



DOCTORAL DISSERTATION, INTERNATIONAL CO-SUPERVISION

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)

---

**A Robust Linearization Method for  
Complementarity Problems**

**A Detour Through Hyperplane Arrangements**

---

Defended by: **BAPTISTE PLAQUEVENT-JOURDAIN**

For the grade of Philosophiae Doctor (Ph.D.) of MATHEMATICS

Supervised by **JEAN-PIERRE DUSSAULT** and **JEAN CHARLES GILBERT**

Presented and publicly defended the 07/16/2025 at Paris, France

With the jury presided by Mounir Haddou and composed by:

Reviewer	MOUNIR HADDOU, PROFESSEUR DES UNIVERSITÉS, INSA Rennes
Reviewer	MIROSLAV RADA, ASSISTANT PROFESSOR, Prague University of Economics and Business
Examiner	IBTIHEL BEN GHARBIA, DR, IFPEN, Rueil-Malmaison
Examiner	VIRGINIE CHARETTE, FULL PROFESSOR, Université de Sherbrooke
Supervisor	JEAN-PIERRE DUSSAULT, FULL PROFESSOR, Université de Sherbrooke
Supervisor	JEAN CHARLES GILBERT, DR, Inria Paris



# Résumé

**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.

---

**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.





# Acknowledgements

*It is done.* The long journey of this PhD comes to its end. Before delving into its topic, I would like to express my gratitude to all the people who made it possible.

First, to Jean-Pierre and Jean Charles, thank you for allowing me to progress on this path during four (and a bit more) years. Thanks to you, I was able to think and work on what interested me a lot, from nonsmooth optimization to computational geometry including combinatorics. Beyond leaving me to explore some curious things on my own, we did work together, discovered the fascinating topic of arrangements and dove into some nonsmooth analysis questions. At the (very) end of the PhD, some proofs link arrangements and complementarity problems in a few places: despite the work being not finished yet, I do believe we have unveiled some mysteries of the minimum function and I hope we can further progress on that topic. Your devotion to research and the quest for knowledge, your passion, your meticulousness, and many others, shall remain a goal to me. Once again, thank you.

The organization of this international PhD required quite some help and I would like to thank those who made it easier: Patricia Zizzo, Josée Lamoureux, Annie Carboneau, Thomas Brüstle, Anne MacKay, Félix Camirand Lemyre, Derya Gök, and many others. I would also like to thank the MITACS and ISM institutes from Canada for their financial support.

To my family, I will never thank you enough. You always helped and supported me, always asked about what was going on in the PhD and how it was progressing, even if the topic was not necessarily the best for it. You are wonderful, I realize the luck I have, and I hope to be able to share more with you in the future.

Thanks to Charles for the advice and support. To the many students from the BISOUS, Luc, Tania, Vivien, Nicolas, Arthur, Josué, Christopher and the (many) interns, thanks for these summers together. I really enjoyed the time with you, the many shared meals especially during Cakesdays, the fascinating conversations and activities. It was necessary between two housing shenanigans (cockroach poison, zoo, dementia...). I wish you the best for what's next. Thank you to Sylvain, and especially Guillaume for the (*very* many) mornings and quick one-to-one discussions – unfortunate I was the only other early bird at the BISOUS! To SERENA, thank you for welcoming us during the second part of my PhD. I know I haven't made the most intense effort to interact, but I hope you appreciated the Cakesdays nonetheless.

---

I would also like to thank, though for a rather specific role and despite the fact we never seen each other, the community members of “yellowhat” and “NumottheNummy”, for these endless discussions, debates, Caves, trivia and this timeless entertainment, all thanks to “Godfield”.

Finally, to friends, thank you for bearing with me, for the support, and for everything. Sincere apologies to people not cited. A particular thank for the Ariane rocket people: Aurore (and in the meantime to the legend of Amogus), Basile, Servane, Ronan, Aliénor, Guillaume, Pauline, Juliette, Julieng, Léopold(ieu), Benoît B, Benoît S, for the moral support, the help and the advice. Elias, for the help in some moments. Louis, Agnès, Yousra, Adrien, for our precious conversations. Étienne et Hugo, for the (clearly too few) numerous board games sessions at House Riquet. For the *gaming*, with Quentin, Lucas and Florian. Also with Olivier and Zoé, but also for wonderful activities despite the sun, the clean air and the outside; Alice, for all we have shared and still share now – you’ll admit enough words have been written... To Tom and Jérémie, thank you for these discussions and moments.





# Contents

<b>Résumé</b>	<b>i</b>
<b>Acknowledgements</b>	<b>v</b>
<b>List of figures</b>	<b>xvii</b>
<b>List of tables</b>	<b>xxi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Starting point and motivations . . . . .	1
1.2 Related problems . . . . .	2
1.3 Equivalent formulations and algorithms . . . . .	3
1.4 Nonsmoothness and combinatorial geometry . . . . .	4
1.5 Outline and contributions . . . . .	6
<b>2 General setting</b>	<b>9</b>
2.1 Notation . . . . .	10
2.1.1 General notation . . . . .	10
2.1.2 Specific notation . . . . .	12
2.2 Reformulations of complementarity problems . . . . .	12
2.2.1 Some types of LCPs . . . . .	12
2.2.2 Some types of NCP(F) . . . . .	13
2.2.3 General complexity . . . . .	14
2.2.4 Generalized equations and normal maps . . . . .	14
2.2.5 Interior-points . . . . .	15
2.2.6 Absolute value equation . . . . .	16
2.2.7 Problems with complementarity in the constraints . . . . .	16
2.3 Nonsmooth setting and algorithms . . . . .	17
2.3.1 Introduction of C-functions . . . . .	17
2.3.2 Some tools of nonsmooth analysis . . . . .	19
2.3.3 First nonsmooth algorithms . . . . .	24
2.3.4 Particular treatment of the minimum . . . . .	31
2.3.5 Other nonsmooth methods . . . . .	34
2.3.6 Smoothing techniques . . . . .	41
2.3.7 A comment about complexity . . . . .	44
2.4 On the combinatorial aspect . . . . .	46
2.4.1 Relation with the previous topics . . . . .	46

---

2.4.2	Classic references . . . . .	47
2.4.3	Some specific tools . . . . .	48
2.4.4	Oriented Matroids (and circuits) . . . . .	48
2.4.5	Algebra software . . . . .	50
2.4.6	Specific algorithms to identify the chambers . . . . .	51
2.4.7	A few examples of applications . . . . .	52
<b>3</b>	<b>B-differential of the minimum of two vectorial affine functions</b>	<b>55</b>
3.1	Introduction . . . . .	56
3.2	Background . . . . .	59
3.3	Equivalent problems . . . . .	61
3.3.1	B-differential of the minimum of two affine functions . . . . .	62
3.3.2	Linear algebra problems . . . . .	62
3.3.3	Convex analysis problems . . . . .	68
3.3.4	Discrete geometry: hyperplane arrangements . . . . .	73
3.4	Description of the B-differential . . . . .	74
3.4.1	Some properties of the B-differential . . . . .	75
3.4.2	Cardinality of the B-differential . . . . .	77
3.4.3	Particular configurations . . . . .	82
3.4.4	A glance at the C-differential . . . . .	83
3.5	Computation of the B-differential . . . . .	84
3.5.1	Computation of a single Jacobian . . . . .	84
3.5.2	Computation of all the Jacobians . . . . .	85
3.6	Discussion . . . . .	103
	Acknowledgments . . . . .	104
	Statements & Declarations . . . . .	104
<b>4</b>	<b>Additional elements on the B-differential of the minimum and hyperplane arrangements</b>	<b>105</b>
4.1	Additional material from the previous chapter . . . . .	106
4.2	Regularity notions and counterexamples . . . . .	108
4.3	B-differential of the minimum of nonlinear F and G . . . . .	109
4.3.1	Differentials of H . . . . .	109
4.4	Differential of the merit function . . . . .	118
4.5	Details on instances and algorithms . . . . .	128
4.5.1	About the permutohedron instances . . . . .	128
4.5.2	About the crosspolytope separability arrangement . . . . .	133
4.5.3	Perfectly symmetric instances . . . . .	137
<b>5</b>	<b>Primal and dual approaches for the chamber enumeration of hyperplane arrangements</b>	<b>141</b>
5.1	Introduction . . . . .	142
5.2	Background . . . . .	144
5.3	Hyperplane arrangements . . . . .	145
5.3.1	Presentation . . . . .	145
5.3.2	Properties . . . . .	148
5.3.3	Stem vectors . . . . .	153

---

5.3.4	Augmented matrix . . . . .	159
5.4	Chamber computation - Primal approaches . . . . .	168
5.4.1	Primal $\mathcal{S}$ -tree algorithm . . . . .	169
5.4.2	Preventing some computations . . . . .	174
5.5	Chamber computation - Dual approaches . . . . .	176
5.5.1	Algorithms using all the stem vectors . . . . .	177
5.5.2	Algorithms using some stem vectors . . . . .	180
5.6	Compact version of the algorithms . . . . .	184
5.6.1	The compact $\mathcal{S}$ -tree . . . . .	185
5.6.2	Compact primal $\mathcal{S}$ -tree algorithm . . . . .	186
5.6.3	Compact primal-dual $\mathcal{S}$ -tree algorithm . . . . .	190
5.7	Numerical results . . . . .	192
5.7.1	Arrangement instances . . . . .	193
5.7.2	Assessed algorithms . . . . .	195
5.7.3	Numerical results . . . . .	195
5.8	Conclusion . . . . .	197
5.9	Appendix: tables with numerical results . . . . .	199
<b>6</b>	<b>Levenberg-Marquardt least-squares globalization of the PNM algorithm</b>	<b>203</b>
6.1	Modifying the Polyhedral Newton-Min algorithm . . . . .	204
6.1.1	Presentation of the method . . . . .	204
6.1.2	Levenberg-Marquardt least-squares variant . . . . .	206
6.1.3	Choice of the weights and stationarity . . . . .	216
6.1.4	Choice of the weights and differentials . . . . .	220
6.1.5	Discussion of regularity of solutions . . . . .	222
6.2	A considered algorithm . . . . .	225
6.2.1	The method and its properties . . . . .	225
6.2.2	Modifications and potential improvements . . . . .	230
<b>Conclusion</b>		<b>235</b>
<b>A</b>	<b>Detailed information on affine hyperplane arrangements: theory, numerics and complements</b>	<b>237</b>
A.1	Details on properties of chapter 5 . . . . .	237
A.2	Instance values . . . . .	240
A.3	Algorithmic behaviors . . . . .	242
A.3.1	Primal heuristics . . . . .	242
A.3.2	Dual heuristics . . . . .	244
A.3.3	Analysis of the compact algorithm . . . . .	246
A.4	Linear instances and other topics . . . . .	250
A.4.1	Circuits computation . . . . .	252
A.4.2	Recursive covering test . . . . .	254
<b>B</b>	<b>Geometric elements on polytopes</b>	<b>261</b>
B.1	Polytopes and their face(t)s . . . . .	261
B.2	Specific properties of zonotopes . . . . .	265

---

<b>C Inclusion of zonotopes</b>	<b>271</b>
<b>D Weights and element of Clarke's differential</b>	<b>279</b>
D.1 A geometric property on sign vector sets . . . . .	279
D.2 Main proof . . . . .	281
D.2.1 Detailed (simpler) counterexamples . . . . .	288
D.2.2 Degeneracies and (theoretical) corrections . . . . .	298
<b>Bibliography</b>	<b>307</b>

# List of Figures

1.1	Illustration with three hyperplanes and 7 chambers, where the hyperplanes are $H_1 = \{x \in \mathbb{R}^2 : x_1 = 0\}$ , $H_2 = \{x \in \mathbb{R}^2 : x_2 = 0\}$ and $H_3 = \{x \in \mathbb{R}^2 : x_1 + x_2 = 1\}$ . . . . .	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 $s_i v_i^\top d > 0$ for all $i \in [1 : 3]$ . Each conic hull of these vectors, namely $\text{cone}\{s_i v_i : i \in [1 : 3]\}$ , 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 $\mathcal{S}$ . . . . .	70
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 $(\emptyset, [1 : p])$ or $([1 : p], \emptyset)$ , so that $ \mathcal{S}  = 2(n_s + 1)$ , where $n_s$ is the number of represented separating lines. We have set $r := \dim(\text{vect}\{\bar{v}_1, \dots, \bar{v}_p\}) + 1$ . . . . .	70
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 $v_i^\top d > 0$ and $s_i = -$ if $v_i^\top d < 0$ . 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). . . . .	73
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 descendance; the sign sets $\mathcal{S}_k^1$ are defined by (3.42). . . . .	86
4.2	Illustration of the idea of the induction process: in purple the two hyperplanes added. The top black dot represents a point with $d_0 = 0$ and arbitrary $d_{[1:n]}$ which has 4 descendants. The bottom black dot represents a point with $d_0 \neq 0$ and $d_{[1:n]} = 0$ which has only 3 descendants (two arrow plus itself). . . . .	136

---

5.1	Arrangements in $\mathbb{R}^2$ specified by the hyperplanes $H_1 := \{x \in \mathbb{R}^2 : x_1 = 0\}$ , $H_2 := \{x \in \mathbb{R}^2 : x_2 = 0\}$ , $H_3(\text{left}) := \{x \in \mathbb{R}^2 : x_1 + x_2 = 0\}$ , $H_3(\text{middle}) := \{x \in \mathbb{R}^2 : x_1 + x_2 = 1\}$ and $H_3(\text{right}) := \{x \in \mathbb{R}^2 : x_1 + x_2 = -1\}$ . The origin is contained in all the hyperplanes but in $H_3(\text{middle})$ and $H_3(\text{right})$ , so that the arrangement in the left-hand side is <i>linear</i> with 6 chambers and the other ones are <i>affine</i> with 7 chambers. . . . .	146
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 $\mathfrak{S}([V; \tau^\top], 0)$ , 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 $\sigma$ and its opposite $-\sigma$ : $\mathfrak{S}_s(V, \tau)$ , $\mathfrak{S}(V, 0)$ , $\mathfrak{S}_0(V, \tau)$ and $\mathfrak{S}([V; \tau^\top], 0)$ 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 $-\mathfrak{S}([V; \tau^\top], 0) = \mathfrak{S}([V; \tau^\top], 0)$ . By propositions 5.3.15 and 5.3.21, the diagram simplifies when $\tau \in \mathcal{R}(V^\top)$ , since then $\mathfrak{S}_a(V, \tau) = \mathfrak{S}_a(V, -\tau) = \mathfrak{S}_0(V, \tau) = \emptyset$ and there is only one set left. . . . .	155
5.3	Symbolic representation of the sets $\mathcal{S}(V, 0)$ , $\mathcal{S}(V, \tau)$ , $\mathcal{S}_a(V, \tau)$ and $\mathcal{S}([V; \tau^\top], 0)$ , 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 $-s$ : $\mathcal{S}(V, 0)$ , $\mathcal{S}([V; \tau^\top], 0)$ and $\mathcal{S}([V; \tau^\top], 0)^c$ are symmetric in the sense of definition 5.3.4. . . . .	160
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 $s \cdot (V^\top x - \tau) > 0$ ). . . . .	169
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 $-\mathcal{S}_a(V, -\tau) = \mathcal{S}_a(V, \tau) \subseteq \mathcal{S}(V, \tau)$ . . . . .	186
5.6	Performance profiles of the RC, P, PD and D algorithms, for the computing time. . . . .	197
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. . . . .	198
5.8	Performance profiles of the RC vs PD/C solvers, for the computing time. . . . .	198
6.1	Left: level sets of $\varphi_x$ with the choice of $\gamma = (1/2, 1/2)$ . Right: level sets of $\theta$ ; the dotted lines are the kinks ( $\theta$ is not smooth). The red dot indicates a local minimum. The level sets reveal too much difference between $\theta$ and $\varphi_x$ , so the direction given by $\varphi_x$ increases $\theta$ . . . . .	210
6.2	Left: level sets of $\varphi_x$ with the choice of $\gamma = (2/3, 1)$ . Right: level sets of $\theta$ ; the dotted lines are the kinks ( $\theta$ is not smooth). The red dot indicates a local minimum. The level sets of $\varphi_x$ are (at least locally) close enough to those of $\theta$ so a descent direction of $\varphi_x$ decreases $\theta$ . . . . .	211

---

6.3	Illustration of example 6.1.4. The level curves of $\theta$ ( $\sqrt{\theta}$ for visibility) are drawn in color, the blue dashed lines are the kinks of nondifferentiability of $H$ (they are defined by $x_i = (Px + q)_i$ for $i \in \{1, 2\}$ ), the red point above is the unique solution $\bar{x} = (0, 1/10)$ to the problem and the blue point is the current point. The arrows in green, blue, red and black correspond to four possible $-g$ for extremal choices of $\gamma$ , the one in magenta to a descent direction. . . . .	211
6.4	Illustration of nonDini stationarity. One has $F(x) = x$ , $G(x) = 1 + (x - 1)^2$ , so the problem has a solution at $x = 0$ . At $x = 1$ , neither $H$ nor $\theta$ are differentiable. Since $G'(1) = 0$ , by taking a sequence $1 + t_k \rightarrow 1$ , one gets that $0 \in \partial\theta(1)$ , but $x = 1$ is clearly not strongly stationary. Such point is sometimes called a “concave kink”, which, as we shall see, may cause some difficulties to algorithms. . . . .	221
6.5	The curves above $\theta$ are the quadratic models $\varphi_{x_k}$ . While there may be fast convergence to $x = 1$ , it may never be reached ( $x_k > 1$ ) so $\forall k, \mathcal{E}(x_k) = \emptyset$ . . . . .	231
6.6	Illustration of a few iterates, for some $\tau > 0$ , of algorithm 6.2.1 using $\mathcal{E}(x) := \{i \in [1 : n] :  F_i(x) - G_i(x)  < \tau\}$ . . . . .	233
A.1	The black lines represent the two hyperplanes already considered, $x$ a point of the current region. It is simple to add first the blue hyperplanes, that lead to only one child, then add the dotted hyperplanes rather than doing the opposite. While the figure is shown for a central arrangement, the principle remains the same for an affine arrangement. . . . .	243
A.2	Illustration of (A.4) (and (A.3)). The correction, i.e., the difference between (A.3) and (A.4), is denoted by (*) and is added to the bottom pictures; for the instances on the top-right of the right graphs, it brings the points on the line corresponding to the formula, given by $y = 1 + x/2$ . The (c) denotes the number of LOPs of the compact variant. for the pictures on the right, four points, corresponding to the PERM instances, are shifted to the right. A possible explanation is proposed. . . . .	252
B.1	Illustration of lemmas B.1.1 (left) and B.1.4 (right). On the left, we see that the relative interior is obtained by removing the parts of $P$ where the equalities $A_{:,i}x = a_i$ hold. However, consider the same polytope in dimension 3 (thus with empty interior), defined by the additional inequalities $e_3^T x \leq 0$ and $-e_3^T x \leq 0$ , one cannot take the strict inequalities since it would result in an empty set. This is because these two inequalities actually form an equality ( $e_3^T x = 0$ ). On the right, we see that the relative interior in blue corresponds to the relative interior of $P$ (seen as a face of itself), whereas the boundary is composed of the relative interiors of the edges in magenta, then the vertices in red. . . . .	263
B.2	Illustration of lemmas B.1.3 (left) and B.1.5 (right). On the left, one can observe that the green face corresponds to a set $I = \{4\}$ of size 1, whereas the vertex in red corresponds to a set $I = \{1, 2\}$ of size 2. On the right, the same green face has a unique (up to multiplicative constant) normal $c$ since the face is of maximal dimension $n - 1$ whereas the vertex in red has multiple noncolinear possible normals $c$ (see remark B.1.6). . . . .	264

---

B.3	Example of a simple zonotope with $V = [e_1 \ e_2 \ e_1 + e_2 \ e_3]$ . The upper face in orange is a face of dimension two, generated by $e_1, e_2, e_1 + e_2$ , with $I^* = \{4\}$ . The face in purple at the front is generated by $e_2$ and $e_3$ , with $I^* = \{1, 3\}$ . The face in green on the right, which is an edge, is generated by $e_3$ with $I^* = \{1, 2, 3\}$ . All the vertices are also faces with no generators. On the right, some hyperplanes corresponding to normals to faces were added. The full dimensional faces have only one hyperplane but the edges have multiple (since the dimension is 3). . . . .	266
B.4	Left: normal fan of a zonotope. Light green: vertices and their normal cones (dashed boundaries since the boundaries are the normals of the edges); purple: edges and their normal cones. Right: illustration of proposition B.2.3. Green sets: some of the faces considered; green arrows: their normals. Red points: the centers of the faces (equal to the faces for the vertices). Grey: vectors generating the zonotope. Black: the dotted arrows are the $V_{:,i} \kappa_i$ for $i \in I^*$ . We translated to the centers of the faces to show more easily the normals $c$ in green from proposition B.1.7 make a positive scalar product with the black dotted arrows. . . . .	267
B.5	Example of a point such that the direction point – projection does not verify strict complementarity. This occurs at points where the projection is not differentiable. (The distance itself is differentiable outside of the boundary of the (closed) convex). . . . .	270
C.1	In this particular example, the (unique) solution $(\Delta, \beta)$ of the problem detailed in example C.0.5 yields $\lambda^* = 6$ . We observe that when dilating $Z_y$ by $\lambda^*$ , one has $Z_x \subseteq \bar{y} + \lambda^* Y[-1, +1]^2$ , but dilating by any $\lambda < \lambda^*$ , the inclusion does not hold (the top point of the green area is not contained in the light purple area). . . . .	273
D.1	Left: $[X \ Y][-1, +1]^4$ , vertices in blue and other points in black (each dot is two sign vectors). Right: $[X \ -Y][-1, +1]^4$ , vertices in blue and other points in black (each dot is two sign vectors). Schematically, the light blue corresponds to the zonotopes with $[X \ Y]$ and $[X \ -Y]$ and the black to $\partial\theta(x)$ . . . . .	285
D.2	Illustration of the zonotope aspect for a situation encountering multiple difficulties. On the left, the teal zonotope on top left is $Z_x$ while $Z_y$ is represented in the bottom right in magenta. The light purple represents the dilated version (by $\lambda^*$ ) of $Z_y$ . On the right, the blue zonotope is the representation of $\partial\theta(x)$ , see (D.8), up to the three other components equal to zero (thus not shown). First, observe on the right picture that among the eight sign vectors corresponding to $\partial_B H(x)$ , only four of them form the convex hull of the C-differential once multiplied by $H$ . Moreover, the “neighbors” in the picture do not correspond to neighboring sign vectors. Finally, as described in a simpler example later, the method from appendix C returns a value of $\mathcal{E}^{0+}(x)$ corresponding to the leftmost point in the teal area (the dilation of the bottom point on the boundary of $Z_y$ , with $\bar{\zeta}$ ) that corresponds to $g$ (the projection is the top point with $\zeta^*$ ) that is the red point in the right picture and is thus outside $\partial\theta(x)$ ; this comes from the fact that the chosen signs are not in $\mathcal{S}(V, 0)$ . . . . .	288

---

D.3	Illustration of the counterexample. Left: corresponding zonotopes (teal for $Z_x$ , magenta for $Z_y$ ), the arrow represents $g$ for $\eta = -1$ which does not belong to $\partial\theta(x)$ . Right: illustration of $\partial\theta(x)$ and the elements in $\partial_B H(x)^\top H(x)$ ; $g_1$ is the center of the differential. . . . .	290
D.4	Illustration of the counterexample. Left: corresponding zonotopes (teal for $Z_x$ , magenta for $Z_y$ ), the arrows represent $g$ for $\eta = -1$ which does not belong to $\partial\theta(x)$ and $\eta = +1$ which does. Right: illustration of $\partial\theta(x)$ (in the plane $x_2 = 4$ ). We observe, as seen in counterexample D.2.3, that not all sign vectors are extremal $((+, -, +, +)$ and $(-, +, +, -)$ both correspond to the middle blue dot) and that, depending on the value of $\eta$ , $g$ may or may not belong to $\partial\theta(x)$ . The remaining 6 sign vectors correspond to the face in the plane $x_2 = 6$ . . . . .	294
D.5	Illustration of the degeneracies, mostly obtained in counterexamples by artificially adding a dimension (the third dimension is omitted in the right picture, it is in the plane $x_1 = -1$ ). On the left, the zonotope aspect described by the previous equations. On the right, the illustration of the gradients $g$ obtained in this situation: the blue square represents the differential of $\theta$ ; in particular, the vertices correspond to sign vectors that are not “adjacent”. Here, it is due to the fact the vectors of $X$ and $Y$ are colinear, but the absence of adjacency can also occur without this particular case. The orange segment represents the possible $g$ obtained by the choices of $\mathcal{E}^{0+}(x)$ , the red points its vertices. In particular, the vertices of the differential in blue correspond to the zonotope expressed in (D.7). . . . .	296
D.6	Illustration of the dilated quantities, with the data from counterexample D.2.3. . . . .	299
D.7	Left: $Z_y$ in magenta, $Z_x$ in teal, and the subset of $Z_x$ in green corresponding to the $\gamma_{\mathcal{E}^{0+}(x)}$ ’s ( $\eta$ ’s) such that the $\gamma_{\mathcal{E}^-(x)}$ ’s ( $\zeta$ ’s) obtained after projection yield a $g \in \partial\theta(x)$ . Right: corresponding values in $[-1, +1]^2$ (i.e., with $\eta$ ). Recall that the topmost point of the left figure corresponds to $\eta = (1, -1)$ and the leftmost point of the left figure corresponds to $\eta = (-1, -1)$ , which explains the change of orientation (to recover a shape similar to the left picture, turn by $+3\pi/4$ counterclockwise then apply axial symmetry vertically). . . . .	305
D.8	In magenta, $Z_y$ and in teal, $Z_x$ . Left: $\partial\theta(x)$ in blue, the intersection with $Z_x$ in darker green corresponds to the $\eta$ with $g(\eta, \zeta = 0) \in \partial\theta(x)$ . Right: values of $\theta'(x, -g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}))$ for the extremal $\eta$ ’s: every $g$ with $\zeta = 0$ is a descent direction. . . . .	305
D.9	In magenta, $Z_y$ and in teal, $Z_x$ . Left: $\partial\theta(x)$ in blue, the intersection with $Z_x$ in darker green corresponds to the $\eta$ with $g(\eta, \zeta = 0) \in \partial\theta(x)$ . Right: values of $\theta'(x, -g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}))$ for the extremal $\eta$ ’s: for some specific $\eta$ , $-g$ is an ascent direction. . . . .	306



# List of Tables

2.1	Summary of the regularity properties for the LCP. . . . .	30
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: “ $(n, p, r, \varsigma)$ ” are the features of the problem ( $V \in \mathbb{R}^{n \times p}$ is of rank $r$ and has $\varsigma$ 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. . . . .	98
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 $v_{k+1}^T d \simeq 0$ ), C (changing the order of the vectors $v_i$ ’s by taking $i_{k+1}$ by (3.49)), D <sub>1</sub> (pre-computation of $2(p-r)$ stem vectors after the QR factorization), D <sub>2</sub> (D <sub>1</sub> and 2 additional stem vectors computed after solving a LOP, whose optimal value is nonnegative), D <sub>3</sub> (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 <sub>4</sub> (all the stem vectors are first computed and no LOP is solved). The “Ratio” (acceleration ratio) columns give for each considered problem the ratio ( <i>LOPs of the considered ISF version</i> ) / ( <i>LOPs of simulated rc</i> ). Note (1): interruption of the run after several days of computation. The Mean/Median rows give the mean and median values of the ratios. . . . .	99
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 ( <i>Time of the considered ISF version</i> ) / ( <i>Time of simulated rc</i> ). Note (1): interruption of the run after several days of computation. The Mean/Median rows give the mean and median values of the ratios. . . . .	102
5.1	Corresponding lines in algorithms 5.6.5 and 5.6.8. . . . .	190
5.2	Cardinality formulas for some instances, when $p > n$ and $\text{rank}(V) = n$ . . . . .	193

---

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 $\mathcal{A}([V; \tau^\top], 0)$ 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).	194
5.4	Computing times (in seconds) for the <i>standard</i> algorithms listed in section 5.7.2. For each algorithm $A := P, PD$ or $D$ , the second column gives the ratios $\text{time}(RC)/\text{time}(A)$	200
5.5	Computing times (in seconds) for the <i>compact</i> 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 $\text{time}(A)/\text{time}(A/C)$ (upper bounded by 2, approximately) and the third column gives the ratios $\text{time}(RC)/\text{time}(A/C)$ .	201
A.1	Known cardinalities for some of the instances. The values for the <b>RAND-N-P</b> and <b>2D-N-P</b> problems are obtained via affine general position, thus propositions 5.3.31, 3.4.6 and remark 5.3.13 6). Recall that the symmetric stem vectors are counted in pairs (thus the factor 1/2) – the number of circuits does <i>not</i> take this factor into account.	240
A.2	Number of subproblems solved, depending on the state of the current node. Moreover, $\pm s \in \mathcal{S}_k \iff \mathbf{s} = 0$ and $(-)s \in \mathcal{S}_k \iff \mathbf{s} \neq 0$ .	247
A.3	Approximate proportions of symmetric chambers. The 2D instances have particularly low proportion of symmetric chambers.	250
A.4	Relevant values for the affine instances. Columns 2 and 3 represent the cardinalities of the sign vector sets, columns 4-5-6 the number of feasible problems solved. Column 7 is the difference of the two previous ones. The last column represents the second term in the right-hand side of (A.4). This table is illustrated below, in figure A.2. The * represent the irregularities mentioned above (no perfect general position in the subarrangement).	251
A.5	Computation times in black and ratio $\text{time}(RC) / \text{time}(A)$ in blue for the linear instances and the different algorithms; bold ratios are the best ones. For algorithm D, the stem vectors and the covering tests are computed in slightly faster ways described in sections A.4.1 and A.4.2.	253
A.6	Computation times of the stem vectors in the regular variants. The second and fourth columns represent the numbers of stem vectors, the third and fifth columns the number of duplicates. The three remaining columns indicate the time of the initial computation, the computation time with echelon form and their ratio: if over 1, it means the echelon form is faster.	255

---

A.7	Computation times of the stem vectors in the compact variants. The second and fourth columns represent the numbers of stem vectors, the third and fifth columns the duplicates. The three remaining columns indicate the time of the initial computation, the computation time with echelon form and their ratio: if over 1, it means the echelon form is faster. . . . .	256
A.8	Illustration of the recursive implementation of the covering test. There are $p = 5$ vectors in $\mathbb{R}^n$ , and the matrix of stem vectors is given on the right in transpose form in the right half. For instance the first column means $[v_1 \ v_2 \ -v_3]$ is of nullity one in $\mathbb{R}_+^{\{1,2,3\}}$ . On the left, on the line with index = 1 and sign = +, the current vector is $+M_{:,1}$ (the first line of the transposed matrix of stem vectors). On the following line, since the sign is also +, the second line of the matrices of stem vectors is added. On line with index = 3 and sign = -, the current vector is thus the first line of $\mathfrak{S}$ plus the second minus the third. In particular, coordinate 1 of the current vector equals 3, which is the size of the first stem vector (first column of $\mathfrak{S}$ ), so the covering test stops the recursion. . . . .	257
A.9	Run times for the linear instances with option D3. Columns 2-3 represent the total and covering times with the full matrix-vector product in the test. Columns 4-5 represent the total and covering times done recursively. Column 6 gives the number of stem vectors. Columns 7-8 show the ratios for the total times and the covering times. . . . .	258
A.10	Run times for the linear instances with option D4. Columns 2-3 represent the total and covering times with the full matrix-vector product in the test. Columns 4-5 represent the total and covering times done recursively. Column 6 gives the number of stem vectors. Columns 7-8 show the ratios for the total times and the covering times. . . . .	259



# Chapter 1

## Introduction

### 1.1 Starting point and motivations

Our main goal is the study of *Complementarity Problems* (CPs) under an algorithmical lens. Often studied in finite dimension, though extensions exist in infinite dimension, they consist in a number of (in)equalities and relations of complementarity that can be written in the following form. For two vector-valued functions  $F, G : \mathbb{R}^n \rightarrow \mathbb{R}^n$ , one searches for an  $x \in \mathbb{R}^n$  such that

$$F(x) \geq 0, \quad G(x) \geq 0, \quad F(x)^\top G(x) = 0, \quad (1.1)$$

where the inequalities are to be understood componentwise and  $F(x)^\top G(x)$  is the Euclidean scalar product between the vectors  $F(x)$  and  $G(x)$ . A more compact expression is given by

$$0 \leq F(x) \perp G(x) \geq 0,$$

where  $\perp$  denotes orthogonality. In particular, this is equivalent to asking that for each index  $i \in [1 : n] := \{1, \dots, n\}$ ,  $F_i(x) \geq 0$ ,  $G_i(x) \geq 0$  and  $F_i(x)G_i(x) = 0$ . This rather general form makes complementarity problems a valuable tool to treat a vast array of situations, ranging from the optimality conditions of inequality-constrained optimization problems to various physical phenomena. Very often, the function  $G$  is taken to be the identity, leading to:

$$\text{NCP}(F) \quad 0 \leq x \perp F(x) \geq 0. \quad (1.2)$$

Furthermore, when  $F$  is affine, say  $F(x) = Mx + q$ , one gets

$$\text{LCP}(M, q) \quad 0 \leq x \perp Mx + q \geq 0, \quad (1.3)$$

which is called a linear complementarity problem [58]. An important part of the literature focuses on this linear case, where the properties of  $M$  are paramount [181, 86, 48].

Starting with the seminal work of Cottle in his PhD thesis in 1964 [56, 57], the literature on complementarity problems has vastly developed. There exists a plethora of methods that can be used to deal with CPs, to which many authors contributed, coming from both

applied and theoretical backgrounds. Let us mention a few of the fields where complementarity arises, sometimes after a space discretization to make dimension finite, alongside with some related contributions: contact problems [2, 9, 62, 129, 128, 91, 261], multiphase flow in numerical simulations [18, 20, 36, 38, 163, 164, 249], PDEs with complementarity constraints [21, 122], in which the system possesses equalities in addition to complementarity constraints, computer graphics [84], finance and economics [100]; see also the references therein. In particular, the surveys of Harker and Pang [120], Pang [193] then Ferris and Pang [91] contain many more references, details and discussions on applications of CPs such as traffic equilibrium, game theory among others. A variant of the LCP was recently considered in the “tropical algebra” setting [8].

## 1.2 Related problems

Before evoking algorithms solving complementarity in the forms discussed above, let us mention some other problems that are more or less closely related to it. The *vertical* LCP [54, 91, 55, 58] consists in the complementarity between multiple affine functions, contrary to (1.3) with the identity and only one affine function. The *horizontal* LCP reads

$$Ax + By = c, \quad x \geq 0, \quad y \geq 0, \quad x^T y = 0. \quad (1.4)$$

In [55], an extended version with  $Ax + By \in K$  for a polyhedral convex set  $K$  is proposed. The generalized CP (GCP) takes the form

$$F(x) \in K, \quad G(x) \in K^*, \quad \langle F(x), G(x) \rangle = 0, \quad (1.5)$$

for some closed convex cone  $K$  and its dual  $K^*$  for the scalar product  $\langle \cdot, \cdot \rangle$  [120, definition 2.3 p. 166]. The choice of  $K = \mathbb{R}_+^n$  reduces the GCP to the standard CP. Other problems may reduce to complementarity problems, such as quadratic inequality-constrained optimization when looking at the optimality conditions. Consider

$$\min \frac{1}{2} x^T M x + c^T x, \quad \text{s.t.} \quad Ax \leq b,$$

with a symmetric  $M$ . Indeed, if  $\lambda$  denotes the nonnegative multipliers, the optimality conditions read

$$Mx + c + A^T \lambda = 0, \quad 0 \leq \lambda \perp (b - Ax) \geq 0,$$

where complementarity applies for part of the system (the number of the constraints). Such systems are *mixed* complementarity problems (MCPs), where a part of the equations are equalities and the other are complementarity conditions and reads for instance [86, §9.4.2]

$$G(u, v) = 0, \quad 0 \leq v \perp H(u, v) \geq 0, \quad (1.6)$$

where  $G : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$  and  $H : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^m$ .

Another popular related framework is the one of variational inequalities, the link with complementarity problems having been unveiled by Karamardian in [139]. In a rather general form, they read

$$x \in C, \quad \langle F(x), (y - x) \rangle \geq 0, \forall y \in C, \quad (1.7)$$

where  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  and  $C$  is a closed convex set in  $\mathbb{R}^n$ . They may also be expressed in the equivalent form

$$0 \in F(x) + N_C(x), \quad (1.8)$$

where  $N_C(x)$  denotes the normal cone to  $C$  at  $x$  [2, 47, 86, 98, 131]. In particular, for  $C = \mathbb{R}_+^n$ , one recovers NCP( $F$ ). The particular case of box constraints,  $C = \{x \in \mathbb{R}^n : l \leq x \leq u\}$  for a lower bound  $l \in (\{-\infty\} \cup \mathbb{R})^n$  and an upper bound  $u \in (\mathbb{R} \cup \{+\infty\})^n$ , is dealt with in [47, 90, 179]. For a reference in infinite dimension, see for instance [247].

### 1.3 Equivalent formulations and algorithms

In the most general case, solving a CP is NP-complete [49]. Even for specific types of matrices  $M$ , the problem may remain difficult [60].

We start by mentioning the algorithm of Lemke [152] for LCPs, which is reminiscent of the simplex algorithm for linear programming. Under a common assumption on  $M$ , the algorithm moves between complementary pairs, shifting the indices for which  $x_i = 0$  or  $(Mx + q)_i = 0$ . It may however require, as the simplex method does, an exponential number of steps [180, 88].

Interior-points algorithms have also been applied to complementarity problems, by the relaxation of the complementarity into  $F_i(x)G_i(x) = \mu > 0$  for all  $i \in [1 : n]$ . This allows the methods to avoid the inherent combinatorial aspect of finding, for each index  $i \in [1 : n]$ , if  $F_i(x)$  and/or  $G_i(x)$  equals zero. Let us mention a few contributions: [145, 174, 144], [44] for a slightly different point of view.

The framework of *generalized equations* introduced by Robinson, can also deal with complementarity problems. They generalize (1.7) and (1.8) and may be expressed as

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

where  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is a mapping and  $T : \mathbb{R}^n \multimap \mathbb{R}^n$  is a multifunction (a function from  $\mathbb{R}^n$  to the power set of  $\mathbb{R}^n$ ). With  $F(x) = Mx + q$  and  $T(x) = N_{\mathbb{R}_+^n}(x)$  the normal cone to the nonnegative orthant [217, section 4], one recovers the LCP. Further material on generalized equations can be found in [220, 216] for instance.

The Primal-Dual Active Set (PDAS) methods represent another branch of the possible algorithms (when one of the functions is the identity). At an iterate, these methods determine which indices are active ( $x_i = 0$ ) then treat differently the indices whether they correspond to the  $x$  part or the  $F(x)$  part.

This method is utilized in [128] on a practical application, in [122] for variational inequalities and in [124] for the infinite dimensional case, where it is also shown that this method may behave identically as the algorithm based on the C-function  $\varphi_{\min}$  discussed below.

A rather popular technique is the reformulation by C-functions, which is the main one discussed in this thesis, see section 2.3.1. Shortly, a C-function  $\varphi$  verifies, for  $a$  and  $b$  in  $\mathbb{R}$

$$\varphi(a, b) = 0 \iff a \geq 0, \quad b \geq 0, \quad ab = 0,$$

and allows one to reformulate a CP into

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

and to the minimization of the so-called merit function  $\Psi(x) := \|\Phi(x)\|^2/2$ , for which a zero, a local minimum or a stationary point is sought depending on the problem's difficulty.

A few important articles about C-functions are for instance [159, 138, 94, 194, 92, 204, 99, 6]. Often, the resulting systems are nondifferentiable, which leads to their *smoothing*:

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

where  $\tilde{\varphi}(\cdot, \cdot, \mu)$  is differentiable when  $\mu > 0$  and  $\tilde{\varphi}(a, b, 0) = \varphi(a, b)$ . This smoothing of the system is for instance considered in [98, 45, 47, 86, 260, 157, 117, 190, 249].

With the use of C-functions, CPs are related to the vaster question of solving systems of (nonsmooth) equations. Some related work includes [217, 216, 51, 195, 197, 205, 46, 127], among many others.

The cited contributions represent only a glimpse of the available literature; additional material may be found therein the given references. A main goal of this PhD thesis is to study the C-function  $\varphi_{\min}$ . It has the particularities of being among the simplest ones, since it is piecewise linear in its component (in  $F$  and  $G$ ), but also the least differentiable. This drawback seems to have caused  $\varphi_{\min}$  to be less studied (but still quite appreciated in practice) than the Fischer function  $\varphi_{FB}$  and its offspring.

## 1.4 Nonsmoothness and combinatorial geometry

Now, we briefly discuss a topic called “hyperplane arrangements”. Its relation with the above topics comes from a computation related to the minimum C-function, as discussed in chapter 3. While it may appear innocuous for researchers with an optimization background, this is an extremely vast, deep and classic topic for specialists in the fields of algebra, combinatorics and discrete geometry.

Let  $n \geq 1$  be an integer representing the dimension,  $v \in \mathbb{R}^n \setminus \{0\}$  and  $t \in \mathbb{R}$ . The hyperplane  $H_{v,t} := \{x \in \mathbb{R}^n : v^\top x = t\}$  clearly splits  $\mathbb{R}^n$  into itself and its two associated open half-spaces:

$$H_{v,t}^+ := \{x \in \mathbb{R}^n : v^\top x > t\} \quad \text{and} \quad H_{v,t}^- := \{x \in \mathbb{R}^n : v^\top x < t\}. \quad (1.10)$$

Now, let  $V = [v_1 \cdots v_p] \in \mathbb{R}^{n \times p}$  and  $\tau = (\tau_1, \dots, \tau_p) \in \mathbb{R}^p$ . Consider the collection of hyperplanes  $\{H_i : i \in [1 : n]\}$  where  $H_i := H_{v_i, \tau_i}$ . Let  $\mathcal{A}(V, \tau)$  be this collection, called

hyperplane arrangement. The goal is to study the geometrical structure formed by the hyperplanes.

Arrangements may be studied under various viewpoints and for various reasons, though one of the main questions is often the number of *chambers*, also called *regions* or *cells*. More precisely, let  $\{H_1, \dots, H_p\}$  be a collection of  $p$  hyperplanes in  $\mathbb{R}^n$ . Thus,  $\mathbb{R}^n \setminus \bigcup_{i=1}^p H_i$  is split into connected components that are called the chambers. An example is given in figure 1.1.

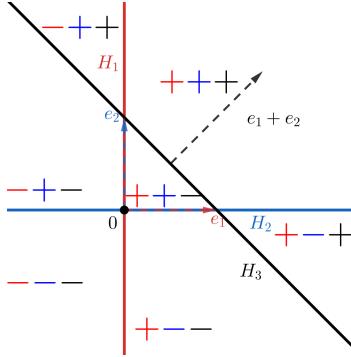


Figure 1.1: Illustration with three hyperplanes and 7 chambers, where the hyperplanes are  $H_1 = \{x \in \mathbb{R}^2 : x_1 = 0\}$ ,  $H_2 = \{x \in \mathbb{R}^2 : x_2 = 0\}$  and  $H_3 = \{x \in \mathbb{R}^2 : x_1 + x_2 = 1\}$ .

Each of these regions is a subset of  $H_i^+$  or  $H_i^-$ , for all  $i \in [1 : p]$ , with which one can associate a sign  $s_i \in \{\pm 1\}$ . Thus, identifying the chambers means obtaining all combinations of  $\{-1, +1\}^p$ , which are called *sign vectors*, that correspond to a nonempty intersection. Counting the chambers is a topic that has been studied very early, with some articles dated from early in the XIX<sup>th</sup> century [239, 227, 215] ([227] was published posthumously; the contribution concerning arrangements is grouped with other papers dealing with extremely various fields). Our goal is to get the set  $\mathcal{S}$  of sign vectors  $s$  described above, not only to know the number  $|\mathcal{S}|$  of such chambers. While the two questions may seem very close – since identifying them gives their number and counting them may be done by an enumeration, our understanding is that both questions actually meaningfully differ. Many contributions provide the number of the chambers without identifying them by using intrinsic combinatorial tools.

The myriad of relevant contributions is too vast to be more than barely glimpsed here. We mention some articles and books that are relevant for counting (since it is much more frequent) and enumerating the chambers of an arrangement.

