in \mathbb{R}^*$, tels que (4.8) soit vérifié avec des vecteurs linéairement indépendants $\{v_i : i \in I_0\}$, mais il n'existe aucun voisinage $\mathcal{U}$ tel que (4.9) soit vérifié.

On déduit de ce dernier fait qu'il existe une suite $\{x_k\} \to x$ telle que $x_k \in \cap_{i \in I_0} \mathcal{V}_i$, mais $x_k \notin \mathcal{V}_{i_0}$ ou $F_{i_0}(x_k) \neq G_{i_0}(x_k)$. En extrayant une sous-suite si nécessaire, on peut supposer qu'on a ($F_{i_0}(x_k) > G_{i_0}(x_k)$ pour tout k) ou ($F_{i_0}(x_k) < G_{i_0}(x_k)$ pour tout k). Supposons que le premier cas se produit (on peut procéder de manière similaire dans le second cas). Ainsi, pour tout $k \to \infty$,

$$\begin{cases} F_i(x_k) = G_i(x_k), & \text{pour } i \in I_0, \\ F_{i_0}(x_k) > G_{i_0}(x_k). \end{cases} \tag{4.13a}$$

Par l'alternative de Gordan, (4.8) implique qu'on ne peut pas trouver une direction d telle que (le sens des inégalités dans (4.13b) ci-dessous est pris en fonction du fait que $F_{i_0}(x_k) < G_{i_0}(x_k)$ dans (4.13a) ; inverser les inégalités si $F_{i_0}(x_k) > G_{i_0}(x_k)$ est vérifié)

$$-v_{i_0}^{\mathsf{T}} d > 0 \qquad \text{et} \qquad \operatorname{sgn}(\alpha_i) v_i^{\mathsf{T}} d > 0, \quad \text{pour } i \in I_0. \tag{4.13b}$$

Cela implique qu'il n'existe pas de $J \in \partial_B(\mathcal{L}_x H))(x)$ satisfaisant

$$J_{i,:} := \begin{cases} G'_i(x) & \text{si } i = i_0, \\ F'_i(x) & \text{si } i \in I_0 \text{ et } \alpha_i > 0, \\ G'_i(x) & \text{si } i \in I_0 \text{ et } \alpha_i < 0. \end{cases} \tag{4.13c}$$

Nous montrons ensuite qu'il existe un J dans $\partial_B H(x)$ qui satisfait (4.13c), ce qui montrera que (iii) n'est pas vérifié, comme attendu.

Puisque les v_i, pour $i \in I_0$, sont linéairement indépendants, on peut trouver une direction $p \in \mathbb{R}^n$ telle que

$$v_i^{\mathsf{T}} p = \operatorname{sgn}(\alpha_i), \quad \text{pour } i \in I_0. \tag{4.13d}$$

Cette direction est utilisée pour définir une suite $\{x'_k\}$, qui est une petite perturbation de $\{x_k\}$, par

$$x'_k := x_k + t_k p + \sigma_k,$$

où les σ_k sont des vecteurs de perturbation (petits) dans $\mathbb{R}$ tels que $\sigma_k = o(t_k)$ et $x'_k \in \mathcal{D}_H$ (cette précaution est possible par le théorème de Rademacher et elle est utile ci-dessous pour les indices $i \in \mathcal{E}^{\neq}(x) \setminus (I_0 \cup \{i_0\})$ s'il y en a) et où $t_k = o(\|x_k - x\|)$ est pris positif et suffisamment petit pour garantir

$$F_{i_0}(x'_k) > G_{i_0}(x'_k) \tag{4.13e}$$

(ceci est possible par $F_{i_0}(x_k) > G_{i_0}(x_k)$ dans (4.13a) et par la continuité de $G_{i_0} - F_{i_0}$). Maintenant, pour $i \in [1:n]$, le théorème des accroissements finis, qui est vrai lorsque F_i est différentiable près de x, garantit que, pour k assez grand

$$\|F_i(x'_k) - F_i(x_k) - F'_i(x)(t_k p + \sigma_k)\| \leqslant \left(\sup_{z \text{ près de } x} \|F'_i(z) - F'_i(x)\| \right) \|t_k p + \sigma_k\|.$$

En procédant de même pour G_i et en utilisant la continuité de $G'_i - F'_i$ en x, on a lorsque $k \to \infty$:

$$\begin{aligned} F_i(x'_k) &= F_i(x_k) + F'_i(x)(t_k p) + o(t_k), \\ G_i(x'_k) &= G_i(x_k) + G'_i(x)(t_k p) + o(t_k). \end{aligned}$$

Pour $i \in I_0$, $F_i(x_k) = G_i(x_k)$ par (4.13a), donc, en utilisant (4.13d) :

$$G_i(x'_k) - F_i(x'_k) = t_k\, v_i^{\mathsf{T}} p + o(t_k) = \operatorname{sgn}(\alpha_i)\, t_k + o(t_k).$$

Par conséquent, pour $i \in I_0$ et k assez grand :

$$\begin{cases} F_i(x'_k) < G_i(x'_k) & \text{si } i \in I_0 \text{ et } \alpha_i > 0, \\ F_i(x'_k) > G_i(x'_k) & \text{si } i \in I_0 \text{ et } \alpha_i < 0. \end{cases} \tag{4.13f}$$

Nous déduisons de (4.13e), (4.13f) et de la différentiabilité de H en x'_k que

$$H'_i(x'_k) = \begin{cases} G'_i(x'_k) & \text{si } i = i_0, \\ F'_i(x'_k) & \text{si } i \in I_0 \text{ et } \alpha_i > 0, \\ G'_i(x'_k) & \text{si } i \in I_0 \text{ et } \alpha_i < 0. \end{cases}$$

À la limite lorsque $k \to \infty$, nous obtenons une jacobienne J dans $\partial_B H(x)$ satisfaisant (4.13c), comme annoncé.

[$(v) \Rightarrow (iv)$] Si v_{i_0} satisfait (4.8) pour certains vecteurs $\{v_i : i \in I_0\}$ qui ne sont pas linéairement indépendants, on peut aussi écrire $v_{i_0} = \sum_{i \in I'_0} \alpha'_i v_i$ avec $\alpha'_i \neq 0$, $I'_0 \subseteq I_0$ et des vecteurs linéairement indépendants $\{v_i : i \in I'_0\}$. Par (v),

$$\bigcap_{i \in I'_0} (\mathcal{V}_i \cap \mathcal{U}) \quad \subseteq \quad \mathcal{V}_{i_0}.$$

Puisque $I'_0 \subseteq I_0$, l'intersection a davantage de termes, est plus petite et (4.9) est valide. □

4.4 Différentiel de la fonction de mérite

Avant de discuter des propriétés de la fonction de mérite θ, montrons une adaptation du lemme 2.3.23, repris de [206, lemme 2.2, p. 356]. Ce lemme indique que, pour une fonction lipschitzienne $F : \mathbb{R}^n \to \mathbb{R}^m$,

$$F'(x;d) = Vd, \quad V \in \partial F(x).$$

Cette propriété, appliquée à la fonction minimum composante par composante H, peut être restreinte à $\partial_B H(x)$: on affaiblit le C-différentiel en le B-différentiel.

Proposition 4.4.1 (dérivée directionnelle et B-différentiel)**.** *Soient F, G deux fonctions $\mathcal{C}^1$ lipschitziennes de $\mathbb{R}^n$ dans $\mathbb{R}^n$. Pour tout x, $d \in \mathbb{R}^n$, on a*

$$H'(x;d) = J^{\mathsf{T}} d \qquad \text{pour un certain } J \in \partial_B H(x).$$

Preuve. Rappelons la formule de $H'(x;d)$

$$H'(x;d) = \begin{pmatrix} F_i'(x)d & \text{si } i \in \mathcal{F}(x) \\ G_i'(x)d & \text{si } i \in \mathcal{G}(x) \\ \min(F_i'(x)d, G_i'(x)d) & \text{si } i \in \mathcal{E}(x) \end{pmatrix}_{i \in [1:n]}.$$

Définissons maintenant les ensembles d'indices suivants

$$\begin{aligned} \mathcal{E}_{\mathcal{F}} &:= \{i \in \mathcal{E}(x) : F_i'(x)d < G_i'(x)d\}, \\ \mathcal{E}_{\mathcal{G}} &:= \{i \in \mathcal{E}(x) : F_i'(x)d > G_i'(x)d\}, \\ \mathcal{E}_{\mathcal{E}} &:= \{i \in \mathcal{E}(x) : F_i'(x)d = G_i'(x)d\}. \end{aligned}$$

Clairement, $H'(x;d) = Jd$, où $J_{\mathcal{E}_{\mathcal{F}},:} = F'(x)_{\mathcal{E}_{\mathcal{F}}}$, $J_{\mathcal{E}_{\mathcal{G}},:} = G'(x)_{\mathcal{E}_{\mathcal{G}}}$ et $J_{i,:} \in \{F_i'(x), G_i'(x)\}$ pour $i \in \mathcal{E}_{\mathcal{E}}$.

Considérons maintenant le (sous-)arrangement défini par les hyperplans H_i avec les indices $i \in \mathcal{E}_{\mathcal{E}}$. Soit $s' \in \mathcal{S}((G'(x) - F'(x))^{\mathsf{T}}_{\mathcal{E}_{\mathcal{E}},:}, 0)$ et d' une direction associée (voir chapitre 3), c'est-à-dire $s' = \operatorname{sgn}((G'(x) - F'(x))d')_{\mathcal{E}_{\mathcal{E}}}$. Pour une suite $\{t_k\}_k \downarrow 0$ et $\varepsilon > 0$ assez petit, considérons la suite $x_k := x + t_k(d + \varepsilon d')$. Montrons que : cette suite n'appartient à aucun des hyperplans H_i pour $i \in \mathcal{E}(x)$ pour k assez grand (t_k assez petit), et que la jacobienne correspondante appartient à $\partial_B H(x)$ et est égale à l'une des J possibles définies ci-dessus.

Par continuité, pour k assez grand, $F_{\mathcal{F}(x)}(x_k) < G_{\mathcal{F}(x)}(x_k)$, $F_{\mathcal{G}(x)}(x_k) > G_{\mathcal{G}(x)}(x_k)$. Ensuite, pour les indices dans $\mathcal{E}(x)$, utilisons les développements suivants

$$\begin{aligned} F_i(x_k) &= F_i(x) + t_k F_i'(x)d + \varepsilon t_k F_i'(x)d' + o(t_k), \\ G_i(x_k) &= G_i(x) + t_k G_i'(x)d + \varepsilon t_k G_i'(x)d' + o(t_k). \end{aligned}$$

Pour les indices dans $\mathcal{E}_{\mathcal{F}}$, puisque ε est assez petit, on a $F_i(x_k) < G_i(x)_k$, et de même $F_i(x_k) > G_i(x_k)$ pour les indices dans $\mathcal{E}_{\mathcal{G}}$. Pour ceux dans $\mathcal{E}_{\mathcal{E}}$,

$$H_i(x_k) = \min(F_i(x_k), G_i(x_k)) = \begin{cases} F_i(x_k) < G_i(x_k) & \text{si } s_i' = +1, \\ G_i(x_k) < F_i(x_k) & \text{si } s_i' = -1. \end{cases}$$

Ainsi, x_k n'appartient à aucun des hyperplans H_i et correspond à la jacobienne J' avec les lignes

$$J'_{i,:} = \begin{cases} F'_i(x) & i \in \mathcal{F}(x) \cup \mathcal{E}_{\mathcal{F}}, \\ G'_i(x) & i \in \mathcal{G}(x) \cup \mathcal{E}_{\mathcal{G}}, \\ F'_i(x) & i \in \mathcal{E}_{\mathcal{E}}, s'_i = +1, \\ G'_i(x) & i \in \mathcal{E}_{\mathcal{E}}, s'_i = -1. \end{cases}$$

Par construction, cette jacobienne appartient à $\partial_B H(x)$ et $J'd = H'(x;d)$. □

Dans la suite, nous considérerons également une fonction de mérite plus générale $\theta_\psi = \psi \circ H(x)$, avec ψ une fonction scalaire $\mathcal{C}^1$ généralisant la norme 2 au carré. Lorsque H est différentiable en x, on a clairement

$$\nabla\theta(x) = \nabla H(x) \times H(x) = \sum_{i=1}^{n} H_i(x)\nabla H_i(x)$$

$$\nabla\theta_\psi(x) = \nabla H(x) \times \nabla\psi(H(x)) = \sum_{i=1}^{n} \nabla\psi(H(x))_i \nabla H_i(x)$$

Pour le C-différentiel, Clarke [51, prop. 2.6.6, pp. 72-73] a montré que pour une fonction différentiable ψ (donc en particulier $||\cdot||^2/2$), on a

$$\partial\theta_\psi(x) = \partial H(x)^\mathsf{T} \nabla\psi(H(x))$$

où le différentiel est vue comme un vecteur colonne (et même des résultats plus forts). Bien que l'on puisse justifier la preuve pour le B-différentiel puis prendre l'enveloppe convexe, la preuve de Clarke développe un raisonnement plus avancé. De plus, on ne peut pas simplement utiliser la relation de Clarke et prendre les points extrémaux (voir les commentaires autour de la proposition 3.4.14) ; nous rappelons ces propriétés.

Proposition 4.4.2 (extrémalité pour H)**.** *On a*

$$\partial_B H(x) = \mathrm{ext}(\partial_C H(x)) = \mathrm{ext}(\mathrm{conv}(\partial_B H(x))).$$ □

Observons que cela vaut également pour des fonctions non linéaires F et G. Notons que la relation $C \supseteq \mathrm{ext}(\mathrm{conv}(C))$ est toujours vraie, mais la réciproque ne l'est pas en général (par exemple une boule), car cela signifie que tout point de C est extrémal.

Remarque 4.4.3. L'observation précédente ne peut pas être utilisée, pour le calcul de $\partial_B\theta(x)$, pour écrire (en omettant la dépendance en (x))

$$\partial_B\theta = \mathrm{ext}(\partial_C\theta) = \mathrm{ext}(\mathrm{conv}(\partial_B\theta))$$

Ce qui empêche d'utiliser simplement ce résultat est que, en général,

$$\mathrm{ext}(S \times v) \neq \mathrm{ext}(S) \times v$$

Cependant, la proposition concernée montre que l'égalité est vraie car $S = \partial_B H^\mathsf{T}$ est suffisamment spécifique. □

Contre-exemple 4.4.4 ($\mathrm{ext}(S \times v) \neq \mathrm{ext}(S) \times v$)**.** Par exemple, considérons

$$S = \mathrm{conv}\left\{ \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} -1 & -1 \\ -1 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} \right\}, \qquad v = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Puisque $\mathrm{ext}(S)$ est composé des cinq matrices données, on a

$$\mathrm{ext}(S) \times v = \left\{ \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} -2 \\ -2 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \end{bmatrix}, \begin{bmatrix} -2 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\},$$
$$\mathrm{ext}(S \times v) = \left\{ \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} -2 \\ -2 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \end{bmatrix}, \begin{bmatrix} -2 \\ 2 \end{bmatrix} \right\}.$$

□

Revenant au B-différentiel de θ, il semble naturel de se demander si une règle de chaîne similaire est valable ou non, puisque le différentiel de Clarke a été introduit en partie pour bénéficier d'une telle propriété. La preuve que nous proposons est plutôt différente de celle de Clarke, reposant sur des propriétés spécifiques de la fonction minimum H et de son différentiel de Bouligand. Contrairement à la C-fonction de Fischer, qui conduit à une fonction de mérite plus différentiable, le résultat n'est pas aussi simple [87], mais on peut retrouver la forme attendue $(\partial_B H)^\mathsf{T} \times H$. Tout d'abord, rappelons une propriété simple.

Proposition 4.4.5 (un seul minimum)**.** *Considérons $\theta : y \mapsto (\min(f(y), g(y)))^2/2$, et x tel que les deux termes du min sont égaux : $f(x) = g(x)$. Alors θ est différentiable en x si et seulement si $\theta(x) = 0$ ou $f'(x) = g'(x)$.* □

Ceci ne peut pas être généralisé à des fonctions à minima multiples simultanément.

Contre-exemple 4.4.6 (minima multiples)**.** Soient $H_1(x, y) = 1 - |x + y|$ et $H_2(x, y) = -1 - |x + y|$. Les fonctions H_1 et H_2 ne sont pas différentiables lorsque $x + y = 0$ mais $\theta(x, y) = (H_1^2 + H_2^2)/2 = 1 + (x + y)^2$. □

Ce phénomène curieux s'explique par la proposition suivante.

Proposition 4.4.7 (minima multiples)**.** *Considérons la fonction suivante*

$$d \in \mathbb{R}^n \mapsto \sum_{i=1}^{q} c_i \min(u_i^\mathsf{T} d, w_i^\mathsf{T} d),$$

où aucun des vecteurs $v_i = (w_i - u_i) \in \mathbb{R}^n$ pour $i \in [1 : q]$ n'est colinéaire à un autre. Si la fonction est linéaire, alors tous les c_i sont nuls.

Preuve. Comme dans [255], récrivons d'abord la fonction sous la forme :

$$\sum_{i=1}^{q} c_i u_i^\mathsf{T} d + \sum_{i=1}^{q} c_i \min(0, v_i^\mathsf{T} d) = \sum_{i=1}^{q} c_i u_i^\mathsf{T} d + \mathcal{L}(d) = \sum_{i=1}^{q} c_i u_i^\mathsf{T} d + l^\mathsf{T} d, \quad l \in \mathbb{R}^n$$

qui est linéaire si et seulement si le terme $\mathcal{L}$ l'est, avec $\mathcal{L}(d) = l^\mathsf{T} d$. Nous nous concentrons donc sur la somme des minima $\mathcal{L}$, $\mathcal{L}(d) := \sum_{i=1}^{q} c_i \min(0, v_i^\mathsf{T} d)$. Considérons maintenant l'arrangement d'hyperplans défini par les $v_i^\perp$ pour $i \in [1:q]$. Soit $\mathcal{S}_q \subseteq \{\pm 1\}^q$ l'ensemble des vecteurs de signes de cet arrangement. Pour chaque vecteur de signes $s \in \mathcal{S}_q$, il existe une direction $d_s \in \mathbb{R}^n$ telle que $\mathrm{Diag}(s) V^\mathsf{T} d_s > 0$, où V est la matrice $[v_1 \ \dots \ v_q]$.

Considérons un vecteur de signes donné $s^0 \in \mathcal{S}_q$ et la direction associée $d^0 \in \mathbb{R}^n$. Séparons les indices : $I^+ = \{i \in \mathcal{E}^{\neq}(x) : v_i^\mathsf{T} d^0 > 0\}$ et $I^- = \{i \in \mathcal{E}^{\neq}(x) : v_i^\mathsf{T} d^0 < 0\}$ de sorte qu'en d^0, $\mathcal{L}$ a l'expression suivante :

$$\mathcal{L}(d^0) = \sum_{i \in I^+} c_i \times 0 + \sum_{i \in I^-} c_i v_i^\mathsf{T} d^0 = l^\mathsf{T} d^0,$$

ainsi la fonction prend la forme $d \mapsto \sum_{i \in I^-} c_i v_i^\mathsf{T} d$. Comme ceci est vrai pour tout d dans la région de d^0, c'est vrai pour une petite boule autour de d^0. En identifiant l'expression de la fonction linéaire, on obtient $l = \sum_{i \in I^-} c_i v_i$. Examinons maintenant ce qui se passe dans une autre région (voisine).

Comme nous avons supposé qu'aucun vecteur n'est colinéaire à un autre, nous savons que l'ensemble des vecteurs de signes est connexe (proposition 3.4.5). Par symétrie (proposition 3.4.1), il existe un chemin $s^0, \dots, s^q$ avec $s^q = -s^0$ tel que deux vecteurs de signes consécutifs diffèrent exactement sur un indice : le long du chemin, le signe de chaque indice change exactement une fois. Examinons de plus près s^0 et s^1.

Sans perte de généralité, supposons que l'indice j modifié appartient à I^+ (le cas $j \in I^-$ est très similaire). Pour une direction d^1 associée à s^1, nous avons :

$$\begin{array}{rcl} d^0 & \to & \sum_{i \in I^+ \setminus \{j\}} c_i \times 0 \quad + c_j \times 0 \quad + \sum_{i \in I^-} c_i v_i^\mathsf{T} d^0 \\ d^1 & \to & \sum_{i \in I^+ \setminus \{j\}} c_i \times 0 \quad + c_j w_j^\mathsf{T} d^1 \quad + \sum_{i \in I^-} c_i v_i^\mathsf{T} d^1 \end{array}$$

En utilisant l'argument d'identification dans les deux régions, nous avons l'égalité suivante :

$$\sum_{i \in I^-} c_i v_i = l = \sum_{i \in I^-} c_i v_i + c_j w_j$$

Comme les v_i ne sont pas colinéaires, ils sont non nuls, donc $c_j = 0$. En répétant ce raisonnement pour chaque indice le long du chemin entre s^0 et $s^q = -s^0$, nous obtenons que tous les coefficients c_i sont nuls. □

Ce lemme illustre qu'une somme de minima de fonctions linéaires non colinéaires ne peut être linéaire que si elle est nulle, en identifiant une expression différente pour la fonction (supposée) linéaire $\mathcal{L}$. Cependant, il ne peut pas être utilisé directement dans le raisonnement principal qui suit, car comme dans le contre-exemple 4.4.6, la fonction peut être non nulle alors que les vecteurs impliqués sont colinéaires. Avant de passer à la proposition principale, détaillons un dernier point.

Lemme 4.4.8 (Agrégation de minima similaires)**.** Soient $v \in \mathbb{R}^n$, I un ensemble d'indices, $\alpha \in \mathbb{R}_*^I$ et $c \in \mathbb{R}^I$. On a la relation suivante :

$$\sum_{i \in I, \alpha_i > 0} c_i \min(0, \alpha_i v^\mathsf{T} x) + \sum_{i \in I, \alpha_i < 0} c_i \min(0, \alpha_i v^\mathsf{T} x) = -\mathcal{C}^- v^\mathsf{T} x + (\mathcal{C}^+ + \mathcal{C}^-) \min(0, v^\mathsf{T} x)$$

où $\mathcal{C}^+ = \sum_{i\in I,\alpha_i>0} c_i\alpha_i$ et $\mathcal{C}^- = \sum_{i\in I,\alpha_i<0} c_i|\alpha_i|$.

Preuve.

$$\begin{aligned}
&\sum_{i\in I,\alpha_i>0} c_i \min(0,\alpha_i v^\mathsf{T} x) + \sum_{i\in I,\alpha_i<0} c_i \min(0,\alpha_i v^\mathsf{T} x)\\
&= \sum_{i\in I,\alpha_i>0} c_i\alpha_i \min(0, v^\mathsf{T} x) - \sum_{i\in I,\alpha_i<0} c_i \max(0,|\alpha_i| v^\mathsf{T} x)\\
&= \mathcal{C}^+ \min(0,v^\mathsf{T} x) - \sum_{i\in I,\alpha_i<0} c_i|\alpha_i| \max(0,v^\mathsf{T} x)\\
&= \mathcal{C}^+ \min(0,v^\mathsf{T} x) - \mathcal{C}^- \max(0,v^\mathsf{T} x) = \mathcal{C}^+ \min(0,v^\mathsf{T} x) - \mathcal{C}^-(v^\mathsf{T} x + \max(-v^\mathsf{T} x,0))\\
&= \mathcal{C}^+ \min(0,v^\mathsf{T} x) - \mathcal{C}^- v^\mathsf{T} x + \mathcal{C}^- \min(0,v^\mathsf{T} x) = -\mathcal{C}^- v^\mathsf{T} x + (\mathcal{C}^+ + \mathcal{C}^-)\min(0,v^\mathsf{T} x). \qquad \square
\end{aligned}$$

L'utilité de ce calcul est la suivante : nous avons une propriété pour les sommes de minima avec des vecteurs qui ne sont pas deux à deux colinéaires. Dans le cas général (comme vu dans le contre-exemple 4.4.6), si certains vecteurs sont colinéaires, alors les indices correspondants peuvent être regroupés à l'aide du lemme 4.4.8, puis nous utilisons la proposition 4.4.7 pour ces indices regroupés ou pour les vecteurs qui ne sont colinéaires à aucun autre.

Proposition 4.4.9 (B-différentiel de $\boldsymbol{\theta}$ dans le cas linéaire)**.** *On a la règle de chaîne suivante :*

$$\widetilde{\partial}_B(\psi\circ H)(x) := \{J^\mathsf{T}\nabla\psi(H(x)); J\in\partial_B H(x)\} = \partial_B(\psi\circ H)(x).$$

En particulier pour $\psi = ||\cdot||^2/2$, on retrouve la fonction de mérite usuelle.

Preuve. [$\subseteq$] Considérons une matrice $J\in\partial_B H(x)$, associée à une suite $\{x_k\}_k \to x$ telle que $\nabla H(x_k)\to J^\mathsf{T}$. Comme H est différentiable aux points de la suite, en utilisant $\nabla\theta = \nabla H\times\nabla\psi(H)$, on obtient :

$$\nabla\theta(x_k) = \nabla H(x_k)\times\nabla\psi(H(x_k)) \to J^\mathsf{T}\nabla\psi(H(x))$$

Ce qui indique que cette limite est dans $\partial_B\theta_\psi(x) = \partial_B(\psi\circ H)(x)$.

[$\supseteq$] Considérons un élément $v\in\partial_B\theta_\psi(x)$, et la suite associée $\{x_k\}_k$ telle que θ est F-différentiable en x_k pour tout k, $x_k\to x$ et $\nabla\theta(x_k)\to v$. Nous voulons montrer que v peut s'écrire comme $J^\mathsf{T}\nabla\psi(H(x))$ pour une certaine $J\in\partial_B H(x)$, c'est-à-dire peut s'exprimer comme une limite de $\nabla H(x_k)\times\nabla\psi(H(x_k))$.

Supposons d'abord que H est différentiable le long de $\{x_k\}_k$: $\nabla\theta(x_k) = \nabla H(x_k)\times\nabla\psi(H(x_k))\to v$ est bien définie. Comme $\nabla H(x_k)$ est constante par morceaux et prend un nombre fini de valeurs possibles (proposition 3.2.2), quitte à extraire une sous-suite, on peut supposer qu'elle a une valeur constante fixe. Ceci montre que v est une limite de la forme désirée.

Considérons maintenant que le long de la suite $\{x_k\}_k$, H a certaines composantes non différentiables :

- si $i \in \mathcal{F}(x)$ ou $i \in \mathcal{G}(x)$, pour k assez grand (donc x_k assez proche de x), on a $F_i(x_k) \neq G_i(x_k)$ donc la composante est différentiable,
- si $i \in \mathcal{E}^{=}(x)$, les fonctions affines sont identiques, donc le minimum disparaît et la composante est toujours différentiable,
- si $i \in \mathcal{E}^{\neq}(x)$; notons $I^k \subseteq \mathcal{E}^{\neq}$ les composantes non différentiables en x_k.

Faisons d'abord une remarque sur I^k. Comme $\mathcal{E}^{\neq}(x)$ est fixe (ne dépend que de x et des données), et $I^k \subseteq \mathcal{E}^{\neq}(x)$ par continuité, il n'y a qu'un nombre fini de I^k possibles (inférieur à $2^{|\mathcal{E}^{\neq}(x)|}$). Ainsi, en extrayant une sous-suite, toujours notée $\{x_k\}_k$, on peut supposer que les indices I^k ne changent pas et sont notés I.

Pour simplifier, nous montrons d'abord le résultat pour $\psi = ||\cdot||^2/2$ avant de considérer le cas général. Comme θ est F-différentiable en x_k par hypothèse (ce qui reste vrai le long de toute sous-suite), nous décomposons :

$$\theta = \frac{1}{2}\sum_{i\in I^c} H_i^2 + \frac{1}{2}\sum_{i\in I} H_i^2$$

où θ et la première somme sont F-différentiables par hypothèse et définition de I, donc la seconde somme l'est aussi, la différentiabilité étant considérée en x_k. Pour tout d de norme unitaire et $\varepsilon > 0$ petit, nous pouvons écrire le développement $\cdot(x_k + \varepsilon d) = \cdot(x_k) + \varepsilon \cdot'(x_k)d + o(\varepsilon)$. Rappelons que H_i pour $i \in I$ n'est pas différentiable en x_k donc $F_i(x_k) = G_i(x_k)$, c'est-à-dire $a_i + A_{i,:}x_k = H_i^k = b_i + B_{i,:}x_k$, abrégé avec $H_I^k = (H_i^k)_{i\in I}$. Ainsi, le développement devient

$$\begin{aligned}
&\frac{1}{2\varepsilon}\left[\sum_{i\in I}[H_i^k + \varepsilon \min(A_{i,:}d, B_{i,:}d)]^2 - \sum_{i\in I}(H_i^k)^2\right] \\
&= \frac{1}{2\varepsilon}\left[\sum_{i\in I}(H_i^k)^2 + 2H_i^k\varepsilon\min(A_{i,:}d, B_{i,:}d) + \varepsilon^2[\min(A_{i,:}d, B_{i,:}d)]^2 - (H_i^k)^2\right] \\
&= \sum_{i\in I} H_i^k \min(A_{i,:}d, B_{i,:}d) + \frac{\varepsilon}{2}[\min(A_{i,:}d, B_{i,:}d)]^2
\end{aligned}$$

Clairement le second terme tend vers 0 quand $\varepsilon \to 0$. De plus, la F-différentiabilité assure que

$$d \mapsto \sum_{i\in I} H_i^k \min(A_{i:}d, B_{i:}d) \quad \text{est linéaire en } d. \tag{4.14}$$

Avant d'appliquer la proposition 4.4.7, notons que si pour un certain indice i, H_i^k est nul pour une infinité de k, alors en extrayant une sous-suite où $H_i^k \equiv 0$, la continuité implique que $H_i(x) = 0$. Ainsi, dans la formule à prouver, l'indice i est sans importance. Comme dans la preuve pour la C-fonction de Fischer, nous avons une expression de la forme "$0\times$ [terme d'indice i]", donc nous pouvons choisir une ligne arbitraire pour $J_{i,:}$. En prenant une sous-suite appropriée, on peut supposer $H_i^k \neq 0$ pour $i \in I$ (en modifiant I si nécessaire

en supprimant les indices pour lesquels $H_i(x) = 0$). [3]

En utilisant $\min(A_{i,:}d, B_{i,:}d) = A_{i,:}d + \min(0, v_i^\mathsf{T} d)$, si les vecteurs v_i ne sont pas deux à deux colinéaires, alors $H_i^k = 0$ pour tout i, ce qui contredit clairement l'hypothèse précédente.

De plus, supposons qu'il existe un v_i qui n'est colinéaire à aucun autre. En utilisant le même argument que dans la preuve de la proposition 4.4.7, des deux côtés de l'hyperplan $v_i^\perp$, nous obtiendrions que $H_i^k = 0$, ce qui contredit l'hypothèse.

En résumé, pour chaque v_i, il existe (au moins) un $v_{i'}$ colinéaire à v_i : les indices peuvent être regroupés en sous-ensembles où tous les v_i sont colinéaires à un seul vecteur w_j. Nous appliquons alors la proposition 4.4.7 aux vecteurs non colinéaires, leurs coefficients étant regroupés via le lemme 4.4.8. Ces relations de colinéarité s'expriment de la manière suivante, en revenant à (4.14) : sans perte de généralité, supposons que les vecteurs $v_1, \dots, v_{p_1}$ sont colinéaires à w_1, puis $v_{p_1+1}, \dots, v_{p_2}$ sont colinéaires à w_2 et ainsi de suite. Nous supposons également que l'ensemble d'indices I s'écrit $I = \cup_{j=1}^q I_j$ avec $I_j := [p_{j-1}+1 : p_j]$ où $p_0 = 0$ et les p_j sont des entiers,

$$
\begin{aligned}
\sum_{i\in I} H_i^k \min(A_{i:}d, B_{i:}d) &= \sum_{i\in I} H_i^k A_{i:}d + \sum_{i\in I} H_i^k \min(0, v_i^\mathsf{T} d) \\
&= \mathcal{L}d + \sum_{i=1}^{p_1} H_i^k \min(0, \alpha_i w_1^\mathsf{T} d) + \cdots + \sum_{i=p_{q-1}+1}^{p_q} H_i^k \min(0, \alpha_i w_q^\mathsf{T} d) \\
&= \left.\begin{array}{l} \mathcal{L}d + \displaystyle\sum_{j=1}^{q} -\mathcal{C}_{(j)}^- w_j^\mathsf{T} d \end{array}\right\} \quad \text{linéaire} \\
&\quad + (\mathcal{C}_{(1)}^+ + \mathcal{C}_{(1)}^-) \min(0, w_1^\mathsf{T} d) + \cdots + (\mathcal{C}_{(q)}^+ + \mathcal{C}_{(q)}^-) \min(0, w_q^\mathsf{T} d)
\end{aligned}
$$

avec $\mathcal{C}_{(j)}^+ = \sum_{p_{j-1}+1}^{p_j} H_i^k \alpha_i$ pour les α_i positifs, et $\mathcal{C}_{(j)}^- = \sum_{p_{j-1}+1}^{p_j} H_i^k |\alpha_i|$ pour les α_i négatifs.

Sous cette forme, la proposition 4.4.7 peut être appliquée : comme les w_j ne sont pas colinéaires, les coefficients devant les minima sont nuls : $\mathcal{C}_{(j)}^+ + \mathcal{C}_{(j)}^- = 0$ pour tout $j \in [1 : q]$. Ceci montre que la dernière ligne s'annule. En revenant à la partie linéaire, nous retrouvons

3. Dans [195, théorème 5.e, p. 232], Pang utilise une approche similaire pour montrer des propriétés de différentiabilité de la reformulation par minimum, mais avec des hypothèses plus fortes. La connaissance acquise sur la structure du minimum exposée au chapitre 3 semble nous permettre d'obtenir des résultats légèrement plus généraux.

une forme liée à une jacobienne potentielle du B-différentiel :

$$\begin{aligned}\sum_{i\in I} H_i^k A_{i,:}d + \sum_{j=1}^{q} -\mathcal{C}_{(j)}^{-} w_j^\intercal d &= \sum_{i\in I} H_i^k A_{i,:}d + \sum_{j=1}^{q}\sum_{i=p_{j-1}+1,\alpha_i<0}^{p_j} H_i^k \alpha_i w_j^\intercal d \\ &= \sum_{i\in I} H_i^k A_{i,:}d + \sum_{j=1}^{q}\sum_{i=p_{j-1}+1,\alpha_i<0}^{p_j} H_i^k v_i^\intercal d \\ &= \sum_{i\in I} H_i^k A_{i,:}d + \sum_{i\in I,\alpha_i<0} H_i^k (B_{i,:} - A_{i,:})d\end{aligned} \tag{4.15}$$

qui est une combinaison de lignes de A et B pour les indices de I. En particulier, cette expression ne dépend pas de d, puisqu'elle ne dépend que du signe des α_i, qui sont liés aux v_i et aux w_j. Pour chaque indice i, le terme correspondant dans la somme est de la forme souhaitée $H_i^k A_{i,:}$ ou $H_i^k B_{i,:}$.

Il nous reste seulement à montrer que la matrice résultante, dont la ligne $i \in I$ vaut $A_{i,:}$ si $\alpha_i > 0$ et $B_{i,:}$ si $\alpha_i < 0$ ($\alpha_i \neq 0$ par l'hypothèse de colinéarité), est bien dans le B-différentiel. Les lignes correspondant aux indices dans $\mathcal{F}(x), \mathcal{G}(x), \mathcal{E}^{=}(x)$ sont nécessairement similaires. Comme nous avons pris une sous-suite appropriée, les lignes avec indices dans $\mathcal{E}^{\neq}(x)\backslash I$ sont déjà connues.

Notons que la trajectoire est telle que H_I est non différentiable, c'est-à-dire que x_k appartient aux hyperplans correspondants ($\{y, v_i^\intercal y = v_i^\intercal x\}$). Nous devons donc justifier que pour l'arrangement d'hyperplans défini par les $v_i^\perp$ pour $i \in I$, la sous-matrice obtenue correspond à une région non vide.

Pour cette dernière partie, nous utilisons que les w_j peuvent être choisis pour former un cône pointé. En fait, l'orientation des w_j détermine la (sous-)matrice. Une direction y du cône formé par les w_j est telle que $x_k + \varepsilon y$ n'appartient à aucun hyperplan pour un $\varepsilon > 0$ approprié, et appartient à la région définie par la dernière ligne de (4.15) (la région avec uniquement des $+1$ correspondant au cône pointé des w_j). Nous avons ainsi montré que la dérivée de θ peut s'exprimer de manière appropriée, ce qui indique que la limite v est dans $\partial_B H(x)^\intercal H(x)$.

De plus, pour les indices tels que $H_i^k \equiv 0$, la direction y peut être modifiée si nécessaire pour assurer que les produits $v_i^\intercal y$ sont non nuls. La sous-matrice peut alors être complétée pour ces indices, car ces termes sont nuls en x_k et x. □

Reprenons maintenant la preuve dans le cas où ψ est une fonction différentiable (pas nécessairement $||\cdot||^2/2$).

Preuve. D'abord, dans le cas séparable, $\psi \circ H = \sum \psi_i(H_i)$, la même preuve peut être faite en remplaçant les H_i^k par $\psi_i'(H_i^k)$ partout.

Dans le cas général, pour les indices $i \notin I$ et ε assez petit, le minimum est déjà connu

et noté $J_{i:}d$. Le développement s'écrit :

$$\begin{aligned}
&\frac{1}{\varepsilon}\left[\psi\begin{pmatrix}(H_i^k+\varepsilon\min(A_{i:}d,B_{i:}d))_{i\in I}\\(H_i^k+\varepsilon J_{i:}d)_{i\in I^c}\end{pmatrix}-\psi\begin{pmatrix}(H_i^k)_{i\in I}\\(H_i^k)_{i\in I^c}\end{pmatrix}\right]\\
&=\psi'((H_i^k)_i)\begin{bmatrix}(\min(A_{i:}d,B_{i:}d))_{i\in I}\\(J_{i:}d)_{i\in I^c}\end{bmatrix}+o(1)\\
&=\sum_{j\in I^c}\partial_j\psi((H_i^k)_i)J_{j:}d+\sum_{j\in I}\partial_j\psi((H_i^k)_i)\min(A_{j:}d,B_{j:}d)+o(1)
\end{aligned}$$

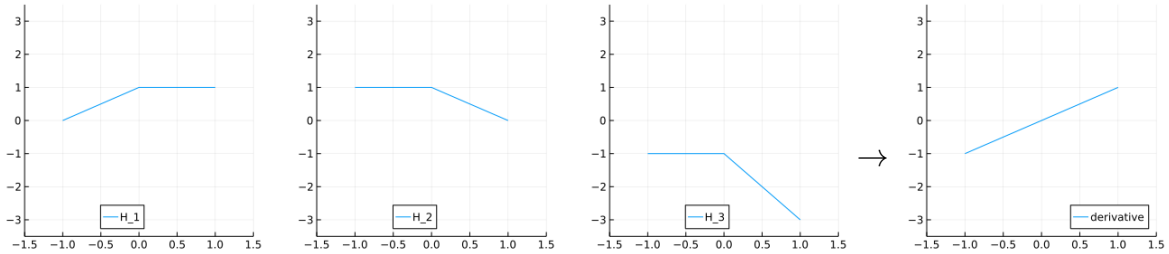
En utilisant à nouveau la F-différentiabilité, cette expression est linéaire en d. En retirant la somme sur I^c, la somme restante est encore linéaire. Le même raisonnement peut alors être appliqué, avec les scalaires $\partial_j\psi((H_i^k)_{i\in I})$. □

Une remarque pertinente est que ce processus exhibe une sous-matrice particulière pour les indices dans I. Cependant, ceci vient du choix de la "direction" des w_j : choisir $-w_j$ au lieu de w_j changera les signes de certains α_i, ce qui modifiera certaines lignes de la sous-matrice.

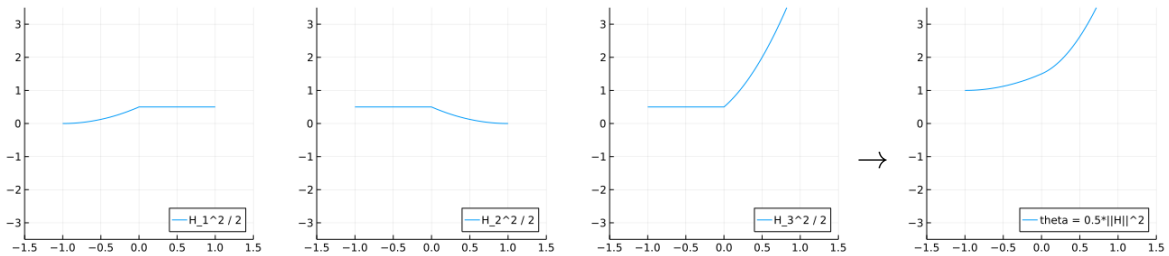
Nous illustrons un exemple où les vecteurs sont deux à deux colinéaires, donc où les valeurs H_i^k ne doivent pas être nulles pour avoir un opérateur linéaire dans (4.14). Considérons les cas suivants (en identifiant vecteurs lignes et colonnes) :

$$H_1=+1,H_2=+1,H_3=-1,\quad A_1=0,B_1=e_1,A_2=0,B_2=-e_1,A_3=0,B_3=-2e_1$$
$$+1\min(0,d_1)+\min(0,-d_1)-\min(0,-2d_1)=d_1.$$

On observe que θ est différentiable à 0 mais pas les H_i.



(a) $\min(0,t)+1$ (b) $\min(0,-t)+1$ (c) $\min(0,-2t)-1$ (d) $\sum H_i^k\min(\dots)$



(e) $H_1^2/2$ (f) $H_2^2/2$ (g) $H_3^2/2$ (h) $\sum H_i^2/2$

Lorsqu'on considère la fonction de Fischer au lieu du minimum, la formule est également vraie, et a l'avantage que les colonnes multivoques sont multipliées par zéro, donc $\partial_B \Psi$ est en fait un seul élément [87]. Comme le montre l'exemple précédent, ce n'est pas ce qui se passe dans le cadre du minimum. Comme montré ci-dessous, cela peut conduire à ce que la fonction de mérite ait plusieurs éléments dans son B-différentiel, lorsque $H_i(x) \neq 0$, alors les matrices jacobiennes sont multipliées par un vecteur non nul, donc les produits sont différents.

Proposition 4.4.10 (Cas non linéaire)**.** *La proposition 4.4.9 reste valable pour des fonctions non linéaires ($\mathcal{C}^1$) F et G.*

Preuve. La question principale est de vérifier ce qui ne s'applique pas depuis le cas linéaire et doit être adapté. L'inclusion $[\subseteq]$ est exactement la même, car la relation $\nabla\theta = \nabla H \times \nabla\psi(H)$ signifie que la suite donnée $\{x_k\}_k$ est suffisante.

Pour l'inclusion inverse, si H est différentiable, on peut faire de même : l'élément $v \in \partial_B\theta$ est la limite de $\nabla H(x_k) \times \nabla\psi(H(x_k))$. En prenant une sous-suite appropriée, le premier terme peut être choisi tel qu'une ligne fixe i soit $F_i'(x_k)$ ou $G_i'(x_k)$ (selon la suite) pour tout k, ce qui correspond à la transposée d'une matrice du B-différentiel de H. Ainsi, le produit a la forme souhaitée.

Si H n'est pas différentiable le long de la suite, en prenant une sous-suite comme précédemment, l'ensemble des indices I pour lesquels H_I n'est pas différentiable est constant. Cependant, les indices peuvent maintenant appartenir à $\mathcal{E}^{\neq}(x) \cup \mathcal{E}^{=}(x)$, alors que $\mathcal{E}^{=}(x)$ était précédemment exclu.

En développant de manière similaire, la F-différentiabilité indique que pour tout k de la sous-suite appropriée, la fonction suivante est linéaire :

$$d \mapsto \sum_{j \in I^c} \partial_j\psi((H_i^k)_i) J_{i:}^k d + \sum_{j \in I} \partial_j\psi((H_i^k)_i) \min(F_i'(x_k)d, G_i'(x_k)d).$$

La principale différence par rapport au cas linéaire est que les $F_i'(x_k), G_i'(x_k)$ ne correspondent pas aux $v_i := G_i'(x) - F_i'(x)$ (puisque $x_k \neq x$). Premièrement, si pour certains i et une sous-suite associée les valeurs $\partial_j\psi((H_i^k)_i)$ sont toutes nulles, alors à la limite, par continuité, on a $\partial_j\psi((H_i)_i) = 0$, donc ce terme n'est pas pertinent dans l'expression finale. Ainsi, on peut compléter la matrice jacobienne à la fin de la preuve.

Néanmoins, on peut toujours appliquer le lemme 4.4.8, car le terme de la dérivée est une somme de "minima pondérés" comme dans le cas linéaire. En effet, en prenant une sous-suite, on peut supposer que l'ensemble des indices I pour lesquels H_i n'est pas différentiable en x_k pour $i \in I$ est constant et indépendant de k.

Cela signifie que pour tout k, les $v_i^k := G_i'(x_k) - F_i'(x_k)$ peuvent être groupés par colinéarité, et les coefficients associés incluant les $||v_i^k||$ (c'est-à-dire les α_i) et les H_i^k s'annulent.

On a, à l'indice k,

$$\begin{cases} \forall\, i \in I_{j_1^k}, v_i^k = \alpha_i^k w_{j_1}^k, \ldots, \forall\, i \in I_{j_{q^k}^k}, v_i^k = \alpha_i^k w_{j_{q^k}}^k \\ \forall\, m \in [1 = j_1^k : j_{q^k}^k], ||w_m^k|| = 1 \\ \forall\, m, \sum_{i \in I_m \alpha_i^k > 0} \partial_i \psi((H_l^k)_l)\alpha_i^k - \sum_{i \in I_m, \alpha_i^k < 0} \partial_i \psi((H_l^k)_l)\alpha_i^k = 0 \end{cases} \tag{4.16}$$

où la dernière ligne s'applique pour chaque groupe de colinéarité I_m (éventuellement dépendant de k). Pour $i \in \mathcal{E}^=(x)$, comme $v_i = 0$ on a $\alpha_i^k \to 0$.

L'ensemble des indices I étant fixé (par les sous-suites précédentes), en particulier il a une taille finie constante, donc le nombre de partitions de I groupant les indices par colinéarité des vecteurs associés est également fini. En extrayant plus de sous-suites, on peut supposer que les groupes de colinéarité sont les mêmes le long d'une bonne sous-suite, c'est-à-dire que les relations $v_i^k \propto w_j^k$ sont vraies pour les mêmes groupes d'indices $[1 : p_1], \ldots, [p_{q-1} + 1 : p_q]$.

De manière similaire, cela peut être fait pour les signes des α_i^k (nombre fini de possibilités). Fixer les signes des α_i^k résulte en une sous-suite pour laquelle l'analogue de (4.15) a, pour chaque indice de la sous-suite, la forme :

$$\sum_{i \in I} \partial_i \psi((H_l^k)_l) F_i'(x_k) d + \sum_{i \in I, \alpha_i^k < 0} \partial_i \psi((H_l^k)_l)(G_i'(x_k) - F_i'(x_k)) d \tag{4.17}$$

Les extractions ci-dessus résultent en chaque point de la sous-suite donnant une (sous-)matrice jacobienne qui a la même composante ($F_i'(x_k)$ ou $G_i'(x_k)$) pour un indice donné i pour tout k. Pour tout k, on peut trouver une direction d_k telle que $d_k^\intercal w_j^k > 0$, conduisant à la matrice jacobienne correspondant à (4.17). Ainsi, la suite $x_k + t_k d_k$ pour un $t_k > 0$ approprié a la propriété désirée.

Pour conclure, la régularité des fonctions fait que $H_i^k \to H_i(x)$ et les lignes de la matrice jacobienne convergent également vers leurs limites $F_i'(x)$ ou $G_i'(x)$. On retrouve la forme souhaitée, donc l'égalité $\partial_B \theta = (\partial_B H)^\intercal \times \nabla\psi(H)$ est vérifiée. □

4.5 Détails sur les instances et algorithmes

Cette section vise à donner des preuves sur certaines valeurs affirmées dans le chapitre 3 concernant certains types d'instances. Premièrement, la section 4.5.1 considère les instances PERM, tandis que la section 4.5.2 discute des instances CROSSPOLYTOPE.

4.5.1 À propos des instances permutahedron

Concentrons-nous maintenant sur les instances PERM. La définition que nous utilisons pour ces arrangements est la suivante : pour un entier positif n, il y a $n(n+1)$ hyperplans,

donnés par :

$$H_i := \{x_i = 0\} \text{ pour } 1 \leqslant i \leqslant n, \qquad H_{ij} = \{x_i - x_j = 0\} \text{ pour } 1 \leqslant i < j \leqslant n. \quad (4.18)$$

De tels arrangements sont bien connus et peuvent être résolus par combinatoire. D'abord, nous détaillons les chambres puis les circuits.

Chambres

Approche analytique Nous montrons qu'il y a $(n+1)!$ vecteurs de signes. D'abord, considérons les $n(n-1)/2$ hyperplans $H_{ij} = (e_i - e_j)^\perp$. Soit x un point donné,

$$x \in \mathbb{R}^n \backslash (\cup_{i,j} H_{ij}) \qquad \Longleftrightarrow \qquad (x_1, \ldots, x_n) \text{ sont tous différents,}$$

puisque s'il existe une paire (i,j) telle que $x_i = x_j$, alors $x \in H_{ij}$. Maintenant, il y a $n!$ façons d'ordonner les coordonnées puisqu'elles sont toutes différentes, les $n!$ permutations. En effet, soit π une permutation de $[1:n]$, telle que $x_{\pi(1)} > \cdots > x_{\pi(n)}$. Si $\pi(i) < \pi(j)$, $x \in H^+_{\pi(i)\pi(j)}$, tandis que $\pi(i) > \pi(j)$ implique $x \in H^-_{\pi(j)\pi(i)}$ par définition des hyperplans considérés. Par conséquent, pour une permutation arbitraire π et les x ayant des coordonnées décroissantes sous π, les signes correspondant aux H_{ij} sont déterminés. Ainsi, il y a $n!$ chambres.

Ensuite, nous montrons que chacune de ces $n!$ régions est exactement divisée en $n+1$ sous-régions par les n hyperplans restants $H_k = \{x_k = 0\}$. Cela conduira à $(n+1) \times n! = (n+1)!$ régions. Soit π une permutation de taille n, telle que $x_{\pi(1)} > \cdots > x_{\pi(n)}$. Alors on peut avoir les configurations suivantes :

$$\begin{array}{lll} x_{\pi(i)} > 0 \quad \forall\, i \in [1:n] & \text{et} \quad x_{\pi(1)} < 0, x_{\pi(i)} > 0 \quad \forall\, i > 1, \\ \ldots & & \ldots, \\ x_{\pi(i)} < 0, x_{\pi(n)} > 0 \; \forall\, i < n & \text{et} \quad x_{\pi(i)} < 0 \quad \forall\, i \in [1:n]. \end{array}$$

Toute autre combinaison est de la forme

$$\{x_{\pi(1)} < 0, \ldots, x_{\pi(i^*)} > 0, \ldots, x_{\pi(j^*)} < 0, \ldots, x_{\pi(n)} > 0\}$$

qui ne respecte pas la définition de π.

Avec l'approche d'Athanasiadis En suivant la suggestion de [35], on peut utiliser la méthode décrite dans le théorème 2.2 et l'exemple 2.3 de [12]. Elle s'énonce comme suit.

Théorème 4.5.1 (2.2 dans [12]). *Soit q un nombre premier suffisamment grand. Un arrangement avec des hyperplans définis par des coordonnées entières (ou rationnelles) vérifie*

$$\chi(q) = \text{card}\left(\mathbb{F}_q^n \setminus \cup_k H_k\right)$$

où χ est le polynôme caractéristique de l'arrangement, $\cup_k H_k$ est l'union des hyperplans, $\mathbb{F}_q := [0 : q-1] \mod q$ donc $\mathbb{F}_q^n$ est identifié à $[0 : q-1]^n$ (toutes les coordonnées $\mod q$). Ainsi, $\chi(q)$ compte le nombre de points avec des coordonnées entières dans $[0 : q-1]^n$ qui ne satisfont ($\mod q$) aucune des équations définissant les hyperplans. □

L'utilité de ce théorème est que, une fois le membre de droite obtenu, cette expression (dépendant de q, n et l'arrangement), peut être évaluée en considérant q comme la variable du polynôme. En particulier, l'évaluation en -1 est liée au nombre de chambres, par l'identité $|\mathcal{S}| = (-1)^n \chi(-1)$ [257]. Obtenons ces expressions pour les arrangements considérés.

Pour un q fixé assez grand et un $x \in [0 : q-1]^n$, il est clair qu'on doit avoir $x_i \neq 0$ pour éviter de vérifier les équations de H_i. Maintenant, pour les H_{ij}, x doit avoir des coordonnées différentes. Ainsi, il y a $q-1$ possibilités pour la première (puisqu'elle ne peut pas être 0), $q-2$ pour la seconde, et finalement, $q-n$ pour la n-ème coordonnée. Par conséquent, le polynôme caractéristique est $\prod_{i=1}^{n}(q-i)$. Alors, le nombre de chambres est

$$(-1)^n \prod_{i=1}^{n}(-1-i) = (-1)^n \prod_{i=1}^{n}(-1)(i+1) = (n+1)!.$$

Notons que cette formule très efficace n'indique pas *quelles* sont les chambres, alors que l'analyse précédente en est capable.

vecteurs souches / circuits

Il est également possible d'exprimer explicitement l'ensemble des circuits / des vecteurs souches. D'abord, rappelons que les nombres de vecteurs souches des instances PERM-N pour N = 5,6,7,8 sont respectivement 197, 1172, 8018, 62814. Ces nombres correspondent à la suite A002807 dans l'OEIS, à un décalage d'indice près (nombre de circuits $= a(n+1)$). La formule donnée correspond précisément aux circuits, comme détaillé dans la proposition suivante. Rappelons qu'on a

$$V = [I_n \; M], \; M = [e_i - e_j]_{1 \leqslant i < j \leqslant n}.$$

Proposition 4.5.2 (circuits des instances PERM). *Le nombre de circuits de PERM-N est donné par*

$$|\mathcal{C}(\textit{PERM-N})| = \sum_{k=3}^{n+1} \frac{(k-1)!}{2} \binom{n+1}{k}$$

En particulier, le nombre de circuits de taille $k \in [3 : n+1]$ est précisément le k-ème terme de la somme.

La preuve repose sur les notions artificielles mais utiles suivantes.

Définition 4.5.3 (coordonnées couvertes par J). Soit $J \subseteq [1 : p]$. On note $c_J := |\{i \in [1 : p] : \exists\, j \in J, (v_j)_i \neq 0\}|$ le nombre de lignes non nulles de $V_{:,J}$. □

Définition 4.5.4 (composantes non nulles dans $\boldsymbol{J}$)**.** Soit $J \subseteq [1:p]$. On note $K_J := \sum_j ||v_j||_1$ le nombre total de composantes non nulles des vecteurs dans J, en sachant que $||v_j||_1 = \sum_1^n |(v_j)_i|$ et $(v_j)_i \in \{-1, 0, +1\}$ pour tout $i \in [1:n]$ et $j \in J$. □

Pour le V définissant PERM-N, on a clairement $c_J \leqslant n$ et $K_J \in [k, 2k]$ pour tout J puisque les colonnes de V appartiennent à $\mathbb{R}^n$ et chacune d'elles a une ou deux composantes non nulles. Dans ce qui suit, nous appelons "coordonnées" un sous-ensemble de $[1:n]$, et "composantes" la(les) valeur(s) d'un (ou plusieurs) vecteur(s) éventuellement à certaines coordonnées spécifiques. Justifions maintenant la proposition ci-dessus.

Preuve. D'abord, le rang de la matrice V est clairement n. Ainsi, les circuits sont de taille au plus $n+1$, et au moins 3, car il n'y a pas de vecteurs colinéaires.

Montrons que pour tout circuit J de taille k, $c_J \in \{k-1, k\}$. Soit $c_J = l$ et notons $i_1, \ldots, i_l$ les coordonnées associées.

- supposons que $l \leqslant k-2$. Par définition, les vecteurs v_j pour $j \in J$ ont des composantes non nulles seulement aux coordonnées $i_1, \ldots, i_l$. Ainsi, V_J est une sous-matrice de $[e_{i_1}, \ldots, e_{i_l}, e_k - e_{k'}]$ (en supposant $i_1 \leqslant k < k' \leqslant i_l$ sans perte de généralité). Cependant, cette matrice est clairement de rang $l \leqslant k-2$ donc ses circuits sont de taille $\leqslant k-2+1 = k-1 < k$.
- supposons que $l \geqslant k+1$. Par définition de V, $k \leqslant K_J \leqslant 2k$ puisque chaque v a une ou deux composantes non nulles. Comme J est un circuit, $\mathrm{null}(V_J) = 1$, signifiant $V_J\eta = 0 \in \mathbb{R}^n$ pour un $\eta \in \mathbb{R}^J_*$. Les indices $k \notin \{i_1, \ldots, i_l\}$ vérifient clairement $(V_J\eta)_k = \sum_{j\in J}(v_j)_k\eta_j = 0$, puisque $(V_{:,j})_k = V_{k,j} = 0$ pour tout $j \in J$. Maintenant, choisissons un indice $i \in i_1, \ldots, i_l$. En utilisant les égalités $V_J\eta = 0$ et $(V_J\eta)_i = \left(\sum_{j\in J} v_j\eta_j\right)_i = \sum_{j\in J}(v_j)_i\eta_j = 0$, il doit y avoir au moins deux vecteurs v_j ayant une composante non nulle i pour avoir $V_J\eta = \sum_j v_j\eta_j = 0 \in \mathbb{R}^n$. Ainsi, pour toutes les l coordonnées $i_1, \ldots, i_l$, il y a au moins $2l$ coordonnées non nulles dans les vecteurs $\{v_j : j \in J\}$, ce qui implique $K_J \geqslant 2l$. Cela contredit $K_J \leqslant 2k = 2|J|$.

Maintenant que les circuits J de taille k vérifient $c_J \in \{k-1, k\}$, il suffit de compter ces deux types de circuits pour tout $k \in [3:n+1]$. Soit $k \in [3:n+1]$, et définissons

$$C_1(n,k) := \frac{(k-1)!}{2}\binom{n}{k-1}, \quad C_2(n,k) := \frac{(k-1)!}{2}\binom{n}{k}.$$

Notons que pour $k = n+1$, $C_2(n,k) = 0$ puisqu'il n'y a que n coordonnées. Nous montrons qu'il y a $C_1(n,k)$ circuits de taille k avec $c_J = k-1$ et $C_2(n,k)$ circuits de taille k avec $c_J = k$. Comme ce sont les seules possibilités pour c_J, le nombre total de circuits sera le résultat annoncé :

$$\sum_{k=3}^{n+1} \frac{(k-1)!}{2}\left[\binom{n}{k-1} + \binom{n}{k}\right] = \sum_{k=3}^{n+1} \frac{(k-1)!}{2}\binom{n+1}{k}.$$

Soit $k \in [3 : n+1]$ et J un circuit de taille k tel que $c_J = k$ (si $k = n+1$ il n'y a rien à faire, il n'y a pas de tels circuits). Soit $i_1, i_2, \ldots, i_k$ les coordonnées associées à J. Comme il y a k coordonnées, on a $K_J \geqslant 2k$. Cependant, on a aussi $K_J \leqslant 2k$: chaque vecteur doit avoir deux composantes non nulles, signifiant $J \subseteq \{e_i - e_j\}_{1 \leqslant i < j \leqslant n}$.

De plus, comme $c_J = k$, il y a exactement deux vecteurs avec une composante non nulle à la coordonnée i_1, deux (autres) vecteurs avec une composante non nulle à la coordonnée i_2, et ainsi de suite. Considérons une suite des k indices, maintenant nommés $j_1 < j_2 < \cdots < j_k$. Clairement $e_{j_1} - e_{j_2}, e_{j_2} - e_{j_3}, \ldots, e_{j_{k-1}} - e_{j_k}, e_{j_1} - e_{j_k}$ est une sous-matrice de nullité un, puisque $(+1, \ldots, +1, -1)$ est dans son noyau et les premiers $k-1$ vecteurs forment une famille de rang $k-1$: les indices correspondants forment un circuit de taille k.

Comme le choix de $i_1, \ldots, i_k$ est arbitraire, cela donne $\binom{n}{k}$ possibilités. Maintenant, pour un choix fixé de $i_1, \ldots, i_k$, justifions qu'il y a $(k-1)!/2$ circuits possibles de taille k pour ces indices. Il y a $k!$ façons possibles d'ordonner les indices, les permutations de $[1 : k]$. Cependant, le paragraphe suivant montre que les circuits résultants sont indépendants par permutation circulaire et par symétrie.

Soit $\pi \in \mathfrak{S}([1 : k])$, et notons $i_{\pi(1)}, \ldots, i_{\pi(k)}$ les indices des coordonnées dans l'ordre modifié par π. Les vecteurs de V qui forment un circuit pour cet ordre sont précisément les vecteurs $\pm(e_{i_{\pi(1)}} - e_{i_{\pi(2)}}), \pm(e_{i_{\pi(2)}} - e_{i_{\pi(3)}}), \ldots, \pm(e_{i_{\pi(1)}} - e_{i_{\pi(k)}})$. Cependant, les indices donnés par une permutation circulaire de π, à savoir, $i_{\pi(1+j_0)}, i_{\pi(2+j_0)}, \ldots, i_{\pi(k+j_0)}$ (pour j_0 un entier fixé), forment un circuit avec les mêmes vecteurs. De même, la suite d'indices $i_{\pi(k)}, i_{\pi(k-1)}, \ldots, i_{\pi(1)}$ forme un circuit avec ces mêmes vecteurs. En résumé ces observations, il y a $C_2(n,k)$ vecteurs souches de cette forme : l'invariance par permutation circulaire et par symétrie divise $k!$ par k et par 2 respectivement ($k \geqslant 3$). Expliquons cela pour $k = 4$. Pour simplifier, nous supposons $i_1 = 1, i_2 = 2, i_3 = 3, i_4 = 4$. On a $4!$ façons d'ordonner l'ensemble $\{1,2,3,4\}$, mais par exemple

$$\begin{aligned}
& \{e_1 - e_2, e_2 - e_3, e_3 - e_4, e_1 - e_4\} && [1-2-3-4] \\
[\text{symétrie}] &= \{e_3 - e_4, e_2 - e_3, e_1 - e_2, e_1 - e_4\} && [4-3-2-1] \\
[\text{circulaire}] &= \{e_2 - e_3, e_1 - e_2, e_1 - e_4, e_3 - e_4\} && [3-2-1-4] \\
[\text{symétrie}] &= \{e_1 - e_4, e_1 - e_2, e_2 - e_3, e_3 - e_4\} && [4-1-2-3] \\
[\text{circulaire}] &= \{e_3 - e_4, e_1 - e_4, e_1 - e_2, e_2 - e_3\} && [3-4-1-2] \\
[\text{symétrie}] &= \{e_1 - e_2, e_1 - e_4, e_3 - e_4, e_2 - e_3\} && [2-1-4-3] \\
[\text{circulaire}] &= \{e_1 - e_4, e_3 - e_4, e_2 - e_3, e_1 - e_2\} && [1-4-3-2] \\
[\text{symétrie}] &= \{e_2 - e_3, e_3 - e_4, e_1 - e_4, e_1 - e_2\} && [2-3-4-1]
\end{aligned}$$

Un ordre différent, par exemple $\{1,3,2,4\}$, impliquerait de nouveaux vecteurs comme $e_1 - e_3$, ce qui signifie qu'un circuit différent est considéré.

Nous avons ainsi identifié que les circuits de taille k avec $c_J = k$ sont de la forme, pour les coordonnées $(i_1, \ldots, i_k)$ dans cet ordre, des vecteurs $\operatorname{sgn}(i_{l+1} - i_l)(e_{i_l} - e_{i_{l+1}})$ pour $l \in [1 : k]$ (où les indices sont compris $\mod k$). Réciproquement, il est clair que de tels

sous-ensembles sont des circuits et qu'il y a $C_2(n,k)$ de tels sous-ensembles.

Maintenant, nous considérons les circuits J tels que $c_J = k-1$. Avec un argument similaire, soit $k' \in [1:k]$ le nombre de vecteurs v_j pour $j \in J$ tels que $v_j \in \{e_i\}_{i\in[1:n]}$. Clairement $K_J = k' \times 1 + (k-k') \times 2 = 2k - k'$. Comme $c_J = k-1$ coordonnées sont impliquées dans les circuits, et qu'il y a $K_J \geqslant 2(k-1) = 2k-2$ coordonnées non nulles totales, signifiant $k' \leqslant 2$. Si $k' = 0$, cela se réduit au sous-cas précédent : le sous-ensemble est composé de deux circuits ou plus (mais deux ont un indice commun, par exemple $e_1 - e_2, e_2 - e_3, e_1 - e_3, e_1 - e_4, e_1 - e_5, e_4 - e_5$). Si $k' = 1$, en notant e_{i^*} le vecteur associé, comme $K_J = 2k-1$, les $k-1$ vecteurs de la forme $e_i - e_j$ sans le vecteur e_{i*} forment déjà un circuit puisqu'ils couvrent $k-1$ coordonnées et sont $k-1$. Autrement dit, la composante de η correspondant à e_{i^*} est nulle : c'est une contradiction. La seule possibilité est $k' = 2$, signifiant qu'il existe une paire d'indices (i^*, j^*) telle que e_{i^*}, e_{j^*} appartiennent aux colonnes de V_J.

Maintenant, comme $c_J = k-1$ et $K_J = 2k-2$, pour chaque coordonnée $i_1, \dots i_{k-1}$ doit avoir deux vecteurs ayant une composante non nulle dans cette coordonnée. Comme e_{i^*} et e_{j^*} n'ont qu'une composante non nulle aux positions i^* et j^*, un vecteur doit être de la forme $\pm(e_{i^*} - e_{i_k})$, un de la forme $\pm(e_{j^*} - e_{i'_k})$, et le raisonnement est poursuivi comme avant dans le cas $c_J = k-1$. Essentiellement, un des vecteurs $e_i - e_j$ est scindé en la paire (e_i, e_j).

En comptant les circuits, comme le choix des $k-1$ coordonnées est arbitraire, il y a $\binom{n}{k-1}$ possibilités. Ensuite, il y a $(k-1)!$ façons d'ordonner les coordonnées et $(k-1)!/2$ en tenant compte de la symétrie. Cependant, il n'y a pas d'invariance par permutation circulaire des $k-1$ indices impliqués. En effet, permuter l'ordre changerait (par exemple) la paire $\{e_{i_j}, e_{i_{j+1}}\}$ en $\{e_{i_{j+1}}, e_{i_{j+2}}\}$, ce qui est un circuit différent. Cela justifie l'hypothèse concernant $C_1(n,k)$.

Nous avons ainsi identifié que les circuits de taille k avec $c_J = k$ sont de la forme, pour les coordonnées $(i_1, \dots, i_k)$ dans cet ordre, des vecteurs $\operatorname{sgn}(i_{l+1} - i_l)(e_{i_l} - e_{i_{l+1}})$ sauf pour un $l_0 \in [1:k]$ tel qu'au lieu de $\operatorname{sgn}(i_{l_0+1} - i_{l_0})(e_{i_{l_0}} - e_{i_{l_0+1}})$ il y a $e_{i_{l_0}}$ et $e_{i_{l_0+1}}$ (où les indices en $i.$ sont compris $\mod k$). Réciproquement, il est clair que de tels sous-ensembles sont des circuits et qu'il y a $C_1(n,k)$ de tels sous-ensembles.

Pour conclure la preuve, considérons les circuits de taille $n+1$. Clairement ils ne peuvent pas avoir $c_J = n+1$ puisque $c_J \leqslant n$. Nécessairement ces circuits sont de la forme décrite par le point $c_J = k-1$, sinon le nombre de coordonnées non nulles est incorrect. Comme toutes les n coordonnées doivent être choisies, et $\binom{n}{n} = 1$, seul l'ordre des coordonnées compte : comme vu précédemment, ce nombre est $(n-1)!/2$, multiplié par n pour choisir quelle paire de coordonnées (i,j) est scindée avec e_i et e_j.

Enfin, les vecteurs souches sont clairement obtenus en regardant la structure des circuits : si $-e_i$ ou $-(e_i - e_j)$ (pour $i < j$) intervient, alors la coordonnée du vecteur souche est $-$, et $+$ sinon. □

4.5.2 À propos de l'arrangement de séparabilité du crosspolytope

Ensuite, nous nous concentrons sur les instances crosspolytope. Dans [35, section 6.4 p. 14], les crosspolytopes sont définis comme des polytopes n-dimensionnels avec les $2n$ sommets $\mathcal{V} = \{\pm e_i\}_{i\in[1:n]}$. À partir de là, l'arrangement de séparabilité associé a les $2n$ hyperplans définis par $\{(1; v)^\perp : v \in \mathcal{V}\}$. En particulier, la "dimension" est égale à $n+1$ et la première coordonnée est notée 0

$$V = \begin{bmatrix} e^\mathsf{T} & e^\mathsf{T} \\ I & -I \end{bmatrix}.$$

Chambres

Nous justifions la formule $|\mathcal{S}| = 2 \times 3^n - 2^n$, vérifiée numériquement pour des n petits et correspondant à la suite A027649 de l'OEIS. D'abord, nous appliquons [12] pour vérifier le cardinal donné dans [35]. Ensuite, une preuve de cette valeur par récurrence, utilisant le formalisme de l'algorithme de l'arbre, est proposée, permettant une énumération explicite des chambres.

Avec l'approche d'Athanasiadis Soit q un grand entier premier et $\mathbb{F}_q^{n+1}$ l'hypercube entier de taille q. Soient les équations définissant les hyperplans écrites comme

$$\begin{array}{llll} x_1 + x_0 = 0, & x_2 + x_0 = 0, & \dots & x_n + x_0 = 0, \\ x_1 - x_0 = 0, & x_2 - x_0 = 0, & \dots & x_n - x_0 = 0. \end{array}$$

Pour le décompte du théorème 4.5.1, on peut résumer les équations comme exigeant que chaque x_i pour $i \in [1 : n]$ soit différent de x_0 et $-x_0 \mod q$. Observons qu'il y a une légère différence si $x_0 = 0$ ou $x_0 \neq 0$. En effet, si $x_0 = 0 = -x_0$, alors tous les x_i pour $i \in [1 : n]$ peuvent prendre une valeur arbitraire dans $[1 : q-1]$, ce qui donne $(q-1)^n$ dans le polynôme caractéristique. Ensuite, quand $x_0 \neq 0$ ($q-1$ choix), les x_i pour $i \in [1 : n]$ peuvent prendre toute valeur différente de x_0 et $-x_0 (= q - x_0 [\mod q])$, ce qui donne $q-2$ choix pour chacune des n valeurs. Ainsi, on a $\chi(q) = (q-1)^n + (q-1)(q-2)^n$. Finalement, on a

$$|\mathcal{S}| = (-1)^{n+1}[(-2)^n + (-2)(-3)^n] = -2^n + 2 * 3^n.$$

Énumération par récurrence

Proposition 4.5.5 (circuits des arrangements crosspolytope). *Soit $n \in \mathbb{N}^*$, les circuits des arrangements crosspolytope sont donnés par les $n(n-1)/2$ uplets $(i, j, i+n, j+n)$ pour $i \neq j \in [1 : n]^2$, donc $|\mathcal{C}(V)| = \binom{n}{2}$.*

Preuve. D'abord, montrons que les sous-ensembles indiqués sont des circuits. En effet, la sous-matrice s'écrit, en supprimant les lignes vides,

$$V_{base} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}.$$

Clairement, $V_{base}(1;-1;+1;-1) = 0 = V_{base}\mathrm{Diag}(+1,-1,+1,-1)(1;1;1;1)$ et toute combinaison de trois vecteurs parmi les quatre est linéairement indépendante.

Maintenant, considérons un circuit arbitraire. Clairement il ne peut pas avoir une taille inférieure à 4. Soit $i \in [1:n]$ et supposons que le circuit contienne v_i ou v_{i+n}. Comme seuls v_i et v_{i+n} ont une coordonnée non nulle i, nécessairement le circuit doit aussi contenir l'autre vecteur. Alors, le circuit est formé de paires $(i, i+n)$ pour certains indices i, mais comme tout tuple $(i, i+n, j, j+n)$ est déjà un circuit, il ne peut pas avoir une taille > 4 sans contredire la minimalité. □

Les vecteurs souches sont $\pm(+,+,-,-)$ pour les indices $(i, j, 1+n, j+n)$. La proposition suivante est une conséquence plutôt claire des propositions 3.3.10 et 4.5.5. Analysons cela de manière "constructive".

Proposition 4.5.6 (chambres des arrangements CROSSPOLYTOPE). *Soit $n \in \mathbb{N}^*$, les $2*3^n - 2^n$ chambres des arrangements CROSSPOLYTOPE correspondent à tous les vecteurs de signes de taille $2n$ tels que, pour toute paire $i \neq j \in [1:n]^2$, on n'a pas $s_i = s_{i+n} = -s_j = -s_{j+n}$.*

Dans ce qui suit, les chambres de l'arrangement CROSSPOLYTOPE-N en dimension $n+1$ sont notées $\mathcal{S}_n$ (également vues comme chambres en dimension $n+2$ pour les besoins de la récurrence).

Preuve. La fin de l'énoncé est claire en utilisant la proposition 4.5.5. Maintenant, procédons par récurrence sur n. Quand $n = 1$, les deux hyperplans sont $(1,1)^\perp$, $(1,-1)^\perp$: il y a quatre régions (c'est la seule dimension pour laquelle l'arrangement est complet, il n'y a pas de circuits) ce qui est $2 \times 3^1 - 2^1 = 4$.

Justifions d'abord la cardinalité des chambres. Supposons le résultat pour n. D'abord, remarquons que les matrices pour la dimension n et $n+1$ peuvent être écrites comme

$$V^n = \begin{bmatrix} 1_n & 1_n \\ I_n & -I_n \end{bmatrix}, \quad V^{n+1} = \begin{bmatrix} V^n_{1,:} & 1 & 1 \\ V^n_{[2:n+1],:} & 0_n & 0_n \\ 0_{2n} & 1 & -1 \end{bmatrix}.$$

L'idée principale de la preuve est la suivante : comme seuls les deux nouveaux vecteurs ont une coordonnée non nulle dans la $n+1$-ème dimension, $(1, e_{n+1})$ n'est pas engendré par les autres donc chaque nœud a un descendant. Mais ajouter ensuite $(1, -e_{n+1})^\perp$ ne duplique pas à nouveau (il est engendré par les autres $2n+1$ vecteurs). Avec la formule $2 \times 3^n - 2^n$,

nous montrons deux choses. Pour une partition spécifique (S^1, S^2) de $\mathcal{S}_n$ avec $|S^1| = 2^n$ et $|S^2| = 2 \times (3^n - 2^n)$, après avoir ajouté les deux nouveaux hyperplans, tout $s \in S^1 \subseteq \mathcal{S}_n$ a 4 descendants et tout $s \in S^2 \subseteq \mathcal{S}_n$ a 3 descendants. Cela signifie que le nombre total de descendants est $4 \times 2^n + 3 \times 2(3^n - 2^n) = 2 \times 3^{n+1} - (6-4)2^n = 2 \times 3^{n+1} - 2^{n+1}$.

Pour cela, rappelons qu'en ajoutant un hyperplan $v^\perp$ à un arrangement $\mathcal{A}(\{v_i\}_i)$, un vecteur de signes s a deux descendants si et seulement si la région associée est divisée par l'hyperplan, c'est-à-dire qu'il existe d^s tel que $s_i v_i^\mathsf{T} d^s > 0, v^\mathsf{T} d^s = 0$. En suivant le même raisonnement, en ajoutant un second hyperplan, s peut avoir 4 descendants si et seulement s'il existe un d^s à l'intérieur de la région sur l'intersection des deux hyperplans ajoutés.

Concentrons-nous sur les deux nouveaux hyperplans, $(1; 0_n; 1)^\perp$ et $(1; 0_n; -1)^\perp$, c'est-à-dire $\{d \in \mathbb{R}^{n+2} : d_0 + d_{n+1} = 0\}$,= et $\{d \in \mathbb{R}^{n+2} : d_0 - d_{n+1} = 0\}$. Leur intersection est $\{d \in \mathbb{R}^{n+2} : d_0 = 0 = d_{n+1}\}$. Justifions qu'il y a précisément 2^n chambres de $\mathcal{S}_n$ autour de cette intersection. Soit $s \in \mathcal{S}_n$, son système d'inéquations est

$$\begin{cases} \forall\, i \in [1:n], & s_i(d_0 + d_i) > 0, \\ \forall\, i \in [1:n], & s_{i+n}(d_0 - d_i) > 0. \end{cases}$$

Maintenant, les chambres qui sont autour de l'intersection des deux nouveaux hyperplans doivent avoir des directions d avec $d_0 = 0$ et $d_{n+1} = 0$. La coordonnée d_{n+1} n'intervient pas dans le système ci-dessus – c'est à cause de l'indépendance des deux vecteurs ajoutés. Cependant, en ajoutant la contrainte $d_0 = 0$, le système ci-dessus devient

$$\begin{cases} \forall\, i \in [1:n], & s_i d_i > 0, \\ \forall\, i \in [1:n], & -s_{i+n} d_i > 0, \end{cases}$$

ce qui signifie que s_i et s_{i+n} doivent être opposés. Ainsi, $s \in \mathcal{S}_n \subseteq \{\pm 1\}^{2n}$ a quatre descendants si et seulement si $s_{n+1:2n} = -s_{1:n}$: il y a 2^n possibilités. Pour résumer, nous avons montré que "s a 4 descendants $\Rightarrow$ $s_{n+1:2n} = -s_{1:n}$" ; la réciproque est directe en inversant les calculs. Pour terminer cette partie de la preuve, nous devons justifier que les 2^n chambres décrites sont bien dans $\mathcal{S}_n$: leurs systèmes correspondants sont vérifiés dans $\mathbb{R}^{n+1}$ par les vecteurs $(0; w)$ pour $w \in \{\pm 1\}^n$, et par les vecteurs $(0; w; 0)$ dans $\mathbb{R}^{n+2}$.

Maintenant, il faut justifier que les $2(3^n - 2^n)$ chambres restantes sont divisées en 3. Remarquons que ces chambres, par le raisonnement ci-dessus, ne vérifient pas $s_{1:n} = -s_{n+1:2n}$. Ainsi, il existe i tel que $s_i = s_{i+n}$, donc les équations correspondantes $s_i(d_0 + d_i) > 0, s_i(d_0 - d_i) > 0$ signifient que d_0 ne peut pas être 0.

Considérons un tel s pour lequel un $d \in \mathbb{R}^{n+1}$ réalisable a $d_0 > 0$ (par symétrie, le même est vrai si $d_0 < 0$). Maintenant, dans $\mathbb{R}^{n+2}$, la ligne $\{(d; t) : t \in \mathbb{R}\}$ vérifie les $2n$ équations de s, qui sont indépendantes de la coordonnée $n+1$. Mais pour les deux hyperplans ajoutés, le système avec $(-, -)$ ne peut pas être vérifié puisque le système

$$\begin{cases} -d_0 - d_{n+1} > 0, \\ -d_0 + d_{n+1} > 0, \end{cases}$$

est sans solution si $d_0 > 0$. Le système $(+,+)$ est vérifié pour $t = 0$, le système $(+,-)$ pour t assez positif et $(-,+)$ pour t assez négatif.

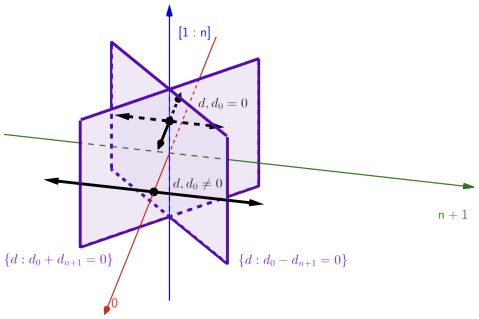


FIGURE 4.2 – Illustration de l'idée du processus de récurrence : en violet les deux hyperplans ajoutés. Le point noir en haut représente un point avec $d_0 = 0$ et $d_{[1:n]}$ arbitraire qui a 4 descendants. Le point noir en bas représente un point avec $d_0 \neq 0$ et $d_{[1:n]} = 0$ qui n'a que 3 descendants (deux flèches plus lui-même).

Pour conclure, la preuve a montré que parmi les $2 \times 3^n - 2^n$ chambres, celles correspondant à des vecteurs de signes de la forme $(s, -s)$ pour un $s \in \{\pm 1\}^n$ ont toutes exactement 4 descendants (dont deux ont la même forme, conduisant à retrouver 2^{n+1}). Pendant ce temps, les $2 \times (3^n - 2^n)$ restantes ne sont divisées qu'en 3 chambres dans $\mathbb{R}^{n+2}$. Cela donne le $2 \times 3^{n+1} - 2^{n+1}$ annoncé. □

Remarque 4.5.7. On pourrait aussi utiliser un argument de comptage similaire, utilisant purement les vecteurs souches. Clairement, les 2^n vecteurs de signes de la forme $(s, -s)$ ne couvrent aucun vecteur souche. Ce qui reste est de compter le nombre de vecteurs de signes non de la forme $(s, -s)$ qui ne couvrent pas de vecteur souche.

D'abord, notons que s'il y a $n+1$ et $n-1$, soit le vecteur de signes est de la forme $(s, -s)$ et donc déjà compté, soit il y a des paires $(i, i+n)$ et $(j, j+n)$ couvrant un vecteur souche (si non de la forme $(s, -s)$, alors il existe un indice i avec $s_i = s_{i+n}$, et comme il y a $n+1$ et $n-1$ il y a un j avec $s_j = s_{j+n}$ avec $s_i = -s_j$). Ainsi, on peut compter les vecteurs de signes ayant $k < n-1$, et par symétrie multiplier par 2 à la fin sera suffisant.

Ensuite, par symétrie, nous considérons le cas où il y a un plus petit nombre de -1. S'il y a $k = 0$ -1, la seule possibilité est 1_{2n}. Pour $k = 1$, il y a $2n$ possibilités – choisir n'importe quel indice pour mettre le -1. Pour $k = 2$, toute possibilité sauf avoir s_i et s_{i+n} égaux à -1, c'est-à-dire $2 \times (d-1) \times (d)$. En poursuivant ce raisonnement, pour k valeurs à -1 on doit dispatcher les k indices à des positions telles qu'il n'y ait aucune paire d'indices $(i, i+n)$ tous deux avec un -1. Cela signifie qu'on choisit k des paires $(i, i+n)$, étiquetées $(i_1, i_1+n), (i_2, i_2+n) \ldots, (i_k, i_k+n)$, et parmi eux change un des signes s_{i_j} ou s_{i_j+n}. Cela donne, pour k changements, $\binom{n}{k} 2^k$.

À partir de là, générer les chambres est direct. Vérifions que nous retrouvons le nombre

total de chambres :

$$\sum_{k=0}^{k=n-1} \binom{n}{k} 2^k = \sum_{k=0}^{k=n} \binom{n}{k} 2^k - 2^n = 3^n - 2^n.$$

Maintenant, par symétrie, on obtient $2 \times (3^n - 2^n)$; en ajoutant les 2^n vecteurs de signes de la forme $(s, -s)$, on obtient $2 \times 3^n - 2^n$ et tout vecteur de signes a été considéré. □

4.5.3 Instances parfaitement symétriques

Dans le chapitre 3 (ainsi que dans le code Julia lié au chapitre suivant), nous rapportons des résultats encourageants particulièrement sur certaines instances, qui tendent à avoir des structures très spécifiques – la matrice V est construite d'une manière très précise, sans valeur aléatoire. Pour terminer ce chapitre, nous discutons de certains travaux futurs possibles qui combineraient des techniques observées dans [35, 212] pour améliorer l'algorithme ISF. Ils sont basés sur la définition suivante, où ici et dans le reste de ce chapitre, π désigne une permutation de $[1 : n]$.

Définition 4.5.8 (instance parfaitement symétrique)**.** Une matrice $V \in \mathbb{R}^{n \times p} = [v_1 \ \dots \ v_p]$ est dite parfaitement symétrique si pour tout $i \in [1 : p]$ et toute permutation π de $[1 : n]$, il existe un indice $j \in [1 : p]$ et un scalaire $\delta_{i,j,\pi}$ tels que $v_i^\pi = \delta_{i,j,\pi} v_j$ où $(v_i^\pi)_k := (v_i)_{\pi(k)}$ pour $k \in [1 : n]$.

Un arrangement d'hyperplans gouverné par les colonnes d'une telle matrice V est également dit parfaitement symétrique. □

Définition 4.5.9 (groupe de symétrie)**.** Soit V une matrice parfaitement symétrique. Pour $i \in [1 : p]$, le groupe de symétrie de i est défini par tous les indices $j \in [1 : p]$ (comptant i lui-même avec la permutation identité) décrits dans la définition 4.5.8. □

En d'autres termes, toutes les dimensions sont "symétriques" ou "équivalentes". Cela se produit par exemple pour les instances PERM, RESONANCE ou DEMICUBE, bien que les instances THRESHOLD aient des propriétés similaires.

Exemple 4.5.10 (instances PERM et RESONANCE)**.** Considérons les données suivantes

$$\text{PERM-4} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & -1 & -1 \end{bmatrix},$$

$$\text{RESONANCE-4} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix},$$

qui sont parfaitement symétriques par vérification routinière. Pour PERM-4, il y a deux groupes de symétrie : $\{1, 2, 3, 4\}$ et $\{5, 6, 7, 8, 9, 10\}$. Pour RESONANCE-4, il y a quatre groupes de symétrie : $\{1, 2, 4, 8\}$, $\{3, 5, 6, 9, 10, 12\}$, $\{7, 11, 13, 14\}$, $\{15\}$. □

Bien qu'il semble difficile d'établir une règle générale, il semble que les types d'instances vérifiant la définition 4.5.8 pourraient être traités d'une manière particulière. Déclinons des possibilités pour des algorithmes utilisant des directions et d'autres utilisant des vecteurs souches.

Pour les vecteurs souches

Cette partie s'inspire directement de [212, section 9]. Quand on veut calculer les circuits d'instances parfaitement symétriques, on peut utiliser les groupes de symétrie de l'arrangement, pour éviter de nombreux calculs de circuits (équivalents).

Illustrons cela sur PERM-4. Considérons le sous-ensemble formé par les colonnes 1, 2 et 5 : c'est clairement un circuit, mais par le groupe de symétrie, pour toute permutation π, il en est de même pour le sous-ensemble avec j^1, j^2 et j^5 (les indices donnés pour 1, 2 et 5 par la définition 4.5.8). Par exemple, avec $\pi = [4; 2; 3; 1]$, c'est-à-dire l'échange des coordonnées 1 et 4, on a

$$e_1^\pi = e_4, e_2^\pi = e_2, (e_1 - e_2)^\pi = -(e_2 - e_4) = -v_9,$$

signifiant que $\{2, 4, 9\}$ est un circuit.

Une fois les circuits obtenus, dans l'algorithme arborescent, on pourrait ainsi vérifier si un sous-vecteur de s est un vecteur souche mais dans une liste beaucoup plus courte qui prendrait en compte les symétries – par exemple, on pourrait garder en mémoire la structure $\{e_i, e_j, e_i - e_j\}$ pour tous ces circuits au lieu des $\binom{n}{2}$ circuits de ce type.

Pour les points témoins

Pour les algorithmes basés sur l'arbre $\mathcal{S}$ et l'optimisation linéaire, de telles instances peuvent avoir une structure exploitable dans $\mathbb{R}^n$. Par exemple, les instances permutahedron divisent l'espace en $n!$ régions "identiques" de la forme (voir section 4.5.1)

$$R_\pi := \{x \in \mathbb{R}^n : x_{\pi(1)} > \cdots > x_{\pi(n)}\}.$$

Ainsi, on pourrait calculer un arbre $\mathcal{S}$ partiel pour une région R_π et obtenir les autres chambres par permutation. Certes, la structure doit être analysée pour chaque type d'instance, mais son idée générale peut être utilisée pour d'autres types d'instances.

Par exemple, pour les instances RESONANCE, il y a une certaine symétrie dans les orthants (au sens classique, $\mathbb{R}^n_+$ et les $2^n - 1$ autres). Il est clair que $\mathbb{R}^n_{++}$ n'est traversé par aucun des hyperplans, donc c'est une chambre (et $\mathbb{R}^n_{--} = -\mathbb{R}^n_{++}$ aussi).

Ensuite, considérons les orthants voisins $\{x \in \mathbb{R}^n : x_i < 0, x_{[1:n]\setminus\{i\}} > 0\}$ pour $i \in [1 : n]$. Leurs décompositions par les hyperplans sont équivalentes car les dimensions peuvent être échangées. Le même raisonnement peut être appliqué pour les orthants avec 2 coordonnées négatives, puis 3 coordonnées négatives et ainsi de suite. Par symétrie, il suffit d'aller jusqu'à $\lfloor \frac{n}{2} \rfloor$, signifiant qu'au lieu de 2^n (2^{n-1} par symétrie des instances linéaires) orthants, seulement $\lfloor \frac{n}{2} \rfloor$ doivent être considérés.

Naturellement, les techniques évoquées dans cette section sont plutôt spécialisées et adaptées à des instances particulières. De plus, elles nécessiteraient des mécanismes avancés pour être implémentées – voir [212, 35].

Chapitre 5

Approches primales et duales pour énumérer les chambres d'arrangements d'hyperplans

Ce chapitre est constitué d'un article en préparation (initialement soumis dans *SIAM Journal on Discrete Mathematics* [79]). Il décrit l'extension du chapitre 3 aux arrangements non centrés.

On y discute de propriétés comme la caractérisation des chambres symétriques, les vecteurs souches (et les circuits, reliés aux matroïdes [151, 191, 141]), les notions de position générale, de bornes sur le nombre de chambres…

Une autre partie du chapitre présente les algorithmes "compacts", qui symétrisent l'arbre $\mathcal{S}$ pour améliorer le code. Ces notions sont implémentées et comparées numériquement.

Ce chapitre va de pair avec l'annexe A, qui contient des détails comme des preuves ou des commentaires supplémentaires. En particulier, on y détaille : certaines identités de la section 5.3, la preuve de la connectivité – comme c'est un cas plus général que dans la proposition 3.4.5, des propriétés sur les instances testées ainsi que des résultats additionnels sur le code Julia.

Par ailleurs, pour un souci d'harmonisation avec le reste, les références sont groupées avec celles de la thèse, et ne sont donc pas ajoutées à la fin de l'article (certaines dates des références comme les classiques de la littérature peuvent être différences de la version publiée). De même, la mise en page, les polices et tailles d'écriture sont différentes.

Note : l'Université de Sherbrooke demande que, pour les articles insérés, la contribution du doctorant soit précisée. La majeure partie du travail a été réalisée de façon commune, en discutant fréquemment pour coordonner les contributions et points de vue. Cet article soumis a été rédigé de façon conjointe via un dépôt Git, de façon à laisser chacun contribuer. Le code Julia obtenant les résultats à la fin du chapitre a été écrit par moi.

Primal and dual approaches for the chamber enumeration of real hyperplane arrangements

Jean-Pierre Dussault [1], Jean Charles Gilbert [2] and Baptiste Plaquevent-Jourdain [3]

Hyperplane arrangements is a problem that appears in various theoretical and applied mathematical contexts. This chapter focuses on the enumeration of the chambers of an arrangement, a task that most often requires algebraic or numerical computation. Among the recent numerical methods, Rada and Černý's recursive algorithm outperforms previous approches, by relying on a specific tree structure and on linear optimization. This chapter presents modifications and improvements to this algorithm. It also introduces a dual approach solely grounded on matroid circuits and its associated concepts of *stem vectors*, thus avoiding the need to solve linear optimization problems. Along the way, theoretical properties of arrangements, such as their cardinality and conditions for their symmetry, completeness and connectivity, as well as properties of their various stem vector sets are presented with an analytic viewpoint. It is shown, in particular, that the set of the chambers of an affine arrangement is located between those of two related linear arrangements. This leads to compact forms of the algorithms, which solve less subproblems. The proposed methods have been implemented in Julia and their efficiency is assessed on various instances of arrangements and manifests itself by speed-up ratios in the range [1.4, 19.3] with an average value of 3.9.

Key words. Duality, hyperplane arrangement, matroid circuit, Motzkin's alternative, Schläfli's bound, stem vector, strict linear inequality system, tree algorithm, Winder's formula.

MSC codes. 05B35, 05C05, 14N20, 49N15, 52B40, 52C35, 52C40, 90C05.

5.1 Introduction

A *hyperplane* of $\mathbb{R}^n$ is a set of the form $H := \{x \in \mathbb{R}^n : v^\mathsf{T} x = \tau\}$, where $v \in \mathbb{R}^n$, $\tau \in \mathbb{R}$ and $v^\mathsf{T} x = \sum_{i=1}^n v_i x_i$ denotes the Euclidean scalar product of v and x. Usually, v is asked to be nonzero, but we have allowed v to vanish to make some formulas more compact below. For $v_1, \ldots, v_p \in \mathbb{R}^n$ and $\tau_1, \ldots, \tau_p \in \mathbb{R}$, consider the collection of hyperplanes

1. J.-P. Dussault, Département d'Informatique, Fac. des Sciences, Univ. de Sherbrooke, Québec, Canada
Jean-Pierre.Dussault@Usherbrooke.ca, ORCID 0000-0001-7253-7462

2. J.Ch. Gilbert, Inria Paris (Serena team), 48 rue Barrault, CS 61534, 75647 Paris Cedex, France
Département de Mathématiques, Fac. des Sciences, Univ. de Sherbrooke, Québec, Canada
Jean-Charles.Gilbert@inria.fr, ORCID 0000-0002-0375-4663

3. B. Plaquevent-Jourdain, Département de Mathématiques, Fac. des Sciences, Univ. de Sherbrooke, Québec, Canada, Inria Paris (Serena team), 48 rue Barrault, CS 61534, 75647 Paris Cedex, France
Baptiste.Plaquevent-Jourdain@USherbrooke.ca, ORCID 0000-0001-7055-4568

$H_i := \{x \in \mathbb{R}^n : v_i^\mathsf{T} x = \tau_i\}$ for $i \in [1 : p]$. The connected parts of the complement of their union, that is $\mathbb{R}^n \setminus (\cup_{i=1}^p H_i)$, are open polyhedrons, called *chambers* or *cells*. *Hyperplane arrangements* is the name given to the discipline that describes this structure [236]. Its study started at least in the 19th century [227, 215, 239] and has continued until the present with theoretical contributions [257, 4, 151, 187], algorithmic developments [81, 118, 208, 77] as well as applications [33] ; see also the references therein. Arrangements can also be stated for complex numbers [186] or over finite fields [61] ; arrangements of circles on a sphere is also a subject of interest, with application in biology [41]. A powerful tool to study arrangements is the characteristic polynomial, which contains much information and provides one way of computing the number of chambers (see for instance [257, 12, 238] ; see the proof of formula (5.27a) below, for a different analytic approach).

This paper focuses on the numerical *enumeration* of the chambers of an arrangement. Several approaches have been designed for that purpose. The algorithm of Bieri and Nef [27] recursively sweeps the space with hyperplanes, decreasing the dimension of the current space in order to explore arrangements in affine spaces of smaller dimension. Edelsbrunner, O'Rourke and Seidel [83] have designed an asymptotically optimal algorithm. The approach of Avis, Fukuda and Sleumer [13, 232] starts with an arbitrary chamber and moves from chamber to neighboring chamber, using a "reverse search" paradigm, thanks to the connectivity of the graph structure of the chambers (proposition 5.3.10 below). Rada and Černý [208] use a more efficient tree, called the $\mathcal{S}$-tree below, obtained by adding hyperplanes incrementally, whereas in previous approaches all hyperplanes are considered from the start. This tree algorithm possesses various interesting properties such as output-polynomiality, meaning that each individual chamber is obtained in polynomial time, and compactness, meaning that the required memory storage is low.

Several pieces of software in algebra or combinatorics, able to deal with arrangements, have been developed : polymake [140], Sagemath [68], Macaulay2 [111], OSCAR [189]. Some related works, such as the package CountingChambers.jl [35], focus on the use of combinatorial symmetries, eventually alongside the deletion-restriction paradigm (see also [253]), to treat arrangements with underlying symmetries and many more hyperplanes. Similar considerations also appear in TOPCOM [214, 212] and yield very good results on particular instances.

Improvements to Rada and Černý's algorithm are proposed and benchmarked in [77]. The authors first present heuristics to bypass some computations. Then, they introduce a dual approach based on Gordan's theorem of the alternative [108], by introducing the notion of *stem vector*, closely related to the *circuits* of a *vector matroid* associated with the arrangement. These modifications allow the authors to significantly reduce the number of linear optimization problems (LOPs) to solve, therefore lowering the computing time, or even to completely remove the need of linear optimization. This paper extends the scope of [77] to arrangements with hyperplanes not necessarily containing the origin. We shall see that the heuristics introduced in [77] have natural extensions in this general case. The

same is true for the dual approach, which is here grounded on Motzkin's alternative [178] ; this one is indeed naturally associated with affine arrangements. These modifications are compared in the penultimate section of the paper.

This contribution is organized as follows. Section 5.2 presents some notation used throughout the paper as well as Motzkin's theorem of the alternative [178], crucial in this paper, which contributes to both theoretical and algorithmic aspects. Section 5.3 starts with the introduction of the concept of *hyperplane arrangements*. Then, it gives conditions ensuring some propreties of the associated sign vector set, like its symmetry and its connectivity. Next, the section introduces the notion of *stem vector*, describes its set, gives its properties and shows how the stem vectors can be used to detect the infeasibility of sign vectors (covering test of proposition 5.3.16). Finally, the role of the *augmented matrix* is discussed. It is shown, in particular, that the sign vector set of an affine arrangement is located between the sign vector sets of two linear arrangements. Information on the number of chambers is also given or recalled, in particular when this one is in *affine general position*.

The rest of the paper focuses on algorithmic issues. Section 5.4 first describes the algorithm of [208], its recursion process and its use of linear optimization. Then, we adapt the heuristic ideas proposed in [77] to affine arrangements, which improves the efficiency of the previous algorithm. Section 5.5 focuses on dual algorithms, which use the stem vectors and often require less computing time. Section 5.6 shows how a compact form of the algorithms can be constructed, taking advantage of the fact that only half of the symmetric sign vectors need to be stored. Often, this technique also allows the compact algorithms to save computing time. Finally, section 5.7 presents the instances used to test the algorithms, their features and some numerical results.

Our presentation is more based on linear algebra and (convex) analysis rather than on discrete geometry or algebra. More specifically, the notion of circuit of a vector matroid and the duality concepts of convex analysis are prominent in sections 5.3, 5.5 and 5.6. In some places, new proofs to known results are proposed with these points of view. This allows the readers with an analytic bent to have easier access to these results.

This paper is an abridged version of the more detailed report [80].