For rather general books on combinatorics, arrangements, going far beyond the scope of what is needed in this thesis, see for example the works by Crapo and Rota [61], Orlik and Terao [187], Stanley [237, 238] and an introductory content [236] at MIT. Edelsbrunner wrote a book specialized on algorithms [81]. Aguiar and Mahajan propose a more recent monograph [4]. Halperin and Sharir summarize many contributions in a survey [118].

Among contributions that address the problem of computing the *number* of chambers (but not the chambers themselves), let us mention the groundbreaking work of Zaslavsky [257], using a tool called “characteristic polynomial”. Other sources using it include [12, 238].

Since we focus on enumerating the chambers of an arrangement, most of the software from the field of combinatorics are much more general, dealing with very wide ranges of topics beyond arrangements. Nonetheless, let us mention some of them: `Sagemath` [68], `Macaulay2` [111], `OSCAR` [67, 189], `TOPCOM` [214], `polymake` [140], `Counting_Chambers` [35] (which, despite its name, does not simply count the chambers).

Finally, let us mention some specific algorithms. Bieri and Nef, in [27], compute the full structure of the arrangement, and not only the chambers of dimension  $n$ . The same is done by Edelsbrunner, O'Rourke and Seidel [83], where the algorithm has an optimal theoretical complexity. To focus on the chambers, which are our main interest, we mention Avis, Fukuda and Sleumer [14, 232]. A better algorithm was designed by Rada and Černý in [208], which is the one we improve and develop in this thesis.

## 1.5 Outline and contributions

This thesis is organized as follows. Chapter 2 consists in a detailed introduction, composed of the notation used throughout the text, a discussion on relevant literature, and an explanation of the goal of this thesis. First, complementarity problems (CPs) are presented more in-depth, focusing on reformulations and algorithms, especially nonsmooth ones. Then, the introduction ends up with a more thorough presentation on arrangements.

The first part, chapters 3 and 4, discusses the B-differential of the minimum of affine functions and centered arrangements. Indeed, the computation of this B-differential, i.e., a nonsmooth analysis question arising in the context of a particular algorithm for solving CPs, results in identifying the chambers of a centered arrangement, in which all the hyperplanes have a common point. These chapters present properties of arrangements and equivalent topics, as well as algorithms that identify the chambers. The second part, chapter 5, focuses on the more general case where the hyperplanes do not necessarily all have a point in common, which implies adapting the notions and algorithms of chapter 3.

Both parts essentially present improvements and algorithmical variants on a state-of-the-art tree algorithm introduced by Rada and Černý [208], in order to reduce the amounts of computation. Some of these improvements are based on analytical or heuristical remarks on tree structure, to prune earlier (sub)branches of the tree.

Another rather different method is discussed, which we call the “dual” approach. In particular, it involves the use of the circuits of the underlying (oriented) matroid [191]. We have not found an explicit (at least algorithmically) use of the circuits to know which sign vectors do (*not*) correspond to chambers. Beyond this conceptual discovery, this rather surprising appearance of duality leads to significant improvements on the algorithm of Rada and Černý, especially on instances having a combinatorial nature coming from applications. This contribution [77] is published in Mathematical Programming Computation; the corresponding Matlab code and its documentation are available online [75, 76].

Chapter 4 presents additional details and complements on the previous topic: case of nonlinear functions and role of the linearization, B-differential of the merit function, details

on some tested instances.

Though centered arrangements occur when one looks at the B-differential of the componentwise minimum of affine functions, the more general case of arbitrary arrangements was not considered. It was therefore natural to know whether the techniques introduced for centered arrangements could be extended to more general arrangements. This is the topic of chapter 5, in preparation (initially submitted to SIAM Journal on Discrete Mathematics [79]).

The last part, chapter 6, discusses the globalization of a variant of the Newton-min algorithm, a method using the minimum C-function  $\varphi_{\min}$  and solves a linear system at each iteration, for solving complementarity problems [72]. In this contribution, the search direction was obtained by finding a point in a certain polyhedron, which was guaranteed to be nonempty close to a point satisfying some regularity assumptions at and everywhere around the solution. Our goal in this thesis was to bypass the stringent regularity assumption by using a Levenberg-Marquardt technique, which guarantees the existence of a solution at a likely higher computational cost.

However, without such assumption, a simple question arose: since the minimum reformulation has a piecewise structure, “What piece should be chosen?”. A similar question is: at an iterate where the function is not smooth, “What element in the differential should be chosen?” in a Newton-type method. In particular, we show a relation between detecting “strong” (Dini) stationarity of an iterate (no better point around it) and the choice of the “piece”, which, in turn, states that detecting “strong” (Dini) stationarity is co-NP-complete in general. Nonetheless, we discuss various properties surrounding this question, such as what the Levenberg-Marquardt globalization technique may yield. These contributions form the bases of a publication in project.



# Chapter 2

## General setting

This chapter aims at discussing the goals of this thesis and its position with respect to the literature. The two issues at stake are complementarity problems and hyperplane arrangements. Actually, the two fields in which these problems arise are rather distant from each other. Consequently, the two main parts of the thesis are not that closely related.

At first, this thesis aimed at designing a globally convergent method for complementarity problems based on the minimum C-function  $\varphi_{\min}$ . Notably, this reformulation is highly nonsmooth. Therefore, unless some strong regularity hypotheses are made, nonsmoothness considerably hinders the globalization of local algorithms such as the Newton-min algorithm. In [72], the authors modify this standard local method by considering polyhedral systems instead of linear ones at each iteration. The initial motivation was to use a Levenberg-Marquardt technique on these systems to avoid regularity assumptions despite some more expensive iterations.

This topic is presented in chapter 6. However, the Levenberg-Marquardt approach was somehow related to the question of choosing one *suitable* element in a differential in order to obtain a descent property on the merit function. This question led us to try to better understand better the involved differential, i.e., the differential of the (componentwise) minimum of two functions. It turned out that even with affine functions involved, computing this differential is equivalent to identify the chambers of a certain hyperplane arrangement, where all the planes intersect.

This is the topic of chapter 3, where other equivalent problems, properties and algorithms to solve this question are discussed. In particular, a new method based on duality and matroid circuits is designed. Complements and details are proposed in chapter 4. Chapter 5 completes the previous discussion by studying the case of arrangements with affine hyperplanes, and extends the work of the preceding chapters. Appendix A discusses various complements on arrangements.

Finally, chapter 6 discusses the initial motivation of CPs and globalization. In particular, we shall see some form of “symbiosis” with the topic of the other chapters: arrangements were considered since they represent the hidden side to a differentiability issue that arises in CPs, and were useful in developing some new insight on the algorithmical approach for

CPs. Appendices B, C, D detail technical precisions completing chapter 6.

After precising the notation used throughout the remaining parts, this chapter mostly develops the introduction and discusses with more details some of the choices made that lead to this work.

## 2.1 Notation

The first part, section 2.1.1, presents relatively common notation, while less common terms are introduced in section 2.1.2.

### 2.1.1 General notation

- $\mathbb{N}$  and  $\mathbb{N}^*$  denote the sets of nonnegative and positive integers.
- $[1 : n] := \{1, \dots, n\}$  the first  $n$  positive integers;  $[n_1 : n_2] := \{n_1, \dots, n_2\}$  for  $n_1 \leq n_2$  in  $\mathbb{N}$ .
- $\mathbb{R}$  and  $\mathbb{R}_+$  denote the sets of real and nonnegative real numbers.
- $\mathbb{R}^n$ ,  $\mathbb{R}_+^n$ , and  $\mathbb{R}_{++}^n$  denote the real space of dimension  $n$ , the nonnegative and positive orthants in  $\mathbb{R}^n$ ;  $\mathbb{R}_+^n = \{x \in \mathbb{R}^n : x \geq 0\}$  where the inequalities apply componentwise;  $\mathbb{R}_{++}^n = \{x \in \mathbb{R}^n : x > 0\}$ .
- $t_+ := \max(t, 0)$  for  $t \in \mathbb{R}$  (also named “ReLU”);  $t_+^2$  is a compact notation for  $(t_+)^2$  (since  $(t^2)_+ = t^2$ ); for vectors  $v \in \mathbb{R}^n$ ,  $v = v_+ - v_-$  with  $(v_+)_i = \max(v_i, 0)$  and  $(v_-)_i = \max(-v_i, 0) = -\min(v_i, 0)$ .
- $\text{sgn}$  is the function  $\mathbb{R} \rightarrow \mathbb{R}$  defined by  $\text{sgn}(t) = -1$  if  $t < 0$ ,  $\text{sgn}(t) = +1$  if  $t > 0$  and  $\text{sgn}(0) = 0$ ;  $\text{sgn}(v) := (\text{sgn}(v_i))_{i \in [1:n]}$  for some  $v \in \mathbb{R}^n$ .
- $u \cdot v$  is the Hadamard product of two vectors  $u$  and  $v$  of  $\mathbb{R}^n$  defined by  $(u \cdot v)_i = u_i v_i$  for  $i \in [1 : n]$ .
- $|v| := (|v_i|)_{i \in [1:n]} = \text{sgn}(v) \cdot v$  is the componentwise absolute value of a vector  $v \in \mathbb{R}^n$ .
- $e$  denotes the vector of all 1’s, with a size understandable from the context; in particular,  $e \cdot u = u$  for  $u \in \mathbb{R}^n$ ;  $\{e_1, \dots, e_n\}$  is the canonical basis of  $\mathbb{R}^n$ .
- $\|\cdot\| := \|\cdot\|_2$  is the 2-norm in  $\mathbb{R}^n$ :  $\|x\|_2^2 = \sum_{i=1}^n x_i^2$ ;  $\|\cdot\|_1$  denotes the 1-norm:  $\|x\|_1 := \sum_{i=1}^n |x_i|$ ;  $\|\cdot\|_\infty$  denotes the  $\infty$ -norm:  $\|x\|_\infty := \max_i\{|x_i|\}$ .
- $\text{supp}(v) := \{i \in [1 : n] : v_i \neq 0\}$  for some vector  $v \in \mathbb{R}^n$ ; the support may also be defined for a set of indices (not necessarily identified with  $[1 : n]$  for some  $n \in \mathbb{N}^*$ ).
- $I$  denotes the identity matrix whose size is clear from the context. It may be called Id.
- $A_{I,J}$  denotes, for a matrix  $A \in \mathbb{R}^{m \times n}$ , subsets  $I \subseteq [1 : m]$  and  $J \subseteq [1 : n]$ , the submatrix  $(A_{ij})_{i \in I, j \in J}$ ; “ $:$ ” denotes no selection ( $I = [1 : m]$  or  $J = [1 : n]$ ), hence the  $i$ th row of  $A$  is  $A_{i,:}$  and its  $j$ th column is  $A_{:,j}$ .

- $\mathcal{N}(A), \mathcal{R}(A)$  denote the null space and the range space of  $A \in \mathbb{R}^{m \times n}$ ;  $\text{rank}(A) := \dim \mathcal{R}(A)$  is its rank,  $\text{null}(A) := \dim \mathcal{N}(A)$  its nullity; by the rank-nullity theorem,  $\text{rank}(A) + \text{null}(A) = n$ .
- $\cdot^\top$  is used to denote the transpose of a vector or a matrix.
- $[A; B]$  is the vertical concatenation of matrices  $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{m' \times n}$  and belongs to  $\mathbb{R}^{(m+m') \times n}$ ;  $[A \ B] := [A^\top; B^\top]^\top$  is the horizontal concatenation for matrices  $A \in \mathbb{R}^{m \times n}$  and  $B \in \mathbb{R}^{m \times n'}$ , belonging to  $\mathbb{R}^{m \times (n+n')}$ ; it also applies to vectors.
- $\det(A)$  is the determinant of  $A$  when  $A$  is square.
- $\text{sp}(A)$  is the spectrum of  $A$ , the set of eigenvalues of  $A$  when  $A$  is square.
- $A \in \mathbb{R}^{n \times n}$  is said to be positive (semi)definite if for all  $x \in \mathbb{R}^n, x^\top A x > 0 (\geq 0)$ .
- $\text{Diag}(v)$  for  $v \in \mathbb{R}^n$  is the diagonal matrix with diagonal equal to  $v$ .
- $\mathcal{S}^n, \mathcal{S}_+^n, \mathcal{S}_{++}^n$ : sets of symmetric, symmetric positive semi-definite, symmetric positive definite matrices.
- $|S|$  denotes the cardinality of a set  $S$ .
- $\cup$  denotes the disjoint union of sets.
- $S^c$  is the complement of  $S$  in a (larger) set that is clear from the context.
- $S^J$  denotes the set of vectors whose elements are in  $S$  and are indexed by the elements of  $J$ ; equivalently it is the set of maps from  $J$  to  $S$ .
- $2^S$  denotes the power set of  $S$ , the set of all subsets of  $S$  (including  $S$  and  $\emptyset$ ); equivalently it is the set of maps from  $S$  to  $\{0, 1\}$ .
- $\text{vect}(S)$  the vector subspace spanned by a subset  $S$  of a vector space.
- $V^\perp := \{x \in \mathbb{R}^n : x^\top v = 0, \forall v \in V\}$  is the orthogonal of a subspace  $V \subseteq \mathbb{R}^n$ .
- $\text{conv}(S)$  denotes the convex hull of the subset  $S$  of a vector space.
- $P_C(x)$  denotes the orthogonal projection of  $x$  on the convex closet set  $C$ , defined by  $P_C(x) = \underset{y \in C}{\text{argmin}} \|y - x\|^2 / 2$ .
- $C^* := \{v \in \mathbb{R}^n : \langle v, c \rangle \geq 0, \forall c \in C\}$  for some subset  $C \subseteq \mathbb{R}^n$  denotes the dual cone of  $C$  for the Euclidean scalar product.
- $N_C(x) := \{v : \forall y \in C, \langle v, y - x \rangle \leq 0\}$  for a closed convex set  $C$  denotes the normal cone at  $x$  to  $C$ . The tangent cone is the dual cone of the normal cone, denoted and defined by  $T_C(x) := N_C(x)^*$ .
- $F'(x)$  is the derivative of function  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  at  $x$ .
- $F'_I(x)$  for a smooth function  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is to be understood as  $(F_I)'(x) \in \mathbb{R}^{I \times n}$  for  $I \subseteq [1 : m]$ .
- $\nabla f(x) \in \mathbb{R}^n$  is the gradient of a (smooth) function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  for the Euclidean scalar product.
- $q(t) = o(t)$  denotes a quantity such that  $\lim_{t \neq 0, t \rightarrow 0} q(t)/t = 0$ ;  $q(t) = O(t)$  denotes a quantity such that  $q(t)/t$  is bounded for  $t \neq 0$ .

### 2.1.2 Specific notation

Now, some more specific notions are introduced.

- $\mathfrak{B}(S)$  denotes the set of bipartitions of  $S$ , i.e., all different pairs of subsets  $I$  and  $J$  such that  $I \cap J = \emptyset$ ,  $I \cup J = S$ , considering  $(I, J)$  and  $(J, I)$  as different.
- $H_{v,t} := \{x \in \mathbb{R}^n : v^\top x = t\}$  is the hyperplane orthogonal to  $v$  containing  $tv/\|v\|^2$ .
- $\mathcal{D}_H$  is used to denote the differentiable domain of the (vector-valued) function  $H$ , the points where the function  $H$  is differentiable.
- $\partial_B H(x) := \{J \in \mathcal{L}(\mathbb{E}, \mathbb{F}) : H'(x_k) \rightarrow J, \mathcal{D}_H \ni x_k \rightarrow x\}$  ( $B$  for Bouligand).
- $\partial_C H(x) := \text{conv}(\partial_B H(x))$ , also denoted  $\partial H(x)$ .
- $\partial_X H(x) := \partial_C H_1(x) \times \cdots \times \partial_C H_n(x)$ . It is often denoted by  $\partial_C$  in the literature, but “ $C$ ” could be understood as a reference to Clarke.
- $\varphi$  denotes a C-function,  $\Phi : \mathbb{R}^n \mapsto \mathbb{R}^n$  the mapping using the C-function  $\varphi$  and  $\Psi := \|\Phi\|^2/2$  the associated merit function.

## 2.2 Reformulations of complementarity problems

Recall from section 1.1 that the considered problems are, in order, the GCP with functions  $F$  and  $G$ , the NCP with function  $F$  and the LCP with matrix  $M$  and vector  $q$ :

$$\text{GCP}(G, F) \quad 0 \leqslant F(x) \perp G(x) \geqslant 0, \tag{2.1a}$$

$$\text{NCP}(F) \quad 0 \leqslant x \perp F(x) \geqslant 0, \tag{2.1b}$$

$$\text{LCP}(M, q) \quad 0 \leqslant x \perp Mx + q \geqslant 0. \tag{2.1c}$$

### 2.2.1 Some types of LCPs

This section aims at presenting a few of the many classes of matrices that are involved in LCPs, which distinguishes the types and variants of LCPs. The definitions are mostly taken from [58, 86, 181]. None of the matrices are necessarily symmetric. First, the class of P-matrices ensures that there exists a unique solution to  $\text{LCP}(M, q)$  for any  $q \in \mathbb{R}^n$  (see also [225]).

**Definition 2.2.1 (P-matrices).**  $M$  is a P-matrix, denoted by  $M \in \mathbf{P}$ , if it verifies any of the following equivalent conditions:

- (i)  $\forall q \in \mathbb{R}^n$ ,  $\text{LCP}(M, q)$  has exactly one solution,
- (ii)  $\forall I \subseteq [1 : n]$ ,  $\det(M_{I,I}) > 0$ ,
- (iii)  $\forall I \subseteq [1 : n]$ ,  $\text{sp}(M_{I,I}) \cap \mathbb{R} \subseteq \mathbb{R}_{++}$ ,
- (iv) any  $x$  verifying  $x \cdot (Mx) \leqslant 0$  vanishes. □

As suggested by points (ii) and (iii), verifying if  $M$  belongs to  $\mathbf{P}$  is not easy. Coxson [60] showed that it is a co-NP-complete problem, by using a result on the NP-completeness of a problem dealing with the nonsingularity of an interval of matrices containing a non-singular matrix. Many other characterizations of the  $\mathbf{P}$ -matrices exist. Ben Gharbia and Gilbert, in [23, 24], give an equivalent definition based on the absence of cycling between two points in a specific algorithm (see section 2.3.3). Rump in [222] proposes another characterization based on matrix spectra and the Cayley transform. The paper also shows that no subset can be ignored in points (ii) and (iii) of definition 2.2.1.

The discussed properties are also, though not stated in these terms, observed in the article of Samelson, Thrall and Wesler [225], which gives a property about how the pair  $(I, -M)$  may decompose the space. Their theorem is stated with 2 matrices but one can be the identity.

The following class, for which the LCP always has (at least) a solution, has no known equivalent (algebraic) characterizations.

**Definition 2.2.2 (Q-matrices).**  $M$  is a  $\mathbf{Q}$ -matrix, denoted by  $M \in \mathbf{Q}$ , if for any  $q \in \mathbb{R}^n$ ,  $\text{LCP}(M, q)$  has (at least) a solution.  $\square$

The following class discuss a particular subclass of nonsingular matrices.

**Definition 2.2.3 (ND-matrices).**  $M$  is a ND-matrix, denoted by  $M \in \mathbf{ND}$  if it verifies any of the following equivalent conditions:

- (i)  $\forall I \subseteq [1 : n], \det(M_{I,I}) \neq 0,$
- (ii) any  $x \in \mathbb{R}^n$  such that  $x \cdot (Mx) = 0$  vanishes.

$\square$

An overview of relations between the presented classes and many more may be found in [20, p. 37, fig. 2.2.1, in French]. An overview of many complexity issues on classes of matrices may be found in an article by Tseng [245]: determining  $\mathbf{P}$ -matricity, strict monotonicity, column sufficiency (corollary 1 p. 187),  $\mathbf{P}_0$ -matricity, semi-monotonicity (corollary 2 p. 190),  $\mathbf{R}_0$  and nondegeneracy (corollary 3 p. 191) are all co-NP-complete problems.

## 2.2.2 Some types of NCP( $F$ )

Akin to the previous section, this one aims at presenting some classes of functions that are involved in NCP( $F$ )s, the definitions being taken from [243] for instance. Some of them are analogous to some matrix classes encountered for the LCP, where  $F'(x)$  (for all  $x \in \mathbb{R}^n$ ) plays the role of the matrix  $M$ .

**Definition 2.2.4** (types of function  $F$ ). Let  $x$  and  $y \in \mathbb{R}^n$ . It is said that  $F$  is a ...function

$\mathbf{P}_0$	if $\exists i \quad (x_i - y_i)(F_i(x) - F_i(y)) \geq 0$	$\Leftrightarrow F'(x) \in \mathbf{P}_0,$
$\mathbf{P}$	if $\exists i \quad (x_i - y_i)(F_i(x) - F_i(y)) > 0,$	
<i>uniform – P</i>	if $\exists i \quad (x_i - y_i)(F_i(x) - F_i(y)) \geq \mu \ x - y\ ^2 \Leftrightarrow F'(x) \in \mathbf{P},$	
monotone	if $(x - y)^T(F(x) - F(y)) \geq 0,$	
strictly monotone	if $(x - y)^T(F(x) - F(y)) > 0,$	
strongly monotone	if $(x - y)^T(F(x) - F(y)) \geq \mu \ x - y\ ^2$	

□

In these properties, instead of  $\exists i \in [1 : n]$ , a maximum over the  $i$  is sometimes used, and  $x_i \neq y_i$  for the first three and  $x \neq y$  for the fifth. One clearly has

$$\text{monotone} \Rightarrow \mathbf{P}_0, \quad \text{strictly monotone} \Rightarrow \mathbf{P}, \quad \text{strongly monotone} \Rightarrow \text{uniform } \mathbf{P}$$

using that  $(x - y)^T(F(x) - F(y)) = \sum_i (x_i - y_i)(F_i(x) - F_i(y))$ . Monotonicity is used for instance by Subramanian [240]. Megiddo [168] found an example of a NCP( $F$ ) without solution despite  $F$  being monotone.

### 2.2.3 General complexity

Before moving on to reformulations, we evoke the complexity of complementarity problems. Since LCPs are a particular case of NCPs, complexity results on LCPs already show the difficulty of the problems at stake. As discussed in section 2.2.1, even determining the type of the matrix involved in a LCP may be difficult.

Chung [49] showed that the LCP is, in general, NP-complete, and even strong-NP-complete.

When  $M \in \mathbf{P}$ , Megiddo have shown that the problem was probably not “hard” [169]: a theorem shows that if LCP( $M, q$ ) is NP-hard, then NP = co-NP. Furthermore, when  $M$  is symmetric and positive (semi)definite, the ellipsoid’s method can be used to solve the LCP in polynomial time [181]. When  $M \in \mathbf{P}_0$ , the complexity becomes NP-complete as described by Kojima, Megiddo, Noma and Yoshisa [144].

### 2.2.4 Generalized equations and normal maps

Complementarity problems can be expressed and formulated in many various ways, which underline the importance and relevance of their study. LCPs can be stated, for instance, as a generalized equation

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

Using a Cartesian product formula for the normal cone [221], [105, proposition 2.30 2) p. 49], one gets:

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

which indicates solutions of the LCP and of (2.2) are the same. Studied in detail by Robinson [217, 220, 216], they represent a wider framework with many applications such as optimality conditions or other equilibrium problems.

Ralph [211] presents a similar reformulation for the  $\text{NCP}(F)$ , a *normal map*, which takes the form (recall that  $z = z_+ + z_-$ ):

$$F(z_+) + (z - z_+) = 0 = F(z_+) + z_-. \quad (2.3)$$

The solutions of both problems are related by  $z = x - F(x)$  for  $x$  a solution of (2.1b) and  $x = z_+$  for  $z$  a solution of (2.3). Indeed, if  $x$  solves (2.1b), for some  $i \in [1 : n]$ ,  $z_i = x_i - F_i(x)$ . By complementarity of  $x$  and  $F(x)$ , one has  $z_+ = x$  and the normal map equation becomes  $F(x) + (x - F(x) - x) = 0$  which clearly vanishes. Conversely, if  $z$  solves (2.3),  $x = z_+ \geq 0$ ,  $F(x) = z_+ - z = -z_- \geq 0$ . Then,  $x_i F_i(x) = (z_+)_i (z_-)_i = 0$ , meaning complementarity is respected.

In addition to the method of Ralph, we mention the contribution of Sun and Qi [243] where they modify the normal map by a smoothing technique (see section 2.3.6).

Normal maps have been studied by Robinson in [219], where a linear transformation is considered but with a polyhedral convex set  $C$  instead of the nonnegative orthant. After him, Josephy [134] proposes a scheme that linearize the function, which transforms  $\text{NCP}(F)$  into a sequence of LCPs to solve, with same dimension.

In (2.3), one could also use  $F(z_+) + z_- = 0$ . This form is employed for instance by Harker and Xiao [121], where it is called a Minty map after [173]. Their algorithm solves an equation with the B-derivative to obtain a direction before launching a linesearch. Therefore, it also deals with a mixed LCP during each of its iterations. It is however different during the linesearch phase, where they suggest it may behave better than the regular reformulation with the minimum.

## 2.2.5 Interior-points

Here, we briefly mention some algorithms dealing with complementarity problems by the use of interior-type methods. For optimization under inequality constraints of the form

$$\min f(x), \quad \text{s.t.} \quad g(x) \leq 0,$$

interior-point methods deal with the system of optimality conditions and replace the complementarity conditions of the form  $0 \leq \lambda \perp (-g(x)) \geq 0$  by  $\lambda_i(-g(x))_i = \mu$  for some  $\mu > 0$ . For general properties on interior-point methods, see for instance the book by Wright [254].

Clearly, interior-point methods can be adapted for complementarity problems. The book by Kojima, Megiddo, Noma and Yoshise [144] discusses the case of LCPs. Kojima and Yoshise, alongside Mizuno [145], still for LCPs, have improved the complexity from  $O(Ln^{7/2})$  ( $O(Ln^{1/2})$ ) iterations, since each divides the merit function by  $(1 - \eta/\sqrt{n})$ , and

each solves a linear system thus in  $O(n^3)$  to  $O(Ln^3)$  with  $L$  the input size and  $n$  the dimension. The reduction to  $n^3$  is based on shrewd manipulations on the systems and subproblems to solve. A similar approach is considered in [44].

An application arising from numerical analysis can be found in [21], where the discretization leads to a system with some complementarity constraint, i.e., a “mixed” system. The interior-point method is compared to other popular approaches discussed below. Additional material may be found in the references therein.

### 2.2.6 Absolute value equation

LCPs are also equivalent, up to some change of variables, to a type of problems called “absolute value equations” (AVE). They take the form

$$Ax - |x| = b \quad \text{or} \quad x - A|x| = b \quad (2.4)$$

where  $|x| = (|x_i|)_{i \in [1:n]}$  is the componentwise absolute value of  $x$ . The derivation relating LCP and absolute value equations can be found for instance in a paper by Mangasarian and Meyer [160], with an affine (but nontrivial) transformation between the variable of the LCP and the one of the AVE. They discuss properties about the existence of solutions by studying the spectra of some matrices. Radons and Tonelli-Cueto [210] also discuss properties of the AVE and the related mapping  $x \mapsto x - A|x|$  by studying some adapted spectrum of  $A$ . Their work uses degree theory, which also intervenes in the study of LCP (especially existence properties), see [58, chapter 6] and [86, section 2.1, pp. 126-145]. The nonlinear case  $F(x) - |x| = 0$ , related to the NCP, was recently considered in [63].

### 2.2.7 Problems with complementarity in the constraints

Now, we mention contributions dealing with what may be called Mathematical Programs with Equilibrium|Complementarity Constraints, MPECs|MPCCs. For instance [226], Scheel and Scholtes deal with problems of the form

$$\min f(z), \quad \text{s.t.} \quad g(z) \leq 0, \quad h(z) = 0, \quad \min(F^1(x), \dots, F^l(x)) = 0$$

with  $F^1, \dots, F^l$  smooth functions from  $\mathbb{R}^n$  to  $\mathbb{R}^m$ . They discuss ways to decompose the constraints and various forms of stationarity for such problems. MPECs were also considered as an application in [197].

Hintermüller and Kopacka [125] discuss the infinite-dimensional case, where stationarity notions must be adapted to this difficulty. Their algorithm solves regularized subproblems, which are nonsmooth, and uses a linesearch approach to globalize the convergence, though “there is no guaranteed descent along such a path” (see their remark 5.4 p. 888).

## 2.3 Nonsmooth setting and algorithms

This section is devoted to the nonsmooth framework arising from (most of) the C-functions which are presented in section 2.3.1. We recall some important related notions in section 2.3.2. Then, sections 2.3.3 and 2.3.5 discuss *some* of the existing methods: hypotheses, assumptions at the solution, subproblems solved etc. Between them, section 2.3.4 discusses a few selected contributions which have motivated this thesis. Following this, the notion of *smoothing*, i.e., to change slightly the system in order to obtain a smooth system, is presented in section 2.3.6 alongside some related contributions. Finally, section 2.3.7 evokes some questions of complexity, to get a better perspective of the different results.

### 2.3.1 Introduction of C-functions

A C-function (C for complementarity) is a scalar function of two variables defined as follows.

**Definition 2.3.1** (C-functions). A function  $\varphi : \mathbb{R}^2 \rightarrow \mathbb{R}$  is said to be a C- function if it verifies the following condition for all  $a, b$  in  $\mathbb{R}^n$ :

$$\varphi(a, b) = 0 \iff a \geq 0, \quad b \geq 0, \quad ab = 0. \quad (2.5)$$

□

The main use of C-functions is the following property:

$$0 \leq F(x) \perp G(x) \geq 0 \iff \Phi(x) := \begin{pmatrix} \varphi(F_1(x), G_1(x)) \\ \vdots \\ \varphi(F_n(x), G_n(x)) \end{pmatrix} = 0, \quad (2.6)$$

where the C-function is applied componentwise. Equivalently, minimizing the merit function  $\Psi(x) := \|\Phi(x)\|^2/2$  (with optimal value 0 if the initial CP has a solution) is another reformulation. While there exists many C-functions, two seem to have a particular role as basic functions that inspired most others.

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

In the first one, FB stands for Fischer-Burmeister, though sometimes only Fischer is cited [92]. The minimum was first used by Kostreva [147] in 1976, and the other two expressions were first observed by Wierzbicki in [252]. See below for further comments.

Many C-functions are nonsmooth, i.e., the CP becomes an equation  $H(x) = 0$  with  $H : \mathbb{R}^n \rightarrow \mathbb{R}^n$  nonsmooth. Despite the existence of smooth reformulations, they may not be as appropriate, as detailed in [86, prop. 9.1.1, pp. 794-795]. In short, algorithms resulting from a smooth (C-function) reformulation cannot benefit from fast local convergence around *degenerate* solutions (see the end of this section).

Mangasarian [159] proposes a more general framework to obtain C-functions.

**Definition 2.3.2** (Mangasarian’s reformulation). Let  $\rho : \mathbb{R} \rightarrow \mathbb{R}$  be a (strictly) increasing function with  $\rho(0) = 0$ . Then  $x$  solves the complementarity problem 2.1a (and similarly 2.1b or 2.1c) if and only if

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

□

Said differently, Mangasarian’s framework states that

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

for  $\rho$  verifying the properties of the definition is a C-function. In particular, the choice  $\rho(t) = t$  gives the C-function  $-2\varphi_{\min}$ . Mangasarian also discusses the nonsingularity of the Jacobian at a solution for  $\text{NCP}(F)$  of (2.1b). In particular (p. 91), if  $\rho'(0) = 0$ , the system becomes differentiable.

An example of contribution based on Mangasarian’s smooth reformulation [159] is by Subramanian [241], which requires strong assumptions on the solution  $x^*$  to obtain convergence results: nondegeneracy (or strict complementarity,  $x_i^* + F_i(x^*) > 0$ ) of the solution,  $\nabla F(x^*)$  is a nondegenerate matrix. The nondegeneracy of the solution is particularly strong since is purely dependent on the problem itself and may not often be verified.

Another construction framework is given by Luo and Tseng in [156], also used by Kanzow, Yamashita and Fukushima in [138]. It consists in combining other functions in the following form (they use an opposite condition on the sign of the variables in the C-functions)

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

where the  $\psi$  functions are continuous and equal to 0 on the negative orthant. For instance,  $\psi_0(t) = (t_+)^p$ ,

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

with  $p \geq 1$  a positive integer.

Many of the C-functions are not smooth (everywhere), but they are said to be *semi-smooth*, a property inbetween Lipschitzianity and smoothness, see section 2.3.2. We essentially focus on “symmetric” C-functions (in  $a$  and  $b$ ), though many asymmetric ones exist, which can be interesting to treat differently  $x$  and  $F(x)$  (resp.  $Mx + q$ ) in  $\text{NCP}(F)$  (resp.  $\text{LCP}(M, q)$ ) for instance.

Additional information on C-functions can be found in the survey from Fischer and Jiang [94] discussing the main properties of C-functions, such as smoothness, forms of sublevel sets, efficiency of the directions ... see also [138] and the references therein. Despite decades of innovation on the C-functions, there are still new functions designed: Galántai [99] discusses how C-functions can be built or *not* built. A way to decompose C-functions is also evoked. He gives a bestiary containing around 30 C-functions, with many being

based on the FB function in a way or another. In [6], the authors discuss more involved constructions of C-functions based on a generalization of the Fischer function, using the  $p$ -norm instead of the 2-norm of  $(a; b)$ . Algorithmic considerations are further detailed in section 2.3.3.

The term “degeneracy” is often used in various mathematical contexts, ranging from linear optimization to combinatorial geometry (see below) including complementarity problems, to emphasize on some relevant technical difficulties. In complementarity problems, a solution  $x^*$  is said to be degenerate if there exists some indices  $i$  for which  $x_i^* = 0 = F_i(x^*)$  (with similar definitions for the LCP and other variants of CPs). These indices, which depend solely on the problem itself, cause most C-functions to be nondifferentiable, which explains the difficulty: one must have some regularity conditions so that *every* Jacobian matrix at the solution is nonsingular.

### 2.3.2 Some tools of nonsmooth analysis

Since the C-function reformulation (2.6) often leads to a nonsmooth system of equations, nonsmooth analysis is thus an important tool to study these systems. The main reference we start with is Clarke’s book [51], which assumes the functions are Lipschitz, a property that is uniformly verified in this thesis, though there exist extensions without Lipschitzianity (done by Rockafellar for instance). For simplicity, we assume that the current space is  $\mathbb{R}^n$ , thus a finite-dimensional space which simplifies some notations.

**Definition 2.3.3** (Lipschitz function). A function  $f$  is locally  $(L)$ -Lipschitz at  $x \in \mathbb{R}^n$  if for all  $y$  and  $z$  in a neighborhood of  $x$ ,  $|f(z) - f(y)| \leq L\|z - y\|$ .  $\square$

It is said to be locally Lipschitz if it is Lipschitz on each bounded subset of its domain, and (globally) Lipschitz if the inequality holds for any  $x$  and  $y$ . Usually, for smooth functions, one has the Taylor expansion

$$f(x + d) = f(x) + \nabla f(x)^\top d + o(\|d\|) = f(x) + f'(x)d + o(\|d\|). \quad (2.9)$$

However, when  $f$  is nonsmooth at  $x$ , one cannot write such expansion as easily. We shall see some generalizations of this expansion. Nonetheless, thanks to Rademacher’s theorem [209] (see also [123]), Lipschitz functions (in finite dimension) benefit from the following property. Recall that  $\mathcal{D}_H$  is the differentiable domain of  $H$ , the points at which  $H$  is differentiable.

**Theorem 2.3.4** (Rademacher’s theorem). *Lipschitz functions are differentiable almost everywhere. In other terms, if  $H$  is a Lipschitz function,  $\text{meas}(\mathcal{D}_H^c) = 0$  where  $\text{meas}$  is the Lebesgue measure.*  $\square$

When  $f$  is nonsmooth at  $x$ , one can generalize the term  $f'(x)d = \nabla f(x)^\top d$  in (2.9).

**Definition 2.3.5** (directional derivative). Let  $f$  be Lipschitz near  $x \in \mathbb{R}^n$ , the one-sided directional derivative at  $x$  in direction  $d \in \mathbb{R}^n$  is denoted and defined by

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

$\square$

Since this limit may not be defined everywhere, sometimes another definition, with a limsup, is used, the Dini upper directional derivative.

**Definition 2.3.6** (Dini upper directional derivative). Let  $f$  be Lipschitz near  $x \in \mathbb{R}^n$ , the Dini upper directional derivative at  $x$  in direction  $d \in \mathbb{R}^n$  is denoted and defined by

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

□

In particular,  $f^D(x; d) = f'(x; d)$  when the latter exists. Clarke [51, section 2.1, p. 25] uses mostly the following definition, the lim sup also applying to the quantity in  $\mathbb{R}^n$  (and not only the scalar).

**Definition 2.3.7** (generalized directional derivative). Let  $f$  be Lipschitz near  $x \in \mathbb{R}^n$ , the generalized directional derivative at  $x$  in direction  $d \in \mathbb{R}^n$  is denoted and defined by

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

□

Since in (2.12) the  $y$  may be taken equal to  $x$ , one clearly has  $f^\circ(x; d) \geq f^D(x; d)$ . As shown by the following example, the minimum is a function that exposes difference between these directional derivatives.

**Example 2.3.8** (case of the minimum). Let  $f(x) = -|x| = \min(x, -x)$ , which is clearly Lipschitz. Then  $f'(0; d) = -|d|$  and  $f^\circ(0; d) = +|d|$ .

Indeed, (2.10) reads  $\lim_t -|td|/t = \lim_t -|d| = -|d|$ , whereas in the lim sup, taking  $y = -td \rightarrow 0 = x$  yields  $+|d|$ . □

Clarke uses definition 2.3.7 to define a certain differential:

**Definition 2.3.9** (generalized gradient, [51, p. 27]).

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

In particular: [51, proposition 2.1.2, p. 27] for any  $d \in \mathbb{R}^n$ ,  $f^\circ(x; d) = \max\{\zeta^\top d; \zeta \in \partial f(x)\}$ ; [51, proposition 2.1.5, p. 29]  $\zeta \in \partial f(x) \Leftrightarrow f^\circ(x; d) \geq \zeta^\top d$  for all  $d \in \mathbb{R}^n$ ,  $\partial f(x)$  is closed (weak\*-closed without the finite dimension assumption). □

Some notions of stationarity related to these directional derivatives are discussed below in section 2.3.7.

For instance, one has that  $\partial(-|\cdot|)(0) = \partial(|\cdot|)(0) = [-1, +1]$ . In what follows, the “usual” differentials, the differentials in the sense of Clarke, are denoted  $\partial$  or  $\partial_C$  (the differential built with the componentwise Cartesian product of the differentials of  $f_i$  with  $f = (f_i)_i$  is denoted with  $\times$  to avoid the possible confusion). Another type of differentials are the B-differentials, where the “B” stands for Bouligand.

**Definition 2.3.10** (scalar B-differential, [51, p. 63]). The B-differential of  $f$  at  $x$  is denoted and defined by

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

In particular,  $f$  is differentiable at the points of  $x_k$  of the sequence.  $\square$

For instance, one has  $\partial_B(|\cdot|)(0) = \{-1, +1\} = \partial_B(-|\cdot|)(0)$ . As detailed in [51, theorem 2.5.1 p. 63], the points  $x_k$  can also be taken out of a set  $S$  of measure zero (since we consider finite dimension), and a key relation between  $\partial_B$  and  $\partial := \partial_C$  is the following.

**Proposition 2.3.11** (scalar C-differential). *One has the following equality*

$$\partial_C f(x) = \text{conv } \partial_B f(x). \quad (2.15)$$

$\square$

In particular [51, proposition 2.2.7], for Lipschitz convex single-valued functions  $f$ ,  $\partial f(x)$  is the usual convex subdifferential, and  $f^\circ = f'$ . For nonconvex functions, a function  $f$  verifying  $f' = f^\circ$  is said to be regular by Clarke.

**Definition 2.3.12** (regularity [51, p. 39]). The function  $f$  is said to be regular at  $x \in \mathbb{R}^n$  if  $f'(x; d)$  exists for all  $d \in \mathbb{R}^n$  and  $f'(x; d) = f^\circ(x; d)$ .  $\square$

In many contributions, this notion is called “subdifferential regularity”, since “regularity” is already used for conditions which ensure that accumulation points of algorithms are solutions of the considered problem (for instance in [119]).

In particular,  $-|\cdot|$  is not regular at 0. The difficulties of the  $-|\cdot|$  function, related to the minimum by  $-|x| = \min(x, -x)$ , have been observed for instance by Pang, Han and Rangaraj in [197, p. 60]. Some of the definitions can be adapted to vector-valued functions [51, sections 2.2 p. 30 and 2.6 p. 70].

**Definition 2.3.13** (vectorial directional derivative [51, p. 30]). Let  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , the directional derivative of  $F$  at  $x$  in direction  $d$  is denoted and defined by

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

It is said that  $F$  has a Gâteaux derivative if this limit is equal to  $DF(x)d$  for some element  $DF(x) \in \mathbb{R}^{m \times n}$  and every  $d$ .  $\square$

This can be used to define the notion of BD-regularity by Qi [204, p. 232]. Regularity conditions are often used at/around a solution to ensure good local properties of algorithms.

**Definition 2.3.14** (BD-regularity). Let  $F$  be directionally differentiable at  $x$ ,  $F$  is said to be BD-regular at  $x$  if the following condition holds:

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

In particular, this implies  $\|h\| \leq c\|F'(x; h)\|$  for all  $h$  and a constant  $c > 0$ .  $\square$

**Definition 2.3.15** (vectorial B-differential). Let  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  be a Lipschitz function, the B-differential of  $F$  at  $x$  is denoted and defined by

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

□

Once again, taking the convex hull of the B-differential yields the C-differential.

**Definition 2.3.16** (vectorial C-differential). Let  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  be a Lipschitz function, the C-differential of  $F$  at  $x$  is denoted and defined by

$$\partial_C F(x) := \text{conv}(\partial_B F(x)) \quad (2.19)$$

The C-differential is nonempty closed convex compact [51, proposition 2.6.2a-b].

□

The B-differential of definition 2.3.15, while less popular than its convexified counterpart from definition 2.3.16, is used by Qi to define another form of regularity. It is often named “BD-regularity”, acknowledging that Qi initially used that term for definition 2.3.14 (by Qi himself as well!).

**Definition 2.3.17** (strong BD-regularity [204, p. 233]). A function  $F$  is said to be strongly BD-regular at  $x$  if for all  $V \in \partial_B F(x)$ ,  $V$  is nonsingular.

In particular [204, lemma 2.6], strong BD-regularity is a diffusing property, in the sense all Jacobians at neighboring points of  $x$  are also nonsingular, and all the inverses at every neighboring point can be bounded above by a common constant.

□

The following (most often strict) inclusions shall be relevant in the subsequent chapters.

**Proposition 2.3.18** (inclusion in the product-differential, [51, proposition 2.6.2e]).

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

Moreover, by definition,  $\partial_B F(x) \subseteq \partial_C F(x)$  with equality if  $F$  is continuously differentiable.

□

Among the most useful properties, one finds the chain rules. We quote only the one of three versions of Clarke (the other applying for some different cases but in a similar spirit).

**Proposition 2.3.19** (chain rule, [51, proposition 2.6.6, pp. 72-73]). Let  $f = g \circ F$ , where  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is Lipschitz near  $x$  and where  $g : \mathbb{R}^m \rightarrow \mathbb{R}$  is Lipschitz near  $F(x)$  for some  $x \in \mathbb{R}^n$ . Then  $f$  is Lipschitz near  $x$  and one has

$$\partial f(x) \subseteq \text{conv}\{\partial g(F(x))\partial F(x)\}. \quad (2.21)$$

If in addition  $g$  is strictly differentiable at  $F(x)$ , then equality holds (and conv is superfluous).

□

In particular, continuous differentiability implies strict differentiability, which is less restrictive. While the book of Clarke [51] remains primordial work, for the specific case of nonsmooth equations arising from C-functions, some additional notions were introduced based on the work of Clarke. The first one is the B-derivative, used by Robinson in [218, appendix, p. 62 onwards], in the context of studying the local structure of the space around solutions of optimization problems. Actually, thanks to a result by Shapiro [230], in finite dimension, directional derivability and B-differentiability are equivalent.

**Definition 2.3.20** (B-derivative and B-differentiability). Let  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , it is said to be B-differentiable at  $x \in \mathbb{R}^n$  if there exists some function  $BH(x) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , called the B-derivative of  $F$  at  $x$ , verifying  $BH(x)(\lambda v) = \lambda BH(x)(v)$  for  $\lambda \in \mathbb{R}_+$  such that

$$F(x + v) = F(x) + BF(x)(v) + o(\|v\|). \quad (2.22)$$

It is B-differentiable if this property holds for all  $x \in \mathbb{R}^n$ . Thanks to Shapiro's result, (2.22) may also be written (finite dimension is assumed in this thesis)

$$F(x + v) = F(x) + F'(x; v) + o(\|v\|). \quad \square$$

Another crucial notion is semismoothness. These functions are often said to lie between Lipschitz and  $C^1$  functions. Mifflin [171] introduced them for the minimisation of  $f$  with the constraint  $h(x) \leq 0$  but with nonsmooth nonconvex functions.

**Definition 2.3.21** (semismooth scalar function). The function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is semismooth at  $x$  if

- $f$  is Lipschitz around  $x$ ,
- for each  $d \in \mathbb{R}^n$ , any sequences  $\mathbb{R}_+ \ni \{t_k\} \searrow 0$ ,  $\{y_k\} \subseteq \mathbb{R}^n$ ,  $\{g_k\} \subseteq \mathbb{R}^n$  with  $y_k/t_k \rightarrow 0$ ,  $g_k \in \partial f(x + t_k d + y_k)$ , then  $\{g_k^\top d\}$  has exactly one accumulation point.  $\square$

To be applied to nonsmooth systems of equations, one needs a vectorial version of nonsmoothness. This was achieved by Qi and Sun [206, p. 354 onwards].

**Definition 2.3.22** (semismooth vector-valued function). The mapping  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is semismooth at  $x$  if the following limit exists for any  $d \in \mathbb{R}^n$

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

Equivalent definitions are given in [206, theorem 2.3 p. 356].  $\square$

Whether considering scalar or vector-valued functions, the class of semismooth functions includes functions that are smooth, piecewise smooth, convex, is stable by addition, multiplication, composition as shown by Mifflin. Qi and Sun also show a simple property [206, lemma 2.2, p. 356].

**Lemma 2.3.23 ( $F'$  and  $\partial F$ ).** If  $F$  is Lipschitz and  $F'(x; d)$  exists for points  $x$  and  $d$ , then one has  $F'(x; d) = Vd$  for some  $V \in \partial F(x)$ .  $\square$

A strong version of semismoothness also exist: such functions are more “smooth” thus enjoy stronger convergence properties [204, lemma 2.3].

**Definition 2.3.24** (strong semismoothness). The mapping  $F$  is said to be strongly semismooth at  $x \in \mathbb{R}^n$  if the following condition holds

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

$\square$

In particular, most of the C-functions from section 2.3.5 are strongly semismooth. Finally, let us consider a class of functions that “lie between  $C^1$  and  $C^2$  functions” ([65, p. 412]).

**Definition 2.3.25 ( $SC^1$  functions).** The mapping  $F$  is said to be  $SC^1$  if it is  $C^1$  and its gradient is semismooth.  $\square$

Other viewpoints on nonsmooth analysis, employing different notions, exist, such as the theory developed by Mordukhovich [176, 177, 175].

### 2.3.3 First nonsmooth algorithms

In the framework of C-functions we will develop in section 2.3.5, complementarity problems become equations of the form  $H(x) = 0$ . When  $H$  is smooth, a classic method among many others [188] to solve such systems of equations is Newton’s method. When started close enough to the solution, it is known to converge at a quadratic rate if the derivative of  $H$  at the solution is nonsingular. Newton’s equation, which yields iterate  $k + 1$  from iterate  $k$ , reads:

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

Without the smoothness assumption, two questions arise: how to replace the derivative of  $H$  at  $x_k$ , and what ensures fast local convergence since the derivative at the solution may not be defined? Many variants of Newton’s method have been designed for the nonsmooth equations at stake, arising from complementarity problems or related equivalent problems. In this section, we discuss some contributions about those equations, using the tools defined in the previous section.

#### General idea: adapting Newton’s method

Let us start by mentionning Kummer’s observation in [150], where a certain nonsmooth piecewise affine function is constructed and Newton’s method fails even if started infinitely close to the solution. This example justifies the need to take into account the absence of nonsmoothness.

The general framework is as follows [86, section 7.2, p. 638]. First, consider Newton's equation (in the smooth case) in the form

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

and let  $d^k = x^{k+1} - x^k$ . Since the second term may not be properly defined, consider an *approximation*

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

where the quantity  $A$  remains vague for now. Around a point  $\bar{x}$  fixed, if the approximation verifies  $A(x, 0) = 0$  and

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

it is a Newton approximation. If  $O(\|x - \bar{x}\|^2)$  replaces the term  $o(\|x - \bar{x}\|)$ , the approximation is strong (it yields faster convergence). After solving the approximated Newton equation and getting a solution  $d^k$ , one updates  $x^{k+1} = x^k + d^k$ . Many possibilities for  $A$  are discussed in [86], such as inexact solve of the approximated equation, case when  $H$  is a piecewise selection among several (smooth) functions or is a composition...

One particular choice of the approximation  $A(x^k, d^k)$  is given by the following choice:

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

where the differential of  $H$  at  $x_k$  is introduced in definitions 2.3.15 and 2.3.16. This rather natural choice leads however to an important question, further discussed in chapter 6: how to choose the element  $J$  of the differential? Often, assumptions are made so that any  $J$  is suitable (see below).

## Some general improvements

In this section, we briefly mention some techniques that are often used in the contributions evoked below. The first one is the linesearch, an extremely classic method, presented for instance in Armijo's form, named after Armijo's seminal paper [11]. In an iterative algorithm minimizing function  $\Psi$ , assuming a direction  $d^k$  is obtained at iterate  $x^k$ , instead of updating  $x^{k+1} = x^k + d^k$ , one computes the smallest integer  $i$  such that

$$\Psi(x^k + 2^{-i}d^k) \leq \Psi(x^k) + \beta 2^{-i} \nabla \Psi(x^k)^\top d^k, \quad \text{then } x^{k+1} = x^k + 2^{-i}d^k, \quad (2.26)$$

where  $\beta \in (0, 1/2)$  is a fixed constant. Many of the papers below use a modified linesearch, a "nonmonotone" variant introduced by Grippo, Lampariello and Lucidi [113]. Though not initially designed for nonsmooth algorithms, it is applicable nonetheless. Its concept is rather straightforward: instead of using  $\Psi(x^k)$  in the right-hand side of (2.26), one uses  $\max\{\Psi(x^k), \Psi(x^{k-1}, \dots, \Psi(x^{k-m(k)}))\}$  where  $m(k)$  is some function verifying, according to the original paper,  $m(0) = 0$  and  $0 \leq m(k) \leq \min(M, 1 + m(k-1))$ ; in short, we the values of the last few iterates are considered instead of only the current one  $x^k$ . This prevents the algorithm for being forced to have too strong of a descent property, which may lead to small stepsizes, i.e., large  $i$ 's in (2.26) (see [65, p. 431, last paragraph]).

The other main technique we evoke is the classical Levenberg-Marquardt method, for which we present a simple framework. Assuming one wants to minimize  $\|H\|^2/2$  with  $H$  smooth, we use the following notation

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

A descent direction is obtained from the iterate  $x$  by solving

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

where  $\lambda \geq 0$  is the parameter to adapt and  $S \succ 0$  is a symmetric semidefinite positive matrix (often the  $S = I$  the identity). The main difference between this globalization technique and the linesearch is that the parameter  $\lambda$  defined a *curve* of solutions, contrary to a linesearch along a halfline. An iteration of Levenberg-Marquardt can for instance be presented as follows [105, algorithm 19.10].

**Algorithm 2.3.26 (ITERATION OF LEVENBERG-MARQUARDT).** The algorithm uses the following constants:  $0 < \tau_1 < 1 < \tau_2$  for the update of  $\lambda$  and  $0 < \kappa_1 < \kappa_2 < 1$  as satisfaction thresholds for the decrease of  $f$ .

1. *Stopping criterion.* If  $g(x) \simeq 0$ , stop.
2. *Displacement.* Take  $\lambda_0 = \lambda$  and repeat the following operations for  $i \in \mathbb{N}$ .

2.1. Compute the solution  $d_i$  of the linear system

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

2.2 If

$$f(x + d_i) \leq f(x) + \kappa_1 g(x)^T d_i,$$

exit the current loop with  $d = d_i$ , otherwise  $\lambda_{i+1} = \tau_2 \lambda_i$ .

3. *New penalty parameter.* If  $f(x + d) \leq f(x) + \kappa_2 g(x)^T d$ ,  $\lambda_+ = \tau_1 \lambda_i$ , otherwise  $\lambda_+ = \lambda_i$ .
4. *New iterate.*  $x_+ = x + d$ .
5. *New matrix.* Choose  $S_+ \succ 0$ .

A possible update rule is of the form  $\lambda_k = \kappa \|H(x^k)\|^\delta$  for a constant  $\kappa > 0$  and  $\delta \in [0, 2]$ . The method was initially presented by Levenberg [153] and later rediscovered by Marquardt [165]. Other options such as penalization are considered for instance in [183].

## Contributions without C-functions

Here, we discuss some papers which consider nonsmooth equations  $H(x) = 0$  not necessarily arising from CPs and C-functions. Kojima and Shindo [146] discuss the case of  $PC^1$  mappings, for piecewise  $C^1$ , i.e.,  $H$  is selected among a finite number of  $C^1$  functions  $H_i$ , which may include CPs and normal maps. Their algorithm essentially chooses a piece corresponding to  $x^k$  and applies a Newton step with this piece. The algorithm converges

if, at a solution  $z$ , for all pieces containing  $z$ , the Jacobians at  $z$   $H'_i(z)$  are nonsingular and the derivative  $H'_i$  are Lipschitz around  $z$ . They also consider a quasi-Newton variant with Broyden's update, by storing an approximation on each piece.

Let us discuss further the case of (2.25). Qi and Sun [206] use a nonsmooth Newton method of the form

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

Since  $H$  is locally Lipschitz,  $\partial H$  is well-defined. One key property ensuring the local convergence is the fact that nonsingularity is a “diffusing” property.

**Proposition 2.3.27** ([206, proposition 3.1], [204, lemma 2.6]). *Let  $x$  be such that each  $V \in \partial H(x)$  is nonsingular. Then there exists  $C > 0$  and a neighborhood  $N(x)$  of  $x$  such that for any  $y \in N(x)$  and  $W \in \partial H(y)$ ,  $W$  is nonsingular and  $\|W^{-1}\| \leq C$ .  $\square$*

This property is essential since, if  $x^*$  is a solution verifying the assumption of proposition 2.3.27, then in a neighborhood of  $x^*$  the Newton-type equation (2.25) can be solved, so the algorithm is well-defined. If in addition the function is semismooth, the algorithm converges locally. A global convergence property is shown under the assumptions that proposition 2.3.27 and several additional Lipschitz-type inequalities all hold globally.

In [204], Qi replaces  $\partial$  by  $\partial_B$  (still under the assumption  $H$  is Lipschitz and semismooth). The main interest of this change is to reduce the “size” of the considered differential, which thus makes the assumption or strong BD-regularity (“all Jacobians at the solution are non-singular”, definition 2.3.17) weaker and more easily verified. This is one of the main motivations of this thesis, in particular since the article deals with (2.1a) as an application.

Alongside Pang [198], they design a method to show that superlinear convergence holds for an iteration of the form (2.27), which also applies for more general iterations not using a differential. It also requires semismoothness and strong BD-regularity (see definition 2.3.17). They also discuss globalization techniques in a similar ideas to those discussed below in section 2.3.4.

Now, we mention an approach developed by Śmietański [233], which combines (2.25) and a technique in the one-dimensional case

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

Observe this equation reduces to Newton's one by taking  $e^t \simeq 1 + t$ . This becomes, in the  $n$ -dimensional setting,

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

where  $V_k \in \partial_B H(x^k)$ . It converges under the same hypotheses as the classic method without the exponential update, semismoothness of the mapping and strong BD-regularity (definition 2.3.14) at the solution.

Finally, let us mention an algorithm in three substeps by Solodov and Svaiter [235], focused on  $\text{NCP}(F)$  with  $F$  monotone. Each iteration is composed of: an inexact solve of the regularized linearization, then some linesearch on the obtained direction, and finally a projection. Their algorithm does not use a particular merit function but instead the quantity  $(F(x + \cdot), \cdot - x)$  arising from the variational formulation of CPs. The main cost of each iteration is a LCP, though no exact solution is required. The algorithm is globally convergent under the assumptions that  $\nabla F(x^*) \succ 0$ , and that  $\nabla F$  verifies some additional regularity around  $x^*$ . This work is based on a similar paper by the same authors [234], in particular providing an illustration (p. 767) of the intuition underlying the algorithm.

### Reformulation with the minimum

**Definition 2.3.28** (min C-function). The minimum C-function is denoted and defined by

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

It is sometimes written as  $\min(a, b) = a - (a - b)_+ = (a + b - |a - b|)/2$ .  $\square$

Since the minimum is a C-function, to solve CPs one may consider

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

The minimum has been used by many authors [3, 195, 192, 204, 207, 95, 66, 20, 73]. It is the least differentiable C-function but is piecewise linear in its arguments, and is also often called “natural residual” ([243, p. 202], [138, p. 116], [5, p. 1]). This name seems to come from the fact the minimum often serves as an error bound or stopping criterion even for other C-functions.

It seems to us that the Fischer function (and other C-functions inspired by it) have been studied more in-depth than the minimum, in particular since its square (thus the associated merit function) is differentiable. This is particularly relevant for the globalization of local algorithms. Nonetheless, some of its properties such as the finite termination for LCPs (see below) or the presence of the minimum in error bounds are particular motivations to study the minimum below (see also section 2.3.4) and in later chapters.

Let us start with [194], where Pang first states an error bound

$$\|x - x^*\| \leq \kappa \|\min(x, Mx + q)\| \quad (2.29)$$

around  $x^*$  whenever it is a regular solution, which is stated in terms of perturbed LCPs around  $x^*$ . It is also proved (lemma 3 p. 59) that, for  $\text{NCP}(F)$ , the mapping  $\min(x, F(x))$  is Lipschitz with constant  $\sqrt{n} \max(1, \alpha)$  where  $\alpha$  is the Lipschitz constant of  $F$ . An approximate scheme, using a linearization of  $F$  is shown to be convergent under regularity of the solution  $x^*$ . In 1990, Pang [195] uses the notion of B-derivative (or directional derivative, thanks to [230]) to consider the equation  $H(x) = 0$  and its merit function  $\|H\|^2/2$ , solved by the iteration

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

Assuming the mapping  $H$  is differentiable at the solution and the derivative is nonsingular and “strong” (i.e., more regular), local convergence is obtained, and global convergence by applying a linesearch. However, when using (2.30), either the function  $H$  is differentiable at  $x^k$  and this reduces to a traditional Newton step or the iteration becomes a mixed LCP since the equation is not linear in  $d$ . Note that, when applied to NCPs, the subproblems are LCPs with less complementarity indices (compare for instance with Josephy’s approach [134] evoked in section 2.2.4). Thus, either further hypotheses are required to ensure simple solvability of the iterations or one must accept to solve mixed LCPs.

Pang’s work leads to the algorithm called “Newton-min” to solve the nonsmooth equation  $H(x) = \min(x, F(x)) = 0$ . At each iteration, it consists in solving a simpler version of (2.30) by splitting the indices for which  $x_i = F_i(x)$ , which (may) cause  $H$  to be nondifferentiable, into the  $x$  or the  $F$  part of the system, to obtain a linear system to solve. A sketch is given below.

**Algorithm 2.3.29 (NEWTON-MIN).** Consider a starting point  $x^0 \in \mathbb{R}^n$ .

1. *Stopping criterion.* If  $H(x^k) = 0$ , stop.
2. *Index decomposition.* Define the following index sets:

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

Let  $(I^k, J^k) \in \mathfrak{B}(\beta(x^k))$  be a partition of  $\beta^k$ , then let  $\bar{\alpha}^k := \alpha(x^k) \cup I$  and  $\bar{\gamma}^k := \gamma(x^k) \cup J$ . Solve the linear system with variable  $d$ .

$$\begin{cases} x_{\bar{\gamma}^k}^k + d_{\bar{\gamma}^k} = 0 \\ F(x^k)_{\bar{\alpha}^k} + F'(x^k)_{\bar{\alpha}^k} d = 0 \end{cases} \quad (2.31)$$

3. *Update.* Let  $x^{k+1} = x^k + d^k$  with  $d^k$  the solution of (2.31).

Clearly, (2.31) can be reduced to a system with  $|\bar{\gamma}^k|$  variables by the direct substitution of  $d_{\bar{\gamma}^k}$ . Its local convergence is related to questions of regularity discussed below, and the class of matrices  $M$  for which the Newton-min algorithm converges has been studied by Ben Gharbia and Gilbert in [23, 24]. While it is further discussed in chapter 6, let us mention that the allocation of the indices in  $\beta(x^k)$  is not innocuous: “wrong” choices may lead to  $\|H(x^k + d^k)\| > \|H(x^k)\|$ , see [20, example 5.8].

One particular property of the algorithm 2.3.29 based on the minimum reformulation was first observed by Fischer and Kanzow [95]: for LCPs with a suitable regularity assumption, we get finite termination. According to [86, p. 853], such property cannot be obtained for the Fischer C-function. More precisely, let  $x^*$  be a solution of (2.1c), and define

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

If for any bipartition  $(I, J) \in \mathfrak{B}(\beta^*)$ , the matrix

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

is nonsingular, the number of iterations is finite (see also [86, theorem 7.2.18 p. 660]). They consider two regularities for the LCP.

**Definition 2.3.30** (regular point for  $\text{LCP}(M, q)$  [95, def 2 p. 281]). A point  $x^*$  is said to be  $b$ -regular if  $M_{\delta, \delta}$  is nonsingular for  $\alpha^* \subseteq \delta \subseteq \alpha^* \cup \beta^*$ . It is said to be  $R$ -regular (for Robinson [220]) if

$$\det(M_{\alpha^*, \alpha^*}) \neq 0 \text{ and } M_{\beta^*, \beta^*} - M_{\beta^*, \alpha^*} M_{\alpha^*, \alpha^*}^{-1} M_{\alpha^*, \beta^*} \in \mathbf{P}. \quad \square$$

While these properties are relevant at a solution  $x^*$ , some more general properties hold at any point; they are summarized in the table below.

condition	property	reference (in [95])
$M \in \mathbf{P}$ (for any $x$ )	all $V \in \partial_C H(x)$ are nonsingular	theorem 5, p. 284
$M \in \mathbf{ND}$ (for any $x$ )	all $V \in \partial_B H(x)$ are nonsingular	theorem 6, p. 285
solution $x^*$ is $R$ -regular	all $V \in \partial_C H(x^*)$ are nonsingular	theorem 9, p. 286
solution $x^*$ is $b$ -regular	all $V \in \partial_B H(x^*)$ are nonsingular	theorem 10, p. 287

Table 2.1: Summary of the regularity properties for the LCP.

This is related to the work of Qi [204] on the semismooth Newton method with the B-differential, since it requires weaker assumptions on the differential of the mapping  $H$  at the solution, because there are less elements in it, than its counterpart with Clarke's differential. Similarly, let us mention the B-differential of the Fischer function contains more elements than the one of the minimum [132, p. 151].

The notion of regular point (not the one of definition 2.3.12) for  $x \mapsto \min(x, F(x))$  is defined as follows [195, definition 2 p. 327], [87, definition 2.1 p. 230].

**Definition 2.3.31** (regular point for  $\text{NCP}(F)$ ). Let  $x \in \mathbb{R}^n$ , define  $\alpha := \{i \in [1 : n] : F_i(x) < x_i\}$  and  $\beta := \{i \in [1 : n] : x_i = F_i(x)\}$ . A point  $x$  is said to be a  $b$ -regular vector of  $\min(x, F(x))$  if, for  $\delta \subseteq \beta$ , the matrix  $\nabla F(x)_{\alpha \cup \delta, \alpha \cup \delta}$  is nonsingular. It is said to be a  $R$ -regular vector if  $[\nabla F(x)]_{\alpha, \alpha}$  is nonsingular and its Schur complement in  $\alpha \cup \beta$ ,  $\nabla F(x)_{\beta, \beta} - \nabla F(x)_{\beta, \alpha} [\nabla F(x)_{\alpha, \alpha}]^{-1} \nabla F(x)_{\alpha, \beta}$  is a  $\mathbf{P}$ -matrix (see definition 2.2.1).  $\square$

Other types of regularity, named semistability and hemiregularity, also exist [30]. Numerically, Pang [195] suggests, to ease the computational cost, to employ an inexact method or a least-squares approach. Pang also adapts this scheme to variational inequalities in [192], by using the componentwise minimum on the related “optimality conditions”. See also section 2.3.4 for a more specific discussion on some contributions. These results were improved in the aforementioned paper of Qi and Sun [206, p. 362, penultimate paragraph].

The differential-based approximation was further investigated in [204], where the B-differential of definition 2.3.15 replaces Clarke's differential. Under strong BD-regularity from definition 2.3.17, *any* element of the B-differential suffices for the algorithm, which is why in the “Final Remarks” (p. 243), only one element is computed. In [255], a specific element of  $\partial H$  is used. This question is further developed in chapter 6. Qi also derives a convergence result without semismoothness and strong BD-regularity by replacing these hypotheses with conditions of sufficient descent (theorem 5.1). An algorithm mixing the differential-based method and the B-differentiable-based method, thus named “hybrid”, is shown to be globally convergent.

### 2.3.4 Particular treatment of the minimum

This section aims at discussing a particular method or type of methods arising with the minimum reformulation. We have observed it in articles by Pang [192], then Han, Pang and Rangaraj [119, 197] or Pang and Gabriel [196], Qi and Sun [207], as well as the book of Facchinei and Pang [86]. It is also discussed in length in [72], which is one of the starting points of this thesis. For the presentation we consider the problem NCP( $F$ ): recall that when using the minimum C-function (definition 2.3.28), often three subsets of  $[1 : n]$  (we omit “ $i \in [1 : n]$ ” below) of indices arise at a given point  $x$ :

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

In particular  $H(x) = \min(x, F(x))$  is differentiable at  $x$  whenever  $F'(x)_{\beta(x),:} = I_{\beta(x),:}$  or  $\beta(x) = \emptyset$ . However, observe  $F_i(x) = x_i > 0$  violates the complementarity but  $F_i(x) = x_i < 0$  violates the complementarity *and* the nonnegativity, which can explain a different treatment for both index sets. A similar comment can be made for indices in  $\alpha(x)$  and  $\gamma(x)$ . Though Pang considers variational inequalities, we state his formalism adapted to CPs, where closely related issues are evoked in [192, section 4, p. 109]. First, define

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

then group these subsets of  $\alpha(x)$  and  $\gamma(x)$  with  $\beta(x)$ :

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

In [72], the authors use a similar idea, except  $\alpha_-(x)$  and  $\gamma_-(x)$  take only the indices for which  $F_i(x)$  and  $x_i$  are close (with the same inequalities). In Pang's paper [192], this makes some equalities of the subproblem become inequalities, to avoid “reduc[ing] to just one single system of linear equations” (p. 110 line 4). Pang then defines a sufficient conditions for regularity, which reads, in this setting,  $\nabla F(x)_{\alpha_+, \alpha_+}$  nonsingular and

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

where  $\alpha_+ := \{i \in [1 : n] : x_i > F_i(x), x_i > 0\}$  and  $\tilde{\beta}(x) := \{i \in [1 : n] : x_i \leq 0, F_i(x) \leq 0\}$ . When compared to definition 2.3.31, we see that  $\alpha(x)$  has become  $\alpha_+(x)$ , and  $\beta(x)$  has increased with a part of  $\alpha(x)$  and a part from  $\gamma(x)$ , the latter being absent from the initial definition of  $R$ -regularity.

Pang, Han and Rangaraj [197] consider a general nonsmooth minimization problem  $\min f(x)$ , with subproblems of the form

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

where  $\psi$  replaces the first-order term (since the gradient may not be defined) and  $B_k$  approximates the second-order term. They obtain global convergence to a Dini stationary point under technical assumptions on  $\psi$  (and  $\alpha I \preceq B_k \preceq \beta I$  for  $0 < \alpha \leq \beta$ ). We now detail their choice of  $\psi$  when minimizing  $\|H\|^2/2$ . It is quite reminiscent of what was detailed for [192], since it employs very similar sets of indices (removing the dependence on  $x$ ).

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

Then, they define

$$\psi(x, d) = H(x)^\top G(x), \quad G_i(x) = \begin{cases} d_i & \text{if } i \in \gamma_{0+}, \\ \nabla F_i(x)^\top d & \text{if } i \in \alpha_{0+}, \\ \min(d_i, \nabla F_i(x)^\top d) & \text{if } i \in \hat{\beta}, \end{cases}$$

which is a modification of  $(\|H\|^2/2)'(x, d)$  by taking the formula of the minimum on a larger set of indices. They obtain global convergence to a solution if the limit point is regular in the sense of definition 2.3.31. Shortly after, the same authors [119] discuss this framework for nonsmooth equations, and also propose a scheme where the second-order term may be defined as desired.

The article of Pang and Gabriel [196] presents a method where the constraint  $x \geq 0$  is added, which simplifies the different intervening index sets. In particular, for the indices  $i$  where  $x_i = F_i(x)$ , the subproblem minimizes the quantity  $(x_i + d_i)^2/2$ . Though the additional constraint makes the notions of regularity slightly more complicated, it leads to convex quadratic subproblems. The method benefits from strong convergence properties, observed on multiple numerical examples, under two types of regularity.

The book of Facchinei and Pang [86] also evokes this technique in a simpler context, modifying only the terms with indices in  $\beta(x)$  [86, section 8.3, p. 767]. They propose to use, as a majorizer of the first-order term,

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

where the second max became a min since  $H_i(x) = x_i = F_i(x) < 0$ . This modification of the terms of “nonnegative equality” actually makes the problem convex (a similar observation is made in [197, p. 72, line 23], where it is said convexity holds if there are no such indices). We shall further discuss this point in chapter 6.

Qi and Sun [207] consider a more general problem of minimizing a nonsmooth function by approximated trust-region subproblems. For NCP( $F$ ), they use the same first-order approximation as described just above. In particular, the algorithm solves convex piecewise quadratic bounded subproblems. Let us propose an explanation of their result when applied to NCP( $F$ ). Indeed, for their algorithm to find a Dini-stationary point (see section 2.3.7 and chapter 6), the approximation  $\psi$  must verify  $\liminf \psi(x, td)/t \leq f^D(x; d)$  which is clearly not verified for the above, since  $\psi$  is positively homogeneous in  $d$  and was built as a majorant of  $f^D$ , with  $f = \|H\|^2/2$ . Therefore, they cannot show directly that cluster points of their algorithm are Dini-stationary (which is the case under the assumption on  $\psi$ ). They show the result holds under the assumption that  $\nabla F(x^*)_{\alpha, \alpha}$  is nonsingular and some matrix (more complicated than the usual Schur complement) is a Q-matrix.

Finally, in [72], the precursory work of chapter 6, the authors discuss improvements of Newton-type directions that may be used in the Newton-min method, algorithm 2.3.29. Recall that this one solves the linear system

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

where  $\alpha \subseteq \bar{\alpha}$  and  $\gamma \subseteq \bar{\gamma}$  with  $(\bar{\alpha}, \bar{\gamma})$  a partition of  $[1 : n]$ . The authors mention an observation from [20] where, for an inappropriate repartition of indices, the obtained direction is not a descent direction for  $\|H\|^2/2$ , which comes from its nonsmoothness. The question of choosing the correct indices (which is related to the question of choosing the correct Jacobian in a differential-based algorithm) is further discussed in chapter 6. They propose modifications of the above system, by replacing some equalities with pairs of inequalities: let  $\mathcal{E}_{\mathcal{F}} \cup \mathcal{E}_{\mathcal{G}}$  be a partition of  $\mathcal{E}(x) := \{i \in [1 : n] : F_i(x) = G_i(x)\}$  and  $\mathcal{E}^-(x) := \{i \in [1 : n] : F_i(x) = G_i(x) < 0\}$ , we search  $d$  such that

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

This particular change is done only for indices corresponding to the “negative kinks”, i.e., such that  $x_i = F_i(x) < 0$ . This method ensures, whenever the polyhedron (since the equations come from linearizations, there are affine constraints to verify) is nonempty, that the direction obtained is a descent direction. It is however conceivable that the systems may not have solutions for instance due to the number of (in)equations that become larger than the dimension. This transformation of “one equality into two inequalities” is also observed in [196]. A version of the system using two inequalities for indices such that  $|F_i(x) - G_i(x)| < \tau$  for some  $\tau \in \mathbb{R}_+^*$  is also used, which avoids employing equalities, for instance for the numerical aspect.

To ensure the polyhedral systems obtained have solutions, some regularity assumptions must be made around the solution. In particular, a Mangasarian-Fromovitz-type constraint qualification must hold for every possible partition that may exist around the concerned point, which is likely to be difficult to verify without a global assumption on the functions. This one considers all possible partitions  $\mathcal{E}_{\mathcal{F}} \cup \mathcal{E}_{\mathcal{G}}$  of  $\mathcal{E}(x) = \{i \in [1 : n] : F_i(x) = G_i(x)\}$

where  $x$  is close to a regular point  $\bar{x}$ , for which one requires

$$\sum_{i \in \mathcal{F} \cup \mathcal{E}_{\mathcal{F}}} \alpha_i \nabla F_i(\bar{x}) + \sum_{i \in \mathcal{G} \cup \mathcal{E}_{\mathcal{G}}} \beta_i \nabla G_i(\bar{x}) + \sum_{i \in \mathcal{E}^-(x)} [\alpha_i \nabla F_i(\bar{x}) + \beta_i \nabla G(\bar{x})] = 0$$

and  $(\alpha_{\mathcal{E}^-(x)}, \beta_{\mathcal{E}^-(x)}) \geq 0$  imply that  $(\alpha, \beta) = 0$ .

By using a linesearch, this method benefits from global convergence, and combining the polyhedral procedure with the classic Newton-min direction (when it is suitable) ensures the good local convergent properties are obtained. Numerically, the algorithm shows very good performance, including on problems randomly generated or coming from applications.

What is discussed in chapter 6 is trying to circumvent these difficult assumptions, replacing the polyhedral system by a least-squares problems. These always have a solution (no hypotheses needed) although naturally have an increased cost.

### 2.3.5 Other nonsmooth methods

#### Glimpse of contributions for the Fischer function

Fischer initially introduced his function for inequality-constrained optimization in [92]. This results in a minor difference in a sign convention since multipliers and constraints have opposite signs whereas in CPs both quantities have same sign. Burmeister is mentioned for defining the function as a “distance to the [nonnegative orthant]”. We may use both names or only Fischer in the rest of the thesis.

**Definition 2.3.32** (Fischer C-function [92]). The Fischer (Fischer-Burmeister) C-function is denoted and defined by

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

□

It may be expressed differently, for instance avoid the loss of numerical precision when subtracting positive quantities [179, p. 306]. The main appeal of the Fischer function is that, while nondifferentiable everywhere, it concentrates the nondifferentiability at the origin, in the sense that it is differentiable outside the origin, but its differential at the origin contains more element than the minimum for instance ([132, p. 151]). Nonetheless, its square is differentiable everywhere, meaning a merit function associated to  $\varphi_{FB}$  is differentiable, which leads to simpler algorithms. Fischer’s initial contribution [92] details algorithmical concerns. This function is “equivalent” to the minimum in the following sense [86, lemma 9.1.3 p. 798]

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

The differential of the Fischer function, when applied to complementarity, was identified in an article by Facchinei and Soares [87, section 3, pp. 232-236] (though Fischer’s paper also

discusses part of it). The mapping  $H(x) = 0$  with the C-function  $\varphi_{FB}$  of definition 2.3.32 is semismooth and the merit function is smooth everywhere. It is also  $SC^1$  provided  $F$  (each  $F_i$ ) is  $SC^1$ . Moreover, when  $x$  is a  $R$ (resp.  $b$ )-regular (definition 2.3.31), all  $J \in \partial_C H(x)$  (resp.  $\partial_B H(x)$ ) are nonsingular and are of the form (up to transposition)

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

where the vectors  $a(x)$  and  $b(x)$  are defined by

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

as long as  $(x_i, F_i(x)) \neq 0$ ; whenever  $x_i = 0 = F_i(x)$ ,  $(a_i(x), b_i(x)) = (\xi_i, \rho_i)$  for any pair  $(\xi_i, \rho_i)$  with  $\sqrt{\xi_i^2 + \rho_i^2} \leq 1$ .

Furthermore (section 4, pp. 236-237), if  $F$  is a  $\mathbf{P}_0$ -function (def 2.2.4), the stationary points of  $\Psi := \|H_{FB}\|^2/2$  are solutions and if it is a uniform P-function (def 2.2.4), the level sets of  $\Psi$  are bounded. Using these properties, they design an hybrid algorithm that computes a direction either by a Newton-min approach or by using  $-\nabla\Psi$ , which is well-defined and a descent direction since the function is smooth. Then, a linesearch procedure may be used.

Fischer also considers the NCP( $F$ ) with  $F$  Lipschitz semismooth but monotone in [93], and develops two algorithms. The first is a descent-based method that avoids computing the derivative of  $F$  and converges under usual assumptions related to the Fischer function. The second is a Newton-type method (inspired from the work of Qi and Sun in the previous sections), and converges to a solution  $x^*$  under a regularity condition on  $F$  at  $x^*$  (slightly different from the previous ones).

The strongly monotone case was for instance analyzed by Geiger and Kanzow in [103], where a descent-based method with linesearch is presented; in particular, it computes only derivatives of the Fischer function but not the derivatives of  $F$  itself, though  $F$  must have Lipschitz Jacobian. When  $F$  is only monotone, they use a linesearch method with second-order information stored in a BFGS fashion.

Facchinei, De Luca and Kanzow [65] propose a “Qi-ed” nonsmooth algorithm that computes a direction by solving a Newton-type equation with an element from  $\partial_B H(x^k)$  or, if it is not satisfactory, uses  $-\nabla\Psi$ . This algorithm generates stationary points of  $\Psi$  (since it is smooth) and if a cluster point is strongly BD-regular and  $F$  is  $SC^1$ , the full sequence converges to it, using the Newton-type direction with unit stepsize, superlinearly (quadratically if  $\nabla^2 F$  is Lipschitz). Note that the stopping criterion used is  $\|\min(x, F(x))\|$  and not a quantity directly based on  $\varphi_{FB}$ . The numerical experiments reported indicate that the alternative direction  $-\nabla\Psi(x^k)$  happens rarely. In [85], Facchinei and Kanzow consider an inexact Levenberg-Marquardt version (also using the opposite of the gradient if needed) with rather similar convergence properties.

In 2000, De Luca, Facchinei and Kanzow [66] compare several algorithms. Each one computes a direction from a Newton-type equation first; if no solution can be found or a descent property does not hold, one backs down to  $-\nabla\Psi(x^k)$ . They compare, giving a convergence theorem for each,

- $H_{FB}(x^k) + J_k d = 0$ ,  $J_k \in \partial H_{FB}(x^k)$ , an exact variant and an inexact Levenberg-Marquardt variant (which always has a solution);
- $H_{\min}(x^k) + J_k d = 0$ ,  $J_k \in \partial H_{\min}(x^k)$ , an exact variant and an inexact Levenberg-Marquardt variant (which always has a solution).

Before launching the algorithms, they also use a crude minimization of  $\Psi_{FB}$  by gradient projections on the nonnegative orthant, which increases the robustness of the algorithm (p. 195) especially for the exact algorithms, which may underperform otherwise. They observe that the exact min-based algorithm requires less runtime than the exact FB-based version (p. 199), which is due to the simpler linear systems produced by the minimum. However, this may be reversed in case function evaluations were costly (section 5.3 p. 202) since the minimum-based versions require more function evaluations during the linesearch: since it is using  $\Psi_{FB}$ , the Fischer reformulation is better tuned than the minimum.

The C-functions  $\min$  and  $\varphi_{FB}$  were also compared by Pieraccini, Gasparo and Pasquali in [200]. They report results on several algorithms using  $\varphi_{\min}$  and/or  $\varphi_{FB}$ . Their observations are as follows: when all algorithms converge, the ones using the minimum tend to require less iterations (p. 379), except for highly singular problems, where the lack of smoothness of the minimum was costly (p. 380). They also mention that the FB merit function may lack efficiency (p. 381).

Liao, Qi and Qi [154] develop a framework minimizing the “energy” represented by  $\Psi_{FB}$  through the means of ODEs and neural networks. A similar work using the minimum function can be found in [5].

In the next parts, we discuss contributions and results on some existing C-functions, though the Fischer function shall return when we evoke some smoothing techniques in section 2.3.6.

### The Kanzow-Kleinmichel C-function

In [137], Kanzow and Kleinmichel propose a family of C-functions.

**Definition 2.3.33** (Kanzow-Kleinmichel C-function [137]). The Kanzow-Kleinmichel C-function is denoted and defined by

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

where  $\lambda \in (0, 4)$  or  $\delta \in (-2, +2)$ .  $\square$

For  $\lambda = 2$  ( $\delta = 0$ ), one obtains the FB function, and for  $\lambda \rightarrow 0$  ( $\delta \rightarrow -2$ ), one recovers the minimum function up to a factor  $-2$ . It clearly benefits similar differentiability properties as  $\varphi_{FB}$ , in particular it is strongly semismooth when  $F$  is differentiable with locally Lipschitz derivative. An overestimate of its Clarke differential is naturally derived from the one of the FB function, as well as regularity properties. The following error bound holds [137, lemma 3.6 p. 241]

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

which, combined with (2.29), means the merit function of  $\varphi_{KK}$  upper bounds  $\|x - x^*\|^2$  when  $F$  is a uniform P-function. A nonsmooth Newton-type method is proposed, with rather similar properties as for the FB function. They report that  $\lambda = 2$  (Fischer) gives better global convergence and  $\lambda \simeq 0$  better local convergence, and propose a version with an updating  $\lambda$  to combine the benefits of both main C-functions.

### The Chen-Chen-Kanzow C-function

The next C-function was introduced by Chen, Chen and Kanzow [43].

**Definition 2.3.34** (C-C-K C-function [43]). The Chen-Chen-Kanzow C-function is denoted and defined by

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

□

Designed to avoid the flatness of  $\varphi_{FB}$  in the positive orthant (see below for a commentary from the Mangasarian-Solodov function), the formulas for its generalized gradient are easily obtained from the one of  $\varphi_{FB}$ , despite being not differentiable on the set  $\{(a, b) \in \mathbb{R}_+^2 : ab = 0\}$ . Its regularity properties are similar to the Fischer-Burmeister or Kanzow-Kleinmichel reformulations.

The authors use nonmonotone a linesearch-based scheme for the globalization. The cluster points of the algorithm are stationary since the merit function is differentiable, and with classic additional assumptions the algorithm converges to solutions. Despite choosing  $\lambda = 0.95$ , so having a function rather close to the Fischer-Burmeister function, they observe particularly good result, likely due to the relevance of improving the landscape and levelsets of the function [p. 25 of the detailed report of the article]: “In fact, we believe that our new NCP function is close to be an *optimal* C-function.”

### The Sun-Qi C-functions

In a similar vein, Sun and Qi [243] proposed the following C-functions.

**Definition 2.3.35** (Sun-Qi variants of the FB C-function [243]). The Sun-Qi C-functions are denoted and defined by ( $\alpha > 0$ )

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

□

These four variants of the FB function share semismoothness, differentiable merit functions and conditions for bounded sublevel sets or solutions at stationary points. Sun and Qi compare the four functions with an algorithm using linesearch: the variants do not perform significantly differently.

### The Fukushima C-function

This C-function, initially designed (only the associated merit function) for variational inequalities, is also smooth. Thanks to this property, Fukushima [98] develops a globally convergent descent method.

**Definition 2.3.36** (Fukushima's gap C-function). The Fukushima C-function is denoted and defined, on the set  $\{(a, b) : a \geq 0\}$ , by

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

□

### The Implicit Lagrangian (Mangasarian-Solodov C-function)

Though initially not stated as a C-function, the implicit Lagrangian of Mangasarian and Solodov [162, 161] is defined as follows, generalizing Fukushima's work.

**Definition 2.3.37** (M-S C-function [162]). The Mangasarian-Solodov C-function is denoted and defined, for  $\alpha > 1$ , by (note that compared to the previous C-function (2.37), no assumption on  $a$  is needed)

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

In particular, it is differentiable when  $F$  is, has vanishing gradient at solutions, and, under a qualification-like condition at a nondegenerate solution, its minimizer is a locally unique global minimum, in addition to being sharper around them (see figure 9.1 p. 797 in [86]). It leads in particular to the possibility of a standard Newton method on  $\nabla M = 0$ . Yamashita and Fukushima [256] show that  $\nabla F$  positive definite implies that stationary points of  $M(\cdot, \alpha)$  are solutions. When  $F$  is  $\mu$ -strongly monotone and  $F$  has Lipschitz gradient, they propose a descent-based method. In particular, the derivatives of  $F$  are not used in the method, which makes it derivative-free, and is shown to converge linearly when the Lipschitz assumption holds around the initial iterate and its sublevel set.

Moreover, the implicit Lagrangian benefits from the following error bounds, first expressed by Mangasarian, Solodov, Luo and Ren [155] for some  $x$  close to a solution  $x^*$  and a constant  $\kappa$

$$\begin{aligned} \frac{\alpha - 1}{\alpha} \| \min(x, F(x)) \|_+^2 &\leq M(x, \alpha) \leq (\alpha - 1) \| \min(x, F(x)) \|_+^2 \\ \|x - x^*\|_+^2 &\leq \frac{2\kappa^2\alpha}{\alpha - 1} M(x, \alpha) \\ \|x - x^*\|_+^2 &\leq \frac{2\kappa^2\alpha}{\alpha - 1} \max(1, \|x\|) M(x, \alpha) \end{aligned} \quad (2.39)$$

where the second bound is obtained via Pang's bound (2.29), and the third is a version global for any  $x$ . When  $F$  is  $\mu$ -strongly monotone and  $L$ -locally Lipschitz, one can take  $\kappa = (L + 1)/\mu$ ; this is also expressed by Yamashita and Fukushima.

Subsequently, in [161], the authors designed a method that only derivates the C-function but not the function of the complementarity problem, and guarantees convergence when the function  $F$  in (2.1b) is strongly monotone.

### Generalization of merit functions

Let us mention a particular contribution of Kanzow [135]. In this paper, several (squared) C-functions are regrouped into a single framework, including  $\varphi_{\min}$ ,  $\varphi_{FB}$  and  $\varphi_{MS}$  (see defns. 2.3.28, 2.3.32, 2.3.37). A simple Newton-type algorithm based on solving  $\nabla\Psi(x) = 0$  for a merit function  $\Psi$  is proposed. It is well-defined, in particular the Hessian  $\nabla^2\Psi$  is shown to be positive definite close to solutions. However, this framework comes with rather strong assumptions: strict complementarity / nondegeneracy holds at the solution as well as a linear independence condition on the gradients of the active components of  $x$  and  $F(x)$  in addition to  $F$  having Lipschitz Hessian. Such assumptions are somewhat expected since the minimum (a highly nonsmooth C-function) is also considered.

Now, we consider a contribution by Tseng, Yamashita and Fukushima [246], dealing with the general CP of (1.5). They reformulate an approximate problem as follows, with  $\alpha < 1$

$$\begin{aligned}\tilde{m}_\alpha(x) := & F(x)^T G(x) - G(x)^T P_K(F(x) + \alpha G(x)) - F(x)^T P_{K^*}(G(x) + \alpha F(x)) \\ & + \frac{1}{2\alpha} (||F(x) - P_K(F(x) + \alpha G(x))||^2 + ||G(x) - P_{K^*}(G(x) + \alpha F(x))||^2).\end{aligned}$$

The framework they develop generalizes the regularized gap of Fukushima (see below) and the implicit Lagrangian of Mangasarian and Solodov. This function benefits from many interesting properties such as differentiability.

### Specific methods for variational inequalities

Let us mention some contributions focused on variational inequalities, problems with the following form: find an  $x^*$  such that

$$x^* \in S, \quad (F(x^*), x - x^*) \geq 0, \quad \forall x \in S.$$

Fukushima [98] studies the following function, related to the merit function called “regularized gap” later.