5.2 Background

One denotes by $\mathbb{Z}$, $\mathbb{N}$ and $\mathbb{R}$ the sets of integers, nonnegative integers and real numbers and one sets $\mathbb{N}^* := \mathbb{N} \setminus \{0\}$ and $\mathbb{R}^* := \mathbb{R} \setminus \{0\}$ ($r \in \mathbb{R}$ is said to be *positive* if $r > 0$ and *nonnegative* if $r \geqslant 0$). For two integers $n_1 \leqslant n_2$, $[n_1 : n_2] := \{n_1, \ldots, n_2\}$ is the set of the integers between n_1 and n_2. We denote by $\mathbb{R}^n_+ := \{x \in \mathbb{R}^n : x \geqslant 0\}$ and $\mathbb{R}^n_{++} := \{x \in \mathbb{R}^n : x > 0\}$ the nonnegative and positive orthants, where the inequalities apply componentwise. For a set S, one denotes by $|S|$ its cardinality, by S^c its complement in a set that will be clear from the context and by S^J, for an index set $J \subseteq \mathbb{N}^*$, the set of

vectors, whose elements are in S and are indexed by the indices in J. The vector e denotes the vector of all ones, whose size depends on the context. The Hadamard product of u and $v \in \mathbb{R}^n$ is the vector $u \bullet v \in \mathbb{R}^n$, whose ith component is $u_i v_i$. The sign function $\mathrm{sgn} : \mathbb{R} \to \mathbb{R}$ is defined by $\mathrm{sgn}(t) = +1$ if $t > 0, \mathrm{sgn}(t) = -1$ if $t < 0$ and $\mathrm{sgn}(0) = 0$. The sign of a vector x or a matrix M is defined componentwise : $\mathrm{sgn}(x)_i = \mathrm{sgn}(x_i)$ and $[\mathrm{sgn}(M)]_{i,j} = \mathrm{sgn}(M_{i,j})$ for i and j. For $u \in \mathbb{R}^n$, $|u| \in \mathbb{R}^n$ is the vector defined by $|u|_i = |u_i|$ for all $i \in [1 : n]$. The dimension of a space $\mathbb{E}$ is denoted by $\dim(\mathbb{E})$, the range space of a matrix $A \in \mathbb{R}^{m \times n}$ by $\mathcal{R}(A)$, its null space by $\mathcal{N}(A)$, its rank by $\mathrm{rank}(A) := \dim \mathcal{R}(A)$ and its nullity by $\mathrm{null}(A) := \dim \mathcal{N}(A) = n - \mathrm{rank}(A)$ thanks to the rank-nullity theorem. The ith row (resp. column) of A is denoted by $A_{i,:}$ (resp. $A_{:,i}$). Transposition operates after a row and/or column selection : $A_{i,:}^\mathsf{T}$ is a short notation for $(A_{i,:})^\mathsf{T}$ for instance. The vertical concatenation of matrices $A \in \mathbb{R}^{n_1 \times m}$ and $B \in \mathbb{R}^{n_2 \times m}$ is denoted by $[A; B] \in \mathbb{R}^{(n_1+n_2) \times m}$. For $u \in \mathbb{R}^n$, $\mathrm{Diag}(u) \in \mathbb{R}^{n \times n}$ is the square diagonal matrix with $\mathrm{Diag}(u)_{i,i} = u_i$. The orthogonal of a subspace $Z \subseteq \mathbb{R}^n$ is denoted by $Z^\perp := \{x \in \mathbb{R}^n : x^\mathsf{T} z = 0, \text{ for all } z \in Z\}$.

This article makes extensive use of the so-called (there have been many contributors) Motzkin theorem of the alternative [178] [115, theorem 3.17], abbreviated as *Motzkin's alternative* below, whose following simplified expression is appropriate for our purpose (the general version also includes affine equalities and non strict affine inequalities). Let us write it as an equivalence, rather than an alternative : for a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $a \in \mathbb{R}^m$,

$$\exists\, x \in \mathbb{R}^n \;:\; Ax > a \quad \Longleftrightarrow \quad \nexists\, \alpha \in \mathbb{R}^m_+ \setminus \{0\} \;:\; A^\mathsf{T}\alpha = 0,\; a^\mathsf{T}\alpha \geqslant 0. \tag{5.1}$$

Gordan's theorem of the alternative [108, p. 1873] is recovered when $a = 0$:

$$\exists\, x \in \mathbb{R}^n \;:\; Ax > 0 \quad \Longleftrightarrow \quad \nexists\, \alpha \in \mathbb{R}^m_+ \setminus \{0\} \;:\; A^\mathsf{T}\alpha = 0. \tag{5.2}$$

The latter equivalence satisfies the needs in [77] because the inequality systems encountered in that paper are homogeneous. It will also be helpful below.

The next lemma will be applied several times. It is taken from [77, lemma 2.6] and is a refinement of [255, lemma 2.1]. It is useful to get a discriminating property by a small perturbation of a point.

Lemma 5.2.1 (discriminating covectors)**.** *Suppose that $(\mathbb{E}, \langle\cdot,\cdot\rangle)$ is a Euclidean vector space, $p \in \mathbb{N}^*$ and $v_1, \ldots, v_p$ are p distinct vectors of $\mathbb{E}$. Then, the set of vectors $\xi \in \mathbb{E}$ such that $|\{\langle \xi, v_i\rangle : i \in [1 : p]\}| = p$ is dense in $\mathbb{E}$.*

5.3 Hyperplane arrangements

5.3.1 Presentation

Let $n \in \mathbb{N}^*$. A *hyperplane* of $\mathbb{R}^n$ is a set of the form $H := \{x \in \mathbb{R}^n : v^\mathsf{T} x = \tau\}$, where $v \in \mathbb{R}^n$ and $\tau \in \mathbb{R}$. This hyperplane H is said to be *proper* if $v \neq 0$ and *improper* otherwise.

A proper hyperplane H partitions $\mathbb{R}^n$ into three subsets : H itself and its negative and positive open halfspaces, respectively defined by

$$H^- := \{x \in \mathbb{R}^n : v^\mathsf{T} x < \tau\} \qquad \text{and} \qquad H^+ := \{x \in \mathbb{R}^n : v^\mathsf{T} x > \tau\}.$$

If H is improper and $\tau = 0$, then $H = \mathbb{R}^n$ and $H^- = H^+ = \varnothing$. If H is improper and $\tau \neq 0$, then $H = \varnothing$ and $H^+ = \mathbb{R}^n$ or $\varnothing$, while $H^- = (H^+)^c$.

A *hyperplane arrangement* is a collection of $p \in \mathbb{N}^*$ hyperplanes $H_i := \{x \in \mathbb{R}^n : v_i^\mathsf{T} x = \tau_i\}$, for $i \in [1:p]$, where $v_1, \ldots, v_p \in \mathbb{R}^n$ and $\tau_1, \ldots, \tau_p \in \mathbb{R}$. It is denoted by

$\mathcal{A}(V, \tau)$, where $V := [v_1 \ \cdots \ v_p] \in \mathbb{R}^{n \times p}$ is the matrix made of the vectors v_i's and $\tau := [\tau_1; \ldots; \tau_p] \in \mathbb{R}^{p \times 1}$. The arrangement $\mathcal{A}(V, \tau)$ is said to be *proper* if V has no zero column (i.e., its hyperplanes are proper) and *improper* otherwise (in proposition 5.4.4, a construction may yield a harmless improper arrangement, which is the reason why we introduce this concept). The arrangement is said to be *linear* if $\tau = 0$ and *affine* in general (therefore, a linear arrangement is just a particular affine arrangement). The arrangement is said to be *centered* if all the hyperplanes have a point in common [12], which is the case if and only if $\tau \in \mathcal{R}(V^\mathsf{T})$ (proposition 5.3.5).

Whilst a proper hyperplane divides $\mathbb{R}^n$ into two nonempty open halfspaces, a proper hyperplane arrangement splits $\mathbb{R}^n$ into nonempty polyhedral convex open sets, called *chambers* (the precise definition of a chamber is given below). This is illustrated in figure 5.1 by two elementary examples that will accompany us throughout the paper.

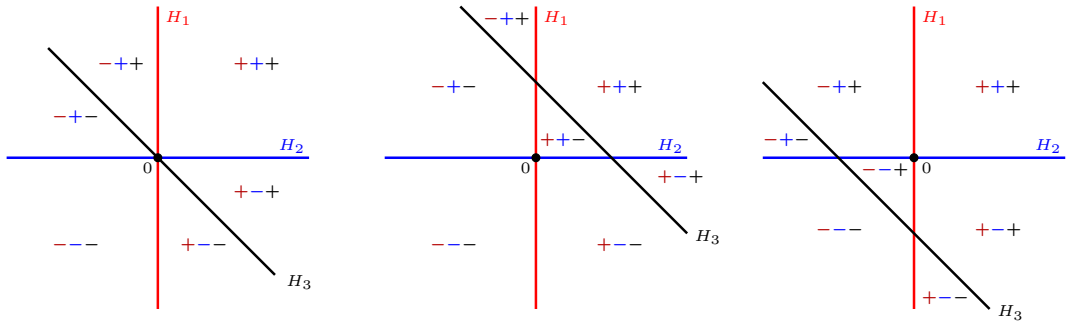


Figure 5.1 – Arrangements in $\mathbb{R}^2$ specified by the hyperplanes $H_1 := \{x \in \mathbb{R}^2 : x_1 = 0\}$, $H_2 := \{x \in \mathbb{R}^2 : x_2 = 0\}$, H_3(left) $:= \{x \in \mathbb{R}^2 : x_1 + x_2 = 0\}$, H_3(middle) $:= \{x \in \mathbb{R}^2 : x_1 + x_2 = 1\}$ and H_3(right) $:= \{x \in \mathbb{R}^2 : x_1 + x_2 = -1\}$. The origin is contained in all the hyperplanes but in H_3(middle) and H_3(right), so that the arrangement in the left-hand side is *linear* with 6 chambers and the other ones are *affine* with 7 chambers.

Enumerating the chambers is the problem at hand in this paper and it can be made precise in the following way. Let

$$\mathfrak{B}([1:p]) := \{(I_+, I_-) \in [1:p]^2 : I_+ \cap I_- = \varnothing,\ I_+ \cup I_- = [1:p]\}$$

be the collection of *bipartitions* (i.e., partitions into two subsets) of $[1:p]$. With each bipartition $(I_+, I_-) \in \mathfrak{B}([1:p])$, one can associate the set

$$C(I_+, I_-) := (\cap_{i\in I_+} H_i^+) \cap (\cap_{i\in I_-} H_i^-). \tag{5.3}$$

Some of these 2^p sets may be empty, while we are interested in enumerating the nonempty ones, which are called the *chambers* of the arrangement. The collection of these chambers, indexed by the bipartitions of $[1:p]$, is denoted by

$$\mathfrak{C}(V,\tau) := \{(I_+, I_-) \in \mathfrak{B}([1:p]) : C(I_+, I_-) \neq \varnothing\}. \tag{5.4}$$

When $\mathfrak{C}(V,\tau) = \mathfrak{B}([1:p])$, the arrangement $\mathcal{A}(V,\tau)$ is said to be *complete*.

As shown by the following proposition, this problem is equivalent to determining the sign vectors $s \in \{\pm 1\}^p$ that make a set of strict inequalities feasible. The collection of these sign vectors is denoted by

$$\mathcal{S}(V,\tau) := \{s \in \{\pm 1\}^p : s \bullet (V^\mathsf{T} x - \tau) > 0 \text{ for some } x \in \mathbb{R}^n\}, \tag{5.5}$$

where "$\bullet$" denotes the Hadamard product. A sign vector $s \in \{\pm 1\}^p$ in $\mathcal{S}(V,\tau)$ is said to be *feasible*, while it said to be *infeasible* if it is in the complementary set

$$\mathcal{S}(V,\tau)^c := \{\pm 1\}^p \setminus \mathcal{S}(V,\tau).$$

For $s \in \mathcal{S}(V,\tau)$, a point x verifying the system of strict inequalities in (5.5) is called a *witness point* of s [208]. It is often more convenient to work with these sign vectors $s \in \{\pm 1\}^p$ rather than with the bipartitions (I_+, I_-) of $[1:p]$ and we shall do so in the rest of the paper. To establish the correspondence between the bipartitions (I_+, I_-) of $[1:p]$ and the sign vectors s of $\{\pm 1\}^p$, one uses the following bijection

$$\phi : (I_+, I_-) \in \mathfrak{B}([1:p]) \mapsto s \in \{\pm 1\}^p, \quad \text{where } s_i = \begin{cases} +1 & \text{if } i \in I_+ \\ -1 & \text{if } i \in I_-, \end{cases} \tag{5.6}$$

whose inverse is given by

$$\phi^{-1} : s \in \{\pm 1\}^p \mapsto (\{i \in [1:p] : s_i = +1\}, \{i \in [1:p] : s_i = -1\}) \in \mathfrak{B}([1:p]).$$

Proposition 5.3.1 (chambers and sign vectors)**.** *The map ϕ given by* (5.6) *is a bijection between the chamber set $\mathfrak{C}(V,\tau)$ and the sign vector set $\mathcal{S}(V,\tau)$.*

Proof. Let $(I_+, I_-) \in \mathfrak{B}([1:p])$ and $s := \phi((I_+, I_-))$. One has

$$\begin{aligned} (I_+, I_-) \in \mathfrak{C}(V,\tau) \quad &\iff \quad \exists\, x \in \mathbb{R} :\ v_i^\mathsf{T} x > \tau_i \text{ for } i \in I_+ \text{ and } v_i^\mathsf{T} x < \tau_i \text{ for } i \in I_- \\ &\iff \quad \exists\, x \in \mathbb{R} :\ s \bullet (V^\mathsf{T} x - \tau) > 0 \\ &\iff \quad s \in \mathcal{S}(V,\tau). \end{aligned}$$

This proves the bijectivity of $\phi : \mathfrak{C}(V,\tau) \to \mathcal{S}(V,\tau)$ and concludes the proof. □ A

consequence of this proposition is that it is equivalent to determine the chamber set $\mathfrak{C}(V,\tau)$ (geometric viewpoint) or the sign vector set $\mathcal{S}(V,\tau)$ (analytic viewpoint).

By this proposition, an arrangement $\mathcal{A}(V,\tau)$ is complete if and only if $\mathcal{S}(V,\tau) = \{\pm 1\}^p$. Note that the proposition does not assume that the hyperplanes are different. Observe also that an arrangement with identical hyperplanes is not complete.

When the hyperplanes are linear (i.e., $\tau = 0$), the description of $\mathcal{S}(V,\tau)$ has various reformulations, sometimes in very different areas of mathematics : see [77] for some of them and the references therein for others.

5.3.2 Properties

Hyperplane arrangements benefit from a myriad of properties. In this section, we mention and prove some of them, which are relevant for the enumeration of chambers. Most of these properties extend, with adjustments, to affine arrangements those that are valid for linear arrangements, in particular those presented in [77]. For further developments and different viewpoints, see for instance [237, 257, 4, 187].

Let $V = [v_1 \ \cdots \ v_p] \in \mathbb{R}^{n\times p}$, $\tau = (\tau_1, \ldots, \tau_p) \in \mathbb{R}^p$ and $r := \operatorname{rank}(V)$. In the sequel, we consider the arrangement $\mathcal{A}(V,\tau)$ formed by the p hyperplanes $H_i := \{x \in \mathbb{R}^n : v_i^\mathsf{T} x = \tau_i\}$ for $i \in [1:p]$.

The next proposition gives conditions characterizing the fact that two hyperplanes are parallel or identical (two hyperplanes H and $\tilde{H}$ are said to be *parallel* if they have the same parallel subspace, that is, if $H - H = \tilde{H} - \tilde{H}$). Discarding identical hyperplanes is important to simplify the task of the algorithms enumerating the chambers and the proposition explains how to detect them from the columns of the matrix $[V;\tau^\mathsf{T}]$. Identical hyperplanes prevent an arrangement from being connected (proposition 5.3.10) and from being in general position (definitions 5.3.25 and 5.3.29). Below, we say that two vectors v and $\tilde{v} \in \mathbb{R}^n$ are *colinear* if there is an $\alpha \in \mathbb{R}^*$ such that $\tilde{v} = \alpha v$ (hence v and $\tilde{v}$, or $x \mapsto x^\mathsf{T} v$ and $x \mapsto x^\mathsf{T}\tilde{v}$ vanish simultaneously).

Proposition 5.3.2 (parallel and identical hyperplanes)**.** *Let $H = \{x \in \mathbb{R}^n : v^\mathsf{T} x = \tau\}$ and $\tilde{H} = \{x \in \mathbb{R}^n : \tilde{v}^\mathsf{T} x = \tilde{\tau}\}$ be two nonempty hyperplanes. Then,*

1) *H and $\tilde{H}$ are parallel if and only if v and $\tilde{v}$ are colinear in $\mathbb{R}^n$,*
2) *$H = \tilde{H}$ if and only if (v,τ) and $(\tilde{v},\tilde{\tau})$ are colinear in $\mathbb{R}^n \times \mathbb{R}$.*

The next proposition identifies some modifications of (V,τ) that have no effect on the sign vector set $\mathcal{S}(V,\tau)$. For a matrix M, we denote by $\operatorname{sgn}(M)$ the matrix defined by $[\operatorname{sgn}(M)]_{i,j} = \operatorname{sgn}(M_{i,j})$ for all i, j. Point 1 of the next proposition is related to proposition 5.3.2(2). A proof of the proposition is given in [80].

Proposition 5.3.3 (equivalent arrangements)**.** *Let $V \in \mathbb{R}^{n\times p}$ and $\tau \in \mathbb{R}^p$.*

1) *If $D \in \mathbb{R}^{p\times p}$ is a nonsingular diagonal matrix, then $\mathcal{S}(VD, D\tau) = \operatorname{sgn}(D)\mathcal{S}(V,\tau)$.*
2) *If $M \in \mathbb{R}^{m\times n}$, then $\mathcal{S}(MV,\tau) \subseteq \mathcal{S}(V,\tau)$ with equality if M is injective.*

A consequence of proposition 5.3.3(2) is that, as far as the sign vector set $\mathcal{S}(V,\tau)$ is concerned, the rank r of $V \in \mathbb{R}^{n\times p}$ is more relevant than its row dimension n. Indeed, $r \leqslant n$ and when $r < n$, one can ignore $n-r$ dependent rows of V, without modifying $\mathcal{S}(V,\tau)$. More specifically, assuming that the last $n-r$ rows of V are linearly dependent of its first r rows, one can write

$$V = \begin{bmatrix} I_r \\ A \end{bmatrix} V_{[1:r],:},$$

for some matrix $A \in \mathbb{R}^{(n-r)\times p}$. Since $[I_r; A]$ is injective, one has $\mathcal{S}(V,\tau) = \mathcal{S}(V_{[1:r],:},\tau)$ by proposition 5.3.3(2). Now, the dimensions of $V_{[1:r],:}$ do not involve n, so that this presentation indicates that the role of n is not very relevant and explains why many results below show r instead of n.

Now that we have identified the chamber set $\mathfrak{C}(V,\tau)$ with the sign vector set $\mathcal{S}(V,\tau)$ (proposition 5.3.1), we are led to the introduction of the notion of symmetry, which naturally presents itself in $\{\pm 1\}^p$.

Definition 5.3.4 (symmetric sign vector set)**.** A set of sign vectors $S \subseteq \{\pm 1\}^p$ with $p \in \mathbb{N}^*$ is said to be *symmetric* if $-S = S$; otherwise, it is said *asymmetric*. For a given set $S \subseteq \{\pm 1\}^p$, one says that $s \in \{\pm 1\}^p$ is *symmetric in* S if $\pm s \in S$.

This notion of symmetry intervenes in the design of the algorithms computing $\mathcal{S}(V,\tau)$. In particular, when this set is symmetric, only half of it need to be computed. Using the definition (5.5) of $\mathcal{S}(V,\tau)$, it follows immediately that

$$\mathcal{S}(V,0) \text{ is symmetric.} \tag{5.7}$$

The next proposition shows that this symmetry property occurs for $\mathcal{S}(V,\tau)$ if and only if the arrangement is centered.

Proposition 5.3.5 (symmetry characterization)**.** *Let $\mathcal{A}(V,\tau)$ be a proper arrangement. Then, the following properties are equivalent :*

(i) *$\mathcal{S}(V,\tau)$ is symmetric,*
(ii) *$\tau \in \mathcal{R}(V^{\mathsf{T}})$,*
(iii) *the arrangement is centered.*

Proof. $[(i) \Rightarrow (ii)]$ One can decompose τ as follows :

$$\tau = \tau^0 + V^{\mathsf{T}}\hat{x}, \tag{5.8a}$$

where $\tau^0 \in \mathcal{N}(V)$ and $\hat{x} \in \mathbb{R}^n$. We pursue by contraposition, assuming that $\tau^0 \neq 0$. Hence $I := \{i \in [1:p] : \tau^0_i \neq 0\}$ is nonempty. Define $s \in \{\pm 1\}^p$ by

$$s_I := \operatorname{sgn}(\tau^0_I), \tag{5.8b}$$

while s_{I^c} is defined below in order to get

$$s \notin \mathcal{S}(V,\tau) \qquad \text{and} \qquad -s \in \mathcal{S}(V,\tau).$$

These properties suffice to prove the implication "$(i) \Rightarrow (ii)$". Set $S_I := \mathrm{Diag}(s_I)$.

To prove that $s \notin \mathcal{S}(V,\tau)$, whatever $s_{I^c} \in \{\pm 1\}^{I^c}$ is, observe that $\alpha_I := |\tau_I^0| \in \mathbb{R}_+^I \setminus \{0\}$ verifies

$$V_{:,I} S_I \alpha_I = 0 \qquad \text{and} \qquad (\tau_I^0)^\mathsf{T} S_I \alpha_I = \|\tau_I^0\|_2^2 \geqslant 0.$$

By Motzkin's alternative (5.1) with $A = S_I V_{:,I}^\mathsf{T}$ and $a = S_I \tau_I^0$, this is equivalent to

$$\nexists\, x \in \mathbb{R}^n : \quad S_I V_{:,I}^\mathsf{T} x > S_I \tau_I^0.$$

Hence, whatever $s_{I^c} \in \{\pm 1\}^{I^c}$ is, there is no $x \in \mathbb{R}^n$ such that $s \cdot (V^\mathsf{T} x - \tau^0) > 0$. Now, using (5.8a), we see that there is no $x \in \mathbb{R}^n$ such that $s \cdot (V^\mathsf{T} x - \tau) > 0$, which proves $s \notin \mathcal{S}(V,\tau)$.

Let us now show that $-s \in \mathcal{S}(V,\tau)$, for some s_{I^c} to specify. Observe that there is no $\alpha_I \in \mathbb{R}_+^I \setminus \{0\}$ such that

$$-V_{:,I} S_I \alpha_I = 0 \qquad \text{and} \qquad -|\tau_I^0|^\mathsf{T} \alpha_I \geqslant 0$$

because the last inequality, with $\alpha_I \geqslant 0$ and $|\tau_I^0| > 0$, implies that $\alpha_I = 0$. By Motzkin's alternative (5.1) with $A = -S_I V_{:,I}^\mathsf{T}$ and $a = -|\tau_I^0| = -S_I \tau_I^0$, this is equivalent to

$$\exists\, x \in \mathbb{R}^n : \quad -S_I V_{:,I}^\mathsf{T} x > -S_I \tau_I^0. \tag{5.8c}$$

Since the columns of V are nonzero, a small perturbation of x can maintain (5.8c) and ensures that the components of $V_{:,I^c}^\mathsf{T} x$ are nonzero (use, for example, the discriminating lemma 3.2.6 with the zero vector and the distinct v_i's with $i \in I^c$). Next, choosing $s_{I^c} := -\operatorname{sgn}(V_{:,I^c}^\mathsf{T} x)$ and setting $S_{I^c} := \mathrm{Diag}(s_{I^c})$ leads to

$$-S_{I^c} V_{:,I^c}^\mathsf{T} x > 0 = -S_{I^c} \tau_{I^c}^0. \tag{5.8d}$$

Thanks to (5.8c) and (5.8d), there is an $x \in \mathbb{R}^n$ such that $-s \cdot (V^\mathsf{T} x - \tau^0) > 0$. Now, using (5.8a), we see that there is an $x \in \mathbb{R}^n$ such that $-s \cdot (V^\mathsf{T} x - \tau) > 0$, which proves that $-s \in \mathcal{S}(V,\tau)$.

$[(ii) \Leftrightarrow (iii)]$ Property (ii) is equivalent to the existence of $\hat{x} \in \mathbb{R}^n$ such that $V^\mathsf{T} \hat{x} = \tau$, which is itself equivalent to the fact that the hyperplanes have the point $\hat{x}$ in common, which means that the arrangement is centered.

$[(ii) \Rightarrow (i)]$ Let $\hat{x} \in \mathbb{R}^n$ be such that $V^\mathsf{T} \hat{x} = \tau$. If $s \in \mathcal{S}(V,\tau)$, then $s \cdot (V^\mathsf{T} x - \tau) > 0$ for some $x \in \mathbb{R}^n$, hence $-s \cdot (V^\mathsf{T}(2\hat{x} - x) - \tau) = -s \cdot (V^\mathsf{T}(-x) + \tau) > 0$, implying that $-s \in \mathcal{S}(V,\tau)$. □

It is short to prove that

$$\mathcal{S}(V,-\tau) = -\mathcal{S}(V,\tau). \tag{5.9}$$

Let us define the *symmetric* and *asymmetric parts* of $\mathcal{S}(V,\tau)$ by

$$\mathcal{S}_s(V,\tau) := \mathcal{S}(V,\tau) \cap \mathcal{S}(V,-\tau) \qquad \text{and} \qquad \mathcal{S}_a(V,\tau) := \mathcal{S}(V,\tau) \setminus \mathcal{S}_s(V,\tau). \tag{5.10}$$

Clearly, by (5.9), $\pm s \in \mathcal{S}(V,\tau)$ when $s \in \mathcal{S}_s(V,\tau)$, while $-s \notin \mathcal{S}(V,\tau)$ when $s \in \mathcal{S}_a(V,\tau)$. This justifes the names given to $\mathcal{S}_s(V,\tau)$ and $\mathcal{S}_a(V,\tau)$. One also observes that (see [80])

$$\mathcal{S}_s(V,-\tau) = -\mathcal{S}_s(V,\tau) = \mathcal{S}_s(V,\tau) \qquad \text{and} \qquad \mathcal{S}_a(V,-\tau) = -\mathcal{S}_a(V,\tau). \tag{5.11}$$

Proposition 5.3.6 (symmetry in $\boldsymbol{\mathcal{S}(V,\tau)}$)**.** 1) $\mathcal{S}(V,0) \subseteq \mathcal{S}(V,\tau)$*, with equality if and only if* $\mathcal{S}(V,\tau)$ *is symmetric,*
2) $\mathcal{S}_s(V,\tau) = \mathcal{S}(V,0)$.

Proof. 1a) Let us first show that $\mathcal{S}(V,0) \subseteq \mathcal{S}(V,\tau)$. Let $s \in \mathcal{S}(V,0)$, so that $s \bullet (V^{\mathsf{T}}x) > 0$ for some $x \in \mathbb{R}^n$. Then, $s \bullet (V^{\mathsf{T}}(tx) - \tau) > 0$ for t large enough, implying that $s \in \mathcal{S}(V,\tau)$.

2) [$\subseteq$] If $s \in \mathcal{S}_s(V,\tau)$, one has $\pm s \in \mathcal{S}(V,\tau)$ and there are points x and $\tilde{x}$ such that $s \bullet (V^{\mathsf{T}}x - \tau) > 0$ and $-s \bullet (V^{\mathsf{T}}\tilde{x} - \tau) > 0$. After adding these inequalities side by side, one gets $s \bullet (V^{\mathsf{T}}(x - \tilde{x})) > 0$, i.e., $s \in \mathcal{S}(V,0)$. [$\supseteq$] Use $\mathcal{S}(V,0) \subseteq \mathcal{S}(V,\tau)$ (1a) for τ and $-\tau$.

1b) If $\mathcal{S}(V,0) = \mathcal{S}(V,\tau)$, $\mathcal{S}(V,\tau)$ is symmetric since so is $\mathcal{S}(V,0)$ by (5.7). Conversely, if $\mathcal{S}(V,\tau)$ is symmetric, then $\mathcal{S}(V,\tau) = -\mathcal{S}(V,\tau) = \mathcal{S}(V,-\tau)$ by (5.9), so that $\mathcal{S}(V,\tau) = \mathcal{S}_s(V,\tau)$; next $\mathcal{S}(V,\tau) = \mathcal{S}(V,0)$ follows from point 2. □

As a corollary of the previous proposition, one has for a matrix $V \in \mathbb{R}^{n\times p}$ of rank r and a vector $\tau \in \mathbb{R}^p$ such that $\mathcal{S}(V,\tau) \neq \varnothing$:

$$2^r \leqslant |\mathcal{S}(V,\tau)| \leqslant 2^{\check{p}}, \tag{5.12}$$

where $\check{p} := |\{i \in [1:p] : V_{:,i} \neq 0\}|$.

Proof. Let $I := \{i \in [1:p] : V_{:,i} = 0\}$. By the assumption $\mathcal{S}(V,\tau) \neq \varnothing$, τ_I has no zero component and any $s \in \mathcal{S}(V,\tau)$ verifies $s_I := -\operatorname{sgn}(\tau_I)$. Therefore, one can only consider the components of the sign vectors that are not in I, which amounts to assuming that $I = \varnothing$. For the lower bound, one has $\mathcal{S}(V,\tau) \supseteq \mathcal{S}(V,0)$ by proposition 5.3.6(1) and $|\mathcal{S}(V,0)| \geqslant 2^r$ by (3.36a). The upper bound is clear since, by definition, coordinates I^c of $\mathcal{S}(V,\tau)$ are included in $\{\pm 1\}^{\check{p}}$, which has cardinality $2^{\check{p}}$. □

More precise lower and upper bounds on $|\mathcal{S}(V,0)|$ and $|\mathcal{S}(V,\tau)|$ are given in propositions 5.3.26, 5.3.31 as well as in (5.39) below. Proposition 5.3.6 tells us that $\mathcal{S}(V,\tau)$ is symmetric if and only if it is invariant with respect to $\tau \in \mathbb{R}^p$ (since it is then equal to $\mathcal{S}(V,0)$ whatever τ is). In the same spirit, one can give yet another characterization of the symmetry of $\mathcal{S}(V,\tau)$, now in terms of the invariance of the chamber existence with respect to $\tau \in \mathbb{R}^p$ and $s \in \mathcal{S}(V,\tau)$, provided that V has no vanishing column. Let us introduce the possibly empty sets, associated with $s \in \{\pm 1\}^p$, denoted by

$$C_\tau(s) := \{x \in \mathbb{R}^n : s \bullet (V^{\mathsf{T}}x - \tau) > 0\},$$

which are in one to one correspondence with the sets denoted by $C(I_1, I_2)$ in (5.3), thanks to the map ϕ defined in proposition 5.3.1. Denote by $C_\tau(s)^\infty := \{d \in \mathbb{R}^n : x + \mathbb{R}_+ d \subseteq C_\tau(s)\}$ the asymptotic cone of $C_\tau(s)$ (called recession cone in [221, § 8]). One has

$$C_\tau(s)^\infty = \{d \in \mathbb{R}^n : s \bullet (V^\mathsf{T} d) \geqslant 0\}.$$

Its interior reads

$$\operatorname{int} C_\tau(s)^\infty = \{d \in \mathbb{R}^n : s \bullet (V^\mathsf{T} d) > 0\} = C_0(s), \tag{5.13}$$

The announced invariance results is the following. For a proof, see [80].

Proposition 5.3.7 (symmetry/chamber existence invariance)**.** *Let $\mathcal{A}(V, \tau)$ be a proper arrangement. Then, the following properties are equivalent :*

(i) *$\mathcal{S}(V, \tau)$ is symmetric,*
(ii) *$\operatorname{int} C_\tau(s)^\infty \neq \varnothing$ for all $s \in \mathcal{S}(V, \tau)$,*
(iii) *$C_0(s) \neq \varnothing$ for all $s \in \mathcal{S}(V, \tau)$,*
(iv) *$C_{\tau'}(s) \neq \varnothing$ for all $\tau' \in \mathbb{R}^p$ and all $s \in \mathcal{S}(V, \tau)$.*

The notions of adjacency and connectivity presented below are crucial in some approaches for computing $\mathcal{S}(V, \tau)$ [13, 232] and are related to *graph theory*.

Definition 5.3.8 (adjacency in $\{\pm 1\}^p$)**.** Two sign vectors s^1 and $s^2 \in \{\pm 1\}^p$ are said to be *adjacent* if they differ by a single component. □

Definition 5.3.9 (connectivity in $\{\pm 1\}^p$)**.** A *path* of *length* l in $S \subseteq \{\pm 1\}^p$ is a finite set of sign vectors $s^0, \ldots, s^l \in S$ such that s^k and s^{k+1} are adjacent for all $k \in [0 : l-1]$; in which case the path is said to be joining s^0 to s^l in S. One says that a subset $S \subseteq \{\pm 1\}^p$ is connected if any pair of elements of S can be joined by a path in S. □

One can transfer the notions of adjacency and connectivity in $\{\pm 1\}^p$ (resp. $\mathcal{S}(V, \tau)$) to $\mathfrak{B}([1 : p])$ (resp. $\mathfrak{C}(V, \tau)$), thanks to the bijection ϕ defined by (5.6) and proposition 5.3.1, thus providing a geometric point of view : two chambers are adjacent if their sets I_+ (or I_-) differ by a single index, which means that they are on either side of (or separated by) a single hyperplane ; while connectivity means that one can join any two chambers by a continuous path in $\mathbb{R}^n$ that never crosses two or more hyperplanes simultaneously.

The next proposition indicates that, provided the hyperplanes are all different (see proposition 5.3.2(2) for an anytical expression of this property), the sign vectors of an arrangement form a connected set. The proof is similar to the one [77, proposition 4.5], see [80].

Proposition 5.3.10 (connectivity of $\mathcal{S}$)**.** *The set $\mathcal{S}(V, \tau)$ of sign vectors of a proper affine arrangement is connected if and only if its hyperplanes are all different. In this case, any elements s and $\tilde{s}$ of $\mathcal{S}(V, \tau)$ can be joined by a path in $\mathcal{S}(V, \tau)$ of length $l := \sum_{i \in [1:p]} |\tilde{s}_i - s_i|/2 \leqslant p$ and there is no path in $\mathcal{S}(V, \tau)$ joining s and $\tilde{s}$ of smaller length.*

5.3.3 Stem vectors

The notion of *stem vector* has been rediscovered in [77] for a linear arrangement $\mathcal{A}(V, 0)$ (a similar notion is presented in [263, § 6.2] under the name of *signed circuit*) and it is extended in this section to an affine arrangement $\mathcal{A}(V, \tau)$. It is based on the notion of *circuit* of the vector matroid formed by the columns of V and its subsets of linearly independent columns. It is useful to determine algebraically the complement

$$\mathcal{S}^c := \{\pm 1\}^p \setminus \mathcal{S}$$

of the sign vector set $\mathcal{S} \equiv \mathcal{S}(V, \tau)$ in $\{\pm 1\}^p$. Indeed, as we shall see, a stem vector is a particular sign vector in $\{\pm 1\}^J$ for some $J \subseteq [1 : p]$ and proposition 5.3.16 below will tell us that a sign vector s is in $\mathcal{S}^c$ if and only if s_J is a stem vector for some $J \subseteq [1 : p]$. This property is used throughout sections 5.5 and 5.6. It also results immediately that, knowing the stem vectors, it is possible to generate completely $\mathcal{S}^c$ (algorithm 5.5.3). Here are the details.

Recall that a *circuit* of the *vector matroid* defined by the columns of $V \in \mathbb{R}^{n\times p}$ and its subsets of linear independent columns [191, proposition 1.1.1] is formed of the indices of a set of columns of V that are linearly dependent, whose strict subsets are the indices of linearly independent columns of V [191, proposition 1.3.5(iii)] In compact mathematical terms, the collection $\mathcal{C} \equiv \mathcal{C}(V)$ of the circuits associated with the matrix $V \in \mathbb{R}^{n\times p}$ is defined by

$$\mathcal{C}(V) := \{J \subseteq [1 : p] : J \neq \varnothing,\ \mathrm{null}(V_{:,J}) = 1,\ \mathrm{null}(V_{:,J_0}) = 0 \text{ for all } J_0 \subsetneq J\}, \tag{5.14}$$

where "null" denotes the nullity (i.e., the dimension of the null space) and "$\subsetneq$" denotes strict inclusion. The stem vectors are defined from the circuits of V, with the desire to validate proposition 5.3.16 below. Recall that, with our notation, a sign vector $\sigma \in \{\pm 1\}^J$ for some $J := \{j_1, \ldots, j_{|J|}\} \subseteq [1 : p]$ is a vector $(\sigma_{j_1}, \ldots, \sigma_{j_{|J|}})$ where the σ_j's are in $\{-1, +1\}$.

Note that an index set $J \subseteq [1 : p]$ verifying $\mathrm{null}(V_{:,J}) = 1$ is not necessarily a circuit of V but we have, nevertheless, the following property (see [77, proposition 3.11]), which will be used several times.

Lemma 5.3.11 (matroid circuit detection)**.** *Suppose that $I \subseteq [1 : p]$ is such that $\mathrm{null}(V_{:,I}) = 1$ and that $\alpha \in \mathcal{N}(V_{:,I}) \setminus \{0\}$. Then, $J := \{i \in I : \alpha_i \neq 0\}$ is a matroid circuit of V and the unique one included in I.*

Definition 5.3.12 (stem vector)**.** A *stem vector* of the arrangement $\mathcal{A}(V, \tau)$ is a sign vector $\sigma \in \{\pm 1\}^J$ for some $J \subseteq [1 : p]$ satisfying

$$\begin{cases} J \in \mathcal{C}(V) \\ \sigma = \mathrm{sgn}(\eta) \text{ for some } \eta \in \mathbb{R}^J \text{ verifying } \eta \in \mathcal{N}(V_{:,J}) \setminus \{0\} \text{ and } \tau_J^{\mathsf{T}}\eta \geqslant 0. \end{cases} \tag{5.15}$$

A stem vector is said to be *symmetric* if $\tau_J^{\mathsf{T}}\eta = 0$ and *asymmetric* otherwise (these properties do not depend on the chosen vector η, as shown in remark 5.3.13(3) below). We denote

respectively by

$$\mathfrak{S}(V,\tau), \qquad \mathfrak{S}_s(V,\tau) \qquad \text{and} \qquad \mathfrak{S}_a(V,\tau) := \mathfrak{S}(V,\tau) \setminus \mathfrak{S}_s(V,\tau)$$

the sets of stem vectors, symmetric stem vectors and asymmetric stem vectors of the arrangement $\mathcal{A}(V,\tau)$. We denote by $\mathfrak{J} : \mathfrak{S}(V,\tau) \to \mathcal{C}(V)$ the map that associates with a stem vector $\sigma \in \{\pm 1\}^J$ its circuit $J := \mathfrak{J}(\sigma)$. □

These definitions deserve some explanations and comments.

Remarks 5.3.13. 1) When the arrangement is linear ($\tau = 0$), one recovers definition 3.9 in [77].

2) The circuits are defined from V, while the stem vectors are defined from $[V;\tau^{\mathsf{T}}]$; the latter depend on τ, which is not the case of the former.

3) *A calculation method from* $\mathcal{C}(V)$. One can associate with a circuit $J \in \mathcal{C}(V)$, either one asymmetric stem vector or two symmetric stem vectors (there are no other possibilities). Take indeed a circuit $J \in \mathcal{C}(V)$. Then, by (5.14), $\mathrm{null}(V_{:,J}) = 1$ and any $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ has no zero component (since $\mathrm{null}(V_{:,J_0}) = 0$ for all $J_0 \subsetneq J$). Therefore, $\mathrm{sgn}(\eta) \in \{\pm 1\}^J$ for any $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$. Now, there may be two complementary cases.

- (a) Either $\tau_J \in \mathcal{N}(V_{:,J})^{\perp}$, in which case $\tau_J^{\mathsf{T}}\eta = 0$ for all $\eta \in \mathcal{N}(V_{:,J})$ and, according to (5.15), there are two symmetric and opposite stem vectors associated with J, namely $\pm\,\mathrm{sgn}(\eta_0)$ for some arbitrary $\eta_0 \in \mathcal{N}(V_{:,J}) \setminus \{0\}$.
- (b) Or $\tau_J \notin \mathcal{N}(V_{:,J})^{\perp}$, in which case $\tau_J^{\mathsf{T}}\eta \neq 0$ for some $\eta \in \mathcal{N}(V_{:,J})$ and, actually, for all $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ since $\mathrm{null}(V_{:,J}) = 1$. In this case, there is a single asymmetric stem vector associated with J, namely $\mathrm{sgn}(\eta_+)$, for some $\eta_+ \in \mathcal{N}(V_{:,J})$ such that $\tau_J^{\mathsf{T}}\eta_+ > 0$.

We have shown, in particular, that the symmetry (resp. asymmetry) property of a stem vector does not depend on the choice of $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ (resp. satisfying $\tau_J^{\mathsf{T}}\eta > 0$).

4) The stem vectors may have different sizes, because the circuits may have different sizes.

5) The sets $\mathfrak{S}(V,\tau)$, $\mathfrak{S}_s(V,\tau)$ and $\mathfrak{S}_a(V,\tau)$ are neither vector spaces nor groups. However, given a stem vector $\sigma \in \{\pm 1\}^J$, one can consider $-\sigma$ as the opposite of σ in $\{\pm 1\}^J$, so that $-\sigma \in \{\pm 1\}^J$ (with the same J). With this meaning given to $-\sigma$, one defines

$$-\,\mathfrak{S}(V,\tau) := \{-\sigma \in \{\pm 1\}^J : \sigma \in \mathfrak{S}(V,\tau) \text{ and } J := \mathfrak{J}(\sigma)\}. \tag{5.16}$$

Proposition 5.3.14(1) below claims that $\sigma \in \mathfrak{S}_s(V,\tau)$ when $\pm\sigma \in \mathfrak{S}(V,\tau)$, which justifies a posteriori the qualifier "symmetric" given to the stem vectors in $\mathfrak{S}_s(V,\tau)$.

6) A matrix $V \in \mathbb{R}^{n\times p}$ of rank r has at most $\binom{p}{r+1}$ circuits and this bound is reached if and only if the columns of V are in linear general position (definition 5.3.25 below) [70]; in that case, the circuits are exactly the selections of $r+1$ columns of V. This number can be exponential in p.

7) For $j \in [1:p]$, one has $\{j\} \in \mathcal{C}(V)$ if and only if $V_{:,j} = 0$. If $J \in \mathcal{C}(V)$ and $|J| \geqslant 2$, $V_{:,J}$ has no zero column. □

It is easy to see that [80]

$$-\mathfrak{S}(V,\tau) = \mathfrak{S}(V,-\tau) \qquad \text{and} \qquad -\mathfrak{S}_a(V,\tau) = \mathfrak{S}_a(V,-\tau). \tag{5.17}$$

Here are some more properties of the stem vectors, which are direct consequences of their definition. The properties stated in proposition 5.3.14 can be symbolically represented like in figure 5.2. In the next proposition, we use the symbol " $\uplus$ " for the disjoint union of sets.

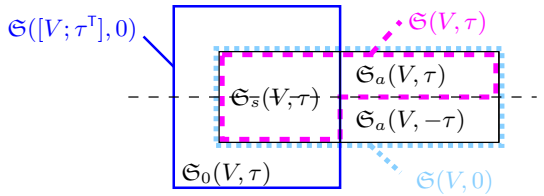


FIGURE 5.2 – Symbolic representation of the sets $\mathfrak{S}(V,\tau)$, $\mathfrak{S}_s(V,\tau)$, $\mathfrak{S}_a(V,\tau)$, $\mathfrak{S}(V,0)$, $\mathfrak{S}_0(V,\tau)$ and $\mathfrak{S}([V;\tau^{\mathsf{T}}],0)$, respecting propositions 5.3.14, 5.3.21 and 5.3.23. The horizontal dashed line aims at representing the reflexion between a stem vector σ and its opposite $-\sigma$: $\mathfrak{S}_s(V,\tau)$, $\mathfrak{S}(V,0)$, $\mathfrak{S}_0(V,\tau)$ and $\mathfrak{S}([V;\tau^{\mathsf{T}}],0)$ are symmetric in the sense that $-\mathfrak{S}_s(V,\tau) = \mathfrak{S}_s(V,\tau)$, $-\mathfrak{S}(V,0) = \mathfrak{S}(V,0)$, $-\mathfrak{S}_0(V,\tau) = \mathfrak{S}_0(V,\tau)$ and $-\mathfrak{S}([V;\tau^{\mathsf{T}}],0) = \mathfrak{S}([V;\tau^{\mathsf{T}}],0)$. By propositions 5.3.15 and 5.3.21, the diagram simplifies when $\tau \in \mathcal{R}(V^{\mathsf{T}})$, since then $\mathfrak{S}_a(V,\tau) = \mathfrak{S}_a(V,-\tau) = \mathfrak{S}_0(V,\tau) = \varnothing$ and there is only one set left.

Proposition 5.3.14 (stem vector properties). *Let $V \in \mathbb{R}^{n\times p}$ and $\tau \in \mathbb{R}^p$. Then,*

1) $\mathfrak{S}(V,\tau) \cap \mathfrak{S}(V,-\tau) = \mathfrak{S}_s(V,\tau) = \mathfrak{S}_s(V,-\tau)$,
2) $\mathfrak{S}(V,\tau) \cup \mathfrak{S}(V,-\tau) = \mathfrak{S}(V,0)$,
3) $\mathfrak{S}_s(V,\tau) \uplus \mathfrak{S}_a(V,\tau) \uplus \mathfrak{S}_a(V,-\tau) = \mathfrak{S}(V,\tau) \uplus \mathfrak{S}_a(V,-\tau) = \mathfrak{S}(V,0)$.

Proof. 1) The last equality can be deduced from the first one, so that only the latter needs to be proved.

[$\subseteq$] Let $\sigma \in \mathfrak{S}(V,\tau) \cap \mathfrak{S}(V,-\tau)$. Then, on the one hand, $\sigma = \operatorname{sgn}(\eta) \in \{\pm 1\}^J$ for some $J \in \mathcal{C}(V)$ and some $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ verifying $\tau_J^{\mathsf{T}}\eta \geqslant 0$ and, on the other hand, $-\sigma = \operatorname{sgn}(\tilde{\eta}) \in \{\pm 1\}^J$ (the same J, see remark 5.3.13(5)) for some $\tilde{\eta} \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ verifying $\tau_J^{\mathsf{T}}\tilde{\eta} \geqslant 0$. Since $\operatorname{null}(V_{:,J}) = 1$ by (5.14), $\tilde{\eta} = \alpha\eta$ for some $\alpha \in \mathbb{R}^*$. Then, $-\sigma = \operatorname{sgn}(\tilde{\eta}) = \operatorname{sgn}(\alpha)\operatorname{sgn}(\eta) = \operatorname{sgn}(\alpha)\sigma$ shows that $\operatorname{sgn}(\alpha) = -1$, so that $0 \leqslant \tau_J^{\mathsf{T}}\tilde{\eta} = \alpha(\tau_J^{\mathsf{T}}\eta)$. Hence $\tau_J^{\mathsf{T}}\eta \leqslant 0$, so that $\tau_J^{\mathsf{T}}\eta = 0$ and σ is symmetric.

[$\supseteq$] Since $\mathfrak{S}_s(V,\tau) \subseteq \mathfrak{S}(V,\tau)$, it suffices to show that $\mathfrak{S}_s(V,\tau) \subseteq \mathfrak{S}(V,-\tau)$ or $-\mathfrak{S}_s(V,\tau) \subseteq \mathfrak{S}(V,\tau)$ by (5.17). If $\sigma \in \mathfrak{S}_s(V,\tau)$ and $J := \mathfrak{J}(\sigma)$, one has $J \in \mathcal{C}(V)$,

$\sigma = \operatorname{sgn}(\eta)$ for some $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ verifying $\tau_J^\mathsf{T}\eta = 0$. Then, clearly, $-\sigma = \operatorname{sgn}(-\eta)$ with $-\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ verifying $\tau_J^\mathsf{T}(-\eta) = 0$. Therefore, $-\sigma \in \mathfrak{S}(V,\tau)$.

2) [$\subseteq$] It suffices to show that $\mathfrak{S}(V,\tau) \subseteq \mathfrak{S}(V,0)$ for an arbitrary τ, which is quite clear since a stem vector $\sigma := \operatorname{sgn}(\eta) \in \mathfrak{S}(V,\tau)$ with $J := \mathfrak{J}(\sigma)$ must satisfy one more property (namely $\tau_J^\mathsf{T}\eta \geqslant 0$) than those in $\mathfrak{S}(V,0)$.

[$\supseteq$] Let $\sigma \in \mathfrak{S}(V,0)$ and $J = \mathfrak{J}(\sigma)$. Then, $\sigma := \operatorname{sgn}(\eta) \in \{\pm 1\}^J$ for some $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$. We see that $\sigma \in \mathfrak{S}(V,\tau)$ if $\tau_J^\mathsf{T}\eta > 0$, $\sigma \in \mathfrak{S}(V,-\tau)$ if $\tau_J^\mathsf{T}\eta < 0$ and both sets $\mathfrak{S}(V,\tau) \cap \mathfrak{S}(V,-\tau) = \mathfrak{S}_s(V,\tau)$ if $\tau_J^\mathsf{T}\eta = 0$.

3) Let us first show that the sets in the left-hand side are disjoint. The sets $\mathfrak{S}_s(V,\tau)$ and $\mathfrak{S}_a(V,\tau)$ are disjoint by their definition. By point 1, $\mathfrak{S}_s(V,-\tau) = \mathfrak{S}_s(V,\tau)$, so that $\mathfrak{S}_s(V,\tau)$ and $\mathfrak{S}_a(V,-\tau)$ are disjoint by their definition. Next, one cannot find an $s \in \mathfrak{S}_a(V,\tau) \cap \mathfrak{S}_a(V,-\tau)$, since s would be in $\mathfrak{S}_s(V,\tau)$ by point 1, which is in contradiction with $s \in \mathfrak{S}_a(V,\tau)$.

Now, the first equality follows from $\mathfrak{S}(V,\tau) = \mathfrak{S}_s(V,\tau) \uplus \mathfrak{S}_a(V,\tau)$ and the second equality follows from $\mathfrak{S}_s(V,\tau) \cup \mathfrak{S}_a(V,\tau) = \mathfrak{S}(V,\tau)$, $\mathfrak{S}_s(V,\tau) \cup \mathfrak{S}_a(V,-\tau) = \mathfrak{S}_s(V,-\tau) \cup \mathfrak{S}_a(V,-\tau) = \mathfrak{S}(V,-\tau)$ and point 2. □

In complement to the characterizations of the centered arrangements in propositions 5.3.5 (symmetry of $\mathcal{S}$) and 5.3.7 (chamber existence invariance), the following characterization is given in terms of the absence of asymmetric stem vector.

Proposition 5.3.15 (centered arrangement and symmetric stem vector set)**.** *For an affine hyperplane arrangement, the following properties are equivalent :*

(i) *the arrangement is centered,*
(ii) *all the stem vectors are symmetric.*

Proof. For $J \subseteq [1:p]$ and $\eta \in \mathbb{R}^J$, we denote by $\overline{\eta} \in \mathbb{R}^p$ the extended vector associated with η that is defined by $\overline{\eta}_j = \eta_j$ for $j \in J$ and $\overline{\eta}_j = 0$ for $j \notin J$.

[$(i) \Rightarrow (ii)$] If the arrangement is centered, one has $\tau \in \mathcal{R}(V^\mathsf{T}) = \mathcal{N}(V)^\perp$ by proposition 5.3.5. A stem vector is of the form $\operatorname{sgn}(\eta)$ with $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ and $\tau_J^\mathsf{T}\eta \geqslant 0$ for some $J \in \mathcal{C}(V)$. Then, the extended vector $\overline{\eta}$ is in $\mathcal{N}(V)$, so that $\tau_J^\mathsf{T}\eta = \tau^\mathsf{T}\overline{\eta} = 0$ (since $\tau \in \mathcal{N}(V)^\perp$ and $\overline{\eta} \in \mathcal{N}(V)$), showing that the stem vector is symmetric.

[$(ii) \Rightarrow (i)$] If all the stem vectors are symmetric, then $\tau_J^\mathsf{T}\eta = 0$ for all $J \in \mathcal{C}(V)$ and $\eta \in \mathcal{N}(V_{:,J})$. If $\overline{\eta}$ extends such an η and if we show that these $\overline{\eta}$'s generate $\mathcal{N}(V)$, we will have $\tau \in \mathcal{N}(V)^\perp = \mathcal{R}(V^\mathsf{T})$, implying that the arrangement is centered (proposition 5.3.5).

Let $r := \operatorname{rank}(V)$, so that $\dim \mathcal{N}(V) = p - r$. To conclude the proof, it suffices to find $p - r$ vectors $\eta \in \mathcal{N}(V_{:,J})$, with $J \in \mathcal{C}(V)$, so that their extensions $\overline{\eta}$ are linearly independent.

By definition of the rank, one can find a set of r linearly independent columns of V ; let J_0 denote the set of their indices. Denote the other column indices by $\{j_1, \dots, j_{p-r}\} := [1 : p] \setminus J_0$ and set $J'_i := J_0 \cup \{j_i\}$ for $i \in [1 : p-r]$. From lemma 5.3.11, J'_i contains a unique circuit, which is denoted by $J_i \in \mathcal{C}(V)$, and $\eta_i \in \mathcal{N}(V_{:,J_i}) \setminus \{0\}$ has no zero component. Necessarily, j_i is in J_i (since, otherwise, $J_i \subseteq J_0$, in which case $V_{:,J_i}$ would be injective and J_i would not be a circuit) and j_i is not in $J_{i'}$ for $i' \in [1 : p-r] \setminus \{i\}$ (j_i is not in $J'_{i'}$ by construction, hence not in $J_{i'} \subseteq J'_{i'}$). Hence, the vectors $\overline{\eta}_i$ extending $\eta_i \in \mathcal{N}(V_{:,J_i})$, $i \in [1 : p]$, are linearly independent in $\mathbb{R}^p$. □

The previous proposition can be rephrased in many ways, in particular as the following equivalence

$$\tau \in \mathcal{R}(V^\mathsf{T}) \qquad \Longleftrightarrow \qquad \mathfrak{S}(V, \tau) = \mathfrak{S}(V, 0). \tag{5.18}$$

The following proposition extends naturally [77, proposition 3.16] to the affine arrangements considered in this paper. The possibility of having the equivalence (5.19) was a certificate for the appropriateness of the proposed definition 5.3.12 of stem vector. The role of this equivalence is important in the design of algorithms having a dual aspect, like those developed in section 5.5. The proof of the proposition is grounded on duality, via Motzkin's alternative (5.1).

Proposition 5.3.16 (generating $\mathcal{S}^c$ from the stem vectors)**.** *For $s \in \{\pm 1\}^p$,*

$$s \in \mathcal{S}(V, \tau)^c \quad \Longleftrightarrow \quad s_J \in \mathfrak{S}(V, \tau) \text{ for some } J \subseteq [1 : p]. \tag{5.19}$$

Proof. [⇒] Take $s \in \mathcal{S}(V, \tau)^c$. Our goal is to show that the index set $J \subseteq [1 : p]$ in the right-hand side of (5.19) can be determined as one satisfying the following two properties :

$$\{x \in \mathbb{R}^n : s_j(v_j^\mathsf{T} x - \tau_j) > 0 \text{ for all } j \in J\} = \varnothing, \tag{5.20a}$$

$$\forall\, J_0 \subsetneq J,\ \{x \in \mathbb{R}^n : s_j(v_j^\mathsf{T} x - \tau_j) > 0 \text{ for all } j \in J_0\} \neq \varnothing. \tag{5.20b}$$

To show that a J satisfying (5.20a) and (5.20b) exists, let us start with $J = [1 : p]$, which verifies (5.20a), since $s \in \mathcal{S}(V, \tau)^c$. Next, remove one index j from $[1 : p]$ if (5.20a) holds for $J = [1 : p] \setminus \{j\}$. Pursuing the elimination of indices j in this way, one finally obtain an index set J satisfying (5.20a) and $\{x \in \mathbb{R}^n : s_j(v_j^\mathsf{T} x - \tau_j) > 0 \text{ for all } j \in J \setminus \{j_0\}\} \neq \varnothing$ for all $j_0 \in J$. Then, (5.20b) clearly holds.

We claim that, for a J satisfying (5.20a) and (5.20b), s_J is a stem vector, which will conclude the proof of the implication.

To show that s_J, with J verifying (5.20a)-(5.20b), is a stem vector, we stick to definition 5.3.12 and start by showing that J is a matroid circuit. By (5.20a), $J \neq \varnothing$. Next, by Motzkin's alternative (5.1) with $A := \mathrm{Diag}(s_J) V_{:,J}^\mathsf{T}$ and $a := s_J \cdot \tau_J$, (5.20a) and (5.20b) read

$$\exists\, \alpha \in \mathbb{R}_+^J \setminus \{0\} \text{ such that } V_{:,J}(s_J \cdot \alpha) = 0,\ \tau_J^\mathsf{T}(s_J \cdot \alpha) \geqslant 0, \tag{5.20c}$$

$$\forall\, J_0 \subsetneq J,\ \nexists\, \alpha' \in \mathbb{R}_+^{J_0} \setminus \{0\} \text{ such that } V_{:,J_0}(s_{J_0} \cdot \alpha') = 0,\ \tau_{J_0}^\mathsf{T}(s_{J_0} \cdot \alpha') \geqslant 0. \tag{5.20d}$$

From these properties, one deduces that $\alpha > 0$ ($\alpha \geqslant 0$ by (5.20c) and α has no zero component since otherwise (5.20d) would not hold) and that $\mathrm{null}(V_{:,J}) \geqslant 1$ ($s_J \cdot \alpha \in \mathcal{N}(V_{:,J}) \setminus \{0\}$). To show that $\mathrm{null}(V_{:,J}) = 1$, we proceed by contradiction. Suppose that there is a nonzero $\alpha'' \in \mathbb{R}^J$ that is not colinear with α and that verifies $V_{:,J}(s_J \cdot \alpha'') = 0$. Since α and α'' are nonzero and not colinear, they have at least two components and one can find $r \in \mathbb{R}$ such that $\beta := \alpha'' - r\alpha \in \mathbb{R}^J$ has at least one positive and one negative component (take for instance $r := (r_1 + r_2)/2$, where $r_1 := \max\{r \in \mathbb{R} : r\alpha \leqslant \alpha''\} < r_2 := \min\{r \in \mathbb{R} : \alpha'' \leqslant r\alpha\}$). One can also assume that $\tau_J^\mathsf{T}(s_J \cdot \beta) \geqslant 0$ (otherwise replace β by $-\beta$, which also has at least one positive and one negative component; one can check below that this sign inversion has no unpleasant impact on the reasoning). Now, set $t := 1/\max\{-\beta_j/\alpha_j : j \in J\}$, which is positive, and $J_0 := \{j \in J : \alpha_j + t\beta_j > 0\}$, so that $J \setminus J_0 = \{j \in J : \alpha_j + t\beta_j = 0\}$. Using the fact that β has positive components and the definition of t, we see that $\varnothing \neq J_0 \subsetneq J$. Let us introduce $\alpha' := \alpha + t\beta \geqslant 0$, which verifies $\alpha'_j > 0$ for $j \in J_0 \neq \varnothing$ and $\alpha'_j = 0$ for $j \in J \setminus J_0 \neq \varnothing$. Therefore,

$$\begin{aligned}
&V_{:,J_0}(s_{J_0} \cdot \alpha'_{J_0}) \\
&\quad = V_{:,J}(s_J \cdot \alpha') \qquad [\alpha'_{J\setminus J_0} = 0] \\
&\quad = V_{:,J}(s_J \cdot \alpha) + t\, V_{:,J}(s_J \cdot \beta) \qquad [\alpha' := \alpha + t\beta] \\
&\quad = tV_{:,J}(s_J \cdot \alpha'') - rt\, V_{:,J}(s_J \cdot \alpha) \qquad [V_{:,J}(s_J \cdot \alpha) = 0,\ \beta = \alpha'' - r\alpha] \\
&\quad = 0 \qquad [V_{:,J}(s_J \cdot \alpha'') = V_{:,J}(s_J \cdot \alpha) = 0]
\end{aligned}$$

and

$$\begin{aligned}
\tau_{J_0}^\mathsf{T}(s_{J_0} \cdot \alpha'_{J_0}) &= \tau_J^\mathsf{T}(s_J \cdot \alpha') \qquad [\alpha'_{J\setminus J_0} = 0] \\
&= \tau_J^\mathsf{T}(s_J \cdot \alpha) + t\, \tau_J^\mathsf{T}(s_J \cdot \beta) \qquad [\alpha' := \alpha + t\beta] \\
&\geqslant 0 \qquad [\tau_J^\mathsf{T}(s_J \cdot \alpha) \geqslant 0,\ \tau_J^\mathsf{T}(s_J \cdot \beta) \geqslant 0,\ t > 0].
\end{aligned}$$

These last two outcomes are in contradiction with (5.20d), as expected.

To show that $J \in \mathcal{C}$ defined by (5.14), we still have to prove that $V_{:,J_0}$ is injective when $J_0 \subsetneq J$. Equivalently, it suffices to show that any $\beta \in \mathcal{N}(V_{:,J})$ with some zero component vanishes. We proceed by contradiction. If there is a $\beta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ with a zero component, $s_J \cdot \alpha$ and β would be two linearly independent vectors in $\mathcal{N}(V_{:,J})$ (since $s_J \cdot \alpha$ has no zero component), contradicting $\mathrm{null}(V_{:,J}) = 1$.

Now, since $s_J = \mathrm{sgn}(s_J \cdot \alpha)$, since $s_J \cdot \alpha \in \mathcal{N}(V_{:,J})$ and $\tau_J^\mathsf{T}(s_J \cdot \alpha) \geqslant 0$ by (5.20c) and since J is a matroid circuit of V, s_J is a stem vector (definition 5.3.12).

[$\Leftarrow$] Since s_J is a stem vector, it reads $s_J := \mathrm{sgn}(\eta) \in \{\pm 1\}^J$ for some $J \in \mathcal{C}$ and some $\eta \in \mathbb{R}^J$ satisfying $V_{:,J}\eta = 0$ and $\tau_J^\mathsf{T}\eta \geqslant 0$. Then, $\alpha := s_J \cdot \eta = |\eta|$ is in $\mathbb{R}^J_+ \setminus \{0\}$ and verifies $V_{:,J}(s_J \cdot \alpha) = 0$ and $\tau_J^\mathsf{T}(s_J \cdot \alpha) \geqslant 0$. By Motzkin's alternative (5.1), there is no $x \in \mathbb{R}^n$ such that $s_J \cdot (V_{:,J}^\mathsf{T} x - \tau_J) > 0$. Hence, there is certainly no $x \in \mathbb{R}^n$ such that $s \cdot (V^\mathsf{T} x - \tau) > 0$. This means that $s \in \mathcal{S}(V, \tau)^c$. □

We say that $s \in \mathcal{S}(V,\tau)^c$ *covers* a sign vector $\sigma \in \{\pm 1\}^J$ for some $J \subseteq [1:p]$ if $s_J = \sigma$. Given a set of stem vecors $\mathfrak{S}$, checking whether a sign vector s covers some $\sigma \in \mathfrak{S}$ is called below a *covering test*. This operation is an essential step of the dual algorithms of section 5.5.

Remark 5.3.17. One might wonder whether having a sign vector $s \in \{\pm 1\}^p$ such that $\pm s \in \mathcal{S}(V,\tau)^c$ would imply that one has $\pm s_J \in \mathfrak{S}(V,\tau)$ for some $J \in [1:p]$. This implication does not hold. Equivalently, for a given $s \in \{\pm 1\}^p$, the two nonempty sets $\{J \subseteq [1:p] : s_J \in \mathfrak{S}(V,\tau)\}$ and $\{J \subseteq [1:p] : s_J \in -\mathfrak{S}(V,\tau)\}$ may have an empty intersection (note that its union may also differ from $\mathcal{C}$). This is the case, for example, when $V := \left[\begin{smallmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{smallmatrix}\right]$ and $\tau^\mathsf{T} := [0\ 0\ 1\ 2]$ (add one hyperplane perpendicular to $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$ in figure 5.1(middle)). More is said on this situation in the equivalence (5.22) below. □

5.3.4 Augmented matrix

According to definition 3.12, verifying whether $s \in \mathcal{S}(V,\tau)$ amounts to checking whether there is an $x \in \mathbb{R}^n$ such that $s \cdot (V^\mathsf{T} x - \tau) > 0$ or, equivalently, whether there is a pair $(x,\xi) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$s \cdot ([V;\tau^\mathsf{T}]^\mathsf{T}[x;\xi]) > 0 \qquad \text{and} \qquad \xi = -1.$$

The first condition above reads $s \in \mathcal{S}([V;\tau^\mathsf{T}],0)$ and refers to the linear arrangement in $\mathbb{R}^{n+1}$ governed by the *augmented matrix* $[V;\tau^\mathsf{T}]$. This presentation of the problem shows that there must be links between the following sign vector sets and between the following stem vector sets

$$\mathcal{S}(V,0), \qquad \mathcal{S}(V,\tau) \qquad \text{and} \qquad \mathcal{S}([V;\tau^\mathsf{T}],0), \tag{5.21a}$$
$$\mathfrak{S}(V,0), \qquad \mathfrak{S}(V,\tau) \qquad \text{and} \qquad \mathfrak{S}([V;\tau^\mathsf{T}],0). \tag{5.21b}$$

For example, we already know the inclusions $\mathcal{S}(V,0) \subseteq \mathcal{S}(V,\tau)$ and $\mathfrak{S}(V,\tau) \subseteq \mathfrak{S}(V,0)$ from propositions 5.3.6 and 5.3.14(2).

This section aims at identifying a few properties where the augmented matrix $[V;\tau^\mathsf{T}]$ intervenes. In section 5.3.4, some links between the sets in (5.21a) are highlighted. Section 5.3.4 establishes some connections between the circuits of V and $[V;\tau^\mathsf{T}]$, as well as between the stem vector sets in (5.21b). In section 5.3.4, one observes that the identity obtained in proposition 5.3.18(4) makes it easy to deduce a formula for $|\mathcal{S}(V,\tau)|$ (proposition 5.3.24) and a known bound on $|\mathcal{S}(V,\tau)|$ (proposition 5.3.31), which is reached if and only if the arrangement is in *affine general position* (proposition 5.3.28 and definition 5.3.29). The role of the augmented matrix in the specification of the notion of *affine general position* is also pointed out.

Viewing an affine arrangement in $x \in \mathbb{R}^n$ as the intersection of a linear arrangement in $(x,\xi) \in \mathbb{R}^{n+1}$ with the affine space $\{(x,\xi) \in \mathbb{R}^{n+1} : \xi = -1\}$ is called the *method of coning* in [187, definition 1.15].

Sign vectors of the augmented matrix

Recall the definition of $\mathcal{S}_s(V,\tau)$ and $\mathcal{S}_a(V,\tau)$ in (5.10) and the properties (5.11). Some of the properties stated in proposition 5.3.18(1) can be symbolically represented like in figure 5.3. In the next proposition, we use the symbol " $\cup$ " for the disjoint union of sets.

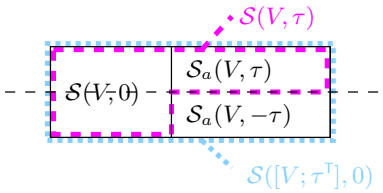


FIGURE 5.3 – Symbolic representation of the sets $\mathcal{S}(V,0)$, $\mathcal{S}(V,\tau)$, $\mathcal{S}_a(V,\tau)$ and $\mathcal{S}([V;\tau^\mathsf{T}],0)$, respecting (5.9), (5.10), (5.11) and propositions 5.3.6 and 5.3.18. The horizontal dashed line aims at representing the reflection between a sign vector s and its opposite $-s$: $\mathcal{S}(V,0)$, $\mathcal{S}([V;\tau^\mathsf{T}],0)$ and $\mathcal{S}([V;\tau^\mathsf{T}],0)^c$ are symmetric in the sense of definition 5.3.4.

Proposition 5.3.18 (properties with $\boldsymbol{\mathcal{S}([V;\tau^\mathsf{T}],0)}$). *Let $\mathcal{A}(V,\tau)$ be an arrangement with $V \in \mathbb{R}^{n\times p}$ and $\tau \in \mathbb{R}^p$. Then, the following properties hold.*

1) $\mathcal{S}(V,\tau) \cap \mathcal{S}(V,-\tau) = \mathcal{S}(V,0) \subseteq \mathcal{S}(V,\tau) \subseteq \mathcal{S}([V;\tau^\mathsf{T}],0) = \mathcal{S}(V,\tau) \cup \mathcal{S}(V,-\tau)$.
2) $\mathcal{S}(V,\tau)^c \cup \mathcal{S}(V,-\tau)^c = \mathcal{S}(V,0)^c \supseteq \mathcal{S}(V,\tau)^c \supseteq \mathcal{S}([V;\tau^\mathsf{T}],0)^c = \mathcal{S}(V,\tau)^c \cap \mathcal{S}(V,-\tau)^c$.
3) $\mathcal{S}(V,0) \cup \mathcal{S}_a(V,\tau) \cup \mathcal{S}_a(V,-\tau) = \mathcal{S}([V;\tau^\mathsf{T}],0)$.
4) $2|\mathcal{S}(V,\tau)| = |\mathcal{S}(V,0)| + |\mathcal{S}([V;\tau^\mathsf{T}],0)|$.
5) $2|\mathcal{S}(V,\tau)^c| = |\mathcal{S}(V,0)^c| + |\mathcal{S}([V;\tau^\mathsf{T}],0)^c|$.

Proof. 1) The first equality repeats proposition 5.3.6(2), using (5.10), the first inclusion repeats proposition 5.3.6(1) and the second inclusion is straightforward : if $s \in \mathcal{S}(V,\tau)$, one has $s \cdot (V^\mathsf{T}x - \tau) > 0$ for some $x \in \mathbb{R}^n$ or $s \cdot ([V;\tau^\mathsf{T}]^\mathsf{T}[x;-1]) > 0$, implying that $s \in \mathcal{S}([V;\tau^\mathsf{T}],0)$.

Consider now the last equality. [$\subseteq$] Let $s \in \mathcal{S}([V;\tau^\mathsf{T}],0)$, so that $s \cdot (V^\mathsf{T}x + \tau\xi) > 0$ for some $(x,\xi) \in \mathbb{R}^n \times \mathbb{R}$. By homogeneity, it follows that $s \in \mathcal{S}(V,\tau)$ if $\xi < 0$, that $s \in \mathcal{S}(V,-\tau)$ if $\xi > 0$ and that $s \in \mathcal{S}(V,0) = \mathcal{S}(V,\tau) \cap \mathcal{S}(V,-\tau)$ if $\xi = 0$. [$\supseteq$] By the second inclusion, $\mathcal{S}(V,\tau) \subseteq \mathcal{S}([V;\tau^\mathsf{T}],0)$ and $\mathcal{S}(V,-\tau) = -\mathcal{S}(V,\tau) \subseteq -\mathcal{S}([V;\tau^\mathsf{T}],0) = \mathcal{S}([V;\tau^\mathsf{T}],0)$.

2) Take the complement of the sets in point 1.

3) Let us first show that the sets are disjoint. By (5.10) and proposition 5.3.6(2), one has $\mathcal{S}_a(V,\pm\tau) = \mathcal{S}(V,\pm\tau) \setminus \mathcal{S}(V,0)$, so that $\mathcal{S}(V,0) \cap \mathcal{S}_a(V,\pm\tau) = \varnothing$. Use the same arguments

to get that

$$\begin{aligned}\mathcal{S}_a(V,\tau)\cap\mathcal{S}_a(V,-\tau) &= [\mathcal{S}(V,\tau)\cap\mathcal{S}(V,0)^c]\cap[\mathcal{S}(V,-\tau)\cap\mathcal{S}(V,0)^c]\\ &= [\mathcal{S}(V,\tau)\cap\mathcal{S}(V,-\tau)]\cap\mathcal{S}(V,0)^c\\ &= \mathcal{S}(V,0)\cap\mathcal{S}(V,0)^c=\varnothing\end{aligned}$$

Consider now the identity :

$$\begin{aligned}\mathcal{S}(V,0)\cup\mathcal{S}_a(V,\tau)\cup\mathcal{S}_a(V,-\tau) &= \mathcal{S}(V,0)\cup[\mathcal{S}(V,\tau)\cap\mathcal{S}(V,0)^c]\cup[\mathcal{S}(V,-\tau)\cap\mathcal{S}(V,0)^c]\\ &= \mathcal{S}(V,0)\cup\mathcal{S}(V,\tau)\cup\mathcal{S}(V,-\tau)\\ &= \mathcal{S}([V;\tau^{\mathsf{T}}],0)\qquad[\text{point 1}].\end{aligned}$$

4) The identity results from

$$\begin{aligned}|\mathcal{S}([V;\tau^{\mathsf{T}}],0)| &= |\mathcal{S}(V,\tau)|+|\mathcal{S}(V,-\tau)|-|\mathcal{S}(V,\tau)\cap\mathcal{S}(V,-\tau)|\qquad[\text{point 1}]\\ &= 2|\mathcal{S}(V,\tau)|-|\mathcal{S}(V,0)|\qquad[\text{(5.9), (5.10), proposition 5.3.6(2)}].\end{aligned}$$

5) Each set in point 4 is a part of $\{\pm1\}^p$ of cardinality 2^p. Hence, the identity in point 4 gives

$$2(2^p-|\mathcal{S}(V,\tau)^c|)=(2^p-|\mathcal{S}(V,0)^c|)+(2^p-|\mathcal{S}([V;\tau^{\mathsf{T}}],0)^c|).$$

Point 5 follows after subtracting 2^{p+1} from both sides. □

Remarks 5.3.19. 1) Let us emphasize the meaning of the inclusions in proposition 5.3.18(1) : the *affine* arrangement $\mathcal{A}(V,\tau)$ has its sign vectors (in bijection with its chambers, by proposition 5.3.1) containing those of the *linear* arrangement $\mathcal{A}(V,0)$ and contained in those of the *linear* arrangement $\mathcal{A}([V;\tau^{\mathsf{T}}],0)$.

2) As a corollary of proposition 5.3.18(2), one has (see remark 5.3.17) :

$$\pm\, s\in\mathcal{S}(V,\tau)^c \qquad\Longleftrightarrow\qquad s_J\in\mathfrak{S}([V;\tau^{\mathsf{T}}],0)\text{ for some }J\subseteq[1:p]. \tag{5.22}$$

Indeed $\pm s\in\mathcal{S}(V,\tau)^c$ if and only if $s\in\mathcal{S}(V,\tau)^c\cap\mathcal{S}(V,-\tau)^c=\mathcal{S}([V;\tau^{\mathsf{T}}],0)^c$ (proposition 5.3.18(2)), which is equivalent to $s_J\in\mathfrak{S}([V;\tau^{\mathsf{T}}],0)$ for some $J\in[1:p]$ (proposition 5.3.16). Note that $\mathfrak{S}([V;\tau^{\mathsf{T}}],0)$ is symmetric, so that the properties in (5.22) are also equivalent to the fact that $s_J\in-\mathfrak{S}([V;\tau^{\mathsf{T}}],0)$ for some $J\subseteq[1:p]$. □

Circuits and stem vectors of the augmented matrix

The next propositions highlight connections between the circuits and the stem vectors of V and those of the augmented matrix $[V;\tau^{\mathsf{T}}]$. Recall from proposition 5.3.5 that an arrangement is centered if and only if $\tau\in\mathcal{R}(V^{\mathsf{T}})$. Note also that

$$\operatorname{rank}([V;\tau^{\mathsf{T}}])=\begin{cases}\operatorname{rank}(V) & \text{if }\tau\in\mathcal{R}(V^{\mathsf{T}})\\ \operatorname{rank}(V)+1 & \text{otherwise,}\end{cases} \tag{5.23a}$$

$$\operatorname{null}([V;\tau^{\mathsf{T}}])=\begin{cases}\operatorname{null}(V) & \text{if }\tau\in\mathcal{R}(V^{\mathsf{T}})\\ \operatorname{null}(V)-1 & \text{otherwise.}\end{cases} \tag{5.23b}$$

The formula of $\mathrm{rank}([V;\tau^\mathsf{T}])$ is clear and the one of $\mathrm{null}([V;\tau^\mathsf{T}])$ can be deduced from (5.23a) by the rank-nullity theorem.

Proposition 5.3.20 (circuits of $\boldsymbol{V}$ and $[\boldsymbol{V};\boldsymbol{\tau}^\mathsf{T}]$)**.** *Let $V \in \mathbb{R}^{n\times p}$ and $\tau \in \mathbb{R}^p$. Then, the following properties are equivalent :*

(i) $\mathcal{C}(V) = \mathcal{C}([V;\tau^\mathsf{T}])$,
(ii) $\mathcal{C}(V) \subseteq \mathcal{C}([V;\tau^\mathsf{T}])$,
(iii) $\tau \in \mathcal{R}(V^\mathsf{T})$, *meaning that that the arrangement $\mathcal{A}(V,\tau)$ is centered.*

Proof.$[(i) \Rightarrow (ii)]$ Clear.

$[(ii) \Rightarrow (iii)]$ Let $J \in \mathcal{C}(V)$. Then, $\mathrm{null}(V_{:,J}) = 1$ by (5.14). By assumption, $J \in \mathcal{C}([V;\tau^\mathsf{T}])$, so that $\mathrm{null}([V;\tau^\mathsf{T}]_{:,J}) = 1$, as well. By (5.23b), $\tau_J \in \mathcal{R}(V_{:,J}^\mathsf{T}) = \mathcal{N}(V_{:,J})^\perp$. According to remark 5.3.13$(3.a)$, the stem vectors associated with J are symmetric. Since J is arbitrary in $\mathcal{C}(V)$, one has $\mathfrak{S}(V,\tau) = \mathfrak{S}_s(V,\tau)$, implying that the arrangement is centered (proposition 5.3.15).

$[(iii) \Rightarrow (i)]$ Let $J \subseteq [1:p]$ and $J_0 \subsetneq J$. When $\tau \in \mathcal{R}(V^\mathsf{T})$, one has $\tau_J \in \mathcal{R}(V_{:,J}^\mathsf{T})$ and $\tau_{J_0} \in \mathcal{R}(V_{:,J_0}^\mathsf{T})$, so that (5.23b) yields

$$\mathrm{null}([V;\tau^\mathsf{T}]_{:,J}) = \mathrm{null}(V_{:,J}) \qquad \text{and} \qquad \mathrm{null}([V;\tau^\mathsf{T}]_{:,J_0}) = \mathrm{null}(V_{:,J_0}).$$

It follows that $\mathrm{null}([V;\tau^\mathsf{T}]_{:,J}) = 1$ and $\mathrm{null}([V;\tau^\mathsf{T}]_{:,J_0}) = 0$ for all $J_0 \subsetneq J$ if and only if $\mathrm{null}(V_{:,J}) = 1$ and $\mathrm{null}(V_{:,J_0}) = 0$ for all $J_0 \subsetneq J$. In other words, $J \in \mathcal{C}([V;\tau^\mathsf{T}])$ if and only if $J \in \mathcal{C}(V)$. We have shown that $\mathcal{C}([V;\tau^\mathsf{T}]) = \mathcal{C}(V)$. □

The implication $(ii) \Rightarrow (i)$ of lemma 5.3.20 is not based on the fact that one would always have $\mathcal{C}(V) \supseteq \mathcal{C}([V;\tau^\mathsf{T}])$, which is not true. As a counter-example, take $V = \left[\begin{smallmatrix}1&0&1&1\\0&1&1&1\end{smallmatrix}\right]$ and $\tau^\mathsf{T} = [0\ 0\ 1\ 2]$, in which case one has $\mathcal{C}(V) = \{\{3,4\},\{1,2,3\},\{1,2,4\}\}$, while $\mathcal{C}([V;\tau^\mathsf{T}]) = \{\{1,2,3,4\}\}$. Actually, the property $\mathcal{C}(V) \supseteq \mathcal{C}([V;\tau^\mathsf{T}])$, which is therefore weaker than those in lemma 5.3.20, has various other equivalent interesting formulations, including $\mathfrak{S}_s(V,\tau) = \mathfrak{S}([V;\tau^\mathsf{T}],0)$, as shown by the following proposition. Recall figure 5.2 for a symbolic representation of the stem vector sets.

Proposition 5.3.21 (stem vectors of $\boldsymbol{\mathcal{A}(V,\tau)}$ and $\boldsymbol{\mathcal{A}([V;\tau^\mathsf{T}],0)}$)**.** *For any V and τ,*

$$\mathfrak{S}_a(V,\tau) \cap \mathfrak{S}([V;\tau^\mathsf{T}],0) = \varnothing \qquad \textit{and} \qquad \mathfrak{S}_s(V,\tau) \subseteq \mathfrak{S}([V;\tau^\mathsf{T}],0), \tag{5.24}$$

with equality in the last inclusion if the arrangement is centered. More precisely, the following properties are equivalent :

(i) $\mathfrak{S}_s(V,\tau) = \mathfrak{S}([V;\tau^\mathsf{T}],0)$,
(ii) $\mathfrak{S}_s(V,\tau) \supseteq \mathfrak{S}([V;\tau^\mathsf{T}],0)$,
(iii) $\mathcal{C}(V) \supseteq \mathcal{C}([V;\tau^\mathsf{T}])$,

(iv) $\tau_J \in \mathcal{R}(V_{:,J}^{\mathsf{T}})$, *for all* $J \in \mathcal{C}([V;\tau^{\mathsf{T}}])$.

Proof. 1) [(5.24)$_1$] Let $\sigma \in \mathfrak{S}([V;\tau^{\mathsf{T}}],0)$ and $J := \mathfrak{J}(\sigma)$. Then, $\sigma = \operatorname{sgn}(\eta)$ for some $\eta \in \mathcal{N}([V;\tau^{\mathsf{T}}]_{:,J}) \setminus \{0\}$. This η verifies $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ and $\tau_J^{\mathsf{T}}\eta = 0$. It follows that $\sigma \notin \mathfrak{S}_a(V,\tau)$, either because $J \notin \mathcal{C}(V)$ or because $J \in \mathcal{C}(V)$, in which case $\sigma \in \mathfrak{S}_s(V,\tau)$ by the properties of η.

[(5.24)$_2$] Let $\sigma \in \mathfrak{S}_s(V,\tau)$ and $J := \mathfrak{J}(\sigma)$. Then, $J \in \mathcal{C}(V)$ and $\sigma = \operatorname{sgn}(\eta)$ for some $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ verifying $\tau_J^{\mathsf{T}}\eta = 0$. It follows that $\eta \in \mathcal{N}([V;\tau^{\mathsf{T}}]_{:,J}) \setminus \{0\}$. To show that $\sigma \in \mathfrak{S}([V;\tau^{\mathsf{T}}],0)$, one still has to verify that $J \in \mathcal{C}([V;\tau^{\mathsf{T}}])$ (definition 5.14). First, $J \neq \varnothing$, since $J \in \mathcal{C}(V)$. Next, $\operatorname{null}([V;\tau^{\mathsf{T}}]_{:,J}) = 1$, since $\eta \in \mathcal{N}([V;\tau^{\mathsf{T}}]_{:,J}) \setminus \{0\}$ and $\operatorname{null}([V;\tau^{\mathsf{T}}]_{:,J}) \leqslant \operatorname{null}(V_{:,J}) = 1$ (because $J \in \mathcal{C}(V)$). Finally, for all $J_0 \subsetneq J$, one has $\operatorname{null}([V;\tau^{\mathsf{T}}]_{:,J_0}) = 0$, since $\operatorname{null}([V;\tau^{\mathsf{T}}]_{:,J_0}) \leqslant \operatorname{null}(V_{:,J_0}) = 0$ (because $J \in \mathcal{C}(V)$).

2) Suppose now that the arrangement $\mathcal{A}(V,\tau)$ is centered (or $\tau \in \mathcal{R}(V^{\mathsf{T}})$) and let us show that $\mathfrak{S}_s(V,\tau) \supseteq \mathfrak{S}([V;\tau^{\mathsf{T}}],0)$ (this could also be viewed as a consequence of the implication $(iv) \Rightarrow (i)$ proved below). Let $\sigma \in \mathfrak{S}([V;\tau^{\mathsf{T}}],0)$ and $J := \mathfrak{J}(\sigma)$. Then, $J \in \mathcal{C}([V;\tau^{\mathsf{T}}])$ and $\sigma = \operatorname{sgn}(\eta)$ for some $\eta \in \mathcal{N}([V;\tau^{\mathsf{T}}]_{:,J}) \setminus \{0\}$. Then, $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ and $\tau_J^{\mathsf{T}}\eta = 0$. We see that it suffices now to observe that $J \in \mathcal{C}(V)$, which results from the implication $(iii) \Rightarrow (i)$ of proposition 5.3.20.

3) Consider now the equivalences (i)-(iv).

[$(i) \Leftrightarrow (ii)$] By (5.24).

[$(ii) \Rightarrow (iii)$] Let $J \in \mathcal{C}([V;\tau^{\mathsf{T}}])$. By remark 5.3.13(3.a), there is a stem vector $\sigma \in \{\pm 1\}^J$ that is in $\mathfrak{S}([V;\tau^{\mathsf{T}}],0)$, hence in $\mathfrak{S}_s(V,\tau)$ by (ii). This latter fact implies that $J \in \mathcal{C}(V)$.

[$(iii) \Rightarrow (iv)$] Let $J \in \mathcal{C}([V;\tau^{\mathsf{T}}])$. By (iii), $J \in \mathcal{C}(V)$, so that $\operatorname{null}([V;\tau^{\mathsf{T}}]_{:,J}) = \operatorname{null}(V_{:,J})$ (these nullities $= 1$). Then, (5.23b) implies that $\tau_J \in \mathcal{R}(V_{:,J}^{\mathsf{T}})$.

[$(iv) \Rightarrow (ii)$] Let $\sigma = \operatorname{sgn}(\eta) \in \mathfrak{S}([V;\tau^{\mathsf{T}}],0)$ and $J := \mathfrak{J}(\sigma)$. Since $\eta \in \mathcal{N}([V;\tau^{\mathsf{T}}]_{:,J}) \setminus \{0\} \subseteq \mathcal{N}(V_{:,J}) \setminus \{0\}$ and $\tau^{\mathsf{T}}\eta = 0$, it suffices to show that $J \in \mathcal{C}(V)$. This property follows from $J \in \mathcal{C}([V;\tau^{\mathsf{T}}]_{:,J})$, since $J \in \mathcal{C}([V;\tau^{\mathsf{T}}])$, from $\mathcal{C}([V;\tau^{\mathsf{T}}]_{:,J}) = \mathcal{C}(V_{:,J})$, by $\tau_J \in \mathcal{R}(V_{:,J}^{\mathsf{T}})$ and the implication $(iii) \Rightarrow (i)$ of proposition 5.3.20, and from $\mathcal{C}(V_{:,J}) \subseteq \mathcal{C}(V)$. □

Examples 5.3.22. 1) An example in which the equivalent properties of proposition 5.3.21 hold, but not those of proposition 5.3.20, is given by $V = \left[\begin{smallmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{smallmatrix}\right]$ and $\tau^{\mathsf{T}} = [0\ 0\ 0\ 1]$. One has

$$\mathcal{C}(V) = \{\{3,4\},\{1,2,3\},\{1,2,4\}\} \qquad \text{and} \qquad \mathcal{C}([V;\tau^{\mathsf{T}}]) = \{\{1,2,3\}\}.$$

We see that $\mathcal{C}([V;\tau^{\mathsf{T}}])$ is included in $\mathcal{C}(V)$ but is not equal to it. Note also that $\tau \notin \mathcal{R}(V^{\mathsf{T}})$,

$$\mathfrak{S}_s(V,\tau) = \mathfrak{S}([V;\tau^{\mathsf{T}}],0) = \left\{ \pm \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} \in \{\pm 1\}^{\{1,2,3\}} \right\} \quad \text{and}$$

$$\mathfrak{S}_a(V,\tau) = \left\{ \begin{bmatrix} -1 \\ 1 \end{bmatrix} \in \{\pm 1\}^{\{3,4\}}, \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix} \in \{\pm 1\}^{\{1,2,4\}} \right\}.$$

All this is in agreement with propositions 5.3.20 and 5.3.21.

2) Property (iv) of proposition 5.3.21 does not imply that $\tau_K \in \mathcal{R}(V^{\mathsf{T}}_{:,K})$, for $K := \cup\{J \in \mathcal{C}([V;\tau^{\mathsf{T}}])\}$. To see this, take $V = \left[\begin{smallmatrix} 1&0&1&1&0&1 \\ 0&1&1&0&1&1 \end{smallmatrix}\right]$ and $\tau^{\mathsf{T}} = [0\ 0\ 1\ 0\ 0\ 1]$. One has $\mathcal{C}([V;\tau^{\mathsf{T}}]) = \{\{1,4\},\{2,5\},\{3,6\}\}$, $\tau_J \in \mathcal{R}(V^{\mathsf{T}}_{:,J})$ for all $J \in \mathcal{C}([V;\tau^{\mathsf{T}}])$ and $K = [1:6]$, but $\tau \notin \mathcal{R}(V^{\mathsf{T}})$. □

Recall (5.3.21)$_2$. In the algorithm of section 5.6.3, it will be interesting to partition $\mathfrak{S}([V;\tau^{\mathsf{T}}],0)$ in $\mathfrak{S}_s(V,\tau)$ and $\mathfrak{S}_0(V,\tau)$, where

$$\mathfrak{S}_0(V,\tau) := \mathfrak{S}([V;\tau^{\mathsf{T}}],0) \setminus \mathfrak{S}_s(V,\tau).$$

The stem vectors of $\mathfrak{S}_0(V,\tau)$ can be recognized thanks to the following proposition.

Proposition 5.3.23 ($\mathfrak{S}_0(V,\tau)$ and $\mathfrak{S}_s(V,\tau)$ characterizations). *Let $\mathcal{A}(V,\tau)$ be an arrangement with $V \in \mathbb{R}^{n\times p}$ and $\tau \in \mathbb{R}^p$. Let $\sigma \in \mathfrak{S}([V;\tau^{\mathsf{T}}],0)$ and $J = \mathfrak{J}(\sigma)$. Then,*

$$\sigma \in \mathfrak{S}_0(V,\tau) \iff \begin{cases} J \in \mathcal{C}([V;\tau^{\mathsf{T}}]) \setminus \mathcal{C}(V) \\ \sigma = \operatorname{sgn}(\eta) \text{ for some } \eta \in \mathcal{N}([V;\tau^{\mathsf{T}}]_{:,J}) \setminus \{0\}. \end{cases} \tag{5.25a}$$

$$\sigma \in \mathfrak{S}_s(V,\tau) \iff J \in \mathcal{C}(V) \iff \operatorname{null}(V_{:,J}) = 1 \iff \tau_J \in \mathcal{R}(V^{\mathsf{T}}_{:,J}). \tag{5.25b}$$

Sign vector set cardinality

Proposition 5.3.18(4) and Winder's formula of $|\mathcal{S}(V,0)|$ (linear arrangement) make it possible to give expressions of $|\mathcal{S}(V,\tau)|$ and $|\mathcal{S}_a(V,\tau)|$ having the flavor of Winder's. Formula (5.27a) below is given by Zaslavsky [257, corollary 5.9, p. 68], who makes its connection with a cardinality formula using a characteristic polynomial of the arrangement [257, theorem A, p. 18] ; an approach that looks rather different from ours.

Recall that, for a matrix $V \in \mathbb{R}^{n\times p}$ without zero column, *Winder's formula* of the cardinality of $\mathcal{S}(V,0)$ reads [253, p. 1966] (see also [77, §4.2.1])

$$|\mathcal{S}(V,0)| = \sum_{J\subseteq[1:p]} (-1)^{\operatorname{null}(V_{:,J})}, \tag{5.26}$$

where the term in the right-hand side corresponding to $J = \varnothing$ is 1 (one takes the convention that $\operatorname{null}(V_{:,\varnothing}) = 0$).

Proposition 5.3.24 (cardinality of $\mathcal{S}(V,\tau)$). *Consider a proper affine arrangement $\mathcal{A}(V,\tau)$ with $V \in \mathbb{R}^{n\times p}$ and $\tau \in \mathbb{R}^p$. Then,*

$$|\mathcal{S}(V,\tau)| = \sum_{\substack{J\subseteq[1:p]\\ \tau_J\in\mathcal{R}(V_{:,J}^\mathsf{T})}} (-1)^{\mathrm{null}(V_{:,J})}, \tag{5.27a}$$

$$|\mathcal{S}_a(V,\tau)| = \sum_{\substack{J\subseteq[1:p]\\ \tau_J\notin\mathcal{R}(V_{:,J}^\mathsf{T})}} (-1)^{\mathrm{null}(V_{:,J})-1}, \tag{5.27b}$$

where the term in the right-hand side of (5.27a) *corresponding to $J=\varnothing$ is* 1 (*it is considered that $\tau_J \in \mathcal{R}(V_{:,J}^\mathsf{T})$ for $J=\varnothing$*).

Proof. Using proposition 5.3.18(4), Winder's formula (5.26) and (5.23b), one gets

$$\begin{aligned}
2|\mathcal{S}(V,\tau)| &= |\mathcal{S}(V,0)| + |\mathcal{S}([V;\tau^\mathsf{T}],0)| \\
&= \sum_{J\subseteq[1:p]} (-1)^{\mathrm{null}(V_{:,J})} + \sum_{J\subseteq[1:p]} (-1)^{\mathrm{null}([V;\tau^\mathsf{T}]_{:,J})} \\
&= 2\sum_{\substack{J\subseteq[1:p]\\ \tau_J\in\mathcal{R}(V_{:,J}^\mathsf{T})}} (-1)^{\mathrm{null}(V_{:,J})} + \sum_{\substack{J\subseteq[1:p]\\ \tau_J\notin\mathcal{R}(V_{:,J}^\mathsf{T})}} \left[(-1)^{\mathrm{null}(V_{:,J})} + (-1)^{\mathrm{null}(V_{:,J})-1}\right] \\
&= 2\sum_{\substack{J\subseteq[1:p]\\ \tau_J\in\mathcal{R}(V_{:,J}^\mathsf{T})}} (-1)^{\mathrm{null}(V_{:,J})},
\end{aligned}$$

since $(-1)^{\mathrm{null}(V_{:,J})}+(-1)^{\mathrm{null}(V_{:,J})-1}=0$ (note that $\mathrm{null}(V_{:,J})>0$ if $\tau_J\notin\mathcal{R}(V_{:,J}^\mathsf{T})$). Formula (5.27a) follows. Formula (5.27b) of $|\mathcal{S}_a(V,\tau)|$ comes from $(5.10)_2$, $\mathcal{S}_s(V,\tau)=\mathcal{S}(V,0)$ and proposition 5.3.18(1), which implies that $|\mathcal{S}_a(V,\tau)| = |\mathcal{S}(V,\tau)| - |\mathcal{S}(V,0)|$. □

Note that one recovers (5.26) from (5.27a) when the arrangement is centered (i.e., $\tau \in \mathcal{R}(V^\mathsf{T})$). By proposition 5.3.18(1), $\mathcal{S}(V,0) \subseteq \mathcal{S}(V,\tau)$, so that $|\mathcal{S}(V,0)| \leqslant |\mathcal{S}(V,\tau)|$, but this inequality is not easy to deduce from (5.26) and (5.27a), since the terms of the sums in the right-hand sides of these formulas may be negative and positive.

Formula (5.27a) is usually not easy to evaluate because the number of terms in the sum can be large. It is therefore sometimes useful to have a bound on $|\mathcal{S}(V,\tau)|$ that is easier to compute than the exact formula. Proposition 5.3.18(4), joined to Schläfli's bound on $|\mathcal{S}(V,0)|$ (linear arrangement), makes it possible to recover a known bound on $|\mathcal{S}(V,\tau)|$ and to clarify the conditions under which this bound is reached. Recall that, for a matrix $V\in\mathbb{R}^{n\times p}$ of rank r, *Schläfli's bound* on the cardinality of $\mathcal{S}(V,0)$ reads [227, p. 211] (see also [77, proposition 4.15])

$$|\mathcal{S}(V,0)| \leqslant 2\sum_{i\in[0:r-1]} \binom{p-1}{i}. \tag{5.28}$$

Winder [253, 1966, corollary] showed that the upper bound in (5.28) is reached if the arrangement $\mathcal{A}(V,0)$ is in *linear general position*, a concept defined as follows.

Definition 5.3.25 (linear general position)**.** Let be given $V \in \mathbb{R}^{n\times p}$ of rank r, without zero column. The linear arrangement $\mathcal{A}(V,0)$ is (or the columns of V are) said to be in *linear general position* if the following equivalent properties hold

$$\begin{aligned} &\forall\, I \subseteq [1:p]: \quad \dim(\cap_{i\in I} H_i^0) = n - \min(|I|, r), \\ &\forall\, I \subseteq [1:p]: \quad \operatorname{rank}(V_{:,I}) = \min(|I|, r), \end{aligned}$$

where $H_i^0 := \{x \in \mathbb{R}^n : V_{:,i}^{\mathsf{T}} x = 0\}$ for $i \in [1:p]$. □

The first condition has a geometric nature, while the second one has an algebraic flavor. Their equivalence comes from the fact that $\dim(\cap_{i\in I} H_i^0) = n - \operatorname{rank}(V_{:,I})$. Observe that the inequality $\operatorname{rank}(V_{:,I}) \leqslant \min(|I|, r)$ always holds. This linear general position property is clearly less restrictive than the injectivity of V, since it holds for an injective V but does not impose $p \leqslant n$. In proposition 3.4.10, it is shown analytically that the linear general position is also necessary to have equality in (5.28). Let us summarize these facts in a proposition.

Proposition 5.3.26 (bound on $|\mathcal{S}(V,0)|$)**.** *Consider a proper linear arrangement $\mathcal{A}(V,0)$, with $V \in \mathbb{R}^{n\times p}$ of rank r. Then,* (5.28) *holds. Furthermore, equality holds in* (5.28) *if and only if the arrangement is in linear general position.*

For instance, the arrangement on the left-hand side pane in figure 5.1 is in linear general position and verifies (5.28) with equality. Also, linear general position generally occurs for a matrix V of rank r that is randomly generated, since then one usually has $\operatorname{rank}(V_{:,I}) = \min(|I|, r)$ for all $I \subseteq [1:p]$.

The general position for an affine arrangement, like the one in the middle and right-hand side panes of figure 5.1, is different from that specified by definition 5.3.25. It is usually given a definition in geometric terms. We are also going to give it an algebraic expression (definition 5.3.29), since this one easily provides necessary and sufficient conditions to have equality in a bound on $|\mathcal{S}(V,\tau)|$ (proposition 5.3.31). The definition is grounded on proposition 5.3.28 below; lemma 5.3.27 will also be useful. The proofs and more discussions can be found in [80].

Lemma 5.3.27 (contribution to the affine general position)**.** *Let $V \in \mathbb{R}^{n\times p}$ of rank r with no zero column and and $\tau \in \mathbb{R}^p$. For the proper affine arrangement $\mathcal{A}(V,\tau)$, the following two conditions are equivalent :*

(i) $|\mathcal{S}([V;\tau^{\mathsf{T}}],0)| = 2\sum_{i\in[0:r]} \binom{p-1}{i}$,
(ii) $\forall\, I \subseteq [1:p]:\ \operatorname{rank}([V;\tau^{\mathsf{T}}]_{:,I}) = \min(|I|, r+1)$.

Proposition 5.3.28 (affine general position)**.** *Let be given $V \in \mathbb{R}^{n\times p}$ of rank r, without zero column, and $\tau \in \mathbb{R}^p$. Set $H_i := \{x \in \mathbb{R}^n : V_{:,i}^{\mathsf{T}} x = \tau_i\}$ for $i \in [1:p]$. Then, the following*

properties are equivalent :

$$\forall\, I \subseteq [1:p]: \quad \begin{cases} \cap_{i\in I} H_i \neq \varnothing \text{ and } \dim(\cap_{i\in I} H_i) = n - |I| & \text{if } |I| \leqslant r \\ \cap_{i\in I} H_i = \varnothing & \text{if } |I| \geqslant r+1, \end{cases} \tag{5.29a}$$

$$\forall\, I \subseteq [1:p]: \quad \begin{cases} \operatorname{rank}(V_{:,I}) = |I| & \text{if } |I| \leqslant r \\ \operatorname{rank}([V;\tau^{\mathsf{T}}]_{:,I}) = r+1 & \text{if } |I| \geqslant r+1, \end{cases} \tag{5.29b}$$

$$\forall\, I \subseteq [1:p]: \quad \begin{cases} \operatorname{rank}(V_{:,I}) = \min(|I|, r) \\ \operatorname{rank}([V;\tau^{\mathsf{T}}]_{:,I}) = \min(|I|, r+1). \end{cases} \tag{5.29c}$$

Definition 5.3.29 (affine general position)**.** A proper affine arrangement $\mathcal{A}(V,\tau)$ is said to be in *affine general position* if the equivalent properties of proposition 5.3.28 hold. □

Remarks 5.3.30. 1) The two conditions in (5.29c) are independent of each other. For the arrangement in the left-hand side pane of figure 5.1, the first condition in (5.29c) holds (linear general position) but not the second one For the arrangement defined by $V = \left[\begin{smallmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{smallmatrix}\right]$ and $\tau^{\mathsf{T}} = [0\ 0\ 1]$, the second condition in (5.29c) holds but not the first one.