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

From that formulation, a descent algorithm with linesearch is proposed, under the assumption the mapping  $F$  is monotone. Later, Majig and Fukushima [158], without the assumption of  $F$  monotone, propose a Josephy-Newton-type method, solving the VI with  $F$

linearized around the iterate  $x^k$ . Since even with the linearization the subproblem remains difficult, they restrict the variable point in the VI to a ball around  $x^k$ . In addition, a line-search technique is employed as well as projection on the set  $S$  if the Josephy-Newton part fails to provide a satisfying direction. The convergence relies on the positive definiteness of the Jacobian of  $F$  at the solution. The authors propose another algorithm, based on an evolutionary technique, meaning a population of candidate points is studied and modified alongside the iterations.

From the regularized gap of (2.40), one can consider the “D(ifference)-gap”  $f_\alpha(x) - f_\beta(x)$  for  $\alpha < \beta$ . It was for instance analyzed by Sun, Fukushima and Qi [242], where they propose a second-order approximation for this smooth function that is not twice differentiable. Observe that the regularized gap of (2.40) requires a projection on the set  $S$ , which must be somewhat tractable. When  $S = \{x : h(x) \leq 0\}$ , a constant rank constraint qualification (on the gradients of active constraints) can be used to analyze the projection. Then, a generalized Newton-type method is presented, which is based on the assumptions that the constraint qualification holds at the solution and all the elements in the generalized Hessian are nonsingular. It is then globalized by trust-regions, which converges to a solution when  $F$  is strongly monotone and either Lipschitz or  $S$  compact. Moreover, when the elements of the generalized Hessian are positive definite, the speed of convergence is super-linear/quadratic. The D-gap is used by Fukushima and Kanzow [136] to design an hybrid algorithm that computes a direction by a Newton-type equation or defaults to the opposite of the gradient to verify some descent property. The particular case with  $\beta = 1/\alpha > 1$  was studied by Peng [199], who derives some upper and lower bound on the merit function.

In a similar approach, Polak and Qi [202] introduce the notion of Newtonian operator, which aims at generalizing a second-order derivative for the merit function of a VI (or a nonsmooth equation) when these are not properly twice differentiable. From there, they adapt Newton’s method to their Newtonian operators; convergence is obtained assuming all the elements in the operator at the solution are nonsingular.

The paper of Ito and Kunisch [131] also discusses nonsmooth equations arising from VIs, especially with box constraints:

$$l \leq x^* \leq u, \quad (F(x^*), x - x^*) \geq 0 \quad \forall x \in [l, u].$$

Let us observe that assumption A.4 on p. 351 is that the merit function is “subdifferentially regular” or regular in the sense of Clarke (definition 2.3.12); recall that the (componentwise) minimum does not verify this property, which is an issue with regard to the illustrations they give (equations (1.3) and (1.5) p. 348). Then, they discuss a method based on index sets reminiscent of the PDAS or Newton-min algorithm 2.3.29. Indeed, as observed in a paper alongside Hintermüller [124], both methods may lead to the same iterates (section 2). They also discuss the infinite-dimensional situation, coming from other fields like optimal control applications coming from obstacle problems.

This viewpoint is also considered by He and Yang [122], where  $F$  is  $T$ -monotone, i.e.,  $(F(v) - F(w), (v - w)_+) \geq 0$ . The problems they study, having box constraints, leads to a framework similar to the ones mentioned below in section 2.3.6, where they use smoothing techniques.

Munson, Facchinei, Ferris, Fischer and Kanzow [179] also discuss the box-constrained

VI, and propose a componentwise double C-function reformulation, depending on the values of  $l_i$  and  $u_i$ :

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

where  $\phi_1$  and  $\phi_2$  are C-functions. They suggest two options:  $\phi_1 = \varphi_{FB} = \phi_2$  and  $\phi_1 = \varphi_{CCK}, \phi_2 = \varphi_{FB}$ . They discuss many numerical aspects such as restarts, regularizing perturbations, least-squares, preconditioning, precision errors, nonmonotone linesearch... and obtain good performance from their algorithm.

### Other notions in nonsmooth analysis

To conclude, let us mention briefly some articles using other notions of nonsmooth analysis, such as the ones introduced by Mordukhovich. Alongside Hoheisel, Kanzow and Phan, in [127], an adaptation of Newton's method is analyzed, where the subproblem is of the form  $-H(x^k) \in DH(x^k)(d^k)$  where  $DH$  is a generalized derivative and the update is  $x^{k+1} = x^k + d^k$ .

Gfrerer and Outrata [104] define an adapted version of semismoothness for sets and set-valued mapping, in order to deal with linearizations of both terms in a generalized equation  $0 \in F(x) + T(x)$ . This allows them to setup a Newton-type method.

### 2.3.6 Smoothing techniques

#### General smoothing

As discussed in the previous sections, there exist many methods to deal with nonsmooth formulations. Here, we mention some references using smoothing, which, in its general form, introduces an additional positive parameter  $\varepsilon$  (also called  $\mu, \tau, \dots$ ), modifies the non-smooth equation (sometimes directly the C-function used) to obtain a system

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

where  $\tilde{H}(x, \varepsilon)$  is differentiable if  $\varepsilon > 0$  and  $h(\varepsilon) = 0 \Leftrightarrow \varepsilon = 0$  (most often,  $h(\varepsilon) = \varepsilon$  and  $\varepsilon$  intervenes as  $\varepsilon^2$  in  $\tilde{H}$ ). Observe that this does not solve the nonsingularity issues at solutions, which, as mentioned before, is not desirable ([86, prop. 9.1.1, pp. 794-795]), but avoids nondifferentiability in the rest of the space.

For instance, Chen, Qi and Sun [47] deal with the following variational inequality, where  $X$  is closed convex:

$$q(x) \in X, \quad (y - q(x))^\top p(x) \geq 0, \forall y \in X.$$

They focus on the case  $X := \{x : l \leq x \leq u\}$ , which leads to the nonsmooth formulation

$$F(x) := q(x) - \text{mid}(l, u, q(x) - p(x)) = 0,$$

$$\text{mid}(a, b, c) := (\text{mid}(a_i, b_i, c_i))_i = \min(b_i, \max(a_i, c_i)) = P_{[a_i, b_i]}(c_i) \quad a \leq b.$$

They directly approximate  $F$  by  $f(x, \varepsilon)$  such that  $\|f(x, \varepsilon) - F(x)\| \leq \mu\varepsilon$ ,  $f$  verifies the “Jacobian consistency property”:

$$\lim_{\varepsilon} \text{dist}(\nabla_x f(x, \varepsilon)^T, \partial_x F(x)) = 0,$$

which means the approximation is suitable for the derivative as well, where  $\partial_x$  is the componentwise product defined in proposition 2.3.18. The smoothing is done by some integral formula which smoothes the mid operator. Then, they discuss a linesearch-based globalization technique, convergent when  $\nabla_x f$  is nonsingular and converging to solutions when  $F$  is strongly regular, i.e., all its Jacobians at the solution are nonsingular (definition 2.3.17).

This “splitting” of  $F$  into  $f$  smooth and  $F - f$  nonsmooth small was used by Chen in [45], where the smoothing verifies the “directional derivative consistence property”

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

In particular [45, lemma 2.3 p. 110], for semismooth mappings  $F$ , “Jacobian consistency” implies “directional derivative consistence property”. This framework is used to develop a quasi-Newton method with Broyden’s update rule. The same author with Nashed and Qi [46] discusses similar issues but in Banach spaces, where “slant differentiability” is used to adapt semismoothness, allowing to deal with infinite dimension arising from nonsmooth PDEs for instance.

In [243], Sun and Qi propose a smoothing of the normal map of the form  $F(z_+) + z - z_+ + \alpha(z_+) \cdot [F(z_+)]_+$  for  $\alpha > 0$ , which possess better levelsets (theorem 5) than the usual  $F(z_+) + z - z_+$  map. Their smoothing function is given by

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

Then, for a smoothing parameter  $u \in \mathbb{R}^n$  ( $n$  variables instead of 1), extend  $\psi$  to  $\mathbb{R}^n$  by  $\psi(u, z)_i = \psi(u_i, z_i)$ . The following function is studied:

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

This mapping is (strongly) semismooth when  $F'$  is locally Lipschitz and differentiable when  $u > 0$ . As for the case of one smoothing variable, the notions or regularity apply to matrices in  $\partial\Psi(0, z^*)$ , though their blockwise structure makes conditions simpler to verify.

Another type of smoothing framework was developed used by Haddou and coauthors. It has been applied to LCPs [190], NCPs [117], MPECs [116, 249, 248] (see section 2.2.7) and

Absolute Value Equations [1, 63] (see section 2.2.6). Consider  $\omega$  a nondecreasing scalar function negative on  $\mathbb{R}_-$ , equal to 0 in 0 and tending to 1 at  $+\infty$ . Examples include  $\omega(t) = t/(t+1)$  or  $\omega(t) = 1 - e^{-t}$ . Define  $\omega_r(t) = \omega(t/r)$ . The complementarity (expressed for NCP( $F$ ) for instance) is replaced by the smooth equation  $\omega_r(x_i) + \omega_r(F_i(x)) = 1$ . Additional conditions are required to ensure this reformulation (and the equivalent ones) have suitable properties, in particular to ensure convergence of the method, which may employ for instance interior-points-like type methods [190].

### Smoothing alongside Levenberg-Marquardt

Now, we discuss some references who employ Levenberg-Marquardt globalization techniques in addition to smoothing methods. Qi [205] also decomposes  $H$  into “smooth + nonsmooth small”, splitting  $H(x) = \min(F(x), G(x))$  into  $H = p + q$ , where, for  $w > 0$  fixed,

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

where the expressions of  $\hat{p}$  and  $\hat{q}$  are given by

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

In particular,  $|q_i(x)| \leq w/4$  and  $p$  is smooth. The derivative of  $p$  is used in a LM-trust region method with global convergence under the assumption that  $q$  does not vary “too fast”.

Zhang and Chen [259] use a smoothing of the Fischer-Burmeister function of the form  $\varphi_\varepsilon(a, b) = a + b - \sqrt{a^2 + b^2 + 2\varepsilon^2}$  and solve the reformulation of the LCP

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

Since the merit function is smooth, cluster points are stationary, which become solutions when  $M$  is a  $\mathbf{P}_0$ -matrix. Assuming the computed sequence is bounded and the algorithm converges to a strictly complementary solution, the rate is quadratic (or superlinear, depending on the LM update). The second assumption may be replaced by assuming the inverse of the mapping is uniformly bounded. In a similar vein, Zhang with Zhang [260] treat the case of NCP( $F$ ).

A variant of the LCP is treated by Tian, Yu and Yuan in [244]: they consider

$$x \geq 0, \quad s \geq 0, \quad x \cdot s = w, \quad F(x, s, y) = 0$$

with  $w \geq 0$  given and  $F(x, s, y) = Px + Qs + Ry - a$ . They use an appropriate modification of a C-function to introduce a weight (for  $w$ ):  $\varphi^c(a, b) = 0 \Leftrightarrow a \geq 0, b \geq 0, ab = c$ . For instance, one can use one of the following, where  $c \geq 0$  and  $\mathbb{Q} := \{3, 5, \dots\}$ .

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

Their algorithm, employing the fourth choice, updates the LM parameter as a multiple of  $\|F(x^k, s^k, y^k)\|^2$  and eventually does some linesearch.

Ma, Tang and Chen [157] discuss NCP( $F$ ) by addressing the minimization of

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

Despite  $t \mapsto (t)_-^2$  being smooth, it is not twice differentiable, thus another smoothing is used, which is twice differentiable and has strongly semismooth gradient:

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

At the price of some computations, they obtain a Levenberg-Marquardt algorithm which possesses local and global convergence assuming the sublevel sets are compact. Their update rules use the function-model ratio similar to trust-region methods and the LM parameter is updated as the norm of the gradient.

### 2.3.7 A comment about complexity

As discussed in the previous parts, complementarity problems are often formulated as systems of (nonsmooth) equations, which may be solved by Newton-type methods. Since smooth problems may already be hard, nonsmooth problems are in general difficult as well. In nonsmooth *optimization*, where the cost function  $f$  or the constraints are nonsmooth, finding solutions is often too demanding. This leads to the research of stationary points, i.e., points verifying some “optimality condition”, often of the form  $0 \in \partial f(x^*)$ , where a certain differential is used (mostly the one of definition 2.3.9 in section 2.3.2). Beck and Hallak [19, p. 57] call this “criticality”, whereas they name stationarity “the lack of feasible directions”. Let us state these properties.

**Definition 2.3.38** (stationarity). Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a Lipschitz function, a point  $x$  is said to be stationary if  $0 \in \partial f(x) = \partial_C f(x)$ .  $\square$

In the definition, the differential used is Clarke’s from definition 2.3.9.

**Definition 2.3.39** (“strong” stationarity). Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a Lipschitz function, a point  $x$  is said to be strongly stationary if for all  $d \in \mathbb{R}^n$ ,  $f'(x; d) \geq 0$ .  $\square$

It is sometimes called “Dini stationarity” (see for instance [197, p. 60]), since this definition makes sense when  $f'$  is defined, thus  $f' = f^D$  which explains the reference to Dini.

Combining the observation that  $f^\circ \geq f'$ , that  $(\mathbb{R} \ni) 0 = 0^\top d$ , and definition 2.3.9, strong stationarity implies stationarity. It is rather clear that in the general setting of non-convex nonsmooth optimization, finding a stationary point is not easy. In chapter 6, we

detail the fact that, in the specific case of the minimum reformulation (which leads to a nonmooth equation and nonsmooth optimization), even verifying strong stationarity at any point may be hard. This may be compared to the work of Murty and Kabadi [182], where many quadratic problems are shown to be NP-complete, and in particular checking the local optimality in smooth nonconvex optimization is also NP-complete. Moreover, in the nonsmooth case as well, stationary points may not be local optima.

Now, we evoke some contributions which discuss some complexity issues in a smooth optimization setting. Though they concern the smooth case, these recent works expose interesting techniques and results. The first pair of papers, by Carmon, Duchi, Hinder and Sidford [39, 40], discusses the difficulty of obtaining an  $\varepsilon$ -stationary point, i.e., some  $x$  such that  $\|\nabla f(x)\| \leq \varepsilon$ . They obtain a lower bound on the worst-case complexity of any algorithm of the form  $c_p \Delta L_p^{1/p} \varepsilon^{-\frac{p+1}{p}}$  where  $f$  is Lipschitz up to order  $p$  with constant  $L_p$ ,  $\Delta$  is a proximity measure of the initial iterate, and  $c_p$  is a constant depending on  $p$ . In particular, they recover for the gradient method  $p = 1$  the rate in  $\varepsilon^{-2}$ , in  $\varepsilon^{-3/2}$  for Newton's method  $p = 2$ , and the general formula  $\varepsilon^{-\frac{p+1}{p}}$  for methods of order  $p$ . Essentially, the idea is to start from Nesterov's complicated function [184],

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

then to tweak progressively the function by successive compositions to make its landscape tortuous enough. In particular, the construction is such that algorithms “discover” the value of exactly one coordinate of the solution at each iteration. Therefore, by choosing the dimension of the space appropriately, the lower bound on the number of iterations can be obtained.

They also present the case of first-order methods [40], for functions which may have (unused in the algorithm due for instance to implementation and practical issues) Lipschitz higher-order derivatives. For instance, when the function has Lipschitz Hessian (second-order), the convergence order in the accuracy  $\varepsilon$  is between  $\varepsilon^{-12/7}$  and  $\varepsilon^{-7/4}$ . This interval becomes  $[\varepsilon^{-8/5}, \varepsilon^{-5/3}]$  for Lipschitz derivatives beyond second-order. More details can be found in the articles and the references therein.

Later, Carmon and Duchi alongside Arjevani, Foster, Srebro and Woodworth present similar questions for stochastic optimization [10].

Let us now evoke some contributions for the nonsmooth case. Jordan, Lin and Zampetakis [133], state that even obtaining a Clarke-stationary point, i.e.,  $0 \in \partial f(x)$ , is not very realistic. This leads to a more feasible problem  $\min\{\|g\| : g \in \partial f(x)\} \leq \varepsilon$ , which is an analogue of the smooth case. Then, since this goal remains difficult, Clarke's differential is replaced by the  $\delta$ -Goldstein differential, denoted and defined by

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

Introduced by Goldstein in [107], it possesses similar properties to Clarke's differential, and also leads to a descent method. They report (see the references therein), to obtain an  $\varepsilon$ -approximate point in the  $\delta$ -Goldstein subdifferential, convergence speeds of order:

$O(\delta^{-1}\varepsilon^{-3})$  for  $f$  directionally differentiable using randomization,  $O(d^{3/2}\delta^{-1}\varepsilon^{-4})$  without using gradients. They also prove that without randomization, algorithms always need a number of iterations proportional to the dimension.

In the nonsmooth nonconvex Lipschitz case, their table 1 regroups a comparison between the deterministic and the randomized case, alongside comments on algorithms using *only* the first-order oracle. In particular, deterministic algorithms using solely the gradient may not converge. The complicated functions involved are different from the ones used in [39], involving minima and/or maxima. They also obtain a complexity bound for a smoothing method, using a smooth replacement of the maximum:

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

which is 1-Lipschitz with has a  $a/2$ -Lipschitz gradient.

The Goldstein differential is also used by Gebken in [102], where descent methods are analyzed and the following convergence speed is obtained: if the algorithm converges to a minimum  $x^*$  verifying  $f(x) \geq f(x^*) + c\|x - x^*\|^p$ , then

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

where  $\varepsilon_k$  and  $\delta_k$  are the constants of the Goldstein differential at step  $k$ . Simple examples and some numerical experiments are presented.

## 2.4 On the combinatorial aspect

This section is planned as follows: first, we explain the main link with the topic of complementarity. Then, we mention some classic references and some important notions of this domain. We finish by a brief mention of oriented matroids and some of their most basic properties, followed by software in discrete geometry and combinatorics and finally algorithms closest to our topic, the identification of chambers.

### 2.4.1 Relation with the previous topics

In the previous sections, we discussed complementarity problems, with an emphasis on nonsmooth methods, especially those arising from the reformulations using C-functions. Let us detail how this can be related, under a certain lens, to a specific problem from computational discrete geometry. We summarize the main elements, discussed below in chapter 3.

- The B-differential of the componentwise minimum mapping may intervene in algorithms solving complementarity problems.
- The computation of the B-differential of the minimum of two affine functions is equivalent to *identifying* the chambers of a centered arrangement of hyperplanes.

- We propose a new approach of the problem of identifying the chambers of an arrangement, which we call the “dual approach”, via Gordan’s theorem of the alternative.
- This dual method is based on the circuits of the (oriented) matroid underlying the arrangement.

As we shall see, computing entirely the B-differential is hard, since a possibly exponential number of elements must be computed. Recall that the Newton-type nonsmooth method requires only one element of the B-differential and not all of them, though the corresponding algorithm must assume suitable hypotheses to be able to use any element and function nonetheless.

However, arrangements being an extremely developed field on their own, efficient algorithms on arrangements are interesting.

#### 2.4.2 Classic references

In their more general statement, when the hyperplanes do not all intersect in a (common) point, arrangements represent the way lines split the plane or planes split the space (dimension 2 or 3). This very general question has been considered as far back as early in the XIX<sup>th</sup> century, with Steiner [239] and Roberts [215]; for instance, the paper of Alexander and Wetzel [7] discusses the case of dimension 3. See also Schläfli [227] (posthumously, among contributions in many other fields), who contributed to the sometimes assumed hypothesis of *general position*, which states

“In an  $n$ -dimensional space,  $k$  hyperplanes intersect in a subspace of dimension  $n - k$ .”

where a negative dimension means the empty set. Sometimes, a configuration *not* in general position is called degenerate. In particular, Schläfli gave the general position bound for arbitrary dimension, which states that, for  $p$  hyperplanes in dimension  $n$ , the number of chambers is upper bounded by

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

General position was also used by Buck in [37] to derive the number of objects of any dimension (intersections of hyperplanes and/or halfspaces) in an arrangement.

Since then, the field has developed and thrived beyond what can be mentioned here. Let us mention some classic books, which cover far evoked than what may be useful in this thesis: Crapo and Rota [61], where they also discuss the case of hyperplanes in a finite field (see [186] for complex numbers), a more recent work by Stanley [237, 238] and its introductory part (based on a MIT course) [236]. Another classic is the book of Orlik and Terao [187]. Edelsbrunner dedicated one to algorithmic aspects [81], see also the book of De Loera, Rambau and Santos [64]. Aguiar and Mahajan [4] propose a more recent treatment a bit more oriented towards current research. See also a survey on the topic by Halperin and Sharir [118].

### 2.4.3 Some specific tools

Some contributions employ a very useful and potent but involved notion called the characteristic polynomial. It is defined as follows: for an arrangement  $\mathcal{A}(H_1, \dots, H_p)$ , let  $L(\mathcal{A})$  be the set of all intersections of any number of hyperplanes, one has

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

where  $\mu$  is the Möbius function of the arrangement, a rather involved function defined recursively on the set  $L(\mathcal{A})$ . This recursive definition means that a direct computation may be difficult. One of the main formula, due to Zaslavsky [257], states that the number of chambers equals  $(-1)^n \chi_{\mathcal{A}}(-1)$ . Also discussed in the aforementioned books, since it encodes plenty of information on the arrangement, it is for instance used by Athanasiadis in [12] to obtain the number of chambers of many “classical” arrangements, by using some shrewd combinatorics reasonings. The code of [35] does in fact compute the characteristic polynomial, by using underlying symmetries to simplify the total computational charge. As the authors explain (p. 1357, third paragraph), if one specifically wants to *identify* the chambers, the options are limited.

Nonetheless, the characteristic polynomial remains extremely useful to analyze some particular arrangements, when the normal vectors to hyperplanes are of the form  $e_i - e_j$  for  $1 \leq i < j \leq p$  [12, 203].

A technique in the field of combinatorics is the deletion-restriction principle. Essentially used in induction reasonings, it decomposes the considered structure into two smaller ones, one with a smaller dimension. It may be reminiscent of Pascal’s formula, since it reads, in a very crude form,

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

Winder applied this principle to the number of chambers in [253], obtaining a formula in terms of the degeneracies of subsets of hyperplanes. Brysiewicz, Eble and Kühne [35] also use this principle in their algorithm.

Efficient algorithms may also be obtained by using randomization techniques [53, 52].

### 2.4.4 Oriented Matroids (and circuits)

Matroids and oriented matroids are abstract notions that generalize linear (in)dependence of vectors. For a given set of objects, it consists in giving the subsets that are independent (or other, see below). While linear (in)dependence is clear with vectors in  $\mathbb{R}^n$ , it may in fact be generalized to more abstract objects. While some of the aforementioned books discuss matroids, since they are a very central concept in combinatorics, we mention two books more “oriented” towards them: the one of Oxley [191] and the classic of Björner, Las Vergnas, Sturmfels, White and Ziegler [28]. A recently updated review on the field may be found in [264]. All the following definitions consider the notations of Ziegler’s

book [263], which is more oriented towards geometry. In what follows, a set of vectors  $V = \{v_1, \dots, v_p\} \subseteq \mathbb{R}^n$  is identified with the matrix  $V = [v_1 \dots v_p]; e_p \in \mathbb{R}^p$  is the vector of size  $p$  composed of 1's. Recall that the support  $\text{supp}$  denotes the indices of nonzero components.

**Definition 2.4.1** (“vectors”). The set of vectors  $\mathcal{V}(V)$  is composed of the (componentwise) signs of the vectors of affine dependencies, i.e.,

$$\mathcal{V}(V) := \{\text{sgn}(z) : z \in \mathbb{R}^p, Vz = 0, e_p^\top z = 0\}. \quad (2.42)$$

□

**Definition 2.4.2** (“circuits”). The set of circuits  $\mathcal{C}(V)$  is composed of the (componentwise) signs of the vectors of minimal dependencies, i.e.,

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

□

The next two definitions are somewhat dual to the previous two.

**Definition 2.4.3** (“covectors”). The set of (signed) covectors  $\mathcal{V}^*(V)$  is composed of the (componentwise) signs of the affine functions of the vectors, i.e.,

$$\mathcal{V}^*(V) := \{\text{sgn}(c^\top V - c_0 e_p^\top) : c \in \mathbb{R}^n, c_0 \in \mathbb{R}\}. \quad (2.44)$$

□

**Definition 2.4.4** (“cocircuits”). The set of (signed) cocircuits  $\mathcal{C}^*(V)$  is composed of the (componentwise) signs of the affine functions of minimal support of the vectors, i.e.,

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

□

**Definition 2.4.5** (oriented matroid). Let  $V = \{v_1, \dots, v_p\} \subseteq \mathbb{R}^n$  a set of vectors with  $\text{span}(V) = \mathbb{R}^n$ . The oriented matroid of  $V$  is a structure given by the set of circuits, the set of vectors, the set of cocircuits and the set of covectors. □

In fact, any of these four quantities can determine the other three, which means only one of them is necessary. The bases, which are the largest independent subsets, may be used too.

**Definition 2.4.6** (bases). The set of bases  $\mathcal{B}(V)$  of bases is composed of the largest independent subsets of  $V$ , i.e.,

$$\mathcal{B}(V) := \{B \subseteq [1 : p] : V_{:,B} \text{ injective, } \text{null}(V_{:,B \cup \{i\}}) = 1 \forall i \in [1 : p] \setminus B\}. \quad (2.46)$$

□

Whether oriented or not, matroids have been studied under many angles. Let us mention some contributions related to our concerns. Minieka proposes algorithms to swap

between bases and circuits [172]. Dósa, Szalkai and Laflamme indicate, under very mild assumptions, the maximal and minimal numbers of circuits and bases [70] (see also chapters 3, 4, 5 and appendix A).

In general, operations on matroids tend to have an exponential complexity. Nonetheless, algorithms in “polynomial incremental time” exist, i.e., the cost of the progress between one element and the next is polynomial (and often short). An example may be found in [229] and in the algorithms defined in section 2.4.6. Contributions discussing complexity of various problems include [141, 143, 166]. Matroids, alongside hyperplanes, can be used for various applications such as the analysis of the efficiency of a set of cameras [213].

Let us conclude this brief glimpse of oriented matroids by stating that they intervene in chapters 3 and 5 (and their complements chapters 4 and A). The dual approach stems from the dual nature of two early theorems of convex analysis, Gordan’s alternative [108] and Motzkin’s alternative [178] (though other names or formulations may be more appropriate).

#### 2.4.5 Algebra software

There exist many software options dedicated to algebra which can, among many other functionalities, deal with matroids and arrangements. Most often, since the softwares deal with much more than only obtaining the chambers, we have not compared with numerical results from the different software; moreover, comparing a research code and a fully implemented software is not always meaningful.

First, SageMath [68] is a large-scope software which can deal with matroids or arrangements, among many other functionalities. The main page can be found [here](#), and the documentation [here](#).

Then, created in 1992, Macaulay2 is a more specialized software emphasizing on research in geometry and algebra (main page) [111]. It possesses packages that treat arrangements or matroids.

Next, polymake [101] is more focused on geometry, as suggested by its name. It is able to make computations on matroids; a package dedicated to hyperplanes was designed by Kastner and Panizzut, documented in [140].

To focus a bit more on matroids, let us mention the dedicated project by Kingan and Kingan, Oid [142], and another software, by Rambau, TOPCOM (Triangulations Of Points Configurations and Oriented Matroids). Besides the main page [214], it recently served to solve some specific problems [212].

Finally, the OSCAR project is coded in Julia. It aggregates many packages and previous tools (such as polymake for instance). It was used in [35], alongside other related packages (see the references therein), to get some new combinatorial results.

### 2.4.6 Specific algorithms to identify the chambers

Overall, computing the chambers of an arrangement is #P-hard [250, chapter 6]<sup>1</sup>. This is not very surprising since most problems in combinatorics require the enumeration of a list of (possibly) exponential size. Nonetheless, many algorithms have been designed for the myriad of problems existing.

For instance, in [83], Edelsbrunner, O'Rourke and Seidel present an algorithm with theoretical optimal complexity to build an arrangement. It is rather involved, but constructs completely the arrangement, by adding the hyperplanes one by one in an incremental fashion. It consists in a topological sweep of the space, a widespread technique that scans the space to find intersections (easily since it is done by linear algebra) between the sweeping plane and the current construction. Their elaborate algorithm is (p. 361, second paragraph of section 5) better than the a previous similar sweeping-plane algorithm, the one of Bieri and Nef [27], which also uses a sweeping plane but in an inductive fashion on the dimension. In the plane, i.e., in dimension 2, instead of a line (a hyperplane), one may use a curve line [82]; the authors give some applications and also adapt the method to be used in higher dimension.

Another very interesting algorithm uses the underlying property stating the graph of chambers is connected (see chapter 3). Therefore, one may wonder if there is a way to explore the graph of chambers. However, since the nodes are unknown (we want to identify them!), one requires a special type of exploration. One answer to that question is the reverse search, an algorithm introduced by Avis and Fukuda in [13, 14]. The idea is as follows: consider the chamber containing the origin, then test its potential neighbors by trying to go on the other side of one hyperplane, then recursively call the procedure when a chamber is found. Among other properties, this algorithm can be done in a way it does not store all the chambers, so requires limited memory (it is “compact”), is “incremental-time polynomial”, meaning finding one additional chamber is done in polynomial time, and output-sensitive, meaning the computational complexity is bounded above by a polynom in the number of chambers. Its nature makes it also easily implementable in parallel. By improving on one of the components, Sleumer [232, 231] decreased the required complexity.

Reverse search also appears in other combinatorial problems, see for instance the applications to triangulations, circuits and cocircuits in [212]. It is related to the simplex method in the sense the algorithm moves from a vertex to another (the vertices of a polytope form a connected graph) but without computing them explicitly in advance. See [34] for some examples, or [97] for an application to the computation of the sum of polytopes. A new approach, though maybe not useful for the enumeration of chambers but proposing a way to build Hamilton paths can be found in [170].

The final algorithm we discuss was designed by Rada and Černý [208]. It is rather simple: with one hyperplane, the space is split into its two half-spaces. Now, consider the second hyperplane, check if the two half-spaces of the first hyperplane are themselves split in two by the newly added plane, and so on. The authors have shown that, after some refor-

---

<sup>1</sup>This reference discusses only a rather specific type of arrangements, which can then be generalized to arbitrary arrangements; on an online post, Timothy Chow confirmed this complexity.

mulation, their algorithm is better than the reverse search (even with Sleumer’s speedup). It benefits from the same properties, albeit having a narrower scope. It is discussed in length, as well as some tuning of it, in chapters 3 and 5.

### 2.4.7 A few examples of applications

Let us finish this section on arrangements by evoking some situations where algorithms computing the chambers of an arrangement can be useful. Some additional elements are discussed in chapter 3 section 3. First, there is one “dual” use: computing the vertices of zonotopes, particular polytopes which shall intervene in chapter 6.

**Definition 2.4.7** (zonotopes). Let  $V = \{v_1, \dots, v_p\} \subseteq \mathbb{R}^n$  be a collection of vectors and let  $c \in \mathbb{R}^n$  be an additional vector. The zonotope defined by  $V$  and  $c$  is the polytope denoted and defined by:

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

In this expression,  $v_i[-1, +1]$  is the segment  $[-v_i, +v_i]$ ,  $V$  is the matrix  $[v_1 \dots v_p]$  and  $V[-1, +1]^p$  is the compact expression of the sum. The zonotope  $Z(V, c)$  is centrally symmetric around its *center*  $c$ , which reads:

$$z \in Z(V, c) \iff 2c - z \in Z(V, c).$$

□

Zonotopes are (surprisingly) useful objects in some fields of research. They can be used in control, where the possible states of the system may form a zonotope, updated from one time-step to another with the possible variations of the system. Since they are affine images (often linear with  $c = 0$ ) of unit hypercubes, they possess rather interesting combinatorial properties. Some elements are given in appendix B, otherwise see the book of Grünbaum [114], the one of Ziegler [263] or dedicated articles, such as the one from McMullen [167]. Let us briefly mention the relation between arrangements and zonotopes (see corollary 7.17, section 7.3, p. 205 of Ziegler’s book with some slightly different notations; we name “faces” the faces of all dimensions).

**Proposition 2.4.8** (link between zonotopes and arrangements). *Let  $V \in \mathbb{R}^{n \times p}$  be a matrix and  $Z(0, V)$  be the associated zonotope. The sign vectors of the faces of  $Z(0, V)$  are in bijection with the sign vectors of the objects in the arrangement formed by the vectors of  $V$ .* □

A surprising use of zonotope can for instance be found in the convex maximization of a convex quadratic over the hypercube (using either the  $\{0, 1\}$  or the  $\{-1, +1\}$  hypercube, which differ only by an affine change of variables) under a hypothesis of fixed rank, where the combinatorial nature can be solved by zonotopes [89]. In light of proposition 2.4.8 above, applications of zonotopes are thus related to arrangements.

Another application is the estimation of the rank in robust regression, where the question of ordering of residuals leads to a specific form of arrangements [42].

Arrangements are also related to threshold functions, a notion inbetween theoretical computer science and mathematics:  $f(x) = \text{sgn}(a_0 + a^\top x)$ , with  $x \in \{-1, +1\}^n$  for some  $a_0 \in \mathbb{R}$  and  $a \in \mathbb{R}^n$ . See for instance [253, 130]. Similar considerations are treated in [251].

In [16] for instance, the authors view the above expression of  $f$  with the evaluation of a neuron on the input  $x$ , which relates our topic with neural network and deep learning, when we restrict the evaluation function to be linear (for general overview on deep learning itself, see the celebrated survey of Schmidhuber [228]).

Some specific types of arrangements have also received dedicated papers, such as arrangements with hyperplanes of the form  $H_{ij} := \{x : x_i \pm x_j = \dots\}$ , see for instance [12, 203]. The *resonance* arrangements, with hyperplanes of the form  $H_v := \{x : c^\top x = 0\}$  for  $c \in \{0, 1\}^n \setminus \{0\}$ , are related to quantum theory [148, 35]. For further details, see the articles mentioned and their references.



# Chapter 3

## B-differential of the minimum of two vectorial affine functions

This chapter is composed of an article published in Mathematical Programming Computation [77]. It describes a specific question of nonsmooth analysis, the computation of the full B-differential (see definition 2.3.15) of the componentwise minimum of two affine functions. This question comes from the study of the Newton-min algorithm 2.3.29, described at section 2.3.3.

We show that this seemingly restricted question of nonsmooth analysis is equivalent to many other problems, including (centered) hyperplane arrangements. This link with the fields of combinatorics and computational geometry is useful to design algorithms that answer numerically. Several improvements on a state-of-the-art algorithm are proposed and benchmarked.

This chapter goes along chapter 4, which contains details such as proofs or additional comments, as well as complements on neighboring questions.

Furthermore, for harmonization purposes with the rest, the references are unified with those of the thesis, and are thus not added at the end of the article (some dates of references such as literature classics may differ from the published version). Similarly, page layout, fonts and font sizes are different.

Note: The Université de Sherbrooke asks that, for inserted articles, the contribution of the doctoral student is precised. This topic was mentioned in a list of questions to consider, that I shortly studied before we worked on it together. Most of the work was done collaboratively, with very frequent discussions to coordinate contributions and viewpoints, the code and the redaction part mostly done by my advisors. This published paper was written jointly through a Git repository, so that all authors could edit and contribute.

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

Jean-Pierre Dussault<sup>1</sup>, Jean Charles Gilbert<sup>2</sup> and Baptiste Plaquevent-Jourdain<sup>3</sup>

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  $H : \mathbb{E} \rightarrow \mathbb{F}$  is the set denoted and defined by

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

---

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

<sup>2</sup>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

<sup>3</sup>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

$$\partial_C H(x) := \text{co } \partial_B H(x), \quad (3.2)$$

where  $\text{co } S$  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  $H : \mathbb{R}^n \rightarrow \mathbb{R}^m$  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

$$H(x) = \min(Ax + a, Bx + b), \quad (3.3)$$

where the minimum operator “min” acts componentwise (for two vectors  $u, v \in \mathbb{R}^m$  and  $i \in [1 : m] := \{1, \dots, m\}$ :  $[\min(u, v)]_i := \min(u_i, v_i)$ ). 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

$$H(x) = 0 \quad (3.4)$$

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

$$0 \leqslant (Ax + a) \perp (Bx + b) \geqslant 0. \quad (3.5)$$

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  $(Ax + a)^\top (Bx + b) = 0$  (the superscript “ $\top$ ” 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  $\alpha$  and  $\beta$ ,  $\min(\alpha, \beta) = 0$  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  $\tilde{H} : \mathbb{E} \rightarrow \mathbb{R}^m$  is defined at  $x \in \mathbb{E}$  by

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

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

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  $H_i : x \in \mathbb{E} \rightarrow H_i(x) \in \mathbb{R}$ , for  $i \in [1 : m]$ , which is trivial in the present context. Indeed, it is known that [51, proposition 2.6.2(e)]

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

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  $i$ th 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  $I \cap J = \emptyset$ . 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  $\text{sgn} : \mathbb{R} \multimap \mathbb{R}$  defined by  $\text{sgn}(t) = \{1\}$  if  $t > 0$ ,  $\text{sgn}(t) = \{-1\}$  if  $t < 0$  and  $\text{sgn}(0) = [-1, 1]$ . We note  $\mathbb{R}_+^n := \{x \in \mathbb{R}^n : x \geq 0\}$  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  $\text{vect}(S)$  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  $u \cdot v \in \mathbb{R}^n$  whose  $i$ th 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  $\text{rank}(A) := \dim \mathcal{R}(A)$  and its *nullity* is  $\text{null}(A) := \dim \mathcal{N}(A) = n - \text{rank}(A)$  by the rank-nullity theorem. The  $i$ th row (resp. column) of  $A$  is denoted by  $A_{i,:}$  (resp.  $A_{:,i}$ ). Transposition operates after a row/column selection:  $A_{i,:}^\top$  is a short notation for the column vector  $(A_{i,:})^\top$  and  $A_{:,i}^\top$  is a short notation for the row vector  $(A_{:,i})^\top$ . For a vector  $\alpha$ ,  $\text{Diag}(\alpha)$  is the square diagonal matrix with the  $\alpha_i$ 's on its diagonal.

## 3.2 Background

Recall that  $F : \mathbb{E} \rightarrow \mathbb{F}$  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  $g : \mathbb{E} \rightarrow \mathbb{R}$  be two functions and  $h : \mathbb{E} \rightarrow \mathbb{R}$  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:

$$\mathcal{A}(x) := \{i \in [1 : m] : (Ax + a)_i < (Bx + b)_i\}, \quad (3.8a)$$

$$\mathcal{B}(x) := \{i \in [1 : m] : (Ax + a)_i > (Bx + b)_i\}, \quad (3.8b)$$

$$\mathcal{E}(x) := \{i \in [1 : m] : (Ax + a)_i = (Bx + b)_i\}, \quad (3.8c)$$

$$\mathcal{E}^=(x) := \{i \in \mathcal{E}(x) : A_{i,:} = B_{i,:}\}, \quad (3.8d)$$

$$\mathcal{E}^\neq(x) := \{i \in \mathcal{E}(x) : A_{i,:} \neq B_{i,:}\}. \quad (3.8e)$$

To simplify the presentation, we assume in the sequel that

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

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

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

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

**Proposition 3.2.2** (superset of  $\partial_B H(x)$ ). *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,  $|\partial_B H(x)| \leq 2^p$ .*

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  $n = 2, m = 2, A = \begin{pmatrix} -1 & 1 \\ -1 & -1 \end{pmatrix}, B = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$  and  $a = b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ . One has  $\mathcal{A}(0) = \mathcal{B}(0) = \emptyset, \mathcal{E}(0) = \mathcal{E}^\neq(0) = \{1, 2\}, \partial_B H(0) = \{A, B\}$ , while  $\partial_B^\times H(0) = \{A, B, \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}\}$ . This example also shows that all the Jacobians of  $\partial_B H(0)$  can be nonsingular, while the Jacobian  $\begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$  of  $\partial_B^\times H(0)$  is singular and the central Jacobian (3.41), namely  $\frac{1}{2}(A + B) = \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} \in \partial_C H(0)$ , 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].  $\square$

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  $|\partial_B H(x)| = 2^p$ .  $\square$

**Definitions 3.2.5** (symmetry in  $\partial_B H(x)$ ). 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  $\partial_B H(x)$ .  $\square$

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  $v_1, \dots, v_p$  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  $\Xi$  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  $\Xi$  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,  $\xi_1$  maintains strict the inequalities that are strict with  $\xi_0$ ). 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  $\xi_1$  than with  $\xi_0$ . Pursuing like this, one can finally obtain a vector  $\xi$  in  $\Xi$ . This vector is arbitrarily close to  $\xi_0$  by taking the  $\varepsilon_i$ 's positive and sufficiently small.  $\square$

### 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  $H : \mathbb{R}^n \rightarrow \mathbb{R}^m$  defined by (3.3).  $\square$

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

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

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

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

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  $\mathcal{A}(x) = \mathcal{B}(x) = \mathcal{E}^=(x) = \emptyset$  and  $\mathcal{E}(x) = \mathcal{E}^\neq(x) = [1 : 3]$ , so that  $p = 3$  and

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

### 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

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

□

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  $s \cdot (V^\top d) > 0$  are given by the corresponding columns of the matrix  $D$ :

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

The sign vectors  $\pm e := \pm(1, 1, 1)$  are not in  $\mathcal{S}$  since  $Ve = 0$  (there is not  $d_\pm$  such that  $(\pm e) \cdot (V^\top d_\pm) > 0$ , since this would imply that  $0 < \pm e^\top V^\top d_\pm = 0$ , 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:

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

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,  $\sigma$  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

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

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

**Proposition 3.3.4** (bijection  $\partial_B H(x) \leftrightarrow \mathcal{S}$ ). *Let  $H : \mathbb{R}^n \rightarrow \mathbb{R}^m$  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  $\sigma$  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  $\sigma(J) \cdot (V^\top d) > 0$ .
- 2) If  $s \in \{\pm 1\}^p$  and  $\exists d \in \mathbb{R}^n$  such that  $s \cdot (V^\top d) > 0$ , then  $\sigma^{-1}(s) \in \partial_B H(x)$ .
- 3) Let  $J \in \partial_B^\times H(x)$ . Then,  $J \in \partial_B H(x) \iff \sigma(J) \cdot (V^\top d) > 0$  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  $\sigma : \partial_B H(x) \rightarrow \mathcal{S}$ . 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

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

[ $\subseteq$  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  $s \cdot (V^\top d) > 0$ . By  $J \in \partial_B H(x)$ , there exists a sequence  $\{x_k\} \subseteq \mathcal{D}_H$  converging to  $x$  such that

$$H'(x_k) \rightarrow J. \quad (3.15b)$$

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  $k \in \mathcal{K}$ :

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

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

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

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  $\sigma$ ,  $(J_{i,:} = A_{i,:}$  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  $i \in [1 : p]$ :

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

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

$$s \cdot (V^\top d) > 0. \quad (3.15e)$$

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,  $x_k \rightarrow x$ . 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  $i \in \mathcal{E}^\neq(x)$ , 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

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

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  $s_i = -1$ .  $\square$

**Equivalence 3.3.5. (B-differential  $\leftrightarrow$  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  $\sigma : \partial_B H(x) \rightarrow \mathcal{S}$  defined by (3.14a). Therefore, knowing  $\sigma$  by its definition (3.14), determining  $\partial_B H(x)$  or  $\mathcal{S}$  are equivalent problems.  $\square$

### 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  $|\partial_B^X H(x) \setminus \partial_B H(x)|$  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  $\mathbb{R}^p$ :

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

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

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

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

$$\iota : s \in \{\pm 1\}^p \rightarrow \iota(s) := \{i \in [1 : p] : s_i = +1\} \in \mathfrak{P}([1 : p]), \quad (3.16)$$

whose reverse map is  $\iota^{-1} : I \in \mathfrak{P}([1:p]) \rightarrow s \in \{\pm 1\}^p$ , 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  $A \in \mathbb{R}^{m \times n}$ ,

$$\exists x \in \mathbb{R}^n : Ax > 0 \iff \nexists \alpha \in \mathbb{R}_+^m \setminus \{0\} : A^\top \alpha = 0. \quad (3.17)$$

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

*Proof.* Let  $s \in \{\pm 1\}^p$  and set  $I := \iota(s) = \{i \in [1:p] : s_i = +1\}$ . Define  $A := \text{Diag}(s)V^\top$  to make the link with Gordan's alternative (3.17). One has the equivalences

$$\begin{aligned} s \in \mathcal{S}^c &\iff \nexists x \in \mathbb{R}^n : Ax > 0 \quad [\text{definition of } \mathcal{S} \text{ in (3.12)}] \\ &\iff \exists \alpha \in \mathbb{R}_+^m \setminus \{0\} : A^\top \alpha = 0 \quad [\text{Gordan's alternative (3.17)}] \\ &\iff \exists \alpha \in \mathbb{R}_+^m \setminus \{0\} : s \cdot \alpha \in \mathcal{N}(V) \\ &\iff \mathcal{N}(V) \cap \mathcal{O}_I^p \neq \{0\} \quad [\text{see below}] \\ &\iff I \in \mathcal{I} \quad [\text{definition of } \mathcal{I}]. \end{aligned} \quad (3.18)$$

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

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

**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  $\iota : \mathcal{S}^c \rightarrow \mathcal{I}$ , established in proposition 3.3.7: to determine  $\mathcal{S}$ , it suffices to determine  $\mathcal{S}^c = \iota^{-1}(\mathcal{I})$ , hence to determine  $\mathcal{I}$ , and vice versa.  $\square$

In example 3.3.2, one has  $\mathcal{N}(V) = \mathbb{R}e$ , which only encounters the orthants  $\mathcal{O}_\emptyset^3$  and  $\mathcal{O}_{[1:3]}^3$  outside the origin; hence  $\mathcal{I} = \{\emptyset, [1:3]\}$ . We have seen that  $\mathcal{S}^c = \{\pm(1, 1, 1)\}$  for this problem. Clearly,  $\iota$  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  $\text{null}(A)$ , 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):

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

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:

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

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  $\tilde{s} := \text{sgn}(\alpha)$  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  $\text{null}(V_{:,J}) = 1$ . 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  $\tilde{s} = \text{sgn}(\alpha)$ , where  $\alpha \in \mathcal{N}(V_{:,J})$  for some  $J \in \mathcal{C}$ .  $\square$

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  $s \in \{\pm 1\}^p$ ,

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

*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:

$$\{d \in \mathbb{R}^n : s_j v_j^\top d > 0 \text{ for all } j \in J\} = \emptyset, \quad (3.22a)$$

$$\forall J_0 \subsetneq J, \{d \in \mathbb{R}^n : s_j v_j^\top d > 0 \text{ for all } j \in J_0\} \neq \emptyset. \quad (3.22b)$$

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  $\{d \in \mathbb{R}^n : s_j v_j^\top d > 0 \text{ for all } j \in J \setminus \{j_0\}\} \neq \emptyset$  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),  $J \neq \emptyset$ . By Gordan's alternative (3.17), (3.22a) and (3.22b) read

$$\exists \alpha \in \mathbb{R}_+^J \setminus \{0\} \text{ such that } \sum_{j \in J} s_j v_j \alpha_j = 0, \quad (3.22c)$$

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

From these properties, one deduces that  $\alpha > 0$  and that  $\text{null}(V_{:,J}) \geq 1$ . To show that  $\text{null}(V_{:,J}) = 1$ , we proceed by contradiction. Suppose that there is a nonzero  $\alpha'' \in \mathbb{R}^J$  that is not colinear with  $\alpha$  and that verifies  $\sum_{j \in J} s_j v_j \alpha''_j = 0$ . One can assume that  $t := \max\{\alpha''_j / \alpha_j : j \in J\} > 0$  (take  $-\alpha''$  otherwise). Set  $J_0 := \{j \in J : \alpha''_j / \alpha_j < t\}$ . By the non-colinearity of  $\alpha$  and  $\alpha''$ , on the one hand, and the definition of  $t$ , on the other hand, one has  $\emptyset \subsetneq J_0 \subsetneq J$ . Furthermore,  $\alpha' := \alpha - \alpha''/t \geq 0$ ,  $\alpha'_j > 0$  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  $\beta$  would be two linearly independent vectors in  $\mathcal{N}(V_{:,J})$  (since  $s_J \cdot \alpha$  has no zero component), contradicting  $\text{null}(V_{:,J}) = 1$ .

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

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  $\text{null}(V_{:,I}) = 1$  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  $J \neq \emptyset$ .

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

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  $\alpha$ . Since the components of  $\alpha$  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  $\text{null}(V_{:,J'}) = 1$  and there is a nonzero  $\alpha' \in \mathcal{N}(V_{:,J'})$ . By (3.20),  $\alpha'$  has nonzero components. Furthermore,  $(\alpha', 0_{I \setminus J'}) \in \mathcal{N}(V_{:,I})$ , which has unit dimension and contains  $\alpha$ . Therefore,  $\alpha$  and  $(\alpha', 0_{I \setminus J'})$  are colinear. Since the components of  $\alpha$  are  $\neq 0$ , we get that  $J' = J$ .  $\square$

### 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 “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  $v_1, \dots, v_p \in \mathbb{R}^n \setminus \{0\}$ . It is requested to determine all the sign vectors  $s \in \{\pm 1\}^p$  such that  $\text{cone}\{s_i v_i : i \in [1 : p]\}$  is pointed.  $\square$

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

**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

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

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

### 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  $\bar{v}_1, \dots, \bar{v}_p \in \mathbb{A}$ . Let  $\mathbb{A}_0 := \mathbb{A} - \mathbb{A}$  be the vector space parallel to  $\mathbb{A}$ , endowed with

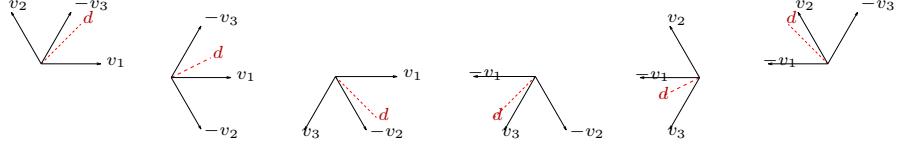


Figure 3.1: The figure is related to the linear complementarity problem defined by example 3.3.2: the  $v_i$ 's are the columns of the matrix  $V$  (their third zero components are not represented). Each of the 6 sets of vectors plots the 3 vectors  $\{s_i v_i : i \in [1:3]\}$ , for each of the 6 sign 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  $s_i v_i^T d > 0$  for all  $i \in [1:3]$ . Each conic hull of these vectors, namely  $\text{cone}\{s_i v_i : i \in [1:3]\}$ , 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  $\mathcal{S}$ .

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

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

□

Of course, if  $(I, J)$  is an appropriate ordered bipartition to which a separating covector  $\xi$  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

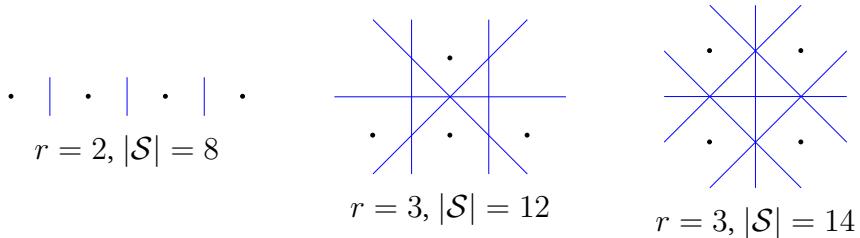


Figure 3.2: Linearly separable bipartitions of a set of  $p = 4$  points  $\bar{v}_i$  in  $\mathbb{R}^2$  (the dots in the figure). Possible separating hyperplanes are the drawn lines. We have not represented any separating line associated with the partition  $(\emptyset, [1 : p])$  or  $([1 : p], \emptyset)$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

solution to this problem by drawing the separating hyperplanes  $\{\bar{v} \in \mathbb{A} : \xi^T \bar{v} = t\}$  corresponding to some separating covector  $\xi$  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  $r := \dim(\text{vect}\{\bar{v}_1, \dots, \bar{v}_p\}) + 1$ , 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  $v_1, \dots, v_p \in \mathbb{R}^n$  such that  $K := \text{cone}\{v_k : k \in [1 : p]\}$  is a pointed cone. From proposition 3.3.13, there is a direction  $d \in \mathbb{R}^n$  such that

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

Define

$$\begin{aligned} \mathbb{A} &:= \{\bar{v} \in \mathbb{R}^n : d^\top \bar{v} = 1\}, & \mathbb{A}_0 &:= \mathbb{A} - \mathbb{A} = \{v \in \mathbb{R}^n : d^\top v = 0\}, \\ \forall k \in [1 : p] : \bar{v}_k &:= v_k / (v_k^\top d) \in \mathbb{A}. \end{aligned}$$

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

$$K_I := \text{cone}\{v_i : i \in I\} \quad \text{and} \quad K_J := \text{cone}\{v_j : j \in J\}, \quad (3.24a)$$

$$C_I := K_I \cap \mathbb{A} \quad \text{and} \quad C_J := K_J \cap \mathbb{A}, \quad (3.24b)$$

with the convention  $K_\emptyset = \{0\}$  and  $C_\emptyset = \emptyset$ . □

**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)  $\text{cone}((-K_I) \cup K_J)$  is pointed,
- (ii)  $K_I \cap K_J = \{0\}$ ,
- (iii)  $C_I \cap C_J = \emptyset$ ,
- (iv) there exists a vector  $\xi \in \mathbb{A}_0$  such that  $\max_{i \in I} \xi^\top \bar{v}_i < \min_{j \in J} \xi^\top \bar{v}_j$ .

*Proof.* [(i)  $\Rightarrow$  (ii)] We show the contrapositive. If there is  $v \in (K_I \cap K_J) \setminus \{0\}$ , then  $-v \in (-K_I) \subseteq \text{cone}((-K_I) \cup K_J)$  and  $v \in K_J \subseteq \text{cone}((-K_I) \cup K_J)$ . Therefore,  $\text{cone}((-K_I) \cup K_J)$  is not pointed.

[(ii)  $\Rightarrow$  (iii)]  $\emptyset = \mathbb{A} \cap \{0\} = \mathbb{A} \cap K_I \cap K_J$  [(ii)]  $= (\mathbb{A} \cap K_I) \cap (\mathbb{A} \cap K_J) = C_I \cap C_J$ .

[(iii)  $\Rightarrow$  (iv)] We claim that

$C_I$  is nonempty, convex and compact.

Indeed, since  $C_I$  is nonempty (it contains the vectors  $\bar{v}_i$  for  $i \in I \neq \emptyset$ ), 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  $v^\top d > 0$  for all  $v \in K_I \setminus \{0\}$ . For the same reason,

$C_J$  is nonempty, convex and compact.

Now, since  $C_I \cap C_J = \emptyset$  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  $\xi^\top v < \xi^\top w$ , for all  $v \in C_I$  and all  $w \in C_J$ . This shows that (iv) holds.

[(iv)  $\Rightarrow$  (i)] Since  $\text{cone}((-K_I) \cup K_J) = \text{cone}(\{-v_i : i \in I\} \cup \{v_j : j \in J\})$ , by proposition 3.3.13, it suffices to find  $d_{(I,J)} \in \mathbb{R}^n$  such that

$$\left(-v_i^\top d_{(I,J)} > 0, \quad \forall i \in I\right) \quad \text{and} \quad \left(v_j^\top d_{(I,J)} > 0, \quad \forall j \in J\right). \quad (3.25)$$

By (iv) and the fact that  $\theta \in (0, \pi) \rightarrow \cot \theta \in \mathbb{R}$  is surjective, one can determine  $\theta \in (0, \pi)$  such that

$$\max_{i \in I} \frac{\xi^\top v_i}{v_i^\top d} < -\cot \theta < \min_{j \in J} \frac{\xi^\top v_j}{v_j^\top d}. \quad (3.26)$$

Since  $\sin \theta > 0$  for  $\theta \in (0, \pi)$  and since  $v_k^\top d > 0$  for all  $k \in [1 : p]$ , this is equivalent to

$$\max_{i \in I} v_i^\top [(\cos \theta)d + (\sin \theta)\xi] < 0 < \min_{j \in J} v_j^\top [(\cos \theta)d + (\sin \theta)\xi].$$

Therefore, (3.25) is satisfied with  $d_{(I,J)} := (\cos \theta)d + (\sin \theta)\xi$ .  $\square$

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  $v_1, \dots, v_p$  and take  $s \in \mathcal{S}$ , which is nonempty. By (3.23),  $\text{cone}\{s_i v_i : i \in [1 : p]\}$  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

$$\text{cone}\{\tilde{s}_i v_i : i \in [1 : p]\} \text{ is pointed} \iff \exists \xi \in \mathbb{A}_0 : \max_{i \in I} \xi^\top \bar{v}_i < \min_{j \in J} \xi^\top \bar{v}_j. \quad (3.27)$$

Indeed, one has

$$\begin{aligned} \text{cone}\{\tilde{s}_i v_i : i \in [1 : p]\} \text{ is pointed} \\ \iff \text{cone}\{\tilde{s}_i s_i (s_i v_i) : i \in [1 : p]\} \text{ is pointed} \\ \iff \text{cone}((-K_I) \cup K_J) \text{ is pointed} \\ \iff \exists \xi \in \mathbb{A}_0 : \max_{i \in I} \xi^\top \bar{v}_i < \min_{j \in J} \xi^\top \bar{v}_j, \end{aligned}$$

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  $\text{cone}\{\tilde{s}_i v_i : i \in [1 : p]\}$  is pointed) and the linearly separable bipartitioning problem 3.3.15 of the vectors  $\bar{v}_i = s_i v_i / (s_i v_i^\top d) = v_i / (v_i^\top d)$ ,  $i \in [1 : p]$ , where  $d$  is associated with the pointed cone  $\text{cone}\{s_i v_i : i \in [1 : p]\}$  by the equivalence (i)  $\Leftrightarrow$  (iii) of proposition 3.3.13.  $\square$

### 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  $v_1, \dots, v_p \in \mathbb{R}^n$ . Consider the hyperplanes containing the origin:

$$\mathcal{H}_i := \{d \in \mathbb{R}^n : v_i^\top d = 0\}. \quad (3.28)$$

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

$$\mathcal{H}_i^+ := \{d \in \mathbb{R}^n : v_i^\top d > 0\} \quad \text{and} \quad \mathcal{H}_i^- := \{d \in \mathbb{R}^n : v_i^\top d < 0\}.$$

The problem is to determine the following set of open sectors or cells of  $\mathbb{R}^n$ , indexed by the bipartitions  $(I_+, I_-)$  of  $[1 : p]$ :

$$\mathfrak{C} := \{(I_+, I_-) \in \mathfrak{B}([1 : p]) : (\cap_{i \in I_+} \mathcal{H}_i^+) \cap (\cap_{i \in I_-} \mathcal{H}_i^-) \neq \emptyset\}, \quad (3.29)$$

where  $\mathfrak{B}([1 : p])$  denotes the set of bipartitions of  $[1 : p]$ . □

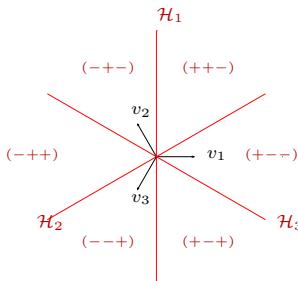


Figure 3.3: Illustration of problem 3.3.19 (arrangement of hyperplanes containing the origin) for the 3 vectors that are the columns on the matrix  $V$  in example 3.3.2 (since the last components of these  $v_i$ 's vanish, only the first two ones are represented above). The hyperplanes  $\mathcal{H}_i$  are defined by (3.28). The regions to determine are represented by the sign vectors here denoted  $(s_1 s_2 s_3)$  with  $s_i = \pm 1$ : if  $d \in \mathbb{R}^2$  belongs to the region  $(s_1 s_2 s_3)$ , then  $s_i = +1$  if  $v_i^\top d > 0$  and  $s_i = -1$  if  $v_i^\top d < 0$ . 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

$$\eta : (I_+, I_-) \in \mathfrak{B}([1:p]) \mapsto s \in \{\pm 1\}^p, \text{ where } s_i = \begin{cases} +1 & \text{if } i \in I_+, \\ -1 & \text{if } i \in I_- \end{cases} \quad (3.30)$$

and the setting  $V = (v_1 \ \cdots \ v_p)$ , 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  $\eta$  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,

$$\begin{aligned} (I_+, I_-) \in \mathfrak{C} &\iff \exists d \in (\cap_{i \in I_+} \mathcal{H}_i^+) \cap (\cap_{i \in I_-} \mathcal{H}_i^-) \\ &\iff \exists d \in \mathbb{R}^n : (v_i^\top d > 0 \text{ for } i \in I_+) \text{ and } (v_i^\top d < 0 \text{ for } i \in I_-) \\ &\iff \exists d \in \mathbb{R}^n : s \cdot (V^\top d) > 0 \\ &\iff s \in \mathcal{S}. \end{aligned}$$

These equivalences show the bijectivity of  $\eta$  from  $\mathfrak{C}$  onto  $\mathcal{S}$ .  $\square$

**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  $\eta : \mathfrak{C} \rightarrow \mathcal{S}$  claimed in proposition 3.3.20.  $\square$

## 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  $v_1, \dots, v_p \in \mathbb{R}^n$  and are nonzero by construction. When the properties are given for  $\mathcal{S}$ , one may have  $p \geq n$  and the referenced matrix  $V \in \mathbb{R}^{n \times p}$  is *assumed* to have nonzero columns, which implies that  $\mathcal{S} \neq \emptyset$ . 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  $K(x) := P_{\mathbb{R}_+^n}[Mx + q]$ ,  $P_{\mathbb{R}_+^n}$  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  $\partial_B P_{\mathbb{R}_+^n}$ , one can look for conditions ensuring that the chain rule applies for the composition defining the map  $K : \mathbb{R}^n \rightarrow \mathbb{R}^n$ . 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  $\partial_B H(x)$ ). *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  $p \leq n$ .*

*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  $s \in \text{sgn}(\alpha)$ , 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  $s \cdot (V^\top d) > 0$ . 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  $V^\top$  is surjective, the system  $V^\top d = s$  holds for some  $d \in \mathbb{R}^n$ . For this  $d$ ,  $s \cdot (V^\top d) = e$ , so that  $s \cdot (V^\top d) > 0$  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  $\partial_B H(x) = \sigma^{-1}(\{\pm 1\}^p)$  ( $\sigma^{-1}$  is defined by (3.14b)) is complete.  $\square$

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  $\sigma$  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  $\{\pm 1\}^p$ ). 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  $\text{co}\{\pm 1\}^p$  can be joined by a single edge).  $\square$

**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  $s^0, \dots, s^l \in S$  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  $S$ .  $\square$

**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  $l := \sum_{i \in [1:p]} |\tilde{s}_i - s_i|/2 \leq p$  in  $\mathcal{S}$ .*

*Proof.*  $[ \Rightarrow ]$  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  $s \in \mathcal{S} \neq \emptyset$ , one can find  $d \in \mathbb{R}^n$  such that  $s \cdot (V^\top d) > 0$ , 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.

$[ \Leftarrow ]$  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.  $\square$

For  $k \in [1 : p]$ , we introduce

$$\mathcal{S}_k := \{s \in \{\pm 1\}^k : \exists d \in \mathbb{R}^n \text{ such that } s_i v_i^\top d > 0 \text{ for } i \in [1 : k]\}. \quad (3.33)$$

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,  $|\mathcal{S}_{k+1}^c| \geq 2|\mathcal{S}_k^c|$ .*  
 2) *If  $v_{k+1} \notin \text{vect}\{v_1, \dots, v_k\}$ , 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  $v_1, \dots, v_k$ , then,  $[(s, \pm 1) \text{ and } (-s, \pm 1) \in \mathcal{S}_{k+1} \text{ for one } s \in \mathcal{S}_k] \text{ and } [(s', +1) \text{ or } (s', -1) \in \mathcal{S}_{k+1} \text{ for any } s' \in \mathcal{S}_k]$ . In particular,  $|\mathcal{S}_{k+1}| \geq |\mathcal{S}_k| + 2$ .*

*Proof.* 1) If  $s \in \mathcal{S}_k^c$ , there is no  $d \in \mathbb{R}^n$  such that  $s_i v_i^\top d > 0$  for  $i \in [1 : k]$ . Therefore, there is no  $d \in \mathbb{R}^n$  such that  $(s_i v_i^\top d > 0)$  for  $i \in [1 : k]$  and  $\pm v_{k+1}^\top d > 0$ . Hence,  $(s, \pm 1) \in \mathcal{S}_{k+1}^c$ . This implies that  $|\mathcal{S}_{k+1}^c| \geq 2|\mathcal{S}_k^c|$ .

2) Let  $P$  be the orthogonal projector on  $\text{vect}\{v_1, \dots, v_k\}^\perp$  for the Euclidean scalar product. By assumption,  $P v_{k+1} \neq 0$ . Let  $s \in \mathcal{S}_k$ , so that there is a direction  $d \in \mathbb{R}^n$  such that  $s_i v_i^\top d > 0$  for  $i \in [1 : k]$ . For any  $t \in \mathbb{R}$  and  $i \in [1 : k]$ , the directions  $d_\pm := d \pm t P v_{k+1}$  verify  $s_i v_i^\top d_\pm = s_i v_i^\top d > 0$  (because  $v_i^\top P v_{k+1} = 0$ ). In addition, for  $t > 0$  sufficiently large, one has  $\pm v_{k+1}^\top d_\pm = \pm v_{k+1}^\top d + t \|P v_{k+1}\|^2 > 0$  (because  $P^2 = P$  and  $P^\top = P$ ). We have shown that both  $(s, +1)$  and  $(s, -1)$  are in  $\mathcal{S}_{k+1}$ . Therefore,  $|\mathcal{S}_{k+1}| \geq 2|\mathcal{S}_k|$ .

Now,  $|\mathcal{S}_k| + |\mathcal{S}_k^c| = 2^k$ ,  $|\mathcal{S}_{k+1}| + |\mathcal{S}_{k+1}^c| = 2^{k+1}$  and  $|\mathcal{S}_{k+1}^c| \geq 2|\mathcal{S}_k^c|$  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

$$\left( \forall i \in [1 : k] : v_i^\top d \neq 0 \right) \quad \text{and} \quad v_{k+1}^\top d = 0. \quad (3.34)$$

Indeed, let  $\mathbb{E} := \{d \in \mathbb{R}^n : v_{k+1}^\top d = 0\}$  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  $v_{k+1}^\top d = 0$ ) such that  $|\{(P v_i)^\top d : i \in [1 : k+1]\}| = |\{P v_i : i \in [1 : k+1]\}|$ . Since  $P v_{k+1} = 0$  and  $P v_i \neq 0$  for  $i \in [1 : k]$  (because the  $v_i$ 's are not colinear with  $v_{k+1}$ ), one has  $(P v_i)^\top d \neq 0$  for  $i \in [1 : k]$ . Since,  $0 \neq (P v_i)^\top d = v_i^\top P d = v_i^\top d$ , (3.34) follows.

Taking  $s_i := \text{sgn}(v_i^\top d)$  for  $i \in [1 : k]$ , one deduces from (3.34) that there is a direction  $d \in \mathbb{R}^n$  such that

$$\left( \forall i \in [1 : k] : s_i v_i^\top d > 0 \right) \quad \text{and} \quad v_{k+1}^\top d = 0.$$

It follows that, for  $\varepsilon > 0$  sufficiently small, the directions  $d_\pm := d \pm \varepsilon v_{k+1}$  satisfy

$$\left( \forall i \in [1 : k] : s_i v_i^\top d_\pm > 0 \right) \quad \text{and} \quad \pm v_{k+1}^\top d_\pm > 0.$$

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  $|\mathcal{S}_{k+1}| \geq 4 + (|\mathcal{S}_k| - 2) = |\mathcal{S}_k| + 2$ .  $\square$

### 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)

$$|\partial_B H(x)| = \sum_{I \subseteq [1:p]} (-1)^{\text{null}(V_{:,I})}, \quad (3.35)$$

where  $\text{null}(V_{:,I})$  is the nullity of  $V_{:,I}$  and the term in the right-hand side corresponding to  $I = \emptyset$  is 1 (one takes the convention that  $\text{null}(V_{:,\emptyset}) = 0$ ). 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  $\text{rank}(V_{:,I})$  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  $r := \text{rank}(V)$ . Then,*

$$2^r \leq |\partial_B H(x)| \leq 2^p. \quad (3.36a)$$

*If  $V$  has no colinear columns, then,*

$$\max(2p, 2^r) \leq 2^r + 2(p - r) \leq |\partial_B H(x)|. \quad (3.36b)$$

*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,  $|\mathcal{S}_{k+1}| \geq |\mathcal{S}_k|$ , 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  $p \geq r \geq 1$  and  $2r \leq 2^r$ . 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  $V$ ,  $|\mathcal{S}_{r+1}| \geq 2^r + 2$ . By induction, The inequality follows.  $\square$

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  $\varphi : x \in \mathbb{T} \rightarrow$

$\varphi(x) \in \mathbb{R}$ , 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  $\varphi(\tilde{x}) \leq \varphi(x) + \varepsilon$ . 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  $\sigma$  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  $s \cdot (V^\top d_s) > 0$ . Clearly, one still has  $s \cdot (\tilde{V}^\top d_s) > 0$ , 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  $s \cdot (\tilde{V}^\top d) > 0$  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  $|\mathcal{S}(V)| \leq |\mathcal{S}(\tilde{V})|$  for  $\tilde{V}$  near  $V$ . The lower semicontinuity of  $V \mapsto |\mathcal{S}(V)|$  follows.  $\square$

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

$$\forall I \subseteq [1 : p] : \quad \text{rank}(V_{:,I}) \leq \min(|I|, r). \quad (3.37)$$

**Definition 3.4.9** (general position). The vectors  $v_1, \dots, v_p \in \mathbb{R}^n$  are said to be in *general position*, if the matrix  $V := (v_1 \ \cdots \ v_p)$  verifies

$$\forall I \subseteq [1 : p] : \quad \text{rank}(V_{:,I}) = \min(|I|, r), \quad (3.38)$$

where  $r := \text{rank}(V)$ .  $\square$

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  $r := \text{rank}(V) = 3$  while for the 3 bottom vectors, with indices in  $I$  say, one has  $\min(|I|, r) - \text{rank}(V_{:,I}) = 3 - 2 \neq 0$ .

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  $2[({}^3_0) + ({}^3_1) + ({}^3_2)] = 14$  (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  $|\partial_B H(x)|$ ). *For  $V$  given by (3.11) and  $r := \text{rank}(V)$ , one has*

$$|\partial_B H(x)| \leq 2 \sum_{i \in [0:r-1]} \binom{p-1}{i}, \quad (3.39)$$

*with equality if and only if (3.38) holds.*

*Proof.* 1) The proof of the implication “(3.38)  $\Rightarrow$  (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  $\beta$  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

$$|\mathcal{S}(V)| > \beta. \quad (3.40a)$$

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

$$|\mathcal{S}(\tilde{V})| = \beta, \quad (3.40b)$$

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  $\text{rank}(V_{:,I}) < \min(|I|, r)$ . By linear algebra arguments (see [78] for more details), one can get an arbitrarily small perturbation  $\tilde{V}_{:,I}$  of  $V_{:,I}$ , such that  $\text{rank}(\tilde{V}_{:,I}) = \min(|I|, r)$  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  $\text{rank}(\tilde{V}_{:,I'}) \leq \min(|I'|, r)$  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  $\text{rank}(V_{:,I'}) \leq$

rank( $\tilde{V}_{:,I'}$ )  $\leq \min(|I'|, r)$  for any  $I' \subseteq [1 : p]$ . Therefore, rank( $V_{:,I'}$ ) =  $\min(|I'|, r)$  implies that 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 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 rank( $\tilde{V}_{:,I''}$ )  $< \min(|I'', r|)$ , one finally obtains a matrix  $\tilde{V}$ , arbitrarily close to  $V$ , such that (3.38) holds: rank( $\tilde{V}_{:,I}$ ) =  $\min(|I|, r)$  for all  $I \subseteq [1 : p]$ .

3) One still has to show that “(3.39) with equality  $\Rightarrow$  (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

$$\text{rank}(V_{:,I}) < \min(|I|, r). \quad (3.40c)$$

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 rank( $\tilde{V}$ )  $\leq r$  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

$$V_{:,J} \text{ is not injective} \quad \text{and} \quad |J| \leq r. \quad (3.40d)$$

Indeed, if  $|I| \leq r$ , one can take  $J = I$  to satisfy (3.40d), since  $\text{rank}(V_{:,I}) < |I|$  by (3.40c), so that  $V_{:,I}$  is not injective. If  $|I| > r$ , then  $\text{rank}(V_{:,I}) < r$  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  $\tilde{s}_j \in \text{sgn}(\alpha_j)$  for all  $j \in J$ . Then, by Gordan’s alternative (3.17),

$$\nexists d \in \mathbb{R}^n : \quad \tilde{s}_j v_j d > 0, \quad \text{for all } j \in J.$$

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  $J^c := [1 : p] \setminus J$ . By (3.40d),  $|J| \leq r \leq n$  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  $\tilde{v}_j^\top d_0 = \tilde{s}_j$  for  $j \in J$ , hence

$$\tilde{s}_j \tilde{v}_j^\top d_0 > 0, \quad \forall j \in J. \quad (3.40e)$$

Set  $\tilde{s}_j = 1$  for  $j \in J^c$ . Let  $d$  be a discriminating covector given by lemma 3.2.6 (there denoted  $\xi$ ) for the vectors  $\{0\} \cup \{\tilde{s}_i v_i : i \in [1 : p]\}$  sufficiently close to  $d_0$ . It results that  $\tilde{s}_j \tilde{v}_j^\top d > 0$  for  $j \in J$  (by (3.40e)) and that  $\tilde{s}_j \tilde{v}_j^\top d \neq 0$  for  $j \in J^c$ . Finally, we see that the sign vector  $s \in \{\pm 1\}^p$  defined by  $s_i = \text{sgn}(\tilde{v}_i^\top d)$  for all  $i \in [1 : p]$  is in  $\mathcal{S}(\tilde{V})$  and satisfies  $s_J = \tilde{s}_J$ , as desired.  $\square$

**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  $v_1, \dots, v_p \in \mathbb{R}^n$  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  $\beta$  is the right-hand side of Schläfli's bound (3.39) with  $r = \text{rank}(V)$ . Now, by the fact that  $\text{rank}(\tilde{V}) = r$ , proposition 3.4.10 ensures that  $|\mathcal{S}(\tilde{V})| \leq \beta$ . Therefore, one must have  $\mathcal{S}(\tilde{V}) = \mathcal{S}(V)$ .  $\square$

### 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  $r := \text{rank}(V)$  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  $|\partial_B H(x)| = 2^p$ .  $\square$

**Proposition 3.4.13** (fan arrangement). *If  $p \geq 2$  and the vectors  $v_i$ 's are not two by two colinear, one has  $\text{rank}(V) = 2$  if and only if  $|\partial_B H(x)| = 2p$ .*

*Proof.*  $[ \Rightarrow ]$  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  $I \subseteq [1 : p]$ :

$$\text{rank}(V_{:,I}) = \begin{cases} |I| & \text{if } |I| \leq 2 \\ 2 & \text{if } |I| > 2. \end{cases}$$

Therefore (3.38) holds. By proposition 3.4.10, this implies that equality holds in (3.39), that is, with  $r := \text{rank}(V) = 2$ :  $|\partial_B H(x)| = 2 \sum_{i \in [0 : 1]} \binom{p-1}{i} = 2p$ .

[ $\Leftarrow$ ] If  $|\partial_B H(x)| = 2p$ , (3.36b) yields  $2p \leq \max(2p, 2^r) \leq 2^r + 2(p - r) \leq 2p$ , so that equality holds in these inequalities. By the last one,  $2^r = 2r$ , which only occurs for  $r \in \{1, 2\}$ . Since  $p \geq 2$  and the vectors are not colinear, one has  $r = 2$ .  $\square$

### 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).  $\partial_B H(x) = \text{ext } \partial_C H(x)$ .

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  $i$ th row,  $i \in [1 : m]$ , is defined by*

$$J_{i,:} = \begin{cases} A_{i,:} & \text{if } i \in \mathcal{A}(x), \\ \frac{1}{2}[A_{i,:} + B_{i,:}] & \text{if } i \in \mathcal{E}(x), \\ B_{i,:} & \text{if } i \in \mathcal{B}(x). \end{cases} \quad (3.41)$$

*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  $\text{co } \partial_B H(x) = \partial_C H(x)$ , by (3.2). This is the formula of  $J$  given in the statement of the proposition.  $\square$

Instead of taking  $J_{1/2} := \frac{1}{2}(M + M')$  in the preceding proof, one could also have taken  $J_t := (1 - t)M + tM'$ , which is also in  $\text{co } \partial_B H(x) = \partial_C H(x)$  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  $p \leq n$  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  $v_1, \dots, v_p$ , which may be identical or colinear, a direction  $d$  such that  $v_i^T d \neq 0$  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  $d \in \mathbb{R}^n$  such that  $v_i^T d \neq 0$  for all  $i$ ).

Let be given  $p$  nonzero vectors  $v_1, \dots, v_p$  in  $\mathbb{R}^n$  and take  $d \in \mathbb{R}^n \setminus \{0\}$ .

Repeat:

1. If  $I := \{i \in [1 : p] : v_i^T d = 0\} = \emptyset$ , exit.
2. Let  $i \in I$ .
3. Take  $t > 0$  sufficiently small such that, for all  $j \notin I$ ,  $(v_j^T d)(v_j^T [d + tv_i]) > 0$ .
4. Update  $d := d + tv_i$ .

*Explanation.* In step 3, any sufficiently small  $t > 0$  is appropriate (the proof of [255, lemma 2.1] computes bounds explicitly), since  $(v_j^T d)(v_j^T [d + tv_i])$  is positive for  $t = 0$ . The new direction  $d$  set in step 4 is such that  $v_i^T (d + tv_i) = t\|v_i\|^2 > 0$ , so that this direction makes at least one more  $v_j^T d$  nonzero than the previous one. This implies that the algorithm finds an appropriate direction in at most  $p$  loops.  $\square$

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  $\sigma$  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  $\partial_B H(x)$ ).

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  $v_1, \dots, v_p \in \mathbb{R}^n$ .
2. By algorithm 3.5.1, compute  $d \in \mathbb{R}^n$  such that  $v_i^\top d \neq 0$  for all  $i \in [1 : p]$ .
3. Define  $s \in \mathcal{S}$  by  $s_i := \text{sgn}(v_i^\top d)$ , 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  $s_i v_i^\top d > 0$  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  $\partial_B H(x)$ .  $\square$

### 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

$$\mathcal{S}_k^1 := \{s \in \mathcal{S}_k : s_1 = +1\}, \quad (3.42)$$

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.

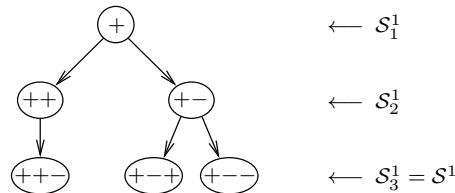


Figure 3.4: Half of the  $\mathcal{S}$ -tree for example 3.3.2 (the other half is obtained by swapping the +'s and the -'s). Top-down arrows indicate descendence; the sign sets  $\mathcal{S}_k^1$  are defined by (3.42).

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  $s \in \mathcal{S}_k^1$ , it solves a single problem (LOP) to determine whether  $(s, +1)$  or  $(s, -1)$  is in  $\mathcal{S}_{k+1}^1$ .

The RC algorithm succeeds in solving only one LOP to determine whether  $(s, +1)$  and  $(s, -1)$  are in  $\mathcal{S}_{k+1}^1$ , at the node  $s \in \mathcal{S}_k^1$ , thanks to the memorization of a direction  $d$  such that  $s \cdot (V_k^\top d) > 0$  (we note  $V_k := V_{:, [1:k]}$ ). Indeed, one has

$$\begin{aligned} v_{k+1}^\top d < 0 &\implies (s, -1) \in \mathcal{S}_{k+1}^1, \\ v_{k+1}^\top d > 0 &\implies (s, +1) \in \mathcal{S}_{k+1}^1, \end{aligned}$$

and one of these two cases takes place if we exclude the case where  $v_{k+1}^\top d = 0$ . In [208, Algorithm 1], the case where  $v_{k+1}^\top d = 0$  is not dealt with completely since  $(s, +1)$  is declared to belong to  $\mathcal{S}_{k+1}^1$  in that case, while it is clear that  $(s, -1)$  is also in  $\mathcal{S}_{k+1}^1$ . 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  $v_{k+1}^\top d \simeq 0$ , so that both  $(s, \pm 1)$  are accepted in  $\mathcal{S}_{k+1}^1$  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  $\mathcal{S}_{k+1}^1$ , the question of whether  $(s, -s_{k+1})$  also belongs to  $\mathcal{S}_{k+1}^1$  arises. In the RC algorithm, the answer to this question is obtained by solving a LOP similar to

$$\left\{ \begin{array}{l} \min_{(d,t) \in \mathbb{R}^n \times \mathbb{R}} t \\ s_i v_i^\top d \geq 1, \quad \forall i \in [1 : k] \\ -s_{k+1} v_{k+1}^\top d \geq -t \\ t \geq -1. \end{array} \right. \quad (3.43)$$

When  $s \in \mathcal{S}_k^1$ , this problem is feasible (take  $d$  satisfying  $s_i v_i^\top d \geq 1$ , for all  $i \in [1 : k]$ , and  $t$  sufficiently large) and bounded (its optimal value is  $\geq -1$ ), 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  $\text{RC-REC}(V, v_1, +1)$ .

**Algorithm 3.5.6** ( $\text{RC-REC}(V, d, s)$ ). 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  $s_i v_i^\top d > 0$  for all  $i \in [1 : k]$ .

1. If  $k = p$ , print  $s$  and return.
2. If  $v_{k+1}^\top d \simeq 0$ , then
  - 2.1. Execute  $\text{RC-REC}(V, d_+, (s, +1))$ , where  $d_+ := d + t_+ v_{k+1}$  with  $t_+ > 0$  chosen in the nonempty open interval
 
$$\left( 0, \min_{\substack{i \in [1:k] \\ s_i v_i^\top v_{k+1} < 0}} \frac{-v_i^\top d}{v_i^\top v_{k+1}} \right).$$
  - 2.2. Execute  $\text{RC-REC}(V, d_-, (s, -1))$ , where  $d_- := d + t_- v_{k+1}$  with  $t_- < 0$  chosen in the nonempty open interval
 
$$\left( \max_{\substack{i \in [1:k] \\ s_i v_i^\top v_{k+1} > 0}} \frac{-v_i^\top d}{v_i^\top v_{k+1}}, 0 \right).$$
3. Else  $s_{k+1} := \text{sgn}(v_{k+1}^\top d)$ .
  - 3.1. Execute  $\text{RC-REC}(V, d, (s, s_{k+1}))$ .
  - 3.2. Solve the LOP (3.43) and denote by  $(d, t)$  a solution.  
If  $t = -1$ , execute  $\text{RC-REC}(V, d, (s, -s_{k+1}))$ .

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  $s_i v_i^\top d_\pm > 0$  for  $i \in [1 : k+1]$  and  $s_{k+1} = \pm 1$ , provided  $|v_{k+1}^\top d|$  is sufficiently small, which justifies the recursive call to  $\text{RC-REC}$  with the given arguments. The test  $v_{k+1}^\top d \simeq 0$  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  $\text{RC}$  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  $\text{RC}$  algorithm comes from the possible numerous LOPs to solve in step 3.2 of  $\text{RC-REC}$ .

### 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<sub>4</sub> of the ISF algorithm (see paragraphs A and D of section 3.5.2).

Note that, this algorithm need not generate directions  $d$  satisfying  $s \cdot (V_k^\top d) > 0$ , 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  $\geq 2$ , 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<sub>2</sub> 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  $t \geq 0$  (equivalently,  $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}$ ) or  $t = -1$  (equivalently,  $(s, -s_{k+1}) \in \mathcal{S}_{k+1}$ ).

Let  $\sigma_i$ ,  $i \in [1 : k + 1]$ , be the multipliers associated with the first  $k + 1$  constraints of (3.43) and  $\tau$  be the multiplier associated with its last constraint. Then, the Lagrangian dual of (3.43) reads [31, 26, 29, 105]

$$\begin{cases} \max_{(\sigma, \tau) \in \mathbb{R}^{k+1} \times \mathbb{R}} \sum_{i \in [1:k]} \sigma_i - \tau \\ \sigma \geq 0 \\ \tau \geq 0 \\ \sigma_{k+1} + \tau = 1 \\ \sigma_{k+1}s_{k+1}v_{k+1} = \sum_{i \in [1:k]} \sigma_i s_i v_i. \end{cases} \equiv \begin{cases} \max_{\sigma \in \mathbb{R}^{k+1}} \sum_{i \in [1:k+1]} \sigma_i - 1 \\ \sigma \geq 0 \\ \sigma_{k+1} \leq 1 \\ \sigma_{k+1}s_{k+1}v_{k+1} = \sum_{i \in [1:k]} \sigma_i s_i v_i, \end{cases} \quad (3.44)$$

where the second form of the dual is obtained by eliminating  $\tau$  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

$$(s, -s_{k+1}) \in \mathcal{S}_{k+1} \iff \sigma = 0.$$

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  $\sigma$  by a factor going to  $+\infty$ ; if  $\sigma_{k+1} \in (0, 1)$ , the dual objective would be increased by replacing  $\sigma$  by  $\sigma/\sigma_{k+1}$ ; in both cases the optimality of  $\sigma$  would be contradicted) if and only if  $\tau = 0$ . We have shown that

$$(s, -s_{k+1}) \notin \mathcal{S}_{k+1} \iff s_{k+1}v_{k+1} \in \text{cone}\{s_i v_i : i \in [1 : k]\}.$$

The next proposition shows how a matroid circuit can be detected from the dual solution  $\sigma$  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  $(\sigma, \tau)$  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.  $\square$

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<sub>2</sub> (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<sub>4</sub> (modifications A and D<sub>4</sub>) 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  $r := \text{rank}(V)$  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  $r := \text{rank}(V)$  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.,  $Q^\top Q = I_n$ ) 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  $v_1, \dots, v_r$  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):

$$\mathcal{S}_r = \{\pm 1\}^r. \quad (3.45)$$

Furthermore, for each  $s \in \mathcal{S}_r$ , we have, using  $S := \text{Diag}(s)$ ,  $Q_r := Q_{:, [1:r]}$  and  $R_r := R_{[1:r], [1:r]}$ , that the vector

$$d_s = Q_r R_r^{-\top} s \quad (3.46)$$

is such that  $s \cdot (V_{:, [1:r]}^\top d_s) = e > 0$ , 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  $\text{RC-REC}(V, d_s, s)$  (see algorithm 3.5.11 below).