2) Even for a linear proper arrangement, we make a difference between *linear general position* (definition 5.3.25) and *affine general position* (definition 5.3.29), since an arrangement $\mathcal{A}(V,0)$ can be considered as a linear arrangement or an affine arrangement $\mathcal{A}(V,\tau)$ with $\tau = 0$. We see on (5.29c) and from the first remark that the notion of affine general position is more restrictive than the notion of linear general position since it requires one more independent condition.

3) A centered proper arrangement $\mathcal{A}(V,\tau)$, with $V \in \mathbb{R}^{n\times p}$ of rank r, can be in affine general position, only if $p = r$ (take $I = [1:p]$ in (5.29c)$_2$).

4) Affine general position usually holds when V of rank r and τ are randomly generated. □

Condition (5.29a) is the one that is usually given to define the affine general position of an affine arrangement $\mathcal{A}(V,\tau)$ [237, p. 287] ; it has a geometric nature. Condition (5.29c) is the form that suits the needs of the proof of the following proposition. We have not found elsewhere the fact that the affine general position is necessary to have equality in (5.30) (for equality in (5.30) when the arrangement is in general position, see [257, (5.7)$_1$]). The right-hand side of (5.30) is sequence A008949 in [185].

Proposition 5.3.31 (bound on $|\mathcal{S}(\boldsymbol{V},\boldsymbol{\tau})|$)**.** *Let $\mathcal{A}(V,\tau)$ be a proper arrangements with $V \in \mathbb{R}^{n\times p}$ of rank r and $\tau \in \mathbb{R}^p$. Then,*

$$|\mathcal{S}(V,\tau)| \leqslant \sum_{i\in[0:r]} \binom{p}{i}, \tag{5.30}$$

with equality if and only if the arrangement is in affine general position, in the sense of definition 5.3.29.

Proof. Observe first that $\mathrm{rank}([V;\tau^{\mathsf{T}}]) \leqslant r+1$. Then, by (5.28), one has

$$|\mathcal{S}(V,0)| \leqslant 2 \sum_{i\in[0\,:\,r-1]} \binom{p-1}{i}, \tag{5.31a}$$

$$|\mathcal{S}([V;\tau^{\mathsf{T}}],0)| \leqslant 2 \sum_{i\in[0\,:\,r]} \binom{p-1}{i}. \tag{5.31b}$$

Using these estimates in proposition 5.3.18(4) provides

$$\begin{aligned} |\mathcal{S}(V,\tau)| &= \frac{1}{2}\,|\mathcal{S}(V,0)| + \frac{1}{2}\,|\mathcal{S}([V;\tau^{\mathsf{T}}],0)| \\ &\leqslant \binom{p-1}{0} + \binom{p-1}{1} + \cdots + \binom{p-1}{r-1} \\ &\quad \underbrace{\qquad\qquad}_{\binom{p}{0}} + \underbrace{\binom{p-1}{0}}_{\binom{p}{1}} + \cdots + \underbrace{\binom{p-1}{r-2}}_{\binom{p}{r-1}} + \underbrace{\binom{p-1}{r-1} + \binom{p-1}{r}}_{\binom{p}{r}} \\ &= \sum_{i\in[0:r]} \binom{p}{i}, \end{aligned} \tag{5.31c}$$

which is the bound (5.30).

By the previous reasoning, equality holds in (5.30) if and only if equalities hold in (5.31a) and (5.31b). By proposition 5.3.26 and lemma 5.3.27, these last equalities are equivalent to the affine general position condition (5.29c). □

For instance, the arrangements in the middle and right-hand side panes in figure 5.1 are in affine general position and verifies (5.30) with equality ($p=3$ and $r=2$).

Observe from (5.31c) that $2\sum_{i\in[0:r-1]}\binom{p-1}{i} = \sum_{i\in[0:r]}\binom{p}{i} - \binom{p-1}{r}$, so that the bound (5.28) on $|\mathcal{S}(V,0)|$ is lower that the bound (5.30) on $|\mathcal{S}(V,\tau)|$, unless $r=p$, in which case the affine arrangement is centered (necessarily $\tau \in \mathcal{R}(V^{\mathsf{T}})$) and can be viewed as a translated linear arrangement.

5.4 Chamber computation - Primal approaches

This section starts the algorithmic part of the paper, which focuses on the computation of the sign vector set $\mathcal{S} \equiv \mathcal{S}(V,\tau)$, defined by (5.5), of the considered arrangement $\mathcal{A}(V,\tau)$. By proposition 5.3.1, the bijection ϕ, defined by (5.6), establishes a one to one correspondence between these sign vectors and the chambers of the arrangement. In this section, we assume that the arrangement is proper, which means that V has only nonzero columns :

$$\forall\, j \in [1:p]: \quad V_{:,j} \neq 0. \tag{5.32}$$

Section 5.6 describes compact versions of some algorithms. Finally, section 5.7 compares these different algorithms on various instances of arrangements.

Many algorithms have been designed to list the chambers of an arrangement (see the introduction). Most of them adopt a *primal* strategy, in the sense that they focus on the realization of the inequality system $s \cdot (V^\mathsf{T} x - \tau) > 0$ in (5.5), by trying to compute witness points $x \in \mathbb{R}^n$. Section 5.4.1 describes the $\mathcal{S}$-tree mechanism of [208], while section 5.4.2 adapts to affine arrangements some of the enhancements brought to this algorithm in [77] for linear arrangements.

5.4.1 Primal $\mathcal{S}$-tree algorithm

For $k \in [1 : p]$, define the partial sign vector set $\mathcal{S}_k \subseteq \{\pm 1\}^k$ and its complement $\mathcal{S}_k^c$ in $\{\pm 1\}^k$ by

$$\mathcal{S}_k \equiv \mathcal{S}_k(V, \tau) := \mathcal{S}(V_{:,[1:k]}, \tau_{[1:k]}) \qquad \text{and} \qquad \mathcal{S}_k^c := \{\pm 1\}^k \setminus \mathcal{S}_k. \tag{5.33}$$

Hence, $\mathcal{S}_k$ is the sign vector set of the arrangement associated with the matrix $V_{:,[1:k]} \in \mathbb{R}^{n \times k}$ and the vector $\tau_{[1:k]} \in \mathbb{R}^k$. Let us denote by v_i the ith column of V, by τ_i the ith component of τ and by $H_i := \{x \in \mathbb{R}^n : v_i^\mathsf{T} x = \tau_i\}$ the ith hyperplane. The $\mathcal{S}$-*tree* is a tree structure, whose k level contains the sign vectors in $\mathcal{S}_k$. Therefore, in addition to its empty root, the complete $\mathcal{S}$-tree has p levels and the bottom one is $\mathcal{S}_p = \mathcal{S}(V, \tau)$. The first level is $\mathcal{S}_1 = \{+1, -1\}$, because the inequalities $(+1)(v_1^\mathsf{T} x_+ - \tau_1) > 0$ and $(-1)(v_1^\mathsf{T} x_- - \tau_1) > 0$ are satisfied by the following two witness points, located on either side of the hyperplane H_1 :

$$x_+ := (\tau_1 + 1) v_1 / \|v_1\|^2 \qquad \text{and} \qquad x_- := (\tau_1 - 1) v_1 / \|v_1\|^2. \tag{5.34}$$

The level $k+1$ is obtained by considering the additional pair $(v_{k+1}, \tau_{k+1}) \in \mathbb{R}^n \times \mathbb{R}$, which defines the hyperplane H_{k+1}. It can be constructed from the level k as follows. By the general assumption (5.32), every node $s \in \mathcal{S}_k$ may have one or two children, namely $(s, +1)$ and/or $(s, -1)$. Geometrically, there are two children if and only if the chamber associated with s is divided in two parts by the hyperplane H_{k+1}, but this geometric view is not easy to detect algebraically in terms of sign vectors (see below). Figure 5.4 shows the three levels of the $\mathcal{S}$-tree corresponding to the arrangement in the middle pane of figure 5.1. Now,

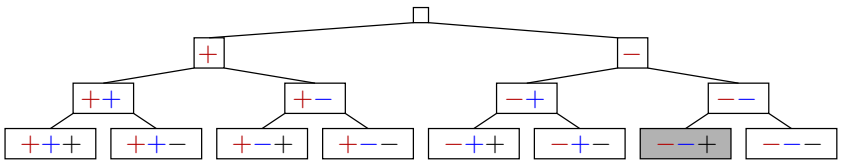


FIGURE 5.4 – $\mathcal{S}$-tree of the arrangement in the middle pane of figure 5.1. The gray node is actually absent from the tree, since there is no chamber associated with $s = (-1, -1, +1)$ (no x such that $s \cdot (V^\mathsf{T} x - \tau) > 0$).

instead of searching the children of every $s \in \mathcal{S}_k$ in order to obtain $\mathcal{S}_{k+1}$, the $\mathcal{S}$-tree will be

constructed by a depth-first search [208], in order to avoid having to keep $\mathcal{S}_k$ in memory, which can be large. In this approach, at most p nodes along a path from the root node to a leaf node must be stored at a time. Note that, in the case of a linear arrangement (i.e., $\tau = 0$) or, more generally, a centered arrangement (i.e., $\tau \in \mathcal{R}(V^\mathsf{T})$), $\mathcal{S}(V, \tau)$ is symmetric (proposition 5.3.5) and only half of the sign vectors must be computed.

In the algorithm descriptions, it is assumed that the problem data (V, τ) is known and we do not repeat this data on entry of the functions. A function can modify its arguments. Let us now outline the algorithm exploring the $\mathcal{S}$-tree, called P_STREE (algorithm 5.4.1, "P" for "primal"), which uses for this purpose a recursive procedure called P_STREE_REC (algorithm 5.4.2).

Algorithm 5.4.1 (P_STREE). // primal $\mathcal{S}$-tree algorithm

1. P_STREE_REC($+1, x_+$) // x_+ given by $(5.34)_1$
2. P_STREE_REC($-1, x_-$) // x_- given by $(5.34)_2$

Algorithm 5.4.2 (P_STREE_REC(s $\in \{\pm 1\}^k$, $x \in \mathbb{R}^n$)).

1. IF ($k = p$)
2. Output s and RETURN // s is a leaf of the $\mathcal{S}$-tree ; end the recursion
3. ENDIF
4. IF ($v_{k+1}^\mathsf{T} x = \tau_{k+1}$)
5. P_STREE_REC($(s, +1), x + \varepsilon v_{k+1}$) // $(s, +1) \in \mathcal{S}_{k+1}$
6. P_STREE_REC($(s, -1), x - \varepsilon v_{k+1}$) // $(s, -1) \in \mathcal{S}_{k+1}$
7. RETURN
8. ENDIF
9. $s_{k+1} := \operatorname{sgn}(v_{k+1}^\mathsf{T} x - \tau_{k+1})$ // $(s, s_{k+1}) \in \mathcal{S}_{k+1}$
10. P_STREE_REC($(s, s_{k+1}), x$)
11. IF ($(s, -s_{k+1})$ is feasible with witness point $\tilde{x}$)
12. P_STREE_REC($(s, -s_{k+1}), \tilde{x}$) // $(s, -s_{k+1}) \in \mathcal{S}_{k+1}$
13. ENDIF

The algorithm P_STREE executes the recursive algorithm P_STREE_REC for constructing the descendants of the nodes "+1" and "-1" of the first level of the $\mathcal{S}$-tree. For its part, the algorithm P_STREE_REC constructs the descendants of a node $s \in \mathcal{S}_k$, knowing a witness point, that is a point $x \in \mathbb{R}^n$ in the chamber associated with s, hence $s_i(v_i^\mathsf{T} x - \tau_i) > 0$ for $i \in [1 : k]$. Let us examine its instructions.

- If $k = p$ (instructions 1..3), the node s is a leaf of the $\mathcal{S}$-tree and has no child. Then, P_STREE_REC just outputs s (it prints it or stores it, depending on the user's wish) and returns to the calling procedure.
- Instructions 4..8 consider the case when x is exactly in the hyperplane H_{k+1}, that is when $v_{k+1}^\mathsf{T} x = \tau_{k+1}$ (in section 5.4.2, the mechanism used in that case will also be applied

when x is sufficiently closed to H_{k+1}) : then s has two children $(s, \pm 1)$, since, for an easily computable sufficiently small $\varepsilon > 0$, $x^{\varepsilon}_{\pm} := x \pm \varepsilon v_{k+1}$ satisfies $s_i(v_i^{\mathsf{T}} x^{\varepsilon}_{\pm} - \tau_i) > 0$, for all $i \in [1 : k]$ and $\pm(v_{k+1}^{\mathsf{T}} x^{\varepsilon}_{\pm} - \tau_{k+1}) > 0$. Note that if $x \in H_{k+1}$, then H_{k+1} is not identical to a previous hyperplane H_i, for $i \in [1 : k]$, since x does not belong to any of these H_i's.

- In the sequel $v_{k+1}^{\mathsf{T}} x \neq \tau_{k+1}$, so that $s_{k+1} := \operatorname{sgn}(v_{k+1}^{\mathsf{T}} x - \tau_{k+1}) \in \{\pm 1\}$ and $(s, s_{k+1}) \in \mathcal{S}_{k+1}$ with witness point x. The instructions 9..10 deal with that situation, asking to compute the descendants of (s, s_{k+1}).
- Instructions 11..13 examine whether $(s, -s_{k+1})$ is also a child of s, which amounts to determining whether the following system has a solution $\tilde{x} \in \mathbb{R}^n$ (see below how this can be done) :

$$\begin{cases} s_i(v_i^{\mathsf{T}} \tilde{x} - \tau_i) > 0, & \text{for } i \in [1 : k] \\ -s_{k+1}(v_{k+1}^{\mathsf{T}} \tilde{x} - \tau_{k+1}) > 0. \end{cases} \tag{5.35}$$

  If this is the case, the descendants of $(s, -s_{k+1})$ are searched using P_STREE_REC.

To determine whether the strict inequalities (5.35) are compatible, one can, like in [208], recast the problem as a linear optimization problem (LOP) and check whether its optimal value is negative. The linear optimization problem reads

$$\begin{array}{rl} \min_{(x,\alpha) \in \mathbb{R}^n \times \mathbb{R}} & \alpha \\ \text{s.t.} & s_i(v_i^{\mathsf{T}} x - \tau_i) + \alpha \geqslant 0, \quad \text{for } i \in [1 : k] \\ & -s_{k+1}(v_{k+1}^{\mathsf{T}} x - \tau_{k+1}) + \alpha \geqslant 0 \\ & \alpha \geqslant -1. \end{array} \tag{5.36}$$

This optimization problem is feasible (by taking an arbitrary $x \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ sufficiently large) and bounded (i.e., its optimal value is bounded below, here by -1), so that it has a solution [29, theorem 19.1]. Denote it by $(\bar{x}, \bar{\alpha})$. It is clear that (5.35) is feasible if and only if $\bar{\alpha} < 0$. This equivalence can then be used as a feasibility criterion for (5.35).

For future reference, we quote in a proposition an observation, which is deduced from the algorithm.

Proposition 5.4.3 (binary $\mathcal{S}$-tree)**.** *Let $\mathcal{A}(V, \tau)$ be a proper arrangement. Then, the $\mathcal{S}$-tree is a binary tree.*

Proof. This fact is obtained by construction of the $\mathcal{S}$-tree in algorithm 5.4.1–5.4.2. Note that to have two children in lines 5..6 of algorithm 5.4.2, one must have $v_{k+1} \neq 0$, hence the assumption of a *proper* arrangement. If $v_{k+1} = 0$, either s has no child (if $\tau_{k+1} = 0$) or the single child $(s, -\operatorname{sgn}(\tau_{k+1}))$. □

By proposition 5.4.3, any sign vector in $\mathcal{S}_k$, with $k \in [1 : p-1]$, has either one or two children in $\mathcal{S}_{k+1}$. The next proposition characterizes the sign vectors of $\mathcal{S}_k$ that have two children in $\mathcal{S}_{k+1}$. It extends to affine arrangements proposition 4.9 in [78], which is there used to give an analytic version of Winder's proof of (3.35), giving the cardinality of $\mathcal{S}(V, 0)$.

Below, proposition 5.4.4 will be useful to get the lower bound (5.39) on $|\mathcal{S}(V,\tau)|$, improving (5.12). In the statement of proposition 5.4.4, $\mathrm{P}_{k+1} : \mathbb{R}^n \to H_{k+1} - H_{k+1}$ denotes the orthogonal projector on the subspace $v_{k+1}^{\perp}$ that is parallel to the affine space $H_{k+1} := \{x \in \mathbb{R}^n : v_{k+1}^{\mathsf{T}} x = \tau_{k+1}\}$; while $\hat{x}_{k+1} := \tau_{k+1} v_{k+1}/\|v_{k+1}\|^2$ is the unique point in $\mathcal{N}(\mathrm{P}_{k+1}) \cap H_{k+1}$. We also denote by P_{k+1} the transformation matrix of the projector, so that $\mathrm{P}_{k+1}\, V_{:,[1:k]}$ can be viewed as the product of two matrices (its jth column is $\mathrm{P}_{k+1}\, v_j$, for $j \in [1:k]$). Note that in (5.37b), $\tilde{V} := \mathrm{P}_{k+1}\, V_{:,[1:k]}$ may have zero columns, in which case, by its definition (5.5), the set $\mathcal{S}(\tilde{V},\tilde{\tau})$ will be nonempty if the corresponding components of $\tilde{\tau}$ do not vanish.

Proposition 5.4.4 (two child criterion). *Let $V \in \mathbb{R}^{n\times p}$, $s \in \{\pm 1\}^k$ for some $k \in [1:p-1]$ and $\hat{x}_{k+1}$ be the unique point in $\mathcal{N}(\mathrm{P}_{k+1}) \cap H_{k+1}$. Then,*

$$
\begin{aligned}
&(s,+1) \text{ and } (s,-1) \in \mathcal{S}_{k+1} \\
&\quad\iff \exists\, x \in \mathbb{R}^n : s_i(v_i^{\mathsf{T}} x - \tau_i) > 0, \text{ for } i \in [1:k], \text{ and } v_{k+1}^{\mathsf{T}} x - \tau_{k+1} = 0 \qquad (5.37a)\\
&\quad\iff s \in \mathcal{S}(\mathrm{P}_{k+1}\, V_{:,[1:k]}, \tau_{[1:k]} - V_{:,[1:k]}^{\mathsf{T}} \hat{x}_{k+1}). \qquad (5.37b)
\end{aligned}
$$

Proof. To simplify the notation, set $V_k := V_{:,[1:k]}$. The following equivalences prove the result ((5.38a) and (5.38b) are justified afterwards) :

$$
\begin{aligned}
&(s,+1) \text{ and } (s,-1) \in \mathcal{S}(V_{k+1}, \tau_{[1:k+1]}) \\
&\quad\iff \begin{cases} \exists\, x_+ \in \mathbb{R}^n : s_i(v_i^{\mathsf{T}} x_+ - \tau_i) > 0, \text{ for } i \in [1:k], \\ \quad \text{and } +(v_{k+1}^{\mathsf{T}} x_+ - \tau_{k+1}) > 0 \\ \exists\, x_- \in \mathbb{R}^n : s_i(v_i^{\mathsf{T}} x_- - \tau_i) > 0, \text{ for } i \in [1:k], \\ \quad \text{and } -(v_{k+1}^{\mathsf{T}} x_- - \tau_{k+1}) > 0 \end{cases} \\
&\quad\iff \exists\, x \in \mathbb{R}^n : s_i(v_i^{\mathsf{T}} x - \tau_i) > 0, \text{ for } i \in [1:k], \text{ and } v_{k+1}^{\mathsf{T}} x - \tau_{k+1} = 0 \qquad (5.38a)\\
&\quad\iff \exists\, x \in \mathbb{R}^n : s_i([\mathrm{P}_{k+1}\, v_i]^{\mathsf{T}} x - [\tau_i - v_i^{\mathsf{T}} \hat{x}_{k+1}]) > 0, \text{ for } i \in [1:k] \qquad (5.38b)\\
&\quad\iff s \in \mathcal{S}(\mathrm{P}_{k+1}\, V_k, \tau_{[1:k]} - V_k^{\mathsf{T}} \hat{x}_{k+1}).
\end{aligned}
$$

The equivalence in (5.38a) is shown as follows.

[⇒] Define $t_- := +(v_{k+1}^{\mathsf{T}} x_+ - \tau_{k+1}) > 0$, $t_+ := -(v_{k+1}^{\mathsf{T}} x_- - \tau_{k+1}) > 0$, $\alpha_- := t_-/(t_- + t_+) \in (0,1)$ and $\alpha_+ := t_+/(t_- + t_+) \in (0,1)$. Then, $x = \alpha_+ x_+ + \alpha_- x_-$ is appropriate since

$$
\begin{aligned}
\text{for } i \in [1:k] : \quad & s_i(v_i^{\mathsf{T}} x - \tau_i) = \alpha_+ s_i[v_i^{\mathsf{T}} x_+ - \tau_i] + \alpha_- s_i[v_i^{\mathsf{T}} x_- - \tau_i] > 0, \\
& v_{k+1}^{\mathsf{T}} x - \tau_{k+1} = \alpha_+[v_{k+1}^{\mathsf{T}} x_+ - \tau_{k+1}] + \alpha_-[v_{k+1}^{\mathsf{T}} x_- - \tau_{k+1}] = \alpha_+ t_- - \alpha_- t_+ = 0.
\end{aligned}
$$

[⇐] Take $x_\pm = x \pm \varepsilon v_{k+1}$ for a sufficiently small $\varepsilon > 0$.

The equivalence in (5.38b) is shown as follows.

[⇒] The point x in (5.38a) satisfies $x \in H_{k+1}$, so that $x = \mathrm{P}_{k+1}\, x + \hat{x}_{k+1}$, since $x - \mathrm{P}_{k+1}\, x \in \mathcal{N}(\mathrm{P}_{k+1}) \cap H_{k+1} = \{\hat{x}_{k+1}\}$. Furthermore, by (5.38a), one has for $i \in [1:$

$k]$:

$$0 < s_i(v_i^\mathsf{T} x - \tau_i) = s_i(v_i^\mathsf{T}[\mathrm{P}_{k+1}\, x + \hat{x}_{k+1}] - \tau_i) = s_i([\mathrm{P}_{k+1}\, v_i]^\mathsf{T} x - [\tau_i - v_i^\mathsf{T}\hat{x}_{k+1}]),$$

where we have used $\mathrm{P}_{k+1}^\mathsf{T} = \mathrm{P}_{k+1}$ (P_{k+1} is an orthogonal projector).

[$\Leftarrow$] Take $x := \mathrm{P}_{k+1}\, x_0 + \hat{x}_{k+1} \in H_{k+1}$ in (5.38a), where x_0 is the x given by (5.38b). □

Note that if the equivalences in proposition 5.4.4 hold, then $s \in \mathcal{S}_k$.

The next proposition improves and extends to affine arrangement proposition 4.6 in [77]. We denote by $\mathrm{vect}\{v_1, \ldots, v_k\}$ the vector space spanned by the vectors $v_1, \ldots, v_k$.

Proposition 5.4.5 (incrementation). *Let $V = [v_1 \cdots v_p] \in \mathbb{R}^{n\times p}$.*

1) *If $s \in \mathcal{S}_k^c$, then $(s, \pm 1) \in \mathcal{S}_{k+1}^c$. Consequently, $|\mathcal{S}_{k+1}^c| \geqslant 2|\mathcal{S}_k^c|$.*
2) *If $v_{k+1} \notin \mathrm{vect}\{v_1, \ldots, v_k\}$, then, $(s, \pm 1) \in \mathcal{S}_{k+1}$ for all $s \in \mathcal{S}_k$, $|\mathcal{S}_{k+1}| = 2|\mathcal{S}_k|$ and $|\mathcal{S}_{k+1}^c| = 2|\mathcal{S}_k^c|$.*
3) *If $v_{k+1} \in \mathrm{vect}\{v_1, \ldots, v_k\}$, $V_{:,[1:k+1]}$ has no zero column, $H_{k+1} \neq H_i$ for $i \in [1:k]$, and $r_k := \dim \mathrm{vect}\{v_1, \ldots, v_k\}$, then $|\mathcal{S}_{k+1}| \geqslant |\mathcal{S}_k| + 2^{r_k - 1}$.*

Proof. 1) If $s \in \mathcal{S}_k^c$, there is no $x \in \mathbb{R}^n$ such that $s_i(v_i^\mathsf{T} x - \tau_i) > 0$ for $i \in [1:k]$. Therefore, there is certainly no $x \in \mathbb{R}^n$ satisfying $s_i(v_i^\mathsf{T} x - \tau_i) > 0$ for $i \in [1:k+1]$, with $s_{k+1} \in \{\pm 1\}$. Therefore, $(s, \pm 1) \in \mathcal{S}_{k+1}^c$. This observation implies that $|\mathcal{S}_{k+1}^c| \geqslant 2|\mathcal{S}_k^c|$.

2) Let Q be the orthogonal projector on $\mathrm{vect}\{v_1, \ldots, v_k\}^\perp$ for the Euclidean scalar product. By assumption, $\mathrm{Q}\, v_{k+1} \neq 0$. Let $s \in \mathcal{S}_k$, so that there is an $x \in \mathbb{R}^n$ such that $s_i(v_i^\mathsf{T} x - \tau_i) > 0$ for $i \in [1:k]$. For any $t \in \mathbb{R}$ and $i \in [1:k]$, the points $x_\pm := x \pm t\, \mathrm{Q}\, v_{k+1}$ verify $s_i(v_i^\mathsf{T} x_\pm - \tau_i) = s_i(v_i^\mathsf{T} x - \tau_i) > 0$ (because $v_i^\mathsf{T}\, \mathrm{Q}\, v_{k+1} = 0$). In addition, for $t > 0$ sufficiently large, one has $\pm(v_{k+1}^\mathsf{T} x_\pm - \tau_{k+1}) = \pm(v_{k+1}^\mathsf{T} x - \tau_{k+1}) + t\|\, \mathrm{Q}\, v_{k+1}\|^2 > 0$ (because $\mathrm{Q}^2 = \mathrm{Q}$ and $\mathrm{Q}^\mathsf{T} = \mathrm{Q}$). We have shown that both $(s, +1)$ and $(s, -1)$ are in $\mathcal{S}_{k+1}$. Therefore, $|\mathcal{S}_{k+1}| \geqslant 2|\mathcal{S}_k|$.

Now, $|\mathcal{S}_k| + |\mathcal{S}_k^c| = 2^k$, $|\mathcal{S}_{k+1}| + |\mathcal{S}_{k+1}^c| = 2^{k+1}$ and $|\mathcal{S}_{k+1}^c| \geqslant 2|\mathcal{S}_k^c|$ by point 1. Therefore, one must have $|\mathcal{S}_{k+1}| = 2|\mathcal{S}_k|$ and $|\mathcal{S}_{k+1}^c| = 2|\mathcal{S}_k^c|$. 3) One has $\mathrm{null}(\mathrm{P}_{k+1}) = 1$ (since $\mathcal{N}(\mathrm{P}_{k+1}) = \mathbb{R}v_{k+1}$) and $\mathrm{rank}(V_{:,[1:k]}) = r_k$ (by definition). Then, $\mathrm{rank}(\mathrm{P}_{k+1} V_{:,[1:k]}) \geqslant r_k - 1$. To apply (5.12) to the arrangement $\mathcal{A}(\mathrm{P}_{k+1} V_{:,[1:k]}, \tau_{[1:k]} - V_{:,[1:k]}^\mathsf{T}\hat{x}_{k+1})$, one must show that $v_i^\mathsf{T}\hat{x}_{k+1} \neq \tau_i$ when $i \in [1:k]$ and $\mathrm{P}_{k+1}\, v_i = 0$ (i.e., v_{k+1} and v_i are colinear or H_{k+1} and H_i are parallel by proposition 5.3.2(1)). This is indeed the case, since $v_i^\mathsf{T}\hat{x}_{k+1} = \tau_i$ would imply that $H_{k+1} = H_i$ (because then $\hat{x}_{k+1}$ would belong to both H_i and H_{k+1}, which are parallel), in contradiction with the assumption. By (5.12), $|\mathcal{S}(\mathrm{P}_{k+1} V_{:,[1:k]}, \tau_{[1:k]} - V_{:,[1:k]}^\mathsf{T}\hat{x}_{k+1})| \geqslant 2^{r_k - 1}$.

By proposition 5.4.4, there are at least 2^{r_k-1} sign vectors in $\mathcal{S}_k$ with two children. Since any $s \in \mathcal{S}_k$ has at least one child when $v_{k+1} \neq 0$, one gets $|\mathcal{S}_{k+1}| \geqslant |\mathcal{S}_k| + 2^{r_k-1}$. □

Corollary 5.4.6 (lower bound of $|\mathcal{S}(V,\tau)|$). *For a matrix V of rank r without zero column and a vector τ such that all the hyperplanes H_i are different, one has*

$$2^r + 2^{r-1}(p-r) \leqslant |\mathcal{S}(V,\tau)|. \tag{5.39}$$

Proof. By a possible change of column order, which only affects the chamber numbering, not the cardinality of $\mathcal{S}(V,\tau)$, one can assume that the first r columns of V are linearly independent. Then, by proposition 5.4.5(2), $|\mathcal{S}_r| = 2^r$. Next, for $k \in [r : p-1]$, $\dim \operatorname{vect}\{v_1, \ldots, v_k\} = r$, so that proposition 5.4.5(3) implies that $|\mathcal{S}_{k+1}| \geqslant |\mathcal{S}_k| + 2^{r-1}$. By induction, one gets (5.39). □

5.4.2 Preventing some computations

The main computation cost of algorithm 5.4.1 comes from solving the LOPs (5.36) at some inner nodes. This section describes three ways of bypassing LOPs. They are adapted from [77], where only linear arrangements are considered, and are identified by the letters A, B and C (the letters appearing in the section titles), which will also be used to label more efficient variants of algorithm 5.4.1. These variants significantly speed up the algorithm by reducing the numbers of LOPs to solve (see section 5.7).

A - Rank of the arrangement

Instead of starting the $\mathcal{S}$-tree with the two nodes of $\mathcal{S}_1 = \{+1, -1\}$, like in algorithm 5.4.1, one can start it with the 2^r nodes of $\mathcal{S}_r = \{\pm 1\}^r$, by considering first a selection of $r := \operatorname{rank}(V)$ linearly independent vectors whose $\mathcal{S}$-tree is easy to construct without having to solve any LOP. Here are the details.

Numerically, r linearly independent vectors can be found by a QR factorization of V :

$$VP = QR,$$

with $P \in \{0,1\}^{p\times p}$ is a permutation matrix, $Q \in \mathbb{R}^{n\times n}$ is orthogonal and $R \in \mathbb{R}^{n\times p}$ is upper triangular with $R_{[r+1\,:\,n],:} = 0$. To simplify the presentation, let us assume that P is the identity matrix, in which case the first r vectors v_1, …, v_r (or columns of V) are linearly independent, and let us note $V_r := V_{:,[1:r]}$, $Q_r := Q_{:,[1:r]}$ and $R_r := R_{[1:r],[1:r]}$. By proposition 5.4.5(2),

$$\mathcal{S}_r = \{\pm 1\}^r.$$

To launch the recursive algorithm 5.4.2, one still need to compute a witness point x_s associated with any $s \in \mathcal{S}_r$. For this purpose, one computes a point $\hat{x} \in \cap_{i=1}^r H_i$, hence verifying $V_r^{\mathsf{T}}\hat{x} = \tau_{[1:r]}$, by $\hat{x} := V_r(V_r^{\mathsf{T}}V_r)^{-1}\tau_{[1:r]}$. Next, for any $s \in \{\pm 1\}^r$, one computes

$d_s := Q_r R_r^{-\mathsf{T}} s \in \mathbb{R}^n$. Let us show that $x_s := \hat{x} + d_s$ is a witness point of the considered s. One has

$$V_r^{\mathsf{T}} x_s - \tau_{[1:r]} = V_r^{\mathsf{T}}[V_r (V_r^{\mathsf{T}} V_r)^{-1} \tau_{[1:r]} + Q_r R_r^{-\mathsf{T}} s] - \tau_{[1:r]} = (Q_r R_r)^{\mathsf{T}} Q_r R_r^{-\mathsf{T}} s = s.$$

Therefore, $s \bullet (V_r^{\mathsf{T}} x_s - \tau_{[1:r]}) = s \bullet s = e > 0$, as desired.

B - Handling of a hyperplane proximity

In the description of algorithm 5.4.2, it is shown why a witness point x of a sign vector $s \in \mathcal{S}_k$ that belongs to H_{k+1}, i.e., $v_{k+1}^{\mathsf{T}} x = \tau_{k+1}$, allows the algorithm to certify that both $(s, +1)$ and $(s, -1)$ are in $\mathcal{S}_{k+1}$, without having to solve a LOP. We show with the next proposition that this is still true when x is near H_{k+1}, in the sense (5.40). Note that this proximity to H_{k+1} is measured by strict inequalities, which is more stable with respect to numerical perturbations than an equality.

Proposition 5.4.7 (two children without LOP). *Let $s \in \mathcal{S}_k$ and $x \in \mathbb{R}^n$ verifying $s \bullet (V_{:,[1:k]}^{\mathsf{T}} x - \tau_{[1:k]}) > 0$. Suppose that $v_{k+1} \neq 0$ and*

$$\underbrace{\max_{s_i v_i^{\mathsf{T}} v_{k+1} > 0} \frac{\tau_i - v_i^{\mathsf{T}} x}{v_i^{\mathsf{T}} v_{k+1}}}_{=: t_{\min}} < \underbrace{\frac{\tau_{k+1} - v_{k+1}^{\mathsf{T}} x}{||v_{k+1}||^2}}_{=: t_0} < \underbrace{\min_{s_i v_i^{\mathsf{T}} v_{k+1} < 0} \frac{\tau_i - v_i^{\mathsf{T}} x}{v_i^{\mathsf{T}} v_{k+1}}}_{=: t_{\max}}. \tag{5.40}$$

Then, for $t_- \in (t_{\min}, t_0)$, $x_- := x + t_- v_{k+1}$ is a witness point of $(s, -1)$ and, for $t_+ \in (t_0, t_{\max})$, $x_+ := x + t_+ v_{k+1}$ is a witness point of $(s, +1)$.

Proof. Note first that, in (5.40), the arguments of the maximum are negative and the arguments of the minimum are positive. Therefore, both inequalities are verified if $v_{k+1}^{\mathsf{T}} x = \tau_{k+1}$ ($t_0 = 0$), that is, when $x \in H_{k+1}$. One has, for $i \in [1 : k]$,

$$s_i(v_i^{\mathsf{T}}(x + t v_{k+1}) - \tau_i) > 0 \quad \Longleftrightarrow \quad \begin{cases} t < \frac{s_i(\tau_i - v_i^{\mathsf{T}} x)}{s_i v_i^{\mathsf{T}} v_{k+1}} & \text{if } s_i v_i^{\mathsf{T}} v_{k+1} < 0 \\ t \in \mathbb{R} & \text{if } s_i v_i^{\mathsf{T}} v_{k+1} = 0 \\ t > \frac{s_i(\tau_i - v_i^{\mathsf{T}} x)}{s_i v_i^{\mathsf{T}} v_{k+1}} & \text{if } s_i v_i^{\mathsf{T}} v_{k+1} > 0. \end{cases}$$

Since the conditions imposed on t in the right-hand side of the equivalence above are satisfied by any $t \in (t_{\min}, t_{\max})$, it follows that $x_\pm$ are witness points of s. One has

$$t > t_0 := \tfrac{\tau_{k+1} - v_{k+1}^{\mathsf{T}} x}{||v_{k+1}||^2} \quad \Longleftrightarrow \quad v_{k+1}^{\mathsf{T}}(x + t v_{k+1}) - \tau_{k+1} > 0,$$
$$t < t_0 := \tfrac{\tau_{k+1} - v_{k+1}^{\mathsf{T}} x}{||v_{k+1}||^2} \quad \Longleftrightarrow \quad v_{k+1}^{\mathsf{T}}(x + t v_{k+1}) - \tau_{k+1} < 0.$$

Since t_+ (resp. t_-) verifies the condition in the left-hand side of the first (resp. second) equivalence above, it follows that x_+ (resp. x_-) is a witness point of $(s, +1)$ (resp. $(s, -1)$). □

C - Choosing the order of the vectors

Every inner node of the $\mathcal{S}$-tree has one or two children and this number is sometimes detected in algorithm 5.4.2 by solving a LOP, which is a time consuming operation. Therefore, a way of decreasing the computation time is to reduce the number of inner nodes of the $\mathcal{S}$-tree. This property can be obtained by choosing wisely the order in which the pairs (v_i, τ_i), or hyperplanes H_i, are taken into account when constructing the branches of the $\mathcal{S}$-tree (this order can be different from one branch to another), with the goal of placing the nodes with a single child close to the root of the tree. This strategy has been investigated in [77, §5.2.4.C] and we adapt the heuristic to the present context of affine arrangements.

Denote by $T_s := \{i_1^s, \ldots, i_k^s\}$ the set of the indices of the hyperplanes selected to reach node $s \in \mathcal{S}_k$ (T_s depends on s). At this node, the algorithm must choose the next hyperplane to consider, whose index is among the index set $T_s^c := [1:p] \setminus T_s$. With the goal of preventing, as much as possible, the node s from having two children, a natural idea is to ignore the indices of T_s^c, for which proposition 5.4.7 ensures two children. In the remaining index set, denote it by T_s^b, the chosen index is the one maximizing the quantity $|v_i^{\mathsf{T}} x - \tau_i| / \|[v_i; \tau_i]\|$ for $i \in T_s^b$ (x is the witness point associated by the algorithm with the current node s), since the larger this quantity is, the further x is from the chosen hyperplane, which should increase the chances that s will have only one child.

5.5 Chamber computation - Dual approaches

The listing of the chambers of an arrangement $\mathcal{A}(V, \tau)$ can be tackled by an approach different from those presented in section 5.4 (algorithm 5.4.1 and its improvements A, B and C), sometimes (or always) replacing optimization phases by algebra techniques. More specifically, we say that an algorithm has a *dual aspect* when it uses the concept of stem vector (definition 5.3.12) by means of proposition 5.3.16. Such a dual approach was introduced in [77, SS5.2.2-5.2.3] for linear arrangements.

Section 5.5.1 deals with algorithms computing $\mathcal{S}(V, \tau)$ that assume the availability of the full stem vector set $\mathfrak{S}(V, \tau)$ and do not use optimization. A method for computing $\mathfrak{S}(V, \tau)$ is presented in section 5.5.1. Section 5.5.1 describes a crude dual approach for computing $\mathcal{S}(V, \tau)$ that is only efficient for small arrangements. The algorithm proposed in section 5.5.1 has the structure of algorithm 5.4.1, in the sense that it constructs the $\mathcal{S}$-tree, but its optimization phases are replaced by the duality technique mentioned above.

In section 5.5.2, we present a way of obtaining circuits in the *primal* algorithm of section 5.4. Indeed, proposition 5.5.6 indicates that in the encountered infeasible LOPs, the *dual* variables contain a circuit. This can lead to an algorithm mixing the primal and dual aspects. This section also assumes (5.32), i.e., V has no zero columns.

5.5.1 Algorithms using all the stem vectors

Proposition 5.3.16 establishes a link between the *infeasible* sign vectors, those in $\mathcal{S}(V, \tau)^c$, and the stem vectors. This section presents two algorithms that start with the computation of the complete stem vector set $\mathfrak{S}(V, \tau)$ (section 5.5.1). The first one uses these stem vectors to compute $\mathcal{S}(V, \tau)^c$, from which the feasible sign vector set $\mathcal{S}(V, \tau) = \{\pm 1\}^p \setminus \mathcal{S}(V, \tau)^c$ can be deduced (section 5.5.1). The second one computes directly $\mathcal{S}(V, \tau)$ like in the $\mathcal{S}$-tree primal algorithm of section 5.4.1, but without solving linear optimization problems and without computing witness points (section 5.5.1).

Stem vector computation

Let us start with the presentation of a plain algorithm that computes the disjoint sets $\mathfrak{S}_s(V, \tau)$ and $\mathfrak{S}_a(V, \tau)$ of the symmetric and asymmetric stem vectors ; recall that the set of all stem vectors is $\mathfrak{S}(V, \tau) = \mathfrak{S}_s(V, \tau) \cup \mathfrak{S}_a(V, \tau)$. This algorithm is based on the detection of the circuits of V and remark 5.3.13(3). It is rudimentary (the one used in [141] is valid for an arbitrary matroid and yields an interesting complexity property, but, in our experience with vector matroids, it is much less efficient than the method used in algorithm 5.5.1 ; see also [214] and the pieces of software mentioned in the introduction). The algorithm can be significantly improved in particular cases, see [212].

Algorithm 5.5.1 (STEM_VECTORS($\mathfrak{S}_s$, $\mathfrak{S}_a$)). // stem vector calculation

1. $\mathfrak{S}_s = \varnothing$ and $\mathfrak{S}_a = \varnothing$
2. FOR $i \in [1 : p]$ DO
3. STEM_VECTORS_REC($\mathfrak{S}_s$, $\mathfrak{S}_a$, $\{i\}$)
4. ENDFOR
5. Remove duplicate stem vectors in $\mathfrak{S}_s$ and $\mathfrak{S}_a$

Algorithm 5.5.2 (STEM_VECTORS_REC($\mathfrak{S}_s$, $\mathfrak{S}_a$, I_0)).

1. FOR $i \in [\max(I_0) + 1 : p]$ DO
2. $I := I_0 \cup \{i\}$
3. IF ($\mathcal{N}(V_{:,I}) \neq \{0\}$)
4. Let $\eta_I \in \mathcal{N}(V_{:,I}) \setminus \{0\}$ and $J := \{i \in I : \eta_i \neq 0\}$ // $J \in \mathcal{C}(V)$
5. IF ($\tau_J^{\mathsf{T}} \eta_J = 0$)
6. $\mathfrak{S}_s := \mathfrak{S}_s \cup \{\operatorname{sgn}(\eta_J\} \cup \{-\operatorname{sgn}(\eta_J)\}$
7. ELSE
8. $\mathfrak{S}_a := \mathfrak{S}_a \cup \{\operatorname{sgn}(\tau_J^{\mathsf{T}} \eta_J) \operatorname{sgn}(\eta_J)\}$
9. ENDIF
10. ELSE
11. STEM_VECTORS_REC($\mathfrak{S}_s$, $\mathfrak{S}_a$, I)
12. ENDIF

13. ENDFOR

Here are some explanations and observations on algorithm 5.5.1. Unless otherwise stated, the line numbers refer to algorithm 5.5.2.

- The function STEM_VECTORS_REC($\mathfrak{S}_s$, $\mathfrak{S}_a$, I_0) adds to $\mathfrak{S}_s$ and/or $\mathfrak{S}_a$ stem vectors σ such that $\mathfrak{J}(\sigma)$ is contained in the set formed of I_0 and indices larger that $\max(I_0)$.
- On entry in algorithm 5.5.2, $V_{:,I_0}$ is assumed to be injective : this is the case in line 3 of algorithm 5.5.1 (recall the assumption (5.32)) and in line 11 of algorithm 5.5.2 (since there $\mathcal{N}(V_{:,I}) = \{0\}$). Therefore, in line 4 of algorithm 5.5.2, $\mathrm{null}(V_{:,I}) = 1$ and J is a circuit of V (lemma 5.3.11).
- The algorithm does not explore all the subsets of $[1 : p]$ since, once a circuit $J \subseteq I$ has been found in line 4, an index set $I' \supseteq I$ satisfying $\mathrm{null}(N_{:,I'}) = 1$ contains no circuit different from J (lemma 5.3.11). This explains why there is no recursive call to STEM_VECTORS_REC when $\mathcal{N}(V_{:,I}) \neq \{0\}$ (lines 4..9).
- In line 4, η_I is obtained by a null space computation, so that algorithm 5.5.2 is sensitive to rounding errors. One can use exact arithmetic linear algebra to compute the set of sign vectors when the data is rational or integer, at the expense of slower computation.
- Line 6 corresponds to remark 5.3.13(3.a) and line 7 to remark 5.3.13(3.b). These lines are symbolically written, since, in addition to $\pm \operatorname{sgn} \eta_J$ one also has to store J. In practice, one can only store half of the symmetric stem vectors in $\mathfrak{S}_s$ and obtain the full set, if needed, by gathering the stem vectors of $\mathfrak{S}_s$ and $-\mathfrak{S}_s$.
- The loop 2..4 of algorithm 5.5.1 may find several times the same stem vector. This is the case, for instance, if $V = [e_1, e_2, e_2]$: the circuit $J = \{2, 3\}$ is found twice by the loop of algorithm 5.5.1 (once with $i = 1$ and again with $i = 2$), as well as the associated stem vectors. This justifies the final elimination of duplicates in line 5 of algorithm 5.5.1.

Crude dual algorithm

The algorithm described in this section, algorithm 5.5.3, is a "crude" way of obtaining the sign vector set $\mathcal{S}(V, \tau)$ from the stem vector set $\mathfrak{S}(V, \tau)$. It uses the characterization of proposition 5.3.16. For each stem vector $\sigma \in \mathfrak{S}(V, \tau)$ with associated circuit $J = \mathfrak{J}(\sigma) \subseteq [1 : p]$, the algorithm generates all the infeasible sign vectors $s \in \mathcal{S}(V, \tau)^c$ satisfying $s_J = \sigma$ and $s_{J^c} \in \{\pm 1\}^{J^c}$. This is made by the function STEM_TO_INFEAS_SIGN_VECTORS in a straightforward manner (the precise computation is not detailed). Once $\mathcal{S}(V, \tau)^c$ is computed, $\mathcal{S}(V, \tau)$ is obtained by $\{\pm 1\}^p \setminus \mathcal{S}(V, \tau)^c$.

The STEM_TO_INFEAS_SIGN_VECTORS function, since it is called multiple times, can produce duplicated sign vectors, thus justifying the cleaning operation in line 5 (this one could be done simultaneously with the union in line 4). For example, with $V = [1, 1, 1]$ and $\tau^\mathsf{T} = [1, 0, 1]$, the stem vectors $(1, -1) \in \{\pm 1\}^{\{1,2\}}$ and $(-1, 1) \in \{\pm 1\}^{\{2,3\}}$ produce the same infeasible sign vector $(1, -1, 1)$.

Algorithm 5.5.3 (CRUDE_DUAL($\mathcal{S}$)). // crude dual algorithm

1. Sc = $\varnothing$ // initialization of $\mathcal{S}(V,\tau)^c$
2. STEM_VECTORS($\mathfrak{S}_s$, $\mathfrak{S}_a$) // algorithm 5.5.1
3. FOR $\sigma \in \mathfrak{S}_s \cup \mathfrak{S}_a$ DO
4. Sc = Sc $\cup$ STEM_TO_INFEAS_SIGN_VECTORS(σ, P)
5. Remove duplicates in Sc
6. ENDFOR
7. $\mathcal{S} := \{\pm 1\}^p \setminus$ Sc

Despite its simplicity, algorithm 5.5.3 is usually not very attractive. Indeed, each stem vector $\sigma \in \{\pm 1\}^J$ produces the exponential number $2^{|J^c|}$ of sign vectors s with $s_{J^c} \in \{\pm 1\}^{J^c}$. As a result, for large p, the algorithm handles a large amount of data, which can take much computing time.

Dual $\mathcal{S}$-tree algorithm

Another possibility is to use the $\mathcal{S}$-tree structure introduced in section 5.4.1. Here is the main idea. Assume that a sign vector s in $\mathcal{S}_k$ has been computed (the set $\mathcal{S}_k$ is defined by (5.33)). Then, algorithm 5.4.2 determines whether (s, s_{k+1}) belongs to $\mathcal{S}_{k+1}$, for $s_{k+1} \in \{\pm 1\}$. As explained in the description of algorithm 5.4.2, the belonging of (s, s_{k+1}) to $\mathcal{S}_{k+1}$ can be revealed by solving a linear optimization problem. Algorithm 5.5.4–5.5.5 below does this differently. It uses the computed stem vector set $\mathfrak{S}(V,\tau)$ and is based on the fact that

$$\mathfrak{S}(V_{:,[1:k]}, \tau_{[1:k]}) = \{\sigma \in \mathfrak{S}(V,\tau) : \mathfrak{J}(\sigma) \subseteq [1:k]\}.$$

Therefore, according to proposition 5.3.16, to determine whether (s, s_{k+1}) is in $\mathcal{S}_{k+1}$, it suffices to see whether it covers a stem vector $\sigma \in \mathfrak{S}(V,\tau)$ such that $\mathfrak{J}(\sigma) \subseteq [1:k+1]$. If so $(s, s_{k+1}) \in \mathcal{S}^c_{k+1}$ and any $\tilde{s} \in \{\pm 1\}^p$, extending (s, s_{k+1}) by ± 1, will be in $\mathcal{S}(V,\tau)^c$, so that the $\mathcal{S}$-tree may be pruned at (s, s_{k+1}). Otherwise $(s, s_{k+1}) \in \mathcal{S}_{k+1}$ and the recursive exploration of the $\mathcal{S}$-tree is pursued below (s, s_{k+1}).

Algorithm 5.5.4 (D_STREE). // dual $\mathcal{S}$-tree algorithm

1. STEM_VECTORS($\mathfrak{S}_s$, $\mathfrak{S}_a$) // get the stem vectors by algorithm 5.5.1
2. D_STREE_REC($\varnothing$,$\mathfrak{S}_s \cup \mathfrak{S}_a$)

Algorithm 5.5.5 (D_STREE_REC(s $\in \{\pm 1\}^k$,$\mathfrak{S}$)).

1. IF ($k = p$)
2. Output s and RETURN // s is a leaf of the $\mathcal{S}$-tree ; end the recursion
3. ENDIF
4. IF ($[s; +1]$ covers a stem vector of $\mathfrak{S}$)
5. D_STREE_REC([s ;-1],$\mathfrak{S}$)

6. ELSE
7. D_STREE_REC([S;+1],$\mathfrak{S}$)
8. IF ($[s;-1]$ does not contain a stem vector of $\mathfrak{S}$)
9. D_STREE_REC([S;-1],$\mathfrak{S}$)
10. ENDIF
11. ENDIF

Here are some explanations and observations on the recursive algorithm 5.5.5.

- If the test in line 4 holds, proposition 5.3.16 tells us that $[s;+1]$ is an infeasible sign vector for the arrangement $\mathcal{A}(V_{:,[1:k+1]}, \tau_{[1:k+1]})$, so that, for any $\tilde{s} \in \{\pm 1\}^{p-k-1}$, the sign vectors $[s;+1;\tilde{s}]$ is also infeasible for the arrangement $\mathcal{A}(V,\tau)$. This has two consequences :
  - there is no point in exploring the descendants of $[s;+1]$ in the $\mathcal{S}$-tree, which explains why there is no recursive call to D_STREE_REC([S;+1],$\mathfrak{S}$) in that case and
  - $[s;-1]$ is necessarily a feasible sign vector for the arrangement $\mathcal{A}(V_{:,[1:k+1]}, \tau_{[1:k+1]})$ since each node of the $\mathcal{S}$-tree has at least one child (proposition 5.4.3), which explains why there is a call to D_STREE_REC([S;-1],$\mathfrak{S}$) in line 5.
- Line 7 is justified since at that point, $[s;+1]$ is a feasible sign vector for the arrangement $\mathcal{A}(V_{:,[1:k+1]}, \tau_{[1:k+1]})$.
- Line 9 is justified since at that point, $[s;-1]$ is a feasible sign vector for the arrangement $\mathcal{A}(V_{:,[1:k+1]}, \tau_{[1:k+1]})$.
- The algorithm does not use witness points, unlike the primal $\mathcal{S}$-tree algorithm 5.4.1–5.4.2.

Let us emphasize the fact that algorithm 5.5.4 does not require to solve linear optimization problems. While this might look enticing, since the LOPs are the main cost of the primal $\mathcal{S}$-tree algorithm 5.4.1, one must be aware of two facts. First, the computation of all the circuits of V by algorithm 5.5.1 can be time consuming, since it requires the exploration of a tree, whose nodes at level k may have up to $p-k$ descendants. Second, determining whether a sign vector covers a stem vector can also take much computing time when the number of stem vectors is large, which is usually the case when p is large (see remark 5.3.13(6)).

5.5.2 Algorithms using some stem vectors

Instead of computing the stem vectors exhaustively like algorithm 5.5.1 does, which is generally a time consuming task, one can get a few stem vectors from the optimal dual variables of some linear optimization problems (LOPs) encountered in algorithm 5.4.1, those that are associated with an infeasible sign vector. This device is described in section 5.5.2. Then, one can design a kind of primal-dual algorithm for computing $\mathcal{S}(V,\tau)$. This one builds the $\mathcal{S}$-tree, but, in order to save running time, it makes use of the stem vectors collected during its construction to prune some unfruitful branches of the $\mathcal{S}$-tree, which avoids having

to solve some LOPs. This algorithm is presented in section 5.5.2.

Getting stem vectors from linear optimization

In line 11 of algorithm 5.4.2, one has to decide whether $(s, -s_{k+1})$ is in $\mathcal{S}_{k+1}$ and it is suggested, after the description of the algorithm, to determine this belonging by solving the linear optimization problem (LOP) (3.43). The Lagrangian dual of this problem [29, 105] reads

$$\begin{array}{rl} \max_{(\lambda,\mu)\in\mathbb{R}^{k+1}\times\mathbb{R}} & \sum_{i\in[1:k]} \lambda_i s_i \tau_i - \lambda_{k+1} s_{k+1} \tau_{k+1} - \mu \\ \text{s.t.} & \lambda \geqslant 0, \mu \geqslant 0 \\ & \sum_{i\in[1:k]} \lambda_i s_i v_i = \lambda_{k+1} s_{k+1} v_{k+1} \\ & \sum_{i\in[1\,:\,k+1]} \lambda_i + \mu = 1, \end{array} \tag{5.41}$$

where $\lambda \in \mathbb{R}^{k+1}$ is the dual variable associated with the first $k+1$ constraints of (3.43) and μ the dual variable associated with its last constraint.

The next proposition gives conditions ensuring that a circuit of V can be obtained from a specific solution to the dual problem (5.41) when $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}$. We denote by $\mathrm{val}\,(3.43)$ (resp. $\mathrm{val}\,(5.41)$) the optimal value of the primal (resp. dual) optimization problem (3.43) (resp. (5.41)). By strong duality in linear optimization [29, 105] and the fact that problem (3.43) has a solution, one has $\mathrm{val}\,(3.43) = \mathrm{val}\,(5.41)$.

Proposition 5.5.6 (matroid circuit detection from optimization)**.**

1) *Problem* (5.41) *has a solution, say* $(\lambda, \mu) \in \mathbb{R}^{k+1}_+ \times \mathbb{R}_+$.
2) *If* $s \in \mathcal{S}_k$ *and* $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}$, *then* $\mathrm{val}\,(3.43) \geqslant 0$, $\lambda_{k+1} > 0$ *and* $\mu = 0$.
3) *If, in addition,* (λ, μ) *is an extreme point of the feasible set of* (5.41)*, then*
   - $J := \{i \in [1:k+1] : \lambda_i > 0\} \in \mathcal{C}(V)$,
   - *if* $\mathrm{val}\,(3.43) = 0$, $\pm(s, -s_{k+1})_J$ *are the two symmetric stem vectors associated with* J,
   - *if* $\mathrm{val}\,(3.43) > 0$, $(s, -s_{k+1})_J$ *is the unique asymmetric stem vector associated with* J.

Proof. 1) By strong duality in linear optimization [224, 29, 105], the fact that the primal problem (3.43) has a solution implies that the dual problem (5.41) has also a solution, say (λ, μ).

2) Suppose that $s \in \mathcal{S}_k$ and that $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}$. Let (x, α) be a solution to (3.43) ($\alpha = \mathrm{val}\,(3.43)$ is uniquely determined). Let us show that

$$\lambda_{k+1} > 0 \qquad \text{and} \qquad \mu = 0. \tag{5.42a}$$

The optimal multiplier μ is associated with the constraint $\alpha \geqslant -1$ of the optimization problem (3.43), which is inactive ($\alpha \geqslant 0$ when $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}$), so that it vanishes. We show that $\lambda_{k+1} > 0$ by contradiction, assuming that $\lambda_{k+1} = 0$. Then, strong duality would imply that $0 \leqslant \alpha = \mathrm{val}\,(3.43) = \mathrm{val}\,(5.41) = \sum_{i\in[1:k]} \lambda_i s_i \tau_i$, while the third constraint of (5.41) would read $\sum_{i\in[1:k]} \lambda_i s_i v_i = 0$. Then, Motzkin's alternative (5.1) would imply that

there is no $x \in \mathbb{R}^n$ such that $s_i(v_i^\mathsf{T} x - \tau_i) > 0$, for $i \in [1:k]$, in contradiction with the assumption $s \in \mathcal{S}_k$.

3) Let $I := \{i \in [1:k],\, \lambda_i > 0\}$. By assumption and $\mu = 0$, $(\lambda, 0)$ is an extreme point of the feasible set of problem (5.41), which implies that the vectors [50, 224, 105]

$$\left\{ \begin{pmatrix} s_i v_i \\ 1 \end{pmatrix}_{i \in I}, \begin{pmatrix} -s_{k+1} v_{k+1} \\ 1 \end{pmatrix} \right\} \text{ are linearly independent,} \tag{5.42b}$$

where we used the fact that $\lambda_{k+1} > 0$ and $\mu = 0$ by (5.42a).

One can deduce from this property that the vectors

$$\{s_i v_i\}_{i \in I} \text{ are linearly independent.} \tag{5.42c}$$

Suppose indeed that $\sum_{i \in I} \alpha_i s_i v_i = 0$ for some real numbers $(\alpha_i)_{i \in I}$. It suffices to show that these numbers vanish and we do so in two steps.

- We first show by contradiction that $\sum_{i \in I} \alpha_i = 0$. If this were not the case, one could find $t \in \mathbb{R}$ such that $\sum_{i \in I}(\lambda_i + t\alpha_i) + \lambda_{k+1} = 0$. Now, using the third constraint of problem (5.41), we would have that $\eta := ((\lambda_i + t\alpha_i)_{i \in I}, \lambda_{k+1})$ is in the null space of the nonsingular matrix whose columns are the vectors in (5.42b), which would imply that $\eta = 0$, in contradiction with $\lambda_{k+1} > 0$ imposed by (5.42a).
- Using $\sum_{i \in I} \alpha_i = 0$ and $\sum_{i \in I} \alpha_i s_i v_i = 0$, we have that the vector $((\alpha_i)_{i \in I}, 0)$ is in the null space of the nonsingular matrix whose columns are the vectors in (5.42b). Hence all the α_i's vanish.

Now, set $J := \{i \in [1\,{:}\,k+1] : \lambda_i > 0\}$, which is $I \cup \{k+1\}$ by the definition of I and (5.42a), and introduce the diagonal matrix $D \in \mathbb{R}^{J \times J}$ defined by $D_{i,i} = s_i$ if $i \in I$ and $D_{k+1,k+1} = -s_{k+1}$. Using (5.42c), we see that

$$\operatorname{null}(V_{:,J} D) = 1.$$

By the third constraint of (5.41), we have that $\lambda_J \in \mathcal{N}(V_{:,J} D) \setminus \{0\}$. Since $\lambda_J > 0$, proposition 5.3.11 tells us that J is a circuit of $V_{:,J} D$, hence a circuit of V.

Since $\eta := D\lambda_J \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ is such that $\tau_J^\mathsf{T} \eta = \operatorname{val}$ (3.43), we see that the number of stem vectors associated with J is governed by val (3.43), as described in remark 5.3.13(3). In addition, $\operatorname{sgn}(\eta) = (s, -s_{k+1})_J$, because $\lambda_J > 0$, showing that $(s, -s_{k+1})_J$ is a stem vector. □

A solution to problem (5.41) that is an extreme point of its feasible set can be obtained by the dual-simplex algorithm. Note that, since $\lambda_{k+1} > 0$, $k+1$ always belongs to the selected circuit J of V.

Primal-dual $\mathcal{S}$-tree algorithm

Proposition 5.5.6(3) shows how circuits and their associated stem vectors can be obtained when the $\mathcal{S}$-tree primal algorithm 5.4.1 solves a LOP (3.43) with an appropriate solver and observes that the sign vector $(s, -s_{k+1})$ is infeasible. Now, with the *partial* list of stem vectors so computed, which grows throughout the iterations, the algorithm can detect *some* infeasible sign vectors by using proposition 5.3.16, like in the crude dual algorithm 5.5.3 or in the $\mathcal{S}$-tree dual algorithm 5.5.4, but without having to solve a LOP. In practice, this technique saves much computing time. Here is this *primal-dual $\mathcal{S}$-tree algorithm*, based on the just presented idea, which has many similarities with the primal $\mathcal{S}$-tree algorithm 5.4.1.

Algorithm 5.5.7 (PD_STREE). // primal-dual $\mathcal{S}$-tree algorithm

1. $\mathfrak{S} = \varnothing$
2. PD_STREE_REC($+1$, x_+, $\mathfrak{S}$) // x_+ given by $(5.34)_1$
3. PD_STREE_REC(-1, x_-, $\mathfrak{S}$) // x_- given by $(5.34)_2$

Algorithm 5.5.8 (PD_STREE_REC(s $\in \{\pm 1\}^k$, $x \in \mathbb{R}^n$, $\mathfrak{S}$)). 1. IF ($k = p$)
2. Output s and RETURN // s is a leaf of the $\mathcal{S}$-tree ; end the recursion
3. ENDIF
4. IF ($v_{k+1}^{\mathsf{T}} x = \tau_{k+1}$)
5. PD_STREE_REC($(s, +1), x + \varepsilon v_{k+1}$, $\mathfrak{S}$) // $(s, +1) \in \mathcal{S}_{k+1}$
6. PD_STREE_REC($(s, -1), x - \varepsilon v_{k+1}$, $\mathfrak{S}$) // $(s, -1) \in \mathcal{S}_{k+1}$
7. RETURN
8. ENDIF
9. $s_{k+1} := \operatorname{sgn}(v_{k+1}^{\mathsf{T}} x - \tau_{k+1})$ // $(s, s_{k+1}) \in \mathcal{S}_{k+1}$
10. PD_STREE_REC($(s, s_{k+1}), x$, $\mathfrak{S}$)
11. IF ($(s, -s_{k+1})$ covers a stem vector of $\mathfrak{S}$)
12. RETURN
13. ELSEIF ($(s, -s_{k+1})$ is feasible with witness point $\tilde{x}$)
14. PD_STREE_REC($(s, -s_{k+1}), \tilde{x}$, $\mathfrak{S}$) // $(s, -s_{k+1}) \in \mathcal{S}_{k+1}$
15. ELSE
16. Add one or two stem vectors to $\mathfrak{S}$
17. ENDIF

We only comment some instructions of the primal-dual $\mathcal{S}$-tree algorithm 5.5.7 that differ from those of the primal $\mathcal{S}$-tree algorithm 5.4.1.

- Unlike algorithm 5.5.4, which computes all the stem vectors at first, algorithm 5.5.7 initializes the list of stem vectors $\mathfrak{S}$ to the empty set in line 1. This list is next gradually filled by algorithm 5.5.8.
- For more efficiency, one could adapt line 4 of algorithm 5.5.8 and its lines 5..6 by using the improvement described in section 5.4.2.

- Lines 11..12 are new with respect to algorithm 5.4.1. They are used to check whether $(s, -s_{k+1}) \in \mathcal{S}^c_{k+1}$, using the stem vectors collected in $\mathfrak{S}$ and proposition 5.3.16, without having to solve a LOP.
- Lines 15..16 are also new with respect to algorithm 5.4.1. They use proposition 5.5.6(3) to detect a new circuit, hence one or two new stem vectors (they are new, since otherwise the test in line 11 would have been successful and line 16 would not have been executed), which are put in $\mathfrak{S}$. For this, it is necessary to solve the LOP in line 13 by a method computing an extreme point of the dual feasible set of this problem.

Algorithm 5.5.7 can be improved by introducing the modifications A, B and C of sections 5.4.2–5.4.2.

5.6 Compact version of the algorithms

All the algorithms computing the sign vector set $\mathcal{S}(V, \tau)$ presented so far, except algorithm 5.5.3, recursively construct the $\mathcal{S}$-tree introduced in algorithm 5.4.1, namely (recall the definition (5.33) of $\mathcal{S}_k(V, \tau)$)

$$\mathcal{T}(V, \tau) := \bigcup_{k \in [1:p]} \mathcal{S}_k(V, \tau). \tag{5.43}$$

When the arrangement is not centered (equivalently, $\tau \notin \mathcal{R}(V^\mathsf{T})$), some sets $\mathcal{S}_k(V, \tau)$ are asymmetric (proposition 5.3.5), so that the sign vectors of the two subtrees $\mathcal{T}^+(V, \tau) := \{s \in \mathcal{T}(V, \tau) : s_1 = +1\}$ and $\mathcal{T}^-(V, \tau) := \{s \in \mathcal{T}(V, \tau) : s_1 = -1\}$ of the $\mathcal{S}$-tree, rooted at the nodes $\{+1\}$ and $\{-1\}$, respectively, are not opposite to each other. Therefore, one cannot just compute $\mathcal{T}^+(V, \tau)$ or $\mathcal{T}^-(V, \tau)$ to get all $\mathcal{T}(V, \tau)$ (recall that when the arrangement is centered, $\mathcal{S}(V, \tau) = \mathcal{S}(V, 0)$ and only half of the sign vectors needs to be computed, see [77]). Nevertheless, these two subtrees have some opposite sign vectors, the symmetric ones, those in $\mathcal{T}(V, 0) = \cup_{k \in [1:p]} \mathcal{S}_k(V, 0)$. The set of asymmetric sign vectors in $\mathcal{T}(V, \tau)$ is denoted by

$$\mathcal{T}_a(V, \tau) := \bigcup_{k \in [1:p]} \mathcal{S}_{a,k}(V, \tau),$$

where $\mathcal{S}_{a,k}(V, \tau) := \mathcal{S}_k(V, \tau) \setminus \mathcal{S}_{s,k}(V, \tau)$. Therefore, it is natural to look for a way to avoid as much as possible repeating the costly operations (linear optimization problems or stem vector coverings) common to the construction of the two subtrees $\mathcal{T}^+(V, \tau)$ and $\mathcal{T}^-(V, \tau)$. The goal of this section is to propose algorithms having that property; they can have a primal or dual nature.

5.6.1 The compact $\mathcal{S}$-tree

For an arrangement $\mathcal{A}(V,\tau)$, with $V \in \mathbb{R}^{n\times p}$ and $\tau \in \mathbb{R}^p$, and for $k \in [1:p]$, we denote the arrangement associated with the first k columns of V and the first k components of τ by

$$\mathcal{A}_k(V,\tau) := \mathcal{A}(V_{:,[1:k]}, \tau_{[1:k]}).$$

By proposition 5.3.18, we have that

$$\mathcal{S}_k(V,0) \subseteq \mathcal{S}_k(V,\tau) \subseteq \mathcal{S}_k([V;\tau^\mathsf{T}],0) \tag{5.44a}$$
$$\mathcal{S}_k([V;\tau^\mathsf{T}],0) \setminus \mathcal{S}_k(V,0) = \mathcal{S}_{a,k}(V,\tau) \uplus \mathcal{S}_{a,k}(V,-\tau). \tag{5.44b}$$

The algorithms described in this section are based on the following considerations. By (5.10), the set $\mathcal{T}(V,\tau)$ of the feasible sign vectors of the $\mathcal{S}$-tree can be written $\mathcal{T}(V,0) \uplus \mathcal{T}_a(V,\tau)$. Taking the intersection with $\mathcal{T}^+(V,\tau)$ and $\mathcal{T}^-(V,\tau)$ partitions $\mathcal{T}(V,\tau)$ into four sets :

$$\mathcal{T}(V,0)\cap\mathcal{T}^+(V,\tau),\ \mathcal{T}_a(V,\tau)\cap\mathcal{T}^+(V,\tau),\ \mathcal{T}(V,0)\cap\mathcal{T}^-(V,\tau),\ \mathcal{T}_a(V,\tau)\cap\mathcal{T}^-(V,\tau). \tag{5.45}$$

Since

$$\mathcal{T}(V,0) \cap \mathcal{T}^-(V,\tau) = -[\mathcal{T}(V,0) \cap \mathcal{T}^+(V,\tau)], \tag{5.46}$$

only two sets must be computed to be able to retrieve all the sign vectors of $\mathcal{T}(V,\tau)$, namely the union of the first two sets of the partition (5.45) and the last one :

$$\mathcal{T}^+(V,\tau) \qquad \text{and} \qquad \mathcal{T}_a(V,\tau) \cap \mathcal{T}^-(V,\tau).$$

The principle of the algorithms described in this section consists in computing the subtree $\mathcal{T}^+(V,\tau)$ rooting at $s_1 = +1$ and in grafting to it the subtrees of

$$-[\mathcal{T}_a(V,\tau) \cap \mathcal{T}^-(V,\tau)],$$

which is in $\mathcal{T}_a(V,-\tau)$. This forms what we call the *compact $\mathcal{S}$-tree*. More precisely, if $s \in \mathcal{S}_k(V,0) \cap \mathcal{T}^+(V,\tau)$ and $(-s, s_{k+1}) \in \mathcal{S}_{a,k+1}(V,\tau) \cap \mathcal{T}^-(V,\tau)$ for some $s_{k+1} \in \{\pm 1\}$, the subtree of $\mathcal{T}^-(V,\tau)$ rooting at $(-s, s_{k+1})$ is grafted at s in the compact tree (with its sign vectors multiplied by -1, so that $(s, -s_{k+1})$ can be a child of s). As a result, the nodes of the level k of the compact $\mathcal{S}$-tree are in one of the sets

$$\mathcal{S}_k(V,0), \quad \mathcal{S}_{a,k}(V,\tau) \quad \text{or} \quad \mathcal{S}_{a,k}(V,-\tau). \tag{5.47}$$

Eventually, a sign vector $s \in \mathcal{S}_a(V,-\tau)$ must be multiplied by -1 to get it in $-\mathcal{S}_a(V,-\tau) = \mathcal{S}_a(V,\tau) \subseteq \mathcal{S}(V,\tau)$. This principle is illustrated in figure 5.5. Housekeeping is done by attaching a flag $\boxed{\text{S}}$ to each node s of the resulting tree, in order to specify which of the sign vector sets listed in (5.47) s belongs. As claimed in point 5 of the next proposition, the grafting process does not introduce nodes with two different flags : if $(s, s_{k+1}) \in -[\mathcal{S}_{a,k+1}(V,\tau)\cap\mathcal{T}^-(V,\tau)]$ is grafted to the compact $\mathcal{S}$-tree, then (s, s_{k+1}) is not in $\mathcal{T}^+(V,\tau)$.

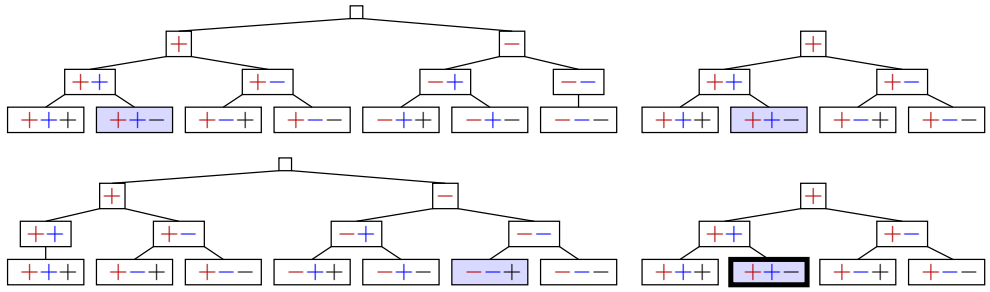


FIGURE 5.5 – Standard $\mathcal{S}$-trees (left) and compact $\mathcal{S}$-trees (right) of the arrangements in the middle pane (above, compare with figure 5.4) and the right-hand side pane (below) of figure 5.1. The sign vectors in the white boxes are in $\mathcal{T}(V,0)$, those in the blue/gray boxes are in $\mathcal{S}_a(V,\tau)$ and the one in the blue/gray box with bold edges is in $\mathcal{S}_a(V,-\tau)$; this last sign vector must be multiplied by -1 to get a sign vector in $-\mathcal{S}_a(V,-\tau) = \mathcal{S}_a(V,\tau) \subseteq \mathcal{S}(V,\tau)$.

Proposition 5.6.1 (compact $\mathcal{S}$-tree). *Let $k \in [1:p-1]$ and let $s \in \mathcal{S}_k(V,0) \uplus \mathcal{S}_{a,k}(V,\tau) \uplus \mathcal{S}_{a,k}(V,-\tau)$ be a sign vector of the compact $\mathcal{S}$-tree. Set $\mathcal{S}_k^{\pm}(V,\tau) := \mathcal{S}_k(V,\tau) \cap \mathcal{T}^+(V,\tau)$.*

1) *If $s \in \mathcal{S}_k(V,0)$, one child of s in the compact $\mathcal{S}$-tree is in $\mathcal{S}_{k+1}(V,0)$.*
2) *If $s \in \mathcal{S}_{a,k}(V,\tau)$, the children of s in the compact $\mathcal{S}$-tree are in $\mathcal{S}_{a,k+1}(V,\tau)$.*
3) *If $s \in \mathcal{S}_{a,k}(V,-\tau)$, the children of s in the compact $\mathcal{S}$-tree are in $\mathcal{S}_{a,k+1}(V,-\tau)$.*
4) *If $(s,s_{k+1}) \in -[\mathcal{S}_{a,k+1}(V,\tau) \cap \mathcal{T}^-(V,\tau)]$ with $s_{k+1} \in \{\pm 1\}$, then $(s,s_{k+1}) \notin \mathcal{T}^+(V,\tau)$.*
5) *Level k of the compact $\mathcal{S}$-tree is formed of $\mathcal{S}_k^+(V,\tau) \uplus (-[\mathcal{S}_{a,k}(V,\tau) \cap \mathcal{T}^-(V,\tau)])$.*

5.6.2 Compact primal $\mathcal{S}$-tree algorithm

In accordance with the presentation of section 5.6.1, the *compact primal $\mathcal{S}$-tree algorithm*, whose reasoned description is given below, ignores the subtree $\mathcal{T}^-(V,\tau)$ rooting at $\{+1\}$, constructs the subtree $\mathcal{T}^+(V,\tau)$ rooting at $\{+1\}$ and grafts to it the opposite of the sign vectors in the subtrees of $\mathcal{T}_a(V,\tau) \cap \mathcal{T}^-(V,\tau)$. Note that $s_1 \in \mathcal{S}_1(V,0)$. Let us describe this algorithm. Its formal statement is given afterwards.

The algorithm identifies each node at level k of the compact $\mathcal{S}$-tree by a triplet $(s,x,\boxed{\mathrm{s}})$, where $s \in \{\pm 1\}^k$ is the sign vector of the node, $\boxed{\mathrm{s}} \in \{-1,0,+1\}$ is a flag specifying to which sign set s belongs and x is some witness point. More specifically,

$$\begin{cases} s \in \mathcal{S}_k(V,0),\ x \text{ is a witness point for } s \text{ in } \mathcal{A}_k(V,0) & \text{if } \boxed{\mathrm{s}} = \ \ 0, \\ s \in \mathcal{S}_{a,k}(V,\tau),\ x \text{ is a witness point for } s \text{ in } \mathcal{A}_k(V,\tau) & \text{if } \boxed{\mathrm{s}} = +1, \\ s \in \mathcal{S}_{a,k}(V,-\tau),\ x \text{ is a witness point for } s \text{ in } \mathcal{A}_k(V,-\tau) & \text{if } \boxed{\mathrm{s}} = -1. \end{cases} \tag{5.48}$$

The flag $\boxed{\mathrm{s}}$ is used below as a scalar, hence $\boxed{\mathrm{s}}\, s$ is the vector whose ith component is $\boxed{\mathrm{s}}\, s_i$. The initialization of the algorithm is done as follows.

0. Take $s_1 = +1 \in \mathcal{S}_1(V, 0)$ and v_1 as witness point for s_1 in $\mathcal{A}_1(V, 0)$.

Consider now a node at level k of the compact $\mathcal{S}$-tree, which is specified by a triplet $(s, x, \boxed{\mathrm{S}})$ satisfying (5.48). We just have to specify how the algorithm determines the children of that node.

1. Suppose that $s \in \mathcal{S}_k(V, 0)$ with x as witness point in $\mathcal{A}_k(V, 0)$, i.e., $\boxed{\mathrm{S}} = 0$.

   Using proposition 5.4.7 with $\tau = 0$, the algorithm can detect whether s has two easily computable children in $\mathcal{A}(V, 0)$ and can find associated witness points. If such is the case, the algorithm pursues recursively from $(s, +1)$ and $(s, -1)$, with appropriate witness points. It returns afterwards.

   Otherwise, $v_{k+1}^\mathsf{T} x \neq 0$, implying that $(s, s_{k+1}) \in \mathcal{S}_{k+1}(V, 0)$ for $s_{k+1} := \mathrm{sgn}(v_{k+1}^\mathsf{T} x)$ and that the algorithm can pursue recursively from (s, s_{k+1}) with x as witness point in $\mathcal{A}_{k+1}(V, 0)$.

   Now, the algorithm specifies to what set $(s, -s_{k+1})$ belongs : $\mathcal{S}_{k+1}(V, 0)$, $\mathcal{S}_{a,k+1}(V, \tau)$, $\mathcal{S}_{a,k+1}(V, -\tau)$ or $\mathcal{S}_{k+1}([V; \tau^\mathsf{T}], 0)^c$ (there are no other possibilities, see proposition 5.3.18 and figure 5.3). For this purpose, the compact primal $\mathcal{S}$-tree algorithm starts by solving the LOP (3.43) with $\tau = 0$, to see whether $(s, -s_{k+1}) \in \mathcal{S}_{k+1}(V, 0)$. Denote by (x_0, α_0) a solution to this LOP.

   1.1. If $\alpha_0 < 0$, then $(s, -s_{k+1}) \in \mathcal{S}_{k+1}(V, 0)$ and the algorithm pursues recursively from $(s, -s_{k+1})$ with x_0 as witness point in $\mathcal{A}_{k+1}(V, 0)$.

   1.2. Otherwise, $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}(V, 0)$ and the algorithm determines if $(s, -s_{k+1}) \in \mathcal{S}_{k+1}([V; \tau^\mathsf{T}], 0)$ by solving the following LOP, which is similar to (3.43) with $\tau = 0$, but for the arrangement $\mathcal{A}([V; \tau^\mathsf{T}], 0)$ instead of $\mathcal{A}(V, 0)$:

$$\begin{array}{rl} \min_{(x,\xi,\alpha)\in\mathbb{R}^n\times\mathbb{R}\times\mathbb{R}} & \alpha \\ \text{s.t.} & s_i(v_i^\mathsf{T} x + \tau_i \xi) + \alpha \geqslant 0, \quad \text{for } i \in [1:k] \\ & -s_{k+1}(v_{k+1}^\mathsf{T} x + \tau_{k+1}\xi) + \alpha \geqslant 0 \\ & \alpha \geqslant -1. \end{array} \tag{5.49}$$

   Denote by (x_1, ξ_1, α_1) a solution to this problem.

   1.2.1. If $\alpha_1 \geqslant 0$, then $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}([V; \tau^\mathsf{T}], 0)$, hence $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}(V, \tau) \cup \mathcal{S}_{k+1}(V, -\tau)$ by (5.44a) and $(s, -s_{k+1})$ can be discarded from the generated compact tree.

   1.2.2. If not, $\alpha_1 < 0$ and $(s, -s_{k+1}) \in \mathcal{S}_{k+1}([V; \tau^\mathsf{T}], 0) \setminus \mathcal{S}_{k+1}(V, 0) = \mathcal{S}_{a,k+1}(V, \tau) \uplus \mathcal{S}_{a,k+1}(V, -\tau)$ by (5.44b). Note that one cannot have $\xi_1 = 0$ in that case, since then one would have $(s, -s_{k+1}) \in \mathcal{S}_{k+1}(V, 0)$, which is excluded in the considered case. Therefore,

   - either $\xi_1 < 0$ and $(s, -s_{k+1}) \in \mathcal{S}_{a,k+1}(V, \tau)$, by (5.49), with $-x_1/\xi_1$ as witness point in $\mathcal{A}_{k+1}(V, \tau)$,

– or $\xi_1 > 0$ and $(s, -s_{k+1}) \in \mathcal{S}_{a,k+1}(V, -\tau)$, by (5.49), with x_1/ξ_1 as witness point in $\mathcal{A}_{k+1}(V, -\tau)$.

In these last two cases, the algorithm pursues recursively from $(s, -s_{k+1})$.

2. Suppose that $s \in \mathcal{S}_{a,k}(V, \tau)$ with x as witness point in $\mathcal{A}_k(V, \tau)$, i.e., $\boxed{\mathrm{S}} = +1$.

   Using proposition 5.4.7, the algorithm can detect whether s has two easily computable children in $\mathcal{A}(V, \tau)$ and can find associated witness points. If such is the case, the algorithm pursues recursively from $(s, +1)$ and $(s, -1)$, with appropriate witness points. It returns afterwards.

   Otherwise, $v_{k+1}^{\mathsf{T}} x \neq \tau_{k+1}$, implying that $(s, s_{k+1}) \in \mathcal{S}_{k+1}(V, \tau)$ for $s_{k+1} := \mathrm{sgn}(v_{k+1}^{\mathsf{T}} x - \tau_{k+1})$ and that the algorithm can pursue recursively from (s, s_{k+1}) with x as witness point in $\mathcal{A}_{k+1}(V, \tau)$.

   Now, the algorithm must determine whether $(s, -s_{k+1})$ is infeasible or is in $\mathcal{S}_{a,k+1}(V, \tau)$ (there are no other possibilities, according to proposition 5.6.1(2)). For this purpose, the algorithm solves (3.43). Let (x_2, α_2) be a solution.

   – If $\alpha_2 < 0$, then $(s, -s_{k+1}) \in \mathcal{S}_{a,k+1}(V, \tau)$ and the compact algorithm can pursue recursively from $(s, -s_{k+1})$.
   – Otherwise, $(s, -s_{k+1})$ is infeasible in $\mathcal{A}_{k+1}(V, \tau)$ and the algorithm can prune the compact $\mathcal{S}$-tree at that node.

3. The last case, when $s \in \mathcal{S}_{a,k}(V, -\tau)$, with x as witness point in $\mathcal{A}_k(V, -\tau)$, i.e., $\boxed{\mathrm{S}} = -1$, is similar to case 2 and is detailed in [80].

One can now present the compact form of the primal $\mathcal{S}$-tree algorithm 5.4.1. To shorten its statement and the one of the next algorithm 5.6.7, we introduce the following function OUTPUT_S, which outputs sign vectors of $\mathcal{S}(V, \tau)$ at a leaf of the compact $\mathcal{S}$-tree (its behavior is more complex than for the standard algorithms and depends on the type $\boxed{\mathrm{S}}$ of the leaf node s, see (5.48)), and C_P_TWO_CHILDREN, which detects whether s has the two children that are given by proposition 5.4.7; if this is the case, it pursues the compact $\mathcal{S}$-tree construction at $(s, \pm 1)$ and returns TRUE; otherwise, it returns FALSE.

Algorithm 5.6.2 (OUTPUT_S(s, $\boxed{\mathrm{S}}$))**.**
It is assumed that $s \in \{\pm 1\}^p$ and that $\boxed{\mathrm{S}} \in \{-1, 0, +1\}$.

1. IF ($\boxed{\mathrm{S}} = 0$)
2. OUTPUT $\pm s$ // $s \in \mathcal{S}(V, 0)$
3. ELSE
4. OUTPUT $\boxed{\mathrm{S}}\, s$ // $s \in \mathcal{S}_a(V, \boxed{\mathrm{S}}\, \tau)$
5. ENDIF

Algorithm 5.6.3 (C_P_TWO_CHILDREN(S, x, $\boxed{\mathrm{S}}$))**.**
It is assumed that $s \in \{\pm 1\}^k$ and that $(s, x, \boxed{\mathrm{S}})$ satisfies (5.48).

1. IF ($v_{k+1}^{\mathsf{T}} x \simeq \boxed{\mathrm{s}}\,\tau_{k+1}$) // two easy children in $\mathcal{A}_{k+1}(V, \boxed{\mathrm{s}}\,\tau)$
2. C_P_STREE_REC($(s, +1)$, $x + t_+ v_{k+1}$, $\boxed{\mathrm{s}}$) for some $t_+ \in (t_0, t_{\max})$
3. C_P_STREE_REC($(s, -1)$, $x + t_- v_{k+1}$, $\boxed{\mathrm{s}}$) for some $t_- \in (t_{\min}, t_0)$
4. RETURN TRUE
5. ELSE
6. RETURN FALSE
7. ENDIF

We have chosen to present the cases when $\boxed{\mathrm{s}} = \pm 1$ jointly in lines 22..25, to save space. An expanded presentation is given in [80].

Algorithm 5.6.4 (C_P_STREE)**.** Let be given $V \in \mathbb{R}^{n \times p}$ and $\tau \in \mathbb{R}^p$.

1. C_P_STREE_REC($+1$, v_1, 0)

Algorithm 5.6.5 (C_P_STREE_REC(S, x, $\boxed{\mathrm{s}}$))**.**
It is assumed that $s \in \{\pm 1\}^k$ and that $(s, x, \boxed{\mathrm{s}})$ satisfies (5.48).

1. IF ($k = p$) // s is a leaf of the compact $\mathcal{S}$-tree
2. OUTPUT_S(S, $\boxed{\mathrm{s}}$)
3. RETURN
4. ENDIF
5. IF (C_P_TWO_CHILDREN(S, x, $\boxed{\mathrm{s}}$)) // two easy children in $\mathcal{A}_{k+1}(V, \boxed{\mathrm{s}}\,\tau)$
6. RETURN
7. ENDIF
8. $s_{k+1} := \operatorname{sgn}(v_{k+1}^{\mathsf{T}} x - \boxed{\mathrm{s}}\,\tau_{k+1})$ // $(s, s_{k+1}) \in \mathcal{S}_{k+1}(V, \boxed{\mathrm{s}}\,\tau)$
9. C_P_STREE_REC((s, s_{k+1}), x, $\boxed{\mathrm{s}}$)
10. IF ($\boxed{\mathrm{s}} = 0$) // $s \in \mathcal{S}_k(V, 0)$ with x as witness point in $\mathcal{A}_k(V, 0)$
11. Solve (3.43) with $\tau = 0$; let (x_0, α_0) be a solution
12. IF ($\alpha_0 < 0$) // $(s, -s_{k+1}) \in \mathcal{S}_{k+1}(V, 0)$
13. C_P_STREE_REC($(s, -s_{k+1})$, x_0, 0)
14. ELSE
15. Solve (5.49); let (x_1, ξ_1, α_1) be a solution // here $\xi_1 \neq 0$
16. IF ($\alpha_1 < 0$) // $(s, -s_{k+1}) \in \mathcal{S}_{a,k+1}(V, -\operatorname{sgn}(\xi_1)\tau)$
17. C_P_STREE_REC($(s, -s_{k+1})$, $x_1/|\xi_1|$, $-\operatorname{sgn}(\xi_1)$)
18. ENDIF
19. ENDIF
20. RETURN
21. ENDIF // here $\boxed{\mathrm{s}} \in \{\pm 1\}$, hence $s \in \mathcal{S}_{a,k}(V, \boxed{\mathrm{s}}\,\tau)$
22. Solve (3.43) with $\tau \curvearrowright \boxed{\mathrm{s}}\,\tau$; let (x_2, α_2) be a solution
23. IF ($\alpha_2 < 0$) // $(s, -s_{k+1}) \in \mathcal{S}_{a,k+1}(V, \boxed{\mathrm{s}}\,\tau)$
24. C_P_STREE_REC($(s, -s_{k+1})$, x_2, $\boxed{\mathrm{s}}$)
25. ENDIF

Observe that, as claimed by proposition 5.6.1(2-3), once $s \in \mathcal{S}_{a,k}(V,\tau)$ (resp. $s \in \mathcal{S}_{a,k}(V,-\tau)$), its descendants in the compact $\mathcal{S}$-tree are all in $\mathcal{S}_{a,l}(V,\tau)$ (resp. $\mathcal{S}_{a,l}(V,-\tau)$) for some $l \in [k+1:p]$. In these cases, the compact algorithm solves at most one LOP per sign vector, like in the standard version of the algorithm, which solves at most one LOP in $\mathcal{T}^+(V,\tau)$ or $\mathcal{T}^-(V,\tau)$, not both since an asymmetric sign vector only appears in one of these subtrees. When $s \in \mathcal{S}_k(V,0)$ has one child in $\mathcal{S}_{a,k+1}(V,\pm\tau)$, the compact algorithm solves two LOPs (in steps 11 and 15), like in the standard algorithm (one LOP in $\mathcal{T}^{\pm}(V,\tau)$ to accept the child in $\mathcal{S}_{a,k+1}(V,\tau)$ and one in $\mathcal{T}^{\mp}(V,\tau)$ to reject a child in $\mathcal{S}_{a,k+1}(V,\tau)$). The only sign vectors at which the compact algorithm solves less LOPs than the standard algorithm are those in $\mathcal{S}(V,0)$ with two symmetric children. In this case, the compact algorithm solves a single LOP (in step 11), while the standard algorithm solves two LOPs (one in each subtree $\mathcal{T}^+(V,\tau)$ and $\mathcal{T}^-(V,\tau)$). Therefore, the compact algorithm 5.6.4 is all the more advantageous with respect to the standard algorithm 5.4.1 as $|\mathcal{T}(V,0)|/|\mathcal{T}(V,\tau)|$ is large (it is always $\leqslant 1$).

5.6.3 Compact primal-dual $\mathcal{S}$-tree algorithm

There are several ways of using the stem vectors, in order to avoid having to solve all or part of the LOPs of the standard primal $\mathcal{S}$-tree algorithms 5.4.1, most of them having a compact form. In this section, we only consider a compact version of the primal-dual $\mathcal{S}$-tree algorithm 5.5.7. The statement of the algorithm is immediate as soon as we know how it collects the stem vectors and how it uses them. This is essentially what we clarify in this section, leaving a complete description to [80].

As shown in section 5.5, stem vectors can be used to detect sign vectors that are not in $\mathcal{S}(V,\tau)$, using proposition 5.3.16. Our goal in this section is to apply this technique to construct the compact primal-dual $\mathcal{S}$-tree, by "dualizing" the compact primal $\mathcal{S}$-tree algorithm 5.6.4 (we have found this approach easier than "compacting" the standard primal-dual $\mathcal{S}$-tree algorithm 5.5.7). The principle is simple. The algorithm manages subsets $\tilde{\mathfrak{S}}_s$ of $\mathfrak{S}_s(V,\tau)$, $\tilde{\mathfrak{S}}_a$ of $\mathfrak{S}_a(V,\tau)$ and $\tilde{\mathfrak{S}}_0$ of $\mathfrak{S}_0(V,\tau)$, named *collectors*, which are initially empty and are progressively filled during the iterations (this is explained below). Then, each group of statements in algorithm 5.6.5 (the lines given by the first column of table 5.1), dealing with a LOP and its consequences, is replaced by other lines in algorithm 5.6.8 (those given by the second column of table 5.1). These latter lines are organized as follows.

- So as to avoid having to solve certain LOPs, a covering test is run to see whether the sign vector $(s,-s_{k+1})$ is in the set in the third column of table 5.1 (recall definition (5.33)). Appropriate stem vectors must be used to realize that operation, namely those in the collectors in the fifth column of table 5.1, which are contained in the sets in the fourth column of table 5.1 (see propositions 5.3.14 and 5.3.21, and figure 5.2).
- If the covering test succeeds (i.e., $(s,-s_{k+1})$ covers an appropriate stem vector), then

Algorithm 5.6.5	Algorithm 5.6.8			
Lines	Lines	Sign vector set	Stem vector set	Collectors
11..14	11..19	$\mathcal{S}_{k+1}(V,0)$	$\mathfrak{S}(V,0)$	$\tilde{\mathfrak{S}}_s \cup \tilde{\mathfrak{S}}_a \cup (-\tilde{\mathfrak{S}}_a)$
15..18	20..27	$\mathcal{S}_{k+1}([V;\tau^{\mathsf{T}}],0)$	$\mathfrak{S}([V;\tau^{\mathsf{T}}],0)$	$\tilde{\mathfrak{S}}_s \cup \tilde{\mathfrak{S}}_0$
22..25	30..37	$\mathcal{S}_{a,k+1}(V,\boxed{\mathbb{S}}\tau)$	$\mathfrak{S}(V,\boxed{\mathbb{S}}\tau)$	$\tilde{\mathfrak{S}}_s \cup (\boxed{\mathbb{S}}\tilde{\mathfrak{S}}_a)$

TABLE 5.1 – Corresponding lines in algorithms 5.6.5 and 5.6.8.

$(s, -s_{k+1})$ is not in the sign vector set in the third column of table 5.1 (proposition 5.3.16) and the compact $\mathcal{S}$-tree is pruned.
- Otherwise, because there is no equality between the stem vector sets in the fourth column of table 5.1 and their collectors in the fifth column of table 5.1, a LOP is solved like in algorithm 5.6.4.
- If this LOP has a negative optimal value, $(s, -s_{k+1})$ is in the sign vector set in the third column of table 5.1 and the recursion is proceeded from that node.
- Otherwise, one or two stem vectors are added to the appropriate collectors in the fifth column of table 5.1.

This yields the following algorithm. One first adapts the C_P_TWO_CHILDREN algorithm 5.6.3, so that it calls the appropriate procedures of the present framework.

Algorithm 5.6.6 (C_PD_TWO_CHILDREN(s, x, $\boxed{\mathbb{S}}$)).
It is assumed that $s \in \{\pm 1\}^k$ and $(s, x, \boxed{\mathbb{S}})$ satisfies (5.48).

1. IF ($v_{k+1}^{\mathsf{T}} x \simeq \boxed{\mathbb{S}}\tau_{k+1}$) // two easy children in $\mathcal{A}_{k+1}(V, \boxed{\mathbb{S}}\tau)$
2. C_PD_STREE_REC($(s,+1)$, $x + t_+ v_{k+1}$, $\boxed{\mathbb{S}}$) for some $t_+ \in (t_0, t_{\max})$
3. C_PD_STREE_REC($(s,-1)$, $x + t_- v_{k+1}$, $\boxed{\mathbb{S}}$) for some $t_- \in (t_{\min}, t_0)$
4. RETURN TRUE
5. ELSE
6. RETURN FALSE
7. ENDIF

We can now present the result of the adaptation of algorithm 5.6.4 along the principle described above.

Algorithm 5.6.7 (C_PD_STREE). Let be given $V \in \mathbb{R}^{n\times p}$ and $\tau \in \mathbb{R}^p$.

1. $\tilde{\mathfrak{S}}_s = \varnothing, \tilde{\mathfrak{S}}_a = \varnothing, \tilde{\mathfrak{S}}_0 = \varnothing$ // initial empty collectors
2. C_PD_STREE_REC($+1$, v_1, 0, $\tilde{\mathfrak{S}}_s$, $\tilde{\mathfrak{S}}_a$, $\tilde{\mathfrak{S}}_0$)

Algorithm 5.6.8 (C_PD_STREE_REC(s, x, $\boxed{\mathbb{S}}$, $\tilde{\mathfrak{S}}_s$, $\tilde{\mathfrak{S}}_a$, $\tilde{\mathfrak{S}}_0$)).
It is assumed that $s \in \{\pm 1\}^k$, that $(x, s, \boxed{\mathbb{S}})$ satisfies (5.48) and that $\tilde{\mathfrak{S}}_s \subseteq \mathfrak{S}_s(V,\tau)$, $\tilde{\mathfrak{S}}_a \subseteq \mathfrak{S}_a(V,\tau)$, $\tilde{\mathfrak{S}}_0 \subseteq \mathfrak{S}_0(V,\tau)$.

1. IF ($k = p$) // s is a leaf of the compact $\mathcal{S}$-tree
2. OUTPUT_S(S, $\boxed{\mathrm{S}}$)
3. RETURN
4. ENDIF
5. IF (C_PD_TWO_CHILDREN(S, x, $\boxed{\mathrm{S}}$)) // two easy children in $\mathcal{A}_{k+1}(V, \boxed{\mathrm{S}}\tau)$
6. RETURN
7. ENDIF
8. $s_{k+1} := \operatorname{sgn}(v_{k+1}^{\mathsf{T}} x - \boxed{\mathrm{S}}\tau_{k+1})$ // $(s, s_{k+1}) \in \mathcal{S}_{k+1}(V, \boxed{\mathrm{S}}\tau)$
9. C_PD_STREE_REC((s, s_{k+1}), x, $\boxed{\mathrm{S}}$, $\tilde{\mathfrak{S}}_s$, $\tilde{\mathfrak{S}}_a$, $\tilde{\mathfrak{S}}_0$)
10. IF ($\boxed{\mathrm{S}} = 0$) // $s \in \mathcal{S}_k(V, 0)$ with x as witness point in $\mathcal{A}_k(V, 0)$
11. IF ($(s, -s_{k+1})$ does not cover a stem vector of $\tilde{\mathfrak{S}}_s \cup \tilde{\mathfrak{S}}_a \cup (-\tilde{\mathfrak{S}}_a)$)
12. Solve (3.43) with $\tau = 0$; let (x_0, α_0) be a solution
13. IF ($\alpha_0 < 0$) // $(s, -s_{k+1}) \in \mathcal{S}_{k+1}(V, 0)$
14. C_PD_STREE_REC($(s, -s_{k+1})$, x_0, 0, $\tilde{\mathfrak{S}}_s$, $\tilde{\mathfrak{S}}_a$, $\tilde{\mathfrak{S}}_0$)
15. RETURN
16. ELSE
17. Add two or one stem vectors to $\tilde{\mathfrak{S}}_s$ or $\tilde{\mathfrak{S}}_a$, respectively
18. ENDIF
19. ENDIF // here $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}(V, 0)$, check if it $\in \mathcal{S}_{k+1}([V; \tau^{\mathsf{T}}], 0)$
20. IF ($(s, -s_{k+1})$ does not cover a stem vector of $\tilde{\mathfrak{S}}_s \cup \tilde{\mathfrak{S}}_0$)
21. Solve (5.49); let (x_1, ξ_1, α_1) be a solution // here $\xi_1 \neq 0$
22. IF ($\alpha_1 < 0$) // $(s, -s_{k+1}) \in \mathcal{S}_{a,k+1}(V, -\operatorname{sgn}(\xi_1)\tau)$
23. C_PD_STREE_REC($(s, -s_{k+1})$, $x_1/|\xi_1|$, $-\operatorname{sgn}(\xi_1)$, $\tilde{\mathfrak{S}}_s$, $\tilde{\mathfrak{S}}_a$, $\tilde{\mathfrak{S}}_0$)
24. ELSE
25. Add two stem vectors to $\tilde{\mathfrak{S}}_s$ or $\tilde{\mathfrak{S}}_0$
26. ENDIF
27. ENDIF
28. RETURN
29. ENDIF // here $\boxed{\mathrm{S}} \in \{\pm 1\}$, hence $s \in \mathcal{S}_{a,k}(V, \boxed{\mathrm{S}}\tau)$
30. IF ($(s, -s_{k+1})$ does not cover a stem vector of $\tilde{\mathfrak{S}}_s \cup (\boxed{\mathrm{S}}\tilde{\mathfrak{S}}_a)$)
31. Solve (3.43) with $\tau \curvearrowright \boxed{\mathrm{S}}\tau$; let (x_2, α_2) be a solution
32. IF ($\alpha_2 < 0$)
33. C_PD_STREE_REC($(s, -s_{k+1})$, x_2, $\boxed{\mathrm{S}}$, $\tilde{\mathfrak{S}}_s$, $\tilde{\mathfrak{S}}_a$, $\tilde{\mathfrak{S}}_0$)
34. ELSE
35. Add two or one stem vectors to $\tilde{\mathfrak{S}}_s$ or $\tilde{\mathfrak{S}}_a$, respectively
36. ENDIF
37. ENDIF

5.7 Numerical results

The goal of this section is to assess the efficiency of a selection of algorithms enumerating the chambers, among those introduced in sections 5.4, 5.5 and 5.6, on a selection of hyperplane arrangements. Section 5.7.1 lists and briefly describes the considered arrangement instances. The chosen algorithms are specified in section 5.7.2. Section 5.7.3 details and discusses the results of this evaluation.

5.7.1 Arrangement instances

This section describes the arrangements that form the test bed for the evaluation of the selected algorithms presented in the next section. These arrangements $\mathcal{A}(V, \tau)$ are specified by their matrix $V \in \mathbb{R}^{n\times p}$ and vector $\tau \in \mathbb{R}^p$ (see section 5.3.1). One always has $p > n$ and $r := \mathrm{rank}(V) = n$. The instance features are given in tables 5.2 (theoretical values, for some of them) and 5.3 (numerical values). More is said on these problems in [80].

The given five problems are affine, with $\tau \neq 0$. Some of them were examined in [208]. Their linear versions, obtained by setting $\tau = 0$, were considered in [77]. Random numbers are generated with the Julia function RAND.

- RAND-N-P : $V \in \mathbb{R}^{n\times p}$ and $\tau \in \mathbb{R}^p$ are randomly generated in $[-5, +5]$.
- SRAND-N-P-Q : One has $V_{[1:n],[1:n]} = I_n$ and each of the remaining $p - n$ columns has q nonzero random integer components, randomly positioned. Each element of $\tau_{[n+1:p]}$ has a $1/2$ probability of being a random integer ; it vanishes otherwise ; $\tau_{[1:n]} = 0$. Random integers are taken in $[-10 : +10] \setminus \{0\}$.
- 2D-N-P : The matrix V is such that : $V_{[1:2],[1:n-2]} = 0$ and $V_{[3:n],[n-1:p]} = 0$. Its remaining elements and τ are randomly generated integers in $[-20, +20]$ [208, 77].
- PERM-N : This problem refers to the hyperplane arrangements that are called *permutahedron* in [208] : one has $p = n(n + 1)/2$, $V_{:,[1:n]}$ is the identity matrix and $V_{:,[n+1:p]}$ is a Coxeter matrix [203] (each column is of the form $e_i - e_j$ for some $i < j$ in $[1 : n]$, where e_k is the kth basis vector of $\mathbb{R}^n$). The vector τ is defined by $\tau_i = 1$ for $i \in [1 : n]$ and $\tau_i = 0$ for $i \in [n+1 : p]$. Since $(1, \ldots, 1)$ belongs to all the hyperplanes, the arrangement is centered.
- RATIO-N-P-T : $V_{[1:n],[1:n]}$, $\tau_{[1:n]}$ are randomly generated in $[-50 : +50]$ and $\mathrm{T} \in [0, 1]$. Then, the remaining columns of $[V; \tau^{\mathsf{T}}]$ can either be random with probability $1 - \mathrm{T}$ or randomly generated linear combinations in $[-4 : 4]$ of the previous vectors. One recovers problem RAND-N-P when $\mathrm{T} = 0$.

Some cardinality formulas are gathered in table 5.2. The numerical values of several cardinalities of the considered instances are given in table 5.3.

Problems	Circuits $\lvert\mathcal{C}(V)\rvert$	Stem vectors $\lvert\mathfrak{S}_s(V,\tau)\rvert/2$	$\lvert\mathfrak{S}_a(V,\tau)\rvert$	$\lvert\mathfrak{S}([V;\tau^{\mathsf{T}}],0)\rvert/2$	Chambers $\lvert\mathcal{S}(V,\tau)\rvert$
RAND-N-P	$\binom{p}{r+1}$	0	$\binom{p}{r+1}$	$\binom{p}{r+2}$	$\sum_{i=0}^{r}\binom{p}{i}$
2D-N-P	$\binom{p-n+2}{3}$	0	$\binom{p-n+2}{3}$	$\binom{p-n+2}{4}$	$2^{n-2}\sum_{i=0}^{2}\binom{p-n+2}{i}$
PERM-N	$\sum_{i=3}^{n+1}\frac{i!}{2i}\binom{n+1}{i}$	$\sum_{i=3}^{n+1}\frac{i!}{2i}\binom{n+1}{i}$	0	$\sum_{i=3}^{n+1}\frac{i!}{2i}\binom{n+1}{i}$	$(n+1)!$

TABLE 5.2 – Cardinality formulas for some instances, when $p > n$ and $\mathrm{rank}(V) = n$.

Remarks 5.7.1 (on table 5.3). 1) As expected, the randomly generated arrangements RAND-* are in affine general position (definition 5.3.29). This is revealed in table 5.3 by a number $\lvert\mathfrak{S}_s(V,\tau)\rvert/2+\lvert\mathfrak{S}_a(V,\tau)\rvert$ of circuits of V (5th and 6th columns, see remark 5.3.13(3)) that reaches its maximum (4th column), see remark 5.3.13(6); by a number $\lvert\mathfrak{S}([V;\tau^{\mathsf{T}}],0)\rvert/2$ of circuits of $[V;\tau^{\mathsf{T}}]$ (7th column, see [77, after definition 3.9]) that reaches its maximum (8th column), see remark 5.3.13(6); and by a number $\lvert\mathcal{S}(V,\tau)\rvert$ of sign vectors (9th column) that reaches its upper bound (10th column), see proposition 5.3.31.

2) Half the number of stem vectors of the linear arrangement $\mathcal{A}([V;\tau^{\mathsf{T}}],0)$ (7th column) is also the number $\lvert\mathcal{C}([V;\tau^{\mathsf{T}}])\rvert$ of circuits of $[V;\tau^{\mathsf{T}}]$ (see [77, after definition 3.9]) and we see that this one is unrelated to the number of circuits of V (sum of columns 5 and 6). This confirms the observation made after proposition 5.3.20, according to which neither $\mathcal{C}(V) \subseteq \mathcal{C}([V;\tau^{\mathsf{T}}])$ nor $\mathcal{C}([V;\tau^{\mathsf{T}}]) \subseteq \mathcal{C}(V)$ must hold. □

5.7.2 Assessed algorithms

In the next section, the following algorithms have been evaluated on the problem instances listed in the previous section. These algorithms are identified by the following labels.

- RC : the original RC algorithm [208].
- P : the primal $\mathcal{S}$-tree algorithm 5.4.1.
- PD : the primal-dual $\mathcal{S}$-tree algorithm 5.5.7.
- D : the dual $\mathcal{S}$-tree algorithm 5.5.4.
- RC/C : the compact version of the RC algorithm.
- P/C : the compact primal $\mathcal{S}$-tree algorithm 5.6.4.
- PD/C : the compact primal-dual $\mathcal{S}$-tree algorithm 5.6.7.
- D/C : the compact dual $\mathcal{S}$-tree algorithm.

All the algorithms, but RC and RC/C, benefit from the enhancements A (section 5.4.2), B (section 5.4.2) and C (section 5.4.2). By want of space, the algorithms RC, RC/C and D/C have not been presented in sections 5.4 and 5.6. Briefly, algorithm RC is algorithm 5.4.2 (with its header 5.4.1) without its steps 4-8; algorithm RC/C is algorithm 5.6.5 (with its header 5.6.4) without its steps 5-7; algorithm D/C is obtained from algorithm 5.5.4 using

Problem	n	p	Circuits of V Bound	Stem vectors of $\mathcal{A}(V,\tau)$ $\lvert\mathfrak{S}_s\rvert/2$	$\lvert\mathfrak{S}_a\rvert$	Stem vectors of $\mathcal{A}([V;\tau^{\mathsf{T}}],0)$ $\lvert\mathfrak{S}\rvert/2$	Bound	Chambers $\lvert\mathcal{S}(V,\tau)\rvert$	Bound
RAND-2-8	2	8	56	0	56	70	70	37	37
RAND-4-8	4	8	56	0	56	28	28	163	163
RAND-4-9	4	9	126	0	126	84	84	256	256
RAND-5-10	5	10	210	0	210	120	120	638	638
RAND-4-11	4	11	462	0	462	462	462	562	562
RAND-6-12	6	12	792	0	792	495	495	2510	2510
RAND-5-13	5	13	1716	0	1716	1716	1716	2380	2380
RAND-7-14	7	14	3003	0	3003	2002	2002	9908	9908
RAND-7-15	7	15	6435	0	6435	5005	5005	16384	16384
RAND-8-16	8	16	11440	0	11440	8008	8008	39203	39203
RAND-9-17	9	17	19448	0	19448	12376	12376	89846	89846
SRAND-8-20-2	8	20	167960	56	321	987	184756	36225	263950
SRAND-8-20-4	8	20	167960	1185	70650	94534	184756	213467	263950
SRAND-8-20-6	8	20	167960	20413	123909	105345	184756	245396	263950
2D-4-20	4	20	15504	1	815	3046	38760	684	6196
2D-5-20	5	20	38760	0	680	2380	77520	1232	21700
2D-6-20	6	20	77520	1	559	1808	125970	2176	60460
2D-7-20	7	20	125970	0	443	1365	167960	3840	137980
2D-8-20	8	20	167960	0	364	1001	184756	6784	263950
PERM-5	5	15	5005	197	0	197	6435	720	4944
PERM-6	6	21	116280	1172	0	1172	203490	5040	82160
PERM-7	7	28	3108105	8018	0	8018	6906900	40320	1683218
PERM-8	8	36	94143280	62814	0	62814	254186856	362880	40999516
RATIO-3-20-7	3	20	4845	19	4614	14043	15504	1119	1351
RATIO-3-20-9	3	20	4845	118	4550	12993	15504	1176	1351
RATIO-4-20-7	4	20	15504	102	15271	36781	38760	6015	6196
RATIO-4-20-9	4	20	15504	2327	11908	19882	38760	4600	6196
RATIO-5-20-7	5	20	38760	97	33945	61452	77520	15136	21700
RATIO-5-20-9	5	20	38760	23514	10954	23514	77520	11325	21700
RATIO-6-20-7	6	20	77250	238	76595	120663	125970	59519	60640
RATIO-6-20-9	6	20	77250	345	71861	106115	125970	53795	60460
RATIO-7-20-7	7	20	125970	125	123792	159956	167960	135064	137980
RATIO-7-20-9	7	20	125970	154	123731	159636	167960	135039	137980

TABLE 5.3 – Description of the 33 considered arrangements. The first column gives the problem names. The next two columns specify the dimensions of $V \in \mathbb{R}^{n\times p}$. The 4th column gives the upper bound on the number of circuits of V, recalled in remark 5.3.13(6); by remark 5.3.13(3), it is also an upper bound on $|\mathfrak{S}_s|/2 + |\mathfrak{S}_a|$, where $|\mathfrak{S}_s|$ (resp. $|\mathfrak{S}_a|$) is the number of symmetric (resp. asymmetric) stem vectors (definition 5.3.12) of the arrangement $\mathcal{A}(V,\tau)$; $|\mathfrak{S}_s|/2$ and $|\mathfrak{S}_a|$ are given in columns 5 and 6. Columns 7 and 8 give half the number of stem vectors of the arrangement $\mathcal{A}([V;\tau^{\mathsf{T}}],0)$ and its Schläfli upper bound, derived from (5.28). The last two columns give the number $|\mathcal{S}(V,\tau)|$ of chambers of the arrangement $\mathcal{A}(V,\tau)$ and its upper bound given by (5.30).

the compaction principles described in section 5.6.

5.7.3 Numerical results

To evaluate the algorithms listed in the previous section, we have implemented them in a Julia code named isf.jl, which extends the Matlab code isf.m used in chapter 3, from linear to general affine arrangements. The implementation has been done in Julia (version "1.8.5") on a MacBookPro18, 2/10cores (parallelism is not used) with the system macOS Monterey, version 12.6.1.

All the solvers, but D and D/C, need to solve linear optimization problems. The linear optimization solver used in the Julia code is Gurobi. This one appears to be more efficient than the Matlab solver Linprog used in [75, 76]. Since the improvement is obtained by a reduction of the number of LOPs, which are solved much faster in the Julia version, we observe a less important improvement (wrt the RC algorithm) in computing time in the present study (Julia code) than reported in [77].

The main computational burden of the "pure primal" variants P and P/C is the solution of the LOPs while, for the "pure dual" variants D and D/C, it is the computation of the stem vectors and their use in the covering tests. These are not comparable. Therefore, counting the number of LOPs or the number of covering tests is not a relevant criterion for comparing the solvers. For this reason, we rely on computing time. Since the RC algorithm was shown in [208] to have better performance in time than earlier methods, a comparison is often made with the RC algorithm. Since this algorithm is implemented in Python, we avoid biases due to the programming language by making the comparison with our Julia version of the RC algorithm, which can be easily simulated from algorithm 5.4.1, as mentioned above.

For ease of reading, the comparison of the solvers' efficiency is carried out by using *performance profiles* [69] (tables with precise numbers are also given in section 5.9) : these are curves in a graph with the *relative efficiency* on the x-axis (sometimes in logarithmic scale) and a *percentage of problems* on the y-axis. There is one graph per performance, which is the computing time in our case, and there is one curve per solver in that graph : a point (e, f) of the curve of a solver tells us that the efficiency of this solver is never worse than e times that of the best solver (this one depends on the considered problem) on a fraction f of the problems. As a result, the solver with the highest curve, if any, can be legitimately considered as the most effective one, while the ranking of the other solvers by the position of their curve in the graph should be taken with caution [109]. The performance profiles only depend on the *relative* performance of the solvers, that is, for a particular problem, their performance divided by the one of the best solver for that problem. Therefore taking the *computing time* or the *computing time per chamber* as performance yield the same performance profiles.

Standard solvers

Let us first compare the standard solvers RC, P, PD and D with each other, on the selected arrangements described in table 5.3. The computing is the compared quantity. This computing times are reported in table 5.4 and the performance profiles are given in figure 5.6.

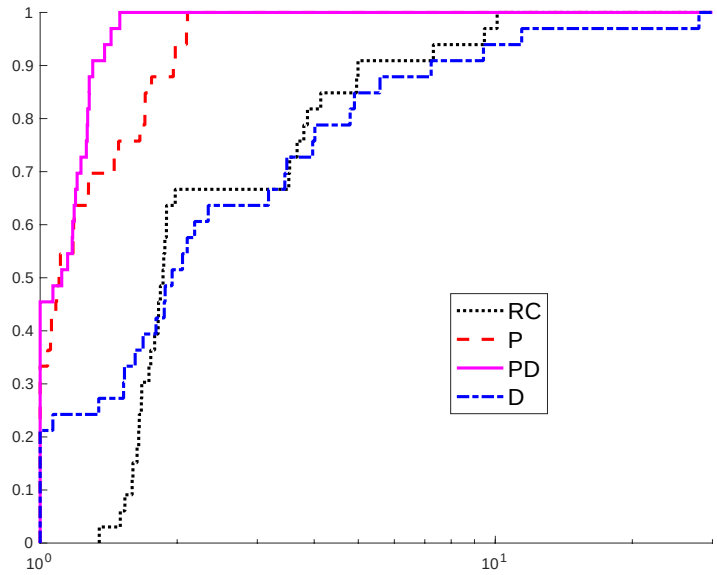


FIGURE 5.6 – Performance profiles of the RC, P, PD and D algorithms, for the computing time.

One observes that the PD algorithm is generally the most efficient one when the *computing time* is taken as a reference. The speed up with respect to the RC algorithm can reach 10 (this can be observed on table 5.4 : ratio 10.14 of the PD algorithm on instance PERM-8) or on figure 5.6 (by the abscissa of the rightmost change in curve of the RC algorithm, whose relative performance is there given relatively to the PD algorithm).

Compact solvers

Computation time To show the interest of the compact versions of the algorithms, introduced in section 5.6, we compare each solver RC, P, PD and D to its compact version RC/C, P/C, PD/C and D/C. The computing times are given in table 5.5 and the performance profiles are given in figure 5.7.

We observe indeed on figure 5.7 that the compact versions improve their standard version on the computing time, particularly for the PD/C, for which the mean (resp. median) improvement is 1.35 (resp. 1.35). The improvement bound 2 is obtained by D/C on the instances SRAND-8-20-4 and PERM-7. This improvement can also be observed on figure 5.7, with more ambiguity, since it is not indicated which algorithm is the best for each problem

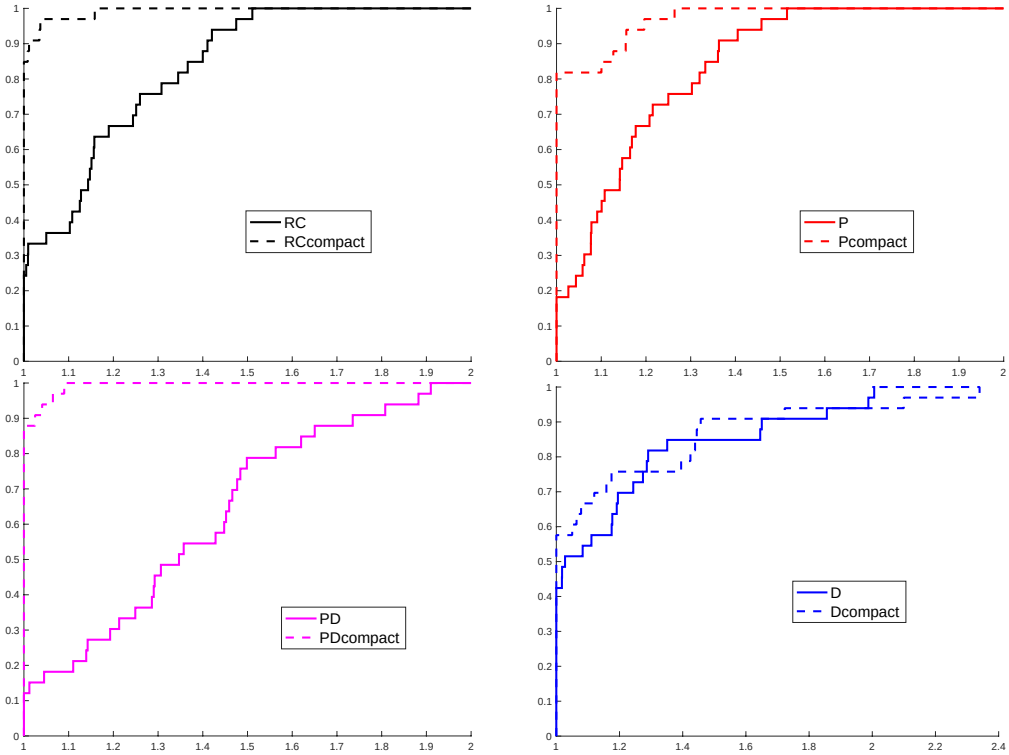


Figure 5.7 – Performance profiles of the RC vs RC/C, P vs P/C, PD vs PD/C and D vs D/C algorithms, for the computing time. The dashed lines refer to the compact versions of the algorithms.

instance (for example the x-axis larger than 2 for the performance profiles D vs. D/C is not due to a performance of D/C that is 2.34 times better than D on some problem, but the opposite : it is D that is 2.34 times faster than D/C on some problem). Nevertheless, it is indeed algorithm A/C that generally outperforms algorithm A when A $=$ RC, P or PD (their curve is higher).

The performance profiles of figure 5.8 compares the most effective solver, namely PD/C, to the RC solver. The former shows a speedup that can reach 19.3.

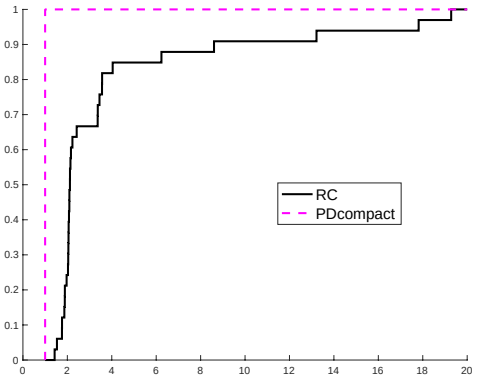


FIGURE 5.8 – Performance profiles of the RC vs PD/C solvers, for the computing time.

5.8 Conclusion

This chapter deals with the enumeration of the chambers of a hyperplane arrangement. It brings improvements to a recursive algorithm proposed by Rada and Černý, and proposes a family of new algorithms having, to various extends, dual aspects based on the Motzkin's alternative, matroid circuits and the introduced notion of *stem vector*. Most algorithms are grounded on a tree of sign vectors that are in one-to-one correspondence with the chambers of the arrangement. Compact versions of the algorithms are also presented, which aim at reducing the size of the sign vector tree, in order to avoid duplicating costly identical subproblems like linear optimization problems or covering tests. The most efficient method of this algorithm anthology is the one that includes primal and dual ingredients, and uses the compact form of the tree, which has been named PD/C in the paper. The speedup it provides, with respect to Rada-Černý's algorithm, much depends on the features of the considered arrangement, in particular its dimensions, and ranges between 1.4 and 19.3, with a mean value of 3.9.

These algorithms are grounded on a theory that is presented before their introduction. This one includes the structure of the sign vector sets and the stem vector sets, in particular conditions for their symmetry, their connectivity, their full cardinality and much more.

Numerous aspects of the presented algorithms can be further improved or developed, covering both conceptual and implementation aspects. Let us mention a few topics. (i) The

linear optimization problems could be solved approximately, hence saving computation time when the tested sign vector is feasible. (ii) The way stem vectors are computed, stored and used could be improved, with specific structures designed for that purpose. (iii) In case the arrangements present combinatorial symmetries, the approaches presented in [212, 35] should increase significantly the algorithm performances. (iv) The proposed approaches could be extended to compute the chambers of the hyperplane arrangement and subarrangements, those recursively included in the hyperplane intersections of any smaller dimension.

5.9 Appendix : tables with numerical results

This appendix gives the tables with the detailed numerical results, comparing the solvers selected in section 5.7.2, on which the performance profiles of figures 5.6, 5.7 and 5.8 are based. Comments on these results can be found in section 5.7.3. Table 5.4 deals with the standard algorithms and table 5.5 is related to the compact versions of these algorithms.

Acknowledgments

We thank Miroslav Rada and Michal Černý for providing us with their code and problem instances, presented in [208].

Funding

The first author's research was partially supported by NSERC grant OGP0005491. The third author was partially supported by a Mitacs-Inria grant.

Problems	RC time	P time	P ratio	PD time	PD ratio	D time	D ratio
RAND-2-8	0.09	0.03	3.76	0.03	3.51	**0.02**	**4.13**
RAND-4-8	0.12	0.07	1.71	0.08	1.47	**0.06**	**1.86**
RAND-4-9	0.20	0.12	1.67	0.13	1.56	**0.10**	**1.99**
RAND-5-10	0.48	0.27	1.79	0.31	1.58	**0.26**	**1.89**
RAND-4-11	0.56	0.35	1.58	0.38	1.47	**0.30**	**1.88**
RAND-6-12	1.89	1.18	1.59	1.34	1.41	**1.09**	**1.73**
RAND-5-13	2.32	1.37	1.69	1.45	1.60	**1.30**	**1.79**
RAND-7-14	6.96	**4.37**	**1.59**	5.14	1.35	4.66	1.50
RAND-7-15	12.40	**7.51**	**1.65**	8.92	1.39	10.10	1.22
RAND-8-16	29.20	**17.50**	**1.67**	21.10	1.38	28.30	1.03
RAND-9-17	63.00	**39.40**	**1.60**	50.50	1.25	70.80	0.89
SRAND-8-20-2	33.30	9.70	3.43	**6.67**	**5.00**	23.00	1.45
SRAND-8-20-4	199.00	**105.00**	**1.90**	137.00	1.45	989.00	0.20
SRAND-8-20-6	238.00	**131.00**	**1.81**	196.00	1.21	947.00	0.25
2D-4-20	2.12	0.89	2.38	**0.60**	**3.53**	1.01	2.11
2D-5-20	3.50	1.58	2.22	**0.95**	**3.67**	1.86	1.88
2D-6-20	5.84	2.82	2.07	**1.66**	**3.53**	3.11	1.88
2D-7-20	9.86	4.48	2.20	**2.55**	**3.86**	5.24	1.88
2D-8-20	16.40	7.35	2.23	**4.32**	**3.80**	8.13	2.02
PERM-5	1.17	0.46	2.53	**0.24**	**4.96**	0.82	1.42
PERM-6	10.60	3.04	3.50	**1.45**	**7.33**	7.11	1.50
PERM-7	106.00	23.60	4.49	**11.20**	**9.50**	128.00	0.83
PERM-8	1070.00	210.00	5.10	**106.00**	**10.14**	2970.00	0.36
RATIO-3-20-0.7	2.53	1.54	1.65	**1.39**	**1.82**	2.12	1.19
RATIO-3-20-0.9	2.59	1.80	1.44	**1.41**	**1.84**	2.16	1.20
RATIO-4-20-0.7	11.10	6.17	1.80	**5.94**	**1.87**	12.50	0.89
RATIO-4-20-0.9	7.07	5.53	1.28	**4.33**	**1.63**	9.46	0.75
RATIO-5-20-0.7	21.00	**12.00**	**1.75**	12.80	1.64	38.10	0.55
RATIO-5-20-0.9	16.00	11.30	1.42	**9.57**	**1.67**	22.40	0.71
RATIO-6-20-0.7	75.90	**46.10**	**1.65**	58.50	1.30	183.00	0.42
RATIO-6-20-0.9	65.40	**43.60**	**1.50**	50.10	1.30	175.00	0.37
RATIO-7-20-0.7	147.00	**109.00**	**1.36**	151.00	0.98	523.00	0.28
RATIO-7-20-0.9	148.00	**96.40**	**1.54**	138.00	1.07	538.00	0.28
Mean			2.11		**2.76**		1.28
Median			**1.71**		1.63		1.23

TABLE 5.4 – Computing times (in seconds) for the *standard* algorithms listed in section 5.7.2. For each algorithm A := P, PD or D, the second column gives the ratios time(RC)/time(A)

Problems	RC/C time	RC/C ratio	RC/C ratio	P/C time	P/C ratio	P/C ratio	PD/C time	PD/C ratio	PD/C ratio	D/C time	D/C ratio	D/C ratio
rand-2-8	0.08	1.25	1.25	0.03	0.89	3.33	0.03	0.96	3.37	**0.03**	0.85	**3.52**
rand-4-8	0.08	1.37	1.37	0.05	1.33	2.28	0.05	1.45	2.14	**0.05**	1.24	**2.32**
rand-4-9	0.16	1.24	1.24	0.10	1.21	2.02	0.11	1.21	1.89	**0.09**	1.11	**2.20**
rand-5-10	0.35	1.40	1.40	**0.22**	1.25	**2.24**	0.24	1.29	2.05	0.22	1.18	2.23
rand-4-11	0.49	1.15	1.15	0.30	1.16	1.84	0.30	1.29	1.89	**0.29**	1.03	**1.93**
rand-6-12	1.34	1.41	1.41	**0.87**	1.36	**2.18**	0.90	1.48	2.09	0.92	1.19	2.07
rand-5-13	1.95	1.19	1.19	**1.20**	1.14	**1.93**	1.11	1.31	2.09	1.20	1.08	1.93
rand-7-14	4.90	1.42	1.42	**3.11**	1.41	**2.24**	3.43	1.50	2.03	3.90	1.19	1.78
rand-7-15	9.22	1.34	1.34	**5.69**	1.32	**2.18**	6.04	1.48	2.05	8.58	1.18	1.45
rand-8-16	19.80	1.47	1.47	**12.00**	1.46	**2.43**	13.50	1.56	2.16	17.20	1.65	1.70
rand-9-17	41.70	1.51	1.51	**26.00**	1.52	**2.42**	30.60	1.65	2.06	42.90	1.65	1.47
srand-8-20-2	30.20	1.10	1.10	9.45	1.03	3.52	**5.34**	1.25	**6.24**	22.60	1.02	1.47
srand-8-20-4	158.00	1.26	1.26	**80.60**	1.30	**2.47**	93.90	1.46	2.12	493.00	2.01	0.40
srand-8-20-6	182.00	1.31	1.31	**96.10**	1.36	**2.48**	121.00	1.62	1.97	701.00	1.35	0.34
2d-4-20	2.10	1.01	1.01	1.03	0.87	2.06	**0.62**	0.98	**3.45**	1.74	0.58	1.22
2d-5-20	3.54	0.99	0.99	1.89	0.84	1.85	**1.04**	0.92	**3.37**	2.71	0.69	1.29
2d-6-20	5.89	0.99	0.99	3.26	0.87	1.79	**1.64**	1.01	**3.56**	4.43	0.70	1.32
2d-7-20	9.81	1.01	1.01	4.93	0.91	2.00	**2.44**	1.05	**4.04**	7.31	0.72	1.35
2d-8-20	16.40	1.00	1.00	9.29	0.79	1.77	**4.60**	0.94	**3.57**	11.70	0.69	1.40
perm-5	1.04	1.12	1.12	0.42	1.11	2.80	**0.14**	1.74	**8.60**	0.64	1.29	1.83
perm-6	9.27	1.14	1.14	2.82	1.08	3.76	**0.80**	1.81	**13.22**	5.58	1.27	1.90
perm-7	94.00	1.13	1.13	22.30	1.06	4.75	**5.95**	1.88	**17.82**	64.40	1.99	1.65
perm-8	1070.00	1.00	1.00	195.00	1.08	5.49	**55.50**	1.91	**19.28**	1600.00	1.86	0.67
ratio-3-20-0.7	2.53	1.00	1.00	1.45	1.06	1.74	**1.22**	1.14	**2.07**	4.96	0.43	0.51
ratio-3-20-0.9	2.68	0.97	0.97	1.65	1.09	1.57	**1.27**	1.11	**2.04**	3.12	0.69	0.83
ratio-4-20-0.7	11.00	1.01	1.01	5.73	1.08	1.94	**4.98**	1.19	**2.23**	13.30	0.94	0.83
ratio-4-20-0.9	8.19	0.86	0.86	5.30	1.04	1.33	**3.79**	1.14	**1.87**	7.33	1.29	0.96
ratio-5-20-0.7	20.00	1.05	1.05	10.90	1.10	1.93	**9.92**	1.29	**2.12**	80.00	0.48	0.26
ratio-5-20-0.9	13.90	1.15	1.15	9.67	1.17	1.65	**6.61**	1.45	**2.42**	22.00	1.02	0.73
ratio-6-20-0.7	68.50	1.11	1.11	**40.20**	1.15	**1.89**	43.10	1.36	1.76	212.00	0.86	0.36
ratio-6-20-0.9	67.80	0.96	0.96	38.20	1.14	1.71	**37.20**	1.35	**1.76**	196.00	0.89	0.33
ratio-7-20-0.7	127.00	1.16	1.16	**89.70**	1.22	**1.64**	103.00	1.47	1.43	564.00	0.93	0.26
ratio-7-20-0.9	128.00	1.16	1.16	**81.90**	1.18	**1.81**	96.60	1.43	1.53	565.00	0.95	0.26
Mean		1.16	1.16		1.14	2.33		1.35	**3.95**		1.09	1.30
Median		1.14	1.14		1.14	2.02		1.35	**2.12**		1.03	1.35

Table 5.5 – Computing times (in seconds) for the *compact* algorithms listed in section 5.7.2. For each algorithm A = RC, P, PD, or D, the first column gives the computing time of A/C in seconds, the second column gives the ratios time(A)/time(A/C) (upper bounded by 2, approximately) and the third column gives the ratios time(RC)/time(A/C).

Chapitre 6

Globalisation de PNM par moindres-carrés et Levenberg-Marquardt

Après cette incursion longue et détaillée dans le monde des hyperplans, ce chapitre détaille des travaux en cours concernant la motivation initiale de cette thèse, la globalisation de l'algorithme Newton-min polyédrique (NMP) de l'article [72], à paraître bientôt dans *Mathematical Programming*. Dans une première partie, la section 6.1 présente une facon de modifier l'algorithme NMP, tandis que la section 6.2 discute de propriétés de convergence de l'algorithme modifié correspondant.

Rappelons que nous considérons le problème de complémentarité général suivant

$$0 \leqslant F(x) \perp G(x) \geqslant 0, \tag{6.1}$$

où F et G sont deux fonctions C^1 de $\mathbb{R}^n$ dans $\mathbb{R}^n$, et éventuellement sa version affine

$$0 \leqslant Ax + a \perp Bx + b \geqslant 0.$$

En utilisant la reformulation par la C-fonction minimum, cela mène à un système non différentiable et la minimisation de sa fonction de mérite

$$H(x) := \min(F(x), G(x)) = 0 \qquad \text{and} \qquad \min \theta(x) := \frac{1}{2}||H(x)||^2. \tag{6.2}$$

Les ensembles d'indices suivants jouent un rôle très important :

$$\begin{array}{rcl} \mathcal{E}(x) & := & \{i \in [1:n] : F_i(x) = G_i(x)\}, \\ \mathcal{F}(x) & := & \{i \in [1:n] : F_i(x) < G_i(x)\}, \\ \mathcal{G}(x) & := & \{i \in [1:n] : F_i(x) > G_i(x)\}. \end{array} \tag{6.3}$$

Le système et la minimisation dans (6.2) sont $\mathcal{C}^1$ par morceaux. Comme mentionné dans la section 2.3.3 à propos de l'algorithme de Newton-min, à un itéré la question de "quel

morceau choisir ?" peut se poser. Le choix d'un morceau et l'obtention d'une direction de descente n'est pas nécessairement facile et, comme on va le voir, même détecter la stationnarité d'un point est en général co-NP-complet. Cela est montré par une reformulation géométrique en partie reliée au chapitre 3.

Néanmoins, en supposant que l'on peut traiter de tels problèmes (par injectivité d'une sous-matrice impliquée par exemple), un algorithme basé sur une courbe de Levenberg-Marquardt tangente à un élément de $\partial\theta(x)$ à chaque itération est proposée, qui bénéficie des propriétés usuelles de Levenberg-Marquardt.

6.1 Modification de Newton-min polyédrique

6.1.1 Présentation de la méthode

Nous rappelons dans un premier temps l'algorithme de Newton-min. L'équation non lisse $H(x) = 0$ est résolue par des équations de type Newton où les indices de $\mathcal{E}(x)$, qui rendent le système non lisse, sont mis arbitrairement dans $\mathcal{F}(x)$ ou $\mathcal{G}(x)$.

Algorithme 6.1.1 (Newton-min)**.** Soit $x^0 \in \mathbb{R}^n$ un point de départ.

1. *Test d'arrêt.* Si $H(x^k) = 0$, arrêt.
2. *Décomposition des indices.* On définit les ensembles d'indices suivants :

$$\mathcal{E}(x^k), \quad \mathcal{F}(x^k), \quad \mathcal{G}(x^k).$$

   Soit $\tilde{F}(x^k)$, $\tilde{G}(x^k)$ une partition de $[1:n]$ telle que $\tilde{F}(x^k) \supseteq \mathcal{F}(x^k)$ et $\tilde{G}(x^k) \supseteq \mathcal{G}(x^k)$. Résoudre le système linéaire en d :

$$\begin{cases} F(x^k)_{\tilde{F}(x^k)} + F'(x^k)_{\tilde{F}(x^k)} d = 0, \\ G(x^k)_{\tilde{G}(x^k)} + G'(x^k)_{\tilde{G}(x^k)} d = 0. \end{cases} \tag{6.4}$$

3. *Mise à jour.* Poser $x^{k+1} = x^k + d^k$ où d^k est une solution de (6.4),

L'algorithme suppose que le système (6.4) a une solution. C'est garanti autour d'un point s'il vérifie une certaine condition de régularité.

Comme présenté dans la section 2.3.4, une variante de cet algorithme résout un système qui utilise des inégalités et pas uniquement des égalités, pour s'assurer d'avoir une direction de descente pour la fonction de mérite θ [72]. On définit la partition suivante de $\mathcal{E}(x)$:

$$\begin{aligned} \mathcal{E}^{0+}(x) &:= \{i \in \mathcal{E}(x) : F_i(x) = G_i(x) \geqslant 0\}, \\ \mathcal{E}^{-}(x) &:= \{i \in \mathcal{E}(x) : F_i(x) = G_i(x) < 0\}. \end{aligned} \tag{6.5}$$

Ensuite, considérons une partition de $\mathcal{E}^{0+}(x)$ en $\mathcal{E}^{0+}_{\mathcal{F}}(x) \cup \mathcal{E}^{0+}_{\mathcal{G}}(x)$. L'idée proposée est de garder une égalité pour les indices de $\mathcal{E}^{0+}(x)$ et d'introduire deux inégalités pour ceux de $\mathcal{E}^{-}(x)$. De fait, on résout

$$\begin{cases} F_i(x) + F_i'(x)d = 0 & \text{si } i \in \mathcal{F}(x) \cup \mathcal{E}^{0+}_{\mathcal{F}}(x), \\ G_i(x) + G_i'(x)d = 0 & \text{si } i \in \mathcal{G}(x) \cup \mathcal{E}^{0+}_{\mathcal{G}}(x), \\ F_i(x) + F_i'(x)d \geqslant 0 & \text{si } i \in \mathcal{E}^{-}(x), \\ G_i(x) + G_i'(x)d \geqslant 0 & \text{si } i \in \mathcal{E}^{-}(x). \end{cases} \tag{6.6}$$

Observons que ce système a $n - |\mathcal{E}^-(x)|$ égalités et $2|\mathcal{E}^-(x)|$ inégalités. On montre que, pour une solution de ce système d, on a $\theta'(x;d) \leqslant -2\theta(x)$; là où pour la méthode de Newton avec H lisse, on a $\nabla\theta(x)^\mathsf{T} d = -2\theta(x)$. En effet, en utilisant $\theta'(x;d) = H(x)^\mathsf{T} H'(x;d)$, on a [1]

$$\begin{aligned} \theta'(x;d) = & \sum_{i\in\mathcal{F}(x)} F_i(x)F_i'(x)d + \sum_{i\in\mathcal{G}(x)} G_i(x)G_i'(x)d + \sum_{i\in\mathcal{E}(x)} H_i(x)\min(F_i'(x)d, G_i'(x)d) \\ = & -||F_{\mathcal{F}(x)}(x)||^2 - ||G_{\mathcal{G}(x)}(x)||^2 - ||H_{\mathcal{E}(x)}(x)||^2 \\ & + H_{\mathcal{E}(x)}(x)^\mathsf{T} \min\left(F_{\mathcal{E}(x)}(x) + F'_{\mathcal{E}(x)}(x)d, G_{\mathcal{E}(x)}(x) + G'_{\mathcal{E}(x)}(x)d\right) \\ = & -2\theta(x) + H_{\mathcal{E}(x)}(x)^\mathsf{T} \min\left(F_{\mathcal{E}(x)}(x) + F'_{\mathcal{E}(x)}(x)d, G_{\mathcal{E}(x)}(x) + G'_{\mathcal{E}(x)}(x)d\right). \end{aligned}$$

Étudions la dernière ligne terme par terme. Pour $i \in \mathcal{E}^{0+}(x)$, soit $F_{\mathcal{E}(x)}(x) + F'_{\mathcal{E}(x)}(x)d = 0$ ou $G_{\mathcal{E}(x)}(x) + G'_{\mathcal{E}(x)}(x)d = 0$, ce qui signifie the leur minimum est $\leqslant 0$. Ensuite, multiplier par $H_i(x) \geqslant 0$ fait que le produit est négatif. Pour $i \in \mathcal{E}^-(x)$, les deux arguments du minimum sont $\geqslant 0$, donc leur minimum aussi, mais est multiplié par un terme strictement négatif. Au final, toutes les quantités sont négatives, donc $\theta'(x;d) \leqslant -2\theta(x)$.

Observons que $\theta'(x;d)$ est convexe (en d) quand $\mathcal{E}^{0+}(x) = \varnothing$ et concave quand $\mathcal{E}^-(x) = \varnothing$ (en d). En effet, en utilisant que le maximum / minimum de fonctions linéaires est convexe / concave : [2]

$$\begin{aligned} &\text{pour } i \in \mathcal{E}^-(x) \quad H_i(x)\min(F_i'(x)d, G_i'(x)d) = \max(H_i(x)F_i'(x)d, H_i(x)G_i'(x)d), \\ &\text{pour } i \in \mathcal{E}^{0+}(x) \quad H_i(x)\min(F_i'(x)d, G_i'(x)d) = \min(H_i(x)F_i'(x)d, H_i(x)G_i'(x)d). \end{aligned}$$

Dans [72], d'autres variantes du système (6.6) sont aussi considérées, en particulier en étendant l'ensemble $\mathcal{E}^-(x)$ à $\mathcal{E}^-_\tau(x)$ pour un $\tau > 0$, les indices tels que

$$\mathcal{E}^-_\tau(x) := \{i \in [1:n] : F_i(x) < 0,\ G_i(x) < 0,\ |F_i(x) - G_i(x)| < \tau\}.$$

Cette tolérance autour des "plis négatifs" ($\mathcal{E}^-(x)$, dans la limite $\tau \to 0$) est aussi hautement pertinent d'un point de vue numérique – cela pourrait concerner, hors de l'algorithme PNM, les "plis positifs" aussi ; on y reviendra plus tard. Voici une version simple de l'algorithme polyédrique.

1. Pour cela, on utilise que $i \in \mathcal{E}(x)$, $F_i(x) = H_i(x) = G_i(x)$ et $H_i(x)\min(F_i'(x)d, G_i'(x)d) = -H_i(x)^2 + H_i(x)\min(F_i(x) + F_i'(x)d, G_i(x) + G_i'(x)d)$.

2. Techniquement, il peut y avoir des indices pour lesquels $F_i(x) = 0 = G_i(x)$ (qui ont arbitrairement été mis avec ceux tels que $F_i(x) = G_i(x) > 0$).

Algorithme 6.1.2 (Algorithme PNM [72])**.** Soit $\tau \in (0, \infty]$ la constante de tolérance des plis, $\omega \in (0, 1/2)$ et $\beta \in (0, 1)$ les deux constantes utilisées dans la recherche linéaire de l'étape 4 ci-dessous. Le prochain itéré x_+ est calculé à partir de l'itéré x comme suit.

1. *Critère d'arrêt.* Si $\theta(x) = 0$, arrêt (x est une solution).
2. *Ensembles d'indices.* Choisir une partition $(\mathcal{E}^{0+}_{\mathcal{F}}(x), \mathcal{E}^{0+}_{\mathcal{G}}(x))$ de $\mathcal{E}^{0+}(x)$ et calculer les ensembles d'indices $\mathcal{E}^{-}_{\tau}(x)$, $\mathcal{F}(x) \setminus \mathcal{E}^{-}_{\tau}(x)$ et $\mathcal{G}(x) \setminus \mathcal{E}^{-}_{\tau}(x)$.
3. *Direction.* Calculer une direction d comme la solution de

$$\min\{||d|| : d \text{ satisfait (6.6)}\} \tag{6.7}$$

   en utilisant les ensembles d'indices du point 2.
4. *Pas.* Fixer $\alpha := \beta^i$ où i est le plus petit entier positif tel que

$$\theta(x + \alpha d) \leqslant (1 - 2\omega\alpha)\theta(x). \tag{6.8}$$

5. *Nouvel itéré.* Poser $x_+ = x + \alpha d$.

6.1.2 Variante moindres-carrés et Levenberg-Marquardt

Le souci principal du système (6.6) est que le polyèdre défini peut être vide. Dans [72], cette vacuité est évitée par des hypothèses de régularité techniques. Récrivons légèrement le système (6.6), avec des variables $\gamma_i \in \{0, 1\}$ pour $i \in \mathcal{E}^{0+}(x)$ et $\overline{\gamma}_i := 1 - \gamma_i$, pour avoir $\gamma_i = +1 \Leftrightarrow i \in \mathcal{E}^{0+}_{\mathcal{F}}(x)$ et $\gamma_i = 0 \Leftrightarrow i \in \mathcal{E}^{0+}_{\mathcal{G}}(x)$:

$$\begin{cases} F_i(x) + F_i'(x)d = 0 & \text{si } i \in \mathcal{F}(x), \\ G_i(x) + G_i'(x)d = 0 & \text{si } i \in \mathcal{G}(x), \\ \gamma_i(F_i(x) + F_i'(x)d) + \overline{\gamma}_i(G_i(x) + G_i'(x)d) = 0 & \text{si } i \in \mathcal{E}^{0+}(x), \\ F_i(x) + F_i'(x)d \geqslant 0 & \text{si } i \in \mathcal{E}^{-}(x), \\ G_i(x) + G_i'(x)d \geqslant 0 & \text{si } i \in \mathcal{E}^{-}(x). \end{cases}$$

De plus, observons que les indices de $\mathcal{E}^{-}(x)$ sont dédoublés : dans une version de moindres-carrés, ces indices peuvent donc avoir un poids supplémentaire comparés aux autres. Nous avons donc essayé d'introduire une pondération convexe des linéarisations de F et G pour ces indices, afin de donner un poids total de 1 à chaque indice. Cela aura un intérêt particulier plus tard. Soit

$$\Gamma = \mathrm{Diag}(\gamma) \in \mathbb{R}^{\mathcal{E}(x)\times\mathcal{E}(x)}, \quad \overline{\Gamma} = I - \mathrm{Diag}(\Gamma) \in \mathbb{R}^{\mathcal{E}(x)\times\mathcal{E}(x)}.$$ [3]

3. Ci-dessous, on utilise $\Gamma_{\mathcal{E}^{0+}(x)} := \Gamma_{\mathcal{E}^{0+}(x),\mathcal{E}^{0+}(x)}$ et $\Gamma_{\mathcal{E}^{-}(x)} := \Gamma_{\mathcal{E}^{-}(x),\mathcal{E}^{-}(x)}$, qui est un léger abus de notation pour une matrice diagonale.

Ensuite, la version moindres-carrés avec poids de (6.6) devient

$$\min_{d\in\mathbb{R}^n}\frac{1}{2}||w(x,d)||_2^2, \quad \text{où } w(x,d) := \begin{bmatrix} F_{\mathcal{F}(x)}(x) + F'_{\mathcal{F}(x)}(x)d \\ G_{\mathcal{G}(x)}(x) + G'_{\mathcal{G}(x)}(x)d \\ (\Gamma_{\mathcal{E}^{0+}(x)})^{1/2}[F_{\mathcal{E}^{0+}(x)}(x) + F'_{\mathcal{E}^{0+}(x)}(x)d] \\ (\overline{\Gamma}_{\mathcal{E}^{0+}(x)})^{1/2}[G_{\mathcal{E}^{0+}(x)}(x) + G'_{\mathcal{E}^{0+}(x)}(x)d] \\ -(\Gamma_{\mathcal{E}^{-}(x)})^{1/2}[F_{\mathcal{E}^{-}(x)}(x) + F'_{\mathcal{E}^{-}(x)}(x)d]^- \\ -(\overline{\Gamma}_{\mathcal{E}^{-}(x)})^{1/2}[G_{\mathcal{E}^{-}(x)}(x) + G'_{\mathcal{E}^{-}(x)}(x)d]^- \end{bmatrix}, \tag{6.9}$$

qui mérite quelques commentaires. D'abord, $||w||^2$ est lisse (mais, en général, pas w) puisque $((\cdot)^-)^2$ l'est. Les poids Γ et $\overline{\Gamma}$ s'appliquent à $||w||^2$, ce qui explique pourquoi, dans w, il y a des racines carrées. Pour les indices dans $\mathcal{E}^-(x)$, le signe moins et $[\cdot]^-$ proviennent des inégalités du système précédent, où seulement les valeurs négatives sont à pénaliser.

Pour $\lambda \in \mathbb{R}_+$ et une matrice S symétrique définie positive (souvent, $S = I$), une variante Levenberg-Marquardt est :

$$\min_{d\in\mathbb{R}^n}\left(\varphi_x(d) := \frac{1}{2}\left(||w(x,d)||_2^2 + \lambda\, d^{\mathsf{T}}Sd\right)\right). \tag{6.10}$$

Pour un itéré donné x, le paramètre λ définit une courbe de solutions $d(\lambda)$ (S est fixée au sein d'une itération) et est modifié pour obtenir des propriétés de descente. Un aspect important, qui est ce que l'on discute à partir de maintenant, est la *tangence* entre θ et φ_x, i.e., l'adéquation entre le modèle φ_x et la vraie fonction θ au premier ordre. Cette tangence est modélisée par le gradient du modèle φ_x en $d = 0$.[4]

Cette proximité est lourdement reliée aux valeurs de γ, qui correspondent à quel morceau est considéré. Observons que la répartition des indices dans l'algorithme 6.1.1 ou la partition de $\mathcal{E}^{0+}(x)$ en $\mathcal{E}^{0+}_{\mathcal{F}}(x)$ et $\mathcal{E}^{0+}_{\mathcal{G}}(x)$ dans (6.6) dont des étapes précurseures des poids convexes γ.

Le gradient de φ_x en $d = 0$ est écrit $g := g(\gamma) = g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^{-}(x)})$ à partir de maintenant (pour éviter des notations lourdes, on omet parfois la dépendance de g en γ). Il joue un rôle central et son expression est donnée ensuite. Dans celle-ci, x et fixé et, pour simplifier la notation, on considère que γ est pris dans $[0,1]^{\mathcal{E}(x)}$ au lieu de $\{0,1\}^{\mathcal{E}^{0+}(x)} \times [0,1]^{\mathcal{E}^{-}(x)}$. On verra ensuite que les valeurs $\{0,1\}^{\mathcal{E}^{0+}(x)}$ ont cependant une place primordiale dans $[0,1]^{\mathcal{E}^{0+}(x)}$.

$$\begin{aligned} g &:= F'_{\mathcal{F}(x)}(x)^{\mathsf{T}}F_{\mathcal{F}(x)}(x) + G'_{\mathcal{G}(x)}(x)^{\mathsf{T}}G_{\mathcal{G}(x)}(x) + [F'_{\mathcal{E}(x)}(x)^{\mathsf{T}}\Gamma + G'_{\mathcal{E}(x)}(x)^{\mathsf{T}}\overline{\Gamma}]H_{\mathcal{E}(x)}(x) \\ &= g_0(x) + [F'_{\mathcal{E}^{0+}(x)}(x) - G'_{\mathcal{E}^{0+}(x)}(x)]^{\mathsf{T}}\,\mathrm{Diag}(H_{\mathcal{E}^{0+}(x)}(x))\gamma_{\mathcal{E}^{0+}(x)} \\ &\quad + [F'_{\mathcal{E}^{-}(x)}(x) - G'_{\mathcal{E}^{-}(x)}(x)]^{\mathsf{T}}\,\mathrm{Diag}(H_{\mathcal{E}^{-}(x)}(x))\gamma_{\mathcal{E}^{-}(x)} \\ &= g_0(x) + \mathcal{M}_+\gamma_{\mathcal{E}^{0+}(x)} + \mathcal{M}_-\gamma_{\mathcal{E}^{-}(x)} \end{aligned} \tag{6.11}$$

4. Pour des moindres-carrés lisses, résoudre l'équation normale donne toujours une direction de descente, à moins que l'itéré ne soit stationnaire.

avec les variables intermédiaires

$$\begin{aligned} g_0(x) &:= F'_{\mathcal{F}(x)}(x)^\mathsf{T} F_{\mathcal{F}(x)}(x) + G'_{\mathcal{G}(x)}(x)^\mathsf{T} G_{\mathcal{G}(x)}(x) + G'_{\mathcal{E}(x)}(x)^\mathsf{T} G_{\mathcal{E}(x)}(x) \\ \mathcal{M}_+ &:= (F'_{\mathcal{E}^{0+}(x)}(x) - G'_{\mathcal{E}^{0+}(x)}(x))^\mathsf{T} \operatorname{Diag}(H_{\mathcal{E}^{0+}(x)}(x)) \\ \mathcal{M}_- &:= (F'_{\mathcal{E}^{-}(x)}(x) - G'_{\mathcal{E}^{-}(x)}(x))^\mathsf{T} \operatorname{Diag}(H_{\mathcal{E}^{-}(x)}(x)) \end{aligned} \tag{6.12}$$

En fait, cette expression de g correspond, lorsque les poids sont arbitraires, à des éléments dans $H(x)^\mathsf{T}\partial_\times H(x)$. En effet, on rappelle que (voir (3.10))

$$\begin{aligned} J_0 &= [F'_{\mathcal{F}(x)}(x); G'_{\mathcal{G}(x)}(x)], \\ \partial_B H(x) \subseteq \partial_B^\times H(x) &:= \{J \in \mathbb{R}^{n\times n} : J_{\mathcal{F}(x)\cup\mathcal{G}(x),:} = J_0, J_{i,:} \in \{F'_i(x), G'_i(x)\}, i \in \mathcal{E}(x)\}, \\ \partial_C H(x) \subseteq \operatorname{conv}(\partial_B^\times H(x)) &:= \partial_\times H(x), \\ \partial_\times H(x) &= \{J \in \mathbb{R}^{n\times n} : J_{\mathcal{F}(x)\cup\mathcal{G}(x),:} = J_0, J_{i,:} = \gamma_i F'_i(x) + \overline{\gamma}_i G'_i(x), i \in \mathcal{E}(x), \gamma_i \in [0,1]\}. \end{aligned}$$

Explicitons ces relations, en se concentrant sur les indices de $\mathcal{E}(x)$:

- $\partial_B H(x)$ correspond à *certaines* valeurs $\gamma \in \{0,1\}^{\mathcal{E}(x)}$ (voir les chapitres 3 et la section 4.3) ;
- $\partial_B^\times H$ correspond à *toutes* les valeurs de $\gamma \in \{0,1\}^{\mathcal{E}(x)}$ (le sur-ensemble du B-différentiel de H) ;
- $\partial_C H(x)$ correspond à *certaines* valeurs de $\gamma \in [0,1]^{\mathcal{E}(x)}$;
- $\partial_\times H(x)$ correspond à *toutes* les valeurs de $\gamma \in [0,1]^{\mathcal{E}(x)}$, ce qui est le cas de valeurs de γ quelconques.

En particulier, $\partial_C H(x)$ est un polytope (enveloppe convexe d'un nombre fini de points). Ensuite, on a

$$\begin{aligned} \partial_\times H(x)^\mathsf{T} H(x) = {} & F'_{\mathcal{F}(x)}(x)^\mathsf{T} F_{\mathcal{F}(x)}(x) + G'_{\mathcal{G}(x)}(x)^\mathsf{T} G_{\mathcal{G}(x)}(x) \\ & + \{[\Gamma F'_{\mathcal{E}(x)}(x) + \overline{\Gamma} G'_{\mathcal{E}(x)}(x)] : \gamma \in [0,1]^{\mathcal{E}(x)}\}^\mathsf{T} H_{\mathcal{E}(x)}(x). \end{aligned}$$

Maintenant que le rôle des poids γ est en partie expliqué, discutons de la pertinence de leurs valeurs. Cela se traduit surtout par la quantité $\theta'(x; -g)$, qui idéalement devrait être négative : cela signifie que l'opposé du gradient du modèle quadratique φ_x est également une direction de descente de θ.[5]

Pour un γ quelconque, il n'y a pas de garantie que $\theta'(x; -g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^{-}(x)})) < 0$, puisque φ_x et θ peuvent largement différer. Cette technicalité est inhérente au caractère non lisse – un phénomène similaire peut se produire pour l'algorithme de Newton-min 2.3.29 lorsque la répartition des indices est "mal" faite, voir [20, exemple 5.8] par exemple.

5. Pour simplifier, on présente le cas $S = I$; autrement, il faut considérer $\theta'(x; -S^{-1}g)$.

Exprimons $\theta'(x,-g)$ pour analyser ses propriétés.

$$\begin{aligned}
\theta'(x;-g) &= F_{\mathcal{F}(x)}(x)^\mathsf{T} F'_{\mathcal{F}(x)}(x)(-g) + G_{\mathcal{G}(x)}(x)^\mathsf{T} G'_{\mathcal{G}(x)}(x)(-g) \\
&\quad + H_{\mathcal{E}(x)}(x)^\mathsf{T} \min(F'_{\mathcal{E}(x)}(x)(-g), G'_{\mathcal{E}(x)}(x)(-g)) \\
&= [F_{\mathcal{F}(x)}(x)^\mathsf{T} F'_{\mathcal{F}(x)}(x) + G_{\mathcal{G}(x)}(x)^\mathsf{T} G'_{\mathcal{G}(x)}(x) \\
&\qquad \pm H_{\mathcal{E}(x)}(x)^\mathsf{T} (\Gamma F'_{\mathcal{E}(x)}(x) + \overline{\Gamma} G'_{\mathcal{E}(x)}(x))](-g) \\
&\quad + H_{\mathcal{E}(x)}(x)^\mathsf{T} \min[F'_{\mathcal{E}(x)}(x)(-g), G'_{\mathcal{E}(x)}(x)(-g)] \\
&= -||g||^2 - [\Gamma F'_{\mathcal{E}(x)}(x) + \overline{\Gamma} G'_{\mathcal{E}(x)}(x)](-g) \\
&\quad + H_{\mathcal{E}(x)}(x)^\mathsf{T} \min[F'_{\mathcal{E}(x)}(x)(-g), G'_{\mathcal{E}(x)}(x)(-g)] \\
&= -||g||^2 + H_{\mathcal{E}(x)}(x)^\mathsf{T} [\min(F'_{\mathcal{E}(x)}(x)(-g), G'_{\mathcal{E}(x)}(x)(-g)) \\
&\qquad - (\Gamma F'_{\mathcal{E}(x)}(x) + \overline{\Gamma} G'_{\mathcal{E}(x)}(x))(-g)]
\end{aligned} \tag{6.13}$$

Dans la dernière expression, observons que le terme $-||g||^2$ est similaire au cas lisse, et que le reste correspond à la non-différentiabilité. Si $\mathcal{E}(x) = \varnothing$, i.e., le système est lisse en x, ce terme compliqué n'intervient pas. On répartir le dernier terme en ses parties $\mathcal{E}^{0+}(x)$ et $\mathcal{E}^{-}(x)$.

$$\begin{aligned}
\theta'(x;-g) = -||g||^2 &+ H_{\mathcal{E}^{0+}(x)}(x)^\mathsf{T} [\min(F'_{\mathcal{E}^{0+}(x)}(x)(-g), G'_{\mathcal{E}^{0+}(x)}(x)(-g)) \\
&\quad - (\Gamma_{\mathcal{E}^{0+}(x)} F'_{\mathcal{E}^{0+}(x)}(x) + \overline{\Gamma}_{\mathcal{E}^{0+}(x)} G'_{\mathcal{E}^{0+}(x)}(x))(-g)] \\
&+ H_{\mathcal{E}^{-}(x)}(x)^\mathsf{T} [\min(F'_{\mathcal{E}^{-}(x)}(x)(-g), G'_{\mathcal{E}^{-}(x)}(x)(-g)) \\
&\quad - (\Gamma_{\mathcal{E}^{-}(x)} F'_{\mathcal{E}^{-}(x)}(x) + \overline{\Gamma}_{\mathcal{E}^{-}(x)} G'_{\mathcal{E}^{-}(x)}(x))(-g)]
\end{aligned} \tag{6.14}$$

Ensuite, observons que les composantes entre crochets sont de la forme $\min(a,b) - \gamma a - \overline{\gamma} b$, i.e., le minimum moins une combinaison convexe de deux valeurs. C'est donc clairement négatif[6]. Ensuite, dans les indices de $\mathcal{E}^{0+}(x)$ on multiplie par $H_{\mathcal{E}^{0+}(x)} \succcurlyeq 0$, donc les lignes 1-2 n'ont que des termes négatifs, alors que pour les lignes restantes (indices dans $\mathcal{E}^{-}(x)$, lignes 3-4), les termes sont positifs. Il est possible d'avoir $\theta'(x;-g) \geqslant 0$.

Exemple 6.1.3 (premier exemple de $-g$ non descendant)**.** Soit $\delta \in \mathbb{R}^+_*$ et $n = 2$, définissons

$$\begin{aligned}
F_1(x) &= x_1 - 2, & G_1(x) &= -2x_1 + 1, \\
F_2(x) &= x_2 - 1 - \delta, & G_2(x) &= 2x_2 - 2 - \delta,
\end{aligned} \tag{6.15}$$

et considérons $x = (1,1)$. On a clairement

$$F_1(x) = -1 = G_1(x), \quad F_2(x) = -\delta = G_2(x), \quad \mathcal{E}^-(x) = \{1,2\}.$$

Avec l'environnement des poids convexes, $g(\gamma)$ devient

$$\begin{aligned}
g(\gamma) &= (\gamma_1 \nabla F_1(x) + \overline{\gamma}_1 \nabla G_1(x)) H_1(x) + (\gamma_2 \nabla F_2(x) + \overline{\gamma}_2 \nabla G_2(x)) H_2(x) \\
&= (\gamma_1 - 2\overline{\gamma}_1) \times (-1) e_1 + (\gamma_2 + 2\overline{\gamma}_2) \times (-\delta) e_2 \\
&= (2\overline{\gamma}_1 - \gamma_1) e_1 - \delta(\gamma_2 + 2\overline{\gamma}_2) e_2.
\end{aligned}$$

6. En effet, $\min(a,b) - \gamma a - \overline{\gamma} b = \min(a - \gamma a - \overline{\gamma} b, b - \gamma a - \overline{\gamma} b) = \min(\overline{\gamma}(a-b), \gamma(b-a)) \leqslant 0$.

Maintenant, pour $\gamma = (1/2, 1/2)$, i.e., $g(1/2, 1/2) = (e_1 - 3\delta e_2)/2$, on a

$$\begin{aligned}\theta'(x; -g) &= (-1)\min(e_1^\mathsf{T}(-g), -2e_1^\mathsf{T}(-g)) + (-\delta)\min(e_2^\mathsf{T}(-g), 2e_2^\mathsf{T}(-g)) \\ &= -\min(-1/2, 1) - \delta\min(3\delta/2, 6\delta/2) \\ &= \frac{1}{2} - \frac{3}{2}\delta^2\end{aligned}$$

qui est positif pour δ assez petit. C'est illustré à la figure 6.1.

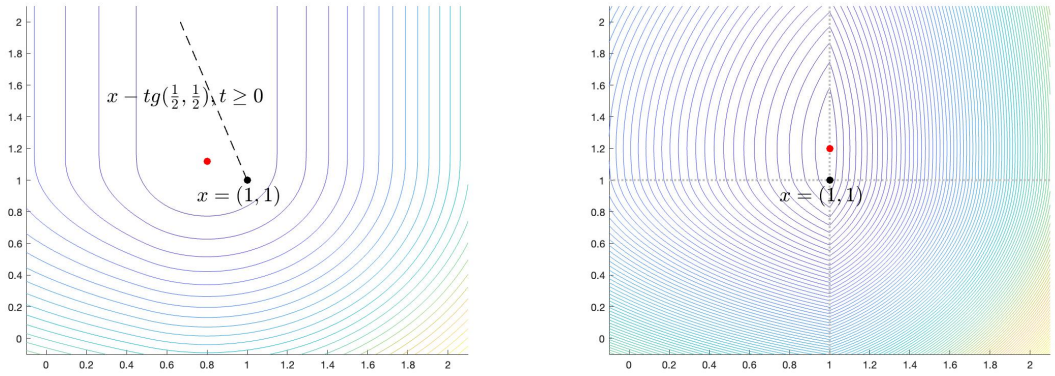


FIGURE 6.1 – Gauche : courbes de niveau de φ_x avec $\gamma = (1/2, 1/2)$. Droite : courbes de niveau de θ ; les lignes pointillées sont les plis (θ n'est pas différentiable). Le point rouge correspond à un minimum local. Les lignes de niveau révèlent de trop grandes différences entre θ et φ_x, donc la direction donnée par φ_x *augmente* θ.

Considérons $\gamma_{\mathcal{E}^-(x)} = (2/3, 1)$ (expliqué plus tard). On a, voir figure 6.2,

$$\begin{aligned}g(\gamma) = g(2/3, 1) &= [2(1 - 2/3) - 2/3]e_1 - \delta[1 + 2(1 - 1)]e_2 = -\delta e_2 \\ \theta'(x; -g) &= (-1)\min(-e_1^\mathsf{T} g, +2e_1^\mathsf{T} g) - \delta\min(-e_2^\mathsf{T} g, -2e_2^\mathsf{T} g) = -\delta\min(\delta_2, 2\delta_2) = -\delta^2\end{aligned}$$

ce qui confirme la décroissance de θ avec ces poids. □

On peut observer cela même dans le cas de $\mathrm{PCL}(M, q)$ avec une $\mathbf{P}$-matrice.

Exemple 6.1.4 (second exemple de $-g$ non descendant)**.** Considérons le PCL suivant (P, q), avec la $\mathbf{P}$-matrice P et le point $\hat{x}$:

$$P = \begin{pmatrix} 1/2 & 1/2 \\ -5 & 1 \end{pmatrix}, \qquad q = \begin{pmatrix} 0 \\ -1/10 \end{pmatrix} \qquad \text{et} \qquad \hat{x} = \begin{pmatrix} -1/50 \\ -1/50 \end{pmatrix}.$$

En $\hat{x}$, $\mathcal{E}(\hat{x}) = \{1, 2\}$; la flèche verte oblique est le gradient $g(\gamma)$ pour $\gamma = (0, 0)$ (son opposé), qui est donc le gradient de θ en $\hat{x}$ dans la partie en bas à droite de la figure où il vaut $x \mapsto \frac{1}{2}\|Px + q\|_2^2$. D'où

$$g = P^\mathsf{T}(P\hat{x} + q) = P^\mathsf{T}\hat{x} = \begin{pmatrix} 9/100 \\ -3/100 \end{pmatrix}.$$

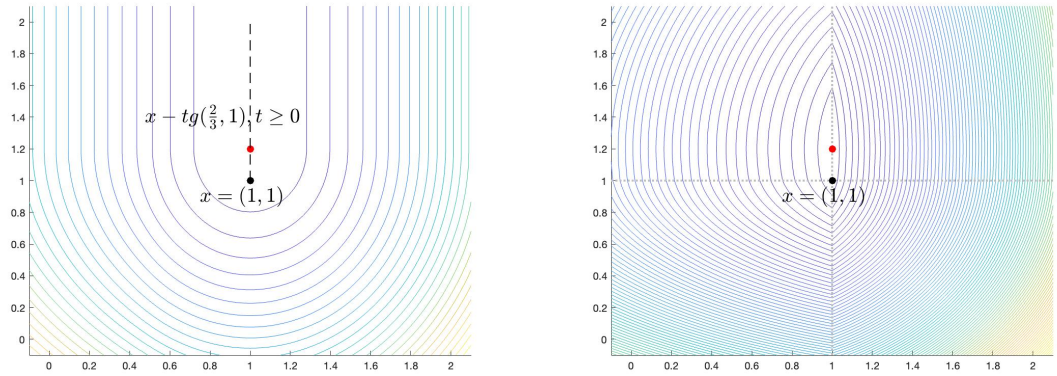


FIGURE 6.2 – Gauche : courbes de niveau de φ_x avec $\gamma = (2/3, 1)$. Droite : courbes de niveau de θ ; les lignes pointillées sont les plis (θ n'est pas différentiable). Le point rouge est un minimum local. Les lignes de niveau de φ_x sont (du moins localement) assez proches de celles de θ pour qu'une direction de descente de φ_x *diminue* θ.

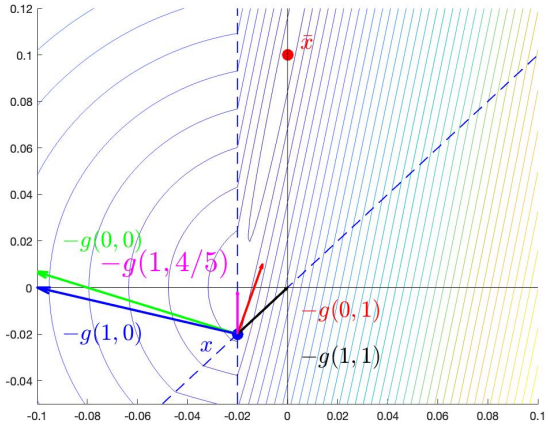


FIGURE 6.3 – Illustration de l'exemple 6.1.4. Les courbes de niveau de θ ($\sqrt{\theta}$ pour la visibilité) sont en couleur, les lignes pointillées en bleu les plis de H (définis par $x_i = (Px + q)_i$ pour $i \in \{1, 2\}$), le point rouge au-dessus est l'unique solution $\bar{x} = (0, 1/10)$ du problème et le point bleu est l'itéré courant. Les flèches en vert, bleu, rouge et noir correspondent à quatre $-g$ possibles pour des choix de γ extrémaux, celle en magenta à une direction de descente.

L'image montre clairement que θ augmente selon $-g$, ce qui est confirmé par le calcul : pour $t > 0$, $\hat{x} - tg \leqslant P(\hat{x} - tg) + q$ et on a

$$\theta(\hat{x} - tg) = \frac{1}{2}\|\hat{x} - tg\|_2^2 = \frac{1}{2}\|\hat{x}\|_2^2 - t\hat{x}^\mathsf{T} g + \frac{t^2}{2}\|g\|_2^2 > \theta(\hat{x}),$$

puisque $-\hat{x}^\mathsf{T} g = 6/5000 > 0$.

Néanmoins, pour $\gamma = (1, 4/5)$, θ décroît. En effet, $g(1, 4/5)$ vaut

$$g = \Gamma\hat{x} + P^{\mathsf{T}}\overline{\Gamma}\hat{x} = \begin{pmatrix} -1/50 \\ -4/250 \end{pmatrix} + \begin{pmatrix} 1/2 & -5 \\ 1/2 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ -1/250 \end{pmatrix} = \begin{pmatrix} 0 \\ -1/50 \end{pmatrix}.$$

Ce gradient est représenté par la flèche magenta verticale dans l'image. On observe que, avec ce g, $-g$ est une direction de descente de θ en $\hat{x}$, puisque θ diminue, et vaut $x \mapsto \frac{1}{2}\|x\|_2^2$ dans cette direction. □

Remarque 6.1.5 (Interaction entre $\boldsymbol{\gamma_{\mathcal{E}^-(x)}}$ et $\boldsymbol{\gamma_{\mathcal{E}^{0+}(x)}}$)**.** On a vu que les indices de $\mathcal{E}^-(x)$ impactaient différemment de ceux dans $\mathcal{E}^{0+}(x)$. Cependant, dans (6.14), la partie $\mathcal{E}^-(x)$ est quadratique en γ (rappelons que x est fixé). En effet, par (6.11) et (6.12), g est affine en γ donc les quantités entre crochets sont quadratiques en γ. De fait, sous cette forme, il n'y a pas de valeur immédiate de $\gamma_{\mathcal{E}^-(x)}$ et $\gamma_{\mathcal{E}^{0+}(x)}$ qui annulent ce terme. Observons que si $\mathcal{E}^-(x) = \varnothing$, tout choix de $\gamma_{\mathcal{E}^{0+}(x)}$ convient. □

Le lemme suivant est un aspect central de ce qui suit : l'annulation du dernier terme dans (6.14). Sa preuve peut être simple mais requiert des notations supplémentaires rappelées ensuite. [7]

Lemme 6.1.6 (annuler le terme de $\boldsymbol{\mathcal{E}^-(x)}$)**.** *Soit $\gamma_{\mathcal{E}^{0+}(x)} \in [0,1]^{\mathcal{E}^{0+}(x)}$, il existe un $\gamma_{\mathcal{E}^-(x)} \in [0,1]^{\mathcal{E}^-(x)}$ tel que le terme lignes 3-4 de* (6.14) *s'annule, i.e.,*

$$\begin{aligned} &H_{\mathcal{E}^-(x)}(x)^{\mathsf{T}}[\min(F'_{\mathcal{E}^-(x)}(x)(-g), G'_{\mathcal{E}^-(x)}(x)(-g)) \\ &-(\Gamma_{\mathcal{E}^-(x)}F'_{\mathcal{E}^-(x)}(x) + \overline{\Gamma}_{\mathcal{E}^-(x)}G'_{\mathcal{E}^-(x)}(x))(-g)] = 0 \\ &\quad \Leftrightarrow \min(F'_{\mathcal{E}^-(x)}(x)(-g), G'_{\mathcal{E}^-(x)}(x)(-g)) = -(\Gamma_{\mathcal{E}^-(x)}F'_{\mathcal{E}^-(x)}(x) + \overline{\Gamma}_{\mathcal{E}^-(x)}G'_{\mathcal{E}^-(x)}(x))g. \end{aligned}$$

□

L'équivalence entre les deux lignes provient du fait que, pour $u < 0$ et $v \leqslant 0$, $u^{\mathsf{T}}v = 0 \Leftrightarrow v = 0$. La preuve de ce lemme a nettement été simplifiée après le détour du chapitre 3, qui mit en lumière le fait que cette valeur de $\gamma_{\mathcal{E}^-(x)}$ est essentiellement obtenue par une projection, i.e., une opération plutôt simple. La difficulté est d'identifier quel point est projeté sur quel ensemble.

Dans la remarque suivante, $V := (G'_{\mathcal{E}(x)}(x) - F'_{\mathcal{E}(x)}(x))^{\mathsf{T}}$ est la différence entre les matrices jacobiennes partielles, qui est reliée au B-différentiel de H (voir chapitre 3).

Remarque 6.1.7 (reformuler le terme de $\boldsymbol{\mathcal{E}^-(x)}$)**.** Dans le lemme 6.1.6, l'équation peut être reformulée comme suit, en isolant le terme $G'_{\mathcal{E}^-(x)}(x)d$, en utilisant la définition de V et en

7. Il est tout à fait possible d'avoir $\theta'(x; -g) \leqslant 0$ même si ce terme est strictement positif, si les autres sont suffisamment négatifs.

multipliant par $\mathrm{Diag}(H_{\mathcal{E}^-(x)}(x))$.

$$\begin{aligned}
\Gamma_{\mathcal{E}^-(x)}F'_{\mathcal{E}^-(x)}(x)(-g) + \overline{\Gamma}_{\mathcal{E}^-(x)}G'_{\mathcal{E}^-(x)}(x)(-g) &= \min(F'_{\mathcal{E}^-(x)}(x)(-g), G'_{\mathcal{E}^-(x)}(x)(-g)) \\
\Gamma_{\mathcal{E}^-(x)}[F'_{\mathcal{E}^-(x)}(x) - G'_{\mathcal{E}^-(x)}(x)] + G'_{\mathcal{E}^-(x)}(x)(-g) &= \min((F'_{\mathcal{E}^-(x)}(x) - G'_{\mathcal{E}^-(x)}(x))(-g), 0) \\
&\quad + G'_{\mathcal{E}^-(x)}(x)(-g) \\
\Gamma_{\mathcal{E}^-(x)}\,\mathrm{Diag}(H_{\mathcal{E}^-(x)}(x))[-V^{\mathsf{T}}_{:,\mathcal{E}^-(x)}](-g) &= \mathrm{Diag}(H_{\mathcal{E}^-(x)}(x))\min(V^{\mathsf{T}}_{:,\mathcal{E}^-(x)}g, 0) \\
\Gamma_{\mathcal{E}^-(x)}{\mathcal{M}_-}^{\mathsf{T}}(-g) &= \max({\mathcal{M}_-}^{\mathsf{T}}(-g), 0)
\end{aligned} \tag{6.16}$$

En utilsant $\tilde{g} = {\mathcal{M}_-}^{\mathsf{T}}(-g)$, la dernière expression s'écrit aussi, pour tout $i \in \mathcal{E}^-(x)$, $\gamma_i\tilde{g}_i = \max(\tilde{g}_i, 0)$.

Maintenant, nous définissont quelques variables intermédiaires. En particulier, on reparamétrise $\gamma_{\mathcal{E}^{0+}(x)}$ et $\gamma_{\mathcal{E}^-(x)}$ qui sont dans $[0,1]$ en η et ζ qui sont dans $[-1,+1]$.

Règle 6.1.8 (correspondance entre variables)**.** Dans la suite, on utilise les quantités suivantes :

$$\begin{aligned}
X = \frac{1}{2}\mathcal{M}_+, \quad Y = -\frac{1}{2}\mathcal{M}_-, \quad \overline{x} - \overline{y} &:= g_0(x) + \frac{\mathcal{M}_+}{2}e + \frac{\mathcal{M}_-}{2}e \\
\eta = 2\gamma_{\mathcal{E}^{0+}(x)} - e, \quad \zeta &= 2\gamma_{\mathcal{E}^-(x)} - e \\
g = g_0(x) + \mathcal{M}_+\gamma_{\mathcal{E}^{0+}(x)} + \mathcal{M}_-\gamma_{\mathcal{E}^-(x)} & \\
= \frac{\mathcal{M}_+}{2}\eta + \frac{\mathcal{M}_-}{2}\zeta + \overline{x} - \overline{y} &= \overline{x} - \overline{y} + X\eta - Y\zeta := g(\eta, \zeta)
\end{aligned} \tag{6.17}$$

Observons que $\overline{x}$ et $\overline{y}$ ne sont pas explicitement définis. En fait, dans la suite, on n'a besoin que de leur différence $\overline{x} - \overline{y}$.[8] Ces variables permettent la reformulation suivante.

Remarque 6.1.9 (reformulation de $\boldsymbol{\theta'(x; -g)}$)**.** On peut exprimer $\theta'(x; -g)$ dans (6.14) comme suit :

$$\begin{aligned}
\theta'(x; (-g)) &= -||g||^2 - \gamma^{\mathsf{T}}_{\mathcal{E}^{0+}(x)}[\mathrm{Diag}(H_{\mathcal{E}^{0+}(x)})(F'_{\mathcal{E}^{0+}(x)}(x) - G'_{\mathcal{E}^{0+}(x)}(x))](-g) \\
&\quad - H^{\mathsf{T}}_{\mathcal{E}^{0+}(x)}\max((G'_{\mathcal{E}^{0+}(x)}(x) - F'_{\mathcal{E}^{0+}(x)}(x))(-g), 0) \\
&\quad - \gamma^{\mathsf{T}}_{\mathcal{E}^-(x)}[\mathrm{Diag}(H_{\mathcal{E}^-(x)})(F'_{\mathcal{E}^-(x)}(x) - G'_{\mathcal{E}^-(x)}(x))](-g) \\
&\quad - H^{\mathsf{T}}_{\mathcal{E}^-(x)}\max((G'_{\mathcal{E}^-(x)}(x) - F'_{\mathcal{E}^-(x)}(x))(-g), 0) \\
&= -||g||^2 - \gamma^{\mathsf{T}}_{\mathcal{E}^{0+}(x)}{\mathcal{M}_+}^{\mathsf{T}}(-g) - e^{\mathsf{T}}\max(-{\mathcal{M}_+}^{\mathsf{T}}(-g), 0) \\
&\quad - \gamma^{\mathsf{T}}_{\mathcal{E}^-(x)}{\mathcal{M}_-}^{\mathsf{T}}(-g) - e^{\mathsf{T}}\min(-{\mathcal{M}_-}^{\mathsf{T}}(-g), 0) \\
&= -||g||^2 - \eta^{\mathsf{T}}X^{\mathsf{T}}(-g) - e^{\mathsf{T}}X^{\mathsf{T}}(-g) - e^{\mathsf{T}}\max(-2X^{\mathsf{T}}(-g), 0) \\
&\quad + \zeta^{\mathsf{T}}Y^{\mathsf{T}}(-g) + e^{\mathsf{T}}Y^{\mathsf{T}}(-g) - e^{\mathsf{T}}\min(2Y^{\mathsf{T}}(-g), 0) \\
&= -||g||^2 - \eta^{\mathsf{T}}X^{\mathsf{T}}(-g) + \zeta^{\mathsf{T}}Y^{\mathsf{T}}(-g) \\
&\quad - e^{\mathsf{T}}\max(X^{\mathsf{T}}(-g), -X^{\mathsf{T}}(-g)) - e^{\mathsf{T}}\min(Y^{\mathsf{T}}(-g), -Y^{\mathsf{T}}(-g)) \\
&= -||g||^2 - \eta^{\mathsf{T}}X^{\mathsf{T}}(-g) + \zeta^{\mathsf{T}}Y^{\mathsf{T}}(-g) - ||X^{\mathsf{T}}(-g)||_1 + ||Y^{\mathsf{T}}(-g)||_1
\end{aligned}$$

8. Comme dans le chapitre 3 où la différence des jacobiennes partielles intervient.

En particulier, on retrouve le fait que le terme d'indices dans $\mathcal{E}^{0+}(x)$ est négatif et celui dans $\mathcal{E}^{-}(x)$ est positif. □

Proposition 6.1.10 (lemme comme projection). *Soit $\gamma_{\mathcal{E}^{0+}(x)} \in [0,1]^{\mathcal{E}^{0+}(x)}$, $\eta = 2\gamma_{\mathcal{E}^{0+}(x)} - e$. Soit $Z_y = Y[-1,+1]^{\mathcal{E}^{-}(x)}$ (le zonotope généré par Y, voir section B.2). Soit $\eta \in [-1,+1]^{\mathcal{E}^{0+}(x)}$, $z = \mathrm{P}_{Z_y}(\overline{x} - \overline{y} + X\eta)$ et $\zeta^* \in [-1,+1]^{\mathcal{E}^{-}(x)}$ tels que $z = Y\zeta^*$.*

Alors $\gamma_{\mathcal{E}^{-}(x)} = (1+\zeta^*)/2$ est une valeur vérifiant le lemme 6.1.6 pour $\gamma_{\mathcal{E}^{0+}(x)}$. □

Preuve. Soit $z_0 = \overline{x} - \overline{y} + X\eta$, et considérons $\mathrm{P}_{Z_y}(\overline{x} - \overline{y} + X\eta)$ (voir remarque B.2.7 et proposition B.2.8). Puisque l'ensemble Z_y est fermé et convexe (transformation affine d'un compact convexe), la projection est bien définie. Ce problème s'écrit

$$\min_{z \in Z_y} \frac{1}{2}||z - z_0||^2 = \min_{\zeta \in [-1,+1]^{\mathcal{E}^{-}(x)}} \frac{1}{2}||Y\zeta - z_0||^2$$

qui a des contraintes clairement qualifiées (affines et ensemble non vide). De fait, les conditions d'optimalité s'écrivent

$$\mathrm{KKT} \quad \begin{cases} Y^\mathsf{T}(Y\zeta - z_0) - \mu + \nu = 0, \\ 0 \leqslant \mu \perp (-e - \zeta) \leqslant 0, \\ 0 \leqslant \nu \perp (\zeta - e) \leqslant 0. \end{cases}$$

Par les contraintes de complémentarité, on a $\zeta_i = -1 \Rightarrow \nu_i = 0$ et $\zeta_i = +1 \Rightarrow \mu_i = 0$. De plus, si $\zeta_i \in (-1,+1)$, $\mu_i = 0 = \nu_i = (Y^\mathsf{T}(Y\zeta - z_0))_i$. Le système de KKT devient (en utilisant y_i pour les colonnes de Y)

$$\begin{cases} \zeta_i = +1, & y_i^\mathsf{T}(Y\zeta - z_0) = -\nu_i \leqslant 0, \\ \zeta_i = -1, & y_i^\mathsf{T}(Y\zeta - z_0) = +\mu_i \geqslant 0, \\ \zeta_i \in (-1,+1), & y_i^\mathsf{T}(Y\zeta - z_0) = 0, \end{cases} \Leftrightarrow \begin{cases} \zeta_i \in \{-1,+1\}, & \zeta_i y_i^\mathsf{T}(z_0 - Y\zeta) \geqslant 0, \\ \zeta_i \in (-1,+1), & y_i^\mathsf{T}(z_0 - Y\zeta) = 0. \end{cases}$$

Maintenant, rappelons que $z_0 = \overline{x} - \overline{y} + X\eta$, donc $z_0 - Y\zeta = g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^{-}(x)}) = g$. De plus, dans la remarque 6.1.7, puisque $Y = -\mathcal{M}_-/2$, on peut récrire (6.16) comme

$$\begin{aligned} \Gamma_{\mathcal{E}^{-}(x)} \mathcal{M}_-{}^\mathsf{T}(-g) &= \max(\mathcal{M}_-{}^\mathsf{T}(-g), 0) \\ \Leftrightarrow \frac{1}{2}(\zeta + e) \cdot (-2Y)^\mathsf{T}(-g) &= \max((-2Y)^\mathsf{T}(-g), 0) \\ \Leftrightarrow (\zeta + e) \cdot Y^\mathsf{T} g &= \max(2Y^\mathsf{T} g, 0) \\ \Leftrightarrow \zeta \cdot Y^\mathsf{T} g &= \max(Y^\mathsf{T} g, -Y^\mathsf{T} g) \\ \Leftrightarrow \zeta \cdot Y^\mathsf{T} g &= |Y^\mathsf{T} g| \\ \Leftrightarrow \forall\, i \in \mathcal{E}^{-}(x), \zeta_i y_i^\mathsf{T} g &= |y_i^\mathsf{T} g| \end{aligned}$$

qui est la même chose que le système KKT précédent. □

Mentionnons un point pertinent : même si la projection z est unique (car bien définie), l'ensemble $\{\zeta \in [-1,+1]^{\mathcal{E}^{-}(x)} : Y\zeta = z\}$ peut ne pas être réduit à un point. Cependant, n'importe lequel de ces ζ produit la même valeur de $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^{-}(x)})$. En effet, rappelons que $g = \overline{x} - \overline{y} + X\eta - Y\zeta = \overline{x} - \overline{y} + X\eta - z$ qui est indépendant du ζ choisi.

Contre-exemple 6.1.11 (non-unicité de $\boldsymbol{\zeta}$)**.** Considérons les données $Y = [e_1\ e_2\ e_1 + e_2]$ en dimension 3 et projetons e_3 sur $Y[-1,+1]^3$. Le système KKT est résolu par $\zeta = (t, t, -t)$ pour tout $t \in [-1,+1]$ et $\mu = 0 = \nu$. En effet, $Y^\mathsf{T}(Y[t;t;-t] - e_3) = Y^\mathsf{T}(0 - e_3) = 0$. □

Nous résumons les propriétés précédentes dans la discussion et la définition qui suivent. Soit X, Y et $\overline{x} - \overline{y}$ comme décrits dans la règle 6.1.8. Pour tout $\gamma_{\mathcal{E}^{0+}(x)} \in [0,1]^{\mathcal{E}^{0+}(x)}$ et son équivalent $\eta = 2\gamma_{\mathcal{E}^{0+}(x)} - e \in [-1,+1]^{\mathcal{E}^{0+}(x)}$, on définit $z := \mathrm{P}_{Y[-1,+1]^{\mathcal{E}^-(x)}}(\overline{x} - \overline{y} + X\eta)$ et $\zeta \in [-1,+1]^{\mathcal{E}^-(x)}$ tels que $z = Y\zeta$ et $\gamma_{\mathcal{E}^-(x)} = (\zeta + e)/2$. Alors on a, par le lemma 6.1.6,

$$\begin{gathered}\min(F'_{\mathcal{E}^-(x)}(x)(-g), G'_{\mathcal{E}^-(x)}(x)(-g)) = (\Gamma_{\mathcal{E}^-(x)}F'_{\mathcal{E}^-(x)}(x) + \overline{\Gamma}_{\mathcal{E}^-(x)}G'_{\mathcal{E}^-(x)}(x))(-g),\\ \theta'(x;-g) \leqslant -||g||^2,\end{gathered}$$

ce qui signifie que $-g$ est une direction de descente de θ s'il est non nul.

Définition 6.1.12 (choisir $\boldsymbol{\gamma_{\mathcal{E}^-(x)}}$ pour avoir la descente)**.** Soit $\gamma_{\mathcal{E}^{0+}(x)} \in [0,1]^{\mathcal{E}^{0+}(x)}$, $\eta = 2\gamma_{\mathcal{E}^{0+}(x)} - e$ et $z = \mathrm{P}_{Y[-1,+1]^{\mathcal{E}^-(x)}}(\overline{x} - \overline{y} + X\eta) = Y\zeta$.

Dans la suite, on définit $\mathbb{G}(\gamma_{\mathcal{E}^{0+}(x)}) \subseteq [0,1]^{\mathcal{E}^-(x)}$ comme le sous-ensemble $\mathbb{G}(\gamma_{\mathcal{E}^{0+}(x)}) := \{\gamma_{\mathcal{E}^{0+}(x)} \in [0,1]^{\mathcal{E}^{0+}(x)} : Y(2\gamma_{\mathcal{E}^-(x)} - e) = Y\zeta\}$. Ensuite, on considère une valeur particulière de cet ensemble, définie par [9]

$$\begin{array}{rlll}\gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^{0+}(x)}) := & \operatorname{argmin} & \frac{1}{2}||\gamma_{\mathcal{E}^-(x)} - e/2||^2 & \\ & \text{s.t.} & \gamma_{\mathcal{E}^-(x)} \in \mathbb{G}(\gamma_{\mathcal{E}^{0+}(x)}) \Leftrightarrow & \gamma_{\mathcal{E}^-(x)} \in [0,1]^{\mathcal{E}^-(x)} \\ & & & Y(2\gamma_{\mathcal{E}^-(x)} - e) = Y\zeta.\end{array}$$ □

Observons que $\gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^{0+}(x)})$ est bien définie en vertu d'être la projection de $e/2$ sur un ensemble non vide (puisque $(\zeta + e)/2$ y appartient) défini par des contraintes affines. On a choisi l'élément le plus proche du centre de l'hypercube ($\gamma_{\mathcal{E}^-(x)} = e/2$ ou $\zeta = 0$ est le centre de l'hypercube), mais on aurait pu choisir une autre convention.

Nous finissons cette section par quelques remarques et illustrations des propriétés précédentes.

- Pour un $\gamma_{\mathcal{E}^{0+}(x)}$ donné, $\mathbb{G}(\gamma_{\mathcal{E}^{0+}(x)})$ est fermé convexe compact.
- L'ensemble des $\gamma_{\mathcal{E}^{0+}(x)}$ tels que $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^{0+}(x)})) = 0$ n'est pas nécessairement de mesure nulle dans $[0,1]^{\mathcal{E}^{0+}(x)}$.

Maintenant, retournons sur l'exemple 6.1.3 : par de simples calculs, on a

$$X = \varnothing, Y = \frac{1}{2}\begin{bmatrix}3 & 0\\ 0 & -\delta\end{bmatrix}, \overline{x} - \overline{y} = \begin{bmatrix}2\\ -2\delta\end{bmatrix} - \frac{1}{2}\begin{bmatrix}3 & 0\\ 0 & -\delta\end{bmatrix} e = \begin{bmatrix}1/2\\ -\delta\end{bmatrix}.$$

La projection du point $(1/2, -\delta)$ sur $Y[-1,+1]^2 = [-3/2, 3/2] \times [-\delta/2, +\delta/2]$ est le point $(1/2, -\delta/2)$ correspondant à $\zeta = (1/3, 1)$ et $\gamma_{\mathcal{E}^-(x)} = (2/3, 1)$.

9. La dernière contrainte s'écrit aussi $\overline{x} - \overline{y} + X\eta - Y(2\gamma_{\mathcal{E}^-(x)} - e) = g$.

De même, dans l'exemple 6.1.4, on a

$$X = \varnothing, Y = \frac{1}{200}\begin{bmatrix}1 & 10\\ 1 & 0\end{bmatrix}, \overline{x} - \overline{y} = \frac{1}{100}\begin{bmatrix}9\\ -3\end{bmatrix} - \frac{1}{200}\begin{bmatrix}11\\ 1\end{bmatrix} = \frac{1}{200}\begin{bmatrix}7\\ -7\end{bmatrix}.$$

La projection du point $(7,-7)/200$ sur $Y[-1,+1]^2$ est $(7,-1)/200$ qui correspond à $\zeta = (1, 3/5)$ et $\gamma_{\mathcal{E}^{0+}(x)} = (1, 4/5)$.

6.1.3 Choix des poids et stationnarité

Conditions nécessaires et suffisantes de stationnarité

L'environnement présenté ci-dessus introduit des poids convexes pour les indices dans $\mathcal{E}(x)$, et propose une façon de choisir une partie des valeurs, $\gamma_{\mathcal{E}^-(x)}$, pour s'assurer que l'algorithme a une direction de descente. Cependant, la choix de $\gamma_{\mathcal{E}^{0+}(x)}$ reste ouvert. Un obstacle important est l'observation suivante. Rappelons que $\mathbb{G}$ est introduit dans la définition 6.1.12.

Remarque 6.1.13 (cas $g = 0$)**.** Pour un certain $\gamma_{\mathcal{E}^{0+}(x)}$, il est possible que pour tout $\gamma_{\mathcal{E}^-(x)} \in \mathbb{G}(\gamma_{\mathcal{E}^{0+}(x)})$, on ait $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}) = 0$. □

C'est en réalité pertinent pour la proposition suivante, qui caractérise la stationnarité forte/Dini de θ (définition 2.3.39) en un point x.

Proposition 6.1.14 (CNS de θ-stationnarité)**.** *Les propriétés suivantes sont équivalentes :*

(i) *x est un point stationnaire fort/Dini de θ, i.e.* $\forall\, h \in \mathbb{R}^n, \theta'(x;h) \geqslant 0$,
(ii) *($\mathcal{C}_\theta^{[]}$) pour tout $\gamma_{\mathcal{E}^{0+}(x)} \in [0,1]^{\mathcal{E}^{0+}(x)}$ et $\gamma_{\mathcal{E}^-(x)} \in \mathbb{G}(\gamma_{\mathcal{E}^{0+}(x)})$, on a $g = 0$,*
(iii) *($\mathcal{C}_\theta^{\{\}}$) pour tout $\gamma_{\mathcal{E}^{0+}(x)} \in \{0,1\}^{\mathcal{E}^{0+}(x)}$ et $\gamma_{\mathcal{E}^-(x)} \in \mathbb{G}(\gamma_{\mathcal{E}^{0+}(x)})$, on a $g = 0$.*

Dans ce qui suit, on utilise $(\mathcal{C}_\theta)$ pour référer à $(\mathcal{C}_\theta^{[]})$ ou $(\mathcal{C}_\theta^{\{\}})$. Il est clair que le point (iii) est un cas particulier de (ii), et qui correspond aux partitions de $\mathcal{E}^{0+}(x)$, comme c'est fait dans les algorithmes de Newton-min et NMP. Observons que même dans ce cas, le $\gamma_{\mathcal{E}^-(x)}$ correspondant donné par la définition 6.1.12 n'est pas nécessairement dans $\{0,1\}^{\mathcal{E}^-(x)}$.

Preuve. $[(i) \Rightarrow (ii)]$ Par contraposée, si (ii) n'est pas vérifié, il existe un certain $\gamma_{\mathcal{E}^{0+}(x)}$ tel que $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^{0+}(x)})) \neq 0$ où $\gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^{0+}(x)}) \in \mathbb{G}(\gamma_{\mathcal{E}^{0+}(x)})$ (définition 6.1.12). De fait, c'est une direction non nulle de descente stricte, donc θ n'est pas Dini-stationnaire à x.

$[(ii) \Rightarrow (iii)]$ Clair puisque (iii) ne considère qu'une partie des cas de (ii).

$[(iii) \Rightarrow (i)]$ Soit $h \in \mathbb{R}^n$ une direction quelconque et montrons que $\theta'(x;h) \geqslant 0$. Considérons la valeur de $\gamma_{\mathcal{E}^{0+}(x)}$ particulière définie comme suit :

$$\begin{cases} \mathcal{E}^{0+}_{\mathcal{F}}(x) := \{i \in \mathcal{E}^{0+}(x) : F_i'(x)h \leqslant G_i'(x)h\}, \\ \mathcal{E}^{0+}_{\mathcal{G}}(x) := \{i \in \mathcal{E}^{0+}(x) : F_i'(x)h > G_i'(x)h\}, \end{cases} \qquad \gamma_i = \begin{cases} 1 & i \in \mathcal{E}^{0+}_{\mathcal{F}}(x), \\ 0 & i \in \mathcal{E}^{0+}_{\mathcal{G}}(x). \end{cases}$$

Clairement, $\gamma_{\mathcal{E}^{0+}(x)} \in \{0,1\}^{\mathcal{E}^{0+}(x)}$. Maintenant, considérons le $\gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^{0+}(x)})$ associé de la définition 6.1.12. Par (iii), $g = 0$ et (6.14) devient

$$\begin{aligned}
\theta'(x;h) &= (F_{\mathcal{F}(x)}(x)^\mathsf{T} F'_{\mathcal{F}(x)}(x) + G_{\mathcal{G}(x)}(x)^\mathsf{T} G'_{\mathcal{G}(x)}(x))h \\
&\quad + H_{\mathcal{E}(x)}(x)^\mathsf{T} \min(F'_{\mathcal{E}(x)}(x)h, G'_{\mathcal{E}(x)}(x)h) \\
[(6.11)] \quad &= (g - [F'_{\mathcal{E}(x)}(x)^\mathsf{T}\Gamma + G'_{\mathcal{E}(x)}(x)^\mathsf{T}\overline{\Gamma}]H_{\mathcal{E}(x)}(x))^\mathsf{T} h \\
&\quad + H_{\mathcal{E}(x)}(x)^\mathsf{T} \min(F'_{\mathcal{E}(x)}(x)h, G'_{\mathcal{E}(x)}(x)h) \\
[g = 0] \quad &= H_{\mathcal{E}(x)}(x)^\mathsf{T}[\min(F'_{\mathcal{E}(x)}(x)h, G'_{\mathcal{E}(x)}(x)h) - (\Gamma F'_{\mathcal{E}(x)}(x) + \overline{\Gamma} G'_{\mathcal{E}(x)}(x))h]
\end{aligned}$$

Maintenant, on répartit les indices entre $\mathcal{E}^{0+}_{\mathcal{F}}(x)$, $\mathcal{E}^{0+}_{\mathcal{G}}(x)$ et $\mathcal{E}^-(x)$:

- si $i \in \mathcal{E}^{0+}_{\mathcal{F}}(x)$, on a $\min(F'_i(x)h, G'_i(x)h) = F'_i(x)h$, $\gamma_i = 1$ et $\overline{\gamma}_i = 0$, donc le crochet s'annule ;
- si $i \in \mathcal{E}^{0+}_{\mathcal{G}}(x)$, on a $\min(F'_i(x)h, G'_i(x)h) = G'_i(x)h$, $\gamma_i = 0$ et $\overline{\gamma}_i = 1$, donc le crochet s'annule ;
- si $i \in \mathcal{E}^-(x)$, le terme entre crochets est négatif car de la forme $\min(a,b) - \gamma a - \overline{\gamma} b$, donc après multiplication par $H_i(x) < 0$, cela devient positif.

En sommant sur tous les indices, $\theta'(x;h) \geqslant 0$ pour tout h, i.e., θ est fortement stationnaire en x. □

La conséquence principale de cette proposition est d'exprimer, avec le formalisme des poids convexes, que soit un point est Dini-stationnaire soit il y a une direction de descente.

Commentaires sur la complexité de $(\mathcal{C}_\theta)$

Bien que $(\mathcal{C}_\theta)$ puisse apparaître comme un critère d'arrêt pratique (l'algorithme cesse ou a une garantie de progression), cette vérification peut ne pas être faisable en temps polynomial. Nous montrons cela en deux étapes simples : la reformulation des points (ii) et (iii) de la proposition 6.1.14 comme un problème géométrique, et l'utilisation d'une paire de références traitant de ce problème géométrique particulier pour montrer sa nature combinatoire.

Proposition 6.1.15 (reformulation polytopique)**.** *Les propositions suivantes sont équivalentes*

(i) x *est un point fortement stationnaire de* θ*,*
(ii) $\forall\, \gamma = (\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)})$ *avec* $\gamma_{\mathcal{E}^-(x)} \in \mathbb{G}(\gamma_{\mathcal{E}^{0+}(x)})$, $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}) = 0$,
(iii) $\forall\, \gamma_{\mathcal{E}^{0+}(x)} \in [0,1]^{\mathcal{E}^{0+}(x)}, \exists\, \gamma_{\mathcal{E}^-(x)} \in [0,1]^{\mathcal{E}^-(x)}$ *tel que* $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}) = 0$,
(iv) $\mathcal{M}_+[0,1]^{\mathcal{E}^{0+}(x)} \subseteq -g_0(x) - \mathcal{M}_-[0,1]^{\mathcal{E}^-(x)}$.
(v) $\overline{x} + X[-1,+1]^{\mathcal{E}^{0+}(x)} \subseteq \overline{y} + Y[-1,+1]^{\mathcal{E}^-(x)}$.

Preuve. [$(i) \Leftrightarrow (ii)$] Clair puisque (ii) est une reformulation du point (ii) de la proposition 6.1.14.

$[(ii) \Leftrightarrow (iii)]$ Le sens $\Rightarrow$ est clair ; le sens $\Leftarrow$ provient du fait que les valeurs de $\gamma_{\mathcal{E}^-(x)}$ telles que $g = 0$ sont solutions.