B. SPECIAL HANDLING OF THE CASE WHERE  $v_{k+1}^\top d \simeq 0$ . Directions  $d_\pm := d + t_\pm v_{k+1}$  ensuring that  $(s, \pm 1) \cdot (V_{k+1}^\top d_\pm) > 0$  can be computed not only when  $v_{k+1}^\top d \simeq 0$  like in step 2 of the RC-REC algorithm 3.5.6, but also when  $v_{k+1}^\top d$  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  $s_i v_i^\top d > 0$  for all  $i \in [1 : k]$ ), so that these inequalities are verified when  $v_{k+1}^\top d = 0$ . 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  $s \cdot (V_k^\top d) > 0$ , that  $v_{k+1} \neq 0$  and that*

$$\max_{\substack{i \in [1:k] \\ s_i v_i^\top v_{k+1} > 0}} \frac{-v_i^\top d}{v_i^\top v_{k+1}} < \frac{-v_{k+1}^\top d}{\|v_{k+1}\|^2} < \min_{\substack{i \in [1:k] \\ s_i v_i^\top v_{k+1} < 0}} \frac{-v_i^\top d}{v_i^\top v_{k+1}}. \quad (3.47)$$

- 1) *The direction  $d_+ := d + t_+ v_{k+1}$  verifies  $s \cdot (V_k^\top d_+) > 0$  and  $v_{k+1}^\top d_+ > 0$  if and only if  $t_+$  is in the nonempty open interval*

$$\left( \frac{-v_{k+1}^\top d}{\|v_{k+1}\|^2}, \min_{\substack{i \in [1:k] \\ s_i v_i^\top v_{k+1} < 0}} \frac{-v_i^\top d}{v_i^\top v_{k+1}} \right). \quad (3.48a)$$

- 2) *The direction  $d_- := d + t_- v_{k+1}$  verifies  $s \cdot (V_k^\top d_-) > 0$  and  $-v_{k+1}^\top d_- > 0$  if and only if  $t_-$  is in the nonempty open interval*

$$\left( \max_{\substack{i \in [1:k] \\ s_i v_i^\top v_{k+1} > 0}} \frac{-v_i^\top d}{v_i^\top v_{k+1}}, \frac{-v_{k+1}^\top d}{\|v_{k+1}\|^2} \right). \quad (3.48b)$$

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  $T_s := \{i_1, \dots, i_k\} \subseteq [1 : p]$ . Hence, we have to choose the next vector  $v_{i_{k+1}}$  by 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

$$i_{k+1} = \operatorname{argmax}_{i \in T_s^b} \frac{|v_i^\top d|}{\|v_i\|}, \quad (3.49)$$

which favors the vectors  $v_i$  for which  $|v_i^\top d|/\|v_i\|$  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<sub>1</sub>) Natural candidates as stem vectors are those obtained from the matroid circuits  $I$  made of  $r + 1$  columns of  $V$  ( $r = \operatorname{rank}(V)$ ) 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<sub>2</sub>) 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<sub>3</sub>) 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  $s_i v_i^\top d > 0$  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  $(s, s_{k+1}) \in \mathcal{S}_{k+1}^1$ . If a single sign  $s_{k+1} \in \{\pm 1\}$  is selected, the stem vectors can decide whether  $(s, -s_{k+1}) \in \mathcal{S}_{k+1}^1$ . If this is the case, this option D<sub>3</sub> 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  $s_i v_i^\top d > 0$  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<sub>4</sub>) Like with the option D<sub>3</sub>, all the stem vectors are computed, before running the recursive process that builds the  $\mathcal{S}$ -tree. But now, unlike with option D<sub>3</sub>, 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  $\pm 1$  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<sub>2</sub>. 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<sub>1</sub>). Then, the hat procedure calls the recursive procedure **ISF-REC**.

**Algorithm 3.5.11 (isf ( $V$ ), with options ABCD<sub>2</sub>).** 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  $r = \text{rank}(V)$  and  $T_r := \{i_1, \dots, i_r\}$  be the indices of  $r$  selected linear independent columns of  $V$ .
2. Compute the  $p - r$  matroid circuits (see option D<sub>1</sub>).
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  $\text{BCD}_2$ ). 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  $s_i v_i^\top d > 0$  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  $s_{k+1} := \text{sgn}(v_{i_{k+1}}^\top d)$ .
  - 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 ( $0 < Q \leq p - N$ ), 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 *permutohedron* 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  $k$ th 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  $\rho$  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  $\rho$ ) 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  $N \geq 2$ , 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  $N \leq 9$ . Our approach, which does not use the particular structure of this arrangement, can get  $|\mathcal{S}|$  in a reasonable time on a laptop for  $N \leq 6$ , which is to be compared to  $N \leq 9$  in [35]. See [148] for applications.
- The problems `BEK-CROSSPOLYTOPE-N` refer to the *cross-polytope arrangements* in [35, § 6.4]: for  $N \geq 2$ , 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  $N \leq 12$  (this observation is made for  $N \leq 21$  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  $\varsigma$  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  $\varsigma$  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  $\varsigma$ , 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	$n$	$p$	$r$	$\varsigma$	$\varsigma_{\max}$	$ \partial_B H(x) $	Schläfli's bound	LOPs solved in		
								Original RC	Simulated RC	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: “( $n, p, r, \varsigma$ )” are the features of the problem ( $V \in \mathbb{R}^{n \times p}$  is of rank  $r$  and has  $\varsigma$  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.

Problem	Simulated RC	Number of linear optimization problems (LOPs) solved and acceleration ratio (Ratio) for various options											
		ISF (A)		ISF (AB)		ISF (ABC)		ISF (ABCD1)		ISF (ABCD2)		ISF (ABCD3)	
		RC	LOP	Ratio	LOP	Ratio	LOP	Ratio	LOP	Ratio	LOP	Ratio	LOP
RAND-8-15-7	9907	9844	<b>1.01</b>	7641	<b>1.30</b>	5210	<b>1.90</b>	5199	<b>1.91</b>	4355	<b>2.27</b>	3638	<b>2.72</b>
RAND-9-16-8	22818	22691	<b>1.01</b>	17586	<b>1.30</b>	13046	<b>1.75</b>	13023	<b>1.75</b>	11185	<b>2.04</b>	9943	<b>2.29</b>
RAND-10-17-9	50642	50387	<b>1.01</b>	38167	<b>1.33</b>	28849	<b>1.76</b>	28839	<b>1.76</b>	25370	<b>2.00</b>	23266	<b>2.18</b>
SRAND-8-20-2	28620	28620	<b>1.00</b>	20207	<b>1.42</b>	6668	<b>4.29</b>	5535	<b>5.17</b>	2881	<b>9.93</b>	2851	<b>10.04</b>
SRAND-8-20-4	135566	136027	<b>1.00</b>	113493	<b>1.19</b>	60066	<b>2.26</b>	59267	<b>2.29</b>	45569	<b>2.97</b>	42445	<b>3.19</b>
SRAND-8-20-6	167262	167351	<b>1.00</b>	137450	<b>1.22</b>	77800	<b>2.15</b>	77752	<b>2.15</b>	62694	<b>2.67</b>	54980	<b>3.04</b>
RC-2D-20-6	1927	1904	<b>1.01</b>	1680	<b>1.15</b>	912	<b>2.11</b>	688	<b>2.80</b>	40	<b>48.17</b>	0	<b>—</b>
RC-2D-20-7	3343	3296	<b>1.01</b>	2912	<b>1.15</b>	2208	<b>1.51</b>	1792	<b>1.87</b>	52	<b>64.29</b>	0	<b>—</b>
RC-2D-20-8	5855	5760	<b>1.02</b>	4992	<b>1.17</b>	2752	<b>2.13</b>	1984	<b>2.95</b>	28	<b>209.11</b>	0	<b>—</b>
RC-PERM-6	9346	9280	<b>1.01</b>	7898	<b>1.18</b>	2076	<b>4.50</b>	1836	<b>5.09</b>	92	<b>101.59</b>	61	<b>153.21</b>
RC-PERM-7	90169	90094	<b>1.00</b>	79049	<b>1.14</b>	17230	<b>5.23</b>	16558	<b>5.45</b>	960	<b>93.93</b>	855	<b>105.46</b>
RC-PERM-8	953009	952597	<b>1.00</b>	856597	<b>1.11</b>	160781	<b>5.93</b>	158989	<b>5.99</b>	9766	<b>97.58</b>	9393	<b>101.46</b>
RC-RATIO-20-5-7	13669	15341	<b>0.89</b>	14028	<b>0.97</b>	7108	<b>1.92</b>	7064	<b>1.94</b>	3644	<b>3.75</b>	2467	<b>5.54</b>
RC-RATIO-20-6-7	322883	35882	<b>0.92</b>	31992	<b>1.03</b>	17797	<b>1.85</b>	17505	<b>1.88</b>	10669	<b>3.08</b>	8765	<b>3.75</b>
RC-RATIO-20-7-7	82447	81428	<b>1.01</b>	72272	<b>1.14</b>	47798	<b>1.72</b>	47748	<b>1.73</b>	30442	<b>2.71</b>	25841	<b>3.19</b>
BEK-THRESHOLD-4	87	79	<b>1.10</b>	54	<b>1.61</b>	46	<b>1.89</b>	37	<b>2.35</b>	26	<b>3.35</b>	16	<b>5.54</b>
BEK-THRESHOLD-5	2757	2884	<b>0.96</b>	2399	<b>1.15</b>	1270	<b>2.17</b>	1180	<b>2.34</b>	502	<b>5.49</b>	370	<b>3.75</b>
BEK-THRESHOLD-6	248521	261728	<b>0.95</b>	236027	<b>1.05</b>	71963	<b>3.45</b>	70410	<b>3.53</b>	21339	<b>11.65</b>	19184	<b>3.19</b>
BEK-RESONANCE-4	635	672	<b>0.94</b>	546	<b>1.16</b>	171	<b>3.71</b>	138	<b>4.60</b>	31	<b>20.48</b>	0	<b>—</b>
BEK-RESONANCE-5	36311	37607	<b>0.97</b>	34056	<b>1.07</b>	6700	<b>5.42</b>	6569	<b>5.53</b>	1141	<b>31.82</b>	810	<b>44.83</b>
BEK-RESONANCE-6	6164040	6269410	<b>0.98</b>	5956586	<b>1.03</b>	760930	<b>8.10</b>	760457	<b>8.11</b>	155555	<b>39.63</b>	(1)	<b>—</b>
BEK-CROSSPOLYTOPE-11	86526	110418	<b>0.78</b>	58954	<b>1.47</b>	17569	<b>4.92</b>	15265	<b>5.67</b>	6085	<b>14.22</b>	6049	<b>14.30</b>
BEK-CROSSPOLYTOPE-12	260601	337910	<b>0.77</b>	182575	<b>1.43</b>	46900	<b>5.56</b>	41780	<b>6.24</b>	18785	<b>13.87</b>	18740	<b>13.91</b>
BEK-CROSSPOLYTOPE-13	788970	1028066	<b>0.77</b>	560013	<b>1.41</b>	124828	<b>6.32</b>	113564	<b>6.95</b>	57299	<b>13.77</b>	57244	<b>13.78</b>
BEK-DEMICUBE-5	99	90	<b>1.10</b>	33	<b>3.00</b>	24	<b>4.12</b>	12	<b>8.25</b>	3	<b>33.00</b>	0	<b>—</b>
BEK-DEMICUBE-6	4719	4761	<b>0.99</b>	3659	<b>1.29</b>	1882	<b>2.51</b>	1741	<b>2.71</b>	665	<b>7.10</b>	588	<b>8.03</b>
BEK-DEMICUBE-7	674663	704553	<b>0.96</b>	623160	<b>1.08</b>	175870	<b>3.84</b>	175595	<b>3.84</b>	60876	<b>11.08</b>	58333	<b>11.57</b>
Mean			<b>0.97</b>		<b>1.28</b>	<b>3.45</b>	<b>3.88</b>		<b>31.54</b>	<b>24.52</b>		<b>—</b>	
Median			<b>1.00</b>		<b>1.17</b>	<b>2.51</b>	<b>2.95</b>		<b>11.65</b>	<b>5.54</b>		<b>—</b>	

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  $v_{k+1}^\top d \simeq 0$ ), C (changing the order of the vectors  $v_i$ 's by taking  $i_{k+1}$  by (3.49)), D<sub>1</sub> (pre-computation of  $2(p-r)$ ) stem vectors after the QR factorization), D<sub>2</sub> (D<sub>1</sub> and 2 additional stem vectors computed after solving a LOP, whose optimal value is nonnegative), D<sub>3</sub> (all the stem vectors are first computed and, for  $(s, \pm 1) \in S_{k+1}$ , a LOP is solved to get a handle  $d$ ), D<sub>4</sub> (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 median 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  $v_{k+1}^\top d \simeq 0$  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<sub>3</sub>) and ISF(AD<sub>4</sub>) 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  $\text{rank}(V) = 5$  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  $2\varsigma$  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.

- $\text{ISF}(\text{ABCD}_2)$  is the most efficient solver that does not compute all the stem vectors.
- The solvers  $\text{ISF}(\text{ABCD}_3)$  and  $\text{ISF}(\text{AD}_4)$  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  $\text{ISF}(\text{AD}_4)$  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 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

Problem	CPU times (in sec)								
	Simulated RC	ISF (ABCD <sub>2</sub> )		ISF (ABCD <sub>3</sub> )		ISF (AD <sub>4</sub> )		Time	Ratio
		Time	Ratio	Time	Ratio	Time	Ratio		
RAND-8-15-7	71.77	33.27	<b>2.16</b>	32.91	<b>2.18</b>	5.62	<b>12.77</b>		
RAND-9-16-8	151.39	75.45	<b>2.01</b>	82.30	<b>1.84</b>	14.43	<b>10.49</b>		
RAND-10-17-9	347.32	185.05	<b>1.88</b>	198.18	<b>1.75</b>	55.96	<b>6.21</b>		
SRAND-8-20-2	174.44	16.91	<b>10.32</b>	19.64	<b>8.88</b>	3.66	<b>47.68</b>		
SRAND-8-20-4	832.74	309.15	<b>2.69</b>	450.35	<b>1.85</b>	349.83	<b>2.38</b>		
SRAND-8-20-6	1011.30	483.97	<b>2.09</b>	732.82	<b>1.38</b>	746.49	<b>1.35</b>		
RC-2D-20-6	11.01	0.32	<b>34.71</b>	0.25	<b>43.53</b>	0.22	<b>50.95</b>		
RC-2D-20-7	19.88	0.50	<b>39.95</b>	0.50	<b>39.68</b>	0.38	<b>52.97</b>		
RC-2D-20-8	35.87	0.41	<b>87.97</b>	0.74	<b>48.56</b>	0.63	<b>56.78</b>		
RC-PERM-6	53.29	0.76	<b>70.05</b>	2.10	<b>25.41</b>	1.90	<b>28.00</b>		
RC-PERM-7	549.04	7.44	<b>73.78</b>	45.62	<b>12.04</b>	67.10	<b>8.18</b>		
RC-PERM-8	6171.22	74.93	<b>82.36</b>	1233.80	<b>5.00</b>	3355.22	<b>1.84</b>		
RC-RATIO-20-5-7	83.34	22.71	<b>3.67</b>	28.36	<b>2.94</b>	18.58	<b>4.49</b>		
RC-RATIO-20-6-7	202.09	72.04	<b>2.81</b>	101.51	<b>1.99</b>	112.12	<b>1.80</b>		
RC-RATIO-20-7-7	504.52	247.99	<b>2.03</b>	351.08	<b>1.44</b>	353.15	<b>1.43</b>		
BEK-THRESHOLD-4	0.61	0.18	<b>3.44</b>	0.11	<b>5.46</b>	0.01	<b>74.64</b>		
BEK-THRESHOLD-5	17.43	3.56	<b>4.89</b>	2.83	<b>6.16</b>	0.35	<b>50.40</b>		
BEK-THRESHOLD-6	1758.16	194.75	<b>9.03</b>	4577.26	<b>0.38</b>	6532.56	<b>0.27</b>		
BEK-RESONANCE-4	3.97	0.22	<b>17.71</b>	0.09	<b>46.12</b>	0.08	<b>48.99</b>		
BEK-RESONANCE-5	228.41	7.90	<b>28.90</b>	44.78	<b>5.10</b>	183.84	<b>1.24</b>		
BEK-RESONANCE-6	38296.20	1988.60	<b>19.26</b>	(1)	—	(1)	—		
BEK-CROSSPOLYTOPE-11	480.07	34.35	<b>13.97</b>	39.27	<b>12.22</b>	7.63	<b>62.95</b>		
BEK-CROSSPOLYTOPE-12	1579.19	108.76	<b>14.52</b>	124.66	<b>12.67</b>	25.22	<b>62.62</b>		
BEK-CROSSPOLYTOPE-13	5017.73	322.43	<b>15.56</b>	404.80	<b>12.40</b>	104.22	<b>48.15</b>		
BEK-DEMICUBE-5	0.55	0.02	<b>25.24</b>	0.01	<b>85.73</b>	0.01	<b>108.69</b>		
BEK-DEMICUBE-6	27.38	4.15	<b>6.59</b>	4.09	<b>6.69</b>	0.43	<b>63.82</b>		
BEK-DEMICUBE-7	4310.35	510.25	<b>8.45</b>	2405.08	<b>1.79</b>	6396.66	<b>0.67</b>		
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 median values of the ratios.

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].

# Chapter 4

## Additional elements on the B-differential of the minimum and hyperplane arrangements

This shorter chapter aims at providing additional material to the previous one. This material consists of rather related properties but is not specifically focused on computing  $\partial_B H(x)$  with  $H : x \mapsto \min(Ax + a, Bx + b)$ . After a few proofs, we start this chapter by giving a few examples of situations that may occur regarding regularity properties introduced in section 2.3.2.

Then, the case of smooth nonlinear functions  $F$  and  $G$  (instead of  $F(x) \equiv Ax + a$  and  $G(x) \equiv Bx + b$ ) is considered: the sets  $\{x \in \mathbb{R}^n : F_i(x) = G_i(x)\}$  for each  $i \in [1 : n]$  are not hyperplanes anymore, which clearly complicates any direct computation. We consider the linearizations of  $F$  and  $G$  at  $x$ , and propose necessary and sufficient conditions to ensure the B-differential of the minimum of these linearizations is equal to the real B-differential ( $\subseteq$  always hold but the converse is unlikely to be true in general).

The following section details a chain rule property on  $\partial_B \theta(x)$ , where  $\theta = \|H\|^2/2$ . Essentially, we show that  $\partial_B \theta(x) = \partial_B H(x)^\top H(x)$ . While this formula is not very surprising, its proof is heavily relying on the inherent structure of the componentwise minimum.

Finally, we give additional precisions concerning some instances tested in chapter 3, such as the structures and cardinals of  $\mathcal{S}$  and  $\mathfrak{S}$ . When we could, we also propose explanations of the behaviors of the tested heuristics/algorithms on some types of instances.

Let us recall that we consider the following general nonlinear complementarity problem

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

where  $F$  and  $G$  are smooth functions from  $\mathbb{R}^n$  to  $\mathbb{R}^n$ , using the reformulation via the minimum C-function, leading to the nonsmooth system and the minimization of its associated

merit function

$$H(x) = \min(F(x), G(x)) = 0, \quad \min \theta(x) := \frac{1}{2} \|H(x)\|^2.$$

We possibly consider the linear/affine case (see chapter 3) with  $F(x) = Ax + a$  and  $G(x) = Bx + b$ ,

$$0 \leqslant Ax + a \perp Bx + b \geqslant 0,$$

In both cases, we consider the following index sets:

$$\begin{aligned} \mathcal{E}(x) &:= \{i \in [1 : n] : F_i(x) = G_i(x)\}, \\ \mathcal{F}(x) &:= \{i \in [1 : n] : F_i(x) < G_i(x)\}, \\ \mathcal{G}(x) &:= \{i \in [1 : n] : F_i(x) > G_i(x)\}. \end{aligned} \tag{4.1}$$

## 4.1 Additional material from the previous chapter

Let us give the proofs of a few properties presented in the previous chapter, but were not deemed important enough for the paper.

**Proposition 4.1.1** (superset of  $\partial_B H(x)$ ). *One has*

$$\partial_B H(x) \subseteq \partial_B H_1(x) \times \cdots \times \partial_B H_m(x) = \partial_B^\times H(x). \tag{4.2}$$

In particular,  $|\partial_B H(x)| \leqslant 2^p$ .

*Proof.* The inclusion in (4.2) is clear since, when  $H'(x_k)$  converges to some  $J$ ,  $H'_i(x_k) \rightarrow J_{i,:}$ , for all  $i \in [1 : m]$ . The equality is also clear as a consequence of lemma 3.2.1 ([78, lemma 2.1.4]).

The last claim is a straightforward consequence of the fact that  $J_{i,:}$  can take two different values,  $A_{i,:}$  or  $B_{i,:}$ , only for the indices  $i \in \mathcal{E}^\neq(x)$  (recall that  $|\mathcal{E}^\neq(x)| = p$ ).  $\square$

**Proposition 4.1.2** (a link with the C-differential).  $\partial_B H(x) = \text{ext } \partial_C H(x)$ .

*Proof.* Observe first that, since  $\mathcal{S}$  given by (3.12) is contained in  $\{\pm 1\}^p$ , one has  $\mathcal{S} = \text{ext}(\text{conv}\mathcal{S})$ . To get the result, it suffices now to carry this identity into  $\mathbb{R}^{p \times n}$  thanks to the affine map  $\tau : \mathbb{R}^p \rightarrow \mathbb{R}^{p \times n}$  defined at  $s \in \mathbb{R}^p$  by

$$\tau(s) = \frac{1}{2} [(I - \text{Diag}(s))B_{\mathcal{E}^\neq(x),:} + (I + \text{Diag}(s))A_{\mathcal{E}^\neq(x),:}] .$$

The restriction of  $\tau$  to  $\mathcal{S}$  is  $\tau|_{\mathcal{S}} = \sigma^{-1}$ , defined by (3.14b). Furthermore,  $\tau$  is injective, since  $A_{i,:} \neq B_{i,:}$  for  $i \in \mathcal{E}^\neq(x)$ . Therefore, by applying  $\tau$  to both sides of the identity

$\mathcal{S} = \text{ext}(\text{conv}\mathcal{S})$ , one gets

$$\begin{aligned}\tau(\mathcal{S}) &= \text{ext}(\tau(\text{conv}\mathcal{S})) && [\text{injectivity of } \tau \text{ [105, prop. 2.12(2)]}] \\ &= \text{ext}(\text{conv}(\tau(\mathcal{S}))) && [\text{affinity of } \tau \text{ [105, prop. 2.5(1)]}].\end{aligned}$$

The result now follows from the fact that  $\tau(\mathcal{S}) = \sigma^{-1}(\mathcal{S}) = \partial_B H(x)$  (proposition 3.3.4) and  $\partial_C H(x) = \text{conv} \partial_B H(x)$ .  $\square$

Finally, let us mention a way, suggested by a reviewer of [77] before publication approval, of proceeding that seems to us to be too restrictive 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  $p > n$ . If  $\partial_B H(x)$  is the main concern, one can write  $H(x) = Ax + a - K(x)$ , where  $K(x) := P_{\mathbb{R}_+^n}[Mx + q]$ ,  $P_{\mathbb{R}_+^n}$  is the orthogonal projector on the positive orthant,  $M = A - B$  and  $q = a - b$ , so that

$$\partial_B H(x) = A - \partial_B K(x). \quad (4.4)$$

To take advantage of the explicit formula of  $\partial_B P_{\mathbb{R}_+^n}$ , one can look for conditions ensuring that the chain rule applies for the composition defining the map  $K : \mathbb{R}^n \rightarrow \mathbb{R}^n$ . We claim 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. Therefore, this approach is of too limited an interest.

Suppose indeed that the chain rule applies for the computation of  $\partial_B K(x)$ . Then, one would have

$$\partial_B K(x) = [\partial_B P_{\mathbb{R}_+^n}(Mx + q)]M.$$

Now,  $\partial_B P_{\mathbb{R}_+^n}(z)$  is known explicitly as the Cartesian product of

$$[\partial_B P_{\mathbb{R}_+^n}(z)]_i = \begin{cases} \{0\} & \text{if } z_i < 0, \\ \{0, 1\} & \text{if } z_i = 0, \\ \{1\} & \text{if } z_i > 0, \end{cases}$$

for  $i \in [1 : n]$ . Therefore,  $\partial_B K(x)$  is the Cartesian product of

$$[\partial_B K(z)]_i = \begin{cases} \{0\} & \text{if } (Ax + a)_i < (Bx + b)_i, \\ \{0, A_{i,:} - B_{i,:}\} & \text{if } (Ax + a)_i = (Bx + b)_i, \\ \{A_{i,:} - B_{i,:}\} & \text{if } (Ax + a)_i > (Bx + b)_i, \end{cases}$$

for  $i \in [1 : n]$ . With (4.4),  $\partial_B H(x)$  reads as the Cartesian product of

$$[\partial_B H(x)]_i = \begin{cases} \{A_{i,:}\} & \text{if } (Ax + a)_i < (Bx + b)_i, \\ \{A_{i,:}, B_{i,:}\} & \text{if } (Ax + a)_i = (Bx + b)_i, \\ \{B_{i,:}\} & \text{if } (Ax + a)_i > (Bx + b)_i, \end{cases}$$

for  $i \in [1 : n]$ . This is the formula of a complete B-differential.

## 4.2 Regularity notions and counterexamples

This section aims at describing a few (counter-)examples to show how the notions of BD-regularity 2.3.14 and strong BD-regularity interact with the B-differential. They show in particular that: BD-regularity does not imply strong BD-regularity, that strong BD-regularity does not imply the B-differential is complete and that the B-differential may be complete despite all its elements being singular.

Let us define  $V := B_{\mathcal{E}(x),:}^T - A_{\mathcal{E}(x),:}^T$ , the Jacobian matrices of  $\partial_B H(x)$ , the B-differential of  $H$ , are the matrices  $J(s)$  defined by

$$J(s)_{\mathcal{F}(x) \cup \mathcal{G}(x),:} = \begin{bmatrix} F'_{\mathcal{F}(x)}(x) \\ G'_{\mathcal{G}(x)}(x) \end{bmatrix}, \quad J(s)_{\mathcal{E}(x),:} = \frac{s+e}{2} \cdot F'_{\mathcal{E}(x)}(x) + \frac{e-s}{2} \cdot G'_{\mathcal{E}(x)}(x).$$

for some vectors  $s$  defined by

$$s \in \{\{\pm 1\}^{\mathcal{E}(x)} : \exists d, s \cdot V^T d > 0\}.$$

These examples show the peculiarity of  $\partial_B H(x)$ , where its elements are determined by the matrix  $F'(x) - G'(x)$  but their (non)singularity is determined by  $F'(x)$  and  $G'(x)$ .

**Example 4.2.1** (BD regular but not strongly BD regular). Let  $\mathcal{F}(x) = \emptyset = \mathcal{G}(x)$  and  $\mathcal{E}(x) = \{1, 2\}$ . Let  $A$  and  $B$  be defined by

$$A = I_2, \quad B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Since  $V = B^T - A^T = [0 \ 1; 1 \ 0]$  is injective, one has  $\mathcal{S} = \{\pm 1\}^2$ . However, for  $s = (-1, -1)$ , one has  $J(s) = B$  which is singular, therefore there is no strong BD regularity. Let us verify that BD regularity holds. This reads

$$\begin{pmatrix} \min(d_1, d_1 + d_2) \\ \min(d_2, d_1 + d_2) \end{pmatrix},$$

which can be zero if and only if  $(d_1, d_2) = 0$ . Indeed, BD-regularity reads (definition 2.3.14)  $d \neq 0 \Rightarrow H'(x; d) \neq 0$ .  $\square$

**Example 4.2.2** (all matrices are invertible but not all in the B-differential). Consider the following example:

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad a = 0, \quad B = \begin{bmatrix} 2 & 1 \\ 2 & 3 \end{bmatrix}, \quad b = 0, \quad x = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

One clearly gets that  $Ax + a = Bx + b = x$ . Moreover, the 4 potential Jacobian matrices are

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 \\ 2 & 3 \end{bmatrix}, \quad \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 2 & 1 \\ 2 & 3 \end{bmatrix},$$

which are all nonsingular. However, at  $x$ , the matrix involved in the determination of  $\partial_B H(x)$  is

$$V = (B - A)^\top = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix},$$

therefore only two out of the four Jacobian matrices belong to  $\partial_B H(x)$ .  $\square$

**Example 4.2.3** (complete B-differential & singular Jacobians). Consider the following data:

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 2 & 2 \end{bmatrix}, \quad a = 0, \quad B = \begin{bmatrix} -1 & 3 & 2 \\ 1 & 0 & 1 \\ -1 & 0 & -1 \end{bmatrix}, \quad b = 0, \quad x = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}.$$

At the point  $x$ , one gets

$$Ax + a = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad Bx + b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

meaning  $\mathcal{E}(x) = \{2, 3\}$ . Now, the matrix involved in the B-differential is

$$V = (B - A)_{\mathcal{E}(x)}^\top = \begin{bmatrix} 1 & -1 \\ -1 & -2 \\ 0 & -3 \end{bmatrix}.$$

Since its two columns are independent, the B-differential is complete and the four Jacobians

$$\begin{bmatrix} -1 & 3 & 2 \\ 0 & 1 & 1 \\ 0 & 2 & 2 \end{bmatrix}, \quad \begin{bmatrix} -1 & 3 & 2 \\ 0 & 1 & 1 \\ -1 & 0 & -1 \end{bmatrix}, \quad \begin{bmatrix} -1 & 3 & 2 \\ 1 & 0 & 1 \\ 0 & 2 & 2 \end{bmatrix}, \quad \begin{bmatrix} -1 & 3 & 2 \\ 1 & 0 & 1 \\ -1 & 0 & -1 \end{bmatrix}$$

are all singular.  $\square$

## 4.3 B-differential of the minimum of nonlinear $F$ and $G$

### 4.3.1 Differentials of $H$

This first section discusses a few properties of  $\partial_B H(x)$  and  $\partial_C H(x)$ . The main element we discuss is the link between  $\partial_B H(x)$  and  $\partial_B(\mathcal{L}_x H)(x)$ , the B-differential of the minimum of the linearizations of  $F$  and  $G$ . More precisely, define

$$(\mathcal{L}_x H)(y) := \min (F(x) + F'(x)(y - x), G(x) + G'(x)(y - x)). \quad (4.5)$$

The main use of the linearization is as follows. Since it is a minimum of affine functions, it is governed by the properties described in chapter 3. Then, we shall see that the differentials of  $\mathcal{L}_x H$  are subsets of the “real” B-differential of  $H$ .

Naturally, if the functions  $F$  and  $G$  are affine, they are equal to their linearizations and equality holds in (4.6a) and (4.6b). Otherwise, the equality may easily not hold, and the cardinal of the B-differential may be odd (counter-example 4.3.2 below).

**Proposition 4.3.1** (subset of  $\partial_B H(x)$ ). *Suppose that  $F$  and  $G$  are continuously differentiable at  $x$ . Let  $\mathcal{L}_x$  be the operator defined by (4.5). Then,*

$$\partial_B(\mathcal{L}_x H)(x) \subseteq \partial_B H(x), \quad (4.6a)$$

$$\partial_C(\mathcal{L}_x H)(x) \subseteq \partial_C H(x). \quad (4.6b)$$

*Proof.* The inclusion (4.6b) can be deduced from (4.6a) by taking the convex hulls of its two sides, so that we just have to focus on (4.6a).

First, observe that the index sets  $\mathcal{F}(x)$ ,  $\mathcal{E}^{\neq}(x)$ ,  $\mathcal{E}^=(x)$  and  $\mathcal{G}(x)$  are identical for  $H$  and  $\mathcal{L}_x H$ . Let  $\tilde{J} \in \partial_B(\mathcal{L}_x H)(x)$ . We want to show that  $\tilde{J} \in \partial_B H(x)$ . By equation (3.10), one gets

$$\tilde{J}_{i,:} = \begin{cases} F'_i(x) & \text{if } i \in \mathcal{F}(x), \\ F'_i(x) = G'_i(x) & \text{if } i \in \mathcal{E}^=(x), \\ F'_i(x) \text{ or } G'_i(x) & \text{if } i \in \mathcal{E}^{\neq}(x), \\ G'_i(x) & \text{if } i \in \mathcal{G}(x). \end{cases} \quad (4.7a)$$

In addition, one can find a direction  $d \in \mathbb{R}^n$  such that, for any  $i \in \mathcal{E}^{\neq}(x)$ , one has

$$F'_i(x)d - G'_i(x)d < 0, \quad \text{if } \tilde{J}_{i,:} = F'_i(x), \quad (4.7b)$$

$$F'_i(x)d - G'_i(x)d > 0, \quad \text{if } \tilde{J}_{i,:} = G'_i(x). \quad (4.7c)$$

Let us now construct a Jacobian  $J \in \partial_B H(x)$  and show that  $J = \tilde{J}$ , which will conclude the proof of the proposition. Consider the sequence  $\{x_k\}$  defined by

$$x_k := x + t_k d + \sigma(t_k), \quad (4.7d)$$

where  $\{t_k\} \downarrow 0$  and  $\sigma : \mathbb{R}_+ \rightarrow \mathbb{R}^n$  is a (small perturbation) function that need not be continuous but is chosen such that  $\sigma(0) = 0$ ,  $\sigma(t_k) = o(t_k)$  and  $x_k \in \mathcal{D}_H$  (this is possible by Rademacher’s theorem [209]). By extracting a subsequence  $\mathcal{K}$  of  $\mathbb{N}$  if necessary, one can assume that, for each index  $i \in [1 : n]$ , one of the following three properties holds

$$\forall k \in \mathcal{K} : \quad F_i(x_k) < G_i(x_k), \quad (4.7e)$$

$$\forall k \in \mathcal{K} : \quad F_i(x_k) = G_i(x_k), \quad (4.7f)$$

$$\forall k \in \mathcal{K} : \quad F_i(x_k) > G_i(x_k). \quad (4.7g)$$

By construction of  $\{x_k\}$ ,  $x_k \in \mathcal{D}_H$ , so that  $H'(x_k)$  exists. Let us show that, for  $i \in [1 : n]$  and when  $k \rightarrow \infty$  in  $\mathcal{K}$ , one has

$$(4.7e) \text{ holds} \implies H'_i(x_k) \rightarrow F'_i(x), \quad (4.7h)$$

$$(4.7f) \text{ holds} \implies H'_i(x_k) \rightarrow F'_i(x) = G'_i(x), \quad (4.7i)$$

$$(4.7g) \text{ holds} \implies H'_i(x_k) \rightarrow G'_i(x). \quad (4.7j)$$

The implication (4.7h) (resp. (4.7j)) is clear, since  $H'_i(x_k) = F'_i(x_k)$  (resp.  $H'_i(x_k) = G'_i(x_k)$ ),  $x_k \rightarrow x$  and  $F'_i$  (resp.  $G'_i$ ) is continuous at  $x$ . The implication (4.7i) comes from the fact that, when (4.7f) holds,  $H'_i(x_k) = F'_i(x_k) = G'_i(x_k)$  by the differentiability of  $H_i$  at  $x_k$  (use lemma 3.2.1), implying again that  $H'_i(x_k) \rightarrow F'_i(x) = G'_i(x)$ . By definition, the limit  $J$  of  $H'(x_k)$  is in  $\partial_B H(x)$ . It remains to show that  $J = \tilde{J}$  to conclude the proof of the proposition.

Let us look at an arbitrary row  $i \in [1 : n]$  of  $J \in \partial_B H(x)$  and  $\tilde{J} \in \partial_B(\mathcal{L}_x H)(x)$  and show that  $J_{i,:} = \tilde{J}_{i,:}$ . By the differentiability property of  $F$  and  $G$  at  $x$  and by (4.7d) with  $\sigma(t_k) = o(t_k)$ , one has for  $k \in \mathcal{K}$ :

$$\begin{cases} F_i(x_k) = F_i(x) + t_k F'_i(x)d + o(t_k), \\ G_i(x_k) = G_i(x) + t_k G'_i(x)d + o(t_k). \end{cases} \quad (4.7k)$$

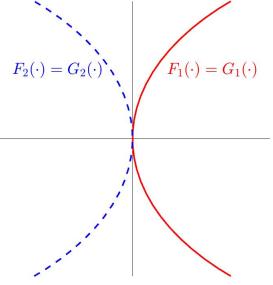
- If  $i \in \mathcal{F}(x)$  (resp.  $i \in \mathcal{G}(x)$ ), one has  $F_i(x) < G_i(x)$  (resp.  $F_i(x) > G_i(x)$ ), so that, by continuity of  $F$  and  $G$  at  $x$ , one also has  $F_i(x_k) < G_i(x_k)$  (resp.  $F_i(x_k) > G_i(x_k)$ ). Therefore, case (4.7e)+(4.7h) (resp. case (4.7g)+(4.7j)) applies and we have  $J_{i,:} = F'_i(x)$  (resp.  $J_{i,:} = G'_i(x)$ ). Hence  $J_{i,:} = \tilde{J}_{i,:}$  by (4.7a).
- If  $i \in \mathcal{E}^\neq(x)$  and  $\tilde{J}_{i,:} = F'_i(x)$  (resp.  $\tilde{J}_{i,:} = G'_i(x)$ ), one has  $F_i(x) = G_i(x)$ , so that (4.7b) (resp. (4.7c)) and (4.7k) yield  $F(x_k) < G(x_k)$  (resp.  $F(x_k) > G(x_k)$ ) for  $k$  large enough. Therefore, case (4.7e)+(4.7h) (resp. case (4.7g)+(4.7j)) applies and we have  $J_{i,:} = F'_i(x)$  (resp.  $J_{i,:} = G'_i(x)$ ). Hence  $J_{i,:} = \tilde{J}_{i,:}$  in this case as well.
- Finally, if  $i \in \mathcal{E}^=(x)$ ,  $\tilde{J}_{i,:} = F'_i(x) = G'_i(x)$ , by (4.7a). Let us now look at the value of  $J_{i,:}$  by considering the three possible cases (4.7e)-(4.7g).
  - If (4.7e) (resp. (4.7g)) occurs, one has  $H'_i(x_k) = F'_i(x_k)$  (resp.  $H'_i(x_k) = G'_i(x_k)$ ) and we get at the limit  $J_{i,:} = F'_i(x)$  (resp.  $J_{i,:} = G'_i(x)$ ). Hence  $J_{i,:} = \tilde{J}_{i,:}$  in these cases.
  - If (4.7f) occurs, one has  $F'_i(x_k) = G'_i(x_k)$  from lemma 3.2.1, because  $F_i(x_k) = G_i(x_k)$  and  $x_k \in \mathcal{D}_H$ . At the limit, we get  $J_{i,:} = F'_i(x) = G'_i(x)$ . Hence  $J_{i,:} = \tilde{J}_{i,:}$  in this case as well.

□

**Counter-example 4.3.2** (no equality in (4.6)). Consider the case where  $n = 2$ ,  $F(x) \equiv x$

and  $G(x) \equiv (x_1^2 + x_2^2 - x_1, x_1^2 + x_2^2 + 2x_1 + x_2)$ :

$$\begin{aligned} H(x) &= \min \left( x, \begin{pmatrix} x_1^2 + x_2^2 - x_1 \\ x_1^2 + x_2^2 + 2x_1 + x_2 \end{pmatrix} \right) \\ (\mathcal{L}_0 H)(x) &= \min \left( x, \begin{pmatrix} -1 & 0 \\ 2 & 1 \end{pmatrix} x \right). \end{aligned}$$



One has  $\mathcal{F}(0) = \mathcal{G}(0) = \emptyset$  and  $\mathcal{E}(0) = \{1, 2\}$ , and the two involved matrices  $A$  and  $B$  are

$$A = I \text{ and } B = \begin{pmatrix} -1 & 0 \\ 2 & 1 \end{pmatrix}, \quad \text{thus } V := (B - A)^\top = \begin{pmatrix} -2 & 2 \\ 0 & 0 \end{pmatrix}.$$

Since  $\text{rank}(V) = 1 < 2$ , there is less than  $2^2$  elements in  $\partial_B(\mathcal{L}_0 H)(0)$ . Actually, there are only two sign vectors  $s \in \{\pm 1\}^2$  such that  $s \cdot V^\top d > 0$  is feasible for  $d$ , since one must have  $s_1 = -s_2$ , namely  $s = (1, -1)$  and  $s = (-1, 1)$ . From this observation, one deduces that

$$\partial_B(\mathcal{L}_0 H)(0) = \left\{ \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \right\}.$$

The jacobians in  $\partial_B(\mathcal{L}_0 H)(0)$  are also in  $\partial_B H(0)$  (proposition 4.3.1), but the latter contains also one other Jacobian. Indeed, take points of the form  $x^t = (0, t)$ , with  $t \downarrow 0$  or  $t \uparrow 0$ . Then  $F_1(x^t) < G_1(x^t)$  and  $F_2(x^t) < G_2(x^t)$  for any  $t \neq 0$ . As a result,  $H'(x^t) = I$  and one has the Jacobian  $I \in \partial_B H(0)$ , which is not in  $\partial_B(\mathcal{L}_0 H)(0)$ . This counter-example also shows that, although  $|\partial_B(\mathcal{L}_x H)_{\mathcal{E}^\neq(x)}(x)|$  is even,  $|\partial_B H_{\mathcal{E}^\neq(x)}(x)|$  may be odd.  $\square$

Let us use the notation:  $(G'(x) - F'(x))_{\mathcal{E}^\neq(x)} = V^\top$ ,  $v_i = \nabla G_i(x) - \nabla F_i(x)$  and  $\mathcal{V}_i := \{y \in \mathbb{R}^n : F_i(y) = G_i(y)\}$  for  $i \in \mathcal{E}^\neq(x)$ . When  $V$  is surjective, the B-differential of the linearization is complete, meaning equality holds in (4.6). There is actually an equivalence between these properties. To have the equality but not necessarily the completeness, one must weaken the surjectivity hypothesis.

**Proposition 4.3.3** (NSC(s) to have equality in (4.6)). *Suppose that  $F$  and  $G$  are  $\mathcal{C}^1$  at  $x$ . Then, the following relations hold: (i)  $\Leftrightarrow$  (ii)  $\Leftrightarrow$  (iii)  $\Rightarrow$  (iv)  $\Leftrightarrow$  (v)*

(i) *for any  $\alpha \in \mathbb{R}^{|\mathcal{E}^\neq(x)|} \setminus \{0\}$  and  $s \in \{\pm 1\}^{|\mathcal{E}^\neq(x)|}$  such that  $V\alpha = 0$  and  $s \cdot \alpha \geq 0$ , there exists a neighborhood  $\mathcal{U}$  of  $x$  such that*

$$\{x' \in \mathcal{U} : s \cdot [G(x') - F(x')]_{\mathcal{E}^\neq(x)} > 0\} = \emptyset, \quad (4.8)$$

(ii) *same condition as (i) except one uses*

$$\{x' \in \mathcal{U} : s_I \cdot [G(x') - F(x')]_I > 0\} = \emptyset, \quad (4.9)$$

*where  $I := \{i \in \mathcal{E}^\neq(x) : \alpha_i \neq 0\}$ .*

- (iii) equality holds in (4.6), that is  $\partial_B(\mathcal{L}_x H)(x) = \partial_B H(x)$ .
- (iv) for any  $i_0 \in \mathcal{E}^\neq(x)$ , such that

$$v_{i_0} = \sum_{i \in I_0} \alpha_i v_i, \quad \text{for some } I_0 \subseteq \mathcal{E}^\neq(x) \setminus \{i_0\} \text{ and } \alpha_i \in \mathbb{R}^*, \quad (4.10a)$$

there is a neighborhood  $\mathcal{U}$  of  $x$  such that

$$\bigcap_{i \in I_0} (\mathcal{V}_i \cap \mathcal{U}) \subseteq \mathcal{V}_{i_0}, \quad (4.10b)$$

- (v) for any  $i_0 \in \mathcal{E}^\neq(x)$ , such that (4.10a) holds with linearly independent vectors  $\{v_i : i \in I_0\}$ , there is a neighborhood  $\mathcal{U}$  of  $x$  such that (4.10b) holds.

*Proof.* [(i)  $\Rightarrow$  (ii)] Suppose that (i) holds and that  $\alpha_i = 0$  for some  $i \in \mathcal{E}^\neq(x)$ . Then, (4.8) holds for  $s_i = \pm 1$ , so that for  $\mathcal{U}$  sufficiently small

$$\{x' \in \mathcal{U} : s_j[G_j(x') - F_j(x')] > 0 \text{ for all } j \neq i \text{ and } G_i(x') - F_i(x') > 0\} = \emptyset, \quad (4.11a)$$

$$\{x' \in \mathcal{U} : s_j[G_j(x') - F_j(x')] > 0 \text{ for all } j \neq i \text{ and } G_i(x') - F_i(x') < 0\} = \emptyset. \quad (4.11b)$$

We claim that, for  $\mathcal{U}$  sufficiently small, one also has

$$\{x' \in \mathcal{U} : s_j[G_j(x') - F_j(x')] > 0 \text{ for all } j \neq i \text{ and } G_i(x') - F_i(x') = 0\} = \emptyset. \quad (4.11c)$$

Indeed, if this was not true, one could find an  $x'$  in the set in the left-hand side of (4.11c), with  $\mathcal{U}$  sufficiently small to have  $F'_i(x') \neq G'_i(x')$  (recall that  $i \in \mathcal{E}^\neq(x)$ ). But, then,  $x_t := x' + t[G'_i(x') - F'_i(x')]$  with  $t > 0$  sufficiently small would be such that  $s_j[G_j(x_t) - F_j(x_t)] > 0$  for all  $j \neq i$  and

$$\begin{aligned} G_i(x_t) - F_i(x_t) &= G_i(x') - F_i(x') + t[G'_i(x') - F'_i(x')]^2 + o(t) \\ &= t[G'_i(x') - F'_i(x')]^2 + o(t) \quad [G_i(x') = F_i(x')] \\ &> 0 \quad [G'_i(x') \neq F'_i(x')]. \end{aligned}$$

Then,  $x_t$  would be in the set in the left-hand side of (4.11a), contradicting its emptiness. Combining (4.11a)-(4.11c), we get

$$\{x' \in \mathcal{U} : s_j[G_j(x') - F_j(x')] > 0 \text{ for all } j \neq i\} = \emptyset.$$

One can pursue in the same way with the other indices giving a zero component to  $\alpha$ . This shows that (i)  $\Rightarrow$  (ii).