$[(iii) \Leftrightarrow (iv)]$ Rappelons (6.11), où $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}) = g_0(x) + \mathcal{M}_+\gamma_{\mathcal{E}^{0+}(x)} + \mathcal{M}_-\gamma_{\mathcal{E}^-(x)}$. De fait, $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}) = 0$ s'écrit

$$g_0(x) + \mathcal{M}_+\gamma_{\mathcal{E}^{0+}(x)} + \mathcal{M}_-\gamma_{\mathcal{E}^-(x)} = 0 \Leftrightarrow \mathcal{M}_+\gamma_{\mathcal{E}^{0+}(x)} = -g_0(x) - \mathcal{M}_-\gamma_{\mathcal{E}^-(x)}.$$

Ensuite, l'équivalence est un résultat de l'inclusion de (iv).

$[(iv) \Leftrightarrow (v)]$ Par la règle 6.1.8, on a les inclusions équivalentes suivantes :

$$\begin{aligned}
\mathcal{M}_+[0,1]^{\mathcal{E}^{0+}(x)} &\subseteq -g_0(x) - \mathcal{M}_-[0,1]^{\mathcal{E}^-(x)} \\
[\mathcal{M}_+ = 2X] \quad 2X[0,1]^{\mathcal{E}^{0+}(x)} &\subseteq -g_0(x) + 2Y[0,1]^{\mathcal{E}^-(x)} \quad [\mathcal{M}_- = -2Y] \\
X([-1,+1]^{\mathcal{E}^{0+}(x)} + e) &\subseteq -g_0(x) + Y([-1,+1]^{\mathcal{E}^-(x)} + e) \\
X[-1,+1]^{\mathcal{E}^{0+}(x)} &\subseteq -g_0(x) - Xe + Ye + Y[-1,+1]^{\mathcal{E}^-(x)} \\
X[-1,+1]^{\mathcal{E}^{0+}(x)} &\subseteq -g_0(x) - \frac{\mathcal{M}_+}{2}e - \frac{\mathcal{M}_-}{2}e + Y[-1,+1]^{\mathcal{E}^-(x)} \\
\overline{x} + X[-1,+1]^{\mathcal{E}^{0+}(x)} &\subseteq \overline{y} + Y[-1,+1]^{\mathcal{E}^-(x)} \quad [g_0(x) + \frac{\mathcal{M}_+}{2}e + \frac{\mathcal{M}_-}{2}e = \overline{x} - \overline{y}]
\end{aligned}$$

ce qui termine la preuve. □

Dans la dernière ligne, on a mis $\overline{x}$ et $\overline{y}$ de côtés différents pour s'accorder avec le formalisme d'un article pertinent traitant (partiellement) la question (voir l'annexe C). Clairement, cela ne change pas si l'inclusion est vraie ou non.

En particulier, les points (iv) et (v) montrent que vérifier la stationnarité est équivalent à un problème d'inclusion de polytopes. C'est là où le problème se cache : la complexité d'une telle inclusion dépend surtout de comment les polytopes sont décrits. En général, les polytopes sont décrits soit comme une liste de sommets, la formulation V, ou comme une intersection de demi-espaces, la formulation H. Par exemple, déterminer si un H-polytope est inclus dans un V-polytope n'est en général pas déterminable en temps polynomial alors que les autres cas le sont. C'est fortement relié au fait que passer d'une formulation V à une formulation H (ou l'inverse) n'est pas réalisable en temps polynomial (voir par exemple [81, 96, 13, 34, 263]).

Les polytopes des points (iv) et (v) ne sont ni en forme H ni en forme V, mais sont des *zonotopes* (pour davantage d'informations sur ces polytopes, voir la section B.2 et par exemple les références [167, 263]). Puisque l'on se concentre sur l'inclusion de zonotopes, nous mentionnons deux articles relativement récents sur le sujet. Le premier [223] discute de méthodes algorithmiques pour résoudre plusieurs problèmes d'inclusions polytopiques, dont le cas particulier des zonotopes. Une méthode simple est proposée, qui peut confirmer que l'inclusion est vraie, en calculant via un problème d'optimisation linéaire qui calcule une valeur scalaire qui joue le rôle d'une sorte de facteur de dilatation.

Lorsque la valeur optimale du problème est $\leqslant 1$, l'inclusion est vérifiée, i.e., le premier zonotope $\overline{x} + X[-1,+1]^{\mathcal{E}^{0+}(x)}$ est inclus dans le second $\overline{y} + Y[-1,+1]^{\mathcal{E}^{-}(x)}$. Cependant, lorsque ce facteur est > 1, l'inclusion peut être valide – voir [223, example 2]. C'est lié avec la propriété principale du second article [149], qui montre que, en général, l'inclusion de zonotopes ("zonotope containment") est co-NP-complet.

Théorème 6.1.16 (corollaire 4 de [149]). *Déterminer une inclusion de zonotopes est co-NP-complet.* □

Remarque 6.1.17 ('discret' versus 'continu'). Dans les propositions 6.1.14 et 6.1.15, on a l'équivalence entre $[0,1]^{\mathcal{E}^{0+}(x)}$ et $\{0,1\}^{\mathcal{E}^{0+}(x)}$ puisque vérifier si un polytope convexe P_1 ($[0,1]$) est contenu dans un polytope P_2 est équivalent à vérifier si les sommets de P_1 ($\{0,1\}$) sont contenus dans P_2. De plus, il est clair que les sommets du zonotope $V[0,1]^m$ pour $V \in \mathbb{R}^{n\times m}$ sont contenus dans $V\{0,1\}^m$. □

Mentionnons un cas particulier, bien que restrictif, qui réduit l'inclusion de zonotopes à un simple POL.

Proposition 6.1.18 (injectivité dans $\boldsymbol{\mathcal{E}^-(x)}$). *Si la matrice $(G'(x) - F'(x))_{\mathcal{E}^-(x),:}$ est surjective, ou de façon équivalente les matrices Y et $\mathcal{M}_-$ sont injectives, alors l'inclusion peut être résolue en temps polynomial, et un $\gamma_{\mathcal{E}^{0+}(x)}$ tel que $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^{0+}(x)})) \neq 0$ peut être obtenu lorsque l'inclusion n'est pas vérifiée.* □

La preuve est donnée dans les annexes, dans la proposition C.0.7 (un cas (très) particulier de [223, thm 3, corollary 2]). Brièvement, lorsque Y est injective, elle a une inverse à gauche, donc le second zonotope est similaire à un hypercube et la paramétrisation ζ est unique (voir l'exemple 6.1.11). Usuellement, cela ne peut se produire pour les zonotopes puisque la matrice Y a plus de colonnes que de lignes. Ici, c'est possible puisque $Y \in \mathbb{R}^{n\times|\mathcal{E}^-(x)|}$ et clairement $|\mathcal{E}^-(x)| \leqslant n$ puisque $\mathcal{E}^-(x) \subseteq [1:n]$.

Dans [149], quelques algorithmes énumératifs sont proposés. L'annexe C donne quelques détails supplémentaires sur la question de l'inclusion de zonotopes et sur l'algorithme simple (mais inexact) proposé dans [223], ainsi qu'une façon d'inclure l'énumération dans un unique POL à variable binaires [10].

La difficulté exposée de trouver une direction de descente / vérifier la stationnarité dans le cas non convexe non différentiable est discutée par exemple dans [19] : "Accordingly, there are very few methods that are guaranteed to converge to stationary points in the case of a nonsmooth and nonconvex objective function." (paragraphe entre les pp. 57 et 58). La méthode proposée, bien que créée pour d'autres problèmes, traite d'un problème non lisse non convexe sous contraintes. En particulier, leur algorithm calcule un "positive spanning set" des directions admissibles, qui peut avoir une complexité exponentielle. Le

10. Qui est possiblement plus simple que de lancer l'énumération combinatoire explicite et peut être résolu par des codes efficaces comme Gurobi.

remède proposé est d'utiliser de l'aléatoire lors de la sélection des directions plutôt que de faire tout le calcul exponentiel.

Dans notre cadre, on pourriat imaginer sélectionner aléatoirement quelques valeurs de $\gamma_{\mathcal{E}^{0+}(x)} \in \{0,1\}^{\mathcal{E}^{0+}(x)}$ jusqu'à ce que l'une retourne un $g \neq 0$ (si un critère d'arrêt approximatif n'est pas vérifé). En effet, puisque l'on cherche des directions de descente (en général, x n'est pas un point stationnaire), il "suffit" d'en trouver une pour démarrer l'itération de Levenberg-Marquardt en x.

En plus, soulignons la chose suivante : la proposition 6.1.14 et le théorème 6.1.16 indiquent que même la *vérification* de la forte stationnarité de θ est en général non polynomiale. De fait, *obtenir* un tel point est vraisemblablement, sans hypothèses fortes, peu réaliste. Pour les PCL, de telles hypothèses sont discutées dans la section 6.1.5.

6.1.4 Choix des poids et différentiels

Les parties précédentes illustrent un lien entre la recherche de directions de descente de θ et le traitement des indices dans $\mathcal{E}(x)$ comme dans l'algorithme de Newton-min par exemple. L'observation suivante semble reliée à ces questions.

Proposition 6.1.19 (projection de zéro sur le sous-différentiel)**.** *Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction locallement lipschitzienne et soit $\partial f(x)$ son différentiel (de Clarke) en x. Soit $g := \mathrm{P}_{\partial f}(0)$* [11]*, alors $-g$ est une direction de descente de f en x.*

Preuve. Par la définition de la projection orthogonale sur un convexe, on a l'inégalité suivante

$$\forall\, \xi \in \partial f(x), (\xi - g, g - 0) \geqslant 0.$$

Qui s'écrit aussi $(\xi, g) \geqslant ||g||^2$ ou $(\xi, -g) \leqslant -||g||^2$ qui est strictement négatif si $g \neq 0$. Cependant, $g = 0$ signifie que 0 est dans le différentiel, donc que f est faiblement stationnaire en x. Ensuite, en rappelant que $f' \leqslant f^\circ$ (voir la remarque sous la définition 2.3.7), on a

$$f'(x; -g) \leqslant f^\circ(x; -g) = \max\{(\xi, -g) : \xi \in \partial f(x)\} \leqslant -||g||^2 \leqslant 0$$

où la seconde égalité provient de la définition 2.3.9. Clairement, $-g$ est une direction de descente stricte si elle est non nulle. □

Cette proposition, inspirée du cas convexe, peut en particulier être utilisée pour θ. Naturellement, elle requiert la connaissance de $\partial\theta = \partial H^\mathsf{T} H$, ce qui sera usuellement difficile à calculer. [12]

11. Le différentiel de Clarke est un convexe fermé, voir définition 2.3.9 et [51, §2].

12. Pas de formule pour ∂H, $\partial H = \mathrm{conv}(\partial_B H)$ requiert la caractérisation non triviale de $\partial_B H$ (chapitres 3-4).

De plus, lorsque le g retourné est 0, i.e., $0 \in \partial\theta(x)$, il peut toujours y avoir des directions de descente – le point est stationnaire mais pas Dini-stationnaire (définitions 2.3.38 et 2.3.39). Cette situation est illustrée à l'exemple suivant, qui reviendra plus tard.

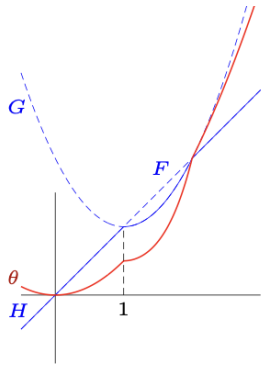


FIGURE 6.4 – Illustration de la faible stationnarité. On a $F(x) = x$, $G(x) = 1 + (x-1)^2$, donc le problème a une solution en $x = 0$. À $x = 1$, ni H ni θ ne sont différentiables. Puisque $G'(1) = 0$, en prenant une suite $1 + t_k \to 1$, on a que $0 \in \partial\theta(1)$, mais $x = 1$ n'est clairement pas fortement stationnaire. De tels points sont parfois appelés "plis concaves", qui, comme on le verra, sont difficiles à traiter.

L'inadéquation de la proposition 6.1.19 peut provenir de l'explication suivante : on a utilisé l'inégalité $f' \leqslant f^\circ$. Cependant, comme remarqué en section 2.3.2, cette inégalité peut être stricte et imprécise lorsque la fonction fait intervenir un minimum. Dit autrement, cette proposition ne peut espérer obtenir mieux que des points Clarke-stationnaires puisque de tels points retournent $g = 0$. En $x = 1$ dans la figure 6.4, $g(\gamma) = \gamma \times 1 + \overline{\gamma} \times 0$ qui est non nul pour $\gamma > 0$, en accord avec la proposition 6.1.14.

Nous concluons cette section par la propriété et discussion suivantes. Résumons :

- remplacement du système polyédrique et de ses hypothèses par les moindres-carrés ;
- utilisation de poids convexes pour équilibrer les indices entre eux ;
- relier ces poids avec le(s) différentiel(s) de θ ;
- donner une condition partielle sur les poids pour avoir des directions de descente ;
- déduire une une CNS de la forte stationnarité de θ et sa complexité.

Avant de discuter d'un algorithme basé sur une régularisation Levenberg-Marquardt des moindres-carrés, mentionnons une dernière propriété. Sa preuve étant plutôt longue et technique, elle est laissée à l'annexe D.

Proposition 6.1.20 ($g \in \partial\theta(x)$ avec un $\gamma_{\mathcal{E}^{0+}(x)}$ approprié). *Supposons que θ n'est pas Dini-stationnaire en x. Alors il existe $\gamma_{\mathcal{E}^{0+}(x)}$ tel que $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^{0+}(x)})) \in \partial\theta(x)$. De plus, un tel $\gamma_{\mathcal{E}^{0+}(x)}$ peut être trouvé en résolvant les problèmes équivalents suivants ($\gamma_{\mathcal{E}^{0+}(x)} =$*

$(\eta + e)/2$)

$$\begin{aligned} & \max_{\eta\in[-1,+1]^{\mathcal{E}^{0+}(x)}} \min_{\zeta\in[-1,+1]^{\mathcal{E}^{-}(x)}} \frac{1}{2}\|g(\eta,\zeta)\|^2 \\ \Leftrightarrow & \max_{\eta\in[-1,+1]^{\mathcal{E}^{0+}(x)}} \frac{1}{2}\mathrm{dist}(\overline{x}-\overline{y}+X\eta, Y[-1,+1]^{\mathcal{E}^{-}(x)})^2 \\ \Leftrightarrow & \max_{\eta\in[-1,+1]^{\mathcal{E}^{0+}(x)}} \frac{1}{2}\|\overline{x}-\overline{y}+X\eta-\mathrm{P}_{Y[-1,+1]^{\mathcal{E}^{-}(x)}}(\overline{x}-\overline{y}+X\eta)\|^2 \end{aligned} \tag{6.18}$$

Des maxima locaux stricts impliquent également $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^{-}(x)}(\gamma_{\mathcal{E}^{0+}(x)})) \in \partial\theta(x)$. □

Rappelons que la distance à un ensemble convexe est une fonction convexe. De fait, le second problème maximise une fonction quadratique convexe sur un polyèdre, le maximum est donc atteint sur un sommet. Ce problème est plus exigeant que sa version précédente, où l'on veut juste trouver un $\gamma_{\mathcal{E}^{0+}(x)}$ tel que $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^{-}(x)}(\gamma_{\mathcal{E}^{0+}(x)})) \neq 0$, ce qui est cohérent puisque l'on a rajouté la requête $g \in \partial\theta(x)$ (en plus de $g \neq 0$).

Mentionnons quel est l'intérêt de cette proposition alambiquée. Les méthodes de Levenberg-Marquardt permettent de prouver que, sous des hypothèses raisonnables, les directions tangentes produites tendent vers zéro (théorème 6.2.4 dessous). Dans le cas lisse, cela est suffisant pour obtenir un point stationnaire. Dans le cas non lisse, c'est là que cette proposition intervient : si l'on peut s'assurer que $g_k \in \partial\theta(x_k)$ pour tout k et x est un point d'accumulation de $\{x_k\}$, par les propriétés de $\partial\cdot$ (voir définition 2.3.9 et [51, proposition 2.1.5b p. 29]), $0 \in \partial\theta(x)$.

6.1.5 Régularité de solutions

Cette section discute, dans le formalisme développé ci-dessus, de conditions de régularité pour s'assurer que des points x sont solutions du problème. Cela peut être vu comme un renforcement de la proposition 6.1.14, puisque les solutions sont des points Dini-stationnaires particuliers. Les propriétés sont possiblement adaptables pour des versions non linéaires, mais pour simplifier on considère le PCL

$$0 \leqslant x \perp Mx + q \geqslant 0 \tag{6.19}$$

où l'on utilise la notation suivante :

$$A := \mathcal{F}(x), \qquad E := \mathcal{E}(x) \qquad \text{et} \qquad I := \mathcal{G}(x), \tag{6.20}$$

où l'on a assumé que $F(x) \equiv x$ et $G(x) \equiv Mx + q$. Dans la suite, pour éviter la confusion avec l'ensemble d'indice I, la matrice identité est notée Id.

Démarrons par rappeler [58, proposition 5.8.4], qui identifie des conditions de régularité pour qu'un point θ-stationnaire x soit une solution du problème. Ces conditions s'écrivent,

en x, (rappelons que $M \in \mathbf{Q}$ signifie que le PCL a une solution pour tout q, voir définition 2.2.2) :

$$\begin{cases} M_{I,I} \text{ est inversible,} \\ \text{le complément de Schur } M_{E,E} - M_{E,I}M_{I,I}^{-1}M_{I,E} \in \mathbf{Q}. \end{cases} \tag{6.21}$$

Par (6.20), les ensembles I et E utilisés dépendent de x. Ces conditions sont satisfaites indépendamment de x quand $M \in \mathbf{P}$, puisqu'une $\mathbf{P}$-matrice a des mineurs principaux > 0 et ses compléments de Schur sont dans $\mathbf{P}$ [192], donc dans $\mathbf{Q}$.

Proposition 6.1.21 (stationnarité de $\boldsymbol{\theta}$ pour le PCL [58])**.** *Si x est un point stationnaire de θ vérifiant la condition de régularité* (6.21)*, alors x est une solution du PCL. En particulier, si $M \in \mathbf{P}$, l'unique point stationnaire de θ est la solution du PCL.* □

De fait, pour un PCL avec $\mathbf{P}$-matrice, "solution" et "point stationnaire" de θ sont deux concepts identiques.

Regardons désormais comment relier la stationnarité forte de points ou le fait d'être une solution en termes des modèles φ_x de (6.10), selon les choix de $\Gamma \in [0, \mathrm{Id}]$. Avec les notations (6.20), l'annulation du gradient du modèle quadratique par morceau φ_x dans (6.10) s'exprime comme suit (en utilisant $x_E = (Mx+q)_E$)

$$I_{A,:}^{\mathsf{T}} x_A + M_{I,:}^{\mathsf{T}}(Mx+q)_I + I_{E,:}^{\mathsf{T}}\Gamma x_E + M_{E,:}^{\mathsf{T}}\overline{\Gamma}(Mx+q)_E = 0. \tag{6.22}$$

Comme évoqué au début de cette section, il faut renforcer les hypothèses pour traiter les solutions, qui sont des points stationnaires particuliers. Deux conditions de régularité sont proposées. La première est comme suit (rappelons la définition 2.2.3 : $M \in \mathbf{ND}$, si ses sous-matrices principales sont toutes inversibles) :

$$\begin{cases} M_{I,I} \text{ est inversible,} \\ \text{le complément de Schur } M_{E,E} - M_{E,I}M_{I,I}^{-1}M_{I,E} \in \mathbf{ND}. \end{cases} \tag{6.23}$$

Ces conditions sont satisfaites quand $M \in \mathbf{P}$ pour des raisons similaires à celles citées pour dire que (6.21) est vérifiée pour une $\mathbf{P}$-matrice.

L'hypothèse de régularité (6.23) utilisée dans dans la proposition suivante 6.1.14 est proche, mais plus faible, que la notions de *R-regularity* de Facchinei et Soares [87, définition 2.1]. Elle est définie à une solution du PCN et s'écrit, dans notre cas :

$$\begin{cases} M_{I,I} \text{ est inversible,} \\ \text{le complément de Schur } M_{E,E} - M_{E,I}M_{I,I}^{-1}M_{I,E} \in \mathbf{P}. \end{cases} \tag{6.24}$$

Cette dernière notion est appelée *régularité forte* dans [58, définition 5.8.3] et est reliée à la régularité forte de Robinson [220] (voir aussi [196]). De même, (6.24) est satisfaite pour tout x si $M \in \mathbf{P}$. Dans la suite, on utilise beaucoup $\Gamma = \mathrm{Diag}(\gamma) \in [0, \mathrm{Id}_{\mathcal{E}(x)}]$ qui intervient dans de multiples calculs.

Proposition 6.1.22 (CNS de $\boldsymbol{\theta}$-stationnarité pour PCL - I). *Pour $x \in \mathbb{R}^n$, considérons les propriétés suivantes :*

(i) *x est une solution du PCL* (6.19),
(ii) *pour tout $\Gamma \in [0, Id]$, on a* (6.22),
(iii) *pour un $\Gamma \in \mathrm{ext}[0, Id]$, on a* (6.22).

Alors, $(i) \Rightarrow (ii) \Rightarrow (iii)$. Si de plus (6.23) *est vérifiée, alors on a aussi $(iii) \Rightarrow (i)$.*

Preuve. $[(i) \Rightarrow (ii)]$ Puisque x est une solution, $F_{\mathcal{F}(x)}(x) = x_{\mathcal{F}(x)} = x_A = 0$, $G_{\mathcal{G}(x)}(x) = (Mx+q)_{\mathcal{G}(x)} = (Mx+q)_I = 0$ et $H_{\mathcal{E}(x)} = x_E = 0 = (Mx+q)_E$, donc $g = 0$ quel que soit Γ.

$[(ii) \Rightarrow (iii)]$ Clair.

$[(iii) \Rightarrow (i)]$ Si (6.23) est vérifiée en x, le système (6.22) s'écrit

$$\begin{cases} x_A + M_{I,A}^{\mathsf{T}}(Mx+q)_I + M_{E,A}^{\mathsf{T}}(\mathrm{Id}-\Gamma)(Mx+q)_E = 0 \\ M_{I,E}^{\mathsf{T}}(Mx+q)_I + \Gamma x_E + M_{E,E}^{\mathsf{T}}(\mathrm{Id}-\Gamma)(Mx+q)_E = 0 \\ M_{I,I}^{\mathsf{T}}(Mx+q)_I + M_{E,I}^{\mathsf{T}}(\mathrm{Id}-\Gamma)(Mx+q)_E = 0. \end{cases} \tag{6.25a}$$

Puisque $M_{I,I}$ est inversible, la dernière équation donne

$$(Mx+q)_I = -M_{I,I}^{-\mathsf{T}} M_{E,I}^{\mathsf{T}}(\mathrm{Id}-\Gamma)(Mx+q)_E. \tag{6.25b}$$

Après substitution dans la seconde équation de (6.25a) et l'usage de $(Mx+q)_E = x_E$, on obtient

$$\left(M_{E,E} - M_{E,I}M_{I,I}^{-1}M_{I,E}\right)^{\mathsf{T}}(\mathrm{Id}-\Gamma)x_E + \Gamma x_E = 0. \tag{6.25c}$$

Puisque $\Gamma \in \mathrm{ext}[0, \mathrm{Id}]$, on a $\Gamma^2 = \Gamma$, donc $(\mathrm{Id}-\Gamma)\Gamma = 0$ et, après multiplication à gauche par $(\mathrm{Id}-\Gamma)$, l'équation précédente devient

$$\Big((\mathrm{Id}-\Gamma)\left(M_{E,E} - M_{E,I}M_{I,I}^{-1}M_{I,E}\right)(\mathrm{Id}-\Gamma)\Big)^{\mathsf{T}} x_E = 0.$$

Par (6.23) et $\Gamma \in \mathrm{ext}[0, \mathrm{Id}]$, la sous-matrice principale $(\mathrm{Id}-\Gamma)(M_{E,E} - M_{E,I}M_{I,I}^{-1}M_{I,E})(\mathrm{Id}-\Gamma)$ de $M_{E,E} - M_{E,I}M_{I,I}^{-1}M_{I,E}$ est inversible, ce qui implique que

$$(\mathrm{Id}-\Gamma)x_E = 0.$$

De là, par (6.25b), (6.25c) et la première équation de (6.25a), on a

$$(Mx+q)_I = 0, \qquad \Gamma x_E = 0 \qquad \text{et} \qquad x_A = 0.$$

Pour montrer que x est une solution du (6.19), il faut montrer que $x_I \geqslant 0$ et $(Mx+q)_A \geqslant 0$. Cela peut s'obtient par les définitions de A and I, puisque l'on a

$$0 = x_A < (Mx+q)_A, \qquad \text{et} \qquad 0 = (Mx+q)_I < x_I. \qquad \square$$

Curieusement, on voit que les propositions 6.1.14 et 6.1.22 donnent des résultats reliés mais différents (indifféremment du fait que l'un considère le PCN et l'autre le PCL) :

- la différence principale est que le point (iii) de la proposition 6.1.14 requiert une condition pour *chaque* $\Gamma_{\mathcal{E}^{0+}(x)}$ dans $[0, \mathrm{Id}_{\mathcal{E}^{0+}(x)}]$, alors que la propriété (iii) de la proposition 6.1.22 requiert une condition pour *un seul* $\Gamma \in \mathrm{ext}[0, \mathrm{Id}]$,
- proposition 6.1.14 n'utilise pas de condition de régularité, mais seul $\Gamma_{\mathcal{E}^{0+}(x)}$ peut être choisi arbitrairement dans $[0, \mathrm{Id}_{\mathcal{E}^{0+}(x)}]$, alors que $\Gamma_{\mathcal{E}^{-}(x)}$ est déterminé par 6.1.12.
- proposition 6.1.14 traite de la forte stationnarité, qui est en général différente, sans hypothèse supplémentaire, du fait d'être solution.

Le prochain résultat renforce la proposition 6.1.22, dans la mesure où Γ est maintenant arbitraire dans $[0, \mathrm{Id}]$, mais il faut la condition de régularité plus forte (6.24) au lieu de (6.23).

Proposition 6.1.23 (CNS de $\boldsymbol{\theta}$-stationnarité pour PCL - II)**.** *Pour $x \in \mathbb{R}^n$, considérons les propriétés suivantes :*

(i) *x est une solution du PCL* (6.19)*,*
(ii) *pour tout $\Gamma \in [0, Id]$, on a* (6.22)*,*
(iii) *pour un $\Gamma \in [0, Id]$, on a* (6.22)*.*

Alors, $(i) \Rightarrow (ii) \Rightarrow (iii)$*. Si de plus* (6.24) *est vérifiée, alors on a aussi* $(iii) \Rightarrow (i)$*.*

Preuve. La preuve de l'implication $(iii) \Rightarrow (i)$ est identique à celle de la proposition 6.1.22, jusqu'à (6.25c), que l'on récrit ici en utilisant $(Mx + q)_E = x_E$ (une conséquence de la définition de E) :

$$\Big((\mathrm{Id} - \Gamma)\left(M_{E,E} - M_{E,I}M_{I,I}^{-1}M_{I,E}\right) + \Gamma\Big)^{\mathsf{T}} x_E = 0.$$

Par (6.24), $M_{E,E} - M_{E,I}M_{I,I}^{-1}M_{I,E} \in \mathbf{P}$. Maintenant, toute combinaison convexe de lignes d'une $\mathbf{P}$-matrice et de Id est inversible [3, lemme 2.1]. De fait, la matrice du système précédent est inversible. Cela fait que $x_E = 0$. Le reste de la preuve est similaire à la proposition 6.1.22. □

La condition de R-régularité (6.24) dépend du point considéré mais est uniforme pour tout $x \in \mathbb{R}^n$ si $M \in \mathbf{P}$.

Clarifions désormais un aspect de la proposition, en utilisant la contraposée de $(iii) \Rightarrow (i)$. Celle-ci nous indique que lorsque x n'est pas une solution du PCL (6.19), alors g (6.22) est non nul quel que soit $\Gamma \in [0, \mathrm{Id}]$. Insistons sur le fait que cela n'implique pas que, pour tout $\Gamma \in [0, \mathrm{Id}]$, $-g$ est une direction de descente de θ en x puisqu'un phénomène similaire à celui rencontré pour les fonctions convexes non lisses (g est un sous-gradient, mais $-g$ n'est pas une direction de descente [126, fin de § VIII.1.1]), voir l'exemple 6.1.4. Pour s'assurer que $-g$ est une direction de descente, on peut par exemple utiliser la définition 6.1.12.

6.2 Un algorithme envisagé

6.2.1 La méthode et ses propriétés

Cette section détaille les propriétés que l'on peut obtenir en globalisant l'algorithme NMP [72] régularisé par une approche Levenberg-Marquardt. Dans cette partie, on ne suppose pas $S = I$ ($S_k = I$), bien que S et S_k ne jouent pas un rôle central (dans la théorie). Toutes les propriétés précédemment citées restent valides (essentiellement, F' et G' sont multipliées à droite par $S^{-1/2}$ qui est bien définie).

Sans hypothèses de régularité, l'algorithme ne peut espérer mieux que de trouver un point (Clarke-)stationnaire de la fonction de mérite des moindres-carrés θ ; de fait, on s'arrête à de tels points. Si l'itéré courant x_k n'est pas stationnaire, l'algorithme détermine un modèle local de θ, de la forme (6.10), par exemple

$$\varphi_k : d \in \mathbb{R}^n \mapsto \varphi_k(d) := \frac{1}{2}\left(\|w_k(d)\|_2^2 + \lambda\, d^{\mathsf{T}} S_k d\right).$$

On utilise w_k pour la fonction $w(x_k, \cdot)$ qui associe à $d \in \mathbb{R}^n$ le vecteur $w_k(d) \in \mathbb{R}^{n+|\mathcal{E}(x_k)|}$, dont l'expression est celle de (6.9) avec $x \equiv x_k$

$$w_k(d) = \begin{bmatrix} F_{\mathcal{F}(x_k)}(x_k) + F'_{\mathcal{F}(x_k)}(x_k)d \\ G_{\mathcal{G}(x_k)}(x_k) + G'_{\mathcal{G}(x_k)}(x_k)d \\ (\Gamma_{\mathcal{E}^{0+}(x_k)})^{1/2}[H_{\mathcal{E}^{0+}(x_k)}(x_k) + F'_{\mathcal{E}^{0+}(x_k)}(x_k)d] \\ (\overline{\Gamma}_{\mathcal{E}^{0+}(x_k)})^{1/2}[H_{\mathcal{E}^{0+}(x_k)}(x_k) + G'_{\mathcal{E}^{0+}(x_k)}(x_k)d] \\ -(\Gamma_{\mathcal{E}^{-}(x_k)})^{1/2}[H_{\mathcal{E}^{-}(x_k)}(x_k) + F'_{\mathcal{E}^{-}(x_k)}(x_k)d]^- \\ -(\overline{\Gamma}_{\mathcal{E}^{-}(x_k)})^{1/2}[H_{\mathcal{E}^{-}(x_k)}(x_k) + G'_{\mathcal{E}^{-}(x_k)}(x_k)d]^- \end{bmatrix}$$

Notons que, puisque $F_i(x_k) = G_i(x_k)$ pour $i \in \mathcal{E}(x_k)$ et $\gamma_k + \overline{\gamma}_k = 1$, on a

$$\varphi_k(0) = \theta(x_k). \tag{6.27}$$

Ce modèle requiert de définir le vecteur de poids $\gamma_k \in [0,1]^{|\mathcal{E}(x_k)|}$ et c'est fait de façon à ce que le gradient de $||w_k||^2/2$ en zéro ne s'annule pas (comme on l'a vu dans la section 6.1.4, c'est possible lorsque x_k n'est pas un point fortement stationnaire de θ). Le modèle requiert également une matrice $S_k \succ 0$ et un paramètre λ_k connus au début de l'itération ; des précisions sont données dans le théorème 6.2.4 plus bas.

Rappelons que φ_k est différentiable, donc un minimiseur d_k vérifie $\nabla\varphi_k(d_k) = 0$ avec $\lambda = \lambda_k$, ou

$$\begin{aligned} &\Big(F'_{\mathcal{F}(x_k)}(x_k)^{\mathsf{T}} F'_{\mathcal{F}(x_k)}(x_k) + G'_{\mathcal{G}(x_k)}(x_k)^{\mathsf{T}} G'_{\mathcal{G}(x_k)}(x_k) \\ &+ F'_{\mathcal{E}^{0+}(x_k)}(x_k)^{\mathsf{T}} (\Gamma_k)_{\mathcal{E}^{0+}(x_k)} F'_{\mathcal{E}^{0+}(x_k)}(x_k) + G'_{\mathcal{E}^{0+}(x_k)}(x_k)^{\mathsf{T}} (\overline{\Gamma}_k)_{\mathcal{E}^{0+}(x_k)} G'_{\mathcal{E}^{0+}(x_k)}(x_k)\Big) d_k \\ &- F'_{\mathcal{E}^{-}(x_k)}(x_k)^{\mathsf{T}} (\Gamma_k)_{\mathcal{E}^{-}(x_k)} \left(F'_{\mathcal{E}^{-}(x_k)}(x_k) d_k\right)^- \\ &- G'_{\mathcal{E}^{-}(x_k)}(x_k)^{\mathsf{T}} (\overline{\Gamma}_k)_{\mathcal{E}^{-}(x_k)} \left(G'_{\mathcal{E}^{-}(x_k)}(x_k) d_k\right)^- + \lambda_k S_k d_k = -g_k, \end{aligned} \tag{6.28}$$

où $\Gamma_k := \mathrm{Diag}(\gamma_k)$, $\overline{\Gamma}_k = \mathrm{Id} - \Gamma_k$ et $g_k := \nabla\varphi_k(0)$ est donné par (voir (6.11) dans $x \equiv x_k$) :

$$\begin{aligned} g_k &= F'_{\mathcal{F}(x_k)}(x_k)^{\mathsf{T}} F_{\mathcal{F}(x_k)}(x_k) + G'_{\mathcal{G}(x_k)}(x_k)^{\mathsf{T}} G_{\mathcal{G}(x_k)}(x_k) \\ &\quad + [F'_{\mathcal{E}(x_k)}(x_k)^{\mathsf{T}}\Gamma_k + G'_{\mathcal{E}(x_k)}(x_k)^{\mathsf{T}}\overline{\Gamma}_k] H_{\mathcal{E}(x_k)}(x_k). \end{aligned} \tag{6.29}$$

Une fois les poids choisis, avec le paramètre de pénalisation $\lambda_k \geqslant 0$, l'algorithme minimise le modèle φ_k

$$\min_{d\in\mathbb{R}^n} \varphi_k(d). \tag{6.30}$$

La solution calculée dépend de λ_k et est notée $d_k(\lambda_k)$ (uniquement déterminée si $\lambda_k > 0$). La pénalisation λ est (éventuellement) modifiée pour avoir une décroissance suffisante de θ, au sens de [13]

$$\theta(x_k + d_k(\lambda_k)) \leqslant \theta(x_k) + \eta_1\, g_k^{\mathsf{T}} d_k(\lambda_k), \tag{6.31}$$

Lorsque cette inégalité est satisfaite, l'itéré suivant est obtenu par $x_{k+1} := x_k + d_k(\lambda_k)$.

On donne désormais une description formelle et précise de l'algorithme LM proposé pour résoudre (6.1), sur lequel les résultats ci-dessous sont basés. Dans la description de l'algorithme, "constant" signifie qu'une quantité ne dépend pas de l'itération.

Algorithme 6.2.1 (algorithme de type LM)**.** L'algorithme utilise les constantes suivantes : $0 < \sigma_1 < 1 < \sigma_2$ pour la mise à jour de λ_k et $0 < \eta_1 < \eta_2 < 1$ comme seuils de satisfaction pour la décroissance de θ. On suppose que $S_k \succ 0$. Une itération de l'algorithme, calculant $(x_{k+1}, \lambda_{k+1}, S_{k+1})$ à partir de (x_k, λ_k, S_k), procède comme suit.

1. *Critère d'arrêt et poids.* Si x_k est un point stationnaire de θ, arrêt. Sinon, fixer $\Gamma_k \in [0, \mathrm{Id}]$ tel que g_k donné par (6.29) est non nul.
2. *Déplacement.* Fixer $\lambda_{k,0} := \lambda_k$ et répéter les opérations suivantes, indicées par $i \in \mathbb{N}$, jusqu'à ce satisfaire (6.31).
   2.1. Calculer une solution $d_{k,i}$ à (6.30).
   2.2. Si
   $$\theta(x_k + d_{k,i}) \leqslant \theta(x_k) + \eta_1\, g_k^{\mathsf{T}} d_{k,i}, \tag{6.32}$$
   est vérifiée, sortir de la boucle avec $d_k := d_{k,i}$ (puis aller à 3.), sinon $\lambda_{k,i+1} = \sigma_2\lambda_{k,i}$ et retourner à 2.1.
3. *Nouvelle pénalisation.* Si
   $$\theta(x_k + g_k) \leqslant \theta(x_k) + \eta_2\, g_k^{\mathsf{T}} d_k, \tag{6.33}$$
   alors $\lambda_k := \sigma_1\lambda_{k,i}$, sinon $\lambda_k := \lambda_{k,i}$.
4. *Nouvel itéré.* Mettre à jour $x_{k+1} := x_k + d_k$.

13. On peut utiliser $\theta(x_k + d_k(\lambda_k)) \leqslant \theta(x_k) + \eta_1\, \theta'(x_k; d_k(\lambda_k))$ sans grande différence.

5. *Nouvelle matrice.* Choisir $S_{k+1} \succ 0$.

Remarques 6.2.2. 1) On suppose que le test de stationnarité peut être résolu à chaque itération ; cela peut être fait par énumération si $|\mathcal{E}(x_k)|$ est petit ou par exemple sous les hypothèses de la proposition 6.1.18.
2) Le coût de calcul de l'algorithme est principalement lié aux problèmes d'optimisation (6.30) à résoudre, possiblement plusieurs par itération.

Un algorithme avec des sous-problèmes quadratiques à chaque itération peut par exemple être trouvé dans [86, §9.2, algorithme 9.2.2]. De même dans [17], où une fonction différentiable fortement convexe quadratique par morceau est minimisée sans contraintes, et pour laquelle une méthode de Newton semi-lisse avec recherche linéaire exacte est utilisée. Le point 4 de la proposition suivante a pour conséquence que la boucle de l'étape 2 de l'algorithme 6.2.1 est traitée un nombre fini de fois par itération. La proposition suivante adapte les propriétés usuelles de Levenberg-Marquardt à notre situation.

Proposition 6.2.3 (décroissance suffisante)**.** *Supposons que le gradient donné par* (6.29) $g_k \neq 0$*, que* $S_k \succ 0$ *et que* $\eta_1 < 1$*. Définissons* $d_k(\lambda)$ *comme une solution de* (6.30)*. Alors,*

1) $d_k(\lambda) \neq 0$ *pour tout* $\lambda \geqslant 0$*,*
2) $d_k(\lambda) \to 0$ *quand* $\lambda \to +\infty$*,*
3) $d_k(\lambda)/\|d_k(\lambda)\| \to -S_k^{-1}g_k/\|S_k^{-1}g_k\|$ *quand* $\lambda \to \infty$*,*
4) *la condition de décroissance suffisante* (6.31) *est vérifiée pour* λ *assez grand.*

Preuve. 1) On procède par contradiction. Si $d_k(\lambda) = 0$, on aurait $\nabla\varphi_k(0) = 0$, puisque $d_k(\lambda)$ résout le problème (6.30) et φ_k est différentiable. De façon équivalente par définition de g_k dans (6.29), on aurait $g_k = 0$, qui contredit l'hypothèse. De fait, $d_k(\lambda) \neq 0$.

2) La convergence de $d_k(\lambda)$ vers zéro résulte du fait que $d_k(\lambda)$ minimise $\varphi_k(\cdot)$, et donc

$$0 \leqslant \frac{\lambda}{2}\, d_k(\lambda)^{\mathsf{T}} S_k d_k(\lambda) \leqslant \varphi_k(d_k(\lambda)) \leqslant \varphi_k(0) = \theta(x_k),$$

par (6.27). En divisant par λ et en prenant la limite quand $\lambda \to +\infty$, on a $d_k(\lambda)^{\mathsf{T}} S_k d_k(\lambda) \to 0$. Puisque $S_k \succ 0$, on a $d_k(\lambda) \to 0$.

3) Puisque $d_k(\lambda)$ minimise $\varphi_k(\cdot)$ et φ_k est différentiable, on a $\nabla\varphi_k(d_k(\lambda)) = 0$. Ensuite, en prenant la limite pour $\lambda \to +\infty$ dans l'équation $\nabla\varphi_k(d_k(\lambda)) = 0$ et en utilisant le point 2, il vient

$$\lambda S_k d_k(\lambda) \to -g_k \qquad \text{ou équivalemment} \qquad \lambda\, d_k(\lambda) \to -S_k^{-1} g_k,$$

où g_k est défini par (6.29). Maintenant, puisque $\lambda\, d_k(\lambda) \neq 0$ par le point 1, on obtient le point 3.

4) On procède par contradiction. Supposons que (6.31) n'est pas vérifiée pour une suite $\lambda \to +\infty$. Ensuite, pour ces $\lambda \to +\infty$, on a

$$\frac{\theta(x_k + d_k(\lambda)) - \theta(x_k) - g_k^T d_k(\lambda)}{\|d_k(\lambda)\|} > (1 - \eta_1) \frac{-g_k^T d_k(\lambda)}{\|d_k(\lambda)\|}.$$

Par le point 3, le membre de droite tend vers $(1-\eta_1) g_k^T S_k^{-1} g_k / \|S_k^{-1} g_k\|$, qui est > 0 puisque $S_k \succ 0$, $g_k \neq 0$ et $\eta_1 < 1$. De plus, montrons que le terme de gauche tend vers un nombre négatif, ce qui justifiera la contradiction.

Rappelons que θ est dérivable directionnellement. De plus, θ est aussi lipschitzienne, donc $\theta'(x;\cdot)$ l'est aussi et θ est dérivable directionnellement au sens de Hadamard (voir par exemple [230] et ses références), qui indique que

$$\theta'(x;d) = \lim_{\substack{t \downarrow 0 \\ d' \to d}} \frac{\theta(x + td') - \theta(x)}{t}.$$

De fait, en choisissant $t := \|d_k(\lambda)\|$, avec $t \to 0$ quand $\lambda \to \infty$, et $d' := d_k(\lambda)/\|d_k(\lambda)\|$, qui tend vers $-S_k^{-1} g_k / \|S_k^{-1} g_k\|$ par le point 3, on déduit par la formule précédente que

$$\lim_{\lambda \to +\infty} \frac{\theta(x_k + d_k(\lambda)) - \theta(x_k)}{\|d_k(\lambda)\|} = \theta'(x_k; -S_k^{-1} g_k / \|S_k^{-1} g_k\|).$$

De fait,

$$\lim_{\lambda \to \infty} \frac{\theta(x_k + d_k(\lambda)) - \theta(x_k) - g_k^T d_k(\lambda)}{\|d_k(\lambda)\|} = \frac{\theta'(x_k; -S_k^{-1} g_k) + g_k S_k^{-1} g_k}{\|S_k^{-1} g_k\|}.$$

Par la définition 6.1.12, le terme de droite est négatif. □

Le prochain résultat requiert la bornitude de certaines quantités générées par l'algorithme, le produit λS, ainsi que F' et G', l'hypothèse sur F et G étant plutôt inoffensive. Il suppose que l'algorithme génère une suite, donc qu'il ne s'arrête pas à un point stationnaire (ou presque).

Théorème 6.2.4 (convergence de l'algorithme LM). *Soit $\{(x_k, \lambda_k)\}$ une séquence générée par l'algorithme* 6.2.1. *Définissons g_k via* (6.29). *Alors,*

1) $\{\theta(x_k)\}$ *converge,*
2) *pour toute sous-suite $\mathcal{K}$ de $\mathbb{N}$ telle que $\{(F'(x_k), G'(x_k), \lambda_k S_k)\}_{k \in \mathcal{K}}$ est bornée, on a $\{g_k\}_{k \in \mathcal{K}} \to 0$ quand $k \to +\infty$ dans $\mathcal{K}$.*

Preuve. 1) La convergence de la suite $\{\theta(x_k)\}$ provient de sa décroissance et de sa positivité.

2) Grâce à la condition de décroissance suffisante (6.31) et à la convergence de $\theta(x_k)$, il vient

$$g_k^T d_k \to 0, \qquad \text{quand } k \to \infty. \tag{6.34}$$

Désormais, il faut montrer que la convergence dans (6.34) est bien liée à un gradient g_k qui tend vers zéro et non pas à la convergence des d_k vers zéro (bien qu'également vrai). En prenant le produit scalaire de (6.28) avec d_k, il vient

$$\begin{aligned} -g_k^{\mathsf{T}} d_k &= \|F'_{\mathcal{F}(x_k)}(x_k)d_k\|^2 + \|G'_{\mathcal{G}(x_k)}(x_k)d_k\|^2 \\ &\quad + \|(\Gamma_k)^{1/2}_{\mathcal{E}^{0+}(x_k)} F'_{\mathcal{E}^{0+}(x_k)}(x_k)d_k\|^2 + \|(\overline{\Gamma}_k)^{1/2}_{\mathcal{E}^{0+}(x_k)} G'_{\mathcal{E}^{0+}(x_k)}(x_k)d_k\|^2 \\ &\quad + \|(\Gamma_k)^{1/2}_{\mathcal{E}^{-}(x_k)} [F'_{\mathcal{E}^{-}(x_k)}(x_k)d_k]^-\|^2 + \|(\overline{\Gamma}_k)^{1/2}_{\mathcal{E}^{-}(x_k)} [G'_{\mathcal{E}^{-}(x_k)}(x_k)d_k]^-\|^2 \\ &\quad + \lambda_k S_k d_k. \end{aligned}$$

On déduit ensuite de (6.34) et du fait que $\lambda_k > 0$ quand $k \to \infty$:

$$\begin{aligned} F'_{\mathcal{F}(x_k)}(x_k)d_k \to 0, \quad & G'_{\mathcal{G}(x_k)}(x_k)d_k \to 0, \\ (\Gamma_k)^{1/2}_{\mathcal{E}^{0+}(x_k)} F'_{\mathcal{E}^{0+}(x_k)}(x_k)d_k \to 0, \quad & (\overline{\Gamma}_k)^{1/2}_{\mathcal{E}^{0+}(x_k)} G'_{\mathcal{E}^{0+}(x_k)}(x_k)d_k \to 0, \\ (\Gamma_k)^{1/2}_{\mathcal{E}^{-}(x_k)} [F'_{\mathcal{E}^{-}(x_k)}(x_k)d_k]^- \to 0, \quad & (\overline{\Gamma}_k)^{1/2}_{\mathcal{E}^{-}(x_k)} [G'_{\mathcal{E}^{-}(x_k)}(x_k)d_k]^- \to 0, \\ \lambda_k^{1/2} S_k^{1/2} d_k \to 0. & \end{aligned}$$

Maintenant, $\{\Gamma_k, \overline{\Gamma}_k\}$ étant bornée, on voit que, si la suite $\{(F'(x_k), G'(x_k), \lambda_k S_k)\}_{k\in\mathcal{K}}$ est bornée, on a, quand $k \to \infty$ in $\mathcal{K}$:

$$\begin{aligned} F'_{\mathcal{F}(x_k)}(x_k)^{\mathsf{T}} F'_{\mathcal{F}(x_k)}(x_k)d_k &\to 0, \\ G'_{\mathcal{G}(x_k)}(x_k)^{\mathsf{T}} G'_{\mathcal{G}(x_k)}(x_k)d_k &\to 0, \\ F'_{\mathcal{E}^{0+}(x_k)}(x_k)^{\mathsf{T}} (\Gamma_k)_{\mathcal{E}^{0+}(x_k)} F'_{\mathcal{E}^{0+}(x_k)}(x_k)d_k &\to 0, \\ G'_{\mathcal{E}^{0+}(x_k)}(x_k)^{\mathsf{T}} (\overline{\Gamma}_k)_{\mathcal{E}^{0+}(x_k)} G'_{\mathcal{E}^{0+}(x_k)}(x_k)d_k &\to 0, \\ F'_{\mathcal{E}^{-}(x_k)}(x_k)^{\mathsf{T}} (\Gamma_k)_{\mathcal{E}^{-}(x_k)} [F'_{\mathcal{E}^{-}(x_k)}(x_k)d_k]^- &\to 0, \\ G'_{\mathcal{E}^{-}(x_k)}(x_k)^{\mathsf{T}} (\overline{\Gamma}_k)_{\mathcal{E}^{-}(x_k)} [G'_{\mathcal{E}^{-}(x_k)}(x_k)d_k]^- &\to 0, \\ \lambda_k S_k d_k &\to 0. \end{aligned}$$

La définition (6.28) de l'itération implique alors que $g_k \to 0$ quand $k \to \infty$ dans $\mathcal{K}$. □

Observons que sans rien de plus, l'algorithme peut s'arrêter à des points qui ne sont pas nécessairement satisfaisants. L'hypothèse sur la bornitude de λ est relativement forte, elle suppose que la partie Levenberg-Maquardt ne fait pas des déplacements trop petits. Avoir une version plus forte du théorème sans une telle hypothèse est une perspective intéressante.

Contre-exemple 6.2.5 (point d'accumulation pas Dini-stationnaire)**.** Prenons la figure 6.4 et le problème associé, où $F(x) \equiv x$, $G(x) \equiv 1 + (x-1)^2$. Si λS est assez grand, i.e., les déplacements sont petits, on peut justifier que l'algorithme n'atteint jamais l'ensemble $[0, 1]$. De fait, le formalisme développé avec les poids γ n'intervient jamais. □

Observons que, par la proposition 6.1.20, le point d'accumulation est Clarke stationnaire.

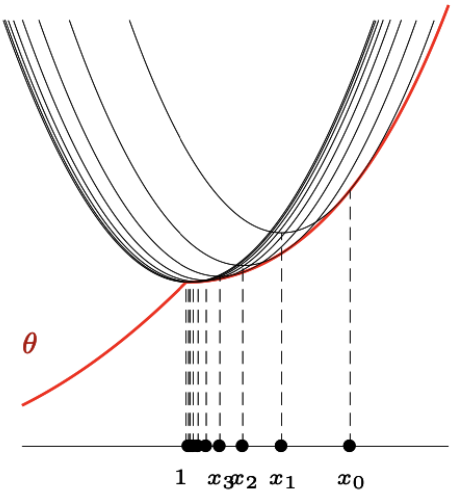


FIGURE 6.5 – Les courbes au-dessus de θ sont les modèles quadratiques φ_{x_k}. Bien qu'il puisse y avoir convergence rapide vers $x = 1$, le point peut ne jamais être atteint ($x_k > 1$) donc $\forall\, k, \mathcal{E}(x_k) = \varnothing$.

6.2.2 Modifications et améliorations potentielles

Maintenant que nous avons vu les propriétés de l'algorithme, nous évoquons des améliorations potentielles qui pourraient y être apportées dans des travaux futurs.

Choix des poids

Rappelons que malgré le fait que la détection de la stationnarité soit non polynomiale dans le cas général (propositions 6.1.14 et 6.1.15 avec le théorème 6.1.16), dans le cadre de ce chapitre il est possible que cela puisse souvent être résolu facilement, par exemple dans les cas suivants (voir l'annexe C pour les détails) :

- les lignes de $G'_{\mathcal{E}^-(x)}(x) - F'_{\mathcal{E}^-(x)}(x)$ sont indépendantes,
- $\mathcal{E}^-(x) = \varnothing$ (pas de projection),
- $\mathcal{E}^{0+}(x) = \varnothing$ (un point à projeter),
- $\mathcal{R}([(G'_{\mathcal{E}^{0+}(x)}(x) - F'_{\mathcal{E}^{0+}(x)}(x))^\mathsf{T}\, \overline{x} - \overline{y}]) \not\subseteq \mathcal{R}((G'_{\mathcal{E}^{0+}(x)}(x) - F'_{\mathcal{E}^{0+}(x)}(x))^\mathsf{T})$ (brièvement, une dimension du premier zonotope n'est pas générée par le second donc n'est pas contenu dedans).

Tous ces cas sont vérifiables et traitables aisément. Dans le cas général, surtout "loin" de solutions, i.e., au début de l'algorithme, on peut imaginer ajouter de l'aléatoire pour le choix des poids (comme fait dans [19]). Par exemple, tester quelques valeurs quelconques pour voir si on peut facilement obtenir une direction de descente. On peut aussi imaginer ne modifier que les poids des indices i qui changent d'ensemble d'indices.

Alternatives au choix des poids

Pour éviter la machinerie lourde des itérations de Levenberg-Marquardt, une option est d'hybrider la méthode : comme décrit en section 2.3.3, d'utiliser, si ça convient, une direction plus simplement que par l'itération complète. Dans notre cas, on pourrait utiliser une expression plus simple du modèle φ_x de (6.10) en utilisant

$$w_k(d) = \begin{bmatrix} F_{\mathcal{F}(x_k)}(x_k) + F'_{\mathcal{F}(x_k)}(x_k)d \\ G_{\mathcal{G}(x_k)}(x_k) + G'_{\mathcal{G}(x_k)}(x_k)d \\ \Gamma^{1/2}[F_{\mathcal{E}(x_k)}(x_k) + F'_{\mathcal{E}(x_k)}(x_k)d] \\ \overline{\Gamma}^{1/2}[G_{\mathcal{E}(x_k)}(x_k) + G'_{\mathcal{E}(x_k)}(x_k)d] \end{bmatrix}.$$

On pourrait aussi utiliser la direction de Newton-min (rappelons que cela revient à un choix de $\gamma \in \{0,1\}^{\mathcal{E}(x)}$), qui a un coût faible. Plusieurs approches dans la section 2.3.3 utilisent cette astuce efficacement.

Dans les algorithmes, cela signifie remplacer l'étape 3 de l'algorithme PNM 6.1.2 par l'étape suivante.

Algorithme 6.2.6 (Étape hybride pour PNM [72, algorithm 3.8])**.** 3.1 Pour une partition $\tilde{\mathcal{F}}(x), \tilde{\mathcal{G}}(x)$ de $[1:n]$ satisfaisant $\tilde{\mathcal{F}}(x) \supseteq \mathcal{F}(x)$ et $\tilde{\mathcal{G}}(x) \supseteq \mathcal{G}(x)$, calculer une direction de Newton-min d^{NM} comme solution de

$$\begin{cases} F_i(x) + F'_i(x)d = 0 & \text{si } i \in \tilde{\mathcal{F}}(x), \\ G_i(x) + G'_i(x)d = 0 & \text{si } i \in \tilde{\mathcal{G}}(x). \end{cases}$$

3.2 *Test de décroissance.* Si (6.8) est vérifiée avec d et $\alpha = 1$, aller à l'étape 5. Sinon, calculer d comme indiqué par (6.7).

De même, on remplace les étapes 1 et 2 de l'algorithme LM 6.2.1 par l'étape suivante.

Algorithme 6.2.7 (Étape hybride pour LM)**.** 1.1 Pour une partition $\tilde{\mathcal{F}}(x), \tilde{\mathcal{G}}(x)$ de $[1:n]$ satisfaisant $\tilde{\mathcal{F}}(x) \supseteq \mathcal{F}(x)$ et $\tilde{\mathcal{G}}(x) \supseteq \mathcal{G}(x)$, calculer une direction de Newton-min d^{NM} comme solution de

$$\begin{cases} F_i(x) + F'_i(x)d = 0 & \text{si } i \in \tilde{\mathcal{F}}(x), \\ G_i(x) + G'_i(x)d = 0 & \text{si } i \in \tilde{\mathcal{G}}(x). \end{cases}$$

1.2 *Test de décroissance.* Si (6.32) est vérifiée avec d, aller à l'étape 4. Sinon, calculer d comme donné par les étapes 1 et 2 de l'algorithme 6.2.1.

Tolérance et précision numérique

Comme discuté au contre-exemple 6.2.5, les "plis concaves" que θ peut avoir sont des difficultés que l'algorithme ne peut traiter tel quel. De plus, surtout numériquement, définir

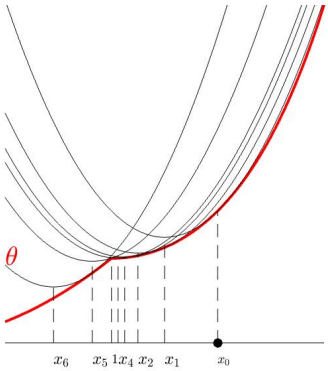


Figure 6.6 – Illustration de quelques itérés, avec $\tau > 0$ et $\mathcal{E}(x) := \{i \in [1:n] : |F_i(x) - G_i(x)| < \tau\}$, de l'algorithme 6.2.1.

$\mathcal{E}$ par $F_i(x) = G_i(x)$ est insatisfaisant. Cela justifie le besoin d'introduire une tolérance $\tau > 0$ afin de définir ces indices par $|F_i(x) - G_i(x)| < \tau$. Cette tolérance est déjà présente dans [72].

Avec une telle tolérance, le contre-exemple 6.2.5 ne risque pas de poser problème : pour k assez large, x_k sera assez proche de 1, signifiant qu'au lieu que l'ensemble soit $\mathcal{G}$, on aura $\mathcal{E}(x)^\tau$: l'autre morceau (correspondant à $x < 1$) sera considéré et, pour γ pas trop proche de 0, $\gamma F'(x_k) + (1-\gamma)G'(x_k) = \gamma + (1-\gamma) \times 2(x_k - 1)$ est assez grand pour "surmonter" le pli à 1, comme schématisé sur l'image suivante.

Pour terminer, les travaux en cours autour de cet algorithme bénéficieraient clairement d'affinage et d'astuces, certaines discutées dans le chapitre 2. En particulier, la question générale du choix des poids γ et de leur impact semble être une piste intéressante de ce que ce chapitre a présenté.

Historique

Finissons par un bref commentaire sur le lemme 6.1.6. Initialement, nous avions une preuve assez laborieuse et capillotractée pour prouver l'existence des $\gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^{0+}(x)})$, mais pas vraiment constructive. La compréhension de la structure du minimum (chapitre 3) semble fournir une approche plus simple (et constructive).

Conclusion et perspectives

Dans cette thèse, on a discuté de plusieurs éléments entourant l'utilisation de la C-fonction minimum pour les problèmes de complémentarité. Des méthodes locales comme globales partant du système non lisse induit par cette C-fonction peuvent conduire à des considérations géométriques fortes, qui résultent de la structure intrinsèque du minimum.

Cela a été initié dans le chapitre 6, où la recherche d'une direction de descente (ou de validation de stationnarité) s'est révélée être un problème NP-complet, du fait de la nature combinatoire du minimum. L'algorithme associé, bien que s'assurant techniquement d'atteindre un point Clarke-stationnaire, reste à tester sur des problèmes classiques.

Pour justifier cela, la géométrie des zonotopes, des polytopes symétriques particuliers, a été essentielle. Cela est fortement relié au chapitre 3 ([77]), où on a montré que certains éléments du B-différentiel de la C-fonction minimum sont reliés aux sommets de zonotopes spécifiques et donc à des arrangements d'hyperplans.

Pour ce problème particulier, on a proposé des améliorations sur un algorithme de l'état de l'art, les plus promettantes étant celles basées sur des heuristiques intéressantes comme l'utilisation de la dualité, par les circuits du matroïde sous-jacent. Ces améliorations ont été utilisées aussi dans le chapitre 5, pour des arrangements généraux – pas nécessairement des arrangements centrés provenant de l'application du minimum sur un PC. L'emploi de ces heuristiques sur le calcul de l'arrangement complet ou leur entremêlement avec d'autres méthodes basées sur les symétries combinatoires, mènerait certainement à des observations intéressantes.

Annexe A

Informations détaillées sur les instances affines et les algorithmes

Cette annexe complète le chapitre 5, en précisant certaines preuves et en donnant des détails, comme dans la section 4.5, sur certaines valeurs numériques observées.

A.1 Détails sur des propriétés du chapitre 5

$$\mathcal{S}(V,-\tau) = -\mathcal{S}(V,\tau). \tag{5.9}$$

En effet, si $s \in \mathcal{S}(V,-\tau)$, $s \cdot (V^{\mathsf{T}}x + \tau) > 0$ pour un $x \in \mathbb{R}^n$. De fait, $-s \cdot (V^{\mathsf{T}}(-x) - \tau) > 0$, montrant que $-s \in \mathcal{S}(V,\tau)$ ou $s \in -\mathcal{S}(V,\tau)$. On a montré l'inclusion $\mathcal{S}(V,-\tau) \subseteq -\mathcal{S}(V,\tau)$. En changeant τ en $-\tau$, on a $\mathcal{S}(V,\tau) \subseteq -\mathcal{S}(V,-\tau) \subseteq \mathcal{S}(V,\tau)$, donc (5.9).

$$\mathcal{S}_s(V,-\tau) = -\mathcal{S}_s(V,\tau) = \mathcal{S}_s(V,\tau) \qquad \text{et} \qquad \mathcal{S}_a(V,-\tau) = -\mathcal{S}_a(V,\tau). \tag{5.11}$$

En effet,

$$\begin{aligned}
\mathcal{S}_s(V,-\tau) &= \mathcal{S}(V,-\tau) \cap \mathcal{S}(V,\tau) = \mathcal{S}_s(V,\tau),\\
-\mathcal{S}_s(V,\tau) &= [-\mathcal{S}(V,\tau)] \cap [-\mathcal{S}(V,-\tau)] = \mathcal{S}(V,-\tau) \cap \mathcal{S}(V,\tau) = \mathcal{S}_s(V,\tau),\\
\mathcal{S}_a(V,-\tau) &= \mathcal{S}(V,-\tau) \setminus \mathcal{S}_s(V,-\tau) = [-\mathcal{S}(V,\tau)] \setminus [-\mathcal{S}_s(V,\tau)] = -\mathcal{S}_a(V,\tau).
\end{aligned}$$

Proposition A.1.1 (connectivité de $\boldsymbol{\mathcal{S}}$)**.** *L'ensemble $\mathcal{S}(V,\tau)$ des vecteurs de signes d'un arrangement affine propre est connexe si et seulement si ses hyperplans sont tous distincts. Dans ce cas, deux éléments quelconques s et $\tilde{s}$ de $\mathcal{S}(V,\tau)$ peuvent être reliés par un chemin dans $\mathcal{S}(V,\tau)$ de longueur $l := \sum_{i\in[1:p]} |\tilde{s}_i - s_i|/2 \leqslant p$ et il n'existe pas de chemin dans $\mathcal{S}(V,\tau)$ reliant s et $\tilde{s}$ de longueur inférieure.*

Preuve. Le fait que tout chemin reliant s et $\tilde{s}$ dans $\mathcal{S} \equiv \mathcal{S}(V, \tau)$ soit de longueur $\geqslant l$ vient du fait que s et $\tilde{s}$ ont l composantes différentes et que deux vecteurs de signes adjacents diffèrent par une seule composante.

[$\Rightarrow$] Nous démontrons la contraposée. Supposons que, pour certains $i \neq j$ dans $[1 : p]$, les hyperplans H_i et H_j soient identiques. Alors, d'après la proposition 5.3.2(2), les paires non nulles (v_i, τ_i) et (v_j, τ_j) sont colinéaires dans $\mathbb{R}^n \times \mathbb{R} : (v_j, \tau_j) = \alpha(v_i, \tau_i)$, pour un certain $\alpha \in \mathbb{R}^*$. Supposons que $\alpha > 0$ (resp. $\alpha < 0$). Pour tout $\tilde{s} \in \mathcal{S}$, il existe un $\tilde{x} \in \mathbb{R}^n$ tel que $\tilde{s} \bullet (V^\mathsf{T}\tilde{x} - \tau) > 0$, ce qui implique que l'on doit avoir $\tilde{s}_i = \tilde{s}_j$ (resp. $\tilde{s}_i = -\tilde{s}_j$). Prenons un $s \in \mathcal{S}(V, 0)$ ($\neq \varnothing$ d'après (5.12)), de sorte que $-s \in \mathcal{S}(V, 0)$ par la symétrie de $\mathcal{S}(V, 0)$, indiquée par (5.7). Nous affirmons qu'il est impossible de trouver un chemin dans $\mathcal{S}$ reliant $s \in \mathcal{S}$ et $-s \in \mathcal{S}$. En effet, toutes les composantes de s doivent changer de signe. Or, les composantes i et j de tout vecteur de signe $\tilde{s}$ sur un tel chemin sont nécessairement égales (resp. opposées), de sorte qu'elles changeraient simultanément, alors que l'adjacence impose de changer un seul signe entre deux vecteurs de signes consécutifs d'un chemin.

[$\Leftarrow$] Soient s et $\tilde{s} \in \mathcal{S}$, supposés distincts (sinon le résultat est évident). Il faut montrer qu'il existe un chemin de longueur l dans $\mathcal{S}$ reliant s à $\tilde{s}$. Par définition de $\mathcal{S}$, on peut trouver x et $\tilde{x}' \in \mathbb{R}^n$ tels que

$$s \bullet (V^\mathsf{T}x - \tau) > 0 \quad \text{et} \quad \tilde{s} \bullet (V^\mathsf{T}\tilde{x}' - \tau) > 0.$$

Le chemin recherché dans $\mathcal{S}$ est déterminé par les vecteurs de signes des chambres, donnés par ϕ dans (5.6), qui sont traversés le long du segment reliant x à une petite modification $\tilde{x}$ de $\tilde{x}'$. La modification est introduite pour que $\tilde{x}$ et $\tilde{x}'$ appartiennent à la même chambre et que le segment les reliant ne croise pas deux hyperplans ou plus simultanément.

Voici comment la modification $\tilde{x}$ de $\tilde{x}'$ est obtenue. Soit $d' := \tilde{x}' - x$, qui est non nul, puisque $s \neq \tilde{s}$. Par la proposition 5.3.2(2), puisque les hyperplans de l'arrangement sont tous distincts, les vecteurs $\{(v_i, \tau_i) \in \mathbb{R}^n \times \mathbb{R} : i \in [1 : p]\}$ ne sont pas colinéaires, de sorte que les vecteurs $\{v_i/(v_i^\mathsf{T}x - \tau_i) \in \mathbb{R}^n : i \in [1 : p]\}$ sont distincts [1]. Ainsi, pour un d arbitrairement proche de d', donc pour un

$$\tilde{x} := x + d \tag{A.1a}$$

arbitrairement proche de $\tilde{x}'$, on peut garantir l'inégalité $\tilde{s} \bullet (V^\mathsf{T}\tilde{x} - \tau) > 0$ et, d'après le lemme 3.2.6,

$$|\{(v_i^\mathsf{T}d)/(v_i^\mathsf{T}x - \tau_i) : i \in [1 : p]\}| = p. \tag{A.1b}$$

Puisque, en appliquant le lemme 3.2.6, on aurait pu ajouter $v_0 = 0$ à la liste des vecteurs non nuls distincts $v_i/(v_i^\mathsf{T}x - \tau_i)$, on peut aussi garantir que

$$v_i^\mathsf{T}d \neq 0, \quad \forall\, i \in [1 : p]. \tag{A.1c}$$

1. Notons d'abord que $v_i^\mathsf{T}x \neq \tau_i$ pour tout $i \in [1 : p]$, puisque x n'appartient à aucun hyperplan, de sorte que les vecteurs $v_i/(v_i^\mathsf{T}x - \tau_i)$ sont bien définis. Maintenant, si on avait $v_i/(v_i^\mathsf{T}x - \tau_i) = v_j/(v_j^\mathsf{T}x - \tau_j)$ pour certains $i \neq j$, il s'ensuivrait que $v_i = \alpha v_j$, pour $\alpha := (v_i^\mathsf{T}x - \tau_i)/(v_j^\mathsf{T}x - \tau_j)$. Alors, $v_i^\mathsf{T}x - \tau_i = \alpha(v_j^\mathsf{T}x - \tau_j)$ ou $\tau_i = \alpha\tau_j$. Cela impliquerait que (v_i, τ_i) et (v_j, τ_j) soient colinéaires dans $\mathbb{R}^n \times \mathbb{R}$.

Pour déterminer les vecteurs de signes des chambres traversées le long du chemin $t \mapsto x+td$ dans $\mathbb{R}^n$, nous déterminons d'abord les t_i auxquels ce chemin rencontre un hyperplan. D'après (A.1c), on peut définir $t_i := -(v_i^\mathsf{T} x - \tau_i)/(v_i^\mathsf{T} d)$, pour $i \in [1:p]$, qui sont p valeurs distinctes d'après (A.1b). Il en résulte que les expressions équivalentes suivantes valent pour tout $i \in [1:p]$: $(v_i^\mathsf{T} x - \tau_i) + t_i(v_i^\mathsf{T} d) = 0$ ou, en utilisant (A.1a),

$$(1-t_i)(v_i^\mathsf{T} x - \tau_i) + t_i(v_i^\mathsf{T} \tilde{x} - \tau_i) = 0 \quad \text{ou} \quad v_i^\mathsf{T}[(1-t_i)x + t_i\tilde{x}] - \tau_i = 0. \tag{A.1d}$$

D'après la dernière expression dans (A.1d), le point $z^i := (1-t_i)x + t_i\tilde{x} = x + t_i d$ appartient au i-ème hyperplan, comme annoncé. Maintenant, d'après la première expression dans (A.1d), $t_i \in (0,1)$ (c'est-à-dire que z^i est dans l'intérieur relatif de $[x,\tilde{x}]$) si et seulement si $(v_i^\mathsf{T} x - \tau_i)$ et $(v_i^\mathsf{T}\tilde{x} - \tau_i)$ sont de signes opposés, ce qui s'écrit aussi $s_i\tilde{s}_i = -1$. Par conséquent, le nombre de t_i dans $(0,1)$ est égal à $l = \sum_{i\in[1:p]} |\tilde{s}_i - s_i|/2 \leqslant p$. Notons-les par

$$0 < t_{i_1} < \cdots < t_{i_l} < 1.$$

Par définition des t_i, pour $t \in (t_{i_j}, t_{i_{j+1}})$, le vecteur de signes $s^{i_j} := \operatorname{sgn}(V^\mathsf{T}[(1-t)x+t\tilde{x}]-\tau)$ est constant, ce qui s'écrit aussi

$$s^{i_j} \bullet (V^\mathsf{T}[(1-t)x + t\tilde{x}] - \tau) > 0, \quad \text{pour } t \in (t_{i_j}, t_{i_{j+1}}).$$

De plus, lorsque $t \in (0,1)$ traverse un $t_{i_j} \in (0,1)$, une seule composante de $V^\mathsf{T}[(1-t)x + t\tilde{x}] - \tau$ change de signe. Nous avons donc défini un chemin de longueur $l \leqslant p$ dans $\mathcal{S}$, à savoir $s^{i_0} = s$, s^{i_1}, …, $s^{i_l} = \tilde{s}$, reliant s à $\tilde{s}$. Ceci prouve l'implication. □

$$(5.17) \qquad -\mathfrak{S}(V,\tau) = \mathfrak{S}(V,-\tau) \qquad \text{et} \qquad -\mathfrak{S}_a(V,\tau) = \mathfrak{S}_a(V,-\tau).$$

Preuve. En effet, soit $\sigma \in \mathfrak{S}(V,\tau)$ avec $J = \mathfrak{J}(\sigma)$. Alors, $J \in \mathcal{C}(V)$ et $\sigma = \operatorname{sgn}(\eta)$ pour un certain $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ vérifiant $\tau_J^\mathsf{T}\eta \geqslant 0$. Puisque $J \in \mathcal{C}(V)$ et $-\sigma = \operatorname{sgn}(-\eta)$ pour un certain $-\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ vérifiant $(-\tau_J)^\mathsf{T}(-\eta) \geqslant 0$, il s'ensuit que $-\sigma \in \mathfrak{S}(V,-\tau)$. Nous avons montré que $-\mathfrak{S}(V,\tau) \subseteq \mathfrak{S}(V,-\tau)$. En changeant τ en $-\tau$ et en utilisant (5.16), on obtient $\mathfrak{S}(V,-\tau) \subseteq -\mathfrak{S}(V,\tau)$. L'identité recherchée $-\mathfrak{S}(V,\tau) = \mathfrak{S}(V,-\tau)$ en découle.

On procède de même pour montrer que $-\mathfrak{S}_s(V,\tau) = \mathfrak{S}_s(V,-\tau)$, en remplaçant les inégalités « $\geqslant$ » par des égalités (en fait, la proposition 5.3.14(1) ci-dessous montrera que $-\mathfrak{S}_s(V,\tau) = \mathfrak{S}_s(V,\tau)$). Enfin, la dernière formule dans (5.17) se déduit directement de la définition $\mathfrak{S}_a(V,\tau) := \mathfrak{S}(V,\tau) \setminus \mathfrak{S}_s(V,\tau)$. □

Le cônage, relié à la section 5.3.4, peut être vu comme suit. Si l'on utilise la matrice $\mathcal{V}_t := [V\ 0; \tau^\mathsf{T}\ t]$ pour un scalaire t, on a

$$\mathcal{S}(\mathcal{V}_t) = [\mathcal{S}(V,\tau) \times \{\operatorname{sgn}(-t)\}] \cup [-\mathcal{S}(V,\tau) \times \{\operatorname{sgn}(+t)\}],$$

et il est clair que prendre la moitié de cet arrangement linéaire donne $\mathcal{S}(V,\tau)$ (à $\operatorname{sgn}(-t)$ près). Des observations très similaires sont réalisables pour les vecteurs souches.

Remarque A.1.2 (D_2 est une heuristique combinatoire)**.** La stratégie discutée aux sections 3.5.2.D_2 et 5.5.2 (la méthode primale-duale) a des proximités avec des heuristiques développées pour d'autres problèmes combinatoires comme le problème SAT avec l'apprentissage de contrainte. □

En effet, D_2 peut se voir comme suit : lorsque l'on arrive à une feuille infaisable de l'arbre, un vecteur souche est obtenu. Un vecteur souche, par définition, est une combinaison de signes infaisable de taille minimale. De fait, cette infaisabilité minimale peut être apprise et réutilisée dans d'autres sous-arbres pour empêcher l'exploration à moindres coût. Le même principe est utilisé en apprentissage de contrainte pour le problème de satisfiabilité de clauses SAT, voir par exemple [201].

A.2 Valeurs des instances

Les types d'instances sont définis dans la section 5.7.1. Les valeurs suivantes sont celles attendues pour les instances RAND, 2D and PERM. Pour les instances SRAND et RATIO, il n'y a pas de formule exacte.

	chambres	circuits / vecteurs souches		
Problèmes	$\lvert\mathcal{S}(V,\tau)\rvert$	$\lvert\mathfrak{S}_s(V,\tau)\rvert/2$	$\lvert\mathfrak{S}_a(V,\tau)\rvert$	$\lvert\mathfrak{S}([V;\tau^{\mathsf{T}}],0)\rvert/2$
RAND-N-P	$\sum_{i=0}^{n}\binom{p}{i}$	0	$\binom{p}{n+1}$	$\binom{p}{n+2}$
2D-N-P	$2^{n-2}\sum_{i=0}^{2}\binom{p-n+2}{i}$	0	$\binom{p-n+2}{3}$	$\binom{p-n+2}{4}$
PERM-N	$(n+1)!$	$\sum_{i=3}^{n+1}\frac{i!}{2i}\binom{n+1}{i}$	0	$\sum_{i=3}^{n+1}\frac{i!}{2i}\binom{n+1}{i}$

TABLE A.1 – Cardinalités connues pour certaines instances. Les valeurs des problèmes RAND-N-P et 2D-N-P sont obtenues par position générale affine, donc les propositions 5.3.31, 3.4.6 et la remarque 5.3.13 6). Rappelons que les vecteurs souches symétriques sont comptés par paires (d'où le facteur $1/2$) – le nombre de circuits *n'a pas* à considérer ce facteur.

Justifions ces valeurs. Pour les instances RAND, puisqu'elles sont générées aléatoirement la position générale (affine) est vérifiée donc il suffit d'utiliser les bornes supérieures. Pour les instances 2D, on peut faire comme suit. Par indépendance des premiers $n-2$ vecteurs, on peut décomposer les vecteurs de signes (de taille p) :

$$
\begin{aligned}
s \in \mathcal{S}(V,\tau) \iff \exists\, x \text{ t.q. } & \begin{cases} s_i \begin{bmatrix} 0 & (v_i)_{[3:n]}^\mathsf{T} \end{bmatrix} x > s_i\tau_i, & \forall\, i \in [1:n-2] \\ s_i \begin{bmatrix} (v_i)_{[1:2]}^\mathsf{T} & 0 \end{bmatrix} x > s_i\tau_i, & \forall\, i \in [n-1:p] \end{cases} \\
\iff & \begin{cases} \exists\, x_{[3:n]} \text{ t.q. } s_i (v_i)_{[3:n]}^\mathsf{T} x_{[3:n]} > s_i\tau_i, & \forall\, i \in [1:n-2] \\ \exists\, x_{[1:2]} \text{ t.q. } s_i (v_i)_{[1:2]}^\mathsf{T} x_{[1:2]} > s_i\tau_i, & \forall\, i \in [n-1:p] \end{cases} \\
\iff & \begin{cases} s_{[1:n-2]} \in \mathcal{S}(V_{[3:n],[1:n-2]}, \tau_{[1:n-2]}) \\ s_{[n-1:p]} \in \mathcal{S}([V_{[1:2],[n-1:p]}], \tau_{[n-1:p]}). \end{cases}
\end{aligned}
$$

De fait, on a :

$$
\begin{array}{rcl}
|\mathcal{S}(V_{:,[n-1:p]}, \tau_{[n-1:p]})| & = & \sum_{i=0}^{2} \binom{n-p+2}{i}, \\
\mathcal{C}(V_{:,[n-1:p]}, \tau_{[n-1:p]}) & = & \{(i,j,k) \in [n-1:p]^3 : i,j,k \text{ différents}\}, \\
\mathcal{C}([V;\tau^\mathsf{T}]_{:,[n-1:p]}, 0) & = & \{(i,j,k,l) \in [n-1:p]^4 : i,j,k,l \text{ différents}\}.
\end{array}
$$

Maintenant, puisque les vecteurs restants (indices dans $[1:n-2]$) sont indépendants des autres et entre eux, les circuits ne contiennent aucun de ces indices et

$$
\mathcal{S}(V,\tau) = \{-1,+1\}^{[1:n-2]} \times \mathcal{S}_{[n-1:p]}(V,\tau), \quad |\mathcal{S}(V,\tau)| = 2^{n-2} \times \sum_{i=0}^{2} \binom{n-p+2}{i},
$$

$$
\mathcal{C}(V,\tau) = \mathcal{C}_{[n-1:p]}(V,\tau), \quad \mathcal{C}([V;\tau^\mathsf{T}],0) = \mathcal{C}_{[n-1:p]}([V;\tau^\mathsf{T}],0).
$$

Pour les instances PERM, le raisonnement est identique à celui de la section 4.5.1. Soit

$$
H_i := \{x : x_i = 1\} \text{ pour } 1 \leqslant i \leqslant n, \quad H_{ij} = \{x : x_i - x_j = 0\} \text{ pour } 1 \leqslant i < j \leqslant n. \tag{A.2}
$$

Ensuite, en utilisant

$$
x \in \mathbb{R}^n \backslash (\cup_{i,j} H_{ij}) \qquad \iff \qquad (x_1, \ldots, x_n) \text{ sont tous différents},
$$

les hyperplans H_{ij} scindent l'espace en $n!$ régions de la forme $x_{\sigma(1)} > \cdots > x_{\sigma(n)}$, une pour chacune des permutations σ de $[1:n]$. Pour σ fixée, on peut avoir les configurations suivantes :

$$
\begin{array}{ll}
x_{\sigma(i)} > 1 \;\forall\, i \in [1:n], & x_{\sigma(1)} < 1, x_{\sigma(i)} > 1 \;\forall\, i > 1, \\
\ldots & \ldots \\
x_{\sigma(i)} < 1, x_{\sigma(n)} > 1 \;\forall\, i < n, & x_{\sigma(i)} < 1 \;\forall\, i \in [1:n].
\end{array}
$$

et toute autre combinaison est de la forme

$$
\{x_{\sigma(1)} < 1, \ldots, x_{\sigma(i^*)} > 1, \ldots, x_{\sigma(j^*)} < 1, \ldots, x_{\sigma(n)} > 1\}
$$

qui ne respecte pas la définition de σ.

Pour les vecteurs souches, rappelons que les circuits sont indépendants de τ, donc par la proposition 4.5.2 on connaît le nombre de circuits et leur structure. Justifions que tous les vecteurs souches sont symétriques.

Pour un circuit donné, s'il ne contient que des vecteurs $e_i - e_j$, puisque $\tau_{ij} = 0$, les vecteurs souches résultants sont symétriques. Autrement, il n'y a que deux vecteurs e_i et e_j dans le circuit, et leurs poids étant opposés, les vecteurs souches résultants sont symétriques.

Pour la matrice augmentée, par la proposition 5.3.20, l'arrangement étant centré, on a $\mathcal{C}(V) = \mathcal{C}([V; \tau^{\mathsf{T}}])$, donc $\mathfrak{S}_0(V, \tau) = \varnothing$.

Pour les instances 2D, il y a quelques irrégularités dans le tableau 5.3. Pour $n = 4$ and $n = 6$, il y a en réalité exactement un vecteur souche symétrique ; on peut montrer que cela réduit le nombre de chambres du sous-arrangment avec indices dans $[n-1 : p]$ d'exactement 1, donc, après multiplication par 2^{n-2}, de 2^{n-2}. Pour $n = 7$, c'est un peu différent. Il y a un vecteur souche asymétrique de taille 2 (disons $\{i, j\}$, donc aucun $\{i, j, k\}$ pour $k \in [n-1 : p] \setminus \{i, j\}$ n'est un circuit), i.e., deux hyperplans sont parallèles ; cela réduit également le nombre de chambres de 2^{n-2}, cela change aussi le nombre de vecteurs souches.

Pour les instances RATIO, il peut y avoir du "mauvais conditionnement" à cause de la façon dont les instances sont générées : puisque les vecteurs sont des combinaisons linéaires des précédents, certains peuvent avoir des coordonnées larges. Cela a parfois résulté en une ou deux chambres en moins détectées par certains algorithmes. [2]

A.3 Comportements algorithmiques

A.3.1 Heuristiques primales

Heuristique B

D'abord, on considère la modification B, qui évite l'utilisation d'optimisation linéaire et/ou de vecteurs souches lorsque le point courant est "proche" du nouvel hyperplan. Bien que pour des instances fortement aléatoires, il est improbable que cette heuristique soit très efficace, pour des instances PERM par exemple elle peut jouer un rôle important.

En effet, rappelons qu'après la modification A les sous-arbres démarrent avec n signes, dont les indices correspondent à la partie de V qui est l'identité. De fait, le point témoin correspondant est de la forme $x^s = s/\sqrt{n}$ (en le normant). Ensuite, beaucoup des hyperplans restants, qui sont de la forme $(e_i - e_j)^\perp$, contiendront x^s. Voir l'explication détaillée plus bas.

2. Même en normant les données V et τ, selon quelle norme a été utilisée, parfois une chambre était détectée et parfois non.

Heuristique C

Maintenant, on évoque la modification C, qui, à un vecteur de signe $s = (s_{i_1}, \ldots, s_{i_k}) \in \{\pm 1\}^k$, suggère l'hyperplan suivant avec un indice dans $[1 : p] \setminus \{i_1, \ldots, i_k\}$ qui idéalement fait que s n'a qu'un descendant. Démarrons par illustrer cela pour les instances 2D.

Les matrices V y ont une forme particulière, avec deux sous-arrangements indépendants. De fait, après la modification A, les hyperplans du sous-arrangements en dimension $n-2$ sont déjà considérés. Les hyperplans restants forment un arrangement en dimension 2, avec deux de ces hyperplans complétant les $n-2$ premiers via la factorisation QR. Lorsque l'on ajoute les hyperplans restants, il est probable que les hyperplans donnant lieu à un seul descendant soient ajoutés d'abord, voir figure A.1.

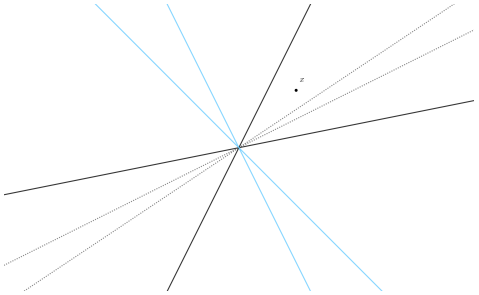


Figure A.1 – Les lignes noires sont les hyperplans déjà considérés et x est un point de la région courante. Il est simple d'ajouter d'abord les deux hyperplans bleus, qui impliquent un seul descendant, puis d'ajouter les hyperplans en pointillés qui impliquent deux descendants. Bien que la figure soit montrée pour un arrangement centré, le principe est similaire pour le cas affine.

Nous poursuivons avec les instances linéaires perm : rappelons que pour tout $s \in \{\pm 1\}^n$, le sous-arbre démarrant avec s a le point témoin $x^s = s/\sqrt{n}$. Cette expression simple, combinée avec la forme de V, rend très simple le choix des hyperplans n'impliquant qu'un seul descendant (selon s). Par exemple, considérons que $s_1 = +1, s_2 = -1$, alors l'hyperplan $(e_1 - e_2)^\perp$ ne peut avoir que le signe $+1$ dans le sous-arbre de s.

Parmi les n premiers indices, notons $I_s^+ = \{i \in [1 : n], s_i = +1\}$ et $I_s^- = \{i \in [1 : n], s_i = -1\}$. Soit $i^+ \in I_s^+, i^- \in I_s^-$, si $i^+ < i^-$ alors le signe de l'hyperplan $(e_{i^+} - e_{i^-})^\perp$ doit être $+1$ pour que les vecteurs de signe soit réalisables, autrement on aurait $x_{i^+} > 0, x_{i^-} < 0$ et $x_{i^+} - x_{i^-} < 0$, qui est un (sous-)système sans solution. Si $i^+ > i^-$, on change juste les signes dans le raisonnement (puisque l'hyperplan s'écrit $(e_{i^-} - e_{i^+})^\perp$). De fait, à chaque sous-arbre démarrant à $s \in \{\pm 1\}^n$, chaque hyperplan de la forme décrite ne peut impliquer qu'un descendant : parmi les $n(n-1)/2$ hyperplans restants, $|I_s^+||I_s^-|$ n'impliquent qu'un descendant et sont détectables facilement. Avec ce raisonnement, il y a

$$\sum_{s \in \{\pm 1\}^n} |I_s^+||I_s^-| = \sum_{i = |I_s^+| \in [0,n]} \binom{n}{i} i(n-i) = n(n-1)2^{n-2}$$

hyperplans ajoutables facilement, ce qui est la moitié du nombre total d'hyperplans car il y a 2^n sous-arbres et chacun a $n(n-1)/2$ hyperplans à ajouter, c'est-à-dire $2^n n(n-1)/2$ hyperplans au total.[3] Cela peut epxliquer les bonnes améliorations observées pour ces instances dans le tableau A.5.

Par ailleurs, si l'on prend les instances PERM affines, les points de départ donc $e + s/\sqrt{n}$, mais pour toute paire $i \neq j \in [1:n]^2$, $\langle e_i - e_j, e + s/\sqrt{n}\rangle = \langle e_i - e_j, s/\sqrt{n}\rangle$ donc le raisonnement reste valide.

Pour les instances CROSSPOLYTOPE, leurs vecteurs souches comme leurs chambres sont connues exactement, voir section 4.5.2. En particulier, les vecteurs souches sont de la forme $s_i = s_{i+n} \neq s_j = s_{j+n}$ pour $i \neq j \in [1:n]$ et les autres coordonnées sont nulles. Puisque la matrice V est de rang $n+1$, les sous-arbres démarrent avec $n+1$ coordonnées. Pour chaque paire $(e_0 + e_i, e_0 - e_i)$, précisément un de ces vecteurs fait partie des $n+1$, mis à part pour un indice, disons i^*, pour lequel les deux vecteurs de la paire correspondante sont dans les $n+1$ premiers.[4] Supposons que $s_{i^*} \neq s_{i^*+n}$, par exemple $s_{i^*} = +1, s_{i^*+n} = -1$, les inégalités correspondantes sont

$$x_0 + x_{i^*} > 0 \qquad \text{et} \qquad -x_0 + x_{i^*} > 0,$$

qui sont vérifiées pour x_{i^*} positif assez grand. Si $s_{i^*} = -1, s_{i^*+n} = +1$, les inégalités sont vérifiées pour x_{i^*} négatif assez grand. Dans les deux cas, si $s_{i^*} \neq s_{i^*+n}$, les deux contraintes i^* et $i^* + n$ peuvent être satisfaites par le choix de x_{i^*}. Si $s_{i^*} = +1 = s_{i^*+n}$, les deux équations correspondantes sont

$$x_0 + x_{i^*} > 0, \qquad x_0 - x_{i^*} > 0 \qquad \Longrightarrow \qquad x_0 > 0$$

De fait, pour chacun des $n-1$ hyperplans restants, on ne peut avoir $s_j = -1 = s_{j+n}$, puisque par un argument similaire cela impliquerait $x_0 < 0$. Dit autrement, le vecteur de signe recouvrirait un vecteur souche. (Si $s_{i^*} = -1 = s_{i^*+n}$, on utilise le même argument en inversant les signes.)

Comptons combien d'hyperplans impliquant un seul descendant peuvent être ajoutés. Parmi les 2^{n-1} sous-arbres de départ avec $s_{i^*} = +1 = s_{i^*+n}$, on peut ajouter, en notant s' les $n-1$ signes des indices différents de i^* et $i^* + n$,

$$\sum_{s' \in \{\pm 1\}^{n-1}} |\{j \in [1:n] \setminus \{i^*\} : s'_i = -1 \text{ or } s'_{i+n} = -1\}| = \sum_{k=0}^{n-1} \binom{n-1}{k} k = (n-1)2^{n-2}$$

hyperplans qui n'impliquent qu'un seul descendant. De fait, il y a $(n-1)2^{n-1}$ hyperplans que l'on peut ajouter ainsi. Comparé aux $(n-1)2^{n+1}$ hyperplans, cela signifie qu'un quart des hyperplans au total peuvent être ajoutés directement avec un seul descendant à chaque fois.

3. Même en prenant la symétrie en compte, les deux valeurs sont divisées par deux donc c'est toujours la moitié des hyperplans qui n'implique qu'un seul descendant.

4. Sinon, si pour plusieurs paires on a les deux vecteurs, il est impossible de générer $\mathbb{R}^{n+1}$.

Maintenant, si l'on a $s_{i^*} \neq s_{i^*+n}$, leurs inégalités peuvent être ignorées en choisissant x_{i^*} judicieusement. Supposons que le prochain indice ajouté est j^* ou $j^* + n$ tel que $s_{j^*} = s_{j^*+n}$, alors on peut réitérer le processus sur les indices restants. Si $s_{j^*} \neq s_{j^*+n}$, on continue jusqu'à ce que tous les signes soient ajoutés ou que deux indices modulo n aient le même signe. Cependant, dans ces sous-arbres le premier hyperplan ajouté produit deux descendants (nécessairement, aucun vecteur souche ne peut être recouvert). De fait, bien que dans *ces* descendants il est facile de trouver des hyperplans n'impliquant qu'un seul descendant, le dénombrement devient laborieux. [5]

A.3.2 Heuristiques duales

Dans cette section, comme on considère les instances linéaires, par "vecteur souche" on comprend une paire $\pm\operatorname{sgn}(\eta)$ de deux vecteurs souches opposés. Donc, on compte le nombre de circuits (voir les facteurs $1/2$ dans le chapitre 5).

Position générale et aléatoire

Mentionnons une observation importante : les instances aléatoires sont en position générale, ce qui signifie un nombre de chambres maximal (proposition 5.3.31) *et* nombre maximal de circuits (remarque 5.3.13 6)). Cela peut sembler contradictoire, puisque les circuits génèrent $\mathcal{S}^c$, donc on peut penser de prime abord que beaucoup de circuits signifie moins de vecteurs souches dans $\mathcal{S}$. L'explication est qu'en position générale, les circuits sont *tous* les sous-ensembles de taille $r+1$. Cependant, puisque c'est leur taille maximale (j vecteurs avec $j \geqslant r+2$ ont une nullité $j - r > 1$), tous ces circuits génèrent "moins" de vecteurs de signes dans $\mathcal{S}^c$.

Cette observation est intéressante pour deux aspects. D'abord, pourquoi l'algorithme complètement dual prend "la double peine" : en plus d'avoir des tests de couverture pas forcément beaucoup plus efficaces que l'optimisation linéaire, il y a beaucoup de vecteurs souches et de tests de couverture à faire. Les instances avec beaucoup d'aléatoire comme les RATIO ou SRAND avec beaucoup de composantes (Q pas trop petit comparé à N) risquent de souffrir de la même observation. Cela peut, dans une moindre mesure, amoindrir les performances de la méthode primale-duale : les vecteurs souches acquis ont un impact limité.

Pertinence de la méthode primale-duale

À la lumière du paragraphe précédent, la méthode primale-duale est intéressante surtout pour les instances avec pas trop de vecteurs souches (i.e., l'instance n'est pas trop

5. On conjecture, avec cette observation, que $n2^n - 2^{n+1} + 2 = \sum_{i=0}^{n-1} i2^i$ parmi les $2^{n+1}(n-1)$ hyperplans peuvent être ajoutés directement. Cela fait environ la moitié au total.

aléatoire). C'est là que les instances avec une certaine structure (combinatoire) sont intéressantes : puisque le nombre d'hyperplans peut augmenter rapidement, calculer tous les vecteurs souches, sans les techniques de [212] (et [35]), prend trop de temps.

Cependant, de telles instances ont plutôt de petits vecteurs souches – pas nécessairement que des petits, comme on l'a vu pour les instances PERM dans la section 4.5.1. Cette petite taille les rend particulièrement efficaces puisque l'on peut élaguer d'autres sous-arbres plus facilement.

En effet, pour les instances PERM, comme vu dans la section 4.5.1, il y a $(k-1)!\binom{n+1}{k}/2$ vecteurs souches de taille k, et approximativement la moitié ont leur premières n composantes nulles. De fait, si ceux-ci sont acquis, comme ils sont indépendants du sous-arbre de départ (correspondant aux hyperplans $e_i^\perp$ pour $i \in [1:n]$), l'algorithme peut les réutiliser partout ; cependant, même les circuits qui utilisent des indices de $[1:n]$ peuvent être réutilisés efficacement puisqu'ils n'ont que deux composantes non nulles sur les n premiers indices.

Par ailleurs, considérons les intances CROSSPOLYTOPE. Nous illustrons ce phénomène sur CROSSPOLYTOPE-9 et ses $\binom{n}{2} = \binom{9}{2} = 36$ circuits. Parmi les $2^{n+1-1} = 2^9 = 512$ sous-arbres de départ, tous les vecteurs souches étaient trouvés après les 130 premiers sous-arbres, soit environ 25%, et 33/36 dès les premiers 7% des sous-arbres de départ. De fait, l'algorithme primal-dual peut bénéficier ici de la connaissance de (presque) tous les vecteurs souches sans les calculer tous explicitement dès le début.

Dans les instances 2D, rappelons que les vecteurs souches sont les $\binom{p+2-n}{3}$ vecteurs avec des coordonnées non nulles sur exactement trois des $p+2-n$ derniers indices. Chaque vecteur souche acquis avec des zéros aux coordonnées $n-1$ et n peut être réutilisé pour élaguer l'arbre dans chaque autre sous-arbre de départ. De fait, chaque vecteur souche acquis peut être beaucoup utilisé, et, comme ils sont de taille 3, "tôt" dans l'arbre. De plus, le pourcentage de tests de couverture utiles est autour de 70-75%, ce qui signifie que malgré avoir peu de vecteurs souches on peut détecter quand l'arbre doit être arrêté.

A.3.3 Analyse de l'algorithme compact

Théorie

Pour conclure cette section, expliquons ce qu'on peut attendre de la technique de compaction présentée. Pour simplifier l'exposition, nous analysons la version compacte purement primale (algorithmes 5.6.4-5.6.5) et la version purement duale (non présentée).

L'analyse se fait en deux étapes. La première expose les situations où les algorithmes compacts (versions purement primale et duale sans heuristiques supplémentaires) résolvent moins de sous-problèmes, et la seconde quantifie ces situations.

Remarque A.3.1 (imprécision de l'estimation duale)**.** Dans l'algorithme purement dual, à un nœud s donné, les situations suivantes peuvent se produire. Le premier descendant couvre un vecteur souche donc le second est toujours faisable (un seul test) ; sinon, le second descendant nécessite un deuxième test. Ainsi, quand tous les descendants ne sont pas faisables, on ne peut pas connaître a priori le nombre de tests.

Proposition A.3.2 (différence du nombre de sous-problèmes)**.** *Les algorithmes primaux et duaux sur l'arbre $\mathcal{S}$, à un nœud $s \in \mathcal{S}_k$ et éventuellement son opposé $-s$, résolvent un nombre de sous-problèmes donné par le tableau suivant.*

	POL					tests de couverture				
	$\pm s \in \mathcal{S}_k$			$(-)s \in \mathcal{S}_k$		$\pm s \in \mathcal{S}_k$			$(-)s \in \mathcal{S}_k$	
nombre de descendants	4	3	2	2	1	4	3	2	2	1
algorithme classique	2	2	2	1	1	4	3 ou 4	3	1 ou 2	1 ou 2
algorithme compact	1	2	2	1	1	2	1 ou 2	1 ou 2	1 ou 2	1 ou 2

Table A.2 – Nombre de sous-problèmes résolus selon l'état du nœud courant. De plus, on note $\pm s \in \mathcal{S}_k \iff \boxed{\mathbb{S}} = 0$ et $(-)s \in \mathcal{S}_k \iff \boxed{\mathbb{S}} \neq 0$.

En particulier, pour l'algorithme primal[6]*, les seules situations où moins de sous-problèmes sont résolus concernent les nœuds symétriques ayant chacun deux descendants, c'est-à-dire des nœuds symétriques avec descendants symétriques (donc 4 descendants au total).*

La preuve justifie brièvement tous les cas, évoquant les 10 colonnes après la première.

Preuve.

[4 descendants, primal] [classique] L'algorithme classique obtient deux des quatre descendants puis nécessite deux POL (un pour s, un pour $-s$, concluant tous deux que le vecteur de signes est faisable). [compact] L'algorithme compact obtient deux descendants faciles et, par symétrie, le POL conclura que $\pm(s, -s_{k+1})$ sont faisables.

[3 descendants, primal] [classique] De même, deux POL sont nécessaires (mais ils donnent des conclusions opposées). [compact] De même, deux descendants faciles, mais le POL échouera et nécessitera un nouveau POL (dans $\mathbb{R}^{n+1}$) qui sera faisable ($\boxed{\mathbb{S}} \neq 0$).

[2 descendants, s et $-s$, primal] [classique] De même, deux POL sont nécessaires (concluant tous deux que le vecteur de signes est infaisable). [compact] De même, après les deux descendants faciles (car $\boxed{\mathbb{S}} = 0$ ici), le premier POL indique que le descendant est infaisable, tout comme le second (dans $\mathbb{R}^{n+1}$).

[2 descendants, seulement s, primal] [classique] Un seul nœud donc un POL (descendant faisable). [compact] Ici, $\boxed{\mathbb{S}} \neq 0$, donc l'algorithme utilise un POL (descendant faisable).

6. Algorithme primal sans heuristiques = algorithme RC.

[1 descendant, primal] [classique] Un seul nœud donc un POL (vecteur de signes infaisable). [compact] Ici, $\boxed{\mathrm{S}} \neq 0$, donc l'algorithme utilise un POL (descendant infaisable).

[4 descendants, dual] [classique] Pour s et $-s$, aucun vecteur souche n'est couvert donc deux tests sont nécessaires pour chacun (4 au total). [compact] Pour s, aucun vecteur souche n'est couvert donc deux tests sont nécessaires.

[3 descendants, dual] [classique] Sans perte de généralité, supposons que s a deux descendants et $-s$ un : s nécessite deux tests de couverture et $-s$ un ou deux (voir remarque A.3.1). [compact] De même, selon l'ordre des descendants potentiels, un test (si le premier test couvre un vecteur souche) ou deux (si le premier échoue) sont nécessaires.

[2 descendants, s et $-s$, dual] En fait, par symétrie on peut montrer que s et $-s$ ont des descendants opposés. S'ils ont des descendants $(s, +1)$ et $(-s, +1)$, cela contredit le fait que les régions de s et $-s$ ont des cônes asymptotiques opposés. [classique] De fait, trois tests sont nécessaires (puisque les descendants sont opposés donc un de deux requiert deux tests, l'autre un). [compact] Un test si le premier descendant potentiel est infaisable (même dans $\mathbb{R}^{n+1}$), deux si c'est le second qui est testé en premier.

[2 descendants, seulement s, dual] [classique] De même, un ou deux tests selon l'ordre. [compact] Identique au cas classique.

[1 descendant, dual] [classique] De même, un ou deux tests selon l'ordre. [compact] Identique au cas classique. □

Remarque A.3.3 (précision sur les tests de couverture)**.** Dans la variante duale, quand $\pm s \in \mathcal{S}_k$ ($\boxed{\mathrm{S}} = 0$), il y a moins de tests mais ils peuvent nécessiter plus de vecteurs souches car ils utilisent ceux de $([V; \tau^{\mathsf{T}}], 0)$ et (V, τ). Quand seul s ou $-s$ reste ($\boxed{\mathrm{S}} \neq 0$), les vecteurs souches considérés par l'algorithme compact sont seulement ceux de $[V; \tau^{\mathsf{T}}]$. □

Passons à la seconde partie, où nous donnons un sens particulier à la partie du tableau A.2 avec 4 descendants.

Définition A.3.4 (lignée directe complète)**.** Soit $\mathcal{A}(V, \tau)$ pour $V \in \mathbb{R}^{n\times p}$ et $\tau \in \mathbb{R}^p$ un arrangement. Considérons le calcul de ses chambres par l'algorithme de l'arbre $\mathcal{S}$. Pour $k \in [1 : p-1]$, un $s \in \mathcal{S}_k$ a une lignée directe complète si $-s \in \mathcal{S}_k$, et si s et $-s$ ont chacun deux descendants, c'est-à-dire que $(s, +1)$, $(s, -1)$, $(-s, +1)$, $(-s, -1)$ appartiennent tous à $\mathcal{S}_{k+1}$.

Naturellement, s a une lignée directe complète $\iff$ $-s$ aussi. □

Remarquons qu'a priori, la lignée directe complète (LDC) dépend de l'ordre des vecteurs de signes considérés. De plus, d'après la proposition A.3.2, un sous-problème de moins est résolu par l'algorithme compact primal à chaque paire de vecteurs opposés avec LDC. Par convention, les feuilles de l'arbre sont considérées comme n'ayant pas de descendants.

Proposition A.3.5 (paires avec 4 descendants et $\mathcal{S}_s$)**.** *Le nombre de vecteurs de signes avec LDC dans l'arbre $\mathcal{S}$ vaut $|\mathcal{S}_s| - 2$.*[7]

Preuve. La preuve procède par récurrence sur p, le nombre de vecteurs et la profondeur de l'arbre. [Initialisation] Pour $p = 1$, on a $\mathcal{S} = \{+1, -1\}$ (si $V \neq 0_n$), et le nombre de paires symétriques est $0 = 2 - 2$.[8]

[Hérédité] Supposons la propriété vraie pour tout arrangement à p vecteurs et ajoutons un $(p+1)$-ème vecteur. Soit $s \in \mathcal{S}_p$. S'il est asymétrique ($-s \notin \mathcal{S}_p$), ses descendants le sont aussi : ni le nombre de vecteurs avec LDC ni $|\mathcal{S}_s|$ n'augmentent.

Si s et $-s$ sont dans $\mathcal{S}_p$ et ont deux ou trois descendants au total, seulement deux sont symétriques (voir la preuve de la proposition A.3.2) : ni le nombre de vecteurs avec LDC ni $|\mathcal{S}_s|$ n'augmentent (s et $-s$ dans $\mathcal{S}_p$ deviennent $(s, +1)$ et $(-s, -1)$ ou $(s, -1)$ et $(-s, +1)$ dans $\mathcal{S}_p$).

Le dernier cas est celui de la LDC, c'est-à-dire $\pm s \in \mathcal{S}_p$ avec quatre descendants dans $\mathcal{S}_{p+1}$. Clairement, ces deux vecteurs augmentent $|\mathcal{S}_s|$ de deux car s et $-s$ dans $\mathcal{S}_p$ deviennent $(s, +1)$, $(-s, -1)$, $(s, -1)$, $(-s, +1)$ dans $\mathcal{S}_{p+1}$. Enfin, il y a deux vecteurs supplémentaires avec LDC, s et $-s$. Ceci complète la preuve. □

Corollaire A.3.6 (LDC et $\mathcal{S}_s$)**.** *Le nombre de vecteurs de signes (partiels) dans l'arbre $\mathcal{S}$ avec lignée directe complète vaut $|\mathcal{S}_s| - 2$. Comme un sous-problème de moins est résolu à chaque paire opposée de tels vecteurs, l'algorithme compact primal résout $|\mathcal{S}_s|/2 - 1$ sous-problèmes de moins.*[9]

Remarquons que seuls les POL avec des vecteurs de signes *faisables* sont évités par l'algorithme compact. Avant d'énoncer un corollaire intéressant, nous avons (en notant « (c) » pour compact) :

$$\#\mathrm{POL(c)} = \#\mathrm{POL} - \frac{|S_s(V,\tau)|}{2}, \quad \frac{\#\mathrm{POL}}{\#\mathrm{POL(c)}} = 1 + \frac{|S_s(V,\tau)|}{2\#\mathrm{POL(c)}}. \tag{A.3}$$

Rappelons que $S_s(V, \tau) = S(V, 0)$, mais nous utilisons la première expression pour distinguer plus facilement entre arrangements affines et linéaires. Cependant, comme les expériences numériques utilisent la modification A (départ avec 2^n sous-arbres issus d'un sous-arrangement complet), la formule prend la forme modifiée :

$$\#\mathrm{POL(c)} = \#\mathrm{POL} - \frac{|S_s(V,\tau)| - 2^n}{2}, \quad \frac{\#\mathrm{POL}}{\#\mathrm{POL(c)}} = 1 + \frac{|S_s(V,\tau)| - 2^n}{2\#\mathrm{POL(c)}}. \tag{A.4}$$

Corollaire A.3.7 (LDC indépendante de l'ordre des vecteurs)**.** *Puisque $|\mathcal{S}_s|$ est indépendant de l'ordre des vecteurs, il en va de même pour le nombre de nœuds avec LDC.*

7. Ce -2 peut paraître curieux, bien qu'il semble intrinsèque. Si on considère la racine vide comme étant à la fois « $+\varnothing$ et $-\varnothing$ » avec leurs descendants $\{(\varnothing, 1), (\varnothing, -1), (-\varnothing, 1), (-\varnothing, -1)\}$), on obtient $|\mathcal{S}_s|$.

8. Si on compte la racine comme 2, on a $2 = |\mathcal{S}_s| = 2$.

9. Le terme -1 pourrait être annulé si, par convention, on compte 1 POL pour obtenir les deux premiers vecteurs $\mathcal{S}_1 = \{-1, +1\}$.

Aspects numériques

Nous observons maintenant de la pertinence des algorithmes compacts introduits dans la section 5.6 et présentés dans le tableau 5.5. La première colonne de temps représente l'algorithme symétrisé avec la modification A, qui est très proche de l'algorithme RC symétrisé (la modification A a peu d'effet sur les temps de calcul). Proposons quelques observations principales.

- Pour les instances RAND, les algorithmes compacts sont toujours meilleurs – les ratios semblent augmenter de 0,5-0,6 en moyenne.
- Pour les instances SRAND, les algorithmes compacts sont légèrement meilleurs.
- Pour les instances 2D, les algorithmes compacts donnent de moins bons résultats.
- Pour les instances PERM, les algorithmes compacts montrent une amélioration plutôt convaincante – ce qui était attendu puisque les instances sont centrées.
- Pour les instances RATIO, les algorithmes compacts sont plus efficaces lorsqu'ils sont primaux mais moins bons lorsqu'ils calculent la liste complète des vecteurs stems.

Une explication possible concernant les instances 2D est la suivante. Considérons le sous-arrangement défini par les indices $[n-1:p]$ et soit $q = p-n+2$. Il possède $\sum_{i=0}^{n} \binom{q}{i}$ chambres et la partie symétrique est $2\sum_{i=0}^{n-1} \binom{q-1}{i} = \sum_{i=0}^{n} \binom{q}{i} - \binom{p-1}{n}$. Cependant, comme ce terme est multiplié par 2^{n-2}, il en résulte que les instances 2D ont une proportion bien plus grande de chambres asymétriques par rapport aux autres instances, voir le tableau A.3. Comme l'objectif de l'algorithme compact est d'éviter les calculs symétriques, il est logique que ses performances se dégradent sur des instances moins symétriques.

type d'instance	RAND	SRAND	2D	PERM	RATIO
estimation $\frac{\lvert\mathcal{S}(V,0)\rvert}{\lvert\mathcal{S}(V,\tau)\rvert}$ (%)	> 80	> 70	25	100	50

TABLE A.3 – Proportions approximatives de chambres symétriques. Les instances 2D ont une proportion particulièrement faible de chambres symétriques.

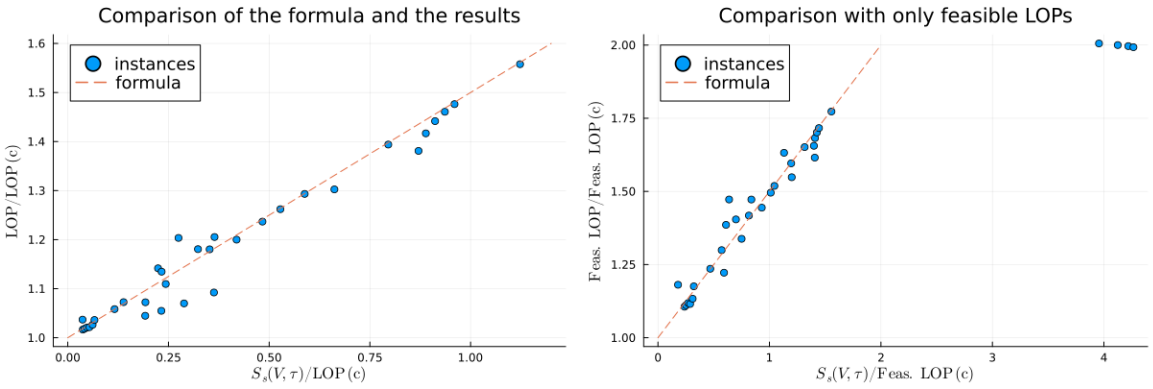
Sous-problèmes Dans le tableau suivant, nous mentionnons les instances testées (leurs valeurs de n et p). Le RC désigne l'algorithme originel [208] tandis que (c) désigne la version "compacte".

Détaillons ce qui se passe pour les instances PERM, qui sont légèrement décalées vers la droite dans les figures A.2b et A.2d. Les coefficients des vecteurs étant dans $\{-1, 0, +1\}$, et que chaque (v_i, τ_i) n'a que deux composantes non nulles, par exemple 0 appartient à la plupart des hyperplans, donc de nombreux problèmes d'optimisation linéaire sont sautés puisqu'il y a clairement deux descendants. Si cela n'avait pas été fait, le nombre de POL faisables et de POL(c) faisables devrait être approximativement doublé : les coordonnées sur l'axe y resteraient donc à la même valeur tandis que celles sur l'axe x seraient divisées

Problèmes	$\lvert S(V,\tau)\rvert$	$\lvert S_s(V,\tau)\rvert$	Réal.(RC)	Réal.	Réal.(c)	Réal. - Réal.(c)	(A.4)
rand-2-8	37	16	35	33	27	6	6
rand-4-8	163	128	161	147	91	56	56
rand-4-9	256	186	254	240	155	85	85
rand-5-10	638	512	636	606	366	240	240
rand-4-11	562	352	560	546	378	168	168
rand-6-12	2510	2048	2508	2446	1454	992	992
rand-5-13	2380	1588	2378	2348	1570	778	778
rand-7-14	9908	8192	9906	9780	5748	4032	4032
rand-7-15	16384	12952	16382	16256	9844	6412	6412
rand-8-16	39203	32768	39201	38947	22691	16256	16256
rand-9-17	89846	78406	89844	89334	50387	38947	38947
2d-4-20	684*	144	682	668	604	64	128
2d-5-20	1232	272	1230	1200	1080	120	120
2d-6-20	2176*	512	2174	2112	1888	224	224
2d-7-20	3840**	960	3838	3712	3328	384	416**
2d-8-20	6784	1792	6782	6528	5760	768	768
srand-8-20-2	36225	24544	36223	35029	24712	10317	12144
srand-8-20-4	213467	157192	213465	212847	140143	72704	78468
srand-8-20-6	245396	186430	245394	244925	153474	91451	93087
perm-5	720	720	718	365	182	183	344
perm-6	5040	5040	5038	2444	1222	1222	2488
perm-7	40320	40320	40318	19080	9557	9523	20096
perm-8	362880	362880	362878	169560	85089	84471	181312
ratio-3-20-0.7	1119	304	1117	1111	945	166	148
ratio-3-20-0.9	1176	178	1174	1168	989	179	85
ratio-4-20-0.7	6015	2278	6013	5999	4855	1144	1131
ratio-4-20-0.9	4600	2016	4598	4584	3528	1056	1000
ratio-5-20-0.7	15136	8470	15134	15104	11289	3815	4219
ratio-5-20-0.9	11325	7826	11323	11293	6920	4373	3897
ratio-6-20-0.7	59519	26194	59517	59455	42906	16549	13065
ratio-6-20-0.9	53795	26758	53793	53731	38261	15470	13347
ratio-7-20-0.7	135064	76790	135061	134935	91654	43281	38331
ratio-7-20-0.9	135039	58468	135036	134910	91630	43280	29170

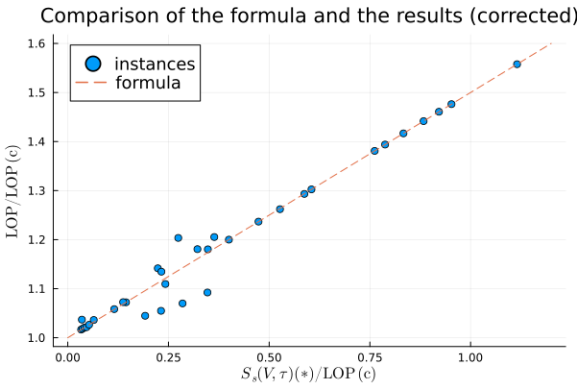
Table A.4 – Valeurs pertinentes pour les instances affines. Les colonnes 2 et 3 représentent les cardinalités des ensembles de vecteurs de signes, les colonnes 4-5-6 les nombres de problèmes réalisables résolus. La colonne 7 est la différence des deux précédentes. La dernière colonne représente le second terme du membre de droite de (A.4). Ce tableau est illustré ci-dessous dans la figure A.2. Les * représentent les irrégularités mentionnées précédemment (pas de position générale parfaite dans le sous-arrangement).

par deux, ce qui signifie que les valeurs seraient beaucoup plus proches des autres et de la ligne.

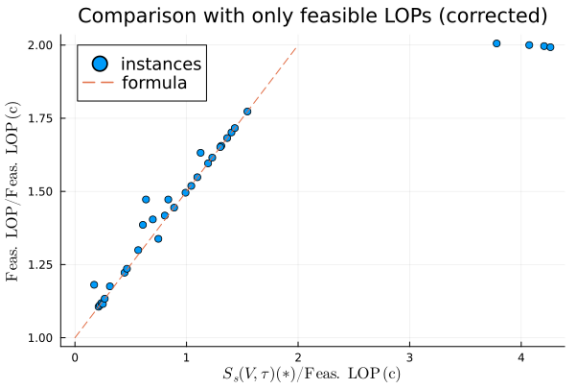


(a) Avec tous les POL

(b) Avec seulement les POL faisables



(c) Avec tous les POL, corrigé



(d) Avec seulement les POL faisables, corrigé

FIGURE A.2 – Illustration de (A.4) (et (A.3)) (“LOP” signifie POL). La correction, c’est-à-dire la différence entre (A.3) et (A.4), est notée $(*)$ et ajoutée aux images du bas ; pour les instances en haut à droite des graphiques de droite, cela place les points sur la ligne correspondant à la formule, donnée par $y = 1 + x/2$. Le (c) désigne le nombre de POL de la variante compacte. Pour les images de droite, quatre points, correspondant aux instances PERM, sont décalés vers la droite. Une explication possible est proposée ci-contre.

A.4 Instances linéaires et autres sujets

Bien que les instances linéaires ne soient pas le but principal du code Julia, c’est important de savoir comment le code s’est comporté dans leur traitement. Les résultats sont donnés dans le tableau A.5. Globalement, on observe des ratios assez similaires. Pour les instances combinatoires présentées dans la dernière partie de la section 3.5.2.A, des résultats encourageants sont observés, surtout pour l’algorithme primal-dual.

Dans ce qui suit, rappelons que D1 utilise quelques vecteurs souches obtenus au début, D2 est la méthode primale-duale, D3 est la méthode intermédiaire avec tous les vecteurs souches et les directions pour avoir un descendant sans coût, et D4 est la méthode complètement duale.

Problem	Temps de calcul (en sec) et ratios d'amélioration						
	RC	P		PD		D	
rand-8-15-7	6.04	**3.36**	**1.80**	3.74	1.62	4.53	1.33
rand-9-16-8	14.3	**8.34**	**1.72**	9.54	1.50	13.6	1.05
rand-10-17-9	32.7	**19.7**	**1.66**	23.3	1.40	33.1	0.987
srand-8-20-2	15.7	3.96	3.97	**2.90**	**5.42**	9.74	1.62
srand-8-20-4	79.6	**36.7**	**2.17**	42.7	1.86	204	0.389
srand-8-20-6	100	**50.2**	**1.99**	67.0	1.49	42	0.238
2d-6-20	1.10	0.551	1.99	**0.20**	**5.49**	0.566	1.94
2d-7-20	1.93	1.34	1.44	**0.444**	**4.36**	1.01	1.92
2d-8-20	3.42	1.72	1.99	**0.580**	**5.90**	1.73	1.98
perm-6	5.11	1.21	4.22	**0.448**	**11.4**	3.55	1.44
perm-7	50.7	10.2	4.97	**3.92**	**13.0**	64.6	0.786
perm-8	551	99.7	5.53	**41.7**	**13.2**	1520	0.362
ratio-5-20-0.7	8.43	4.50	1.87	**3.76**	**2.24**	12.2	0.691
ratio-6-20-0.7	20.7	11.6	1.79	**10.8**	**1.91**	40.1	0.516
ratio-7-20-0.7	53.0	**31.7**	**1.67**	33.6	1.58	165	0.321
threshold-3	0.0464	0.0245	1.89	0.0260	1.78	**0.0233**	**1.99**
threshold-4	1.67	0.784	2.13	**0.557**	**3.00**	0.975	1.71
threshold-5	167	50.2	3.33	**35.8**	**4.67**	1300	0.129
resonance-4	0.367	0.103	3.55	**0.0517**	**7.10**	0.221	1.66
resonance-5	22.5	4.28	5.27	**2.11**	**10.7**	70.5	0.320
resonance-6	4240	**565**	**7.50**	684	6.19	*	*
crosspolytope-10	62.7	11.	5.73	**7.39**	**8.49**	38.6	1.62
crosspolytope-11	193	30.1	6.43	**21.9**	**8.84**	123	1.57
crosspolytope-12	595	83.0	7.16	**65.3**	**9.11**	369	1.61
demicube-4	0.0546	0.0137	3.97	**0.00854**	**6.39**	0.0299	1.83
demicube-5	2.79	1.09	2.55	**0.718**	**3.89**	1.58	1.77
demicube-6	434	119	3.65	**90.1**	**4.82**	3210	0.135

Table A.5 – Temps de calcul en noir et ratios temps(RC) / temps(A) en bleu pour les instances linéaires et les différents algorithmes ; les meilleurs sont en gras. Pour l'algorithme D, les vecteurs souches et les tests de couverture sont calculés de façon un peu plus efficace comme décrit en sections A.4.1 et A.4.2.

A.4.1 Calcul des circuits

Le calcul des circuits peut être effectué avec l'algorithme 5.5.1. Lorsque l'arrangement est linéaire, l'algorithme est essentiellement le même, sauf qu'on obtient toujours $\tau_j^{\mathsf{T}}\eta = 0$ puisque $\tau = 0$. Pour l'algorithme compact, comme les circuits de V et $[V;\tau^{\mathsf{T}}]$ sont nécessaires, il y a un léger changement à vérifier : lorsqu'un circuit de V est trouvé, on vérifie s'il est déjà un circuit de $[V;\tau^{\mathsf{T}}]$. Si oui, la récursion s'arrête, sinon elle continue. Voir la section 5.3 et en particulier la figure 5.2 et la sous-section 5.3.4.

Dans la plupart des cas, l'obtention des circuits prend une petite fraction du temps total – la majeure partie est consacrée aux calculs utilisant les vecteurs souches. Cependant, pour les instances plus grandes, lorsque p dépasse 30, comme dans PERM-8 ou les instances plus grandes de [35], l'algorithme peut être très lent (ne termine pas pour RESONANCE-6). Des méthodes spécialisées [35, 212] se concentrant sur ces instances utilisent des approches différentes puisque l'algorithme 5.5.1 peut ne pas être utilisable. Elles exploitent les symétries combinatoires des instances, via des groupes de symétrie, pour calculer beaucoup moins de circuits tout en obtenant toutes les informations à travers les symétries. Par exemple, dans les instances PERM, lorsque le circuit des vecteurs e_1, e_2 et $(e_1 - e_2)$ est trouvé, par symétrie on pourrait ajouter tous les circuits formés par les triplets $e_i^\perp, e_j^\perp, (e_i - e_j)^\perp$ pour tous $i < j$.

Cette section présente un raffinement simple inspiré de TOPCOM [214, 212] qui peut être appliqué à tout arrangement. Au lieu de calculer un noyau à chaque ajout d'un nouvel indice, et donc de calculer un noyau à chaque niveau de l'arbre, on peut calculer la forme échelonnée de la matrice $V_{:,[1:k]}$ (ou $[V; \tau^\mathsf{T}]_{:,[1:k]}$). Cela signifie que le calcul du noyau est remplacé par la mise à jour de la forme échelonnée, ce qui coûte moins cher. Cependant, comme une matrice supplémentaire est gardée en mémoire, et que le coût d'un noyau est déjà faible, il n'est pas nécessairement plus rapide de calculer les vecteurs souches de cette manière. En effet, si la profondeur de récursion dans l'algorithme 5.5.2 est faible, il peut être préférable de faire quelques calculs d'espaces nuls plutôt que de garder une variable en mémoire. Les comparaisons sont illustrées dans les tableaux A.6 et A.7 sur les instances affines.

Commentons les ratios d'amélioration apportés par l'implémentation de la forme échelonnée, en commençant par le tableau A.6 et les variantes normales. Elle est raisonnablement efficace pour les instances RAND et RATIO et semble s'améliorer lorsque la dimension augmente. Ceci est cohérent puisque les vecteurs souches ont des tailles plus grandes, donc la récursion dans l'algorithme 5.5.2 a plus de niveaux. Inversement, lorsque la récursion est peu profonde comme dans les instances 2D, le calcul d'un espace nul à chaque niveau peut être préférable. Les instances PERM ne bénéficient pas beaucoup du calcul de la forme échelonnée, tandis que les instances SRAND montrent des ratios moyens.

Pour le cas compact dans le tableau A.7, l'amélioration est légèrement moins bonne, ce qui est cohérent puisque les deux temps sont augmentés, donc le ratio diminue. Une règle empirique simple peut être la suivante : lorsque la factorisation QR est calculée, si le rang est très petit ou si la matrice R a beaucoup de blocs (proches de) zéro comme dans les instances 2D, cette caractéristique n'est pas utilisée.

A.4.2 Test de couverture récursive

Remarque A.4.1 (couverture récursive)**.** Les tests de couverture peuvent être effectués de manière récursive. En effet, considérons que le nœud $s \in \mathcal{S}_k$ a été vérifié par un test de couverture : le produit $M_{:,I^s} s$ a été calculé, où M représente une matrice de vecteurs souches

version non compacte							
Problème	$\lvert\mathfrak{S}_s(V,\tau)\rvert$	doublons	$\lvert\mathfrak{S}_a(V,\tau)\rvert$	doublons	temps	échelon	ratio
RAND-8-2	0	0	56	0	0.00131	0.00108	1.21
RAND-8-4	0	0	56	0	0.0127	0.00977	1.30
RAND-9-4	0	0	126	0	0.016	0.0128	1.24
RAND-10-5	0	0	210	0	0.0418	0.0283	1.48
RAND-11-4	0	0	462	0	0.0324	0.0275	1.18
RAND-12-6	0	0	792	0	0.217	0.127	1.71
RAND-13-5	0	0	1716	0	0.177	0.126	1.41
RAND-14-7	0	0	3003	0	0.679	0.358	1.90
RAND-15-7	0	0	6435	0	1.22	0.633	1.93
RAND-16-8	0	0	11440	0	3.12	1.40	2.24
RAND-17-9	0	0	19448	0	7.85	2.95	2.66
SRAND-8-20-2	56	17602	321	53618	6.31	4.88	1.29
SRAND-8-20-4	1185	12044	70650	64704	24.2	14.5	1.67
SRAND-8-20-6	20413	4319	123909	18530	27.8	17.0	1.63
2D-4	1	3	815	2445	0.091	0.0952	0.96
2D-5	0	0	680	4760	0.188	0.20	0.94
2D-6	1	15	559	8385	0.342	0.364	0.94
2D-7	0	0	443	14085	0.614	0.633	0.97
2D-8	0	0	364	22932	1.05	1.08	0.97
PERM-5	197	3179	0	0	0.222	0.182	1.22
PERM-6	1172	56185	0	0	3.49	3.02	1.16
PERM-7	8018	1096176	0	0	77.9	68.1	1.14
PERM-8	62814	23874562	0	0	335	306	1.10
RATIO-3-20-0.7	12	10	3834	0	0.0169	0.0168	1.01
RATIO-3-20-0.9	118	35	4550	0	0.0198	0.0194	1.02
RATIO-4-20-0.7	102	34	15271	0	0.0981	0.0919	1.07
RATIO-4-20-0.9	2327	401	11908	0	0.0911	0.0858	1.06
RATIO-5-20-0.7	58	123	25857	0	0.244	0.206	1.18
RATIO-5-20-0.9	23514	820	10954	0	0.311	0.269	1.16
RATIO-6-20-0.7	238	257	76595	0	0.986	0.767	1.29
RATIO-6-20-0.9	345	317	71861	0	0.887	0.693	1.28
RATIO-7-20-0.7	125	314	123792	0	2.17	1.49	1.45
RATIO-7-20-0.9	154	554	123731	0	2.24	1.55	1.44
Moyenne							1.34
Médiane							1.22

TABLE A.6 – Temps de calcul des vecteurs souches dans les variantes régulières. Les deuxième et quatrième colonnes représentent les nombres de vecteurs souches, les troisième et cinquième colonnes le nombre de doublons. Les trois colonnes restantes indiquent le temps du calcul initial, le temps de calcul avec la forme échelonnée et leur ratio : si supérieur à 1, cela signifie que la forme échelonnée est plus rapide.

(de tout type) et I^s contient les hyperplans déjà traités dans s. Soient $s_+ = (s, +1)$ et $s_- =$

	version compacte						
Problème	$\lvert\mathfrak{S}_s(V,\tau)\rvert$	doublons	$\lvert\mathfrak{S}_a(V,\tau)\rvert$	doublons	temps	échelon	ratio
RAND-8-2	56	0	70	0	0.00194	0.00235	0.82
RAND-8-4	56	0	28	0	0.0142	0.0140	1.01
RAND-9-4	126	0	84	0	0.0191	0.0164	1.16
RAND-10-5	210	0	120	0	0.0527	0.0385	1.37
RAND-11-4	462	0	462	0	0.0526	0.0486	1.08
RAND-12-6	792	0	495	0	0.237	0.174	1.37
RAND-13-5	1716	0	1716	0	0.260	0.225	1.16
RAND-14-7	3003	0	2002	0	0.857	0.511	1.68
RAND-15-7	6435	0	5005	0	1.62	1.03	1.57
RAND-16-8	11440	0	8008	0	3.89	2.16	1.80
RAND-17-9	19448	0	12376	0	9.44	4.33	2.18
SRAND-8-20-2	321	53618	987	57010	10.6	8.00	1.33
SRAND-8-20-4	70650	64704	94534	74917	40.3	29.4	1.37
SRAND-8-20-6	123909	18530	105345	68402	42.9	30.7	1.40
2D-4	815	2445	3046	9182	0.470	0.421	1.12
2D-5	680	4760	2380	16660	0.874	0.730	1.20
2D-6	559	8385	1808	27200	1.60	1.40	1.14
2D-7	443	14085	1365	42315	2.82	2.43	1.16
2D-8	364	22932	1001	63063	4.70	4.19	1.12
PERM-5	0	0	197	3179	0.214	0.185	1.16
PERM-6	0	0	1172	56185	3.54	3.02	1.17
PERM-7	0	0	8018	1096176	77.9	68.4	1.14
PERM-8	0	0	62814	23874562	335	305	1.10
RATIO-3-20-0.7	3834	0	11268	60	0.0708	0.0735	0.96
RATIO-3-20-0.9	4550	0	12993	531	0.0834	0.0850	0.98
RATIO-4-20-0.7	15271	0	36781	192	0.322	0.319	1.01
RATIO-4-20-0.9	11908	0	19882	1993	0.208	0.204	1.02
RATIO-5-20-0.7	25857	0	45278	499	0.60	0.563	1.07
RATIO-5-20-0.9	10954	0	23514	8787	0.374	0.327	1.14
RATIO-6-20-0.7	76595	0	120663	1411	2.19	1.96	1.11
RATIO-6-20-0.9	71861	0	106115	721	1.87	1.68	1.11
RATIO-7-20-0.7	123792	0	159956	1170	4.09	3.34	1.23
RATIO-7-20-0.9	123731	0	159636	2310	4.23	3.51	1.20
Moyenne							1.23
Médiane							1.16

TABLE A.7 – Temps de calcul des vecteurs souches dans les variantes compactes. Les deuxième et quatrième colonnes représentent les nombres de vecteurs souches, les troisième et cinquième colonnes les doublons. Les trois colonnes restantes indiquent le temps du calcul initial, le temps de calcul avec la forme échelonnée et leur ratio : si supérieur à 1, cela signifie que la forme échelonnée est plus rapide.

$(s,-1)$ les descendants de s et i_{k+1} l'indice de l'hyperplan ajouté. Le test de couverture calcule $M_{:,I^{s_+}}s_+ = M_{:,I^s}s + M_{:,i_{k+1}}$ et $M_{:,I^{s_-}}s_- = M_{:,I^s}s - M_{:,i_{k+1}}$. Par conséquent, le

produit matrice-vecteur peut être calculé de manière incrémentale, et à chaque nœud il n'y a qu'une addition/soustraction de vecteurs. Cela est illustré dans le tableau A.8. □

Cependant, il n'est pas évident que cela réduise nécessairement le temps total de calcul. En effet, le produit matrice-vecteur du test de couverture est effectué dans la dimension égale au nombre de signes, l'autre étant le nombre de vecteurs souches qui peut être grand. Le calcul récursif proposé ici doit encore additionner (ou soustraire) deux vecteurs dont la longueur est le nombre de vecteurs souches. Dans les comparaisons numériques ci-dessous, la variante D1 n'est pas montrée car il y a très peu de vecteurs souches. La variante D2 n'est également pas montrée pour la raison suivante : comme le nombre de vecteurs souches augmente, lorsque l'algorithme revient à un sous-arbre non terminé, il faut ajuster la taille du vecteur gardant le produit en mémoire. Bien que cela puisse être implémenté, c'est en effet inefficace et donc non montré ici. Cette méthode récursive est donc uniquement comparée pour les variantes D3 et D4. Avant de discuter des comparaisons, le tableau A.8 présente une illustration de ce calcul récursif.

index	signe	vecteur courant								$\mathfrak{S}$							
1	+	+1	+1	+1	0	+1	+1	0	0	+	+	+	·	+	+	·	·
2	+	+2	+2	+1	+1	+2	+1	+1	0	+	+	·	+	+	·	+	·
3	+	+1	+2	+2	0	+2	+1	+1	+1	−	·	+	−	·	·	+	+
...	...									·	·	·	·	−	+	−	−
...	...									·	−	−	+	·	−	+	·
3	−	+3	+2	0	+2	+2	+1	+1	−1								
...	...																
1	−	−1	−1	−1	0	−1	−1	0	0								
2	+	0	0	−1	+1	0	−1	+1	0								

Table A.8 – Illustration de l'implémentation récursive du test de couverture. Il y a $p = 5$ vecteurs dans $\mathbb{R}^n$, et la matrice des vecteurs souches est donnée à droite sous forme transposée dans la moitié droite. Par exemple, la première colonne signifie que $[v_1\ v_2\ -v_3]$ est de nullité un dans $\mathbb{R}_+^{\{1,2,3\}}$. À gauche, sur la ligne avec index = 1 et signe = +, le vecteur courant est $+M_{:,1}$ (la première ligne de la matrice transposée des vecteurs souches). Sur la ligne suivante, comme le signe est aussi $+$, la deuxième ligne des matrices de vecteurs souches est ajoutée. Sur la ligne avec index = 3 et signe = -, le vecteur courant est donc la première ligne de $\mathfrak{S}$ plus la deuxième moins la troisième. En particulier, la coordonnée 1 du vecteur courant est égale à 3, qui est la taille du premier vecteur souche (première colonne de $\mathfrak{S}$), donc le test de couverture arrête la récursion.

Dans les tableaux A.9 et A.10, nous comparons les temps totaux des variantes D3 et D4 en utilisant les tests de couverture récursifs ou non. Les tests ont été effectués sur certaines des instances linéaires, ce qui explique qu'il n'y ait qu'un seul type de vecteurs souches.

On peut observer que l'efficacité de cette méthode dépend principalement de la cardinalité de $\mathfrak{S}$, ce qui est attendu. Cependant, la nature de l'arrangement importe également : des améliorations très proches sont obtenues pour perm-8 et ratio-7-20-0.7 mais le nombre de

Pour D3	version de base, temps		version récursive, temps		$\lvert\mathcal{C}\rvert$	ratios	
Nom	total	couverture	total	couverture		total	couverture
RAND-4-8-2	0.00313	0.00184	0.00901	0.00548	56	0.35	0.34
RAND-7-8-4	0.0224	0.00507	0.0460	0.0199	56	0.49	0.25
RAND-7-9-4	0.0343	0.00823	0.0749	0.0349	126	0.46	0.24
RAND-7-10-5	0.105	0.0233	0.207	0.0878	210	0.50	0.27
RAND-7-11-4	0.0849	0.0236	0.173	0.0793	462	0.49	0.30
RAND-7-12-6	0.513	0.112	0.894	0.357	792	0.57	0.31
RAND-7-13-5	0.471	0.125	0.805	0.336	1716	0.59	0.37
RAND-7-14-7	2.42	0.662	3.84	1.49	3003	0.63	0.44
RAND-8-15-7	4.30	1.44	6.12	2.43	6435	0.70	0.59
RAND-9-16-8	13.4	5.54	16.2	6.44	11440	0.82	0.86
RAND-10-17-9	40.4	19.8	41.7	17.4	19448	0.97	1.13
2D-4-20	0.0696	0.0397	0.179	0.109	680	0.39	0.36
2D-5-20	0.132	0.0660	0.347	0.195	680	0.38	0.34
2D-6-20	0.270	0.147	0.667	0.391	560	0.41	0.38
2D-7-20	0.621	0.332	1.52	0.889	455	0.41	0.37
2D-8-20	0.795	0.391	2.27	1.37	364	0.35	0.29
SRAND-8-20-2	4.14	0.973	9.35	4.22	540	0.44	0.23
SRAND-8-20-4	202	159	141	85.7	84390	1.44	1.85
SRAND-8-20-6	466	399	258	182	160074	1.81	2.19
PERM-5	0.0853	0.0369	0.307	0.176	197	0.28	0.21
PERM-6	1.03	0.336	2.56	1.29	1172	0.40	0.26
PERM-7	21.2	7.60	28.8	11.1	8018	0.74	0.69
PERM-8	992	645	646	257	62814	1.53	2.51
RATIO-3-20-0.7	0.227	0.119	0.364	0.197	3486	0.62	0.60
RATIO-3-20-0.9	0.141	0.0776	0.305	0.174	1332	0.46	0.45
RATIO-4-20-0.7	2.23	1.44	2.21	1.34	15138	1.01	1.27
RATIO-4-20-0.9	1.85	1.18	1.96	1.01	14052	0.94	1.17
RATIO-5-20-0.7	11.4	8.57	8.64	4.83	34556	1.33	1.77
RATIO-5-20-0.9	10.4	7.52	7.92	4.38	31334	1.31	1.72
RATIO-6-20-0.7	48.9	39.1	28.9	17.0	56184	1.70	2.30
RATIO-6-20-0.9	52.2	42.2	33.9	21.1	36970	1.54	2
RATIO-7-20-0.7	232	201	117	80.3	112576	1.98	2.51
RATIO-7-20-0.9	131	108	73.0	45.4	74970	1.79	2.38

TABLE A.9 – Temps d'exécution pour les instances linéaires avec l'option D3. Les colonnes 2-3 représentent les temps totaux et de couverture avec le produit matrice-vecteur complet dans le test. Les colonnes 4-5 représentent les temps totaux et de couverture effectués de manière récursive. La colonne 6 donne le nombre de vecteurs souches. Les colonnes 7-8 montrent les ratios pour les temps totaux et les temps de couverture.

tests de couverture effectués varie. Une règle empirique possible est choisie comme suit : s'il y a beaucoup de vecteurs (p grand), ou si le ratio du nombre de vecteurs souches divisé par le nombre maximal de vecteurs souches $\binom{p}{r+1}$ est élevé, alors cette méthode est utilisée.

Pour D4	version de base, temps		version récursive, temps		$\lvert\mathcal{C}\rvert$	ratios	
Nom	total	couverture	total	couverture		total	couverture
rand-4-8-2	0.00803	0.00441	0.014	0.00837	56	0.57	0.53
rand-7-8-4	0.0279	0.0171	0.0453	0.0277	56	0.62	0.62
rand-7-9-4	0.0516	0.0320	0.0820	0.0504	126	0.63	0.63
rand-7-10-5	0.122	0.0754	0.203	0.122	210	0.60	0.62
rand-7-11-4	0.132	0.0861	0.206	0.129	462	0.64	0.67
rand-7-12-6	0.565	0.359	0.885	0.552	792	0.64	0.65
rand-7-13-5	0.533	0.346	0.869	0.551	1716	0.61	0.63
rand-7-14-7	2.42	1.68	3.31	2.14	3003	0.73	0.79
rand-8-15-7	4.57	3.4	5.52	3.68	6435	0.83	0.92
rand-9-16-8	15.6	12.6	14.1	9.63	11440	1.10	1.30
rand-10-17-9	49.1	41.9	35.3	25.1	19448	1.39	1.67
2d-20-4	0.173	0.106	0.303	0.190	680	0.57	0.56
2d-20-5	0.359	0.214	0.623	0.391	680	0.58	0.55
2d-20-6	0.652	0.40	1.23	0.779	560	0.53	0.51
2d-20-7	1.11	0.658	2.17	1.41	455	0.51	0.47
2d-20-8	1.92	1.14	4.03	2.60	364	0.48	0.44
srand-8-20-2	10.5	6.30	19.3	12.5	540	0.54	0.50
srand-8-20-4	371	350	202	172	84390	1.83	2.03
srand-8-20-6	890	863	404	364	160074	2.21	2.37
perm-5	0.375	0.227	0.691	0.431	197	0.54	0.53
perm-6	3.81	2.27	6.98	4.34	1172	0.55	0.52
perm-7	63.7	42.6	65.8	37.0	8018	0.97	1.15
perm-8	3518	3081	1593	1058	62814	2.21	2.90
ratio-3-20-0.7	0.364	0.253	0.532	0.339	3486	0.68	0.69
ratio-3-20-0.9	0.177	0.111	0.318	0.196	1332	0.56	0.57
ratio-4-20-0.7	3.26	2.70	3.01	2.06	15138	1.08	1.31
ratio-4-20-0.9	2.87	2.34	3.14	2.16	14052	0.91	1.08
ratio-5-20-0.7	18.6	16.7	12.6	9.35	34556	1.48	1.78
ratio-5-20-0.9	15.1	13.3	10.6	7.77	31334	1.42	1.71
ratio-6-20-0.7	75.7	70.2	40.3	32.1	56184	1.88	2.18
ratio-6-20-0.9	92.9	86.9	48.9	40.1	63970	1.90	2.17
ratio-7-20-0.7	374	360	160	140	112576	2.34	2.58
ratio-7-20-0.9	179	169	87.4	73.0	74970	2.05	2.32

Table A.10 – Temps d'exécution pour les instances linéaires avec l'option D4. Les colonnes 2-3 représentent les temps totaux et de couverture avec le produit matrice-vecteur complet dans le test. Les colonnes 4-5 représentent les temps totaux et de couverture effectués de manière récursive. La colonne 6 donne le nombre de vecteurs souches. Les colonnes 7-8 montrent les ratios pour les temps totaux et les temps de couverture.

Un dernier commentaire

Un autre aspect intéressant serait d'étudier le comportement des algorithmes pour des instances combinatoires *affines*. On pourrait imaginer ajouter des perturbations ($\tau \neq 0$)

aux instances threshold, resonance, or demicube, ou considérer les arrangements traités par exemple dans [203].

Annexe B

Éléments de géométrie sur les polytopes

Cette annexe décrit quelques propriétés utiles sur les polytopes, ainsi que quelques-unes inhérentes aux zonotopes. La plupart sont, aux notations et conventions près, des propriétés basiques et bien connues.

Dans ce qui suit, on nomme "faces" pour décrire les faces de dimension quelconque d'un polytope. Pour insister, on utilise "de dimension maximale" pour décrire ce qui est parfois appelé "facet(te)s" en géométrie combinatoire.

B.1 Polytopes et leurs faces

Dans ce qui suit, on utilise la formulation H des polytopes (convexes), i.e., leur représentaiton comme une intersection finie de demi-espaces de la forme $P = \{x \in \mathbb{R}^n : Ax \leqslant a, Bx = b\}$. Le premier lemme décrit les intérieurs relatifs de tels polyèdres.

Lemme B.1.1 (intérieur relatif des polyèdres)**.** *Soit $P = \{x \in \mathbb{R}^n : Ax \leqslant a, Bx = b\}$ pour $A \in \mathbb{R}^{m\times n}$ et $a \in \mathbb{R}^m$. Supposons que pour tout $i \in [1 : m]$, il existe un $x^i \in P$ tel que $(Ax^i - a)_i < 0$, i.e., aucune inégalité n'est en fait une égalité. Alors* $\mathrm{ri}(P) = \{x \in \mathbb{R}^n : Ax < a, Bx = b\}$.

Preuve. [$\subseteq$] Soit $x \in \mathrm{ri}(P)$, clairement $x \in P$ donc $Bx = b$. Supposons qu'il existe un i tel que $(Ax - a)_i = 0$. Alors, en utilisant le $x^i \in P$ tel que $(Ax^i - a)_i < 0$, on sait qu'il existe un $\delta^i > 0$ tel que $x + \delta^i(x - x^i) \in P$, par définition de l'intérieur relatif. Cependant, en regardant l'indice i, on a

$$\begin{aligned}(A(x + \delta^i(x - x^i)) - a)_i &= (Ax - a + \delta^i(Ax - Ax^i - a + a))_i \\ &= (Ax - a + \delta^i((Ax - a) - (Ax^i - a)))_i \\ &= 0 + \delta^i(0 - (Ax^i - a)_i) > 0,\end{aligned}$$

qui est une contradiction avec la définition de P.

[$\supseteq$] Soit x tel que $Ax < a$ et $Bx = b$, et $x_0 \in P$, on montre qu'il existe un $\delta > 0$ tel que $x + \delta(x - x_0)$ est dans P. Alors, $B(x + \delta(x - x_0)) = b + \delta(b - b) = b$. De plus,

$$A(x + \delta(x - x_0)) - a = (Ax - a) + \delta(Ax - a - (Ax_0 - a))$$

qui est < 0 pour δ assez petit puisque $(Ax - a) < 0$. □

Le lemme suivant décrit les faces, dont on rappelle la définition, d'une façon manipulable.

Définition B.1.2 (face d'un polyèdre)**.** Une face F d'un polytope P est un sous-ensemble de P tel que pour tout $x, y \in P$ et tout $t \in (0, 1)$, $(1 - t)x + ty \in F \Rightarrow x, y \in F$. □

Lemme B.1.3 (faces et sous-ensembles d'indices)**.** *Soit $P = \{x \in \mathbb{R}^n : Ax \leqslant a, Bx = b\}$ un polyèdre convexe, F est une face de P si et seulement s'il existe un sous-ensemble d'indices I tel que $F = \{x \in P : (Ax - a)_I = 0\}$.*

Preuve. Si $F = \{x \in P : (Ax - a)_I = 0\}$, soit $x_1, x_2 \in P$ tel que $(x_1 + x_2)/2 \in F$. Cela s'écrit $(A(x_1 + x_2)/2 - a)_I = 0$, mais puisque $(Ax_1 - a)_I \leqslant 0$ et $(Ax_2 - a)_I \leqslant 0$, les deux quantités doivent s'annuler, indiquant que x_1 et x_2 sont dans F.

Inversement, puisque F est un convexe non vide, prenons x_0 dans son intérieur relatif et définissons $I = \{i : (Ax_0)_i = a_i\}$. Soit $x \in F \subseteq P$, par définition de $\mathrm{ri}(F)$, $(1-t)x_0 + tx \in F$ pour un certain $t > 1$. Maintenant,

$$a_I \geqslant (A((1-t)x_0 + tx))_I = (1-t)(Ax_0)_I + t(Ax)_I \iff 0 \geqslant (1-t)(Ax-a)_I + t(Ax_0 - a)_I,$$

où $(Ax_0 - a)_I = 0$ et $1 - t < 0$. De fait, on doit avoir $(Ax - a)_I = 0$ puisque $(Ax - a)_I \leqslant 0$. De même, si $x \in P$ vérifie $(Ax - a)_I = 0$, montrons qu'il appartient à F. Puisque F est une face, il suffit de montrer que $x_t = (1 - t)x_0 + tx \in P$ pour un t négatif petit, parce qu'alors x_0 est une combinaison convexe de x_t et x (qui est clairement dans F, ce qui implique que x_t et x sont dans F). Clairement, $(Ax_t - a)_I = (1-t)(Ax_0 - a)_I + t(Ax - a)_I = (1-t)0 + t0$. Pour les indices restants, puisque $(Ax_0 - a)_i < 0$ pour t assez proche de zéro on a toujours $(Ax_t - a)_i < 0$. □

Les lemmes B.1.1 et B.1.4, qui vient, sont illustrés en figure B.1.

Lemme B.1.4 (décomposition en intérieurs relatifs des faces)**.** *Soit $P := \{x \in \mathbb{R}^n : Ax \leqslant a, Bx = b\}$ un polyèdre, on a*

$$P = \bigcup_{F \in \mathrm{faces}(P)} \mathrm{ri}(F),$$

où l'union est disjointe et contient P lui-même comme une face.

Preuve. D'abord, justifions que l'union est disjointe. Soit F^1 et F^2 deux faces, en utilsant le lemme B.1.3 notons I_1 et I_2 les ensembles d'indices correspondants, i.e.,

$$F^1 = \{x \in \mathbb{R}^n : Bx = b, (Ax - a)_{I_1} = 0, (Ax - a)_{I_1^c} \leqslant 0\}$$
$$F^2 = \{x \in \mathbb{R}^n : Bx = b, (Ax - a)_{I_2} = 0, (Ax - a)_{I_2^c} \leqslant 0\}.$$

Ensuite, avec le lemme B.1.1, on a

$$\mathrm{ri}(F^1) = \{x \in \mathbb{R}^n : Bx = b, (Ax - a)_{I_1} = 0, (Ax - a)_{I_1^c} < 0\}$$
$$\mathrm{ri}(F^2) = \{x \in \mathbb{R}^n : Bx = b, (Ax - a)_{I_2} = 0, (Ax - a)_{I_2^c} < 0\}.$$

Maintenant, suppsons qu'il existe $x \in \mathrm{ri}(F^1) \cap \mathrm{ri}(F^2)$. Alors, pour les indices dans $I_1 \cap I_2^c$ et $I_2 \cap I_1^c$, on doit avoir $(Ax - a)_i < 0$ et $(Ax - a)_i = 0$, ce qui est une contradiction à moins que $I_1 = I_2$, i.e., $F^1 = F^2$.

[$\subseteq$] Soit $x \in P$ et fixons $I_0^x := \{i : (Ax - a)_i = 0\}$ ainsi que $I_*^x := \{i : (Ax - a)_i < 0\}$. Soit F^x la face définie par le sous-ensemble I_0^x, on a $x \in \mathrm{ri}(F^x)$.

[$\supseteq$] Clair puisque $x \in \mathrm{ri}(F) \subseteq F \subseteq P$. □

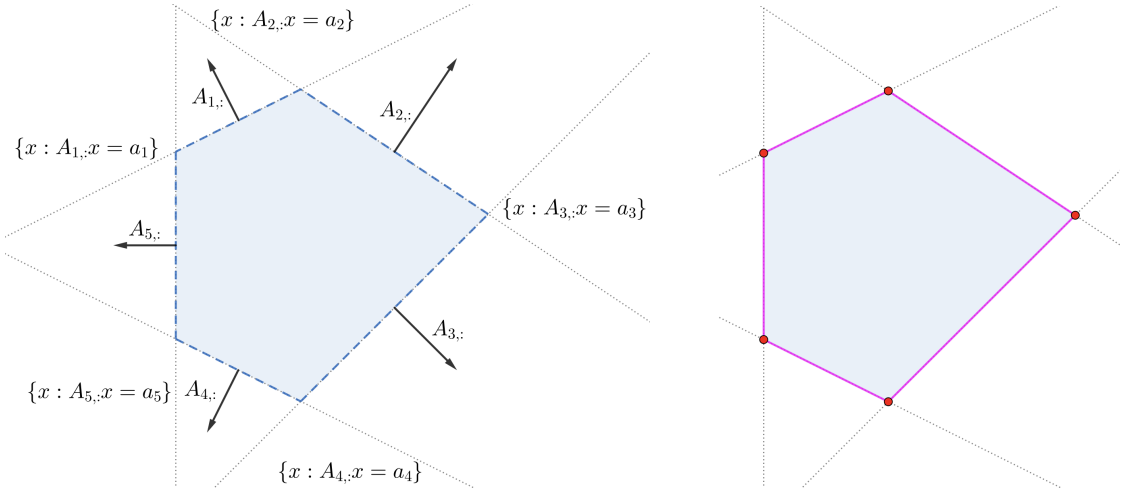


Figure B.1 – Illustration des lemmes B.1.1 (gauche) et B.1.4 (droite). À gauche, on voit que les intérieurs relatifs sont obtenus en retirant les parties de P où des égalités $A_{:,i}x = a_i$ sont vraies. Cependant, imaginons le même polytope en dimension 3 (donc avec un intérieur vide), avec les inégalités supplémentaires $e_3^\mathsf{T} x \leqslant 0$ et $-e_3^\mathsf{T} x \leqslant 0$, on ne peut prendre les inégalités strictes puisque l'on aurait un ensemble vide. C'est parce que ces inégalités forment en fait une égalité. À droite, on voit que l'intérieur relatif en bleu corresond à l'intérieur relatif de P (en tant que face de lui-même), alors que la frontière est composée est intérieurs relatifs des faces en magenta et des sommets en rouge.

Les lemmes suivants définissent et décrivent les propriétés des *vecteurs normaux*, appelés "normales", aux faces.

Lemme B.1.5 (faces et "normales"). *Soit P un polyèdre convexe, et $F \subseteq P$, F est une face (de dimension quelconque) si et seulement s' il existe un $c \in \mathbb{R}^n$ tel que $F = \operatorname{argmin}\{c^\mathsf{T} x : x \in P\}$. En particulier, $F = P \cap H$ où $H = c^\perp + x_F$ pour tout $x_F \in F$.*

Preuve. S'il existe un tel c, pour tout $x_1, x_2 \in P$, on a $c^\mathsf{T} x \geqslant c^\mathsf{T} x_1$ et $c^\mathsf{T} x \geqslant c^\mathsf{T} x_2$ pour tout $x \in P$. Quand $c^\mathsf{T}(x_1 + x_2)/2 = \alpha := \min\{c^\mathsf{T} x : x \in P\}$, on doit avoir $c^\mathsf{T} x_1 = \alpha = c^\mathsf{T} x_2$, donc $x_1, x_2 \in F$.

Inversement, soit $F = \{x \in P, (Ax)_I = a_I\}$ pour un ensemble d'indices I par le lemme B.1.5. On doit trouver un c tel que $F = \operatorname{argmin}\{c^\mathsf{T} x : x \in P\}$. Les conditions d'opitmalité sont $c = -A^\mathsf{T}\mu$, $\mu^\mathsf{T}(Ax - a) = 0$. Pour I^c, puisque $Ax < a$, on a $\mu_{I^c} = 0$. Avec $\mu_I > 0$ soit $\alpha = -a^\mathsf{T}\mu$. Pour $x \in P$, on a $c^\mathsf{T} x = -\mu^\mathsf{T} Ax \geqslant -\mu_I^\mathsf{T}(Ax)_I = \alpha$. Pour $x \in F$, $c^\mathsf{T} x = \alpha$ donc $F \subseteq \operatorname{argmin}\{c^\mathsf{T} x : x \in P\}$. Inversement, pour un $x' \in \operatorname{argmin}\{c^\mathsf{T} x : x \in P\}$, et pour tout $x \in P$, on a $c^\mathsf{T} x \geqslant c^\mathsf{T} x'$, qui s'écrit $0 \geqslant \mu^\mathsf{T} A(x - x')$. Pour $x \in F$, on a $0 \geqslant \mu^\mathsf{T}(a - Ax')$, et $\mu_I > 0$ implique $(Ax')_I = a_I$.

Maintenant, soit $\alpha := \min\{c^\mathsf{T} x : x \in P\}$, $\operatorname{argmin}\{c^\mathsf{T} x : x \in P\} = P \cap \{x : c^\mathsf{T} x = \alpha\} = P \cap H$. □

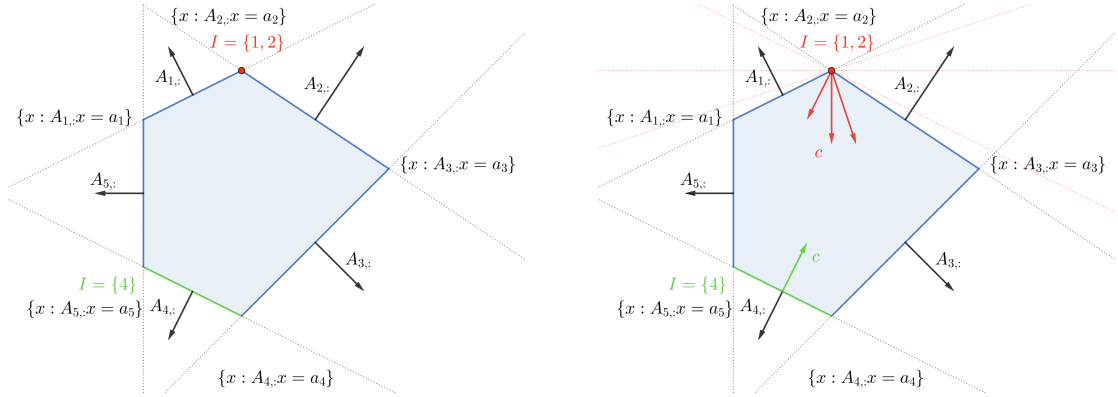


FIGURE B.2 – Illustration des lemmes B.1.3 (gauche) et B.1.5 (droite). À gauche, on peut observer que la face verte correspond à l'ensemble $I = \{4\}$ de taille 1, alors que les sommet en rouge correspond à un ensemble $I = \{1, 2\}$ de taille 2. À droite, la même face verte a une unique (à une constante > 0 multiplicative près) normale c puisque la face est de dimension maximale $n - 1$ alors que le sommet en rouge a de multiples normales non colinéaires possibles c (voir la remarque B.1.6).

Remarque B.1.6 (faces de dimension maximale et autres faces). Soit $P \subseteq \mathbb{R}^n$ un polyèdre, si une face F est de dimension $n - 1$ alors $\operatorname{aff}(F)$ est un hyperplan et c est unique à une constante strictement positive près. Cela provient du fait que $I = \{i\}$ pour un indice i (ou il y a des contraintes redondantes). Lorsque la face est de plus petite dimension, I n'est pas réduit à un singleton, et comme n'importe quel $\mu_I > 0$ dans la preuve du lemme B.1.5 peut convenir, il n'y a pas unicité. □

Observons que dans le lemme précédent, on aurait pu utiliser argmax à la place. Cela rajoute surtout un signe moins aux vecteurs c.

Proposition B.1.7 (normale sortante)**.** *Soit $P \subseteq \mathbb{R}^n$ un polyèdre convexe, $L = \mathrm{aff}(P) - \mathrm{aff}(P)$ le sous-espace parallèle à son enveloppe affine, F une face de P et $L_F = \mathrm{aff}(F) - \mathrm{aff}(F)$. Il existe $c \in L$ telle que $||c|| = 1$, $c \in L_F^\perp$ et c est sortant, i.e., $x_F + tc \notin P$ pour $x_F \in F$ et $t > 0$.*

Preuve. Soit d le vecteur obtenu dans le lemme B.1.5. Définissons $d = c_L + c_\perp$ sa décomposition dans $L + L^\perp$. Puisque $P \subseteq x_P + L$ pour un $x_P \in P$, on a

$$\begin{aligned} F = \mathrm{argmin}\{d^\intercal x : x \in P\} &= \mathrm{argmin}\{c_L^\intercal x + c_\perp^\intercal x : x \in P\} \\ &= \mathrm{argmin}\{c_L^\intercal x + c_\perp^\intercal x_P : x \in P\} = \mathrm{argmin}\{c_L^\intercal x : x \in P\} \end{aligned}$$

Si $c_L = 0$, alors $F = P$. Sinon, soit $c := -c_L/||c_L||$. Clairement $||c|| = 1$ et $c \in L$. Soit $v \in L_F$, on a $v = v_1 - v_2$ pour un v_1 et un v_2 dans $\mathrm{aff}(F)$. Ensuite, en écrivant $v_i = (1 - t_i)x_i + t_i x_i'$ pour un x_i et un x_i' dans F (puisque c'est convexe) pour $i \in \{1, 2\}$. On a alors

$$c^\intercal v = c^\intercal((1-t_1)x_1 + t_1 x_1' - (1-t_2)x_2 - t_2 x_2') = -\frac{1}{||c_L||}[(1-t_1)\alpha + t_1\alpha - (1-t_2)\alpha - t_2\alpha] = 0$$

où α est la valeur optimale de la définition de F. Maintenant, montrons que c est sortant. Soit $x_F \in F$ et $t > 0$, $c_L^\intercal(x^F + tc) = \alpha + tc_L^\intercal(-c_L/||c_L||) = \alpha - t||c_L|| = \alpha - t < \alpha$, donc $x^F + tc \notin P$. □

Essentiellement, cette propriété prend l'opposé, projette sur L puis norme la normale décrite dans le lemme B.1.5. Cette discussion est reliée à la notion de "normal fan" [263, p. 193], qui est la décomposition de $\mathbb{R}^n$ en les (intérieurs des) cônes normaux aux faces.

B.2 Propriétés spécifiques des zonotopes

Les propriétés suivantes se concentrent sur les zonotopes, qui sont des polytopes particuliers. Après avoir rappelé la définition d'un zonotope, on se concentre sur leurs faces, la propriété idoine étant illustrée en figure B.3. La preuve est de [167].

Définition B.2.1 (zonotopes)**.** Soit $V \in \mathbb{R}^{n \times m}$ et $\overline{z} \in \mathbb{R}^n$, le zonotope $Z(V, \overline{z})$ est un polytope convexe défini par

$$Z(V, \overline{z}) := V[-1, +1]^m + \overline{z}.$$

En particulier, c'est un compact centro-symétrique autour de $\overline{z}$ ($z + \overline{z} \in Z(V, \overline{z}) \iff \overline{z} - z \in Z(V, \overline{z})$). Souvent, on considère que $\overline{z} = 0$. □

Proposition B.2.2 (***k***-faces d'un zonotope). *Soit $V \in \mathbb{R}^{n\times m}$ et $Z = V[-1,+1]^m$, $V = [v_1 \dots v_m]$, les faces de Z peuvent s'exprimer comme $F = V_{:,I^F}[-1,+1]^{I^F} + V_{:,I^*}\kappa$ pour un $\kappa \in \{-1,+1\}^{I^*}$ où I^F sont les indices des vecteurs générateurs de la face et $I^* = [1:m]\backslash I^F$. En particulier, les k-faces (faces de dimension k) d'un zonotope sont des zonotopes elles-mêmes.*

Preuve. Puisque Z est un polytope convexe, pour une face F donnée, soit c une normale donnée par la proposition B.1.7, $H = c^\perp + z^F$ pour un $z^F \in F$, i.e., $H = \{x \in \mathbb{R}^n : c^\mathsf{T}x = c^\mathsf{T}z^F\}$. En particulier, on a $F = Z \cap H$ et $Z \subseteq H^-$. Soit $H_0 = H - H$ et $I^F := \{i \in [1:m] : v_i \in H_0\}$. Définissons κ_i tel que $\kappa_i v_i \in \operatorname{int} H_0^+$ pour $i \in I^* := [1:m]\backslash I^F$. Pour $z \in F = Z \cap H$, $c^\mathsf{T}z = \alpha$ s'écrit

$$\alpha = c^\mathsf{T}z = c^\mathsf{T}\sum t_i v_i = c^\mathsf{T}\sum_{i\in I^*} t_i v_i \underset{c^\mathsf{T}t_iv_i \leqslant c^\mathsf{T}\kappa_i v_i}{\leqslant} c^\mathsf{T}\sum_{i\in I^*} v_i\kappa_i = c^\mathsf{T}V_{:,I^*}\kappa$$

Cependant, $V_{:,I^*}\kappa \in Z$ donc $c^\mathsf{T}V_{:,I^*}\kappa \leqslant \alpha$. Cela signifie que $t_i = \kappa_i$, donc z a la forme demandée. Réciproquement, un point $z \in V_{:,I^F}[-1,+1]^{I^F} + V_{:,I^*}\kappa$ est clairement dans Z. De plus, $c^\mathsf{T}z = c^\mathsf{T}V_{:,I^*}\kappa = \alpha$ en utilisant la définition de κ. □

Si le zonotope n'est pas centré, i.e., $Z = \overline{z} + V[-1,+1]^m$, la propriété est valide après translation de $\overline{z}$. C'est vrai pour la prochaine proposition, qui discute d'une propriété technique des normales aux faces (c dans les propositions précédentes).

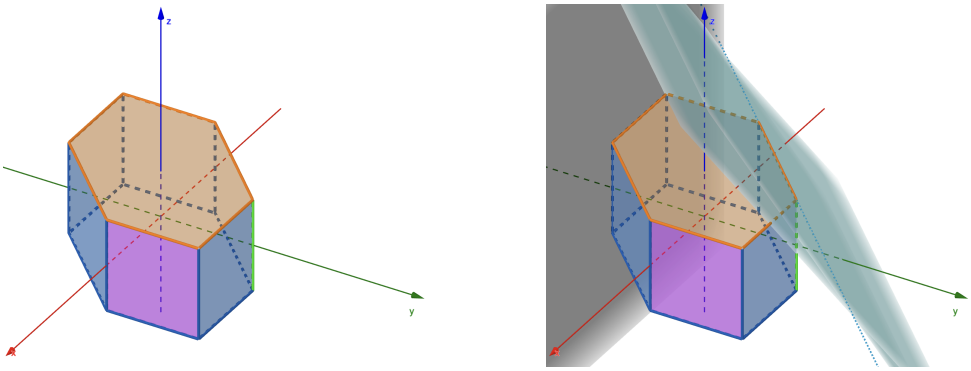


FIGURE B.3 – Exemple d'un zonotope simple avec $V = [e_1\ e_2\ e_1 + e_2\ e_3]$. La face supérieure en orange est une face de dimension deux, générée par $e_1, e_2, e_1 + e_2$, avec $I^* = \{4\}$. La face en violet en avant est générée par e_2 et e_3, avec $I^* = \{1,3\}$. La face verte à droite, qui est une arête, est générée par e_3 avec $I^* = \{1,2,3\}$. tous les sommets sont aussi des faces sans générateurs. À droite, certains hyperplans orthogonaux aux normales sont ajoutés. Les faces de dimension maximale ont un seul hyperplan mais les arêtes en ont plusieurs (puisque la dimension est 3).

La proposition suivante est possiblement nouvelle, mais assez niche et spécifique. Elle utilise une normale particulière donnée par la proposition B.1.7 qui est unique (à une constante près) pour les faces de dimension maximale (dimension $n-1$ dans un espace de dimension n). Une illustration est proposée à la figure B.4.

Proposition B.2.3 (propriété de la normale spécifique). *Soit $V \in \mathbb{R}^{n\times m}$, $Z = V[-1,+1]^m$ un zonotope, F une face de Z, I^F l'ensemble des générateurs de cette face, I^* et $\kappa \in \{\pm 1\}^{I^*}$ tel que le centre de F est $V_{:,I^*}\kappa$ comme indiqué par la proposition B.2.2. Soit c une "normale" à F donnée par la proposition B.1.7, alors*

$$\kappa \cdot V_{:,I^*}^\mathsf{T} c > 0, \quad \text{i.e.,} \quad \forall\, i \in I^*, \kappa_i v_i^\mathsf{T} c > 0.$$

Preuve. Observons que pour $i \in I^*$, $\kappa_i v_i^\mathsf{T} c$ ne peut être zéro puisque cela implique $v_i^\mathsf{T} c = 0$ donc $\mathrm{Vect}(v_i) \subseteq c^\perp = L$ (dans $\mathcal{R}(V)$) donc $i \in I^F$. Suposons par contradiction qu'il existe une partition (I_-^*, I_+^*) de I^* telle que $\kappa_{I_-^*} \cdot V_{:,I_-^*}^\mathsf{T} c < 0$ et $\kappa_{I_+^*} \cdot V_{:,I_+^*}^\mathsf{T} c > 0$ avec I_-^* non vide (sinon le résultat est montré). Alors, le point

$$z_* := V_{:,I_+^*}\kappa_{I_+^*} - V_{:,I_-^*}\kappa_{I_-^*} = V_{:,I_+^*}\kappa_{I_+^*} + V_{:,I_-^*}\kappa_{I_-^*} - 2V_{:,I_-^*}\kappa_{I_-^*} = z^F - 2V_{:,I_-^*}\kappa_{I_-^*}$$

est dans Z car s'écrivant $z_* = V\eta$ pour un $\eta \in [-1,+1]^m$. Cependant, $c^\mathsf{T} z_* = c^\mathsf{T} z^F - 2\sum_{i\in I_-^*} \kappa_i v_i^\mathsf{T} c > c^\mathsf{T} z^F$, qui contredit la définition de c. Alors, I_-^* est vide, ce qui termine la preuve. □

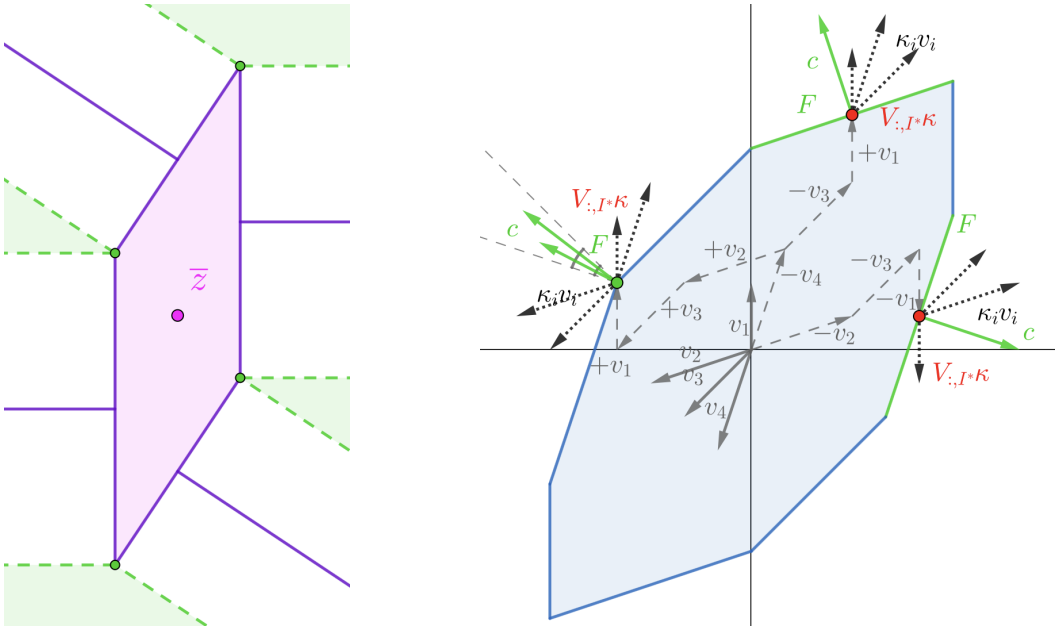


Figure B.4 – Gauche : "normal fan" d'un zonotope. Vert clair : sommets et leurs cônes normaux (frontières en pointillés puisque'elles correspondent aux normales des arêtes) ; violet : arêtes et leurs cônes normaux. Droite : Illustration de la proposition B.2.3. Ensembles et flèches en vert : faces et leurs normales. Points rouges : centres des faces (égaux aux faces pour les sommets). Gris : vecteurs générant le zonotope. Noir : les flèches pointillées sont les $V_{:,i}\kappa_i$ pour $i \in I^*$. On a translaté en le centre des faces pour voir plus facilement que les normales c en vert de la proposition B.1.7 ont un produit scalaire positif avec les flèches pointillées en noir.

Exemple B.2.4 (proposition B.2.3 pour faces quelconques)**.** Pour le zonotope de la figure B.3, et sa face supérieure en orange, sa normale est $c = e_3$, unique à une multiplication par un scalaire strictement positif puisque la face est de dimension maximale. C'est également vrai pour la face violette, de normale $c = e_1$. Cependant, pour l'arête verte, tout c tel que $c_3 = 0$, $c_1 < 0$ et $c_2 + c_1 > 0$ vérifiera les propriétés requises, puisque

$$F_{\text{green}} = [0; 2; 0] + e_3[-1, 1] = -e_1 + e_2 + (e_1 + e_2) + e_3[-1, 1].$$

De fait, $-e_1^\mathsf{T} c > 0$, $(e_1 + e_2)^\mathsf{T} c > 0$ et $e_2^\mathsf{T} c > 0$. □

Remarque B.2.5 (proposition B.2.3 pour les sommets)**.** Lorsque la face F considérée est un sommet, tout point témoin d du sommet est une normale acceptable. Cela provient du fait que les sommets sont des faces sans générateurs, donc $I^* = [1 : m]$. De plus, les normales peuvent être prises dans l'intérieur du cône des directions vérifiant le sommet, puisque cela s'écrit $\kappa_i v_i^\mathsf{T} c > 0$, ou de façon équivalente "$s_i v_i^\mathsf{T} d > 0$" (chapitre 3). □

Finissons par observer une propriété assez naturelle des zonotopes.

Proposition B.2.6 (intérieur relatif des zonotopes)**.** *Soit $Z = \overline{z} + V[-1, +1]^m$ pour $\overline{z} \in \mathbb{R}^n$ et $V \in \mathbb{R}^{n\times m}$. Un point z est dans l'intérieur relatif de Z si et seulement s il s'écrit $z = \overline{z} + V\kappa$, $\kappa \in (-1, +1)^m$.*

Preuve. Bien qu'une preuve directe soit possible, comme les transformations affines et l'intérieur relatif commutent ([105, proposition 2.17 3)$_1$]), on a

$$\text{ri}(c + V[-1, +1]^m) = c + V\text{ri}([-1, +1]^m) = c + V(-1, +1)^m.$$ □

En particulier, la proposition précédente peut être utilisée sur les faces d'un zonotope puisque ce sont des zonotopes. Terminons par discuter de la projection sur un zonotope. Le but de cette explication provient du chapitre 6.

Remarque B.2.7 (projection sur un zonotope)**.** Soit $c \in \mathbb{R}^n$, $Z \in \mathbb{R}^{n\times m}$ et considérons le zonotope $\overline{z} + Z[-1, +1]^m$. La projection du point $x \in \mathbb{R}^n$ sur $\overline{z} + Z[-1, +1]^m$ s'écrit

$$\min_{z\in\overline{z}+Z[-1,+1]^m} \frac{1}{2}||z - x||^2 = \min_{\xi\in[-1,+1]^m} \frac{1}{2}||\overline{z} - x + Z\xi||^2$$

qui est un problème de moindres-carrés avec des contraintes de bornes.

Proposition B.2.8 (projection et normale)**.** *Soit $Z = \overline{z} + V[-1, +1]^m$ pour $\overline{z} \in \mathbb{R}^n$ et $V \in \mathbb{R}^{n\times m}$ et $p \in \mathbb{R}^n$. Définissons $z = \text{P}_Z(p)$ la projection de p sur Z, alors la projection de $p - z$ sur $\mathcal{R}(V)$, $\text{P}_{\mathcal{R}(V)}(p - z)$, vérifie les inégalités des propositions B.1.7 and B.2.3 avec des inégalités larges.*

Preuve. Puisque Z est un convexe non vide, la projection $\text{P}_Z(p)$ est bien définie par le problème

$$\min_{z\in Z} \frac{||z - p||^2}{2} = \min_{\xi\in[-1,+1]^m} \frac{||\overline{z} - p + Z\xi||^2}{2}$$

qui a clairement des contraintes qualifiées. Le système de KKT s'écrit

$$\begin{cases} Z^\mathsf{T}(Z\xi + \overline{z} - p) + \lambda - \mu = 0, \\ \lambda^\mathsf{T}(\xi - 1) = 0, \\ \mu^\mathsf{T}(-1 - \xi) = 0. \end{cases}$$

Ensuite, notons $I_-^* := \{i \in [1:m] : \xi_i = -1\}$ et $I_+^* := \{i \in [1:m] : \xi_i = +1\}$. Par les conditions de complémentarité, on a $i \in I_-^* \Rightarrow \mu_i = 0$ et $i \in I_+^* \Rightarrow \lambda_i = 0$. Le système KKT devient alors

$$\begin{cases} i \in I_+^*, & z_i^\mathsf{T}(p - \overline{z} - Z\xi) = \lambda_i, \\ i \in I_-^*, & z_i^\mathsf{T}(p - \overline{z} - Z\xi) = -\mu_i, \\ \xi_i \in (-1,+1), & z_i^\mathsf{T}(p - \overline{z} - Z\xi) = 0, \end{cases} \Leftrightarrow \begin{cases} i \in I_+^*, & z_i^\mathsf{T}(p - \overline{z} - Z\xi) \geqslant 0, \\ i \in I_-^*, & z_i^\mathsf{T}(p - \overline{z} - Z\xi) \leqslant 0, \\ \xi_i \in (-1,+1), & z_i^\mathsf{T}(p - \overline{z} - Z\xi) = 0. \end{cases}$$

Observons que, comme décrit dans le contre-exemple B.2.9, il n'y a pas nécessairement stricte complémentarité, i.e., on n'a pas $i \in I_+^* \iff \lambda_i > 0$ et $i \in I_-^* \iff \mu_i < 0$. Ensuite, écrivons le système comme

$$\begin{cases} \xi_i \in \{-1,+1\}, & \xi_i z_i^\mathsf{T}(p - \overline{z} - Z\xi) \geqslant 0, \\ \xi_i \in (-1,+1), & z_i^\mathsf{T}(p - \overline{z} - Z\xi) = 0. \end{cases}$$

De plus, soit $c := p - \overline{z} - Z\xi$ la direction de la projection vers le point projeté, c vérifie les inégalités larges de la normale. Pour terminer, observons que diviser le vecteur par sa norme ne change pas les inégalités. De plus, la projection sur $\mathcal{R}(V)$ maintient les inégalités : soit $p - z = \mathrm{P}_{\mathcal{R}(V)}(p - z) + [p - z - \mathrm{P}_{\mathcal{R}(V)}(p - z)]$, par définition de la projection orthogonale sur un sous-espace vectoriel, dans les produits définissant les inégalités le second terme disparaît (puisque $v_i \in \mathcal{R}(V)$ en est perpendiculaire), donc le seul terme restant est $\mathrm{P}_{\mathcal{R}(V)}(p - z)$. Finalement, il est clairement sortant. □

Observons que la projection sur un zonotope est un problème de moindres-carrés avec contraintes. Le contre-exemple suivant illustre le cas où les inégalités ne sont pas strictes, ce qui intervient dans les annexes suivantes.

Contre-exemple B.2.9 (pas de normale / complémentarité stricte)**.** Considérons les données suivantes, où il n'y a pas de complémentarité stricte dans le système KKT, i.e., la différence $p - z = p - \mathrm{P}_Z(p)$ n'est pas une normale stricte. C'est illustré en figure B.5.

$$p = \begin{bmatrix} -2; 6 \end{bmatrix}, \quad V = \begin{bmatrix} 0 & 1 \\ 5/2 & 3/2 \end{bmatrix}, \quad \overline{z} = 0$$

Il est clair que la projection de p sur Z est $z = [1; 4]$ et $\xi = [1; 1]$. Cependant, le système KKT est

$$Z^\mathsf{T}(Z\xi + \overline{z} - z) = \begin{bmatrix} 0 & 5/2 \\ 1 & 3/2 \end{bmatrix} \left(\begin{bmatrix} 0 & 1 \\ 5/2 & 3/2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 0 - \begin{bmatrix} -2 \\ 6 \end{bmatrix} \right) = \begin{bmatrix} 0 & 5/2 \\ 1 & 3/2 \end{bmatrix} \begin{bmatrix} 1+2 \\ 4-6 \end{bmatrix} = \begin{bmatrix} -5 \\ 0 \end{bmatrix}$$

Donc $\mu = 0$ et $\lambda = [5; 0]$ résout le système mais sans stricte complémentarité. □

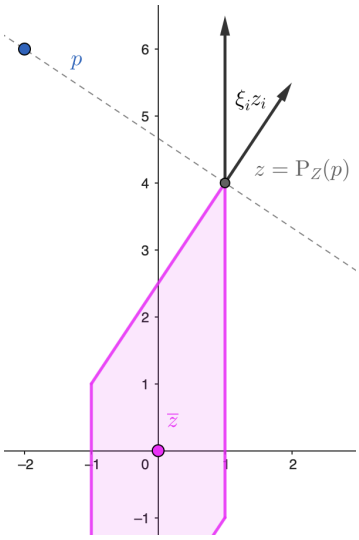


FIGURE B.5 – Exemple d'un point tel que la direction point – projection ne vérifie pas la complémentarité stricte. Cela se produit à un point où la projection n'est pas différentiable. (La fonction distance elle-même est différentiable en-dehors des points sur la frontière du convexe.

Annexe C

Inclusion de zonotopes

Cette annexe détaille des aspects pertinents sur l'inclusion de zonotopes. Bien que cette question ait été essentiellement traitée dans [223] et [149], où il est montré que déterminer si un zonotope est contenu dans un autre est, généralement, co-NP-complet, nous détaillons quelques éléments qui peuvent améliorer la compréhension de ce problème. En particulier, l'algorithme approximatif de [223] ne correspond pas à ce qui est idéalement souhaité dans le chapitre 6, puisque l'on veut plutôt montrer que l'inclusion n'est pas vérifiée – si possible sans énumération combinatoire.

Nous détaillons surtout le travail de [223]. En particulier, mentionnons que l'environnement considéré est tel que la dimension n de l'espace contenant le zonotope est plus grande que le nombre de générateurs m. Ce cas plutôt inhabituel fait que le théorème 3 de [223], le théorème C.0.1 qui suit, peut être une CNS (l'article assume $m > n$ donc ça ne peut se produire).

Théorème C.0.1 (théorème 3 de [223]). *Soit $Z_x = \overline{x} + X[-1, +1]^q$ et $Z_y = \overline{y} + Y[-1, +1]^p$ pour $\overline{x}$, $\overline{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times q}$ et $Y \in \mathbb{R}^{n \times p}$. S'il existe $\Delta \in \mathbb{R}^{p \times q}, \beta \in \mathbb{R}^p$ tel que la condition suivante est vérifiée, alors $Z_x \subseteq Z_y$.*

$$Y[\Delta\ \beta] = [X\ \overline{y} - \overline{x}], |||[\Delta\ \beta]|||_\infty \leqslant 1 \qquad \square$$

Dans le théorème, la norme infinie de la matrice $[\Delta\ \beta]$ est définie comme le maximum des normes 1 des lignes, i.e.,

$$|||[\Delta\ \beta]|||_\infty = \max_{i \in [1:p]} |||[\Delta_{i,:}\ \beta_i]|||_1$$

Deux questions se posent alors. Lorsque la condition n'est pas vérifiée, comment vérifier si l'inclusion est valide ou non. Dans tous les cas, comment obtenir un point indiquant que l'inclusion est fausse. Détaillons ces propriétés. D'abord, observons que, puisque le problème est co-NP-complet et a une nature clairement combinatoire, on peut vérifier si

tous les sommets (pour calculer les sommets, voir chapitre 3) sont contenus dans le second zonotope ou non et s'arrêter au premier qui ne l'est pas.

La condition du théorème C.0.1 peut être vérifiée en résolvant le problème d'optimization suivant :

$$\begin{array}{ll} \min & ||[\Delta\ \beta]||_\infty \\ \text{t.q.} & Y[\Delta\ \beta] = [X\ \overline{y} - \overline{x}] \end{array} \tag{C.1}$$

et en vérifiant si sa valeur optimale est $\leqslant 1$. Lorsque l'optimum est > 1, l'inclusion $Z_x \subseteq Z_y$ peut toujours être vraie : voir l'exemple 2 dans [223] [1]. Mais ce n'est pas dit clairement comment trouver un point de $Z_x \backslash Z_y$ lorsque l'inclusion est fausse. Considérons d'abord le cas où les contraintes ne sont pas réalisables.

Proposition C.0.2 (contraintes irréalisables – 1)**.** *Si $Y\beta = \overline{y} - \overline{x}$ n'est pas réalisable, alors le point $\overline{x} = \overline{x} + X0$ n'appartient pas à $Z_y = \overline{y} + Y[-1, +1]^p$.*

Preuve. Si la contrainte n'est pas réalisable, alors $\overline{y} - \overline{x} \notin \mathcal{R}(Y)$. Cela s'écrit aussi $\overline{x} \notin \overline{y} + \mathcal{R}(Y)$. Puisque $\overline{y} + Y[-1, +1]^p \subseteq \overline{y} + \mathcal{R}(Y)$, cela signifie que $\overline{x}$ n'est pas dans Z_y. □

Proposition C.0.3 (contraintes irréalisables – 2)**.** *Si $Y\Delta = X$ n'est pas réalisable, alors on peut trouver en temps polynomial un point vérifiant que l'inclusion est fausse.*

Preuve. Si ces contraintes ne sont pas réalisables, alors pour un indice $j \in [1 : q]$, $X_{:,j} \notin \mathcal{R}(Y)$ (cet indice est trouvable en temps polynomial). Supposons que $\overline{x}+X_{:,j}$ et $\overline{x}-X_{:,j}$ sont tous deux dans Z_y. Par convexité de Z_y, on a $[\overline{x} - X_{:,j}, \overline{x} + X_{:,j}] \subseteq Z_y$. Cela implique que $\overline{x} + \mathrm{vect}(X_{:,j}) \subseteq \overline{y} + \mathcal{R}(Y)$. En utilisant un argument de dimension, on a que $\mathrm{vect}(X_{:,j}) \subseteq \mathcal{R}(Y)$, ce qui est une contradiction. Donc $\overline{x} + X_{:,j}$ et/ou $\overline{x} - X_{:,j}$ n'est pas dans Z_y. □

Ces deux cas simples, dans lesquels l'optimum est $+\infty$, signifient soit que les centres ne sont pas "alignés" (proposition C.0.2) ou Z_x contient une dimension pas générée par Y (proposition C.0.3). Bien qu'elles ne soit pas pertinentes dans [223], puisque les sonotopes sont de dimension maximale, il est raisonnable de considérer, dans le cadre du chapitre 6, que les hypothèses de ces propositions puissent être vérifiées.

Mentionnons un possible parallèle : la partie du problème avec β et $\overline{y}-\overline{x}$ joue un rôle un peu différent de celui des colonnes de X et Δ, qui peut faire penser aux sections 5.3.4 et 5.6 du chapitre 5, où la partie affine de l'arrangement, τ, joue une sorte de rôle intermédiaire entre la dimension n et la dimension $n + 1$.

Dans ce qui suit, on suppose the (C.1) est réalisable, i.e., par les propriétés de l'optimisation linéaire (voir [105, 29]), qu'il a une valeur optimale finie (sinon les propositions C.0.2 et C.0.3 suffisent) et que la valeur optimale est $\lambda^* > 1$ (sinon il y a inclusion).

1. On conjecture qu'il n'est pas possible d'avoir de contre-exemple en dimension 2 comme les auteurs le suggèrent.

La proposition C.0.4 ci-dessous propose une signification de la valeur de λ^* (un peu différente de celle de $\lambda_{\text{Theorem3}}$ dans [223], davantage dans l'idée de la norme de zonotope de [149]) : λ^* est un facteur de dilatation qui suffit pour que Z_y contienne Z_x (puisque X est générée par Y). C'est illustré dans l'exemple C.0.5.

Proposition C.0.4 (dilatation par $\boldsymbol{\lambda^*}$ fini). *On a* $\overline{x} + X[-1,+1]^q \subseteq \overline{y} + \lambda^* Y[-1,+1]^p$.

Preuve. En utilisant les contraintes, on a

$$\begin{aligned}\overline{x} + X[-1,+1]^q &= \overline{y} - Y\beta + Y\Delta[-1,+1]^q = \overline{y} + Y[\Delta[-1,+1]^q - \beta] \\ &= \overline{y} + Y[\Delta\ \beta]\begin{bmatrix}[-1,+1]^q \\ -1\end{bmatrix} \subseteq \overline{y} + \lambda^* Y[-1,+1]^p\end{aligned}$$

en utilisant la définition $\lambda^* = ||[\Delta\ \beta]||_\infty$. □

En particulier, la preuve montre la légère différence susmentionnée entre les colonnes de X (les indices $[1:q]$) et $\overline{y} - \overline{x}$.

Exemple C.0.5 (la valeur de $\boldsymbol{\lambda^*}$). Considérons les données suivantes, où $\lambda^* = 6$:

$$(\overline{x}, \overline{y}, X, Y) = \left(\begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}2\\0\end{bmatrix}, \frac{1}{2}\begin{bmatrix}0 & 1 & 1\\2 & 2 & -2\end{bmatrix}, \frac{1}{2}\begin{bmatrix}1 & 1\\1 & -1\end{bmatrix}\right), [\Delta\ \beta] = \frac{1}{2}\begin{bmatrix}2 & 3 & -1 & 2\\-2 & -1 & 3 & 6\end{bmatrix}$$

□

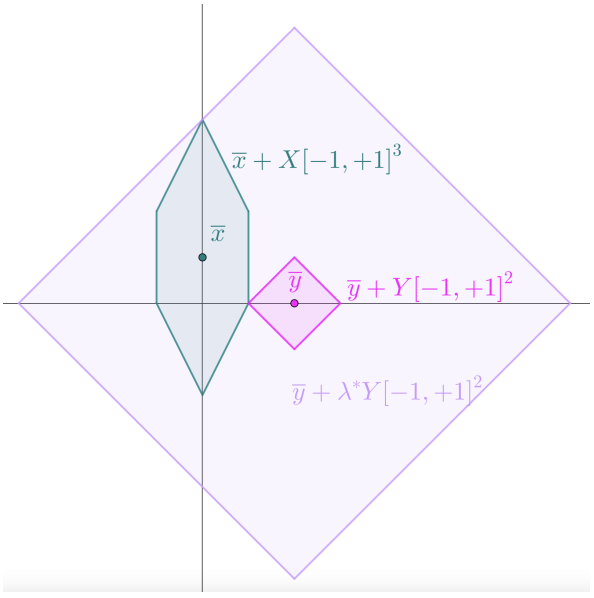


Figure C.1 – Dans cet exemple particulier, la solution (unique) (Δ, β) du problème détaillé dans l'exemple C.0.5 donne $\lambda^* = 6$. Observons que lorsque l'on dilate Z_y par λ^*, on a $Z_x \subseteq \overline{y} + \lambda^* Y[-1,+1]^2$, mais en dilatant par tout $\lambda < \lambda^*$, l'inclusion n'est pas vérifiée (le point en haut de l'aire verte n'est pas contenu dans la l'aire violette).

Lorsque $Z_x \subseteq Z_y$ mais $\lambda^* > 1$, i.e., le théorème C.0.1 se trompe, la valeur de λ^* est imprécise (une sorte de sur-estimation) puisque $Z_y = \overline{y} + Y[-1,+1]^p \subsetneq \overline{y} + \lambda^* Y[-1,+1]^p$. Néanmoins, le théorème devient une condition nécessaire et suffisante si Y a des colonnes indépendantes. Cette observation n'est pas dite dans [223] puisque la dimension n est plus petite que le nombre de générateurs m, ce qui est le cas usuel des zonotopes. Cependant, nous supposons l'inverse. Avant de détailler cela, obtenons le problème dual de (C.1). En utilisant la dualité lagrangienne [2], on a

$$\begin{array}{ll} \max & (\Lambda, [X\ \overline{y} - \overline{x}]) \\ \text{t.q.} & ||Y^\mathsf{T}\Lambda||_1 \leqslant 1 \end{array} \tag{C.2}$$

où la norme utilisée, la norme 1 sur les matrices, est duale à la norme précédente et définit comme la somme des normes infinies des lignes, i.e., $||M||_1 = \sum_{i=1}^{p} ||M_{i,:}||_\infty$. De plus, comme on peut reformuler ces problèmes comme de l'optimisation linéaire, il y a dualité forte.

Remarque C.0.6 (norme 1 des matrices avec forme particulière)**.** Soit $M = ms^\mathsf{T}$ pour $m \in \mathbb{R}^n$ et $s \in \{\pm 1\}^l$, alors $||M||_1 = \sum_{i=1}^{n} ||m_i s^\mathsf{T}||_\infty = \sum_{i=1}^{n} |m_i| = ||m||_1$. □

Proposition C.0.7 (*Y* de rang plein / injective). *Quand Y a des colonnes indépendantes et $\lambda^* > 1$, on peut obtenir un point $\overline{x} + X\eta \notin Z_y$ pour un $\eta \in [-1,+1]^q$.*

Preuve. Si Y a des colonnes indépendantes, puisque $Y \in \mathbb{R}^{n \times p}$ et $n > p$, Y est injective et $Y^\mathsf{T}Y$ est inversible. Alors l'unique solution primale est

$$[\Delta\ \beta] = (Y^\mathsf{T}Y)^{-1}Y^\mathsf{T}[X\ \overline{y} - \overline{x}].$$

Soit $i^* \in [1:p]$ tel que $\lambda^* = ||[\Delta\ \beta]_{i,:}||_1$ et $z := Y(Y^\mathsf{T}Y)^{-1}e_{i^*}$, en particulier $z \in \mathcal{R}(Y)$ et $Y^\mathsf{T}z = e_{i^*}$. Construisons une variable duale de la forme $\Lambda = zs^\mathsf{T}$ qui atteint la valeur λ^* pour un $s \in \{\pm 1\}^{q+1}$ bien choisi. D'abord, observons que

$$||Y^\mathsf{T}\Lambda||_1 = ||Y^\mathsf{T}zs^\mathsf{T}||_1 = ||(Y^\mathsf{T}z)s^\mathsf{T}||_1 = ||e_{i^*}s^\mathsf{T}||_1 = ||s^\mathsf{T}||_\infty = 1$$

où $||s^\mathsf{T}||_\infty$ est la norme infinie sur $\mathbb{R}^{q+1}$ (pour un vecteur ligne). Le coût dual s'écrit :

$$\begin{aligned}(\Lambda, [X\ \overline{y} - \overline{x}]) &= (zs^\mathsf{T}, Y[\Delta\ \beta]) = (Y^\mathsf{T}zs^\mathsf{T}, [\Delta\ \beta]) \\ &= (e_{i^*}s^\mathsf{T}, [\Delta\ \beta]) = (s, [\Delta\ \beta]_{i^*,:}) = \sum_{j=1}^{q} s_j \Delta_{i^*,j} + s_{q+1}\beta_{i^*}\end{aligned}$$

2. En effet, soit $\tilde{\Delta} := [\Delta\ \beta]$,

$$\begin{aligned}\min_{\tilde{\Delta}} \max_{\Lambda} ||\tilde{\Delta}||_\infty + (-Y\tilde{\Delta} + [X\ \overline{y} - \overline{x}], \Lambda) &\iff \max_{\Lambda} \min_{\tilde{\Delta}} (\Lambda, [X\ \overline{y} - \overline{x}]) + ||\tilde{\Delta}||_\infty - (\tilde{\Delta}, Y^\mathsf{T}\Lambda) \\ &\iff \max_{\Lambda} (\Lambda, [X\ \overline{y} - \overline{x}]), \text{t.q. } ||Y^\mathsf{T}\Lambda||_1 \leqslant 1\end{aligned}$$

Maintenant, prenons $s_{q+1} = \text{sgn}(\beta_{i^*})$ et $s_j = \text{sgn}(\Delta_{i^*,j})$ pour $j \in [1 : q]$ (si β_{i^*} ou $\Delta_{i^*,j} = 0$, le signe correspondant est choisi arbitrairement). Le coût dual devient :

$$\sum_{j=1}^{q} s_j \Delta_{i^*,j} + s_{q+1}\beta_{i^*} = \sum_{j=1}^{q} |\Delta_{i^*,j}| + |\beta_{i^*}| = \lambda^*$$

Finalement, considérons $z^* = \overline{x} + X(-\text{sgn}(\beta_{i^*})s_{[1:q]})$ en supposant $\beta_{i^*} \neq 0$ pour l'instant. Montrons que ce point n'appartient pas à Z_y. En utilisant les équations 5 et 6 de [149], on a

$$\overline{x} + Xs_{[1:q]} \in Z_y \iff \begin{array}{ll} \min & ||\zeta||_\infty \\ \text{t.q.} & \overline{x} + Xs_{[1:q]} = \overline{y} + Y\zeta \end{array} \text{ has optimal value} \leqslant 1. \tag{C.3}$$

En effet, lorsque la seconde condition se produit $\overline{x} + Xs_{[1:q]}$ peut s'exprimer comme un point de Z_y. En utilisant les contraintes primales et l'injectivité de Y, on obtient

$$\begin{aligned} \overline{x} + X(-\text{sgn}(\beta_{i^*})s_{[1:q]}) &= \overline{y} + Y\zeta \\ \iff \overline{x} - \overline{y} + X(-\text{sgn}(\beta_{i^*})s_{[1:q]}) &= Y\zeta \\ \iff -Y(\beta + \text{sgn}(\beta_{i^*})\Delta s_{[1:q]}) &= Y\zeta \\ \iff -\beta - \text{sgn}(\beta_{i^*})\Delta s_{[1:q]} &= \zeta \end{aligned}$$

où ζ est unique. Ensuite, on a :

$$\zeta_{i^*} = -\beta_{i^*} - \text{sgn}(\beta_{i^*}) \sum_{j=1}^{q} s_j \Delta_{i^*,j} = -\beta_{i^*} - \text{sgn}(\beta_{i^*}) \sum_{j=1}^{q} |\Delta_{i^*,j}| = -\text{sgn}(\beta_{i^*})\lambda^* \notin [-1, +1].$$

Si $\beta_{i^*} = 0$, soit $z^* = \overline{x} + X(\pm s_{[1:q]})$, on a :

$$\begin{aligned} \overline{x} \pm Xs_{[1:q]} &= \overline{y} + Y\zeta \\ \iff \overline{x} - \overline{y} \pm Xs_{[1:q]} &= Y\zeta \\ \iff -Y\beta \pm Y\Delta s_{[1:q]} &= Y\zeta \\ \iff -\beta \pm \Delta s_{[1:q]} &= \zeta \end{aligned}$$

et $\zeta_{i^*} = 0 \pm \sum_{j=1}^{q} s_j \Delta_{i^*,j} = \pm\lambda^* \notin [-1, +1]$. □

Le cas où Y est de rang plein est par exemple celui de l'exemple C.0.5. Observons que les indices où le signe s_j est choisi arbitrairement n'intervient pas dans le calcul final. L'hypothèse sur Y est relativement forte. Cependant, même lorsqu'elle n'est pas vérifiée, on peut tout de même obtenir une variable duale de cette forme. De là, un raisonnement similaire peut être utilisé.

Proposition C.0.8 (solution duale de la forme $zs^\intercal$)**.** *Soit $\Lambda = zs^\intercal$ pour $z \in \mathbb{R}^n$ et $s \in \{\pm 1\}^{q+1}$ être une solution duale, avec $\lambda^* > 1$ et $s_{q+1} = -1$ (quitte à prendre $(-z, -s)$). Alors $Z_x \nsubseteq Z_y$.*

Preuve. Montrons que le point $\overline{x}+Xs_{1:q}$ n'est pas dans Z_y. Comme dans (C.3), les problèmes primaux et duaux sont

$$(P)\ \begin{cases} \min & ||\zeta||_\infty \\ \text{t.q.} & \overline{x}+Xs_{[1:q]} = \overline{y}+Y\zeta \end{cases} \qquad (D)\ \begin{cases} \max & w^\mathsf{T}Y(\Delta s_{[1:q]}-\beta) \\ \text{t.q.} & ||Y^\mathsf{T}w||_1 \leqslant 1 \end{cases} \tag{C.4}$$

Montrons que, dans le problème dual, z est une variable telle que le coût dual est $\lambda^* > 1$. Par dualité forte, le coût primal sera aussi $\lambda^* > 1$. En effet, l'évaluation du coût dual en z s'écrit

$$(Y^\mathsf{T}z)^\mathsf{T}(\Delta s_{[1:q]}-\beta) = (Y^\mathsf{T}z)^\mathsf{T}(\Delta s_{[1:q]}+\beta s_{q+1}) = (Y^\mathsf{T}z)^\mathsf{T}[\Delta\ \beta]s$$

alors que le coût dual avec Λ s'écrit

$$\lambda^* = (\Lambda, [X\ \overline{y}-\overline{x}]) = (Y^\mathsf{T}z, [\Delta\ \beta]s). \qquad \square$$

En particulier, cela n'est pas vérifié pour l'exemple 2 de [223]. Le tableau suivant résume les cas traités pour l'instant.

cas	$Z_x \subseteq Z_y$?	récupération du paramètre	coût
$\lambda^* \leqslant 1$	oui	rien à faire	POL
$\mathrm{rank}(Y) = p$	non	possible	POL
solution duale $\Lambda = zs^\mathsf{T}$	non	donné par s	POL

Montrons maintenant que l'on a aussi l'implication inverse de la proposition C.0.8, i.e., lorsqu'il n'y a pas inclusion mais que les problèmes sont réalisables il existe une solution duale avec colonnes colinéaires. La preuve utilise certaines propriétés des zonotopes (et polytopes) détaillées dans l'annexe B.