[(ii)  $\Rightarrow$  (i)] This implication is clear since the set in (4.8) is larger than the one in (4.9).

[(i)  $\Rightarrow$  (iii)] We proceed by contraposition, assuming that (iii) does not hold. Then, by the inclusion (4.6), there is  $J \in \partial_B H(x)$  that is not in  $\partial_B(\mathcal{L}_x H)(x)$ . By the membership  $J \in \partial_B H(x)$ , there is a sequence  $\{x_k\} \subseteq \mathcal{D}_H$  converging to  $x$  such that  $H'(x_k) \rightarrow J$ .

The examination of the sequence  $\{x_k\}$  allows us to determine  $J$ . Since,  $\{x_k\} \subseteq \mathcal{D}_H$  and  $F'_i(x) \neq G'_i(x)$  for  $i \in \mathcal{E}^\neq(x)$ , one must have  $F_i(x_k) \neq G_i(x_k)$  for  $i \in \mathcal{E}^\neq(x)$  and  $k$  sufficiently large (lemma 3.2.1). Then, one can find a partition  $(I_-, I_+)$  of  $\mathcal{E}^\neq(x)$  and a subsequence of indices  $k$  such that

$$\begin{cases} G_i(x_k) < F_i(x_k) & \text{for } i \in I_-, \\ G_i(x_k) > F_i(x_k) & \text{for } i \in I_+. \end{cases} \quad (4.12a)$$

This implies that for  $k$  sufficiently large in the selected subsequence (for the indices in  $\mathcal{E}^=(x)$ , use again lemma 3.2.1<sup>1</sup>):

$$H'_i(x_k) = \begin{cases} F'_i(x_k) & \text{if } i \in \mathcal{F}(x), \\ F'_i(x_k) \text{ or } G'_i(x_k) & \text{if } i \in \mathcal{E}^=(x), \\ G'_i(x_k) & \text{if } i \in I_- \subseteq \mathcal{E}^\neq(x), \\ F'_i(x_k) & \text{if } i \in I_+ \subseteq \mathcal{E}^\neq(x), \\ G'_i(x_k) & \text{if } i \in \mathcal{G}(x) \end{cases}$$

and therefore, the Jacobian  $J \in \partial_B H(x)$  quoted above has its  $i$ th row given by

$$J_{i,:} = \begin{cases} F'_i(x) & \text{if } i \in \mathcal{F}(x), \\ F'_i(x) = G'_i(x) & \text{if } i \in \mathcal{E}^=(x), \\ G'_i(x) & \text{if } i \in I_- \subseteq \mathcal{E}^\neq(x), \\ F'_i(x) & \text{if } i \in I_+ \subseteq \mathcal{E}^\neq(x), \\ G'_i(x) & \text{if } i \in \mathcal{G}(x). \end{cases} \quad (4.12b)$$

Now, define  $s \in \{\pm 1\}^{|\mathcal{E}^\neq(x)|}$  by

$$s_i = \begin{cases} -1 & \text{if } i \in I_-, \\ +1 & \text{if } i \in I_+. \end{cases} \quad (4.12c)$$

Since the Jacobian  $J$  given by (4.12b) is not in  $\partial_B(\mathcal{L}_x H)(x)$ , we know that

$$\nexists d \in \mathbb{R}^n : s \cdot V^\top d > 0.$$

Now, Gordan's alternative implies that one can find  $\alpha \in \mathbb{R}^{|\mathcal{E}^\neq(x)|} \setminus \{0\}$  such that

$$V\alpha = 0 \quad \text{and} \quad s \cdot \alpha \geq 0.$$

We see that this pair  $(\alpha, s)$  satisfies the properties in the premise of (i), but, by (4.12a) and (4.12c), there is a sequence  $\{x_k\} \rightarrow x$  such that

$$s \cdot [G(x_k) - F(x_k)]_{\mathcal{E}^\neq(x)} > 0,$$

which is in contradiction with (4.8). Therefore, (i) does not holds, as expected.

---

<sup>1</sup>The possible values of  $H'(x_k)$  are clear if  $F_i(x_k) \neq G_i(x_k)$ . If  $F_i(x_k) = G_i(x_k)$ , one must have  $F'_i(x_k) = G'_i(x_k)$  and  $H'_i(x_k) = F'_i(x_k) = G'_i(x_k)$  by the differentiability of  $H_i$  at  $x_k$  (lemma 3.2.1).

$[(iii) \Rightarrow (i)]$  We also proceed by contraposition, assuming that (i) does not hold. Then, there exists a pair  $(\alpha, s) \in (\mathbb{R}^{|\mathcal{E}^\neq(x)|} \setminus \{0\}) \times \{\pm 1\}^{|\mathcal{E}^\neq(x)|}$  such that

$$V\alpha = 0 \quad \text{and} \quad s \cdot \alpha \geq 0, \quad (4.12d)$$

but no neighborhood  $\mathcal{U}$  of  $x$  such that (4.8) holds. One deduces from this last fact that there is a sequence  $\{x_k\} \rightarrow x$  such that

$$s \cdot [G(x_k) - F(x_k)]_{\mathcal{E}^\neq(x)} > 0, \quad (4.12e)$$

By (4.12d) and Gordan's alternative,

$$\nexists d \in \mathbb{R}^n : \quad s \cdot V^\top d > 0.$$

Then, this implies that there is no  $J \in \partial_B(\mathcal{L}_x H)(x)$  satisfying<sup>2</sup>

$$J_{i,:} := \begin{cases} F'_i(x) & \text{if } s_i = +1, \\ G'_i(x) & \text{if } s_i = -1. \end{cases} \quad (4.12f)$$

Let us now prove that there is a  $J$  in  $\partial_B H(x)$  that satisfies (4.12f), which will show that (i) does not hold, as expected. By (4.12e), the continuity of  $F$  and  $G$  at  $x$  and the fact that  $x_k \rightarrow x$ , one has (note that nothing is said about the indices in  $\mathcal{E}^=(x)$  or the differentiability of  $H$  at  $x_k$ )

$$\begin{cases} F_i(x_k) < G_i(x_k) & \text{if } i \in \mathcal{F}(x), \\ F_i(x_k) < G_i(x_k) & \text{if } i \in \mathcal{E}^\neq(x) \text{ and } s_i = +1, \\ F_i(x_k) > G_i(x_k) & \text{if } i \in \mathcal{E}^\neq(x) \text{ and } s_i = -1, \\ F_i(x_k) > G_i(x_k) & \text{if } i \in \mathcal{G}(x). \end{cases} \quad (4.12g)$$

We claim that one can slightly perturb the sequence  $\{x_k\}$  to get a sequence  $\{x'_k\} \subseteq \mathcal{D}_H$  converging to  $x$  and satisfying (hence, the differentiability of  $H$  at  $x'_k$  is now guaranteed)

$$\begin{cases} F_i(x'_k) < G_i(x'_k) & \text{if } i \in \mathcal{F}(x), \\ F_i(x'_k) < G_i(x'_k) & \text{if } i \in \mathcal{E}^\neq(x) \text{ and } s_i = +1, \\ F_i(x'_k) > G_i(x'_k) & \text{if } i \in \mathcal{E}^\neq(x) \text{ and } s_i = -1, \\ F_i(x'_k) > G_i(x'_k) & \text{if } i \in \mathcal{G}(x). \end{cases}$$

Indeed, to prove this, one only has to say what is done for the indices in  $\mathcal{E}^=(x)$  to ensure that  $\{x'_k\}$  is in  $\mathcal{D}_H$  and converges to  $x$ , since the inequalities in (4.12g) are preserved by small perturbation due to the continuity of  $F$  and  $G$  at  $x_k$  and guarantee that the corresponding

---

<sup>2</sup>Let us prove this claim by contraposition. If such a  $J \in \partial_B(\mathcal{L}_x H)(x)$  would exist, one would have by the previous displayed claim

$$\nexists d \in \mathbb{R}^n : \quad \begin{cases} [G'_i(x) - F'_i(x)]d > 0 & \text{if } J_{i,:} = F'_i(x), \\ [G'_i(x) - F'_i(x)]d < 0 & \text{if } J_{i,:} = G'_i(x). \end{cases}$$

This would be in contradiction with  $J \in \partial_B(\mathcal{L}_x H)(x)$ .

component of  $H$  is differentiable at  $x'_k$ . Take a first index  $i \in \mathcal{E}^=(x)$ . If  $F_i(x') = G_i(x')$  for  $x'$  near  $x_k$ , then  $H_i$  is differentiable at  $x_k$  and no perturbation of  $x_k$  is needed. Otherwise, there is an  $x'$  arbitrary close to  $x_k$  with  $F_i(x') \neq G_i(x')$ ;  $H_i$  is differentiable at this  $x'$ . Proceeding like this for the possible other indices in  $\mathcal{E}^=(x)$ , while preserving the strict inequalities obtained so far, we get a sequence  $\{x'_k\} \subseteq \mathcal{D}_H$ . In this procedure,  $x'_k$  can be taken arbitrarily close to  $x_k$  in order to guarantee that  $\{x'_k\} \rightarrow x$ . Since

$$\begin{cases} H'_i(x'_k) = F'_i(x'_k) & \text{if } i \in \mathcal{E}^\neq(x) \text{ and } s_i = +1, \\ H'_i(x'_k) = G'_i(x'_k) & \text{if } i \in \mathcal{E}^\neq(x) \text{ and } s_i = -1, \end{cases}$$

we see that  $H'(x'_k)$  converges to a Jacobian  $J \in \partial_B H(x)$  satisfying (4.12f), as expected.

$[(iv) \Rightarrow (v)]$  This is clear, since (4.9) must hold for fewer triplets  $(i_0, I_0, \alpha)$  satisfying (4.8).

$[(iii) \Rightarrow (v)]$  We proceed by contraposition, assuming that (v) does not hold. Then, there is an index  $i_0 \in \mathcal{E}^\neq(x)$ , an index set  $I_0 \subseteq \mathcal{E}^\neq(x) \setminus \{i_0\}$  and  $\alpha_i \in \mathbb{R}^*$ , such that (4.8) holds with linearly independent vectors  $\{v_i : i \in I_0\}$ , but there is no neighborhood  $\mathcal{U}$  such that (4.9) holds.

One deduces from this last fact that there is a sequence  $\{x_k\} \rightarrow x$  such that  $x_k \in \cap_{i \in I_0} \mathcal{V}_i$ , but  $x_k \notin \mathcal{V}_{i_0}$  or  $F_{i_0}(x_k) \neq G_{i_0}(x_k)$ . By extracting a subsequence if needed, one may assume that one has  $(F_{i_0}(x_k) > G_{i_0}(x_k) \text{ for all } k)$  or  $(F_{i_0}(x_k) < G_{i_0}(x_k) \text{ for all } k)$ . Suppose that the first case occurs (one can proceed similarly in the second case). Hence, for all  $k \rightarrow \infty$ ,

$$\begin{cases} F_i(x_k) = G_i(x_k), & \text{for } i \in I_0, \\ F_{i_0}(x_k) > G_{i_0}(x_k). \end{cases} \quad (4.13a)$$

By Gordan's alternative, (4.8) implies that one cannot find a direction  $d$  such that (the sense of the inequalities in (4.13b) below is taken according to the fact that  $F_{i_0}(x_k) < G_{i_0}(x_k)$  in (4.13a); reverse the inequalities if  $F_{i_0}(x_k) > G_{i_0}(x_k)$  holds)

$$-v_{i_0}^\top d > 0 \quad \text{and} \quad \operatorname{sgn}(\alpha_i)v_i^\top d > 0, \quad \text{for } i \in I_0. \quad (4.13b)$$

This implies that there is no  $J \in \partial_B(\mathcal{L}_x H)(x)$  satisfying

$$J_{i,:} := \begin{cases} G'_i(x) & \text{if } i = i_0, \\ F'_i(x) & \text{if } i \in I_0 \text{ and } \alpha_i > 0, \\ G'_i(x) & \text{if } i \in I_0 \text{ and } \alpha_i < 0. \end{cases} \quad (4.13c)$$

We prove next that there is a  $J$  in  $\partial_B H(x)$  that satisfies (4.13c), which will show that (iii) does not hold, as expected.

Since the  $v_i$ 's, for  $i \in I_0$ , are linearly independent, one can find a direction  $p \in \mathbb{R}^n$  such that

$$v_i^\top p = \operatorname{sgn}(\alpha_i), \quad \text{for } i \in I_0. \quad (4.13d)$$

This direction is used to define a sequence  $\{x'_k\}$ , which is a small perturbation of  $\{x_k\}$ , by

$$x'_k := x_k + t_k p + \sigma_k,$$

where the  $\sigma_k$ 's are (small) perturbation vectors in  $\mathbb{R}$  such that  $\sigma_k = o(t_k)$  and  $x'_k \in \mathcal{D}_H$  (this precaution is possible by Rademacher's theorem and it is useful below for the indices  $i \in \mathcal{E}^\neq(x) \setminus (I_0 \cup \{i_0\})$  if any) and where  $t_k = o(\|x_k - x\|)$  is taken positive and sufficiently small to ensure

$$F_{i_0}(x'_k) > G_{i_0}(x'_k) \quad (4.13e)$$

(this is possible by  $F_{i_0}(x_k) > G_{i_0}(x_k)$  in (4.13a) and by the continuity of  $G_{i_0} - F_{i_0}$ ). Now, for  $i \in [1 : n]$ , the mean value theorem, which holds when  $F_i$  is differentiable near  $x$ , ensures that, for  $k$  large enough

$$\|F_i(x'_k) - F_i(x_k) - F'_i(x)(t_k p + \sigma_k)\| \leq \left( \sup_{z \text{ near } x} \|F'_i(z) - F'_i(x)\| \right) \|t_k p + \sigma_k\|.$$

Proceeding similarly for  $G_i$  and using the continuity of  $G'_i - F'_i$  at  $x$ , one has when  $k \rightarrow \infty$ :

$$\begin{aligned} F_i(x'_k) &= F_i(x_k) + F'_i(x)(t_k p) + o(t_k), \\ G_i(x'_k) &= G_i(x_k) + G'_i(x)(t_k p) + o(t_k). \end{aligned}$$

For  $i \in I_0$ ,  $F_i(x_k) = G_i(x_k)$  by (4.13a), so that, using (4.13d):

$$G_i(x'_k) - F_i(x'_k) = t_k v_i^\top p + o(t_k) = \operatorname{sgn}(\alpha_i) t_k + o(t_k).$$

Therefore, for  $i \in I_0$  and  $k$  large enough:

$$\begin{cases} F_i(x'_k) < G_i(x'_k) & \text{if } i \in I_0 \text{ and } \alpha_i > 0, \\ F_i(x'_k) > G_i(x'_k) & \text{if } i \in I_0 \text{ and } \alpha_i < 0. \end{cases} \quad (4.13f)$$

We deduce from (4.13e), (4.13f) and the differentiability of  $H$  at  $x'_k$  that

$$H'_i(x'_k) = \begin{cases} G'_i(x'_k) & \text{if } i = i_0, \\ F'_i(x'_k) & \text{if } i \in I_0 \text{ and } \alpha_i > 0, \\ G'_i(x'_k) & \text{if } i \in I_0 \text{ and } \alpha_i < 0. \end{cases}$$

At the limit when  $k \rightarrow \infty$ , we get a Jacobian  $J$  in  $\partial_B H(x)$  satisfying (4.13c), as announced.

$[(v) \Rightarrow (iv)]$  If  $v_{i_0}$  satisfies (4.8) for some vectors  $\{v_i : i \in I_0\}$  that are not linearly independent, one can also write  $v_{i_0} = \sum_{i \in I'_0} \alpha'_i v_i$  with  $\alpha'_i \neq 0$ ,  $I'_0 \subseteq I_0$  and linearly independent vectors  $\{v_i : i \in I'_0\}$ . By (v),

$$\bigcap_{i \in I'_0} (\mathcal{V}_i \cap \mathcal{U}) \subseteq \mathcal{V}_{i_0}.$$

Since  $I'_0 \subseteq I_0$ , the intersection has more terms and is smaller so (4.9) also holds.  $\square$

## 4.4 Differential of the merit function

Before discussing properties of the merit function  $\theta$ , let us show an adaptation of lemma 2.3.23, taken from [206, lemma 2.2, p. 356]. This lemma states that, for some Lipschitz function  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ ,

$$F'(x; d) = Vd, \quad V \in \partial F(x)$$

This property, when applied to the componentwise minimum function  $H$ , can be restricted to  $\partial_B H(x)$ : the C-differential is weakened into the B-differential.

**Proposition 4.4.1** (directional derivative and B-differential). *Let  $F, G$  be two  $C^1$  Lipschitz functions from  $\mathbb{R}^n$  to  $\mathbb{R}^n$ . For any  $x, d \in \mathbb{R}^n$ , one has*

$$H'(x; d) = J^\top d \quad \text{for some } J \in \partial_B H(x).$$

*Proof.* Recall the formula of  $H'(x; d)$

$$H'(x; d) = \begin{pmatrix} F'_i(x)d & \text{if } i \in \mathcal{F}(x) \\ G'_i(x)d & \text{if } i \in \mathcal{G}(x) \\ \min(F'_i(x)d, G'_i(x)d) & \text{if } i \in \mathcal{E}(x) \end{pmatrix}_{i \in [1:n]}.$$

Now, define the following index sets

$$\begin{aligned} \mathcal{E}_\mathcal{F} &:= \{i \in \mathcal{E}(x) : F'_i(x)d < G'_i(x)d\}, \\ \mathcal{E}_\mathcal{G} &:= \{i \in \mathcal{E}(x) : F'_i(x)d > G'_i(x)d\}, \\ \mathcal{E}_\mathcal{E} &:= \{i \in \mathcal{E}(x) : F'_i(x)d = G'_i(x)d\}. \end{aligned}$$

Clearly, one has  $H'(x; d) = Jd$ , where  $J_{\mathcal{E}_\mathcal{F},:} = F'(x)_{\mathcal{E}_\mathcal{F}}, J_{\mathcal{E}_\mathcal{G},:} = G'(x)_{\mathcal{E}_\mathcal{G}}$  and  $J_{i,:} \in \{F'_i(x), G'_i(x)\}$  for  $i \in \mathcal{E}_\mathcal{E}$ .

Now, consider the (sub)arrangement defined by the hyperplanes  $H_i$  with indices  $i \in \mathcal{E}_\mathcal{E}$ . Let  $s' \in \mathcal{S}((G'(x) - F'(x))_{\mathcal{E}_\mathcal{E},:}^\top, 0)$  and  $d'$  be an associated direction (see chapter 3), i.e.,  $s' = \text{sgn}((G'(x) - F'(x))d')_{\mathcal{E}_\mathcal{E}}$ . For a sequence  $\{t_k\}_k \downarrow 0$  and  $\varepsilon > 0$  small enough, consider the sequence  $x_k := x + t_k(d + \varepsilon d')$ . Let us show that: this sequence does not belong to any of the hyperplanes  $H_i$  for  $i \in \mathcal{E}(x)$  for  $k$  large enough ( $t_k$  small enough), and that the corresponding Jacobian belongs to  $\partial_B H(x)$  and is equal to one of the possible  $J$  defined above.

By continuity, for  $k$  large enough,  $F_{\mathcal{F}(x)}(x_k) < G_{\mathcal{F}(x)}(x_k), F_{\mathcal{G}(x)}(x_k) > G_{\mathcal{G}(x)}(x_k)$ . Then, for the indices in  $\mathcal{E}(x)$ , use the following expansions

$$\begin{aligned} F_i(x_k) &= F_i(x) + t_k F'_i(x)d + \varepsilon t_k F'_i(x)d' + o(t_k), \\ G_i(x_k) &= G_i(x) + t_k G'_i(x)d + \varepsilon t_k G'_i(x)d' + o(t_k). \end{aligned}$$

For the indices in  $\mathcal{E}_\mathcal{F}$ , since  $\varepsilon$  is small enough, one has  $F_i(x_k) < G_i(x_k)$ , and similarly  $F_i(x_k) > G_i(x_k)$  for the indices in  $\mathcal{E}_\mathcal{G}$ . For those in  $\mathcal{E}_\mathcal{E}$ ,

$$H_i(x_k) = \min(F_i(x_k), G_i(x_k)) = \begin{cases} F_i(x_k) < G_i(x_k) & \text{if } s'_i = +1, \\ G_i(x_k) < F_i(x_k) & \text{if } s'_i = -1. \end{cases}$$

Thus,  $x_k$  belongs to none of the hyperplanes  $H_i$  and corresponds to the Jacobian  $J'$  with rows

$$J'_{i,:} = \begin{cases} F'_i(x) & i \in \mathcal{F}(x) \cup \mathcal{E}_{\mathcal{F}}, \\ G'_i(x) & i \in \mathcal{G}(x) \cup \mathcal{E}_{\mathcal{G}}, \\ F'_i(x) & i \in \mathcal{E}_{\mathcal{E}}, s'_i = +1, \\ G'_i(x) & i \in \mathcal{E}_{\mathcal{E}}, s'_i = -1. \end{cases}$$

By construction, this Jacobian belongs to  $\partial_B H(x)$  and  $J'd = H'(x; d)$ .  $\square$

In what follows we will also consider a more general merit function  $\theta_\psi = \psi \circ H(x)$ , with  $\psi$  a  $C^1$  scalar function generalizing the squared 2-norm. When  $H$  is smooth at  $x$ , one clearly has

$$\begin{aligned} \nabla \theta(x) &= \nabla H(x) \times H(x) = \sum_{i=1}^n H_i(x) \nabla H_i(x) \\ \nabla \theta_\psi(x) &= \nabla H(x) \times \nabla \psi(H(x)) = \sum_{i=1}^n \nabla \psi(H(x))_i \nabla H_i(x) \end{aligned}$$

For the C-differential, Clarke [51, prop. 2.6.6, pp. 72-73] showed that for a smooth function  $\psi$  (so in particular  $\|\cdot\|^2/2$ ), one has

$$\partial \theta_\psi(x) = \partial H(x)^\top \nabla \psi(H(x))$$

where the differential is seen as a column vector (and even stronger results). While one could justify the proof for the B-differential then take the convex hull, Clarke's proof develops a more advanced reasoning. Moreover, one cannot simply use Clarke's relation and take the extremal points (see the comments around proposition 3.4.14); we recall these properties.

**Proposition 4.4.2** (extremality for  $H$ ). *One has*

$$\partial_B H(x) = \text{ext}(\partial_C H(x)) = \text{ext}(\text{conv}(\partial_B H(x))). \quad \square$$

Observe that it holds for nonlinear functions  $F$  and  $G$  as well. Note that the relation  $C \supseteq \text{ext}(\text{conv}(C))$  always holds, but the converse is not true in general (for instance a ball), since it means every point of  $C$  is extremal.

**Remark 4.4.3.** The previous observation cannot be used, for the computation of  $\partial_B \theta(x)$ , to write (dropping the  $(x)$  dependency)

$$\partial_B \theta = \text{ext}(\partial_C \theta) = \text{ext}(\text{conv}(\partial_B \theta))$$

What unables to simply use this result is that, in general,

$$\text{ext}(S \times v) \neq \text{ext}(S) \times v$$

However, the concerned proposition shows the equality holds because  $S = \partial_B H^\top$  is specific enough.  $\square$

**Counter-example 4.4.4** ( $\text{ext}(S \times v) \neq \text{ext}(S) \times v$ ). For instance, consider

$$S = \text{conv} \left\{ \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} -1 & -1 \\ -1 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} \right\}, \quad v = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Since  $\text{ext}(S)$  is composed of the five matrices given, one has

$$\begin{aligned} \text{ext}(S) \times v &= \left\{ \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} -2 \\ -2 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \end{bmatrix}, \begin{bmatrix} -2 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}, \\ \text{ext}(S \times v) &= \left\{ \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} -2 \\ -2 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \end{bmatrix}, \begin{bmatrix} -2 \\ 2 \end{bmatrix} \right\}. \end{aligned}$$

□

Coming back to the B-differential of  $\theta$ , it seems natural to wonder whether a similar chain rule holds or not, since Clarke's differential was introduced in part to benefit such property. The proof we propose is rather different than Clarke's, relying on specific properties of the minimum function  $H$  and its Bouligand differential. Contrary to the Fischer C-function, which leads to a more differentiable merit function, the result is not as simple [87], but one can recover the expected form  $(\partial_B H)^\top \times H$ . First, we recall a simple property.

**Proposition 4.4.5** (one minimum). *Consider  $\theta : y \mapsto (\min(f(y), g(y)))^2/2$ , and  $x$  such that both terms of the min are equal:  $f(x) = g(x)$ . Then  $\theta$  is differentiable in  $x$  if and only if  $\theta(x) = 0$  or  $f'(x) = g'(x)$ .* □

This cannot be generalized to multiple minima functions simultaneously.

**Counter-example 4.4.6** (multiple minima). Let  $H_1(x, y) = 1 - |x + y|$  and  $H_2(x, y) = -1 - |x + y|$ . The functions  $H_1$  and  $H_2$  are not differentiable when  $x + y = 0$  but  $\theta(x, y) = (H_1^2 + H_2^2)/2 = 1 + (x + y)^2$ . □

This curious phenomenon is explained by the following proposition.

**Proposition 4.4.7** (multiple minima). *Consider the following function*

$$d \in \mathbb{R}^n \mapsto \sum_{i=1}^q c_i \min(u_i^\top d, w_i^\top d),$$

where the none of the vectors  $v_i = (w_i - u_i) \in \mathbb{R}^n$  for  $i \in [1 : q]$  is colinear to another. If the function is linear, then all the  $c_i$  are zero.

*Proof.* First, as in for instance [255] we rewrite the function under the following form:

$$\sum_{i=1}^q c_i u_i^\top d + \sum_{i=1}^q c_i \min(0, v_i^\top d) = \sum_{i=1}^q c_i u_i^\top d + \mathcal{L}(d) = \sum_{i=1}^q c_i u_i^\top d + l^\top d, \quad l \in \mathbb{R}^n$$

which is linear if and only if the term  $\mathcal{L}$  is, with  $\mathcal{L}(d) = l^\top d$ . Thus we focus on the sum of minima  $\mathcal{L}$ ,  $\mathcal{L}(d) := \sum_{i=1}^q c_i \min(0, v_i^\top d)$ . Now, consider the hyperplane arrangement defined by the  $v_i^\perp$  for  $i \in [1 : q]$ . Let  $\mathcal{S}_q \subseteq \{\pm 1\}^q$  be the set of sign vectors of this arrangement, for every sign  $s \in \mathcal{S}_q$ , there exists a direction  $d_s \in \mathbb{R}^n$  such that  $\text{Diag}(s)V^\top d_s > 0$ , with  $V$  being the matrix  $[v_1 \dots v_q]$ .

Consider then a given sign vector  $s^0 \in \mathcal{S}_q$ , and the associated  $d^0 \in \mathbb{R}^n$ . Let us split the indices:  $I^+ = \{i \in \mathcal{E}^\neq(x) : v_i^\top d^0 > 0\}$  and  $I^- = \{i \in \mathcal{E}^\neq(x) : v_i^\top d^0 < 0\}$  so that in  $d^0$ ,  $\mathcal{L}$  has the following expression

$$\mathcal{L}(d^0) = \sum_{i \in I^+} c_i \times 0 + \sum_{i \in I^-} c_i v_i^\top d^0 = l^\top d^0,$$

thus the function takes the form  $d \mapsto \sum_{i \in I^-} c_i v_i^\top d$ . Indeed, as this is true for every  $d$  in the region of  $d^0$ , this is true for a small ball around  $d^0$ . Identifying the expression of the linear function, we get that  $l = \sum_{i \in I^-} c_i v_i$ . Then we look what happens in another (neighboring) region.

As we supposed no vector was colinear to another, we know that the set of sign vectors is connected (proposition 3.4.5). Then, by symmetry (proposition 3.4.1), there exists a path  $s^0, \dots, s^q$  with  $s^q = -s^0$  such that two consecutive sign vectors have exactly one difference: along the path the sign of each index is changed exactly once. We look more closely at  $s^0$  and  $s^1$ .

Without loss of generality, we suppose the index  $j$  changed belongs to  $I^+$  (the case  $j \in I^-$  is very similar). For a  $d^1$  associated to  $s^1$ , we have

$$\begin{aligned} d^0 &\rightarrow \sum_{i \in I^+ \setminus \{j\}} c_i \times 0 + c_j \times 0 + \sum_{i \in I^-} c_i v_i^\top d^0 \\ d^1 &\rightarrow \sum_{i \in I^+ \setminus \{j\}} c_i \times 0 + c_j w_j^\top d^1 + \sum_{i \in I^-} c_i v_i^\top d^1 \end{aligned}$$

Using the identification argument on both regions, we have the following equality

$$\sum_{i \in I^-} c_i v_i = l = \sum_{i \in I^-} c_i v_i + c_j w_j$$

As the  $v_i$ 's are not colinear, they are nonzero, so  $c_j = 0$ . Now, doing exactly the same thing for every index along the path between  $s^0$  and  $s^q = -s^0$ , we have that every coefficient  $c_i$  is zero.  $\square$

This lemma illustrates that a sum of minima of noncolinear linear functions cannot be linear unless it is zero, by identifying a different expression for the (assumed) linear function  $\mathcal{L}$ . However, it cannot be used directly in the main reasoning that follows, since it can be as in counter-example 4.4.6 where the function is nonzero but the involved vectors are colinear. Before moving to the main proposition, we detail one last point.

**Lemma 4.4.8** (aggregation of similar minima). Let  $v \in \mathbb{R}^n$ ,  $I$  be some index set,  $\alpha \in \mathbb{R}_*^I$  and  $c \in \mathbb{R}^I$ . The following relation holds:

$$\sum_{i \in I, \alpha_i > 0} c_i \min(0, \alpha_i v^\top x) + \sum_{i \in I, \alpha_i < 0} c_i \min(0, \alpha_i v^\top x) = -\mathcal{C}^- v^\top x + (\mathcal{C}^+ + \mathcal{C}^-) \min(0, v^\top x)$$

Where  $\mathcal{C}^+ = \sum_{i \in I, \alpha_i > 0} c_i \alpha_i$  and  $\mathcal{C}^- = \sum_{i \in I, \alpha_i < 0} c_i |\alpha_i|$ .

*Proof.*

$$\begin{aligned}
 & \sum_{i \in I, \alpha_i > 0} c_i \min(0, \alpha_i v^\top x) + \sum_{i \in I, \alpha_i < 0} c_i \min(0, \alpha_i v^\top x) \\
 &= \sum_{i \in I, \alpha_i > 0} c_i \alpha_i \min(0, v^\top x) - \sum_{i \in I, \alpha_i < 0} c_i \max(0, |\alpha_i| v^\top x) \\
 &= \mathcal{C}^+ \min(0, v^\top x) - \sum_{i \in I, \alpha_i < 0} c_i |\alpha_i| \max(0, v^\top x) \\
 &= \mathcal{C}^+ \min(0, v^\top x) - \mathcal{C}^- \max(0, v^\top x) = \mathcal{C}^+ \min(0, v^\top x) - \mathcal{C}^- (v^\top x + \max(-v^\top x, 0)) \\
 &= \mathcal{C}^+ \min(0, v^\top x) - \mathcal{C}^- v^\top x + \mathcal{C}^- \min(0, v^\top x) = -\mathcal{C}^- v^\top x + (\mathcal{C}^+ + \mathcal{C}^-) \min(0, v^\top x). \quad \square
 \end{aligned}$$

The use of this computation is the following: we have a property for sums of minima with vectors that are not two by two colinear. In the general case (as seen in counterexample 4.4.6), if some vectors are colinear then the corresponding indices can be grouped together with lemma 4.4.8, then we use the proposition 4.4.7 for these grouped indices or for vectors that are not colinear to any other.

**Proposition 4.4.9** (B-differential of  $\theta$  in the linear case). *One has the following chain rule:*

$$\tilde{\partial}_B(\psi \circ H)(x) := \{J^\top \nabla \psi(H(x)); J \in \partial_B H(x)\} = \partial_B(\psi \circ H)(x).$$

In particular for  $\psi = \|\cdot\|^2/2$ , we recover the usual merit function.

*Proof.*  $[\subseteq]$  Consider a  $J \in \partial_B H(x)$ , associated with a sequence  $\{x_k\}_k \rightarrow x$  such that  $\nabla H(x_k) \rightarrow J^\top$ . Because  $H$  is differentiable in the points of the sequence, then using  $\nabla \theta = \nabla H \times \nabla \psi(H)$ , we get

$$\nabla \theta(x_k) = \nabla H(x_k) \times \nabla \psi(H(x_k)) \rightarrow J^\top \nabla \psi(H(x))$$

Which indicates this limit is in  $\partial_B \theta_\psi(x) = \partial_B(\psi \circ H)(x)$ .

$[\supseteq]$  We consider an element  $v \in \partial_B \theta_\psi(x)$ , and the associated sequence  $\{x_k\}_k$  such that  $\theta$  is F-differentiable in  $x_k$  for every  $k$ ,  $x_k \rightarrow x$  and  $\nabla \theta(x_k) \rightarrow v$ . We want to show that  $v$  can be written as  $J^\top \nabla \psi(H(x))$  for a  $J \in \partial_B H(x)$ , i.e. can be expressed as a limit of  $\nabla H(x_k) \times \nabla \psi(H(x_k))$ .

First, suppose that  $H$  is differentiable along  $\{x_k\}_k$ :  $\nabla \theta(x_k) = \nabla H(x_k) \times \nabla \psi(H(x_k)) \rightarrow v$  is well-defined. As  $\nabla H(x_k)$  is piecewise constant and has a finite number of possible values (proposition 3.2.2), up to extracting a subsequence we can assume it has a fixed constant value. This says  $v$  is a limit of the desired form.

Now, we consider that along the sequence  $\{x_k\}_k$ ,  $H$  has some nondifferentiable components:

- if  $i \in \mathcal{F}(x)$  or  $i \in \mathcal{G}(x)$ , for  $k$  large enough (so  $x_k$  close enough to  $x$ ), we have  $F_i(x_k) \neq G_i(x_k)$  so the component is differentiable,
- if  $i \in \mathcal{E}^=(x)$ , the affine functions are identical, so the minimum disappears and the component is always differentiable,
- if  $i \in \mathcal{E}^\neq(x)$ ; we note  $I^k \subseteq \mathcal{E}^\neq$  the components not differentiable in  $x_k$ .

We start with a remark about  $I^k$ . As  $\mathcal{E}^\neq(x)$  is fixed (depends only on  $x$  and the data), and  $I^k \subseteq \mathcal{E}^\neq(x)$  by continuity, there is only a finite number of different possible  $I^k$  (smaller than  $2^{|\mathcal{E}^\neq(x)|}$ ). Thus, by extracting a subsequence, still noted  $\{x_k\}_k$ , one can assume that the indices  $I^k$  do not change and are noted  $I$ .

For simplicity we first show the result for  $\psi = \|\cdot\|^2/2$  before showing the general result. Now, as  $\theta$  is F-differentiable in  $x_k$  by hypothesis (still true alongside any subsequence), we decompose:

$$\theta = \frac{1}{2} \sum_{i \in I^c} H_i^2 + \frac{1}{2} \sum_{i \in I} H_i^2$$

where  $\theta$  and the first sum are F-differentiable by hypothesis and definition of  $I$ , thus the second sum is also F-differentiable, the differentiability being considered in  $x_k$ . For every  $d$  of unit norm and  $\varepsilon > 0$  small, we can write the expansion  $\cdot(x_k + \varepsilon d) = \cdot(x_k) + \varepsilon \cdot'(x_k)d + o(\varepsilon)$ . We recall that  $H_i$  for  $i \in I$  is not differentiable in  $x_k$  so  $F_i(x_k) = G_i(x_k)$ , i.e.  $a_i + A_{i,:}x_k = H_i^k = b_i + B_{i,:}x_k$ , compactly noted  $H_I^k = (H_i^k)_{i \in I}$ . Thus, the expansion becomes

$$\begin{aligned} & \frac{1}{2\varepsilon} \left[ \sum_{i \in I} [H_i^k + \varepsilon \min(A_{i,:}d, B_{i,:}d)]^2 - \sum_{i \in I} (H_i^k)^2 \right] \\ &= \frac{1}{2\varepsilon} \left[ \sum_{i \in I} (H_i^k)^2 + 2H_i^k \varepsilon \min(A_{i,:}d, B_{i,:}d) + \varepsilon^2 [\min(A_{i,:}d, B_{i,:}d)]^2 - (H_i^k)^2 \right] \\ &= \sum_{i \in I} H_i^k \min(A_{i,:}d, B_{i,:}d) + \frac{\varepsilon}{2} [\min(A_{i,:}d, B_{i,:}d)]^2 \end{aligned}$$

Clearly the second term vanishes for  $\varepsilon \rightarrow 0$ . Besides, F-differentiability ensures that

$$d \mapsto \sum_{i \in I} H_i^k \min(A_{i,:}d, B_{i,:}d) \quad \text{is linear in } d. \quad (4.14)$$

Before applying proposition 4.4.7, note that if for some index  $i$ ,  $H_i^k$  is 0 for an infinite number of  $k$ , then extracting a subsequence where  $H_i^k \equiv 0$ , continuity tells that  $H_i(x) = 0$ . Thus, in the formula we want to prove, index  $i$  is irrelevant. As in the proof for the Fisher C-function, we have an expression of the form " $0 \times [\text{term of index } i]$ ", so we can choose an arbitrary line for  $J_{i,:}$ . Then, taking an appropriate subsequence one can assume  $H_i^k \neq 0$  for  $i \in I$  (changing  $I$  if needed by removing the indices for which  $H_i(x) = 0$ ).<sup>3</sup>

---

<sup>3</sup>In [195, theorem 5.e, p. 323], Pang uses a somewhat similar approach to show differentiability properties of the minimum reformulation, but with stronger assumptions. The knowledge gained on the structure of the minimum exhibited in chapter 3 seems to allow us to obtain slightly more general results.

Using  $\min(A_{i,:}d, B_{i,:}d) = A_{i,:}d + \min(0, v_i^\top d)$ , if the vectors  $v_i$  are not two by two colinear, then  $H_i^k = 0$  for every  $i$ , which clearly isn't true from the previous assumption.

Moreover, suppose that there exists a  $v_i$  which is noncolinear to any other. Then using the same argument as in the proof of proposition 4.4.7, on both sides of the hyperplane  $v_i^\perp$ , we would get that  $H_i^k = 0$  which is a contradiction with the assumption.

To summarize, for every  $v_i$ , there exists (at least) a  $v_{i'}$  colinear to  $v_i$ : the indices can be grouped into subsets where all the  $v_i$  are colinear to a single vector  $w_j$ . Then, we apply proposition 4.4.7 to the noncolinear vectors, their coefficients grouped through lemma 4.4.8. These colinearity relations are expressed in the following way, coming back to (4.14): without loss of generality, we assume that vectors  $v_1, \dots, v_{p_1}$  are colinear to  $w_1$ , then  $v_{p_1+1}, \dots, v_{p_2}$  are colinear to  $w_2$  and so on. We also assume the index set  $I$  reads  $I = \cup_{j=1}^q I_j$  with  $I_j := [p_{j-1} + 1 : p_j]$  where  $p_0 = 0$  and the  $p_j$  are integers,

$$\begin{aligned} \sum_{i \in I} H_i^k \min(A_{i,:}d, B_{i,:}d) &= \sum_{i \in I} H_i^k A_{i,:}d + \sum_{i \in I} H_i^k \min(0, v_i^\top d) \\ &= \mathcal{L}d + \sum_{i=1}^{p_1} H_i^k \min(0, \alpha_i w_1^\top d) + \dots + \sum_{i=p_{q-1}+1}^{p_q} H_i^k \min(0, \alpha_i w_q^\top d) \\ &= \left. \begin{aligned} &= \mathcal{L}d + \sum_{j=1}^q -\mathcal{C}_{(j)}^- w_j^\top d \\ &+ (\mathcal{C}_{(1)}^+ + \mathcal{C}_{(1)}^-) \min(0, w_1^\top d) + \dots + (\mathcal{C}_{(q)}^+ + \mathcal{C}_{(q)}^-) \min(0, w_q^\top d) \end{aligned} \right\} \text{ linear} \end{aligned}$$

with  $\mathcal{C}_{(j)}^+ = \sum_{i=p_{j-1}+1}^{p_j} H_i^k \alpha_i$  for the positive  $\alpha_i$ , and  $\mathcal{C}_{(j)}^- = \sum_{i=p_{j-1}+1}^{p_j} H_i^k |\alpha_i|$  for the negative  $\alpha_i$ .

Under the above form, proposition 4.4.7 can be used: as the  $w_j$  are noncolinear, the coefficients in front of the minima are zero:  $\mathcal{C}_{(j)}^+ + \mathcal{C}_{(j)}^- = 0$  for all  $j \in [1 : q]$ . This tells that the last line vanishes. Then, coming back to the linear part, we recover a form related to a potential Jacobian from the B-differential:

$$\begin{aligned} \sum_{i \in I} H_i^k A_{i,:}d + \sum_{j=1}^q -\mathcal{C}_{(j)}^- w_j^\top d &= \sum_{i \in I} H_i^k A_{i,:}d + \sum_{j=1}^q \sum_{i=p_{j-1}+1, \alpha_i < 0}^{p_j} H_i^k \alpha_i w_j^\top d \\ &= \sum_{i \in I} H_i^k A_{i,:}d + \sum_{j=1}^q \sum_{i=p_{j-1}+1, \alpha_i < 0}^{p_j} H_i^k v_i^\top d \\ &= \sum_{i \in I} H_i^k A_{i,:}d + \sum_{i \in I, \alpha_i < 0} H_i^k (B_{i,:} - A_{i,:})d \end{aligned} \tag{4.15}$$

which is a combination of lines of  $A$  and  $B$  for indices of  $I$ . In particular, this expression is not dependent on  $d$ , since it only depends on the sign of the  $\alpha_i$ , which are related to the  $v_i$  and the  $w_j$ . For every index  $i$ , the corresponding term in the sum is of the desired form  $H_i^k A_{i,:}$  or  $H_i^k B_{i,:}$ .

We only need to show the resulting matrix, whose line  $i \in I$  equals  $A_{i,:}$  if  $\alpha_i > 0$  and  $B_{i,:}$  if  $\alpha_i < 0$  ( $\alpha_i \neq 0$  by the colinearity hypothesis), indeed in the B-differential. The lines corresponding to indices in  $\mathcal{F}(x)$ ,  $\mathcal{G}(x)$ ,  $\mathcal{E}^=(x)$  are necessarily similar. As we took an appropriate subsequence, the lines with indices in  $\mathcal{E}^\neq(x) \setminus I$  are already known.

Note that the trajectory is such that  $H_I$  is nondifferentiable, i.e.,  $x_k$  belongs to the corresponding hyperplanes ( $\{y, v_i^\top y = v_i^\top x\}$ ). Thus we need to justify that for the arrangement of hyperplanes defined by the  $v_i^\perp$  for  $i \in I$ , the submatrix obtained corresponds to a nonempty region.

For this last part, we use that the  $w_j$  can be taken to form a pointed cone. In fact, the orientation of the  $w_j$ 's is what determines the (sub-)matrix. Then a handle  $y$  of the cone formed by the  $w_j$ 's is such that  $x_k + \varepsilon y$  belongs to no hyperplane for a suitable  $\varepsilon > 0$ , and belongs to the region defined by the last line of (4.15) (the region with only +1's corresponding to the pointed cone of the  $w_j$ 's). We have thus shown that the derivative of  $\theta$  can be expressed suitably, which tells that the limit  $v$  is in  $\partial_B H(x)^\top H(x)$ .

Furthermore, for indices that were such that  $H_i^k \equiv 0$ , the previous direction  $y$  can be modified if necessary to ensure the products  $v_i^\top y$  are nonzero. The submatrix can then be completed for these indices, as these terms are zero in  $x_k$  and  $x$ .  $\square$

Now we resume the proof in the case where  $\psi$  is a smooth function (not necessarily  $\|\cdot\|^2/2$ ).