Proposition C.0.9 (contraposée de la proposition C.0.8)**.** *Si $Z_x \nsubseteq Z_y$ mais que les problèmes sont réalisables, il existe une solution duale de la forme $\Lambda = zs^\mathsf{T}$.*

Preuve. Puisque $Z_x \nsubseteq Z_y$, il existe un $\eta \in \{\pm 1\}^q$ tel que $\overline{x}+X\eta \in \overline{y}+\lambda^* Y[-1,+1]^p$, qui s'écrit

$$\overline{x}+X\eta = \overline{y}+\lambda^* Y\zeta \quad \text{pour un } \zeta \in [-1,+1]^p.$$

Montrons que, pour une paire (z,s) judicieuse, on peut trouver une solution duale $\Lambda = zs^\mathsf{T}$ pour $s_{[1:q]} = \eta$ et $s_{q+1} = -1$ avec un z approprié. D'abord, observons que $\overline{y}+\lambda^* Y\zeta$ appartient à une face de $\overline{y}+\lambda^* Y[-1,+1]^p$. De fait, il existe une certaine normale c comme décrite dans la proposition B.1.7. Puisque c peut être multipliée par une constante strictement positive, supposons que $||Y^\mathsf{T}c||_1 = 0$ et prenons $z = c$. La contrainte duale s'écrit

$$||Y^\mathsf{T}zs^\mathsf{T}||_1 = \sum_i ||y_i^\mathsf{T}zs^\mathsf{T}||_\infty = \sum_i |y_i^\mathsf{T}z| = ||Y^\mathsf{T}c||_1 = 1.$$

De plus, le coût dual s'écrit :

$$(zs^\mathsf{T}, [X\ \overline{y}-\overline{x}]) = (z, [X\ \overline{x}-\overline{y}][\eta;1]) = (c, X\eta+\overline{x}-\overline{y}) = \lambda^*(Y^\mathsf{T}c, \zeta) = \lambda^* \sum_{i=1}^p \zeta_i y_i^\mathsf{T}c$$

et par les propriétés de c dans la proposition B.2.3, on a $\zeta_i y_i^\mathsf{T} c \geqslant 0$. Si $\zeta_i \in (-1, +1)$, y_i est un générateur de la face et $y_i^\mathsf{T} c = 0$, donc soit $y_i^\mathsf{T} c = 0$, soit $\zeta_i \in \{-1, +1\}$ et donc la somme vaut $||Y^\mathsf{T} c|| = 1$, ce qui termine la preuve. □

Pour terminer, résumons ces observations diverses en un algorithme possible.

Algorithme C.0.10 (résoudre l'inclusion de zonotopes)**.** Entrée : $\overline{x}, \overline{y}, X, Y$. Sortie : booléen indiquant si $Z_x \subseteq Z_y$ et point z indiquant s'il n'y a pas inclusion. Quand c'est pertinent, on renvoie aussi η tel que $\overline{x} + X\eta \notin \overline{y} + Y[-1, +1]^p$. Les cas suivants sont disjoints.

1. *Contraintes primales irréalisables.* Si la contrainte $Y\beta = \overline{y} - \overline{x}$ n'est pas réalisable, il n'y a pas inclusion, renvoyer (FAUX, $z = \overline{x} + X0$, $\eta = 0$). Si la contrainte $Y\Delta = X$ n'est pas réalisable, obtenir un indice $j \in [1 : q]$ tel que $X_{:,j} \notin \mathcal{R}(Y)$, vérifier si $\overline{x} \pm X_{:,j} = \overline{x} \pm Xe_j$ n'est pas dans $\overline{y} + Y[-1, +1]^p$, renvoyer (FAUX, $z = \overline{x} \pm Xe_j$, $\eta = e_j$).
2. *Vérification primale.* Si (C.1) est réalisable et $\lambda^* \leqslant 1$, then $Z_x \subseteq Z_y$, renvoyer (VRAI).
3. *Cas de Y avec des colonnes indépendantes.* Si Y est injective, utilisons la proposition C.0.7 pour obtenir le η tel que $z = \overline{x} + X\eta \notin Z_y$, renvoyer (FAUX, z, η).
4. *Vérification duale.* Résoudre le problème dual (C.2). Si la variable duale a des colonnes égales au signe près, i.e., $\Lambda = zs^\mathsf{T}$, soit z et s tels que $s_{q+1} = -1$, et z le point vérifiant que l'inclusion est fausse, $z = \overline{x} + Xs_{[1:q]} \notin \overline{y} + Y[-1, +1]^p$, renvoyer (FAUX, z, η).
5. *Problème dual avec colonnes identiques au signe près.* Résoudre le problème dual en imposant que les colonnes soient égales ou opposées. Si l'optimum est λ^*, alors $Z_x \nsubseteq Z_y$, utiliser $\eta := -\mathrm{sgn}(s_{q+1})s_{[1:q]}$, $z = \overline{x} + X\eta \notin \overline{y} + Y[-1, +1]^p$, renvoyer (FAUX, z, η).
6. *Problème dual sans solution avec colonnes identiques au signe près.* Alors le théorème "échoue" et il y a inclusion, renvoyer (TRUE).

Dans l'algorithme C.0.10, l'étape 5 doit soit utiliser des contraintes non linéaires de la forme $\Lambda_{:,j} = s_j \Lambda_{:,q+1}$ ou des variables binaires par exemple. Dans les deux cas, le problème n'est plus polynomial. Les étapes 4 et 5 diffèrent en le fait que, lorsque l'inclusion n'est pas vérifiée, donc qu'il existe une solution duale de la forme $\Lambda = zs^\mathsf{T}$, le solveur linéaire peut ou peut ne pas l'obtenir [3].

L'algorithme C.0.10 peut sembler laborieux, la raison de sa formulation est que l'on essaie d'éviter le problème combinatoire si possible. On pourrait aussi vérifier chaque sommet du zonotope et vérifier s'il appartient à l'autre zonotope.

3. Des solveurs comme GUROBI peuvent l'obtenir naturellement même sans imposer epxlicitement la contrainte comme dans l'étape 5.

Annexe D

Poids et élément du différentiel de Clarke

Le but de cette annexe est de donner une preuve à la proposition 6.1.20. La preuve s'est révélée être plutôt longue et technique – il est possible que des approches plus simples utilisant des propriétés de la méthode de Levenberg-Marquardt existent. Sur le chemin, on discute aussi de diverses technicalités.

D.1 Géométrie et ensembles de vecteurs de signes

Cette brève section vise à éclaircir un résultat relativement abstrait mais qui explique des difficultés ultérieures, dans la preuve principale de cette annexe. Cela concerne une propriété lorsque deux sous-ensembles de vecteurs sont indépendants et l'impact sur les sous-arrangements correspondants. Dans ce qui suit, X and Y sont des matrices avec le même nombre de lignes, et on rappelle que $[X\ Y]$ (resp. $[X\ -Y]$) est la concaténation horizontale de X et Y (resp. X et $-Y$). Cette proposition semble généraliser la proposition 3.3.17. Puisqu'elle est principalement utlisée pour expliquer des observations ci-dessous, on ne l'a pas mise dans le chapitre 4.

Proposition D.1.1 ($(\boldsymbol{X}, \boldsymbol{Y})$, $(\boldsymbol{X}, \boldsymbol{-Y})$ et images)**.** *Soit $X \in \mathbb{R}^{n\times q}$ et $Y \in \mathbb{R}^{n\times p}$, on a l'équivalence suivante :*

$$\mathcal{R}(X) \cap \mathcal{R}(Y) = \{0\} \qquad \Longleftrightarrow \qquad \mathcal{S}([X\ Y]) = \mathcal{S}([X\ -Y]). \tag{D.1}$$

La signification du membre de gauche est que les vecteurs de X et Y appartiennent à des sous-espaces ayant une intersection vide, tandis que celui de droite signifie que les ensembles de vecteurs de signes sont invariants si on prend l'opposé d'une partie des vecteurs (pour ce dont on a besoin ci-dessous, cela signifie que les zonotopes correspondant à

$[X\ Y]$ et $[X\ -Y]$ ont des sommets avec les mêmes vecteurs de signes, voir plus bas). Rappelons que le "support" fait référence à un sous-ensembles d'indices tels qu'une quantité considérée a des composantes non nuls sur les indices de son support.

Preuve. [⇒] Montrons la contraposée en utilisant les vecteurs souches. Si l'égalité à droite n'est pas vraie, il existe un vecteur de signe dans $\mathcal{S}([X\ Y])\backslash\mathcal{S}([X\ -Y])$ (puisque des vecteurs opposés créent le même hyperplan, les hyperplans dans les deux arrangements sont les mêmes, donc il y a le même nombre de vecteurs de signes dans les deux, et alors, si on suppose qu'ils sont différents, un ne peut être inclus dans l'autre). Soit s ce vecteur de signe, $s \notin \mathcal{S}([X\ -Y])$ s'écrit, avec $s = (s^x, s^y)$ les parties correspondantes aux indices de X et Y,

$$\begin{aligned}\exists\, \alpha = (\alpha_x, \alpha_y) \in \mathbb{R}_+^{q+p}\backslash\{0\},\ \sum(\alpha_x)_i s_i^x X_{:,i} + \sum(\alpha_y)_i s_i^y(-Y_{:,i}) &= 0\\ = X\operatorname{Diag}(s^x)\alpha_x - Y\operatorname{Diag}(s^y)\alpha_y &= 0.\end{aligned}$$

Observons que $X\operatorname{Diag}(s^x) = 0$ est impossible puisque cela impliquerait $Y\operatorname{Diag}(s^y)\alpha_y = 0$, donc $(0, \alpha_y) \neq 0$ serait un vecteur souche, mais cela contredit le fait que $s \in \mathcal{S}([X\ Y])$. De même, on ne peut avoir $Y\operatorname{Diag}(s^y)\alpha_y = 0$. De fait, $d := X\operatorname{Diag}(s^x)\alpha_x = Y\operatorname{Diag}(s^y)\alpha_y$ est non nul, on a trouvé un $d \in \mathcal{R}(X) \cap \mathcal{R}(Y)$ qui est non nul.

[⇐] Supposons que l'on a $\mathcal{S}([X\ Y]) = \mathcal{S}([X\ -Y])$ et $\mathcal{R}(X) \cap \mathcal{R}(Y) \neq \{0\}$, et observons une contradiction. D'abord, montrons une conséquence de l'hypothèse $\mathcal{S}([X\ Y]) = \mathcal{S}([X\ -Y])$. Pour toute paire de sous-ensembles $J_x \subseteq [1:q]$ et $J_y \subseteq [q+1:q+p]$, on a $\mathcal{S}([X_{:,J_x}\ Y_{:,J_y}]) = \mathcal{S}([X_{:,J_x}\ -Y_{:,J_y}])$. En effet, si cette inégalité n'est pas vraie, il existe un vecteur de signe s dans un des deux ensembles qui n'est pas dans l'autre (même raison que l'implication précédente). De fait, en utilisant que (après réordonnancement des indices) [1]

$$\begin{cases} \mathcal{S}([X\ Y]) & \subseteq & \mathcal{S}([X_{:,J_x}\ Y_{:,J_y}]) \times \{\pm 1\}^{J_x^c \cup J_y^c} \\ \mathcal{S}([X\ -Y]) & \subseteq & \mathcal{S}([X_{:,J_x}\ -Y_{:,J_y}]) \times \{\pm 1\}^{J_x^c \cup J_y^c} \end{cases},$$

et que chaque vecteur de signe a au moins un descendant, le vecteur de signe s a un descendant qui n'appartient pas à l'autre ensemble où l'on considère toutes les colonnes de X et Y.

La négation du terme de gauche signifie qu'il existe un $d \in \mathbb{R}^n$ non nul tel que

$$\operatorname{Vect}(d) \subseteq \mathcal{R}(X) \cap \mathcal{R}(Y), \tag{D.2a}$$

qui s'écrit aussi

$$Xd^x = d = Yd^y, \qquad \text{pour un } d^x \in \mathbb{R}^q \text{ et } d^y \in \mathbb{R}^p. \tag{D.2b}$$

De plus, soit $B_x \subseteq [1:q]$ une base de $\mathcal{R}(X)$ et supposons, sans perte de généralité, que $d^x_{B_x^c} = 0$, i.e., le support de d^x est contenu dans B_x. Exprimons (D.2b) comme

$$Xd^x - Yd^y = 0, \qquad \text{ou} \qquad X\operatorname{Diag}(\operatorname{sgn}(d^x))|d^x| - Y\operatorname{Diag}(\operatorname{sgn}(d^y))|d^y| = 0, \tag{D.2c}$$

1. Et en utilisant $J_x^c = [1:q] \setminus J_x$, $J_y^c = [q+1:q+p] \setminus J_y$.

où $0 \times \text{sgn}(0) = 0$. Sous cette forme, on reconnaît, par l'alternative de Gordan, que

$$(s^x, s^y) := (\text{sgn}(d^x), \text{sgn}(d^y)) \tag{D.2d}$$

est un sous-vecteur de signe infaisable (en prenant les composantes non nulles) de $[X_{:,B_x} - Y]$. Par l'hypothèse sur le membre de droite et sa conséquence discutée, il est aussi un sous-vecteur de signe infaisable de $[X_{:,B_x}\ Y]$. Cela signifie qu'il existe un $(\alpha_x, \alpha_y) \geqslant 0$ non nul à support dans $B_x \times \text{supp}(d^y)$ tel que

$$X\,\text{Diag}(s^x)\alpha_x + Y\,\text{Diag}(s^y)\alpha_y = 0. \tag{D.2e}$$

Observons que $Y\,\text{Diag}(s^y)\alpha_y = 0$ est impossible puisque cela impliquerait $X\,\text{Diag}(s^x)\alpha_x = 0$, i.e., que s^x est irréalisable ce qui n'est pas possible puisque son support est dans B_x, i.e., correspond à des vecteurs $X_{:,i}, i \in B_x$ qui sont indépendants. Multiplions (D.2e) par

$$\alpha_0 := \max\left\{\frac{|d^y_j|}{(\alpha_y)_j}, (\alpha_y)_j \neq 0\right\} > 0, \tag{D.2f}$$

qui est bien défini puisque l'on ne peut avoir $\alpha_y = 0$ est non nul par l'hypothèse sur le support. Alors, en sommant α_0(D.2e) et (D.2c), on obtient

$$\begin{aligned} & X\,\text{Diag}(s^x)(|d^x| + \alpha_0\alpha_x) + Y\,\text{Diag}(s^y)(\alpha_0\alpha_y - |d^y|) = 0, \\ \iff & X\,\text{Diag}(s^x)(|d^x| + \alpha_0\alpha_x) - Y[-\,\text{Diag}(s^y)\,\text{Diag}(\text{sgn}(\alpha_0\alpha_y - |d^y|))]|\alpha_0\alpha_y - |d^y|| = 0 \end{aligned} \tag{D.2g}$$

où le support de $\alpha_0\alpha_y - |d^y|$ est strictement contenu dans celui de d^y par définition de α_0 et clairement $|d^x| + \alpha_0\alpha_x \geqslant 0$. Ensuite, on a que la paire $(s^x, -\,\text{Diag}(s^y)\text{sgn}(\alpha_0\alpha_y - |d^y|))$ est infaisable dans $[X_{:,B_x}\ -Y_{:,J^1_y}]$ avec J^1_y le support de $\alpha_0\alpha_y - |d^y|$. Ensuite, il doit aussi être infaisable dans $[X_{:,B_x}\ Y_{:,J^1_y}]$. En utilisant le même argument, il existe un $(\alpha^1_x, \alpha^1_y) \geqslant 0$ non nul avec support dans $B_x \times J^1_y$, toujours tel que α^1_y ne peut être nul, et on peut réutiliser le même argument.

Pour terminer, on peut réduire d'au moins 1 la taille du support de J^l_y à chaque itération l, signifiant qu'à un moment on a $X\,\text{Diag}(s^x)[\dots] = 0$, qui est une contradiction. □

D.2 Preuve principale

Maintenant, on se concentre sur la preuve de la proposition 6.1.20. Avant d'arriver à la preuve, plusieurs propriétés seront discutées et analysées, en partie à travers le prisme des zonotopes, qui a l'avantage d'être plutôt visuel.

L'idée principale est la suivante : d'abord, de choisir un $\gamma_{\mathcal{E}^{0+}(x)}$ extrême, puis, par la projection de la proposition 6.1.10, de trouver un $\gamma_{\mathcal{E}^{-}(x)}$ convenable. Ensuite, ce $\gamma_{\mathcal{E}^{-}(x)}$ correspond (par le lemme B.1.4) à l'intérieur relatif d'une face du zonotope associé. Ce $\gamma_{\mathcal{E}^{-}(x)}$

est naturellement une combinaison convexe des sommets de la face, les $\gamma^{\text{sommet}}_{\mathcal{E}^-(x)}$. Ensuite, grâce aux propriétés des normales aux faces discutées dans l'annexe C, on peut justifier que les couples $(\gamma_{\mathcal{E}^{0+}(x)}, \gamma^{\text{sommet}}_{\mathcal{E}^-(x)})$ correspondent à des vecteurs de signes dans $\mathcal{S}$, i.e., une matrice jacobienne dans $\partial_B H(x)$. Ensuite, par les propriétés des combinaisons convexes, le g appartient au (C-)différentiel de θ. En réalité, on ne considère que les vecteurs de signes des linéarisations de F et G, i.e., la partie de $\partial_B H(x)$ gouvernée par la matrice V (voir les chapitres 3 and 4). Cela signifie que le processus ignore certaines jacobiennes dans $\partial_B H(x)$, donc de certains points dans $\partial\theta(x)$. Cependant, le résultat reste vrai.

Ce chemin est cependant embûché, au-delà de sa technicalité, par ce que l'on appelle "dégénérescences" dans ce qui suit. Visibles en utilisant les zonotopes, elles montrent que pour les valeurs $(\gamma_{\mathcal{E}^{0+}(x)}, \gamma^{\text{sommet}}_{\mathcal{E}^-(x)})$, le $(\eta, \zeta) \in \{-1, +1\}^{\mathcal{E}(x)}$ correspondant ne correspond pas nécessairement à des vecteurs de signes de $\mathcal{S}$. Elles sont aussi reliées à l'absence d'inégalités strictes discutée dans l'annexe C autour du contre-exemple B.2.9. Résumons ces observations.

- Il y a une différence significative entre les éléments de $\partial_B H(x)$ et $\partial\theta(x)$, malgré les relations entre les deux ensembles.
- Les dégénérescences rendent la preuve du cas général plus technique.
- La méthode suggérée de l'annexe C pour l'inclusion de zonotopes n'est pas l'outil adapté : la valeur $\gamma_{\mathcal{E}^{0+}(x)}$ (η) retournée peut être telle que $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^{0+}(x)})) \notin \partial\theta(x)$ pour tout $\gamma_{\mathcal{E}^-(x)} \in \mathbb{G}(\gamma_{\mathcal{E}^{0+}(x)})$.
- Pour cette méthode, on peut corriger les dégénérescences pour obtenir un élément de $\partial\theta(x)$, mais on ne sait pas si c'est toujours une direction de descente puisque l'on ne respecte pas la projection.
- Le point de Z_x le plus éloigné de Z_y (en distance euclidienne), qui n'est pas nécessairement le même que celui de la méthode précédente, après application de la projection, correspond à un élément de $\partial\theta(x)$. C'est également vrai pour des maxima locaux strictes de la distance.

Rappelons quelques notations utiles (l'équation (6.12) et la règle 6.1.8)

$$\begin{aligned} g_0(x) &:= F'_{\mathcal{F}(x)}(x)^\mathsf{T} F_{\mathcal{F}(x)}(x) + G'_{\mathcal{G}(x)}(x)^\mathsf{T} G_{\mathcal{G}(x)}(x) + G'_{\mathcal{E}(x)}(x)^\mathsf{T} G_{\mathcal{E}(x)}(x), \\ \mathcal{M}_+ &:= (F'_{\mathcal{E}^{0+}(x)}(x) - G'_{\mathcal{E}^{0+}(x)}(x))^\mathsf{T} \operatorname{Diag}(H_{\mathcal{E}^{0+}(x)}(x)), \\ \mathcal{M}_- &:= (F'_{\mathcal{E}^-(x)}(x) - G'_{\mathcal{E}^-(x)}(x))^\mathsf{T} \operatorname{Diag}(H_{\mathcal{E}^-(x)}(x)). \end{aligned} \tag{D.3}$$

Dans la règle suivante, λ^* et $\bar{\cdot}$ réfèrent aux valeurs optimales renvoyées par l'algorithme C.0.10 de l'annexe C (quand λ^* est fini).

Règle D.2.1 (correspondance de variables)**.** Dans la suite, on utilise les quantités suivantes

$$\begin{array}{cccc} X = \frac{1}{2}\mathcal{M}_+ & Y = -\frac{1}{2}\mathcal{M}_- & \overline{x} - \overline{y} = g_1 := g_0(x) + \frac{\mathcal{M}_+}{2}e + \frac{\mathcal{M}_-}{2}e \\ \eta = 2\gamma_{\mathcal{E}^{0+}(x)} - e & \zeta = 2\gamma_{\mathcal{E}^-(x)}\gamma_{\mathcal{E}^-(x)} - e & \overline{\zeta} = \frac{-\beta+\Delta\overline{\eta}}{\lambda^*} \end{array}$$
$$\begin{aligned} g &= \frac{\mathcal{M}_+}{2}\eta + \frac{\mathcal{M}_-}{2}\zeta + \overline{x} - \overline{y} = \overline{x} - \overline{y} + X\eta - Y\zeta \\ \overline{g} &= \frac{\mathcal{M}_+}{2}\eta + \frac{\mathcal{M}_-}{2}\overline{\zeta} + \overline{x} - \overline{y} = \overline{x} - \overline{y} + X\eta - Y\overline{\zeta} \\ \overline{g} &= Y(-\beta + \Delta\overline{\eta} - \overline{\zeta}) = Y(-\beta + \Delta\overline{\eta})\frac{\lambda^* - 1}{\lambda^*} = Y\overline{\zeta}(\lambda^* - 1) \end{aligned} \tag{D.4}$$

où $\eta \in [-1,+1]^{\mathcal{E}^{0+}(x)}$ et $\zeta \in [-1,+1]^{\mathcal{E}^-(x)}$ correspondent aux paramétrisations de $[-1,+1]$. □

On démarre avec une observation qui a motivé le résultat de la section D.1. Le sous-ensemble $\mathcal{E}^0(x) := \{i \in \mathcal{E}^{0+}(x) : H_i(x) = 0\}$ joue un rôle légèrement contrariant qui justifie sa démarcation. Soit $x \in \mathbb{R}^n$, $s \in \{\pm 1\}^{\mathcal{E}(x)}$, et $J(s) \in \mathbb{R}^{n\times n}$ la matrice définie par

$$J(s)_{\mathcal{F}(x)\cup\mathcal{G}(x),:} = \begin{bmatrix} F'_{\mathcal{F}(x)}(x) \\ G'_{\mathcal{G}(x)}(x) \end{bmatrix}, \qquad J(s)_{\mathcal{E}(x),:} = \frac{s+e}{2} \cdot F'_{\mathcal{E}(x)}(x) + \frac{e-s}{2} \cdot G'_{\mathcal{E}(x)}(x). \tag{D.5}$$

Elle est telle que $s_i = +1$ signifie $J_{i,:} = F'_i$ et $J_{i,:} = G'_i$ quand $s_i = -1$. Par définition du B-différentiel du minimum composante par composante, $J(s) \in \partial_B H(x) \iff \exists\, d, s{\cdot}V^\mathsf{T}d > 0$ où $V^\mathsf{T} := G'_{\mathcal{E}(x)}(x) - F'_{\mathcal{E}(x)}(x)$. Ensuite, on peut utiliser les équivalences suivantes (en supposant que $\mathcal{E}^0(x) = \{i \in [1:n] : F_i(x) = 0 = G_i(x)\} = \varnothing$, ces indices n'ont pas d'impact sur l'expression finale, donc cette hypothèse n'est pas trop importante) [2]

$$\begin{aligned} J(s) \in \partial_B H(x) \quad &\iff \quad \exists\, d, s \cdot V^\mathsf{T}d > 0 \\ &\iff \quad \exists\, d, s \cdot \mathrm{Diag}(H_{\mathcal{E}(x)})^{-1}\,\mathrm{Diag}(H_{\mathcal{E}(x)})V^\mathsf{T}d > 0 \\ &\iff \quad \exists\, d, s \cdot \mathrm{Diag}(H_{\mathcal{E}(x)})^{-1}[\mathcal{M}_+\ \mathcal{M}_-]^\mathsf{T}d > 0 \\ &\iff \quad \exists\, d, s \cdot \mathrm{Diag}(H_{\mathcal{E}(x)})^{-1}[2X\ -2Y]^\mathsf{T}d > 0 \\ &\iff \quad \exists\, d, s \cdot [X\ Y]^\mathsf{T}d > 0, \end{aligned} \tag{D.6}$$

en utilisant la définition rappelée dans (D.3) et la règle D.2.1, puis que

$$\begin{aligned} s_i H_i(x)^{-1} X_{:,i}^\mathsf{T} d > 0 \quad &\iff \quad s_i X_{:,i}^\mathsf{T} d > 0 \\ s_i H_i(x)^{-1}(-Y_{:,i})^\mathsf{T} d > 0 \quad &\iff \quad s_i Y_{:,i}^\mathsf{T} d > 0 \end{aligned}$$

car $H_i(x) > 0$ pour les indices de $\mathcal{E}^{0+}(x)$ et $H_i(x) < 0$ pour ceux de $\mathcal{E}^-(x)$. De plus, observons que, pour $s \in \{\pm 1\}^{\mathcal{E}(x)}$ (pas nécessairement $s \in \mathcal{S}(V,0) \iff J(s) \in \partial_B H(x)$),

2. Pour simplifier, on écrit $\partial_B H$ malgré ne considérer que les vecteurs de signes des linéarisations, donc indiqués par la matrice V.

on peut exprimer $J(s)^\mathsf{T} H(x)$ comme

$$\begin{aligned} J(s)^\mathsf{T} H(x) &= F'_{\mathcal{F}(x)}(x)^\mathsf{T} F_{\mathcal{F}(x)}(x) + G'_{\mathcal{G}(x)}(x)^\mathsf{T} G_{\mathcal{G}(x)}(x) \\ &\quad + \frac{1}{2}[F'_{\mathcal{E}(x)}(x)^\mathsf{T} F_{\mathcal{E}(x)}(x) + G'_{\mathcal{E}(x)}(x)^\mathsf{T} G_{\mathcal{E}(x)}(x)] \\ &\quad + \frac{1}{2}[F'_{\mathcal{E}(x)}(x)^\mathsf{T} - G'_{\mathcal{E}(x)}(x)^\mathsf{T}] \operatorname{Diag}(s) H_{\mathcal{E}(x)} \\ &= F'_{\mathcal{F}(x)}(x)^\mathsf{T} F_{\mathcal{F}(x)}(x) + G'_{\mathcal{G}(x)}(x)^\mathsf{T} G_{\mathcal{G}(x)}(x) \\ &\quad + \frac{1}{2}[F'_{\mathcal{E}(x)}(x)^\mathsf{T} F_{\mathcal{E}(x)}(x) + G'_{\mathcal{E}(x)}(x)^\mathsf{T} G_{\mathcal{E}(x)}(x)] + \frac{1}{2}[\mathcal{M}_+ \ \mathcal{M}_-]s \\ &= g_1 + [X \ -Y]s = g_1 + V \operatorname{Diag}(H_{\mathcal{E}(x)}(x))s \end{aligned} \tag{D.7}$$

(où les indices de $\mathcal{E}^0(x)$ correspondraient à une colonne de X nulle, donc la valeur de s_i n'a pas d'importance)[3]. Rappelons que

$$\partial\theta(x) = \partial H(x)^\mathsf{T} H(x) = \operatorname{conv}(\partial_B H(x))^\mathsf{T} H(x), \qquad \partial_B H(x) = \operatorname{ext}(\partial_C H(x))$$

où la première égalité est la proposition de Clarke sur la règle de la chaîne, 2.3.19, la seconde est une propriété du différentiel de Clarke et la dernière est la proposition 3.4.14. De fait, les éléments de $\partial_B H$ sont définis par V (de façon équivalente, $[X\ Y]$ si $\mathcal{E}^0(x) = \varnothing$), alors que les éléments de $\partial\theta(x)$ sont gouvernés par $V \operatorname{Diag}(H_{\mathcal{E}(x)}(x)) = [X \ -Y]$).

Ainsi, les éléments extrémaux de $\partial\theta(x)$ (dans $\mathbb{R}^n$) correspondent aux vecteurs de signes appartenant à $\mathcal{S}(V, 0)$ (en restreignant le s dans (D.7) à $\mathcal{S}(V, 0)$) **et** qui correspondent aux sommets du zonotope défini par $[X \ -Y]$ (selon la dernière ligne de (D.7)). En particulier, cela illustre que $\partial\theta(x)$ est un polytope (l'enveloppe convexe d'un nombre fini de points).[4] Cependant, cela n'implique pas que les sommets de $\partial\theta(x)$ correspondent aux vecteurs de signes dans $\mathcal{S}([X\ Y]) \cap \mathcal{S}([X \ -Y])$: en effet, cet ensemble peut même être vide.

Contre-exemple D.2.2 (différences entre $[\boldsymbol{X\ Y}]$ et $[\boldsymbol{X\ -Y}]$)**.** Soit

$$X = I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad Y = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Par vérification directe, on a

$$\begin{aligned} \mathcal{S}([X\ Y]) &= \left\{ \begin{array}{cccc} (+,+,+,+), & (+,-,+,+), & (+,-,-,+), & (+,+,+,-) \\ (-,-,-,-), & (-,+,-,-), & (-,+,+,-), & (-,-,-,+) \end{array} \right\}, \\ \mathcal{S}([X\ -Y]) &= \left\{ \begin{array}{cccc} (-,+,+,+), & (-,-,+,+), & (-,-,+,-), & (+,-,+,-) \\ (+,+,-,-), & (-,+,-,-), & (+,+,-,+), & (-,+,-,+) \end{array} \right\}, \\ \mathcal{S}([X\ Y]) \cap \mathcal{S}([X\ -Y]) &= \varnothing, \qquad \mathcal{S}([X\ Y]) \cup \mathcal{S}([X\ -Y]) = \{\pm 1\}^4. \end{aligned}$$

3. En fait, pour tout $s \in \mathcal{S}(V, 0)$, il existe un $s' \in \mathcal{S}(V_{:,\mathcal{E}(x)\setminus\mathcal{E}^0(x)}, 0)$ avec $s_{\mathcal{E}(x)\setminus\mathcal{E}^0(x)} = s'$ – en utlisant les propriétés de l'arbre-$\mathcal{S}$ avec les indices de $\mathcal{E}^0(x)$ en dernier et en retirant les composantes de $\mathcal{E}^0(x)$.

4. Nous pensons que le C-différentiel de θ avec les linéarisations de F et G est également un zonotope, mais ce n'est qu'une conjecture ; on pourrait être tenté d'utiliser $\operatorname{conv}(\{g_1 + [X \ -Y]s : s \in \mathcal{S}([X\ Y])\}) = g_1 + [X \ -Y]\operatorname{conv}(\{s : s \in \mathcal{S}([X\ Y])\})$, mais le dernier terme est une enveloppe convexe de points (symétriques) de l'hypercube, qui est peu susceptible d'être facile à manipuler [262].

En particulier, les deux zonotopes $Z([X\ Y])$ et $Z([X\ -Y])$ coïncident, mais $[X\ -Y]$ évalué avec des vecteurs de signes de $[X\ Y]$ (ou équivalemment, $[X\ Y]$ évalué avec des vecteurs de signes de $[X\ -Y]$) n'a aucun sommet en commun, voir la figure D.1. □

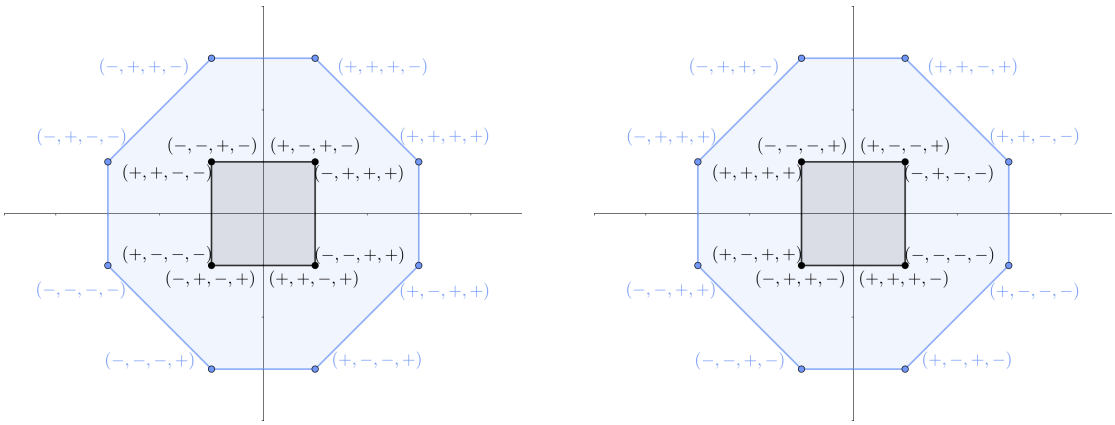


FIGURE D.1 – Gauche : $[X\ Y][-1,+1]^4$, sommets en bleu et autres points en noir (chaque point est deux vecteurs de signes). Droite : $[X\ -Y][-1,+1]^4$, sommets en bleu et autres points en noir (chaque point est deux vecteurs de signes). Schématiquement, le bleu clair correspond aux zonotopes avec $[X\ Y]$ et $[X\ -Y]$ et le noir à $\partial\theta(x)$.

La signification de cette observation est que $\partial\theta(x)$ a, en général, clairement moins de points extrémaux que $|\mathcal{S}(V,0)| = |\partial_B H(x)|$. Naturellement, si $\mathcal{E}^{0+}(x) = \varnothing$ (ou simplement $\mathcal{E}^{0+}(x) \setminus \mathcal{E}^0(x)$), $\mathcal{E}^-(x) = \varnothing$ ou $\mathcal{R}(X) \cap \mathcal{R}(Y) = \{0\}$, les points extrémaux de $\partial\theta(x)$ sont en bijection avec les éléments de $\mathcal{S}(V_{\mathcal{E}(x)\setminus\mathcal{E}^0(x)}, 0)$ [5].

Autrement dit, certaines directions sont multipliées positivement ($\mathcal{E}^{0+}(x)$) et d'autres négativement ($\mathcal{E}^-(x)$), ce qui perturbe l'extrémalité. Ceci est illustré dans le contre-exemple D.2.3 et lié à la proposition D.1.1.

Contre-exemple D.2.3 (la multiplication par $\boldsymbol{H}$ perturbe l'extrémalité)**.** Considérons les données suivantes, illustrées dans la figure D.2 :

$$F(x) = x = I_5 x + 0, \qquad G(x) = \begin{bmatrix} -1 & -16/5 & 0 & 0 & 0 \\ -2 & 21/5 & 0 & 0 & 0 \\ 0 & -5 & 1 & 0 & 0 \\ -2 & -3 & 0 & 1 & 0 \\ 3 & 1 & -1 & -1 & 0 \end{bmatrix} + \begin{bmatrix} 26/5 \\ -6/5 \\ 5 \\ 5 \\ -7 \end{bmatrix}.$$

5. De plus, $s \in \mathcal{S}(V_{\mathcal{E}(x)\setminus\mathcal{E}^0(x)}, 0)$ implique que $(s, s') \in \mathcal{S}(V, 0) \subseteq \{\pm 1\}^{\mathcal{E}(x)}$ par les propriétés récursives de l'arbre $\mathcal{S}$ (chapitre 3).

Pour $x = [1; 1; -1; -1; 0]$, on a clairement

$$F(x) = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 0 \end{bmatrix}, \qquad G(x) = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ -1 \end{bmatrix} \quad \Longrightarrow \quad \begin{cases} \mathcal{E}^{0+}(x) = \{1, 2\}, \\ \mathcal{E}^{-}(x) = \{3, 4\}, \\ \mathcal{F}(x) = \varnothing, \\ \mathcal{G}(x) = \{5\}. \end{cases}$$

D'après (D.3) et la règle 6.1.8, en omettant (x) par simplicité, on obtient :

$$\mathcal{M}_+ = (F'_{\mathcal{E}^{0+}(x)} - G'_{\mathcal{E}^{0+}(x)})^\mathsf{T} \operatorname{Diag}(H_{\mathcal{E}^{0+}(x)}) = \begin{bmatrix} 2 & 2 \\ 16/5 & -16/5 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 \\ \frac{16}{10} & -\frac{16}{10} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\mathcal{M}_- = (F'_{\mathcal{E}^{-}(x)} - G'_{\mathcal{E}^{-}(x)})^\mathsf{T} \operatorname{Diag}(H_{\mathcal{E}^{-}(x)}) = \begin{bmatrix} 0 & -2 \\ -5 & -3 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad Y = \begin{bmatrix} 0 & 1 \\ 5/2 & 3/2 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

et on a :

$$\begin{aligned} g_0(x) &= 0 + G'_5(x)^\mathsf{T} G_5(x) + G'_{\{1,2,3,4\}}(x)^\mathsf{T} G_{\{1,2,3,4\}}(x) \\ &= [-4\ 8\ 0\ 0\ 0]^\mathsf{T}, \end{aligned}$$

ainsi que :

$$\overline{x} - \overline{y} = g_0(x) + Xe - Ye = [-3\ 4\ 0\ 0\ 0]^\mathsf{T}.$$

L'"approche zonotope", utilisant les variables X, Y et $\overline{x} - \overline{y}$, est illustrée dans la figure D.3. Calculons les différentiels associés. Commençons par $\partial_B H(x)$. La matrice correspondante :

$$V := (G'_{\mathcal{E}(x)}(x) - F'_{\mathcal{E}(x)}(x))^\mathsf{T} = \begin{bmatrix} -2 & -2 & 0 & -2 \\ -16/5 & 16/5 & -5 & -3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

a pour vecteurs de signes associés aux matrices jacobiennes :

$$\mathcal{S} = \left\{ \begin{array}{cccc} (+,+,+,+), & (+,+,-,+), & (-,+,-,+), & (-,+,-,-) \\ (-,-,-,-), & (-,-,+,-), & (+,-,+,-), & (+,-,+,+) \end{array} \right\}.$$

Les systèmes correspondants (sans les zéros) se réduisent, à une symétrie près, à :

$$\begin{bmatrix} 1 & 16/10 \\ 1 & -16/10 \\ 0 & 5/2 \\ 1 & 3/2 \end{bmatrix} d > 0, \quad \begin{bmatrix} 1 & 16/10 \\ 1 & -16/10 \\ 0 & -5/2 \\ 1 & 3/2 \end{bmatrix} d > 0, \quad \begin{bmatrix} -1 & -16/10 \\ 1 & -16/10 \\ 0 & -5/2 \\ 1 & 3/2 \end{bmatrix} d > 0, \quad \begin{bmatrix} 1 & 16/10 \\ -1 & 16/10 \\ 0 & 5/2 \\ 1 & 3/2 \end{bmatrix} d > 0,$$

avec $d = \pm[2;1]$, $d = \pm[2;-1]$, $d = \pm[1.55;-1]$, $d = \pm[1;1]$. Les matrices jacobiennes $J(s)$ données par (D.5) pour $s \in \mathcal{S}$ sont :

$$J\begin{pmatrix}+\\+\\+\\+\end{pmatrix} = \begin{bmatrix}1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\3&1&-1&-1&0\end{bmatrix}, \qquad J\begin{pmatrix}-\\-\\-\\-\end{pmatrix} = \begin{bmatrix}-1&-16/5&0&0&0\\-2&21/5&0&0&0\\0&-5&1&0&0\\-2&-3&0&1&0\\3&1&-1&-1&0\end{bmatrix},$$

$$J\begin{pmatrix}+\\+\\-\\+\end{pmatrix} = \begin{bmatrix}1&0&0&0&0\\0&1&0&0&0\\0&-5&1&0&0\\0&0&0&1&0\\3&1&-1&-1&0\end{bmatrix}, \qquad J\begin{pmatrix}-\\-\\+\\-\end{pmatrix} = \begin{bmatrix}-1&-16/5&0&0&0\\-2&21/5&0&0&0\\0&0&1&0&0\\-2&-3&0&1&0\\3&1&-1&-1&0\end{bmatrix},$$

$$J\begin{pmatrix}-\\+\\-\\+\end{pmatrix} = \begin{bmatrix}-1&-16/5&0&0&0\\0&1&0&0&0\\0&-5&1&0&0\\0&0&0&1&0\\3&1&-1&-1&0\end{bmatrix}, \quad J\begin{pmatrix}+\\-\\+\\-\end{pmatrix} = \begin{bmatrix}1&0&0&0&0\\-2&21/5&0&0&0\\0&0&1&0&0\\-2&-3&0&1&0\\3&1&-1&-1&0\end{bmatrix},$$

$$J\begin{pmatrix}-\\+\\-\\-\end{pmatrix} = \begin{bmatrix}-1&-16/5&0&0&0\\0&1&0&0&0\\0&-5&1&0&0\\-2&-3&0&1&0\\3&1&-1&-1&0\end{bmatrix}, \quad J\begin{pmatrix}+\\-\\+\\+\end{pmatrix} = \begin{bmatrix}1&0&0&0&0\\-2&21/5&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\3&1&-1&-1&0\end{bmatrix}.$$

On rappelle que $\partial_B H(x) = \{J(s) : s \in \mathcal{S}\}$ et $\partial\theta(x) = \partial H(x)^\mathsf{T} H(x)$, avec :

$$\partial H(x)^\mathsf{T} H(x) = \mathrm{conv}(\partial_B H(x))^\mathsf{T} H(x) = \mathrm{conv}(\partial_B H(x)^\mathsf{T} H(x))$$

puisque $\cdot^\mathsf{T} H(x)$ est une transformation affine. Les vecteurs $(J^\mathsf{T} H)$ impliqués sont :

$$\begin{bmatrix}-2\\0\\0\\0\\0\end{bmatrix}, \begin{bmatrix}-2\\5\\0\\0\\0\end{bmatrix}, \begin{bmatrix}-4\\9/5\\0\\0\\0\end{bmatrix}, \begin{bmatrix}-2\\24/5\\0\\0\\0\end{bmatrix}, \begin{bmatrix}-4\\8\\0\\0\\0\end{bmatrix}, \begin{bmatrix}-4\\3\\0\\0\\0\end{bmatrix}, \begin{bmatrix}-2\\31/5\\0\\0\\0\end{bmatrix}, \begin{bmatrix}-4\\16/5\\0\\0\\0\end{bmatrix}. \tag{D.8}$$

Détaillons maintenant le processus d'optimalité retournant un $\gamma_{\mathcal{E}^{0+}(x)}$ et un $\gamma_{\mathcal{E}^{-}(x)}$ tels que $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^{-}(x)})$ soit non nul et vérifie la proposition 6.1.10. D'après la proposition C.0.7, le théorème est une équivalence et on a $\lambda^* = 5$, avec $\eta = [-1;-1]$. La projection sur Z_y est donnée par $\overline{y} + Y\zeta^*$ avec $\zeta^* = [1;-11/13]$.

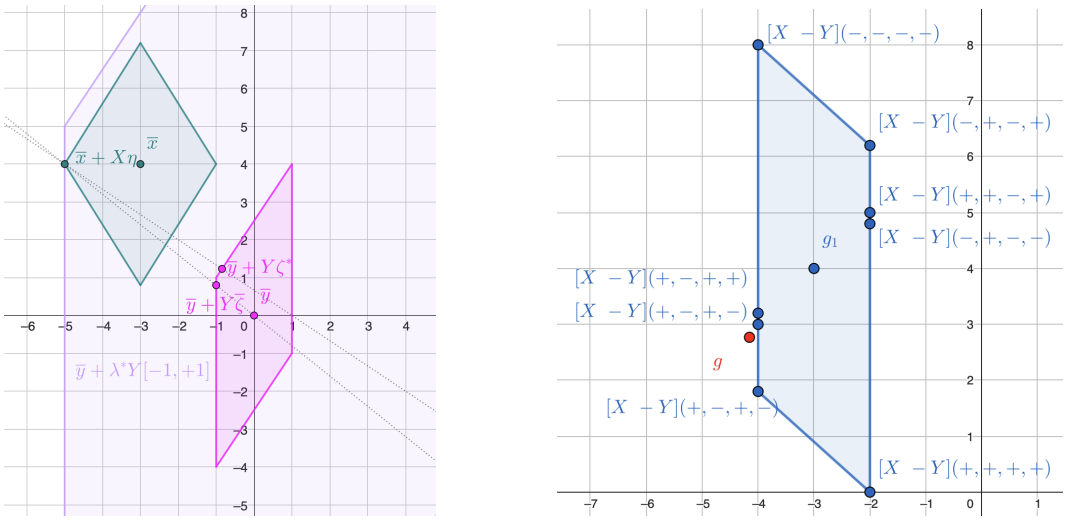


FIGURE D.2 – Illustration de l'aspect zonotope pour une situation présentant plusieurs difficultés. À gauche, le zonotope turquoise en haut à gauche est Z_x, tandis que Z_y est en magenta en bas à droite. Le violet clair représente la version dilatée (par λ^*) de Z_y. À droite, le zonotope bleu correspond à $\partial\theta(x)$ (voir (D.8)), les trois autres composantes nulles n'étant pas représentées. On observe que parmi les huit vecteurs de signes de $\partial_B H(x)$, seuls quatre forment l'enveloppe convexe du C-différentiel après multiplication par H. De plus, les "voisins" dans la figure ne correspondent pas à des vecteurs de signes adjacents. Enfin, comme décrit dans un exemple plus simple, la méthode de l'annexe C renvoie une valeur de $\mathcal{E}^{0+}(x)$ correspondant au point gauche de la zone turquoise (la dilatation du point inférieur de la frontière de Z_y, avec $\overline{\zeta}$), qui correspond à g (la projection est le point plus haut avec ζ^*) qui est le point rouge dans l'image de droite et est hors de $\partial\theta(x)$; cela provient du fait que les signes choisis ne sont pas dans $\mathcal{S}(V,0)$.

Enfin, le g optimal, donné par la proposition 6.1.10, vaut :

$$g\left(\eta = \begin{bmatrix}-1\\-1\end{bmatrix}, \zeta^* = \begin{bmatrix}1\\-\frac{11}{13}\end{bmatrix}\right) = \overline{x} - \overline{y} + X\eta - Y\zeta^*$$
$$= \begin{bmatrix}-3\\4\\0_3\end{bmatrix} + \begin{bmatrix}-2\\0\\0_3\end{bmatrix} - \begin{bmatrix}-\frac{11}{13}\\\frac{16}{13}\\0_3\end{bmatrix} = \begin{bmatrix}-4-\frac{2}{13}\\2+\frac{10}{13}\\0_3\end{bmatrix}$$

et n'est clairement pas une combinaison convexe des vecteurs de (D.8). □

Ce phénomène s'explique en partie par la proposition D.1.1.

D.2.1 Contre-exemples détaillés (simplifiés)

Le contre-exemple D.2.3 montre l'inadéquation de l'algorithme C.0.10 pour trouver un élément de $\partial\theta(x)$ (même s'il permet, par construction, de trouver un g non nul). Le pre-

mier contre-exemple ci-dessous, en dimension réduite, révèle que ce n'est pas une limite intrinsèque de l'algorithme C.0.10. Les suivants abordent des points plus techniques.

Contre-exemple D.2.4 (un $\gamma_{\mathcal{E}^{0+}(x)}$ incorrect conduit à $g \notin \partial\theta(x)$)**.** Considérons les données suivantes :

$$F(x) = x = I_3 x + 0, \qquad G(x) = \begin{bmatrix} 1 & -2 & 0 \\ 0 & -3 & 0 \\ 2 & -4 & 0 \end{bmatrix} + \begin{bmatrix} -2 \\ -2 \\ -7 \end{bmatrix}.$$

Pour $x = [1; -1; 0]$, on a clairement

$$F(x) = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \qquad G(x) = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix} \qquad \Longrightarrow \qquad \begin{cases} \mathcal{E}^{0+}(x) = \{1\}, & \mathcal{E}^{-}(x) = \{2\}, \\ \mathcal{F}(x) = \varnothing, & \mathcal{G}(x) = \{3\}. \end{cases}$$

D'après (D.3) et la règle 6.1.8, on obtient :

$$\begin{aligned} \mathcal{M}_+ = (F'_{\mathcal{E}^{0+}(x)} - G'_{\mathcal{E}^{0+}(x)})^{\mathsf{T}} \operatorname{Diag}(H_{\mathcal{E}^{0+}(x)}) &= \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}, \quad X = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \\ \mathcal{M}_- = (F'_{\mathcal{E}^{-}(x)} - G'_{\mathcal{E}^{-}(x)})^{\mathsf{T}} \operatorname{Diag}(H_{\mathcal{E}^{-}(x)}) &= \begin{bmatrix} 0 \\ -4 \\ 0 \end{bmatrix}, \quad Y = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}. \end{aligned}$$

De plus :

$$\begin{aligned} g_0(x) &= 0 + G'_3(x)^{\mathsf{T}} G_3(x) + G'_{\{1,2\}}(x)^{\mathsf{T}} G_{\{1,2\}}(x) \\ &= [-1\ 5\ 0]^{\mathsf{T}}, \end{aligned}$$

et :

$$\overline{x} - \overline{y} = g_0(x) + Xe - Ye = [-1\ 4\ 0]^{\mathsf{T}}.$$

L'"approche zonotope" utilisant X, Y et $\overline{x} - \overline{y}$ est illustrée dans la figure D.3. Calculons les différentiels. La matrice :

$$V := (G'_{\mathcal{E}(x)}(x) - F'_{\mathcal{E}(x)}(x))^{\mathsf{T}} = \begin{bmatrix} 0 & 0 \\ -2 & -4 \\ 0 & 0 \end{bmatrix}$$

a pour vecteurs de signes associés :

$$\mathcal{S} = \{(+,+), (-,-)\}.$$

Les systèmes correspondants sont :

$$\begin{bmatrix} 0 & -2 & 0 \\ 0 & -4 & 0 \end{bmatrix} d > 0, \qquad \begin{bmatrix} 0 & 2 & 0 \\ 0 & 4 & 0 \end{bmatrix} d > 0,$$

avec $d = [0; -1; 0]$ et $d = [0; 1; 0]$. Les matrices jacobiennes $J(s)$ données par (D.5) sont :

$$J\begin{pmatrix}-\\-\end{pmatrix} = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 2 & -4 & 0\end{bmatrix}, \qquad J\begin{pmatrix}+\\+\end{pmatrix} = \begin{bmatrix}1 & -2 & 0\\ 0 & -3 & 0\\ 2 & -4 & 0\end{bmatrix}.$$

On a $\partial_B H(x) = \{J(s) : s \in \mathcal{S}\}$ et $\partial\theta(x) = \mathrm{conv}(\partial_B H(x)^\intercal H(x))$, avec les vecteurs :

$$\begin{bmatrix}-1\\ 3\\ 0\end{bmatrix} \text{ et } \begin{bmatrix}-1\\ 5\\ 0\end{bmatrix}. \tag{D.9}$$

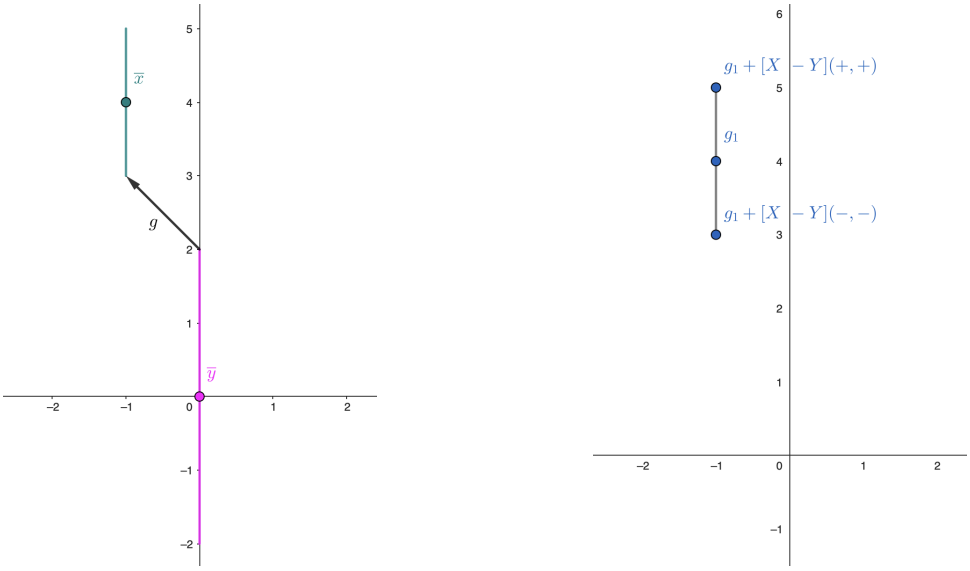


FIGURE D.3 – Illustration du contre-exemple. À gauche : les zonotopes (turquoise pour Z_x, magenta pour Z_y), la flèche représente g pour $\eta = -1$, qui n'appartient pas à $\partial\theta(x)$. À droite : $\partial\theta(x)$ et les éléments de $\partial_B H(x)^\intercal H(x)$; g_1 est le centre du différentiel.

Enfin, pour tout $\gamma_{\mathcal{E}^{0+}(x)}$, la projection est $[0; 2; 0]$ (i.e., $\gamma_{\mathcal{E}^-(x)} = 1$), d'où :

$$g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}) = [-1; 1 + 2\gamma_{\mathcal{E}^{0+}(x)}],$$

qui n'appartient à $\partial\theta(x)$ que si $\gamma_{\mathcal{E}^{0+}(x)} = 1$ (le point maximisant la distance, unique ici). □

Les contre-exemples suivants détaillent les difficultés liées à l'obtention d'un élément du différentiel, notamment des "dégénérescences" similaires aux problèmes d'optimisation linéaire dégénérés lorsque le vecteur coût est orthogonal à une face du domaine réalisable. Le dernier exemple montre une limite de l'algorithme C.0.10 : même sans projection (garantissant que $-g$ est une direction de descente), l'élément peut ne pas appartenir au différentiel.

Contre-exemple D.2.5 (les dégénérescences entraînent $\overline{g} \notin \partial\theta(x)$)**.** Considérons les données suivantes :

$$F(x) = x = I_5 x + 0, \qquad G(x) = \begin{bmatrix} 3 & 0 & -2 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix} + \begin{bmatrix} -4 \\ 2 \\ -2 \\ -2 \\ -15 \end{bmatrix}.$$

Pour $x = [1; -1; -1; -1; 0]$, on a clairement

$$F(x) = \begin{bmatrix} 1 \\ -1 \\ -1 \\ -1 \\ 0 \end{bmatrix}, \qquad G(x) = \begin{bmatrix} 1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix} \qquad \Longrightarrow \qquad \begin{cases} \mathcal{E}^{0+}(x) = \{1\} \\ \mathcal{E}^{-}(x) = \{2, 3, 4\}, \\ \mathcal{G}(x) = \{5\}, \\ \mathcal{F}(x) = \varnothing. \end{cases}$$

D'après (D.3) et la règle 6.1.8, on obtient :

$$\mathcal{M}_+ = (F'_{\mathcal{E}^{0+}(x)} - G'_{\mathcal{E}^{0+}(x)})^\mathsf{T} \operatorname{Diag}(H_{\mathcal{E}^{0+}(x)}) = \begin{bmatrix} -2 \\ 0 \\ 2 \\ 0 \\ 0 \end{bmatrix}, \qquad X = \begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix},$$

$$\mathcal{M}_- = (F'_{\mathcal{E}^{-}(x)} - G'_{\mathcal{E}^{-}(x)})^\mathsf{T} \operatorname{Diag}(H_{\mathcal{E}^{-}(x)}) = \begin{bmatrix} -2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad Y = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

De plus :

$$\begin{aligned} g_0(x) &= 0 + G'_5(x)^\mathsf{T} G_5(x) + G'_{\{1,2,3,4\}}(x)^\mathsf{T} G_{\{1,2,3,4\}}(x) \\ &= [1\ 6\ 3\ 0\ 0]^\mathsf{T}, \end{aligned}$$

ainsi que :

$$\overline{x} - \overline{y} = g_0(x) + Xe - Ye = [-1\ 5\ 3\ 0\ 0]^\mathsf{T}.$$

L'"approche zonotope" utilisant X, Y et $\overline{x} - \overline{y}$ est illustrée dans la figure D.4. Calculons les différentiels associés. Commençons par $\partial_B H(x)$. La matrice correspondante :

$$V := (G'_{\mathcal{E}(x)}(x) - F'_{\mathcal{E}(x)}(x))^\mathsf{T} = \begin{bmatrix} 2 & -2 & 0 & 0 \\ 0 & 0 & -2 & 0 \\ -2 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

a pour vecteurs de signes associés aux matrices jacobiennes :

$$\left\{ \begin{array}{l} (+,+,+,+),(+,+,-,+),(+,-,+,+),(+,-,-,+),(+,-,+,-),(+,-,-,-) \\ (-,-,-,-),(-,-,+,-),(-,+,-,-),(-,+,+,-),(-,+,-,+),(-,+,+,+) \end{array} \right\}.$$

Les systèmes correspondants se réduisent, à une symétrie près, à :

$$\begin{bmatrix} -1 & 0 & 1 & 0_2^{\mathsf{T}} \\ 2 & 0 & 0 & 0_2^{\mathsf{T}} \\ 0 & 2 & 0 & 0_2^{\mathsf{T}} \\ 0 & 0 & 2 & 0_2^{\mathsf{T}} \end{bmatrix} d > 0, \quad \begin{bmatrix} -1 & 0 & 1 & 0_2^{\mathsf{T}} \\ -2 & 0 & 0 & 0_2^{\mathsf{T}} \\ 0 & 2 & 0 & 0_2^{\mathsf{T}} \\ 0 & 0 & 2 & 0_2^{\mathsf{T}} \end{bmatrix} d > 0, \quad \begin{bmatrix} -1 & 0 & 1 & 0_2^{\mathsf{T}} \\ -2 & 0 & 0 & 0_2^{\mathsf{T}} \\ 0 & 2 & 0 & 0_2^{\mathsf{T}} \\ 0 & 0 & -2 & 0_2^{\mathsf{T}} \end{bmatrix} d > 0,$$

$$\begin{bmatrix} -1 & 0 & 1 & 0_2^{\mathsf{T}} \\ 2 & 0 & 0 & 0_2^{\mathsf{T}} \\ 0 & -2 & 0 & 0_2^{\mathsf{T}} \\ 0 & 0 & 2 & 0_2^{\mathsf{T}} \end{bmatrix} d > 0, \quad \begin{bmatrix} -1 & 0 & 1 & 0_2^{\mathsf{T}} \\ -2 & 0 & 0 & 0_2^{\mathsf{T}} \\ 0 & -2 & 0 & 0_2^{\mathsf{T}} \\ 0 & 0 & 2 & 0_2^{\mathsf{T}} \end{bmatrix} d > 0, \quad \begin{bmatrix} -1 & 0 & 1 & 0_2^{\mathsf{T}} \\ -2 & 0 & 0 & 0_2^{\mathsf{T}} 0 \\ 0 & -2 & 0 & 0_2^{\mathsf{T}} \\ 0 & 0 & -2 & 0_2^{\mathsf{T}} \end{bmatrix} d > 0.$$

où l'on peut prendre (en ignorant les coordonnées 4 et 5) $d = \pm[1;1;2]$, $d = \pm[-1;1;1]$, $d = \pm[-2;1;-1]$, $d = \pm[1;-1;2]$, $d = \pm[-1;-1;1]$, $d = \pm[-2;-1;-1]$. Les matrices

jacobiennes $J(s)$ données par (D.5) pour $s \in \mathcal{S}$ sont :

$$J(+,+,+,+) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \; J(-,-,-,-) = \begin{bmatrix} 3 & 0 & -2 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix},$$

$$J(+,+,-,+) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \; J(-,-,+,-) = \begin{bmatrix} 3 & 0 & -2 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix},$$

$$J(+,-,+,+) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \; J(-,+,-,-) = \begin{bmatrix} 3 & 0 & -2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix},$$

$$J(+,-,-,+) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \; J(-,+,+,-) = \begin{bmatrix} 3 & 0 & -2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix},$$

$$J(+,-,+,-) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \; J(-,+,-,+) = \begin{bmatrix} 3 & 0 & -2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix},$$

$$J(+,-,-,-) = \begin{bmatrix} 3 & 0 & -2 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \; J(-,+,+,+) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}.$$

On rappelle que $\partial_B H(x) = \{J(s) : s \in \mathcal{S}\}$ et $\partial\theta(x) = \partial H(x)^\intercal H(x)$, avec :

$$\partial H(x)^\intercal H(x) = \mathrm{conv}(\partial_B H(x))^\intercal H(x) = \mathrm{conv}(\partial_B H(x)^\intercal H(x))$$

puisque $\cdot^\intercal H(x)$ est une transformation affine. Les vecteurs impliqués dans cette enveloppe convexe sont :

$$\begin{bmatrix} -3 \\ 4 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 6 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -3 \\ 6 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 4 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 4 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 6 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 6 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 4 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 4 \\ 5 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 6 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 6 \\ 5 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 4 \\ 1 \\ 0 \\ 0 \end{bmatrix}. \tag{D.10}$$

Détaillons maintenant l'algorithme C.0.10 et le problème correspondant : on a $\lambda^* = 5$, mais il y a dégénérescence car la valeur de η n'est pas pertinente : tout point de Z_x nécessite une dilatation de λ^*.

Prenons par exemple $\eta = -1$, on obtient $z = [0; 5; 2]$ puis $\overline{\zeta} = [0; 1; 2/5]$. La valeur correspondante de $\overline{g}$ est $\overline{g} = [0; 4; 8/5]$, qui n'appartient pas à l'enveloppe convexe des vecteurs ci-dessus. En revanche, pour un autre η (conduisant à la même valeur de λ^*), comme $\eta = +1$, on obtient $\overline{\zeta} = [-2/5; 1; 4/5]$ et $\overline{g} = [-8/5; 4; 16/5]$, qui appartient bien au différentiel.

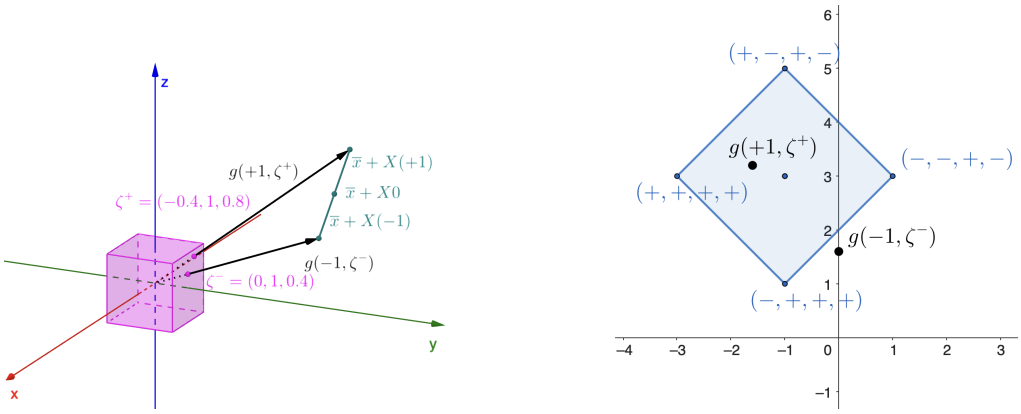


FIGURE D.4 – Illustration du contre-exemple. À gauche : les zonotopes correspondants (turquoise pour Z_x, magenta pour Z_y), les flèches représentent g pour $\eta = -1$ (hors de $\partial\theta(x)$) et $\eta = +1$ (dans $\partial\theta(x)$). À droite : illustration de $\partial\theta(x)$ (dans le plan $x_2 = 4$). Comme dans le contre-exemple D.2.3, tous les vecteurs de signes ne sont pas extrémaux ($(+,-,+,+)$ et $(-,+,+,-)$ correspondent au point bleu central) et selon η, g peut appartenir ou non à $\partial\theta(x)$. Les 6 autres vecteurs de signes correspondent à la face dans le plan $x_2 = 6$.

Puisque ces deux valeurs ont une deuxième composante égale à 4, elles doivent être une combinaison convexe des éléments ayant aussi une deuxième composante égale à 4 (les autres vecteurs l'ayant à 6). Or :

$$\begin{bmatrix} 0 \\ 8/5 \end{bmatrix} \notin \operatorname{conv}\left(\begin{bmatrix} -3 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ 5 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} -1 \\ 3 \end{bmatrix} + B(0,2)_{||\cdot||_1},$$
$$\begin{bmatrix} -8/5 \\ 16/5 \end{bmatrix} \in \operatorname{conv}\left(\begin{bmatrix} -3 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ 5 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} -1 \\ 3 \end{bmatrix} + B(0,2)_{||\cdot||_1},$$

ce qui indique bien qu'un des points est dans le différentiel et pas l'autre. □

Le contre-exemple suivant décrit une difficulté technique rencontrée dans les propositions D.2.7 et D.2.8. Il est obtenu par élévation du contre-exemple D.2.4.

Contre-exemple D.2.6 (dégénérescences)**.** Considérons les données suivantes :

$$F(x) = x = I_5 x + 0, \qquad G(x) = \begin{bmatrix} 1 & -2 & 0 & 0 & 0 \\ 0 & 1 & -2 & 0 & 0 \\ 0 & -4 & 1 & 0 & 0 \\ 0 & 0 & -4 & 1 & 0 \\ 2 & -2 & 0 & 1 & 0 \end{bmatrix} + \begin{bmatrix} 2 \\ -2 \\ 4 \\ -4 \\ -2 \end{bmatrix}.$$

Pour $x = [1; 1; -1; -1; 0]$, on a clairement

$$F(x) = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 0 \end{bmatrix}, \qquad G(x) = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ -1 \end{bmatrix} \qquad \Longrightarrow \qquad \begin{cases} \mathcal{E}^{0+}(x) = \{1, 2\}, \\ \mathcal{E}^{-}(x) = \{3, 4\}, \\ \mathcal{F}(x) = \varnothing, \\ \mathcal{G}(x) = \{5\}. \end{cases}$$

D'après (D.3) et la règle 6.1.8, on obtient :

$$\mathcal{M}_+ = (F'_{\mathcal{E}^{0+}(x)} - G'_{\mathcal{E}^{0+}(x)})^{\mathsf{T}} \operatorname{Diag}(H_{\mathcal{E}^{0+}(x)}) = \begin{bmatrix} 0 & 0 \\ 2 & 0 \\ 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad X = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\mathcal{M}_- = (F'_{\mathcal{E}^{-}(x)} - G'_{\mathcal{E}^{-}(x)})^{\mathsf{T}} \operatorname{Diag}(H_{\mathcal{E}^{-}(x)}) = \begin{bmatrix} 0 & 0 \\ -4 & 0 \\ 0 & -4 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad Y = \begin{bmatrix} 0 & 0 \\ 2 & 0 \\ 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

De plus :

$$\begin{aligned} g_0(x) &= 0 + G'_5(x)^{\mathsf{T}} G_5(x) + G'_{\{1,2,3,4\}}(x)^{\mathsf{T}} G_{\{1,2,3,4\}}(x) \\ &= [-1\ 5\ 1\ 0\ 0]^{\mathsf{T}}, \end{aligned}$$

ainsi que :

$$\overline{x} - \overline{y} = g_0(x) + Xe - Ye = [-1\ 4\ 0\ 0\ 0]^{\mathsf{T}}.$$

L'"approche zonotope" utilisant X, Y et $\overline{x} - \overline{y}$ est illustrée dans la figure D.5. Calculons les différentiels associés. Considérons d'abord $\partial_B H(x)$. La matrice correspondante :

$$V := (G'_{\mathcal{E}(x)}(x) - F'_{\mathcal{E}(x)}(x))^{\mathsf{T}} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -2 & 0 & -4 & 0 \\ 0 & -2 & 0 & -2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

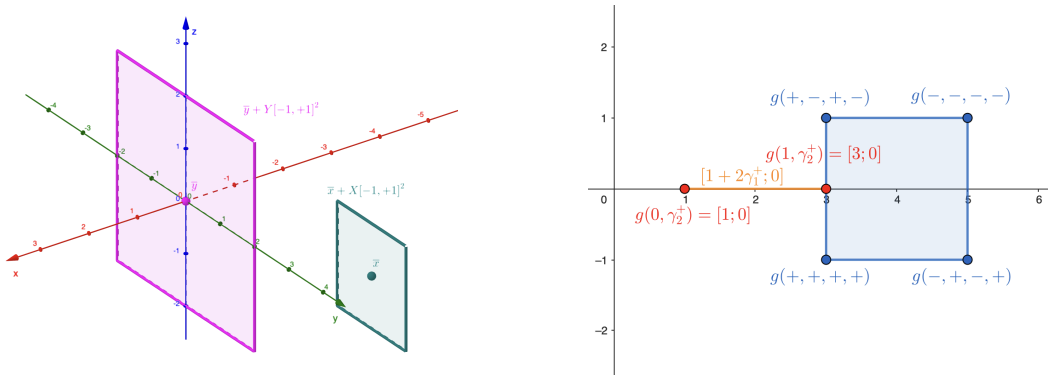


FIGURE D.5 – Illustration des dégénérescences, obtenues en ajoutant artificiellement une dimension (la troisième dimension est omise à droite, dans le plan $x_1 = -1$). À gauche : les zonotopes décrits par les équations précédentes. À droite : les gradients g obtenus ; le carré bleu représente le différentiel de θ, ses sommets correspondent à des vecteurs de signes non "adjacents" (dû ici à la colinéarité des vecteurs de X et Y). Le segment orange représente les g possibles selon $\mathcal{E}^{0+}(x)$, les points rouges ses sommets. Les sommets du différentiel (bleu) correspondent au zonotope exprimé dans (D.7).

a pour vecteurs de signes associés aux matrices jacobiennes :

$$\mathcal{S} = \{(+,+,+,+), (-,-,-,-), (+,-,+,-), (-,+,-,+)\}.$$

Les systèmes correspondants se réduisent, à une symétrie près, à :

$$\pm \begin{bmatrix} 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 \\ 0 & -4 & 0 & 0 & 0 \\ 0 & 0 & -4 & 0 & 0 \end{bmatrix} d > 0, \ \pm \begin{bmatrix} 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & -4 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 \end{bmatrix} d > 0,$$

où l'on peut prendre $d = \pm[0;-1;-1;0;0]$ et $d = \pm[0;-1;1;0;0]$. Les matrices jacobiennes $J(s)$ données par (D.5) pour $s \in \mathcal{S}$ sont :

$$J\begin{pmatrix}+\\+\\+\\+\end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 2 & -2 & 0 & -1 & 0 \end{bmatrix}, \quad J\begin{pmatrix}-\\-\\-\\-\end{pmatrix} = \begin{bmatrix} 1 & -2 & 0 & 0 & 0 \\ 0 & 1 & -2 & 0 & 0 \\ 0 & -4 & 1 & 0 & 0 \\ 0 & 0 & -4 & 1 & 0 \\ 2 & -2 & 0 & -1 & 0 \end{bmatrix},$$

$$J\begin{pmatrix}+\\-\\+\\-\end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -4 & 1 & 0 \\ 2 & -2 & 0 & -1 & 0 \end{bmatrix}, \quad J\begin{pmatrix}-\\+\\-\\+\end{pmatrix} = \begin{bmatrix} 1 & -2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 2 & -2 & 0 & -1 & 0 \end{bmatrix}.$$

On rappelle que $\partial_B H(x) = \{J(s) : s \in \mathcal{S}\}$ et $\partial\theta(x) = \partial H(x)^\intercal H(x)$, avec :

$$\partial H(x)^\intercal H(x) = \mathrm{conv}(\partial_B H(x))^\intercal H(x) = \mathrm{conv}(\partial_B H(x)^\intercal H(x))$$

puisque $\cdot^{\mathsf{T}} H(x)$ est une transformation affine. Les vecteurs impliqués dans cette enveloppe convexe sont :

$$\begin{bmatrix} -1 \\ 3 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 5 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 3 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 5 \\ -1 \\ 0 \\ 0 \end{bmatrix}. \tag{D.11}$$

Détaillons où interviennent les dégénérescences. Soit $\gamma_{\mathcal{E}^{0+}(x)} = (\gamma_1^+, \gamma_2^+)$. Après application de la projection (proposition 6.1.10), on a :

$$g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^{-}(x)}) = [-1; 2 + 2\gamma_1^+ - 1; 0; 0; 0]^{\mathsf{T}} = [-1; 1 + 2\gamma_1^+; 0; 0; 0]^{\mathsf{T}}.$$

Ceci est illustré dans la figure D.5. On observe que le choix de γ_2^+ est sans effet : quel que soit γ_2^+, le même g est obtenu après projection.

Considérons un point extrémal, i.e., $\eta_2 \in \{\pm 1\}$ avec $\eta_1 = 1$. La projection correspond alors à $\zeta = (1, \pm 1/2)$. Les sommets de la face contenant ζ sont $\hat{\zeta}^+ := (+1, +1)$ et $\hat{\zeta}^- := (+1, -1)$ pour $\eta_2 = 1$ ou $\eta_2 = -1$.

Idéalement, les vecteurs de signes correspondants devraient appartenir à $\mathcal{S}$:

$$\eta_2 = +1 \Longrightarrow (+,+,+,+), (+,+,+,-); \qquad \eta_2 = -1 \Longrightarrow (+,-,+,+), (+,-,+,-)$$

mais pour toute valeur de γ_2^+, l'un des deux signes n'est pas dans $\mathcal{S}$. Comme ce phénomène est inévitable (s'il se produit), la preuve procède ainsi : pour le sommet $\hat{\zeta}^+$, une direction $\hat{d}^+ \in \mathcal{R}(Y)$ vérifie $\hat{d}_1^+ = 0$, $\hat{d}_2^+ > 0$ et $\hat{d}_3^+ > 0$. Pour $\hat{\zeta}^-$, une direction $\hat{d}^- \in \mathcal{R}(Y)$ vérifie $\hat{d}_1^- = 0$, $\hat{d}_2^- > 0$ et $\hat{d}_3^- < 0$. De plus, $\overline{c} := [0; 1; 0]$. Ainsi, pour $\hat{\zeta}^+$ et $\hat{d}^+$, on a $\hat{c}_1 = 0$, $\hat{c}_2 > 1 > 0$, et $\hat{c}_3 > 0$ (les deux dernières dimensions sont sans importance), donc $\tilde{\eta}_2 = +1$. De même, pour $\hat{\zeta}^-$, $\hat{d}_1^- = 0$, $\hat{d}_2^- > 0$ et $\hat{d}_3^- < 0$, conduisant à $\tilde{\eta}_2 = -1$.

Le choix de η_2 ne modifie pas g : seule la preuve théorique nécessite une combinaison convexe adéquate. Notons que $\eta_2 = 0$ est la demi-somme de deux éléments du B-différentiel :

$$[-1; 3; 0; 0; 0] = \frac{1}{2}([-1; 3; -1; 0; 0] + [-1; 3; 1; 0; 0]).$$

Toutefois, si le zonotope turquoise était décalé vers le haut/bas, un choix approprié serait nécessaire au lieu de 0. Par exemple, si $\overline{x} = [-1; 4; 1/2]$, g après projection serait inchangé (pour η_1 fixé). Mais $\zeta_2 = (1 + 2\eta_2)/4$, et le "système" à résoudre pour trouver une combinaison convexe devient :

$$\begin{pmatrix} +1 \\ \eta_2 \\ +1 \\ (1+2\eta_2)/4 \end{pmatrix} = t \begin{pmatrix} +1 \\ +1 \\ +1 \\ +1 \end{pmatrix} + (1-t) \begin{pmatrix} +1 \\ -1 \\ +1 \\ -1 \end{pmatrix} \quad \Longleftrightarrow \quad \begin{cases} \eta_2 = 2t - 1 \\ 2\eta_2 + 1 = 4(2t-1) \end{cases}$$

dont la solution est $(\eta_2, t) = (1/2, 3/4)$. □

D.2.2 Dégénérescences et corrections (théoriques)

Premièrement, nous considérons une propriété plus simple concernant les variables $\bar{\cdot}$ du chapitre C.

Proposition D.2.7 (propriété de $\overline{\boldsymbol{\zeta}}$ et $\overline{\boldsymbol{g}}$)**.** *Avec les mêmes notations, il existe une modification $\tilde{\eta}$ de $\overline{\eta}$ telle que $\overline{g} = \overline{x} + X\tilde{\eta} - \overline{y} - Y\overline{\zeta} \in \partial\theta(x)$.*

Preuve. Rappelons que l'on a $\overline{x} + X\eta = \overline{y} + \lambda^* Y\overline{\zeta}$ et posons $\overline{g} := \overline{x} + X\overline{\eta} - \overline{y} - Y\overline{\zeta}$. Il est clair que $\overline{y} + Y\overline{\zeta}$ appartient à la frontière de Z_y, donc à l'intérieur d'une unique face $\overline{F}$, voir la proposition B.1.4. D'après les propositions B.2.2 et B.2.6, soit $I^* := \{i \in \mathcal{E}^-(x) : \overline{\zeta}_i \in \{-1, +1\}\}$ et $I^{\overline{F}} := \{i \in \mathcal{E}^-(x) : \overline{\zeta}_i \in (-1, +1)\}$. En utilisant la proposition B.1.7, soit $\overline{c}$ une "normale" associée à $\overline{F}$. D'après la proposition B.2.3, on a

$$i \in I^* \quad \Longrightarrow \quad \overline{\zeta}_i Y_{:,i}^{\mathsf{T}} \overline{c} > 0, \qquad i \in I^{\overline{F}} \quad \Longrightarrow \quad Y_{:,i}^{\mathsf{T}} \overline{c} = 0.$$

Clairement, $\overline{y} + Y\overline{\zeta}$ est une combinaison convexe des sommets de $\overline{F}$, qui ont pour expression $\overline{y} + Y_{:,I^*}\overline{\zeta}_{I^*} + Y_{:,I^{\overline{F}}}\hat{\zeta}_{I^{\overline{F}}}$, c'est-à-dire

$$\overline{F} = \overline{y} + Y_{:,I^*}\overline{\zeta}_{I^*} + \operatorname{conv}(\{Y_{:,I^{\overline{F}}}\hat{\zeta}_{I^{\overline{F}}},\ \hat{\zeta}_{I^{\overline{F}}} \in \mathcal{S}(Y_{:,I^{\overline{F}}})\})$$

($\overline{y} + Y\overline{\zeta}$ est une combinaison convexe des $\overline{y} + Y\hat{\zeta}$). Ainsi, si l'on pouvait montrer que toutes les matrices jacobiennes correspondant aux signes $[\overline{\eta}; \hat{\zeta}]$ sont dans $\partial_B H(x)$, alors on aurait

$$\begin{bmatrix} \overline{\eta} \\ \overline{\zeta} \end{bmatrix} = \sum t_i \begin{bmatrix} \overline{\eta} \\ \hat{\zeta}^i \end{bmatrix}, \overline{J} := J([\overline{\eta}; \overline{\zeta}]) = \sum t_i J([\overline{\eta}; \hat{\zeta}^i]), \overline{J}^{\mathsf{T}} H(x) = \left[\sum t_i J([\overline{\eta}; \hat{\zeta}^i])\right]^{\mathsf{T}} H(x).$$

Cependant, comme décrit dans le contre-exemple D.2.5, ceci peut ne pas être vérifié en raison des "dégénérescences". Détailons une méthode pour procéder. Tout d'abord, observons les indices de $\mathcal{E}^{0+}(x)$. Montrons d'abord que $\overline{\eta}_i X_{:,i}^{\mathsf{T}} \overline{c} \geqslant 0$. Supposons qu'il existe un $i \in \mathcal{E}^{0+}(x)$ tel que $\overline{\eta}_i X_{:,i}^{\mathsf{T}} \overline{c} < 0$. Avec des arguments similaires à ceux des annexes précédentes, le point $\overline{x} + X\overline{\eta} - 2\overline{\eta}_i X_{:,i} \in Z_x$ serait "au-delà" de l'hyperplan orthogonal à $\overline{c}$ et plus éloigné de Z_y, ce qui contredit l'optimalité de λ^*. Par conséquent, $\overline{c}$ est tel que

$$\forall\, i \in \mathcal{E}^{0+}(x), \overline{\eta}_i X_{:,i}^{\mathsf{T}} \overline{c} \geqslant 0, \qquad \forall\, i \in I^*, \overline{\zeta}_i y_i^{\mathsf{T}} \overline{c} > 0, \qquad \forall\, i \in I^{\overline{F}}, y_i^{\mathsf{T}} \overline{c} = 0.$$

Concentrons-nous maintenant sur les indices tels que $X_{:,i}^{\mathsf{T}} \overline{c} = 0$. Montrons que cela implique $X_{:,i} \in L_{\overline{F}} = \operatorname{aff}(F) - \operatorname{aff}(F)$. En effet, soit $\overline{H} := \overline{c}^{\perp} + z^{\overline{F}}$ pour $z^{\overline{F}} := \overline{y} + Y_{:,I^*}\overline{\zeta}_{I^*}$ le centre de $\overline{F}$, $z^{\overline{F}} \in \overline{F}$, on a $\overline{F} = Z_y \cap \overline{H}$. Soit

$$\begin{aligned} Z_y^* &:= \overline{y} + \lambda^* Y[-1, +1]^{\mathcal{E}^-(x)} = \overline{y} + \lambda^*(Z_y - \overline{y}), \\ \overline{F}^* &:= \overline{y} + \lambda^*(\overline{F} - \overline{y}) \qquad \text{et} \\ \overline{H}^* &:= \overline{c}^{\perp} + \overline{y} + \lambda^*(z^{\overline{F}} - \overline{y}) \end{aligned}$$

les éléments pertinents liés à la dilatation par λ^* du zonotope Z_y. On a $\overline{F}^* = Z_y^* \cap \overline{H}^*$, voir la figure D.6.

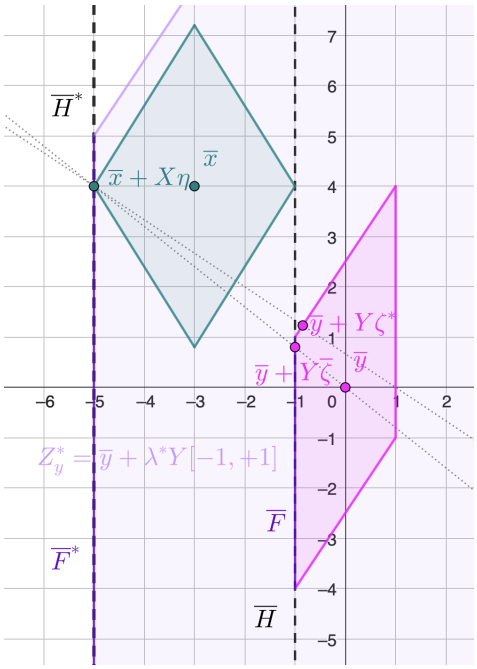


Figure D.6 – Illustration des quantités dilatées, avec les données du contre-exemple D.2.3.

Par conséquent, supposons que pour un certain $i \in \mathcal{E}^{0+}(x)$, $X_{:,i}^\mathsf{T}\overline{c} = 0$ (c'est-à-dire $X_{:,i} \in \overline{c}^\perp = \overline{H} - \overline{H}$) et $X_{:,i} \notin L_{\overline{F}}$. Alors, pour $\delta \in [0, 2]$, le point $z^* := \overline{x} + X\eta - \delta\eta_i X_{:,i}$ dans Z_x par définition, appartient à $\overline{H}^*$ puisque

$$\begin{aligned} z^* = \overline{x} + X\eta - \delta\eta_i X_{:,i} &= \overline{y} + \lambda^* Y\overline{\zeta} - \delta\eta_i X_{:,i} \\ &= \overline{y} + \lambda^* Y_{:,I^*}\overline{\zeta}_{I^*} + \lambda^* Y_{:,I^{\overline{F}}}\overline{\zeta}_{I^F} - \delta\eta_i X_{:,i} \\ &= \overline{y} + \lambda^*(z^{\overline{F}} - \overline{y}) + \lambda^* Y_{:,I^{\overline{F}}}\overline{\zeta}_{I^F} - \delta\eta_i X_{:,i} \\ &\in \overline{y} + \lambda^*(z^{\overline{F}} - \overline{y}) + c^\perp = \overline{H}^*, \end{aligned}$$

où la première égalité provient du problème primal-dual, la seconde sépare les indices de $\mathcal{E}^-(x)$, la troisième réarrange les termes pour faire apparaître $z^{\overline{F}}$, et la dernière ligne vient des propriétés de $\overline{c}$ et de la définition de $\overline{H}^*$. Cependant,

$$\begin{aligned} z^* = \overline{x} + X\eta - \delta\eta_i X_{:,i} = \overline{y} + \lambda^* Y\overline{\zeta} - \delta\eta_i X_{:,i} \\ \in \overline{F}^* - \delta\eta_i X_{:,i}, \end{aligned}$$

donc $z^* \notin \overline{F}^*$ puisque nous avons supposé $X_{:,i} \notin L_{\overline{F}}$. Ainsi, en utilisant $\overline{F}^* = Z_y^* \cap \overline{H}^*$

$$z^* \notin \overline{F}^*, z^* \in \overline{H}^* \quad \Longrightarrow \quad z^* \notin Z_y^*.$$

Ceci signifie qu'il existe un point de Z_x qui n'est pas dans Z_y^*, ce qui contredit l'optimalité de λ^*. Définissons maintenant $\mathcal{E}^{0+}(x)_0 := \{i \in \mathcal{E}^{0+}(x) : X_{:,i}^\mathsf{T}\overline{c} = 0\}$, $i \in \mathcal{E}^{0+}(x)_0 \Rightarrow X_{:,i} \in L_{\overline{F}}$. Cela signifie que les indices de $\mathcal{E}^{0+}(x)_0$ ne sont pas pertinents dans le sens où la valeur des η_i correspondants pourrait être différente sans affecter l'optimalité de λ^*.

Enfin, pour tout $\hat{\zeta}$ correspondant au sommet $\overline{y} + Y\hat{\zeta}$ de $\overline{F}$, soit $\hat{d} \in \mathcal{R}(Y)$ une direction telle que $\hat{\zeta}_i y_i^\mathsf{T} \hat{d} > 0$, c'est-à-dire une direction vérifiant le sommet. Soit $\hat{c} := \overline{c} + \varepsilon\hat{d}$ pour un certain petit $\varepsilon > 0$. Puisque $\hat{d}$ est pris dans un ensemble ouvert, quitte à de petites modifications (puis des ajustements de ε), on peut supposer que $X_{:,i}^\mathsf{T}\hat{c} \neq 0$ pour $i \in \mathcal{E}^{0+}(x)_0$. Soit $\hat{\eta}_i = \operatorname{sgn}(X_{:,i}^\mathsf{T}\hat{c})$ pour $i \in \mathcal{E}^{0+}(x)_0$, on a, en utilisant les propriétés de $\hat{d}$ et $\overline{c}$

$$\begin{cases} i \in I^* & \implies \hat{c}^\mathsf{T}\hat{\zeta}_i y_i = \overline{c}^\mathsf{T}\hat{\zeta}_i y_i + \varepsilon\hat{d}^\mathsf{T}\hat{\zeta}_i y_i > 0, \\ & \qquad \overset{>0}{} \qquad \overset{=\varepsilon\ldots}{} \\ i \in I^{\overline{F}} & \implies \hat{c}^\mathsf{T}\hat{\zeta}_i y_i = \overline{c}^\mathsf{T}\hat{\zeta}_i y_i + \varepsilon\hat{d}^\mathsf{T}\hat{\zeta}_i y_i > 0, \\ & \qquad \underset{=0}{} \qquad \underset{>0}{} \end{cases}$$

et

$$\begin{cases} i \in \mathcal{E}^{0+}(x)_0 & \implies \hat{c}^\mathsf{T}\hat{\eta}_i X_{:,i} = \overline{c}^\mathsf{T}\hat{\eta}_i X_{:,i} + \varepsilon\hat{d}^\mathsf{T}\hat{\eta}_i X_{:,i} > 0, \\ & \qquad \overset{=0}{} \qquad \overset{>0}{} \\ i \in \mathcal{E}^{0+}(x)_+ & \implies \hat{c}^\mathsf{T}\hat{\eta}_i X_{:,i} = \overline{c}^\mathsf{T}\hat{\eta}_i X_{:,i} + \varepsilon\hat{d}^\mathsf{T}\hat{\eta}_i X_{:,i} > 0, \\ & \qquad \underset{>0}{} \qquad \underset{=\varepsilon\ldots}{} \end{cases}$$

ce qui signifie que nous avons justifié que chacune des matrices jacobiennes correspondant aux vecteurs de signes $[\hat{\eta}; \hat{\zeta}]$ sont dans $\partial_B H(x)$. Écrivons maintenant la combinaison convexe

$$\overline{\zeta} = \sum t_j \hat{\zeta}^j, \quad t_j \geqslant 0, \quad \sum t_j = 1$$

et notons $\tilde{\eta} := \sum t_i \hat{\eta}$. Par construction, les vecteurs de signes

$$(\hat{\eta}; \hat{\zeta}) = (\eta_{\mathcal{E}^{0+}(x)_+}; \hat{\eta}_{\mathcal{E}^{0+}(x)_0}; \overline{\zeta}_{I^*}; \hat{\zeta}_{I^{\overline{F}}})$$

correspondent à des jacobiennes de $\partial_B H(x)$, donc en utilisant $(\tilde{\eta}; \overline{\zeta}) = \sum t_j(\hat{\eta}^j; \hat{\zeta}^j)$ on a que $\overline{g} := \overline{x} - \overline{y} + X\tilde{\eta} - Y\overline{\zeta} \in \partial\theta(x)$. □

Remarquons que ni les valeurs "initiales" $(\overline{\eta}, \overline{\zeta})$ ni celles modifiées $(\tilde{\eta}, \overline{\zeta})$ ne vérifient nécessairement la proposition 6.1.10, au sens où il faut calculer explicitement θ' pour s'assurer que g est une direction de descente. La proposition suivante détaille comment une valeur spécifique de η, après projection pour obtenir ζ, peut retourner un élément de $\partial\theta(x)$ (les dégénérescences interviennent dans la preuve mais aucune modification de η n'est requise).

La difficulté est que puisque ζ est obtenu par la projection d'un point dépendant de η, les deux quantités doivent être modifiées simultanément.

Proposition D.2.8 (un élément du différentiel)**.** *Avec les mêmes notations, considérons les problèmes suivants*

$$\max_{\eta \in [-1,+1]^{\mathcal{E}^{0+}(x)}} \min_{\zeta \in [-1,+1]^{\mathcal{E}^-(x)}} ||g(\eta,\zeta)||^2/2 = \max_{\eta \in [-1,+1]^{\mathcal{E}^{0+}(x)}} \operatorname{dist}(\overline{x}+X\eta, \overline{y}+Y[-1,+1]^{\mathcal{E}^-(x)})^2/2$$

et soit (η^{**}, ζ^{**}) *une solution. Alors* $g(\eta^{**}, \zeta^{**}) \in \partial\theta(x)$.

Remarquons que la minimisation interne, qui correspond à une distance/projection, est donnée par la proposition 6.1.10.

Preuve. Premièrement, puisque la fonction considérée est continue et les ensembles sont convexes compacts, les problèmes admettent des solutions. La solution η^{**} du problème externe est obtenue en un point extrémal, i.e., $\eta^{**} \in \{-1,+1\}^{\mathcal{E}^{0+}(x)}$: en effet, comme la fonction à maximiser est convexe, la solution est un sommet du polytope. Si $Z_x \subseteq Z_y$, i.e., le point x est (fortement) θ-stationnaire, on obtient $g(\eta^{**},\zeta^{**}) = 0$ ce qui est cohérent car $0 \in \partial\theta(x)$. Sinon, $g \neq 0$ et le raisonnement est similaire à celui de la preuve de la proposition D.2.7.

La projection de $\overline{x} + X\eta^{**}$ sur Z_y appartient à (la frontière de) Z_y. D'après le lemme B.1.4, cette projection appartient à l'intérieur relatif d'une face F_y^*. Par la proposition B.2.2, cette face F_y^* est un zonotope engendré par les indices de $\mathcal{E}^-(x)_{I^F}$ et centré par ceux de $\mathcal{E}^-(x)_{I^*}$ (notés I^F et I^* pour simplifier, voir figure D.6) :

$$F_y^* = \overline{y} + Y_{:,\mathcal{E}^-(x)_{I^*}}\zeta^{**}_{\mathcal{E}^-(x)_{I^*}} + Y_{:,\mathcal{E}^-(x)_{I^F}}[-1,+1]^{I^F} = \overline{y} + \tilde{y} + Y_{:,\mathcal{E}^-(x)_{I^F}}[-1,+1]^{I^F}.$$

Soit $L_y = \mathrm{aff}(F_y^*) - \mathrm{aff}(F_y^*)$ le sous-espace linéaire engendré par F_y^*. D'après la proposition B.2.6, la projection de $\overline{x}+X\eta^{**}$ est $\overline{y}+Y\zeta^{**}$, avec $\zeta^{**}_{I^*} \in \{\pm 1\}^{I^*}$ et $\zeta^{**}_{I^F} \in (-1,+1)^{I^F}$. Soit $c^* = g := \overline{x} + X\eta^{**} - \overline{y} - Y\zeta^{**}$. Comme $\overline{y} + Y\zeta^{**}$ est la projection de $\overline{x} + X\eta^{**}$ sur Z_y, c^* vérifie la proposition B.2.8, i.e., $\zeta_i^{**} y_i^\mathsf{T} c^* \geqslant 0$. On a aussi $Z_y \subseteq \{z \in \mathbb{R}^n : z^\mathsf{T}c^* \leqslant (\overline{y}+Y\zeta^{**})^\mathsf{T}c^*\}$ puisque c^* est une normale sortante. Montrons que la même propriété vaut pour les indices de $\mathcal{E}^{0+}(x)$.

La pertinence des variables $\cdot^{**}$ vient du fait que $Z_x \subseteq \{z \in \mathbb{R}^n : z^\mathsf{T}c^* \leqslant (\overline{x}+X\eta^{**})^\mathsf{T}c^*\}$. En effet, s'il existait un point de Z_x dans l'autre demi-espace, sa distance à Z_y serait plus grande en utilisant l'inclusion $Z_y \subseteq \{z \in \mathbb{R}^n : z^\mathsf{T}c^* \leqslant (\overline{y} + Y\zeta^{**})^\mathsf{T}c^*\}$.

Supposons qu'il existe $i \in \mathcal{E}^{0+}(x)$ tel que $\eta_i^{**} X_{:,i}^\mathsf{T} c^* < 0$. Alors, le point $\overline{x} + X\eta^{**} - 2\eta_i^{**}X_{:,i} \in Z_x$ vérifierait

$$(\overline{x} + X\eta^{**} - 2\eta_i^{**}X_{:,i})^\mathsf{T}c^* = (\overline{x} + X\eta^{**})^\mathsf{T}c^* - 2\eta_i^{**}X_{:,i}^\mathsf{T}c^* > (\overline{x} + X\eta^{**})^\mathsf{T}c^*,$$

ce qui est une contradiction. Donc, $\eta_i^{**}X_{:,i}^\mathsf{T}c^* \geqslant 0$.

La question des dégénérescences, i.e., les indices $i \in \mathcal{E}^{0+}(x)$ tels que $X_{:,i}^\mathsf{T} g = 0$, ne peut être traitée aussi facilement que dans la proposition D.2.7. En effet, si on modifie seulement le paramètre η, alors le couple η, ζ ne correspond plus à $g = c^*$. Il faut donc justifier que les deux peuvent être modifiés simultanément pour garantir que g ne change pas, car il doit rester identique pour bénéficier de la propriété de projection, i.e., la proposition 6.1.10.

Soit $\mathcal{E}^{0+}(x) := \mathcal{E}^{0+}(x)_+ \cup \mathcal{E}^{0+}(x)_0 = \{i \in \mathcal{E}^{0+}(x) : \eta_i^{**}X_{:,i}^\mathsf{T} g > 0\} \cup \{i \in \mathcal{E}^{0+}(x) : X_{:,i}^\mathsf{T} g = 0\}$. Les indices dégénérés, $\mathcal{E}^{0+}(x)_0$, correspondent à une face F_x^* de Z_x et à des valeurs de $\eta_{\mathcal{E}^{0+}(x)_0}$ qui peuvent être changées sans modifier la valeur optimale (il faut aussi modifier ζ). Comme dans la proposition D.2.7, nous avons $\mathrm{aff}(F_x^*) - \mathrm{aff}(F_x^*) \subseteq \mathrm{aff}(F_y^*) - \mathrm{aff}(F_y^*)$, i.e., l'espace engendré par F_x^* est contenu dans celui de F_y^*, et $F_x^* - g \subseteq F_y^*$.[6] Soit

6. Sinon, en utilisant les directions de X non engendrées par Y, la distance augmente ce qui contredit l'optimalité.

$\tilde{x} := X_{:,\mathcal{E}^{0+}(x)_+}\eta^{**}_{\mathcal{E}^{0+}(x)_+}$, $\tilde{y} := Y_{:,I^*}\zeta^{**}_{I^*}$, on a :

$$F^*_x := \overline{x} + \tilde{x} + X_{:,\mathcal{E}^{0+}(x)_0}[-1,+1]^{\mathcal{E}^{0+}(x)_0},$$
$$g := \overline{x} + \tilde{x} + X_{:,\mathcal{E}^{0+}(x)_0}\eta^{**}_{\mathcal{E}^{0+}(x)_0} - \overline{y} - \tilde{y} - Y_{:,I^F}\zeta^{**}_{I^F} = \tilde{w} + X_{:,\mathcal{E}^{0+}(x)_0}\eta^{**}_{\mathcal{E}^{0+}(x)_0} - Y_{:,I^F}\zeta^{**}_{I^F},$$

où $\tilde{w} = \overline{x} - \overline{y} + \tilde{x} - \tilde{y}$, et soit $\tilde{g} := X_{:,\mathcal{E}^{0+}(x)_0}\eta^{**}_{\mathcal{E}^{0+}(x)_0} - Y_{:,I^F}\zeta^{**}_{I^F}$ de sorte que $g = \tilde{w} + \tilde{g}$. Tout paramètre $\eta = (\eta^{**}_{\mathcal{E}^{0+}(x)_+}, \eta_{\mathcal{E}^{0+}(x)_0})$ pour un $\eta_{\mathcal{E}^{0+}(x)_0}$ arbitraire donne, après projection, la même valeur g, bien que le ζ_{I^F} correspondant dépende de $\eta_{\mathcal{E}^{0+}(x)_0}$. On veut obtenir un certain uplet

$$(\tilde{\eta}^{**}, \tilde{\zeta}^{**}) = ((\eta^{**}_{\mathcal{E}^{0+}(x)_+}, \tilde{\eta}^{**}_{\mathcal{E}^{0+}(x)_0}), (\zeta^{**}_{I^*}, \tilde{\zeta}^{**}_{I^F}))$$

(où $g(\eta^{**}, \zeta^{**}) = g(\tilde{\eta}^{**}, \tilde{\zeta}^{**})$) avec $\eta^{**}_{\mathcal{E}^{0+}(x)_+}$ et $\zeta^{**}_{I^*}$ inchangés, $\tilde{\eta}^{**}_{\mathcal{E}^{0+}(x)_0}$ et $\tilde{\zeta}^{**}_{I^F}$ correspondant à une combinaison convexe d'éléments de $\partial_B H(x)$.

Comme dans la proposition D.2.7, soit les sommets de F^*_y notés $\overline{y} + Y\hat{\zeta}^j$ pour $j \in J$ avec J un ensemble d'indices. En particulier, $\hat{\zeta}^j_{I^*} = \zeta^{**}_{I^*}$ pour tout $j \in J$. Pour un $j \in J$, puisque $\overline{y} + Y\hat{\zeta}^j$ est un sommet de Z_y, soit $\hat{c}^j := c^F + \varepsilon d^j_F$ où c^F est un vecteur normal à F^*_y donné par la proposition B.1.7 et $d^j_F \in L_y$ est une direction vérifiant le sommet $\hat{\zeta}^j_{I^F}$ dans le sous-ensemble I^F [7]. Le vecteur $\hat{c}^j$ vérifie $\hat{\zeta}^j$:

$$\begin{cases} i \in I^* \implies \zeta^{**}_i y_i^\mathsf{T}\hat{c}^j = \overbrace{\zeta^{**}_i y_i^\mathsf{T} c^F}^{>0} + \overbrace{\varepsilon\zeta^{**}_i y_i^\mathsf{T} d^j_F}^{=\varepsilon\dots} > 0, \\ i \in I^F \implies \hat{\zeta}^j_i y_i^\mathsf{T}\hat{c}^j = \underbrace{\hat{\zeta}_i y_i^\mathsf{T} c^F}_{=0} + \underbrace{\varepsilon\hat{\zeta}_i y_i^\mathsf{T} d^j_F}_{>0} > 0, \end{cases}$$

où l'inégalité en haut à gauche vient des propriétés de la normale, l'égalité en bas à gauche du fait que c^F est normal à la face et l'inégalité en bas à droite du choix de d^j_F comme direction vérifiante. Ainsi, pour tout $i \in \mathcal{E}^-(x)$, $\hat{\zeta}^j_i y_i^\mathsf{T}\hat{c}^j > 0$. Dans le sous-espace L_y, quitte à de petites modifications de d^j_F, on peut supposer que $X^\mathsf{T}_{:,i}\hat{c}^j \neq 0$ pour $i \in \mathcal{E}^{0+}(x)_0$ (en gardant $d^j_F \in L_y$ puisque ces $X_{:,i}$ appartiennent aussi à L_y). Alors, pour $i \in \mathcal{E}^{0+}(x)_0$, soit $\hat{\eta}^j_i := \operatorname{sgn}(X^\mathsf{T}_{:,i}\hat{c}^j)$, on a, en utilisant les propriétés de $\hat{d}^j$ et c^F,

$$\begin{cases} i \in \mathcal{E}^{0+}(x)_0 \implies \hat{\eta}_i X^\mathsf{T}_{:,i}\hat{c}^j = \overbrace{\hat{\eta}_i X^\mathsf{T}_{:,i} c^F}^{=0} + \overbrace{\varepsilon\hat{\eta}_i X^\mathsf{T}_{:,i}\hat{d}^j}^{>0} > 0, \\ i \in \mathcal{E}^{0+}(x)_+ \implies \hat{\eta}_i X^\mathsf{T}_{:,i}\hat{c}^j = \underbrace{\hat{\eta}_i X^\mathsf{T}_{:,i} c^F}_{>0} + \underbrace{\varepsilon\hat{\eta}_i X^\mathsf{T}_{:,i}\hat{d}^j}_{=\varepsilon\dots} > 0, \end{cases}$$

ce qui signifie que le vecteur de signes $(\eta^{**}_{\mathcal{E}^{0+}(x)_+}, \hat{\eta}^j_{\mathcal{E}^{0+}(x)_0}, \zeta^{**}_{I^*}, \hat{\zeta}^j_{I^F})$ correspond à un $s \in \mathcal{S}([X\ Y])$ ou de façon équivalente à une matrice de $\partial_B H(x)$ par (D.6) [8]. Pour le sommet symétrique (dans F^*) $\overline{y} + Y_{:,I^*}\hat{\zeta}_{I^*} - Y_{:,I^F}\hat{\zeta}_{I^F}$, on peut prendre la direction $c^F - \varepsilon d^j_F$, donc pour ce sommet symétrique on obtient $-\hat{\eta}^j_{\mathcal{E}^{0+}(x)_0}$.

7. C'est-à-dire $(\hat{\zeta}^j_{I^F})_i y_i^\mathsf{T} d > 0$ pour tout $i \in I^F$.

8. Si $\mathcal{E}^0(x) \neq \varnothing$, les $X_{:,i}$ correspondants sont nuls mais n'interviennent pas dans les problèmes d'optimisation – ces indices peuvent être retirés en utilisant le même argument que sous (D.7).

Cette construction associe (tous) les sommets de F_y^* à certains de F_x^*. Rappelons que $F_x^* - g \subseteq F_y^*$, ce qui s'écrit

$$\begin{aligned} \overline{x} + \tilde{x} + X_{:,\mathcal{E}^{0+}(x)_0}[-1,+1]^{\mathcal{E}^{0+}(x)_0} - \tilde{w} - \tilde{g} &\subseteq \overline{y} + \tilde{y} + Y_{I^F}[-1,+1]^{I^F} \\ \iff \quad -\tilde{g} + X_{:,\mathcal{E}^{0+}(x)_0}[-1,+1]^{\mathcal{E}^{0+}(x)_0} &\subseteq +Y_{:,I^F}[-1,+1]^{I^F} \end{aligned} \tag{D.12}$$

et représente une inclusion de zonotopes en dimension réduite avec moins d'indices. Montrons maintenant une propriété utile des directions $z^j := (X_{:,\mathcal{E}^{0+}(x)_0}\hat{\eta}^j_{\mathcal{E}^{0+}(x)_0} - Y_{:,I^F}\hat{\zeta}^j_{I^F})$ pour $j \in J$: on a $\tilde{g} = X_{:,\mathcal{E}^{0+}(x)_0}\eta^{**}_{\mathcal{E}^{0+}(x)_0} - Y_{:,I^F}\zeta^{**}_{I^F} \in \mathrm{conv}\{z^j : j \in J\}$. Ceci peut être utilisé comme suit, avec $t_j \geqslant 0$, $j \in J$ et $\sum_J t_j = 1$:

$$\begin{aligned} \tilde{g} &= \sum_{j\in J} t_j z^j = \sum_{j\in J} t_j (X_{:,\mathcal{E}^{0+}(x)_0}\hat{\eta}^j_{\mathcal{E}^{0+}(x)_0} - Y_{:,I^F}\hat{\zeta}^j_{I^F}) \\ \tilde{w} + \tilde{g} &= \sum_{j\in J} t_j (\tilde{w} + X_{:,\mathcal{E}^{0+}(x)_0}\hat{\eta}^j_{\mathcal{E}^{0+}(x)_0} - Y_{:,I^F}\hat{\zeta}^j_{I^F}) \\ g &= \sum_{j\in J} t_j (g_1 + X[\eta^{**}_{\mathcal{E}^{0+}(x)_+}; \hat{\eta}^j_{\mathcal{E}^{0+}(x)_0}] - Y[\zeta^{**}_{I^*}; \hat{\zeta}^j_{I^F}]) \\ &= \sum_{j\in J} t_j (g_1 + [X \ -Y][\eta^{**}_{\mathcal{E}^{0+}(x)_+}; \hat{\eta}^j_{\mathcal{E}^{0+}(x)_0}; \zeta^{**}_{I^*}; \hat{\zeta}^j_{I^F}]), \end{aligned}$$

ce qui indique que g est une combinaison convexe d'éléments de la forme $g_1 + [X \ -Y]s$ avec $s \in \mathcal{S}$, comme décrit par (D.7), donc $g \in \partial\theta(x)$. Enfin, justifions que

$$\tilde{g} = X_{:,\mathcal{E}^{0+}(x)_0}\eta^{**}_{\mathcal{E}^{0+}(x)_0} - Y_{:,I^F}\zeta^{**}_{I^F} \in \mathrm{conv}\{z^j : j \in J\}.$$

Supposons que l'inclusion est fausse. Alors, puisque $\tilde{g}$ et l'enveloppe convexe sont des ensembles convexes compacts, il existe un vecteur séparateur strict $\overline{d}$ tel que

$$\forall\, z \in \mathrm{conv}\{z^j : j \in J\}, \overline{d}^\mathsf{T}\tilde{g} > \overline{d}^\mathsf{T} z.$$

Par convexité, prenons $z = z^j$ pour tout $j \in J$, la condition précédente équivaut à

$$\begin{aligned} &\forall\, j \in J, \ \ \overline{d}^\mathsf{T}\tilde{g} \ > \ \overline{d}^\mathsf{T} z^j \\ \iff \quad &\forall\, j \in J, \ \ \overline{d}^\mathsf{T}\tilde{g} \ > \ \overline{d}^\mathsf{T}(X_{:,\mathcal{E}^{0+}(x)_0}\hat{\eta}^j_{\mathcal{E}^{0+}(x)_0} - Y_{:,I^F}\hat{\zeta}^j_{I^F}). \end{aligned}$$

En particulier, puisque $\{z^j, j \in J\}$ est symétrique par construction :

$$\begin{aligned} &\forall\, j \in J, && \overline{d}^\mathsf{T}\tilde{g} \ > \ \overline{d}^\mathsf{T}(Y_{:,I^F}\hat{\zeta}^j_{I^F} - X_{:,\mathcal{E}^{0+}(x)_0}\hat{\eta}^j_{\mathcal{E}^{0+}(x)_0}) \\ \iff \quad &\forall\, j \in J, && \overline{d}^\mathsf{T}(\tilde{g} + X_{:,\mathcal{E}^{0+}(x)_0}\hat{\eta}^j_{\mathcal{E}^{0+}(x)_0}) \ > \ \overline{d}^\mathsf{T}(Y_{:,I^F}\hat{\zeta}^j_{I^F}). \end{aligned}$$

Enfin, puisque les ensembles convexes sont compacts, on peut faire une petite modification de $\overline{d}$ (restant dans L_y) pour garantir que $\overline{d}^\mathsf{T} z^j \neq 0$ pour tout $j \in J$. Alors, définissons le vecteur de signes suivant :

$$\overline{s} := \mathrm{sgn}(Y_{:,I^F}^\mathsf{T}\overline{d}) \in \{\pm 1\}^{I^F},$$

qui est le vecteur de signes (partiel) correspondant au sommet vérifié par $\overline{d}$. Considérons l'inégalité stricte pour l'indice $\overline{j}$ correspondant à $\overline{s}$ (qui existe car J couvre les sommets de F_y^* et $\overline{s}$ est parmi ces sommets). Or, $\overline{z} := Y_{:,I^F}\overline{s}$ maximise $\overline{d}^\mathsf{T} z$ pour $z \in Y_{:,I^F}[-1,+1]^{I^F}$, cela signifie que $\tilde{g} + X_{:,\mathcal{E}^{0+}(x)_0}\hat{\eta}^{\overline{j}}_{\mathcal{E}^{0+}(x)_0}$ n'appartient pas à $Y_{:,I^F}[-1,+1]^{I^F}$. Par symétrie, $-\tilde{g} - X_{:,\mathcal{E}^{0+}(x)_0}\hat{\eta}^{\overline{j}}_{\mathcal{E}^{0+}(x)_0}$ n'appartient pas non plus à $Y_{:,I^F}[-1,+1]^{I^F}$. Ceci contredit l'inclusion des zonotopes dans (D.12) avec la variable $-\eta^{\overline{j}}_{\mathcal{E}^{0+}(x)_0}$, à une translation près par $\pm g$. □

Remarque D.2.9 (minima locaux et globaux)**.** Un maximum local (strict) η de la fonction de distance retourne également un élément $g \in \partial\theta(x)$.

En effet, g est une normale non stricte comme décrit dans la proposition B.2.8. Par maximalité locale de la distance, on peut montrer (comme pour l'optimalité globale) que $\eta_i X^\mathsf{T}_{:,i} g \geqslant 0$. À partir de là, le traitement des indices dégénérés peut être effectué comme dans la preuve principale. □

Exemple D.2.10 (Valeurs correctes pour le différentiel)**.** Considérons l'exemple D.2.3. Les valeurs de $\overline{x} + X\eta \in \mathbb{R}^2$, ou de manière équivalente celles de $\eta \in [-1,+1]^2$ telles que, après projection, $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^{0+}(x)})) \in \partial\theta(x)$ sont indiquées dans la figure D.7.

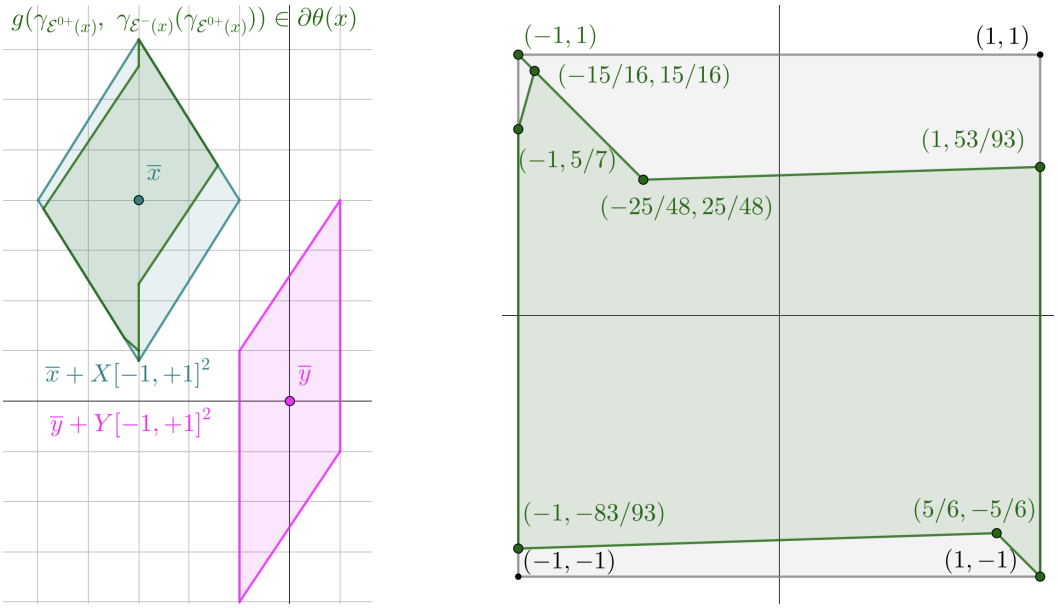


FIGURE D.7 – À gauche : Z_y en magenta, Z_x en bleu-vert, et le sous-ensemble de Z_x en vert correspondant aux $\gamma_{\mathcal{E}^{0+}(x)}$ (η) tels que les $\gamma_{\mathcal{E}^-(x)}$ (ζ) obtenus après projection donnent $g \in \partial\theta(x)$. À droite : valeurs correspondantes dans $[-1,+1]^2$ (c'est-à-dire avec η). Noter que le point le plus haut de la figure à gauche correspond à $\eta = (1,-1)$ et le plus à gauche à $\eta = (-1,-1)$, ce qui explique le changement d'orientation (pour retrouver une forme similaire au graphique de gauche, effectuer une rotation de $+3\pi/4$ dans le sens contre-horaire puis une symétrie axiale verticale).

Sur cet exemple, on constate qu'obtenir un élément du C-différentiel peut ne pas être trivial, car les solutions forment un ensemble complexe.

Cependant, rappelons que $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^{0+}(x)}))$ est obtenu par projection pour garantir que $-g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^{0+}(x)}))$ est une direction de descente. Il peut exister de nombreuses autres combinaisons de $\gamma_{\mathcal{E}^{0+}(x)}$ (η) et $\gamma_{\mathcal{E}^-(x)}$ (ζ) telles que $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)})$ soit une direction de descente et/ou un élément de $\partial\theta(x)$.

Dans la figure suivante, nous fixons arbitrairement $\zeta = 0$: clairement, ζ n'est pas obtenu par projection, donc a priori $-g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)})$ pourrait ne pas être une direction de descente. En réalité, on observe que quel que soit η, $-g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)})$ est une direction de descente, mais les valeurs de η telles que $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}) \in \partial\theta(x)$ changent.

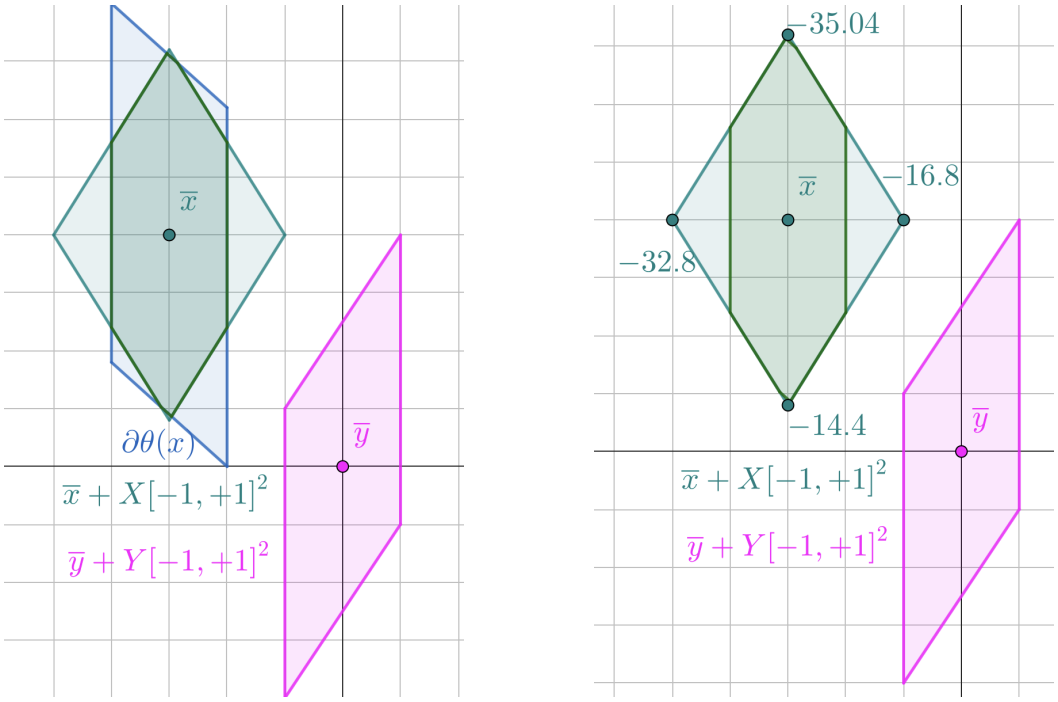


Figure D.8 – En magenta, Z_y et en bleu-vert, Z_x. À gauche : $\partial\theta(x)$ en bleu, l'intersection avec Z_x en vert foncé correspond aux η avec $g(\eta, \zeta = 0) \in \partial\theta(x)$. À droite : valeurs de $\theta'(x, -g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}))$ pour les η extrémaux : tout g avec $\zeta = 0$ est une direction de descente.

Maintenant, fixons arbitrairement $\zeta = e$: ζ n'est pas obtenu par projection, donc a priori $-g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)})$ pourrait ne pas être une direction de descente. En fait, pour certains choix de η, $-g$ n'est pas une direction de descente ; les valeurs de η telles que $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}) \in \partial\theta(x)$ changent également.

Sur les figures, on observe que le plus souvent, $\theta'(x, -g) \leqslant 0$. Cependant, dans une autre situation où les zonotopes seraient beaucoup plus proches (Z_x partiellement contenu dans Z_y), il serait plus probable que davantage de combinaisons de poids conduisent à $-g$ étant une direction de montée. □

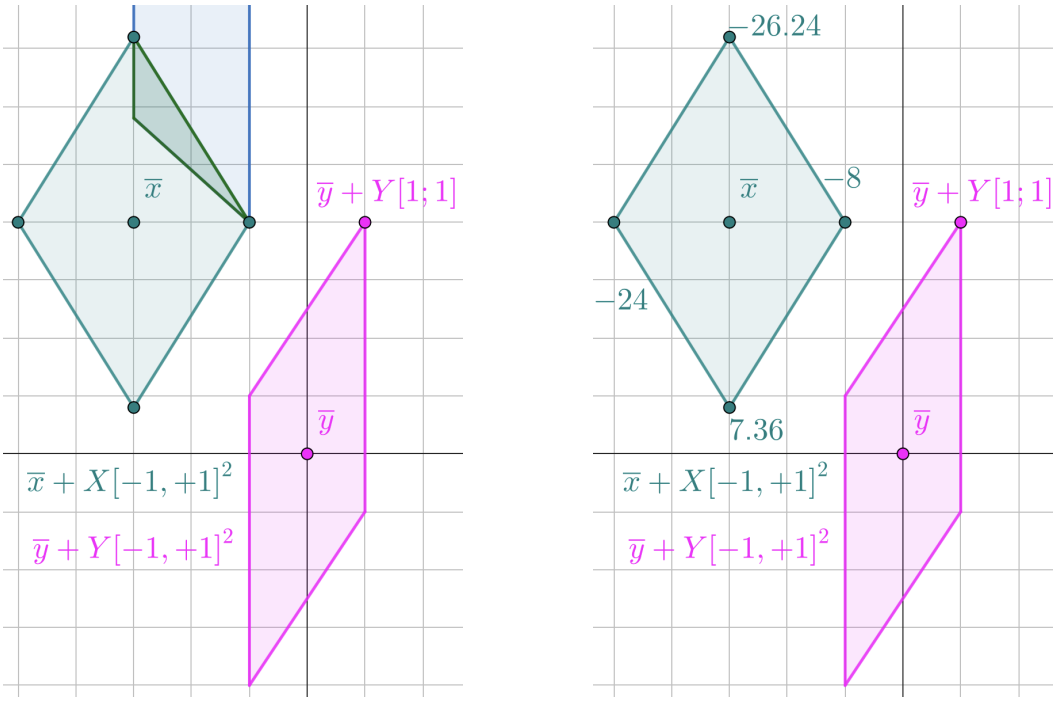


FIGURE D.9 – En magenta, Z_y et en bleu-vert, Z_x. À gauche : $\partial\theta(x)$ en bleu, l'intersection avec Z_x en vert foncé correspond aux η avec $g(\eta, \zeta = 0) \in \partial\theta(x)$. À droite : valeurs de $\theta'(x, -g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}))$ pour les η extrémaux : pour certains η spécifiques, $-g$ est une direction de montée.

Bibliographie

[1] L. Abdallah, M. Haddou et T. Migot. "Solving Absolute Value Equation Using Complementarity and Smoothing Functions". In : *Journal of Computational and Applied Mathematics* 327 (jan. 2018), p. 196-207. ISSN : 03770427. DOI : 10.1016/j.cam.2017.06.019 (cf. p. 44).

[2] Vincent Acary et Bernard Brogliato. *Numerical Methods for Nonsmooth Dynamical Systems : Applications in Mechanics and Electronics*. Lecture Notes in Applied and Computational Mechanics Ser v.v. 35. Berlin, Heidelberg : Springer Berlin / Heidelberg, 2008. ISBN : 978-3-540-75391-9 978-3-540-75392-6 (cf. p. 2, 3).

[3] Muhamed Aganagić. "Newton's Method for Linear Complementarity Problems". In : *Mathematical Programming* 28.3 (oct. 1984), p. 349-362. ISSN : 0025-5610, 1436-4646. DOI : 10.1007/BF02612339 (cf. p. 29, 59, 227).

[4] Marcelo Aguiar et Swapneel Mahajan. *Topics in Hyperplane Arrangements*. T. 226. Mathematical Surveys and Monographs. Providence, Rhode Island : American Mathematical Society, nov. 2017. ISBN : 978-1-4704-3711-4 978-1-4704-4254-5. DOI : 10.1090/surv/226 (cf. p. 49, 75, 77-79, 145, 150).

[5] Jan Harold Alcantara et Jein-Shan Chen. "A New Class of Neural Networks for NCPs Using Smooth Perturbations of the Natural Residual Function". In : *Journal of Computational and Applied Mathematics* 407 (juin 2022), p. 114092. ISSN : 03770427. DOI : 10.1016/j.cam.2022.114092 (cf. p. 29, 37).

[6] Jan Harold Alcantara, Chen-Han Lee, Chieu Thanh Nguyen, Yu-Lin Chang et Jein-Shan Chen. "On Construction of New NCP Functions". In : *Operations Research Letters* 48.2 (mars 2020), p. 115-121. ISSN : 01676377. DOI : 10.1016/j.orl.2020.01.002 (cf. p. 4, 19).

[7] Gerald L. Alexanderson et John E. Wetzel. "Arrangements of Planes in Space". In : *Discrete Mathematics* (1981), p. 219-240 (cf. p. 49, 79).

[8] Xavier Allamigeon, Stéphane Gaubert et Frédéric Meunier. "Tropical Complementarity Problems and Nash Equilibria". In : *SIAM Journal on Discrete Mathematics* 37.3 (sept. 2023), p. 1645-1665. ISSN : 0895-4801, 1095-7146. DOI : 10.1137/21M1446861 (cf. p. 2).

[9] Mihai Anitescu et Florian Alexandru Potra. "Formulating Dynamic Multi-Rigid-Body Contact Problems with Friction as Solvable Linear Complementarity Problems". In : *Nonlinear Dynamics* 14 (1997), p. 231-247 (cf. p. 2).

[10] Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro et Blake Woodworth. “Lower Bounds for Non-Convex Stochastic Optimization”. In : *Mathematical Programming* 199.1-2 (mai 2023), p. 165-214. issn : 0025-5610, 1436-4646. doi : 10.1007/s10107-022-01822-7 (cf. p. 47).

[11] Larry Armijo. “Minimization of Functions Having Lipschitz Continuous First Partial Derivatives”. In : *Pacific Journal of Mathematics* 16.1 (jan. 1966), p. 1-3. issn : 0030-8730, 0030-8730. doi : 10.2140/pjm.1966.16.1 (cf. p. 26).

[12] Christos A. Athanasiadis. “Characteristic Polynomials of Subspace Arrangements and Finite Fields”. In : *Advances in Mathematics* 122.2 (sept. 1996), p. 193-233. issn : 00018708. doi : 10.1006/aima.1996.0059 (cf. p. 6, 50, 55, 131, 136, 145, 148).

[13] David Avis et Komei Fukuda. “A Pivoting Algorithm for Convex Hulls and Vertex Enumeration of Arrangements and Polyhedra”. In : *Discrete Computational Geometry* 8 (1992), p. 295-31. issn : 0179-5376,1432-0444. doi : 10.1007/BF02293050 (cf. p. 53, 145, 154, 220).

[14] David Avis et Komei Fukuda. “Reverse Search for Enumeration”. In : *Discrete Applied Mathematics* 65 (mars 1996), p. 21-46. doi : 10.1016/0166-218X(95)00026-N (cf. p. 6, 53, 63, 75, 77, 87).

[15] Pierre Baldi. “Deep Learning in Biomedical Data Science”. In : *Annual Review of Biomedical Data Science* 1.1 (juill. 2018), p. 181-205. issn : 2574-3414, 2574-3414. doi : 10.1146/annurev-biodatasci-080917-013343 (cf. p. 72).

[16] Pierre Baldi et Roman Vershynin. “Polynomial Threshold Functions, Hyperplane Arrangements, and Random Tensors”. In : *SIAM Journal on Mathematics of Data Science* 1.4 (jan. 2019), p. 699-729. issn : 2577-0187. doi : 10.1137/19M1257792 (cf. p. 55, 63, 72).

[17] Antoine Bambade, Fabian Schramm, Sarah El-Kazdadi, Stéphane Caron, Adrien Taylor et Justin Carpentier. “PROXQP : An Efficient and Versatile Quadratic Programming Solver for Real-Time Robotics Applications and Beyond”. In : (2023), p. 17 (cf. p. 230).

[18] Laurence Beaude, Konstantin Brenner, Simon Lopez, Roland Masson et Farid Smai. “Non-Isothermal Compositional Liquid Gas Darcy Flow : Formulation, Soil-Atmosphere Boundary Condition and Application to High-Energy Geothermal Simulations”. In : *Computational Geosciences* 23.3 (juin 2019), p. 443-470. issn : 1420-0597, 1573-1499. doi : 10.1007/s10596-018-9794-9 (cf. p. 2).

[19] Amir Beck et Nadav Hallak. “On the Convergence to Stationary Points of Deterministic and Randomized Feasible Descent Directions Methods”. In : *SIAM Journal on Optimization* 30.1 (jan. 2020), p. 56-79. issn : 1052-6234, 1095-7189. doi : 10.1137/18M1217760 (cf. p. 46, 221, 233).

[20] Ibtihel Ben Gharbia. “Résolution de Problèmes de Complémentarité. : Application à Un Écoulement Diphasique Dans Un Milieu Poreux”. Thèse de doct. Université Paris Dauphine Paris IX, 2012 (cf. p. 2, 13, 29, 30, 34, 210).

[21] Ibtihel Ben Gharbia, Joëlle Ferzly, Martin Vohralík et Soleiman Yousef. "Semismooth and Smoothing Newton Methods for Nonlinear Systems with Complementarity Constraints : Adaptivity and Inexact Resolution". In : *Journal of Computational and Applied Mathematics* 420 (mars 2023), p. 114765. ISSN : 03770427. DOI : 10.1016/j.cam.2022.114765 (cf. p. 2, 16).

[22] Ibtihel Ben Gharbia et J. Charles Gilbert. "Nonconvergence of the Plain Newton-min Algorithm for Linear Complementarity Problems with a P-matrix". In : *Mathematical Programming* 134.2 (sept. 2012), p. 349-364. ISSN : 0025-5610, 1436-4646. DOI : 10.1007/s10107-010-0439-6 (cf. p. 59).

[23] Ibtihel Ben Gharbia et Jean Charles Gilbert. "An Algorithmic Characterization of P-Matricity". In : *SIAM Journal on Matrix Analysis and Applications* 34.3 (jan. 2013), p. 904-916. ISSN : 0895-4798, 1095-7162. DOI : 10.1137/120883025 (cf. p. 13, 30, 59).

[24] Ibtihel Ben Gharbia et Jean Charles Gilbert. "An Algorithmic Characterization of P-matricity II : Adjustments, Refinements, and Validation". In : *SIAM Journal on Matrix Analysis and Applications* 40.2 (jan. 2019), p. 800-813. ISSN : 0895-4798, 1095-7162. DOI : 10.1137/18M1168522 (cf. p. 13, 30, 59).

[25] Ibtihel Ben Gharbia et Jérôme Jaffré. "Gas Phase Appearance and Disappearance as a Problem with Complementarity Constraints". In : *Mathematics and Computers in Simulation* 99 (mai 2014), p. 28-36. ISSN : 03784754. DOI : 10.1016/j.matcom.2013.04.021 (cf. p. 59).

[26] Dimitri P. Bertsekas. *Nonlinear Programming*. third. Athena Scientific Optimization and Computation Series. Belmont, MA : Athena Scientific, 2016. ISBN : 978-1-886529-05-2 1-886529-05-1 (cf. p. 92).

[27] Hanspeter Bieri et Walter Nef. "A Recursive Sweep-Plane Algorithm, Determining All Cells of a Finite Division of R^d". In : *Computing* 28 (1982), p. 189-198 (cf. p. 6, 53, 87, 145).

[28] Anders Björner, Michel Las Vergnas, Bernd Sturmfels, Neil White et Günter Ziegler. *Oriented Matroids*. Cambridge, UK : Cambridge University Press, 2000. ISBN : 0-521-77750-X (cf. p. 50).

[29] J Frédéric Bonnans, Jean Charles Gilbert, Claude Lemaréchal et Claudia A Sagastizábal. *Numerical Optimization : Theoretical and Practical Aspects*. second. Universitext. Berlin : Springer-Verlag, 2006. ISBN : 3-540-35445-X (cf. p. 89, 92, 173, 183, 274).

[30] J. Frédéric Bonnans. "Local Analysis of Newton-type Methods for Variational Inequalities and Nonlinear Programming". In : *Applied Mathematics & Optimization* 29.2 (mars 1994), p. 161-186. ISSN : 0095-4616. DOI : 10.1007/BF01204181 (cf. p. 31).

[31] J. Frédéric Bonnans, Jean Charles Gilbert, Claude Lemaréchal et Claudia A Sagastizábal. *Optimisation Numérique – Aspects théoriques et pratiques*. Mathématiques et Applications 27. Springer Verlag, Berlin, 1997 (cf. p. 89, 92).

[32] Jonathan M. Borwein et Adrian S. Lewis. *Convex Analysis and Nonlinear Optimization : Theory and Examples*. 2nd ed. CMS Books in Mathematics 3. New York : Springer, 2006. ISBN : 978-0-387-29570-1 (cf. p. 59, 70).

[33] Marie-Charlotte Brandenburg, Jesús De Loera et Chiara Meroni. *The Best Ways to Slice a Polytope*. Juill. 2024. doi : 10.1090/mcom/4006 (cf. p. 145).

[34] David Bremner, Komei Fukuda et Ambros Marzetta. "Primal—Dual Methods for Vertex and Facet Enumeration". In : *Discrete & Computational Geometry* 20.3 (oct. 1998), p. 333-357. issn : 0179-5376. doi : 10.1007/PL00009389 (cf. p. 53, 220).

[35] Taylor Brysiewicz, Holger Eble et Lukas Kühne. "Computing Characteristic Polynomials of Hyperplane Arrangements with Symmetries". In : *Discrete & Computational Geometry* 70.4 (déc. 2023), p. 1356-1377. issn : 0179-5376, 1432-0444. doi : 10.1007/s00454-023-00557-2 (cf. p. 6, 50, 52, 55, 98, 131, 136, 140, 142, 145, 202, 248, 256).

[36] Hannes Buchholzer, Christian Kanzow, Peter Knabner et Serge Kräutle. "The Semismooth Newton Method for the Solution of Reactive Transport Problems Including Mineral Precipitation-Dissolution Reactions". In : *Computational Optimization and Applications* 50.2 (oct. 2011), p. 193-221. issn : 0926-6003, 1573-2894. doi : 10.1007/s10589-010-9379-6 (cf. p. 2).

[37] R. Creighton Buck. "Partition of Space". In : *American Mathematical Monthly* 50 (1943), p. 541-544. issn : 0002-9890,1930-0972. doi : 10.2307/2303424 (cf. p. 49).

[38] Quan M. Bui et Howard C. Elman. "Semi-Smooth Newton Methods for Nonlinear Complementarity Formulation of Compositional Two-Phase Flow in Porous Media". In : *Journal of Computational Physics* 407 (avr. 2020), p. 109163. issn : 00219991. doi : 10.1016/j.jcp.2019.109163 (cf. p. 2).

[39] Yair Carmon, John C. Duchi, Oliver Hinder et Aaron Sidford. "Lower Bounds for Finding Stationary Points I". In : *Mathematical Programming* 184.1-2 (nov. 2020), p. 71-120. issn : 0025-5610, 1436-4646. doi : 10.1007/s10107-019-01406-y (cf. p. 46, 47).

[40] Yair Carmon, John C. Duchi, Oliver Hinder et Aaron Sidford. "Lower Bounds for Finding Stationary Points II : First-Order Methods". In : *Mathematical Programming* 185.1-2 (fév. 2021), p. 315-355. issn : 0025-5610. doi : 10.48550/arXiv.1711.00841. arXiv : 1711.00841 [math] (cf. p. 46, 47).

[41] Frédéric Cazals et Sébastien Loriot. "Computing the Arrangement of Circles on a Sphere, with Applications in Structural Biology". In : *Computational Geometry* 42.6-7 (août 2009), p. 551-565. issn : 09257721. doi : 10.1016/j.comgeo.2008.10.004 (cf. p. 145).

[42] Michal Černý, Miroslav Rada, Jaromír Antoch et Milan Hladík. "A Class of Optimization Problems Motivated by Rank Estimators in Robust Regression". In : *Optimization* 71 (2022), p. 2241-2271. doi : 10.48550/arXiv.1910.05826. arXiv : 1910.05826 [math] (cf. p. 54, 75).

[43] Bintong Chen, Xiaojun Chen et Christian Kanzow. "A Penalized Fischer-Burmeister NCP-function : Theoretical Investigation and Numerical Results". In : *Mathematical Programming* 88.1 (juin 2000), p. 211-216. issn : 0025-5610. doi : 10.1007/PL00011375 (cf. p. 38).

[44] Bintong Chen et Patrick T. Harker. "A Non-Interior-Point Continuation Method for Linear Complementarity Problems". In : *SIAM Journal on Matrix Analysis and Applications* 14.4 (oct. 1993), p. 1168-1190 (cf. p. 3, 16).

[45] Xiaojun Chen. "Superlinear Convergence of Smoothing Quasi-Newton Methods for Nonsmooth Equations". In : *Journal of Computational and Applied Mathematics* 80.1 (avr. 1997), p. 105-126. ISSN : 03770427. DOI : 10.1016/S0377-0427(97)80133-1 (cf. p. 4, 43).

[46] Xiaojun Chen, Zuhair Nashed et Liqun Qi. "Smoothing Methods and Semismooth Methods for Nondifferentiable Operator Equations". In : *SIAM Journal on Numerical Analysis* 38.4 (jan. 2000), p. 1200-1216. ISSN : 0036-1429, 1095-7170. DOI : 10.1137/S0036142999356719 (cf. p. 4, 44).

[47] Xiaojun Chen, Liqun Qi et Defeng Sun. "Global and Superlinear Convergence of the Smoothing Newton Method and Its Application to General Box Constrained Variational Inequalities". In : *Mathematics of Computation* 67.222 (1998), p. 519-540. ISSN : 0025-5718, 1088-6842. DOI : 10.1090/S0025-5718-98-00932-6 (cf. p. 3, 4, 43).

[48] Xiaojun Chen et Shuhuang Xiang. "Sparse Solutions of Linear Complementarity Problems". In : *Mathematical Programming* 159.1-2 (sept. 2016), p. 539-556. ISSN : 0025-5610, 1436-4646. DOI : 10.1007/s10107-015-0950-x (cf. p. 1).

[49] Sung Jin Chung. "NP-Completeness of the Linear Complementarity Problem". In : *Journal of Optimization Theory and Applications* 60.3 (mars 1989), p. 393-399. ISSN : 0022-3239, 1573-2878. DOI : 10.1007/BF00940344 (cf. p. 3, 14).

[50] Vašek Chvátal. *Linear Programming*. A Series of Books in the Mathematical Sciences. New York : W. H. Freeman and Company, 1983. ISBN : 0-7167-1195-8 0-7167-1587-2 (cf. p. 89, 92, 184).

[51] Frank H. Clarke. *Optimization and Nonsmooth Analysis*. second. T. 1. Classics in Applied Mathematics. Philadelphia, PA, USA : Society for Industrial and Applied Mathematics (SIAM), 1990. ISBN : 0-89871-256-4 (cf. p. 4, 19-23, 59-61, 121, 222, 224).

[52] Kenneth L Clarkson et Peter W Shor. "Applications of Random Sampling in Computational Geometry, II". In : *Discrete & Computational Geometry* 4 (1989), p. 387-421. ISSN : 0179-5376,1432-0444. DOI : 10.1007/BF02187740 (cf. p. 50, 79).

[53] Kenneth L. Clarkson. "New Applications of Random Sampling in Computational Geometry". In : *Discrete & Computational Geometry* 2.2 (juin 1987), p. 195-222. ISSN : 0179-5376, 1432-0444. DOI : 10.1007/BF02187879 (cf. p. 50).

[54] Richard W. Cottle et George B. Dantzig. "A Generalization of the Linear Complementarity Problem". In : *Journal of Combinatorial Theory* 8.1 (jan. 1970), p. 79-90. ISSN : 00219800. DOI : 10.1016/S0021-9800(70)80010-2 (cf. p. 2, 59).

[55] Richard Warren Cottle. "Linear Complementarity since 1978". In : *Variational Analysis and Applications*. Nonconvex Optimization and Its Applications 79.1 (2005), p. 239-257 (cf. p. 2).

[56] Richard Warren Cottle. "Nonlinear Programs with Positively Bounded Jacobians". ? Berkeley, USA : University of California, 1964 (cf. p. 1).

[57] Richard Warren Cottle. "Nonlinear Programs with Positively Bounded Jacobians". In : *SIAM Journal on Applied Mathematics* 14.1 (jan. 1966), p. 1-12. DOI : 10.1137/0114012 (cf. p. 1).

[58] Richard Warren Cottle, Jong-Shi Pang et Richard E. Stone. *The Linear Complementarity Problem*. SIAM. Classics in Applied Mathematics 60. Philadelphia, PA, USA : Society for Industrial and Applied Mathematics (SIAM), 2009. isbn : 978-0-89871-686-3 (cf. p. 1, 2, 12, 16, 59, 224, 225).

[59] Thomas M. Cover et Joy A. Thomas. *Elements of Information Theory*. second. Hoboken, NJ : Wiley-Interscience [John Wiley & Sons], 2006. isbn : 978-0-471-24195-9 0-471-24195-4 (cf. p. 91).

[60] Gregory E. Coxson. "The P-matrix Problem Is Co-NP-complete". In : *Mathematical Programming* 64.1-3 (mars 1994), p. 173-178. issn : 0025-5610, 1436-4646. doi : 10.1007/BF01582570 (cf. p. 3, 13).

[61] Henry H. Crapo et Gian-Carlo Rota. *On the Foundations of Combinatorial Theory : Combinatorial Geometries*. The M.I.T. Press, Cambridge, Mass.-London, 1970 (cf. p. 6, 49, 145).

[62] Jad Dabaghi, Vincent Martin et Martin Vohralík. "Adaptive Inexact Semismooth Newton Methods for the Contact Problem Between Two Membranes". In : *Journal of Scientific Computing* 84.2 (août 2020), p. 28. issn : 0885-7474, 1573-7691. doi : 10.1007/s10915-020-01264-3 (cf. p. 2).

[63] Aris Daniilidis, Mounir Haddou, Tri Minh Le et Olivier Ley. "Solving Nonlinear Absolute Value Equations". In : (2024) (cf. p. 16, 44).

[64] Jesús De Loera, Jörg Rambau et Francisco Santos. *Triangulations - Structures for Algorithms and Applications*. Algorithms and Computation in Mathematics 25. Berlin : Springer-Verlag, 2010. isbn : 978-3-642-12970-4 (cf. p. 49).

[65] Tecla De Luca, Francisco Facchinei et Christian Kanzow. "A Semismooth Equation Approach to the Solution of Nonlinear Complementarity Problems". In : *Mathematical Programming* 75.3 (déc. 1996), p. 407-439. issn : 0025-5610, 1436-4646. doi : 10.1007/BF02592192 (cf. p. 24, 26, 36).

[66] Tecla De Luca, Francisco Facchinei et Christian Kanzow. "A Theoretical and Numerical Comparison of Some Semismooth Algorithms for Complementarity Problems". In : *Computational Optimization and Applications* 16 (jan. 2000), p. 173-205. doi : 10.1023/A:1008705425484 (cf. p. 29, 37, 62).

[67] Wolfram Decker, Christian Eder, Claus Fieker, Max Horn et Michael Joswig. *The Computer Algebra System OSCAR : Algorithms and Examples*. 1re éd. T. 32. Algorithms and {C}omputation in {M}athematics. Springer, 2024. isbn : 1431-1550 (issn) (cf. p. 6).

[68] The Sage Developers. *Sagemath, the Sage Mathematics*. 2024 (cf. p. 6, 52, 145).

[69] Elizabeth D. Dolan et Jorge J. Moré. "Benchmarking Optimization Software with Performance Profiles". In : *Mathematical Programming* 91.2 (jan. 2002), p. 201-213. issn : 0025-5610, 1436-4646. doi : 10.1007/s101070100263 (cf. p. 198).

[70] György Dósa, István Szalkai et Claude Laflamme. "The Maximum and Minimum Number of Circuits and Bases of Matroids." In : *Pure Mathematics and Applications*. Mathematics of Optimization 15.4 (sept. 2006), p. 383-392 (cf. p. 51, 99, 156).

[71] Jean-Pierre Dussault, Mathieu Frappier et Jean Charles Gilbert. “A Lower Bound on the Iterative Complexity of the Harker and Pang Globalization Technique of the Newton-min Algorithm for Solving the Linear Complementarity Problem”. In : *EURO Journal on Computational Optimization* 7.4 (déc. 2019), p. 359-380. issn : 21924406. doi : 10.1007/s13675-019-00116-6 (cf. p. 59).

[72] Jean-Pierre Dussault, Mathieu Frappier et Jean Charles Gilbert. “Polyhedral Newton-min Algorithms for Complementarity Problems”. In : *Mathematical Programming* (2025) (cf. p. 7, 9, 32, 34, 59, 205-208, 228, 234, 235).

[73] Jean-Pierre Dussault et Jean Charles Gilbert. “Exact Computation of an Error Bound for the Balanced Linear Complementarity Problem with Unique Solution - The Full Report”. In : *Mathematical Programming* 199 (mai 2023), 1221-1238*. doi : 10.1007/s10107-022-01860-1 (cf. p. 29, 59).

[74] Jean-Pierre Dussault, Jean Charles Gilbert et Baptiste Plaquevent-Jourdain. *Computing the B-differential of the Componentwise Minimum of Two Vector Functions - Partial Description by Linearization*. Rapp. tech. Inria Paris, Université de Sherbrooke, 2025 (cf. p. 76, 77, 85, 86, 106).

[75] Jean-Pierre Dussault, Jean Charles Gilbert et Baptiste Plaquevent-Jourdain. *ISF and BDIFFMIN*. 2023 (cf. p. 7, 86, 98, 106, 198).

[76] Jean-Pierre Dussault, Jean Charles Gilbert et Baptiste Plaquevent-Jourdain. *ISF and BDIFFMIN - MATLAB Functions for Central Hyperplane Arrangements and the Computation of the B-differential of the Componentwise Minimum of Two Affine Vector Functions*. Technical Report. Inria Paris, Université de Sherbrooke, 2023 (cf. p. 7, 86, 198).

[77] Jean-Pierre Dussault, Jean Charles Gilbert et Baptiste Plaquevent-Jourdain. “On the B-differential of the Componentwise Minimum of Two Affine Vector Functions”. In : *Mathematical Programming Computation* (2025) (cf. p. 7, 57, 109, 145-147, 150, 154-156, 159, 166, 167, 171, 175, 176, 178, 186, 195, 196, 198, 237).

[78] Jean-Pierre Dussault, Jean Charles Gilbert et Baptiste Plaquevent-Jourdain. *On the B-differential of the Componentwise Minimum of Two Affine Vector Functions - The Full Report*. Technical Report. Inria Paris, Université de Sherbrooke, 2025, p. 62 (cf. p. 61, 62, 71, 76-78, 80, 82, 85, 94, 108, 173).

[79] Jean-Pierre Dussault, Jean Charles Gilbert et Baptiste Plaquevent-Jourdain. “Primal and Dual Approaches for the Chamber Enumeration of Real Hyperplane Arrangements”. In : *(submitted)* (2025) (cf. p. 7, 87, 97, 99, 106, 143).

[80] Jean-Pierre Dussault, Jean Charles Gilbert et Baptiste Plaquevent-Jourdain. *Primal and Dual Approaches for the Chamber Enumeration of Real Hyperplane Arrangements - The Full Report*. Technical Report (in Preparation). Inria Paris, Université de Sherbrooke, 2025 (cf. p. 146, 150, 153, 154, 157, 168, 190-192, 195).

[81] Herbert Edelsbrunner. *Algorithms in Combinatorial Geometry*. T. 10. EATCS Monographs on Theoretical Computer Science. Berlin : Springer-Verlag, 1987. isbn : 3-540-13722-X (cf. p. 6, 49, 75, 79, 145, 220).

[82] Herbert Edelsbrunner et Leonidas J. Guibas. “Topologically Sweeping an Arrangement”. In : *Journal of Computer and system Sciences* 38 (1989), p. 165-194. issn : 2543-991X, 2080-5519. doi : 10.14708/wm.v48i2.316 (cf. p. 53).

[83] Herbert Edelsbrunner, Joseph O’ROURKE et Raimund Seidel. “CONSTRUCTING ARRANGEMENTS OF LINES AND HYPERPLANES WITH APPLICATIONS”. In : *SIAM Journal on Computation* 15.2 (1986), p. 341-363 (cf. p. 6, 53, 75, 87, 145).

[84] Kenny Erleben et Sarah Niebe. “Numerical Methods for Linear Complementarity Problems in Physics-Based Animation”. In : *Synthesis Lectures on Computer Graphics and Animation* 18 (jan. 2015), 1-159 (?) issn : 978-3-031-79563-3. doi : 10.1007/978-3-031-79564-0 (cf. p. 2).

[85] Francisco Facchinei et Christian Kanzow. “A Nonsmooth Inexact Newton Method for the Solution of Large-Scale Nonlinear Complementarity Problems”. In : *Mathematical Programming* 76.3 (mars 1997), p. 493-512. issn : 0025-5610, 1436-4646. doi : 10.1007/BF02614395 (cf. p. 36).

[86] Francisco Facchinei et Jong-Shi Pang, éd. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research and Financial Engineering. New York, NY : Springer-Verlag New York, Inc, 2003. isbn : 978-0-387-95580-3 978-0-387-21814-4. doi : 10.1007/b97543 (cf. p. 1-4, 12, 16, 18, 25, 30, 32, 33, 35, 39, 43, 59, 60, 230).

[87] Francisco Facchinei et João Soares. “A New Merit Function For Nonlinear Complementarity Problems And A Related Algorithm”. In : *SIAM Journal on Optimization* 7.1 (fév. 1997), p. 225-247. issn : 1052-6234, 1095-7189. doi : 10.1137/S1052623494279110 (cf. p. 31, 36, 59, 122, 129, 225).

[88] Yahya Fathi. “Computational Complexity of LCPs Associated with Positive Definite Symmetric Matrices”. In : *Mathematical Programming* 17.1 (déc. 1979), p. 335-344. issn : 0025-5610, 1436-4646. doi : 10.1007/BF01588254 (cf. p. 3).

[89] J.-A. Ferrez, Komei Fukuda et Thomas M. Liebling. “Solving the Fixed Rank Convex Quadratic Maximization in Binary Variables by a Parallel Zonotope Construction Algorithm”. In : *European Journal of Operational Research* 166.1 (oct. 2005), p. 35-50. issn : 03772217. doi : 10.1016/j.ejor.2003.04.011 (cf. p. 54).

[90] Michael C. Ferris, Christian Kanzow et Todd S. Munson. “Feasible Descent Algorithms for Mixed Complementarity Problems”. In : *Mathematical Programming* 86.3 (déc. 1999), p. 475-497. issn : 0025-5610. doi : 10.1007/s101070050101 (cf. p. 3).

[91] Michael C. Ferris et Jong-Shi Pang. “Engineering and Economic Applications of Complementarity Problems”. In : *SIAM Review* 39.4 (jan. 1997), p. 669-713. issn : 0036-1445, 1095-7200. doi : 10.1137/S0036144595285963 (cf. p. 2).

[92] Andreas Fischer. “A Special Newton-type Optimization Method”. In : *Optimization* 24 (jan. 1992), p. 269-284. doi : 10.1080/02331939208843795 (cf. p. 4, 18, 35).

[93] Andreas Fischer. “Solution of Monotone Complementarity Problems with Locally Lipschitzian Functions”. In : *Mathematical Programming* 76.3 (mars 1997), p. 513-532. issn : 0025-5610, 1436-4646. doi : 10.1007/BF02614396 (cf. p. 36).

[94] Andreas Fischer et Houyuan Jiang. “Merit Functions for Complementarity and Related Problems : A Survey”. In : *Computational Optimization and Applications* 17 (déc. 2000), p. 159-182. doi : 10.1023/A:1026598214921 (cf. p. 4, 19).

[95] Andreas Fischer et Christian Kanzow. “On Finite Termination of an Iterative Method for Linear Complementarity Problems”. In : *Mathematical Programming* 74.3 (sept. 1996), p. 279-292. issn : 0025-5610, 1436-4646. doi : 10.1007/BF02592200 (cf. p. 29-31).

[96] Robert M. Freund et James B. Orlin. “On the Complexity of Four Polyhedral Set Containment Problems”. In : *Mathematical Programming* 33.2 (nov. 1985), p. 139-145. issn : 0025-5610, 1436-4646. doi : 10.1007/BF01582241 (cf. p. 220).

[97] Komei Fukuda. “From the Zonotope Construction to the Minkowski Addition of Convex Polytopes”. In : *Journal of Symbolic Computation* 38.4 (oct. 2004), p. 1261-1272. issn : 07477171. doi : 10.1016/j.jsc.2003.08.007 (cf. p. 53).

[98] Masao Fukushima. “Equivalent Differentiable Optimization Problems and Descent Methods for Asymmetric Variational Inequality Problems”. In : *Mathematical Programming* 53.1-3 (jan. 1992), p. 99-110. issn : 0025-5610, 1436-4646. doi : 10.1007/BF01585696 (cf. p. 3, 4, 39, 41).

[99] Aurél Galántai. “Properties and Construction of NCP Functions”. In : *Computational Optimization and Applications* 52.3 (juill. 2012), p. 805-824. issn : 0926-6003, 1573-2894. doi : 10.1007/s10589-011-9428-9 (cf. p. 4, 19).

[100] Yu Gao, Haiming Song, Xiaoshen Wang et Kai Zhang. “Primal-Dual Active Set Method for Pricing American Better-of Option on Two Assets”. In : *Communications in Nonlinear Science and Numerical Simulation* 80 (jan. 2020), p. 104976. issn : 10075704. doi : 10.1016/j.cnsns.2019.104976 (cf. p. 2).

[101] Ewgenij Gawrilow et Michael Joswig. “Polymake : A Framework for Analyzing Convex Polytopes.” In : *Polytopes—Combinatorics and Computation (Oberwolfach, 1997)*. DMV Sem. 29. Basel : Birkhäuser, 2000, p. 43-73. isbn : 3-7643-6351-7 (cf. p. 52).

[102] Bennet Gebken. *Analyzing the Speed of Convergence in Nonsmooth Optimization via the Goldstein Subdifferential with Application to Descent Methods*. Oct. 2024. arXiv : 2410.01382 [math] (cf. p. 48).

[103] Carl Geiger et Christian Kanzow. “On the Resolution of Monotone Complementarity Problems”. In : *Computational Optimization and Applications* 5.2 (mars 1996), p. 155-173. issn : 0926-6003, 1573-2894. doi : 10.1007/BF00249054 (cf. p. 36).

[104] Helmut Gfrerer et Jiří V. Outrata. “On a Semismooth* Newton Method for Solving Generalized Equations”. In : *SIAM Journal on Optimization* 31.1 (jan. 2021), p. 489-517. issn : 1052-6234, 1095-7189. doi : 10.1137/19M1257408 (cf. p. 42).

[105] Jean Charles Gilbert. *Fragments d’Optimisation Différentiable - Théories et Algorithmes*. /. T. 1. /. / : /, 2021 (cf. p. 15, 26, 92, 109, 183, 184, 270, 274).

[106] Jean Charles Gilbert. *Selected Topics on Continuous Optimization - Version 2*. Lecture Notes of the Master-2 "Optimization" at the University Paris-Saclay. Paris, 2022 (cf. p. 89).

[107] Allen A. Goldstein. “Optimization of Lipschitz Continuous Functions”. In : *Mathematical Programming* 13.1 (déc. 1977), p. 14-22. issn : 0025-5610, 1436-4646. doi : 10.1007/BF01584320 (cf. p. 47).

[108] Paul Gordan. “Über die Auflösung linearer Gleichungen mit reellen Coefficienten.” In : *Mathematische Annalen* (1873), p. 23-28. doi : 10.1007/BF01442864 (cf. p. 52, 67, 145, 147).

[109] Nicholas Ian Mark Gould et Jennifer Scott. “A Note on Performance Profiles for Benchmarking Software”. In : *ACM Transactions on Mathematical Software* 43.2 (juin 2017), p. 1-5. issn : 0098-3500, 1557-7295. doi : 10.1145/2950048 (cf. p. 198).

[110] M. Seetharama Gowda et Roman Sznajder. “The Generalized Order Linear Complementarity Problem”. In : *SIAM Journal on Matrix Analysis and Applications* 15.3 (juill. 1994), p. 779-795. issn : 0895-4798. doi : 10.1137/S0895479892237859 (cf. p. 59).

[111] Daniel R. Grayson et Michael E. Stillman. *Macaulay2, a Software System for Research in Algebraic Geometry*. 2024 (cf. p. 6, 52, 145).

[112] Rick Greer. *Trees and Hills : Methodology for Maximizing Functions of Systems of Linear Relations*. North-Holland Mathematics Studies 96. Amsterdam : North-Holland Publishing Co., 1984. isbn : 0-444-87578-6 (cf. p. 71).

[113] Luigi Grippo, Francesco Lampariello et Stefano Lucidi. “A Nonmonotone Line Search Technique for Newton’s Method”. In : *SIAM Journal on Numerical Analysis* 23.4 (août 1986), p. 707-716. issn : 0036-1429, 1095-7170. doi : 10.1137/0723046 (cf. p. 26).

[114] Branko Grünbaum. *Convex Polytopes*. Interscience Publishers John Wiley & Sons, Inc. T. 16. Pure and Applied Mathematics. New York : AMS, Providence, RI, 1967 (cf. p. 54, 75, 79).

[115] Osman Güler. *Foundations of Optimization*. T. 258. Graduate Texts in Mathematics. New York, NY : Springer New York, 2010. isbn : 978-0-387-34431-7 978-0-387-68407-9. doi : 10.1007/978-0-387-68407-9 (cf. p. 147).

[116] Mounir Haddou. “A New Class of Smoothing Methods for Mathematical Programs with Equilibrium Constraints”. In : *Pacific Journal of Optimization* 5 (2009), p. 87-95 (cf. p. 44).

[117] Mounir Haddou et Patrick Maheux. “Smoothing Methods for Nonlinear Complementarity Problems”. In : *Journal of Optimization Theory and Applications* 160.3 (mars 2014), p. 711-729. issn : 0022-3239, 1573-2878. doi : 10.1007/s10957-013-0398-1 (cf. p. 4, 44).

[118] Dan Halperin et Micha Sharir. “Arrangements”. In : *Handbook of Discrete and Computational Geometry*. Jacob E. Goodman and Joseph O’Rourke and Csaba D. Tòth. CRC Press - Taylor \& Francis Group, 2018 (cf. p. 6, 49, 75, 145).

[119] Shih-Ping Han, Jong-Shi Pang et Narayan Rangaraj. “Globally Convergent Newton Methods for Nonsmooth Equations”. In : *Mathematics of Operations Research* 17.3 (août 1992), p. 586-607. issn : 0364-765X, 1526-5471. doi : 10.1287/moor.17.3.586 (cf. p. 22, 32, 33).

[120] Patrick T. Harker et Jong-Shi Pang. “Finite-Dimensional Variational Inequality and Nonlinear Complementarity Problems : A Survey of Theory, Algorithms and Applications”. In : *Mathematical Programming* 48.1-3 (mars 1990), p. 161-220. issn : 0025-5610, 1436-4646. doi : 10.1007/BF01582255 (cf. p. 2).

[121] Patrick T. Harker et Baichun Xiao. “Newton’s Method for the Nonlinear Complementarity Problem : A B-differentiable Equation Approach”. In : *Mathematical Programming* 48.1-3 (mars 1990), p. 339-357. issn : 0025-5610, 1436-4646. doi : 10.1007/BF01582262 (cf. p. 15).

[122] Xiahui He et Peng Yang. “The Primal-Dual Active Set Method for a Class of Nonlinear Problems with T -Monotone Operators”. In : *Mathematical Problems in Engineering* 2019.1 (jan. 2019). Sous la dir. de Vyacheslav Kalashnikov, p. 1-8. issn : 1024-123X, 1563-5147. doi : 10.1155/2019/2912301 (cf. p. 2, 3, 42).

[123] Juha Heinonen. *Lectures on Lipschitz Analysis*. Report. University of Jyväskylä Department of Mathematics and Statistics 100. University of Jyväskylä, 2005. isbn : 951-39-2318-5 (cf. p. 20).

[124] Michael Hintermüller, Kazufumi Ito et Karl Kunisch. “The Primal-Dual Active Set Strategy as a Semismooth Newton Method”. In : *SIAM Journal on Optimization* 13.3 (jan. 2002), p. 865-888. issn : 1052-6234, 1095-7189. doi : 10.1137/S1052623401383558 (cf. p. 3, 42).

[125] Michael Hintermüller et Ian Kopacka. “Mathematical Programs with Complementarity Constraints in Function Space : C- and Strong Stationarity and a Path-Following Algorithm”. In : *SIAM Journal on Optimization* 20.2 (jan. 2009), p. 868-902. issn : 1052-6234, 1095-7189. doi : 10.1137/080720681 (cf. p. 17).

[126] Jean-Baptiste Hiriart-Urruty et Claude Lemaréchal. *Fundamentals of Convex Analysis*. Grundlehren Text Editions. Berlin, Heidelberg : Springer Berlin Heidelberg, 2001. isbn : 978-3-540-42205-1 978-3-642-56468-0. doi : 10.1007/978-3-642-56468-0 (cf. p. 59, 227).

[127] Tim Hoheisel, Christian Kanzow, Boris S. Mordukhovich et Hung M. Phan. “Generalized Newton’s Method Based on Graphical Derivatives”. In : *Nonlinear Analysis : Theory, Methods & Applications* 75.3 (fév. 2012), p. 1324-1340. issn : 0362546X. doi : 10.1016/j.na.2011.06.039 (cf. p. 4, 42).

[128] Stefan Hüeber, Georg Stadler et Barbara I. Wohlmuth. “A Primal-Dual Active Set Algorithm for Three-Dimensional Contact Problems with Coulomb Friction”. In : *SIAM Journal on Scientific Computing* 30.2 (jan. 2008), p. 572-596. issn : 1064-8275, 1095-7197. doi : 10.1137/060671061 (cf. p. 2, 3).

[129] Stefan Hüeber et Barbara I. Wohlmuth. “A Primal–Dual Active Set Strategy for Non-Linear Multibody Contact Problems”. In : *Computer Methods in Applied Mechanics and Engineering* 194.27-29 (juill. 2005), p. 3147-3166. issn : 00457825. doi : 10.1016/j.cma.2004.08.006 (cf. p. 2).

[130] Anwar A Irmatov. “Arrangements of Hyperplanes and the Number of Threshold Functions”. In : *Acta Applicandae Mathematicae* 68 (2001), p. 211-226. doi : 10.1023/A:1012087813557 (cf. p. 55).

[131] Kazufumi Ito et Karl Kunisch. “On a Semi-Smooth Newton Method and Its Globalization”. In : *Mathematical Programming* 118.2 (mai 2009), p. 347-370. issn : 0025-5610, 1436-4646. doi : 10.1007/s10107-007-0196-3 (cf. p. 3, 42).

[132] Alexey F. Izmailov et Mikhail V. Solodov. *Newton-Type Methods for Optimization and Variational Problems*. Springer Series in Operations Research and Financial Engineering. Cham : Springer International Publishing, 2014. ISBN : 978-3-319-04246-6 978-3-319-04247-3. DOI : 10.1007/978-3-319-04247-3 (cf. p. 31, 35, 59, 62).

[133] Michael I. Jordan, Tianyi Lin et Manolis Zampetakis. *On the Complexity of Deterministic Nonsmooth and Nonconvex Optimization*. Nov. 2022. DOI : 10.48550/arXiv.2209.12463. arXiv : 2209.12463 [math] (cf. p. 47).

[134] Norman H. Josephy. *Newton's Method for Generalized Equations*. Rapp. tech. ADA077096. wisconsin : University of Madison, 1979, p. 37 (cf. p. 15, 29).

[135] C. Kanzow. "Nonlinear Complementarity as Unconstrained Optimization". In : *Journal of Optimization Theory and Applications* 88.1 (jan. 1996), p. 139-155. ISSN : 0022-3239, 1573-2878. DOI : 10.1007/BF02192026 (cf. p. 40).

[136] Christian Kanzow et Masao Fukushima. "Solving Box Constrained Variational Inequalities by Using the Natural Residual with D-gap Function Globalization". In : *Operations Research Letters* 23.1-2 (août 1998), p. 45-51. ISSN : 01676377. DOI : 10.1016/S0167-6377(98)00023-6 (cf. p. 41, 62).

[137] Christian Kanzow et Helmut Kleinmichel. "A New Class of Semismooth Newton-Type Methods for Nonlinear Complementarity Problems". In : *Computational Optimization and Applications* 11 (1998), p. 227-251 (cf. p. 37, 38).

[138] Christian Kanzow, Nobuo Yamashita et Masao Fukushima. "New NCP-Functions and Their Properties". In : *Journal of Optimization Theory and Applications* 94.1 (juill. 1997), p. 115-135. ISSN : 0022-3239. DOI : 10.1023/A:1022659603268 (cf. p. 4, 18, 19, 29).

[139] Stepan Karamardian. "Generalized Complementarity Problem". In : *Journal of Optimization Theory and Applications* 8.3 (sept. 1971), p. 161-168. ISSN : 0022-3239, 1573-2878. DOI : 10.1007/BF00932464 (cf. p. 2).

[140] Lars Kastner et Marta Panizzut. "Hyperplane Arrangements in Polymake". In : *Mathematical Software – ICMS 2020*. Sous la dir. d'Anna Maria Bigatti, Jacques Carette, James H. Davenport, Michael Joswig et Timo De Wolff. T. 12097. Cham : Springer International Publishing, 2020, p. 232-240. ISBN : 978-3-030-52199-8 978-3-030-52200-1. DOI : 10.1007/978-3-030-52200-1_23 (cf. p. 6, 52, 145).

[141] Leonid G Khachiyan, Endre Boros, Khaled M. Elbassioni, Vladimir A. Gurvich et Kazuhisa Makino. "On the Complexity of Some Enumeration Problems for Matroids". In : *SIAM Journal on Discrete Mathematics* 19.4 (jan. 2005), p. 966-984. ISSN : 0895-4801, 1095-7146. DOI : 10.1137/S0895480103428338 (cf. p. 51, 91, 143, 179).

[142] Robert John Kingan et Sandra Reuben Kingan. "A Software System for Matroids". In : *Graphs and Discovery*. DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 69. Providence, Rhode Island : American Mathematical Society, 2005, p. 287-295. ISBN : 0-8218-3761-3 (cf. p. 52).

[143] Kolja Knauer, Luis Pedro Montejano et Jorge Luis Ramírez Alfonsín. "How Many Circuits Determine an Oriented Matroid ?" In : *Combinatorica* 38.4 (août 2018), p. 861-885. ISSN : 0209-9683, 1439-6912. DOI : 10.1007/s00493-016-3556-x (cf. p. 51).

[144] Masakazu Kojima, Nimrod Megiddo, Toshihito Noma et Akiko Yoshise. *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*. Lecture Notes in Computer Science 538. Berlin : Springer, jan. 1991. ISBN : 978-3-540-54509-5 (cf. p. 3, 14, 16).

[145] Masakazu Kojima, Shinji Mizuno et Akiko Yoshise. "A Polynomial-Time Algorithm for a Class of Linear Complementarity Problems". In : *Mathematical Programming* 44.1-3 (mai 1989), p. 1-26. ISSN : 0025-5610, 1436-4646. DOI : 10.1007/BF01587074 (cf. p. 3, 16).

[146] Masakazu Kojima et Susumu Shindo. "Extension of Newton and Quasi-Newton Methods to Systems of PC1 Equations". In : *Journal of the Operations Research Society of Japan* 29.4 (déc. 1986), p. 352-375. DOI : 10.15807/jorsj.29.352 (cf. p. 27, 59).

[147] Michael Martin Kostreva. "Direct Algorithms for Complementarity Problems". Thèse de doct. Ann Arbor, MI : Rensselaer Polytechnic Institute, 1976 (cf. p. 18).

[148] Lukas Kühne. "The Universality of the Resonance Arrangement and Its Betti Numbers". In : *Combinatorica* 43.2 (avr. 2023), p. 277-298. ISSN : 0209-9683, 1439-6912. DOI : 10.1007/s00493-023-00006-x (cf. p. 55, 98).

[149] Adrian Kulmburg et Matthias Althoff. "On the Co-NP-completeness of the Zonotope Containment Problem". In : *European Journal of Control* 62 (nov. 2021), p. 84-91. ISSN : 09473580. DOI : 10.1016/j.ejcon.2021.06.028 (cf. p. 221, 273, 275, 277).

[150] Bernd Kummer. "NEWTON's METHOD FOR NON-DIFFERENTIABLE FUNCTIONS". In : *Advances in Mathematical Optimization*. Sous la dir. de J. Guddat Et Al. De Gruyter, déc. 1988, p. 114-125. ISBN : 978-3-11-247992-6. DOI : 10.1515/9783112479926-011 (cf. p. 25).

[151] Michel Las Vergnas. *Matroïdes Orientables*. Comptes Rendus Hebdomadaires Des Séances de l'Académie Des Sciences. Séries A et B 280. Paris, 1975. ISBN : 0151-0509 (cf. p. 79, 143, 145).

[152] Carlton E. Lemke. "Bimatrix Equilibrium Points and Mathematical Programming". In : *Management Science* 11.7 (mai 1965), p. 681-689. ISSN : 0025-1909, 1526-5501. DOI : 10.1287/mnsc.11.7.681 (cf. p. 3).

[153] Kenneth Levenberg. "A Method for the Solution of Certain Non-Linear Problems in Least Squares". In : *Quarterly of Applied Mathematics* 2.2 (juill. 1944), p. 164-168. ISSN : 0033-569X, 1552-4485. DOI : 10.1090/qam/10666 (cf. p. 27).

[154] Li-Zhi Liao, Houduo Qi et Liqun Qi. "Solving Nonlinear Complementarity Problems with Neural Networks : A Reformulation Method Approach". In : *Journal of Computational and Applied Mathematics* 131.1-2 (juin 2001), p. 343-359. ISSN : 03770427. DOI : 10.1016/S0377-0427(00)00262-4 (cf. p. 37).

[155] Zhi-Quan Luo, Olvi Leon Mangasarian, Jun Ren et Mikhail V. Solodov. "New Error Bounds for the Linear Complementarity Problem". In : *Mathematics of Operations Research* 19.4 (nov. 1994), p. 880-892. ISSN : 0364-765X, 1526-5471. DOI : 10.1287/moor.19.4.880 (cf. p. 40).

[156] Zhi-Quan Luo et Paul Tseng. "A New Class of Merit Functions for the Nonlinear Complementarity Problem". In : *Complementarity and Variational Problems (Baltimore, MD, 1995)*. SIAM, Philadelphia, PA, 1997, p. 204-225. ISBN : 0-89871-391-9 (cf. p. 18).

[157] Changfeng Ma, Jia Tang et Xiaohong Chen. “A Globally Convergent Levenberg–Marquardt Method for Solving Nonlinear Complementarity Problem”. In : *Applied Mathematics and Computation* 192 (2007), p. 370-381 (cf. p. 4, 45).

[158] Mend-Amar Majig et Masao Fukushima. “Restricted-Step Josephy-Newton Method for General Variational Inequalities with Polyhedral Constraints”. In : *Pacific Journal of Optimization* 6 (mai 2010), p. 15 (cf. p. 41).

[159] Olvi Leon Mangasarian. “Equivalence of the Complementarity Problem to a System of Nonlinear Equations”. In : *SIAM Journal on Applied Mathematics* 31.1 (juill. 1976), p. 89-92. issn : 0036-1399, 1095-712X. doi : 10.1137/0131009 (cf. p. 4, 18).

[160] Olvi Leon Mangasarian et Robert R. Meyer. “Absolute Value Equations”. In : *Linear Algebra and its Applications* 419.2-3 (déc. 2006), p. 359-367. issn : 00243795. doi : 10.1016/j.laa.2006.05.004 (cf. p. 16).

[161] Olvi Leon Mangasarian et Michael V Solodov. “A Linearly Convergent Derivative-Free Descent Method for Strongly Monotone Complementarity Problems”. In : *Computational Optimization and Applications* 14 (1999), p. 5-16. issn : 0926-6003,1573-2894. doi : 10.1023/A:1008752626695 (cf. p. 39, 40).

[162] Olvi Leon Mangasarian et Mikhail V. Solodov. “Nonlinear Complementarity as Unconstrained and Constrained Minimization”. In : *Mathematical Programming* 62.1-3 (fév. 1993), p. 277-297. issn : 0025-5610. doi : 10.1007/BF01585171 (cf. p. 39).

[163] Estelle Marchand, Torsten Müller et Peter Knabner. “Fully Coupled Generalised Hybrid-Mixed Finite Element Approximation of Two-Phase Two-Component Flow in Porous Media. Part II : Numerical Scheme and Numerical Results”. In : *Computational Geosciences* 16.3 (juin 2012), p. 691-708. issn : 1420-0597, 1573-1499. doi : 10.1007/s10596-012-9279-1 (cf. p. 2, 59).

[164] Estelle Marchand, Torsten Müller et Peter Knabner. “Fully Coupled Generalized Hybrid-Mixed Finite Element Approximation of Two-Phase Two-Component Flow in Porous Media. Part I : Formulation and Properties of the Mathematical Model”. In : *Computational Geosciences* 17.2 (avr. 2013), p. 431-442. issn : 1420-0597, 1573-1499. doi : 10.1007/s10596-013-9341-7 (cf. p. 2, 59).

[165] Donald W. Marquardt. “An Algorithm for Least-Squares Estimation of Nonlinear Parameters”. In : *Journal of the Society for Industrial and Applied Mathematics* 11.2 (juin 1963), p. 431-441. issn : 0368-4245, 2168-3484. doi : 10.1137/0111030 (cf. p. 27).

[166] Arnaud Mary et Yann Strozecki. “Efficient Enumeration of Solutions Produced by Closure Operations”. In : *Discrete Mathematics \& Theoretical Computer Science. DMTCS.* 21 (2019), 52 :1-52 :13. issn : 1868-8969. doi : 10.4230/LIPICS.STACS.2016.52 (cf. p. 51, 91).

[167] Peter McMullen. “On Zonotopes”. In : *Transactions of the American Mathematical Society* 159 (sept. 1971), p. 91-109. doi : 10.2307/1996000 (cf. p. 54, 220, 267).

[168] Nimrod Megiddo. “A Monotone Complementarity Problem with Feasible Solutions but No Complementary Solutions”. In : *Mathematical Programming* 12.1 (déc. 1977), p. 131-132. issn : 0025-5610, 1436-4646. doi : 10.1007/BF01593775 (cf. p. 14).

[169] Nimrod Megiddo. *A Note on the Complexity of P-matrix LCP and Computing an Equilibrium.* Rapp. tech. RJ 6439 (62557). San Jose, CA, USA : Almaden Research Center, 1988, p. 1-6 (cf. p. 14).

[170] Arturo Merino et Torsten Mütze. "Traversing Combinatorial 0/1-Polytopes via Optimization". In : *SIAM Journal on Computing* 53.5 (oct. 2024), p. 1257-1292. issn : 0097-5397, 1095-7111. doi : 10.1137/23M1612019 (cf. p. 53).

[171] Robert Mifflin. "Semismooth and Semiconvex Functions in Constrained Optimization". In : *SIAM Journal on Control and Optimization* 15.6 (nov. 1977), p. 959-972. issn : 0363-0129, 1095-7138. doi : 10.1137/0315061 (cf. p. 24).

[172] Edward Minieka. "Finding the Circuits of a Matroid". In : *JOURNAL OF RESEARCH of the National Bureau of Standards* 80B.3 (1976) (cf. p. 51).

[173] George J Minty. "Montone (Nonlinear) Operators in Hilbert Spaces". In : *Duke Mathematical Journal* 29 (1962), p. 341-346 (cf. p. 15).

[174] Shinji Mizuno, Akiko Yoshise et Takeshi Kikuchi. "PRACTICAL POLYNOMIAL TIME ALGORITHMS FOR LINEAR COMPLEMENTARITY PROBLEMS". In : *Journal of the Operations Research Society of Japan* 32.1 (1989), p. 75-92. issn : 0453-4514, 2188-8299. doi : 10.15807/jorsj.32.75 (cf. p. 3).

[175] Boris S. Mordukhovich. *Second-Order Variational Analysis in Optimization, Variational Stability, and Control : Theory, Algorithms, Applications.* Springer Series in Operations Research and Financial Engineering. Cham : Springer International Publishing, 2024. isbn : 978-3-031-53475-1 978-3-031-53476-8. doi : 10.1007/978-3-031-53476-8 (cf. p. 24).

[176] Boris S. Mordukhovich. *Variational Analysis and Generalized Differentiation. I.* T. 1. Grundlehren Der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Berlin : Springer-Verlag, 2006. isbn : 978-3-540-25437-9 3-540-25437-4 (cf. p. 24).

[177] Boris S. Mordukhovich. *Variational Analysis and Generalized Differentiation. II.* T. 2. Grundlehren Der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Berlin : Springer-Verlag, 2006. isbn : 978-3-540-25438-6 3-540-25438-2 (cf. p. 24).

[178] Theodore S. Motzkin. *Beiträge zur Theorie der linearen Ungleichungen.* Rapp. tech. Jerusalem, Israel : University Basel, 1936 (cf. p. 52, 146, 147).

[179] Todd S. Munson, Francisco Facchinei, Michael C. Ferris, Andreas Fischer et Christian Kanzow. "The Semismooth Algorithm for Large Scale Complementarity Problems". In : *INFORMS Journal on Computing* 13.4 (nov. 2001), p. 294-311. issn : 1091-9856, 1526-5528. doi : 10.1287/ijoc.13.4.294.9734 (cf. p. 3, 35, 42).

[180] Katta G. Murty. "Computational Complexity of Complementary Pivot Methods". In : *Complementarity and Fixed Point Problems* 7 (1978), p. 61-73. doi : 10.1007/BFb0120782 (cf. p. 3).

[181] Katta G. Murty. *Linear Complementarity, Linear and Nonlinear Programming.* Sigma Series in Applied Mathematics 3. Berlin : Heldermann Verlag, 1988. isbn : 3-88538-403-5 (cf. p. 1, 12, 14, 59).

[182] Katta G. Murty et Santosh N. Kabadi. “Some NP-complete Problems in Quadratic and Nonlinear Programming”. In : *Mathematical Programming* 39 (1987), p. 117-129. ISSN : 0025-5610,1436-4646. DOI : 10.1007/BF02592948 (cf. p. 46).

[183] John Lawrence Nazareth et Liqun Qi. “Globalization of Newton’s Method for Solving Nonlinear Equations”. In : *Numerical Linear Algebra with Applications* 3.3 (mai 1996), p. 239-249. ISSN : 1070-5325, 1099-1506. DOI : 10.1002/(SICI)1099-1506(199605/06)3:3<239::AID-NLA81>3.0.CO;2-U (cf. p. 27).

[184] Yurii Nesterov. *Lectures on Convex Optimization*. second. Springer Optim. Appl. 137. springer, Cham, 2018. ISBN : 978-3-319-91577-7 978-3-319-91578-4 (cf. p. 47).

[185] Foundation Inc. OEIS. *The Online Encyclopedia of Integer Sequences*. 2025 (cf. p. 169).

[186] Peter Orlik et Louis Solomon. “Combinatorics and Topology of Complements of Hyperplanes”. In : *Inventiones Mathematicae* 56.2 (fév. 1980), p. 167-189. ISSN : 0020-9910, 1432-1297. DOI : 10.1007/BF01392549 (cf. p. 49, 145).

[187] Peter Orlik et Hiroaki Terao. *Arrangement of Hyperplanes*. T. 300. Grundlehren Der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Madison : Springer-Verlag Berlin Heidelberg GmbH, 1992. ISBN : 3-540-55259-6 (cf. p. 6, 49, 145, 150, 161).

[188] James M. Ortega et Werner C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. 2e éd. Classics in Applied Mathematics 30. Philadelphia, PA, USA : SIAM, 2000. ISBN : 0-89871-461-3 (cf. p. 25).

[189] OSCAR. *OSCAR – Open Source Computer Algebra Research System, Version 1.0.0*. The OSCAR Team. 2024 (cf. p. 6, 145).

[190] El Hassene Osmani, Mounir Haddou, Lina Abdallah et Naceurdine Bensalem. “A New Approach for Solving the Linear Complementarity Problem Using Smoothing Functions”. In : *2021 7th International Conference on Optimization and Applications (ICOA)*. Wolfenbüttel, Germany : IEEE, mai 2021, p. 1-8. ISBN : 978-1-6654-4103-2. DOI : 10.1109/ICOA51614.2021.9442649 (cf. p. 4, 44).

[191] James G. Oxley. *Matroid Theory*. Second edition. Oxford Graduate Texts in Mathematics 21. Oxford New York, NY : Oxford University Press, 2011. ISBN : 978-0-19-856694-6 978-0-19-960339-8 (cf. p. 6, 50, 68, 81, 143, 155).

[192] Jong-Shi Pang. “A B-differentiable Equation-Based, Globally and Locally Quadratically Convergent Algorithm for Nonlinear Programs, Complementarity and Variational Inequality Problems”. In : *Mathematical Programming* 51.1-3 (juill. 1991), p. 101-131. ISSN : 0025-5610, 1436-4646. DOI : 10.1007/BF01586928 (cf. p. 29, 31-33, 59, 225).

[193] Jong-Shi Pang. “Complementarity Problems”. In : *Handbook of Global Optimization*. T. 2. Nonconvex Optimization and Its Applications. Dordrecht : Kluwer, 1995, p. 271-338 (cf. p. 2, 59).

[194] Jong-Shi Pang. “Inexact Newton Methods for the Nonlinear Complementarity Problem”. In : *Mathematical Programming* 36.1 (oct. 1986), p. 54-71. ISSN : 0025-5610, 1436-4646. DOI : 10.1007/BF02591989 (cf. p. 4, 29).

[195] Jong-Shi Pang. “Newton’s Method for B-Differentiable Equations”. In : *Mathematics of Operations Research* 15.2 (mai 1990), p. 311-341. issn : 0364-765X, 1526-5471. doi : 10.1287/moor.15.2.311 (cf. p. 4, 29, 31, 59, 126).

[196] Jong-Shi Pang et Steven A. Gabriel. “NE/SQP : A Robust Algorithm for the Nonlinear Complementarity Problem”. In : *Mathematical Programming* 60.1-3 (juin 1993), p. 295-337. issn : 0025-5610, 1436-4646. doi : 10.1007/BF01580617 (cf. p. 32-34, 225).

[197] Jong-Shi Pang, Shih-Ping Han et Narayan Rangaraj. “Minimization of Locally Lipschitzian Functions”. In : *SIAM Journal on Optimization* 1.1 (fév. 1991), p. 57-82. issn : 1052-6234, 1095-7189. doi : 10.1137/0801006 (cf. p. 4, 16, 22, 32, 33, 46).

[198] Jong-Shi Pang et Liqun Qi. “Nonsmooth Equations : Motivation and Algorithms”. In : *SIAM Journal on Optimization* 3.3 (août 1993), p. 443-465. issn : 1052-6234, 1095-7189. doi : 10.1137/0803021 (cf. p. 28, 62).

[199] Ji-Ming Peng. “Equivalence of Variational Inequality Problems to Unconstrained Minimization”. In : *Mathematical Programming* 78.3 (sept. 1997), p. 347-355. issn : 0025-5610, 1436-4646. doi : 10.1007/BF02614360 (cf. p. 41).

[200] Sandra Pieraccini, Maria Grazia Gasparo et Aldo Pasquali. “Global Newton-type Methods and Semismooth Reformulations for NCP”. In : *Applied Numerical Mathematics* 44.3 (fév. 2003), p. 367-384. issn : 01689274. doi : 10.1016/S0168-9274(02)00169-1 (cf. p. 37).

[201] Knot Pipatsrisawat et Adnan Darwiche. “On the Power of Clause-Learning SAT Solvers as Resolution Engines”. In : *Artificial Intelligence* 175.2 (fév. 2011), p. 512-525. issn : 00043702. doi : 10.1016/j.artint.2010.10.002 (cf. p. 242).

[202] Elijah Polak et Liqun Qi. “Globally and Superlinearly Convergent Algorithm for Minimizing a Normal Merit Function”. In : *SIAM Journal on Control and Optimization* 36.3 (mai 1998), p. 1005-1019. issn : 0363-0129, 1095-7138. doi : 10.1137/S0363012996310245 (cf. p. 41).

[203] Alexander Postnikov et Richard P. Stanley. “Deformations of Coxeter Hyperplane Arrangements”. In : *Journal of Combinatorial Theory, Series A* 91.1-2 (mars 2000), p. 544-597. issn : 00973165. doi : 10.1006/jcta.2000.3106 (cf. p. 50, 55, 98, 195, 262).

[204] Liqun Qi. “Convergence Analysis of Some Algorithms for Solving Nonsmooth Equations”. In : *Mathematics of Operations Research* 18.1 (1993), p. 227-244 (cf. p. 4, 22-24, 27-29, 31, 59, 61, 62, 86).

[205] Liqun Qi. “Trust Region Algorithms for Solving Nonsmooth Equations”. In : *SIAM Journal of Optimization* 5.1 (1995), p. 219-230 (cf. p. 4, 44).

[206] Liqun Qi et Jie Sun. “A Nonsmooth Version of Newton’s Method”. In : *Mathematical Programming* 58 (1993), p. 353-367 (cf. p. 24, 27, 31, 59, 120).

[207] Liqun Qi et Jie Sun. “A Trust Region Algorithm for Minimization of Locally Lipschitzian Functions”. In : *Mathematical Programming* 66.1-3 (août 1994), p. 25-43. issn : 0025-5610, 1436-4646. doi : 10.1007/BF01581136 (cf. p. 29, 32, 33).

[208] Miroslav Rada et Michal Černý. “A New Algorithm for Enumeration of Cells of Hyperplane Arrangements and a Comparison with Avis and Fukuda’s Reverse Search”. In : *SIAM Journal on Discrete Mathematics* 32.1 (jan. 2018), p. 455-473. ISSN : 0895-4801, 1095-7146. DOI : 10.1137/15M1027930 (cf. p. 6, 53, 60, 61, 87-90, 97, 98, 100, 105, 106, 145, 146, 149, 171-173, 195, 196, 198, 202, 252).

[209] Hans Rademacher. “Über partielle und totale Differenzierbarkeit von Funktionen mehrerer Variablen und über die Transformation der Doppelintegrale”. In : *Matematische Annalen* 79 (jan. 1919). DOI : 10.1007/BF01498415 (cf. p. 20, 59, 112).

[210] Manuel Radons et Josué Tonelli-Cueto. “Generalized Perron Roots and Solvability of the Absolute Value Equation”. In : *SIAM Journal on Matrix Analysis and Applications* 44.4 (déc. 2023), p. 1645-1666. ISSN : 0895-4798, 1095-7162. DOI : 10.1137/22M1517184 (cf. p. 16).

[211] Daniel Ralph. “Global Convergence of Damped Newton’s Method for Nonsmooth Equations via the Path Search”. In : *Mathematics of Operations Research* 19.2 (mai 1994), p. 352-389. ISSN : 0364-765X, 1526-5471. DOI : 10.1287/moor.19.2.352 (cf. p. 15).

[212] Jörg Rambau. “Symmetric Lexicographic Subset Reverse Search for the Enumeration of Circuits, Cocircuits, and Triangulations up to Symmetry”. In : (2023), p. 1-41 (cf. p. 52, 53, 91, 140-142, 145, 179, 202, 248, 256).

[213] Jörg Rambau. “The Visible-Volume Function of a Set of Cameras Is Continuous, Piecewise Rational, Locally Lipschitz, and Semi-Algebraic in All Dimensions”. In : *Discrete & Computational Geometry* 70.3 (oct. 2023), p. 1038-1058. ISSN : 0179-5376, 1432-0444. DOI : 10.1007/s00454-023-00541-w (cf. p. 52).

[214] Jörg Rambau. “TOPCOM : TRIANGULATIONS OF POINT CONFIGURATIONS AND ORIENTED MATROIDS”. In : *Mathematical Software*. Beijing, China : WORLD SCIENTIFIC, juill. 2002, p. 330-340. ISBN : 978-981-238-048-7 978-981-277-717-1. DOI : 10.1142/9789812777171_0035 (cf. p. 6, 52, 145, 179, 256).

[215] Samuel Roberts. “On the Figures Formed by the Intercepts of a System of Straight Lines in a, Plane, and on Analogous Relations in Space of Three Dimensions”. In : *Proceedings of the London Mathematical Society* s1-19.1 (nov. 1887), p. 405-422. ISSN : 00246115. DOI : 10.1112/plms/s1-19.1.405 (cf. p. 5, 49, 74, 79, 145).

[216] Stephen M. Robinson. “Generalized Equations and Their Solutions, Part II. Applications to Nonlinear Programming”. In : *Mathematical Programming Study* (1982), p. 200-221. ISSN : 0303-3929. DOI : 10.1287/88f6b0d7-91b7-4653-a41a-102ab53cf8e3 (cf. p. 3, 4, 15).

[217] Stephen M. Robinson. “Generalized Equations and Their Solutions. Part I. Basic Theory.” In : *Mathematical Programming Study* 10 (1979), p. 128-141. ISSN : 0303-3929. DOI : 10.1007/bfb0120850 (cf. p. 3, 4, 15).

[218] Stephen M. Robinson. “Local Structure of Feasible Sets in Nonlinear Programming, Part III : Stability and Sensitivity”. In : *Mathematical Programming Study* (1987), p. 45-66 (cf. p. 23, 58).

[219] Stephen M. Robinson. “Normal Maps Induced by Linear Transformations”. In : *Mathematics of Operations Research* 17.3 (août 1992), p. 691-714. ISSN : 0364-765X, 1526-5471. DOI : 10.1287/moor.17.3.691 (cf. p. 15).

[220] Stephen M. Robinson. "Strongly Regular Generalized Equations". In : *Mathematics of Operations Research* 5.1 (fév. 1980), p. 43-62. issn : 0364-765X, 1526-5471. doi : 10.1287/moor.5.1.43 (cf. p. 3, 15, 30, 225).

[221] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton 28. Princeton, NJ : Princeton University Press, 1970 (cf. p. 15, 59, 73, 154).

[222] Siegfried M. Rump. "On P-matrices". In : *Linear Algebra and its Applications* 363 (avr. 2003), p. 237-250. doi : 10.1016/S0024-3795(01)00590-0 (cf. p. 13).

[223] Sadra Sadraddini et Russ Tedrake. *Linear Encodings for Polytope Containment Problems*. Mars 2019. arXiv : 1903.05214 [math] (cf. p. 220, 221, 273-276, 278).

[224] Romesh Saigal. *Linear Programming - A Modern Integrated Analysis*. 48. Juin 1995 (cf. p. 183, 184).

[225] Hans Samelson, Robert M. Thrall et Oscar Wesler. "A Partition Theorem for Euclidean N-Space". In : *Proceedings of the American Mathematical Society* 9.5 (oct. 1958), p. 805. issn : 00029939. doi : 10.2307/2033091. JSTOR : 2033091 (cf. p. 12, 13).

[226] Holger Scheel et Stefan Scholtes. "Mathematical Programs with Complementarity Constraints : Stationarity, Optimality, and Sensitivity". In : *Mathematics of Operations Research* 25.1 (fév. 2000), p. 1-22. issn : 0364-765X, 1526-5471. doi : 10.1287/moor.25.1.1.15213 (cf. p. 16).

[227] Ludwig Schläfli. *Gesammelte mathematische Abhandlugen*. Springer, Basel : Birkhäuser, 1950 (cf. p. 5, 49, 81, 145, 167).

[228] Jürgen Schmidhuber. "Deep Learning in Neural Networks : An Overview". In : *Neural Networks* 61 (jan. 2015), p. 85-117. issn : 08936080. doi : 10.1016/j.neunet.2014.09.003 (cf. p. 55, 72).

[229] Paul D. Seymour. "A Note on Hyperplane Generation". In : *Journal of Combinatorial Theory, Series B* 61 (1994), p. 88-91. doi : 10.1006/jctb.1994.1033 (cf. p. 51).

[230] Alexander Shapiro. "On Concepts of Directional Differentiability". In : *Journal of Optimization Theory and Applications* 66.3 (sept. 1990), p. 477-487. issn : 0022-3239, 1573-2878. doi : 10.1007/BF00940933 (cf. p. 23, 29, 231).

[231] Nora Helena Sleumer. "Hyperplane Arrangements : Construction, Visualization and Applications". Thèse de doct. Zurich, Switzerland : Swiss Federal Institute of Technology, 2000 (cf. p. 53, 75).

[232] Nora Helena Sleumer. "Output-Sensitive Cell Enumeration in Hyperplane Arrangements". In : *Algorithm Theory — SWAT'98*. Sous la dir. de Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen, Stefan Arnborg et Lars Ivansson. T. 1432. Berlin, Heidelberg : Springer Berlin Heidelberg, 1998, p. 300-309. isbn : 978-3-540-64682-2 978-3-540-69106-8. doi : 10.1007/BFb0054377 (cf. p. 6, 53, 75, 77, 87, 145, 154).

[233] Marek J. Śmietański. "On a New Exponential Iterative Method for Solving Nonsmooth Equations". In : *Numerical Linear Algebra with Applications* 26.5 (oct. 2019), 1-8 (?) issn : 1070-5325, 1099-1506. doi : 10.1002/nla.2255 (cf. p. 28, 59).

[234] Mikhail V. Solodov et Benar Fux Svaiter. “A New Projection Method for Variational Inequality Problems”. In : *SIAM Journal on Control and Optimization* 37.3 (jan. 1999), p. 765-776. issn : 0363-0129. doi : 10.1137/S0363012997317475 (cf. p. 28).

[235] Mikhail V. Solodov et Benar Fux Svaiter. “A Truly Globally Convergent Newton-Type Method for the Monotone Nonlinear Complementarity Problem”. In : *SIAM Journal on Optimization* 10.2 (jan. 2000), p. 605-625. issn : 1052-6234. doi : 10.1137/S1052623498337546 (cf. p. 28).

[236] Richard P. Stanley. “An Introduction to Hyperplane Arrangements”. In : *Geometric Combinatorics*. 1re éd. T. 13. IAS/Park City Math. Ser. Providence, Rhode Island : Amer. Math. Soc., 2007, p. 389-496. isbn : 978-0-8218-3736-8 0-8218-3736-2 (cf. p. 6, 49, 75, 79, 145).

[237] Richard P. Stanley. *Enumerative Combinatorics*. second. T. 1. Cambridge Studies in Advanced Mathematics. Cambridge, UK : Cambridge University Press, 2012 (cf. p. 6, 49, 150, 169).

[238] Richard P. Stanley. *Enumerative Combinatorics*. second. T. 2. Cambridge Studies in Advanced Mathematics. Cambridge, UK : Cambridge University Press, 2024. isbn : 978-1-009-26249-1 978-1-009-26248-4 (cf. p. 6, 49, 145).

[239] Jakob Steiner. “Einige Gesetze über die Theilung der Ebene und des Raumes.” In : *J. Reine Angew. Math* (1826), p. 349-364 (cf. p. 5, 49, 74, 79, 145).

[240] Pudukkottai K. Subramanian. “A Dual Exact Penalty Formulation for the Linear Complementarity Problem”. In : *Journal of Optimization Theory and Applications* 58.3 (sept. 1988), p. 525-538. issn : 0022-3239, 1573-2878. doi : 10.1007/BF00939395 (cf. p. 14).

[241] Pudukkottai K. Subramanian. “Gauss-Newton Methods for the Complementarity Problem”. In : *Journal of Optimization Theory and Applications* 77.3 (juin 1993), p. 467-482. issn : 0022-3239, 1573-2878. doi : 10.1007/BF00940445 (cf. p. 18).

[242] Defeng Sun, Masao Fukushima et Liqun Qi. “A Computable Generalized Hessian of the D-Gap Function and Newton-Type Methods for Variational Inequality Problems”. In : *Complementarity and variational problems*. International Conference on Complementarity Problems (1997), p. 452-473 (cf. p. 41).

[243] Defeng Sun et Liqun Qi. “On NCP-functions”. In : *Computational Optimization and Applications* 13 (1999), p. 201-220. doi : 10.1023/A:1008669226453 (cf. p. 13, 15, 29, 38, 44).

[244] Panjie Tian, Zhensheng Yu et Yue Yuan. “A Smoothing Levenberg-Marquardt Algorithm for Linear Weighted Complementarity Problem”. In : *AIMS Mathematics* 8.4 (2023), p. 9862-9876. issn : 2473-6988. doi : 10.3934/math.2023498 (cf. p. 45).

[245] Paul Tseng. “Co-NP-completeness of Some Matrix Classification Problems”. In : *Mathematical Programming* 88.1 (juin 2000), p. 183-192. issn : 0025-5610, 1436-4646. doi : 10.1007/s101070000159 (cf. p. 13).

[246] Paul Tseng, Nobuo Yamashita et Masao Fukushima. “EQUIVALENCE OF COMPLEMENTARITY PROBLEMS TO DIFFERENTIABLE MINIMIZATION : A UNIFIED APPROACH”. In : *SIAM Journal of Optimization* 6.2 (mai 1996), p. 446-460 (cf. p. 40).

[247] Michael Ulbrich. *Nonsmooth Newton-like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. SIAM Publications. MPS-SIAM Series on Optimization 11. Philadelphia, PA, USA : Society for Industrial and Applied Mathematics (SIAM), 2012. isbn : 978-1-61197-068-5 (cf. p. 3).

[248] Duc Thach Son Vu. "Numerical Resolution of Algebraic Systems with Complementarity Conditions. Application to the Thermodynamics of Compositional Multiphase Mixtures". Thèse de doct. Université Paris-Saclay, 2020 (cf. p. 44).

[249] Duc Thach Son Vu, Ibtihel Ben Gharbia, Mounir Haddou et Quang Huy Tran. "A New Approach for Solving Nonlinear Algebraic Systems with Complementarity Conditions. Application to Compositional Multiphase Equilibrium Problems". In : *Mathematics and Computers in Simulation* 190 (déc. 2021), p. 1243-1274. issn : 03784754. doi : 10.1016/j.matcom.2021.07.015 (cf. p. 2, 4, 44).

[250] Dominic J. A. Welsh. *Complexity : Knots, Colourings and Counting*. T. 186. University of Oxford : Cambridge University Press, août 1993. isbn : 978-0-521-45740-8 (cf. p. 52).

[251] Walter Wenzel, Nihat Ay et Frank Pasemann. "Hyperplane Arrangements Separating Arbitrary Vertex Classes in N-Cubes". In : *Advances in Applied Mathematics* 25.3 (oct. 2000), p. 284-306. issn : 01968858. doi : 10.1006/aama.2000.0701 (cf. p. 55, 72, 98).

[252] Andrzej P. Wierzbicki. "Note on the Equivalence of Kuhn-Tucker Complementarity Conditions to an Equation". In : *Journal of Optimization Theory and Applications* 37.3 (juill. 1982), p. 401-405. issn : 0022-3239, 1573-2878. doi : 10.1007/BF00935279 (cf. p. 18).

[253] Robert Owen Winder. "Partitions of N -Space by Hyperplanes". In : *SIAM Journal on Applied Mathematics* 14.4 (juill. 1966), p. 811-818. issn : 0036-1399, 1095-712X. doi : 10.1137/0114068 (cf. p. 50, 55, 63, 79-82, 145, 166, 167).

[254] Stephen J. Wright. *Primal-Dual Interior-Point Methods*. Philadelphia, PA, USA : Society for Industrial and Applied Mathematics (SIAM), 1997. isbn : 0-89871-382-X (cf. p. 16).

[255] Shuhuang Xiang et Xiaojun Chen. "Computation of Generalized Differentials in Nonlinear Complementarity Problems". In : *Computational Optimization and Applications* 50.2 (oct. 2011), p. 403-423. issn : 0926-6003, 1573-2894. doi : 10.1007/s10589-010-9349-z (cf. p. 31, 61, 63, 77, 85, 86, 122, 147).

[256] Nobuo Yamashita et Masao Fukushima. "On Stationary Points of the Implicit Lagrangian for Nonlinear Complementarity Problems". In : *Journal of Optimization Theory and Applications* 84.3 (mars 1995), p. 653-663. issn : 0022-3239, 1573-2878. doi : 10.1007/BF02191990 (cf. p. 39).

[257] Thomas Zaslavsky. "Facing up to Arrangements : Face-Count Formulas for Partitions of Space by Hyperplanes". In : *Memoirs of the American Mathematical Society* 1.154 (1975), 1-109 (?) (Cf. p. 6, 49, 71, 79, 132, 145, 150, 166, 169).

[258] Chao Zhang, Xiaojun Chen et Naihua Xiu. "Global Error Bounds for the Extended Vertical LCP". In : *Computational Optimization and Applications* 42.3 (avr. 2009), p. 335-352. issn : 0926-6003, 1573-2894. doi : 10.1007/s10589-007-9134-9 (cf. p. 59).

[259] Ju-liang Zhang et Jian Chen. "A Smoothing Levenberg-Marquardt Type Method for LCP". In : *Journal of Computational Mathematics* (2004), p. 735-752 (cf. p. 45).

[260] Ju-liang Zhang et Xiangsun Zhang. “A Smoothing Levenberg–Marquardt Method for NCP”. In : *Applied Mathematics and Computation* 178.2 (juill. 2006), p. 212-228. ISSN : 00963003. DOI : 10.1016/j.amc.2005.11.036 (cf. p. 4, 45).

[261] Shuzi Zhou et Zhanyong Zou. “A New Iterative Method for Discrete HJB Equations”. In : *Numerische Mathematik* 111.1 (nov. 2008), p. 159-167. ISSN : 0029-599X, 0945-3245. DOI : 10.1007/s00211-008-0166-6 (cf. p. 2).

[262] Günter M. Ziegler. *Lectures on 0/1-Polytopes*. Sept. 1999. arXiv : math/9909177 (cf. p. 286).

[263] Günter M. Ziegler. *Lectures on Polytopes*. 7th. T. 152. Graduate Texts in Mathematics. New York, NY : Springer New York, 2007. ISBN : 978-0-387-94365-7 978-1-4613-8431-1. DOI : 10.1007/978-1-4613-8431-1 (cf. p. 50, 54, 155, 220, 267).

[264] Günter M. Ziegler, Laura Anderson et Kolja Knauer. “Oriented Matroids Today”. In : *The Electronic Journal of Combinatorics* 1000 (2024), p. 73. ISSN : 1077-8926. DOI : 10.37236/25 (cf. p. 50).