*Proof.* First, in the separable case,  $\psi \circ H = \sum \psi_i(H_i)$ , the same proof can be done after replacing the  $H_i^k$  by  $\psi'_i(H_i^k)$  everywhere.

In the general case, for indices  $i \notin I$  and  $\varepsilon$  small enough the minimum is already known and noted  $J_{i:d}$ . The expansion reads

$$\begin{aligned} & \frac{1}{\varepsilon} \left[ \psi \begin{pmatrix} (H_i^k + \varepsilon \min(A_{i:d}, B_{i:d}))_{i \in I} \\ (H_i^k + \varepsilon J_{i:d})_{i \in I^c} \end{pmatrix} - \psi \begin{pmatrix} (H_i^k)_{i \in I} \\ (H_i^k)_{i \in I^c} \end{pmatrix} \right] \\ &= \psi'((H_i^k)_i) \begin{bmatrix} (\min(A_{i:d}, B_{i:d}))_{i \in I} \\ (J_{i:d})_{i \in I^c} \end{bmatrix} + o(1) \\ &= \sum_{j \in I^c} \partial_j \psi((H_i^k)_i) J_{j:d} + \sum_{j \in I} \partial_j \psi((H_i^k)_i) \min(A_{j:d}, B_{j:d}) + o(1) \end{aligned}$$

Then, using once again the F-differentiability, this expression is linear in  $d$ . Removing the sum over  $I^c$ , the remaining sum is still linear. Then the same reasoning can be applied, with scalars  $\partial_j \psi((H_i^k)_{i \in I})$ .  $\square$

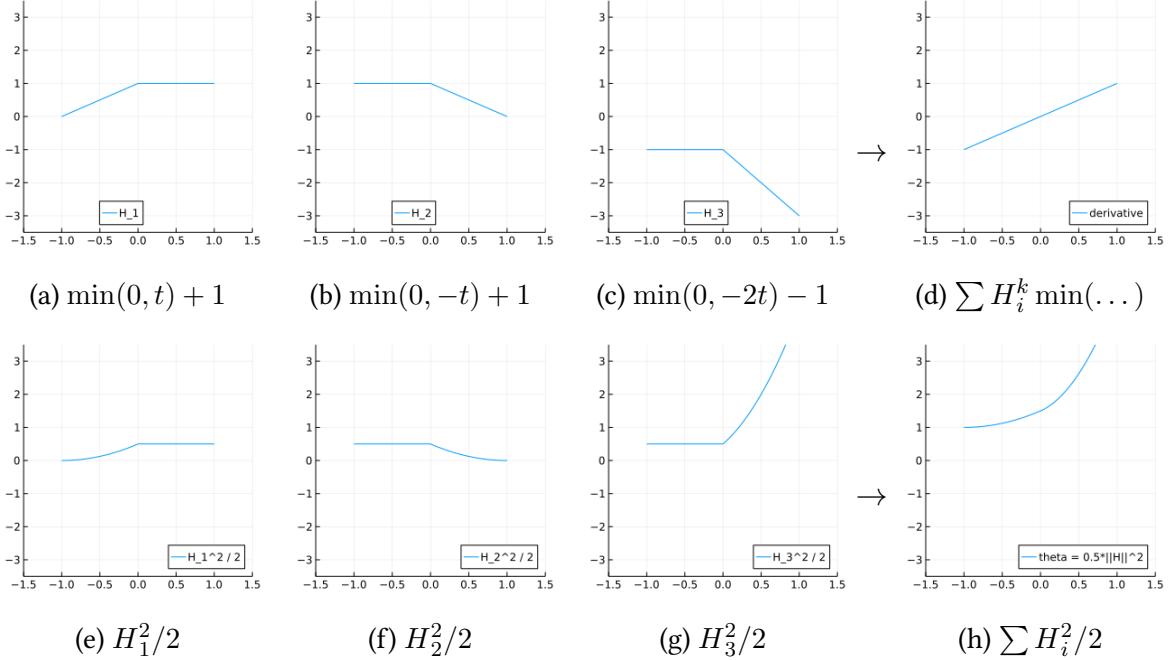
A relevant remark is that this process exhibits one particular submatrix for the indices in  $I$ . However, this comes from the choice of ‘direction’ of the  $w_j$ : choosing a  $-w_j$  instead of  $w_j$  will change the signs of some  $\alpha_i$ , which in turn changes some ligns of the submatrix.

We illustrate an example where vectors are two by two colinear, so where the values  $H_i^k$  need not be zero to have a linear operator in (4.14). Consider the following case (identifying row and column vectors):

$$H_1 = +1, H_2 = +1, H_3 = -1, \quad A_1 = 0, B_1 = e_1, A_2 = 0, B_2 = -e_1, A_3 = 0, B_3 = -2e_1$$

$$+1 \min(0, d_1) + \min(0, -d_1) - \min(0, -2d_1) = d_1.$$

We observe that  $\theta$  is differentiable at 0 but the  $H_i$  are not.



When looking at the Fischer function instead of the minimum, the formula is also true, and has the advantage that the multivalued columns are multiplied by zero, thus  $\partial_B \Psi$  is in fact one element [87]. As shown by the previous example, this is not what happens in the minimum setting. As shown below, this can result in the merit function to have multiple elements in its B-differential, when  $H_i(x) \neq 0$ , then the Jacobian matrices are multiplied with a nonzero vector, so the products are different.

**Proposition 4.4.10** (Nonlinear case). *The proposition 4.4.9 holds for nonlinear ( $\mathcal{C}^1$ ) functions  $F$  and  $G$ .*

*Proof.* The main question is to check what doesn't hold from the linear case and must be adapted. The  $[\subseteq]$  inclusion is exactly the same, as the  $\nabla \theta = \nabla H \times \nabla \psi(H)$  relation means that the given sequence  $\{x_k\}_k$  is enough.

For the reverse inclusion, if  $H$  is differentiable the same can be done: the element,  $v \in \partial_B \theta$ , is the limit of  $\nabla H(x_k) \times \nabla \psi(H(x_k))$ . Up to taking an appropriate subsequence, the first term can be chosen such that a fixed line  $i$  is  $F'_i(x_k)$  or  $G'_i(x_k)$  (depending on the sequence) for every  $k$ , which corresponds to the transpose of a matrix of the B-differential of  $H$ . Thus the product has the wanted form.

If  $H$  is not differentiable along the sequence, taking a subsequence as previously, the set of indices  $I$  such that  $H_I$  is not differentiable is constant. However, the indices can now belong to  $\mathcal{E}^\neq(x) \cup \mathcal{E}^=(x)$ , whereas  $\mathcal{E}^=(x)$  was previously excluded.

Developing in a similar way, the F-differentiability tells that for every  $k$  of the appropriate subsequence, the following function is linear:

$$d \mapsto \sum_{j \in I^c} \partial_j \psi((H_i^k)_i) J_i^k d + \sum_{j \in I} \partial_j \psi((H_i^k)_i) \min(F'_i(x_k) d, G'_i(x_k) d).$$

The main difference compared to the linear case is that the  $F'_i(x_k), G'_i(x_k)$  are not corresponding to the  $v_i := G'_i(x) - F'_i(x)$  (since  $x_k \neq x$ ). First, if for some  $i$  and an associated subsequence the values  $\partial_j \psi((H_i^k)_i)$  are all zero then at the limit, by continuity one has  $\partial_j \psi((H_i)_i) = 0$ , so this term is not relevant in the final expression. Thus one can complete the Jacobian matrix at the end of the proof.

Nonetheless, we can still apply the lemma 4.4.8, as the term of the derivative is a sum of "weighted minima" as in the linear case. Indeed, up to taking a subsequence, one can assume the set of indices  $I$  such that  $H_i$  is not differentiable in  $x_k$  for  $i \in I$  is constant and independent of  $k$ .

This means that for every  $k$ , the  $v_i^k := G'_i(x_k) - F'_i(x_k)$  can be grouped by colinearity, and the associated coefficients including the  $\|v_i^k\|$  (i.e., the  $\alpha_i$ ) and the  $H_i^k$  cancel out. We have, at index  $k$ ,

$$\left\{ \begin{array}{l} \forall i \in I_{j_1^k}, v_i^k = \alpha_i^k w_{j_1}^k, \dots, \forall i \in I_{j_{q^k}^k}, v_i^k = \alpha_i^k w_{j_{q^k}}^k \\ \forall m \in [1 : j_1^k : j_{q^k}^k], \|w_m^k\| = 1 \\ \forall m, \sum_{i \in I_m, \alpha_i^k > 0} \partial_i \psi((H_l^k)_l) \alpha_i^k - \sum_{i \in I_m, \alpha_i^k < 0} \partial_i \psi((H_l^k)_l) \alpha_i^k = 0 \end{array} \right. \quad (4.16)$$

where the last line applies for every group of colinearity  $I_m$  (possibly depending on  $k$ ). For  $i \in \mathcal{E}^=(x)$ , as  $v_i = 0$  we have  $\alpha_i^k \rightarrow 0$ .

The set of indices  $I$  being fixed (by previous subsequences), in particular it has a finite constant size, so the number of partitions of  $I$  grouping the indices by colinearity of the associated vectors is finite as well. Extracting more subsequences we can assume that the colinearity groups are the same along a correct subsequence, i.e., the relations  $v_i^k \propto w_j^k$  are true for the same indices groups  $[1 : p_1], \dots, [p_{q-1} + 1 : p_q]$ .

In a similar fashion, this can be done for the signs of the  $\alpha_i^k$  (finite number of possibilities). Fixing the signs of the  $\alpha_i^k$  results in a subsequence for which the analogue of (4.15) has, for every index of the subsequence, the form:

$$\sum_{i \in I} \partial_i \psi((H_l^k)_l) F'_i(x_k) d + \sum_{i \in I, \alpha_i^k < 0} \partial_i \psi((H_l^k)_l) (G'_i(x_k) - F'_i(x_k)) d \quad (4.17)$$

The above extractions result in every point of the subsequence giving a Jacobian (sub-)matrix that has the same component ( $F'_i(x_k)$  or  $G'_i(x_k)$ ) for a given index  $i$  for every  $k$ .

For every  $k$  one can find a direction  $d_k$  such that  $d_k^T w_j^k > 0$ , leading to the Jacobian matrix corresponding to (4.17). Thus the sequence  $x_k + t_k d_k$  for a suitable  $t_k > 0$  has the desired property.

To conclude, regularity of the functions makes so that  $H_i^k \rightarrow H_i(x)$  and the lines of the Jacobian matrix also converge to their limits  $F'_i(x)$  or  $G'_i(x)$ . We recover the desired form, so the equality  $\partial_B \theta = (\partial_B H)^\top \times \nabla \psi(H)$  holds.  $\square$

## 4.5 Details on instances and algorithms

This section aims at giving proofs on some values claimed in chapter 3 about certain types of instances. First, section 4.5.1 considers the PERM instances, whereas section 4.5.2 discusses the CROSSPOLYTOPE ones.

### 4.5.1 About the permutohedron instances

Now, let us focus on the PERM instances. The definition we use for these arrangements is the following: for some positive integer  $n$ , there are  $n(n+1)$  hyperplanes, given by:

$$H_i := \{x_i = 0\} \text{ for } 1 \leq i \leq n, \quad H_{ij} = \{x_i - x_j = 0\} \text{ for } 1 \leq i < j \leq n. \quad (4.18)$$

Such arrangements are well-known, and can be solved by combinatorics. First, we detail the chambers then the circuits.

#### Chambers

**Analytical approach** We show there are  $(n+1)!$  sign vectors. First, consider the  $n(n-1)/2$  hyperplanes  $H_{ij} = (e_i - e_j)^\perp$ . Let  $x$  be a given point,

$$x \in \mathbb{R}^n \setminus (\cup_{i,j} H_{ij}) \iff (x_1, \dots, x_n) \text{ are all different},$$

since if there exists a pair  $(i, j)$  such that  $x_i = x_j$ , then  $x \in H_{ij}$ . Now, there are  $n!$  ways to order the coordinates since they are all different, the  $n!$  permutations. Indeed, let  $\pi$  be a permutation of  $[1 : n]$ , such that  $x_{\pi(1)} > \dots > x_{\pi(n)}$ . If  $\pi(i) < \pi(j)$ ,  $x \in H_{\pi(i)\pi(j)}^+$ , whereas  $\pi(i) > \pi(j)$  implies  $x \in H_{\pi(j)\pi(i)}^-$  by definition of the hyperplanes considered. Therefore, for an arbitrary permutation  $\pi$  and the  $x$ 's having decreasing coordinates under  $\pi$ , the signs corresponding to the  $H_{ij}$  are determined. Therefore, there are  $n!$  chambers.

Then we show that each of these  $n!$  regions is exactly split in  $n+1$  subregions by the remaining  $n$  hyperplanes  $H_k = \{x_k = 0\}$ . This will lead to  $(n+1) \times n! = (n+1)!$  regions.

Let  $\pi$  be a permutation of size  $n$ , such that  $x_{\pi(1)} > \dots > x_{\pi(n)}$ . Then one can have the following configurations:

$$\begin{aligned} x_{\pi(i)} &> 0 \quad \forall i \in [1 : n] & \text{and} & \quad x_{\pi(1)} < 0, x_{\pi(i)} > 0 \quad \forall i > 1, \\ &\dots & &\dots, \\ x_{\pi(i)} &< 0, x_{\pi(n)} > 0 \quad \forall i < n & \text{and} & \quad x_{\pi(i)} < 0 \quad \forall i \in [1 : n]. \end{aligned}$$

Any other combination is of the form

$$\{x_{\pi(1)} < 0, \dots, x_{\pi(i^*)} > 0, \dots, x_{\pi(j^*)} < 0, \dots, x_{\pi(n)} > 0\}$$

which does not respect the definition of  $\pi$ .

**With Athanasiadis' approach** Following the suggestion of [35], one can use the method described in theorem 2.2 and example 2.3 of [12]. It reads as follows.

**Theorem 4.5.1** (2.2 in [12]). *Let  $q$  be a large enough prime number. An arrangement with hyperplanes defined by integer (or rational) coordinates verifies*

$$\chi(q) = \text{card}(\mathbb{F}_q^n \setminus \cup_k H_k)$$

where  $\chi$  is the characteristic polynomial of the arrangement,  $\cup_k H_k$  is the union of the hyperplanes,  $\mathbb{F}_q := [0 : q - 1] \bmod q$  thus  $\mathbb{F}_q^n$  is identified with  $[0 : q - 1]^n$  (all coordinates mod  $q$ ). Therefore,  $\chi(q)$  counts the number of points with integer coordinates in  $[0 : q - 1]^n$  who do not satisfy ( $\bmod q$ ) any of equations defining the hyperplanes.  $\square$

Where this theorem is useful is that, once the right-hand side is derived, this expression (depending on  $q$ ,  $n$  and the arrangement), can be evaluated considering  $q$  as the variable of the polynom. In particular, the value in  $-1$  is related to the number of chambers, by the identity  $|\mathcal{S}| = (-1)^n \chi(-1)$  [257]. Let us derive these expressions for the considered arrangements.

For a fixed large  $q$  and some  $x \in [0 : q - 1]^n$ , it is clear that one must have  $x_i \neq 0$  to avoid verifying the equations of  $H_i$ . Now, for the  $H_{ij}$ ,  $x$  must have different coordinates. Therefore, there is  $q - 1$  possibilities for the first (since it cannot be 0),  $q - 2$  for the second, and in the end,  $q - n$  for the  $n$ th coordinate. Therefore, the characteristic polynomial is  $\prod_{i=1}^n (q - i)$ . Then, the number of chambers is

$$(-1)^n \prod_{i=1}^n (-1 - i) = (-1)^n \prod_{i=1}^n (i + 1) = (n + 1)!.$$

Note that this very efficient formula does not indicate *what* are the chambers, whereas the previous analysis is able to.

## Stem vectors / circuits

It is also possible to explicitly express the set of circuits and therefore the stem vectors. First, recall that the numbers of stem vectors of the instances PERM-N for  $n = 5, 6, 7, 8$  are respectively 197, 1172, 8018, 62814. These numbers correspond to sequence A002807 in the OEIS, up to some index shift (number of circuits =  $a(n+1)$ ). The given formula corresponds precisely to the circuits, as detailed in the upcoming proposition. Recall that one has

$$V = [I_n \ M], \ M = [e_i - e_j]_{1 \leq i < j \leq n}.$$

**Proposition 4.5.2** (circuits of the PERM instances). *The number of circuits of PERM-N is given by*

$$|\mathcal{C}(\text{PERM-}n)| = \sum_{k=3}^{n+1} \frac{(k-1)!}{2} \binom{n+1}{k}$$

*In particular, the number of circuits of size  $k \in [3 : n+1]$  is precisely the  $k$ th term of the sum.*

The proof relies on some artificial but useful notions defined next.

**Definition 4.5.3** (coordinates covered by  $J$ ). Let  $J \subseteq [1 : p]$ . We denote by  $c_J := |\{i \in [1 : p] : \exists j \in J, (v_j)_i \neq 0\}|$  the number of nonzero lines of  $V_{:,J}$ .  $\square$

**Definition 4.5.4** (nonzero components in  $J$ ). Let  $J \subseteq [1 : p]$ . We denote by  $K_J := \sum_j \|v_j\|_1$  the total number of nonzero components of vectors in  $J$  ( $\|v_j\|_1 = \sum_1^n |(v_j)_i|$  and  $(v_j)_i \in \{-1, 0, +1\}$  for every  $i \in [1 : n]$  and  $j \in J$ ).  $\square$

For the  $V$  defining PERM-N, one clearly has  $c_J \leq n$  and  $K_J \in [k, 2k]$  for any  $J$  since the columns of  $V$  belong to  $\mathbb{R}^n$  and each of them has one or two nonzero components. In what follows, we call "coordinates" a subset of  $[1 : n]$ , and "components" the value(s) of one (or multiple) vector(s) eventually at some specific coordinates. Let us justify the above proposition.

*Proof.* First, the rank of the matrix  $V$  is clearly  $n$ . Therefore, the circuits are of size at most  $n+1$ , and at least 3, because there are no colinear vectors.

Let us show that for any circuit  $J$  of size  $k$ ,  $c_J \in \{k-1, k\}$ . Let  $c_J = l$  and denote by  $i_1, \dots, i_l$  the associated coordinates.

- suppose that  $l \leq k-2$ . By definition, the vectors  $v_j$  for  $j \in J$  have nonzero components only at coordinates  $i_1, \dots, i_l$ . Therefore,  $V_J$  is a submatrix of  $[e_{i_1}, \dots, e_{i_l}, e_k - e_{k'}]$  (assuming  $i_1 \leq k < k' \leq i_l$  without loss of generality). However, this matrix is clearly of rank  $l \leq k-2$  so its circuits are of size  $\leq k-2+1 = k-1 < k$ .
- suppose that  $l \geq k+1$ . By definition of  $V$ ,  $k \leq K_J \leq 2k$  since each  $v$  has one or two nonzero components. Since  $J$  is a circuit,  $\text{null}(V_J) = 1$ , meaning  $V_J \eta = 0 \in \mathbb{R}^n$  for

some  $\eta \in \mathbb{R}_*^J$ . The indices  $k \notin \{i_1, \dots, i_l\}$  clearly verify  $(V_J\eta)_k = \sum_{j \in J} (v_j)_k \eta_j = 0$ , since  $(V_{:,j})_k = V_{k,j} = 0$  for every  $j \in J$ . Now, choose an index  $i \in i_1, \dots, i_l$ . Using the equalities  $V_J\eta = 0$  and  $(V_J\eta)_i = \left(\sum_{j \in J} v_j \eta_j\right)_i = \sum_{j \in J} (v_j)_i \eta_j = 0$ , there must be at least two vectors  $v_j$  having a nonzero component  $i$  to have  $V_J\eta = \sum_j v_j \eta_j = 0 \in \mathbb{R}^n$ . Therefore, for all the  $l$  coordinates  $i_1, \dots, i_l$ , there are at least  $2l$  nonzero coordinates in the vectors  $\{v_j : j \in J\}$ , which implies  $K_J \geqslant 2l$ . This is a contradiction with  $K_J \leqslant 2k = 2|J|$ .

Now that the circuits  $J$  of size  $k$  verify  $c_J \in \{k-1, k\}$ , one only needs to count these two types of circuits for every  $k \in [3 : n+1]$ . Let  $k \in [3 : n+1]$ , and define

$$C_1(n, k) := \frac{(k-1)!}{2} \binom{n}{k-1}, \quad C_2(n, k) := \frac{(k-1)!}{2} \binom{n}{k}.$$

Note that for  $k = n+1$ ,  $C_2(n, k) = 0$  since there are only  $n$  coordinates. We show there are  $C_1(n, k)$  circuits of size  $k$  with  $c_J = k-1$  and  $C_2(n, k)$  circuits of size  $k$  with  $c_J = k$ . Since these are the only possibilities for  $c_J$ , the total number of circuits will be the claimed result:

$$\sum_{k=3}^{n+1} \frac{(k-1)!}{2} \left[ \binom{n}{k-1} + \binom{n}{k} \right] = \sum_{k=3}^{n+1} \frac{(k-1)!}{2} \binom{n+1}{k}.$$

Let  $k \in [3 : n+1]$  and  $J$  be a circuit of size  $k$  such that  $c_J = k$  (if  $k = n+1$  there is nothing to do, there are no such circuits). Let  $i_1, i_2, \dots, i_k$  be the coordinates associated to  $J$ . Since there are  $k$  coordinates, one has  $K_J \geqslant 2k$ . However, one also has  $K_J \leqslant 2k$ : every vector must have two nonzero components, meaning  $J \subseteq \{e_i - e_j\}_{1 \leqslant i < j \leqslant n}$ .

Moreover, since  $c_J = k$ , there are exactly two vectors with a nonzero component at coordinate  $i_1$ , two (other) vectors with a nonzero component at coordinate  $i_2$ , and so on. Consider a sequence of the  $k$  indices, now named  $j_1 < j_2 < \dots < j_k$ . Clearly  $e_{j_1} - e_{j_2}, e_{j_2} - e_{j_3}, \dots, e_{j_{k-1}} - e_{j_k}, e_{j_1} - e_{j_k}$  is a submatrix of nullity one, since  $(+1, \dots, +1, -1)$  is in its null space and the first  $k-1$  vectors form a family of rank  $k-1$ : the corresponding indices form a circuit of size  $k$ .

Since the choice of  $i_1, \dots, i_k$  is arbitrary, this gives  $\binom{n}{k}$  possibilities. Now, for a fixed choice of  $i_1, \dots, i_k$ , let us justify there are  $(k-1)!/2$  possible circuits of size  $k$  for these indices. There are  $k!$  possible ways to order the indices, the permutations of  $[1 : k]$ . However, the next paragraph shows the resulting circuits are independent by circular permutation and by symmetry.

Let  $\pi \in \mathfrak{S}([1 : k])$ , and denote by  $i_{\pi(1)}, \dots, i_{\pi(k)}$  the indices of the coordinates in order modified by  $\pi$ . The vectors of  $V$  that form a circuit for this order are precisely the vectors  $\pm(e_{i_{\pi(1)}} - e_{i_{\pi(2)}}), \pm(e_{i_{\pi(2)}} - e_{i_{\pi(3)}}), \dots, \pm(e_{i_{\pi(1)}} - e_{i_{\pi(k)}})$ . However, the indices given by a circular permutation of  $\pi$ , namely,  $i_{\pi(1+j_0)}, i_{\pi(2+j_0)}, \dots, i_{\pi(k+j_0)}$  (for  $j_0$  a fixed integer), form a circuit with the same vectors. Similarly, the index sequence  $i_{\pi(k)}, i_{\pi(k-1)}, \dots, i_{\pi(1)}$  form a circuit with these same vectors. Summarizing these observations, there are  $C_2(n, k)$  stem

vectors of this form: the invariance by circular permutation and by symmetry divide  $k!$  by  $k$  and by 2 respectively ( $k \geq 3$ ). Let us explain this for  $k = 4$ . For simplicity, we assume  $i_1 = 1, i_2 = 2, i_3 = 3, i_4 = 4$ . One has  $4!$  ways to order the set  $\{1, 2, 3, 4\}$ , but for instance

$$\begin{aligned} & \{e_1 - e_2, e_2 - e_3, e_3 - e_4, e_1 - e_4\} \quad [1 - 2 - 3 - 4] \\ [\text{symmetry}] &= \{e_3 - e_4, e_2 - e_3, e_1 - e_2, e_1 - e_4\} \quad [4 - 3 - 2 - 1] \\ [\text{circular}] &= \{e_2 - e_3, e_1 - e_2, e_1 - e_4, e_3 - e_4\} \quad [3 - 2 - 1 - 4] \\ [\text{symmetry}] &= \{e_1 - e_4, e_1 - e_2, e_2 - e_3, e_3 - e_4\} \quad [4 - 1 - 2 - 3] \\ [\text{circular}] &= \{e_3 - e_4, e_1 - e_4, e_1 - e_2, e_2 - e_3\} \quad [3 - 4 - 1 - 2] \\ [\text{symmetry}] &= \{e_1 - e_2, e_1 - e_4, e_3 - e_4, e_2 - e_3\} \quad [2 - 1 - 4 - 3] \\ [\text{circular}] &= \{e_1 - e_4, e_3 - e_4, e_2 - e_3, e_1 - e_2\} \quad [1 - 4 - 3 - 2] \\ [\text{symmetry}] &= \{e_2 - e_3, e_3 - e_4, e_1 - e_4, e_1 - e_2\} \quad [2 - 3 - 4 - 1] \end{aligned}$$

A different order, for instance  $\{1, 3, 2, 4\}$ , would involve new vectors such as  $e_1 - e_3$ , which means a different circuit is considered.

We have thus identified that circuits of size  $k$  with  $c_J = k$  are of the form, for coordinates  $(i_1, \dots, i_k)$  in this order, of the vectors  $\text{sgn}(i_{l+1} - i_l)(e_{i_l} - e_{i_{l+1}})$  for  $l \in [1 : k]$  (where the indices are understood  $\mod k$ ). Reciprocally, it is clear that such subsets are circuits and that there are  $C_2(n, k)$  such subsets.

Now, we consider the circuits  $J$  such that  $c_J = k - 1$ . With a similar argument, let  $k' \in [1 : k]$  be the number of vectors  $v_j$  for  $j \in J$  such that  $v_j \in \{e_i\}_{i \in [1:n]}$ . Clearly  $K_J = k' \times 1 + (k - k') \times 2 = 2k - k'$ . Since  $c_J = k - 1$  coordinates are involved in the circuits, and there are  $K_J \geq 2(k-1) = 2k - 2$  total nonzero coordinates, meaning  $k' \leq 2$ . If  $k' = 0$ , this reduces to the previous subcase: the subset is composed of two or more circuits (but two have a common index, for instance  $e_1 - e_2, e_2 - e_3, e_1 - e_3, e_1 - e_4, e_1 - e_5, e_4 - e_5$ ). If  $k' = 1$ , denoting by  $e_{i^*}$  the associated vector, since  $K_J = 2k - 1$ , the  $k - 1$  vectors of the form  $e_i - e_j$  without the vector  $e_{i^*}$  already form a circuit since they cover  $k - 1$  coordinate and are  $k - 1$ . In other words, the component of  $\eta$  corresponding to  $e_{i^*}$  equals zero: this is a contradiction. The only possibility is  $k' = 2$ , meaning there exists a pair of indices  $(i^*, j^*)$  such that  $e_{i^*}, e_{j^*}$  belong to the columns of  $V_J$ .

Now, since  $c_J = k - 1$  and  $K_J = 2k - 2$ , for every coordinate  $i_1, \dots, i_{k-1}$  must have two vectors having a nonzero component in this coordinate. Since  $e_{i^*}$  and  $e_{j^*}$  have only one nonzero component in positions  $i^*$  and  $j^*$ , one vector must be of the form  $\pm(e_{i^*} - e_{i_k})$ , one of the form  $\pm(e_{j^*} - e_{i'_k})$ , and the reasoning is pursued as before in the case  $c_J = k - 1$ . Essentially, one of the vectors  $e_i - e_j$  is split into the pair  $(e_i, e_j)$ .

When counting the circuits, since the choice of the  $k - 1$  coordinates is arbitrary, there are  $\binom{n}{k-1}$  possibilities. Then, there are  $(k - 1)!$  ways to order the coordinates and  $(k - 1)!/2$  taking into account the symmetry. However, there is no invariance by circular permutation of the involved  $k - 1$  indices. Indeed, permuting the order would (for instance) change the pair  $\{e_{i_j}, e_{i_{j+1}}\}$  into  $\{e_{i_{j+1}}, e_{i_{j+2}}\}$ , which is a different circuit. This justifies the assumption

about  $C_1(n, k)$ .

We have thus identified that circuits of size  $k$  with  $c_J = k$  are of the form, for coordinates  $(i_1, \dots, i_k)$  in this order, of the vectors  $\text{sgn}(i_{l+1} - i_l)(e_{i_l} - e_{i_{l+1}})$  except for some  $l_0 \in [1 : k]$  such that instead of  $\text{sgn}(i_{l_0+1} - i_{l_0})(e_{i_{l_0}} - e_{i_{l_0+1}})$  there is  $e_{i_{l_0}}$  and  $e_{i_{l_0+1}}$  (where the subscripts in  $i_\cdot$  are understood  $\mod k$ ). Reciprocally, it is clear that such subsets are circuits, and that there are  $C_1(n, k)$  such subsets.

To conclude the proof, let us consider the circuits of size  $n + 1$ . Clearly they cannot have  $c_J = n + 1$  since  $c_J \leq n$ . Necessarily these circuits are of the form described by the point  $c_J = k - 1$ , otherwise the amount of nonzero coordinates is incorrect. Since all  $n$  coordinates must be chosen, and  $\binom{n}{n} = 1$ , only the order of the coordinates matter: as seen previously, this number is  $(n - 1)!/2$ , multiplied by  $n$  to chose which pair of coordinates  $(i, j)$  is split with  $e_i$  and  $e_j$ .

Finally, the stem vectors are clearly obtained by looking at the circuits' structure: if  $-e_i$  or  $-(e_i - e_j)$  (for  $i < j$ ) intervenes, then the coordinate of the stem vector is  $-$ , and  $+$  otherwise.  $\square$

#### 4.5.2 About the crosspolytope separability arrangement

Then, we focus on the CROSSPOLYTOPE instances. In [35, section 6.4 p. 14], the crosspolytopes are defined as  $n$ -dimensional polytope with the  $2n$  vertices  $\mathcal{V} = \{\pm e_i\}_{i \in [1:n]}$ . From there, the associated separability arrangement has the  $2n$  hyperplanes defined by  $\{(1; v)^\perp : v \in \mathcal{V}\}$ . In particular, the "dimension" is equal to  $n + 1$  and the first coordinate is denoted by 0

$$V = \begin{bmatrix} e^\top & e^\top \\ I & -I \end{bmatrix}.$$

#### Chambers

We justify the formula  $|\mathcal{S}| = 2 \times 3^n - 2^n$ , verified numerically for reasonable  $n$ 's and corresponding to sequence First, we apply [12] to verify the cardinal given in [35]. Then, a proof by induction of this value, using the formalism of the tree algorithm, is proposed, allowing for an explicit enumeration of the chambers.

**With Athanasiadis' approach** Let  $q$  be a large prime integer and  $\mathbb{F}_q^{n+1}$  be the integer hypercube of size  $q$ . Let the equations defining the hyperplanes be written as

$$\begin{aligned} x_1 + x_0 &= 0, & x_2 + x_0 &= 0, & \dots & x_n + x_0 &= 0, \\ x_1 - x_0 &= 0, & x_2 - x_0 &= 0, & \dots & x_n - x_0 &= 0. \end{aligned}$$

For the counting of theorem 4.5.1, one can summarize the equations as requiring every  $x_i$  for  $i \in [1 : n]$  to be different from  $x_0$  and  $-x_0 \pmod{q}$ . Observe that there is a slight difference if  $x_0 = 0$  or  $x_0 \neq 0$ . Indeed, if  $x_0 = 0 = -x_0$ , then all the  $x_i$  for  $i \in [1 : n]$  can take an arbitrary value in  $[1 : q - 1]$ , which amounts to  $(q - 1)^n$  in the characteristic polynomial. Then, when  $x_0 \neq 0$  ( $q - 1$  choices), the  $x_i$  for  $i \in [1 : n]$  can take any value different from  $x_0$  and  $-x_0 (= q - x_0 \pmod{q})$ , which amounts to  $q - 2$  choices for each of the  $n$  values. Therefore, one has  $\chi(q) = (q - 1)^n + (q - 1)(q - 2)^n$ . Finally, one has

$$|\mathcal{S}| = (-1)^{n+1}[(-2)^n + (-2)(-3)^n] = -2^n + 2 * 3^n.$$

### Enumeration by induction

**Proposition 4.5.5** (circuits of the CROSSPOLYTOPE arrangements). *Let  $n \in \mathbb{N}^*$ , the circuits of the CROSSPOLYTOPE arrangements are given by the  $n(n - 1)/2$  uplets  $(i, j, i + n, j + n)$  for  $i \neq j \in [1 : n]^2$ , thus  $|\mathcal{C}(V)| = \binom{n}{2}$ .*

*Proof.* First, let us show that the indicated subsets are circuits. Indeed, the submatrix reads, removing empty lines,

$$V_{base} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}.$$

Clearly,  $V_{base}(1; -1; +1; -1) = 0 = V_{base}\text{Diag}(+1, -1, +1, -1)(1; 1; 1; 1)$  and any combination of three vectors among the four are linearly independant.

Now, consider an arbitrary circuit. Clearly it cannot have size smaller than 4. Let  $i \in [1 : n]$  and suppose the circuit contains  $v_i$  or  $v_{i+n}$ . Since only  $v_i$  and  $v_{i+n}$  have a nonzero coordinate  $i$ , necessarily the circuit must also contain the other vector. Then, the circuit is formed of pairs  $(i, i + n)$  for some indices  $i$ , but since any tuple  $(i, i + n, j, j + n)$  is already a circuit, it cannot have size  $> 4$  without contradicting the minimality.  $\square$

The stem vectors are  $\pm(+, +, -, -)$  for indices  $(i, j, 1+n, j+n)$ . The next proposition is a rather clear consequence of propositions 3.3.10 and 4.5.5. Let us “constructively” analyze this.

**Proposition 4.5.6** (chambers of the CROSSPOLYTOPE arrangements). *Let  $n \in \mathbb{N}^*$ , the  $2 * 3^n - 2^n$  chambers of the CROSSPOLYTOPE arrangements correspond to all the sign vectors of size  $2n$  such that, for every pair  $i \neq j \in [1 : n]^2$ , one does not have  $s_i = s_{i+n} = -s_j = -s_{j+n}$ .*

In what follows, the chambers of the CROSSPOLYTOPE-N arrangement in dimension  $n + 1$  are denoted by  $\mathcal{S}_n$  (also seen as chambers in dimension  $n + 2$  for the purpose of the induction).

*Proof.* The end of the statement is clear using proposition 4.5.5. Now, we proceed by induction on  $n$ . When  $n = 1$ , the two hyperplanes are  $(1, 1)^\perp, (1, -1)^\perp$ : there are four regions (it

is the only dimension for which the arrangement is complete, there are no circuits) which is  $2 \times 3^1 - 2^1 = 4$ .

Let us first justify the cardinality of the chambers. Suppose the result is true for  $n$ . First, remark that the matrices for dimension  $n$  and  $n + 1$  can be written as

$$V^n = \begin{bmatrix} 1_n & 1_n \\ I_n & -I_n \end{bmatrix}, \quad V^{n+1} = \begin{bmatrix} V_{1,:}^n & 1 & 1 \\ V_{[2:n+1],:}^n & 0_n & 0_n \\ 0_{2n} & 1 & -1 \end{bmatrix}.$$

The main idea of the proof is as follows: as only the two new vectors have a nonzero coordinate in the  $n + 1$ -th dimension,  $(1, e_{n+1})$  is not spanned by the others so every node has a descendant. But adding afterwards  $(1, -e_{n+1})^\perp$  does not duplicate again (it is spanned by the other  $2n + 1$  vectors). With the formula  $2 \times 3^n - 2^n$ , we show two things. For a specific partition  $(S^1, S^2)$  of  $\mathcal{S}_n$  with  $|S^1| = 2^n$  and  $|S^2| = 2 \times (3^n - 2^n)$ , after adding the two new hyperplanes, every  $s \in S^1 \subseteq \mathcal{S}_n$  has 4 descendants and every  $s \in S^2 \subseteq \mathcal{S}_n$  has 3 descendants. This means the total number of descendants is  $4 \times 2^n + 3 \times 2(3^n - 2^n) = 2 \times 3^{n+1} - (6 - 4)2^n = 2 \times 3^{n+1} - 2^{n+1}$ .

For that purpose, recall that when adding a hyperplane  $v^\perp$  to an arrangement  $\mathcal{A}(\{v_i\}_i)$ , a sign vector  $s$  has two descendants if and only if the associated region is split by the hyperplane, i.e., there exists  $d^s$  such that  $s_i v_i^\top d^s > 0$ ,  $v^\top d^s = 0$ . Following the same reasoning, when adding a second hyperplane,  $s$  can have 4 descendants if and only if there exists a  $d^s$  inside the region on the intersection of both added hyperplanes.

Now, let us focus on the two new hyperplanes,  $(1; 0_n; 1)^\perp$  and  $(1; 0_n; -1)^\perp$ , i.e.,  $\{d \in \mathbb{R}^{n+2} : d_0 + d_{n+1} = 0\}$  and  $\{d \in \mathbb{R}^{n+2} : d_0 - d_{n+1} = 0\}$ . Their intersection is  $\{d \in \mathbb{R}^{n+2} : d_0 = 0 = d_{n+1}\}$ . Let us justify there are precisely  $2^n$  chambers of  $\mathcal{S}_n$  around this intersection. Let  $s \in \mathcal{S}_n$ , its system of inequations is

$$\begin{cases} \forall i \in [1 : n], & s_i(d_0 + d_i) > 0, \\ \forall i \in [1 : n], & s_{i+n}(d_0 - d_i) > 0. \end{cases}$$

Now, the chambers that are around the intersection of the two new hyperplanes need to have directions  $d$  with  $d_0 = 0$  and  $d_{n+1} = 0$ . The coordinate  $d_{n+1}$  does not intervene in the above system – this is because of the independence of the two added vectors. However, when adding the constraint  $d_0 = 0$ , the above system becomes

$$\begin{cases} \forall i \in [1 : n], & s_i d_i > 0, \\ \forall i \in [1 : n], & -s_{i+n} d_i > 0, \end{cases}$$

which means  $s_i$  and  $s_{i+n}$  must be opposite. Therefore,  $s \in \mathcal{S}_n \subseteq \{\pm 1\}^{2n}$  has four descendants if and only if  $s_{n+1:2n} = -s_{1:n}$ : there are  $2^n$  possibilities. To summarize, we have shown that “ $s$  has 4 descendants  $\Rightarrow s_{n+1:2n} = -s_{1:n}$ ”; the converse is straightforward by reversing the computations. To finish this part of the proof, we need to justify the  $2^n$  possible chambers described are indeed in  $\mathcal{S}_n$ : their corresponding systems are verified in  $\mathbb{R}^{n+1}$  by the vectors  $(0; w)$  for  $w \in \{\pm 1\}^n$ , and by the vectors  $(0; w; 0)$  in  $\mathbb{R}^{n+2}$ .

Now, one must justify the remaining  $2(3^n - 2^n)$  chambers are split in 3. Remark that these chambers, by the above reasoning, do not verify  $s_{1:n} = -s_{n+1:2n}$ . Therefore, there exists  $i$  such that  $s_i = s_{i+n}$ , so the corresponding equations  $s_i(d_0 + d_i) > 0, s_i(d_0 - d_i) > 0$  mean  $d_0$  cannot be 0.

Consider such a  $s$  for which a feasible  $d \in \mathbb{R}^{n+1}$  has  $d_0 > 0$  (by symmetry, the same is true if  $d_0 < 0$ ). Now, in  $\mathbb{R}^{n+2}$ , the line  $\{(d; t) : t \in \mathbb{R}\}$  verifies the  $2n$  equations of  $s$ , which are independent of coordinate  $n+1$ . But for the two added hyperplanes, the system with  $(-, -)$  cannot be verified since:

$$\begin{cases} -d_0 - d_{n+1} > 0, \\ -d_0 + d_{n+1} > 0 \end{cases}$$

has no solution if  $d_0 > 0$ . The system  $(+, +)$  is verified for  $t = 0$ , the system  $(+, -)$  for  $t$  positive enough and  $(-, +)$  for  $t$  negative enough.

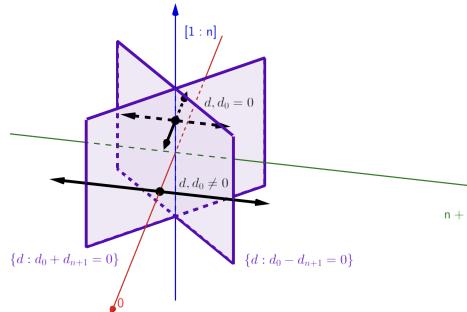


Figure 4.2: Illustration of the idea of the induction process: in purple the two hyperplanes added. The top black dot represents a point with  $d_0 = 0$  and arbitrary  $d_{[1:n]}$  which has 4 descendants. The bottom black dot represents a point with  $d_0 \neq 0$  and  $d_{[1:n]} = 0$  which has only 3 descendants (two arrow plus itself).

To conclude, the proof has shown that among the  $2 \times 3^n - 2^n$  chambers, those corresponding to sign vectors of the form  $(s, -s)$  for some  $s \in \{\pm 1\}^n$  all have exactly 4 descendants (two of which have the same form, leading to recover  $2^{n+1}$ ). Meanwhile, the remaining  $2 \times (3^n - 2^n)$  are only split in 3 chambers in  $\mathbb{R}^{n+2}$ . This amounts to the announced  $2 \times 3^{n+1} - 2^{n+1}$ .  $\square$

**Remark 4.5.7.** One could also use a similar counting argument, purely using the stem vectors. Clearly, the  $2^n$  sign vectors of the form  $(s, -s)$  cover no stem vectors. What remains is to count the number of sign vectors not of the form  $(s, -s)$  that do not cover a stem vector.

First, note that if there are  $n+1$ 's and  $n-1$ 's, either the sign vector is of the form  $(s, -s)$  and therefore already counted, or there are pairs  $(i, i+n)$  and  $(j, j+n)$  covering a stem vector (if not of the form  $(s, -s)$ , then there exists an index  $i$  with  $s_i = s_{i+n}$ , and because there are  $n+1$ 's and  $n-1$ 's there is a  $j$  with  $s_j = s_{j+n}$  with  $s_i = -s_j$ ). Thus one can count

the sign vectors having  $k < n - 1$ 's, and by symmetry multiplying by 2 at the end will be sufficient.

Then, by symmetry, we consider the case where there is a smaller number of  $-1$ . If there is  $k = 0 -1$ 's, the only possibility is  $1_{2n}$ . For  $k = 1$ , there are  $2n$  possibilities – choosing any index to put the  $-1$ . For  $k = 2$ , any possibility except having  $s_i$  and  $s_{i+n}$  equal to  $-1$ , i.e.,  $2 \times (d - 1) \times (d)$ . Continuing this reasoning, for  $k$  values at  $-1$  one needs to dispatch the  $k$  indices at positions such that there is no pair of indices  $(i, i + n)$  both with a  $-1$ . This means one chooses  $k$  of the pairs  $(i, i + n)$ , labelled  $(i_1, i_1 + n), (i_2, i_2 + n) \dots, (i_k, i_k + n)$ , and among them changes one of the signs  $s_{i_j}$  or  $s_{i_j+n}$ . This amounts to, for  $k$  changes,  $\binom{n}{k} 2^k$ .

From there, generating the chambers is straightforward. Let us verify we recover the total number of chambers:

$$\sum_{k=0}^{k=n-1} \binom{n}{k} 2^k = \sum_{k=0}^{k=n} \binom{n}{k} 2^k - 2^n = 3^n - 2^n.$$

Now, by symmetry, one gets  $2 \times (3^n - 2^n)$ ; adding the  $2^n$  sign vectors of the  $(s, -s)$  form, one gets  $2 \times 3^n - 2^n$  and every sign vector has been considered.  $\square$

#### 4.5.3 Perfectly symmetric instances

In chapter 3 (as well in the Julia code related to the next chapter), we report some encouraging results especially on some instances, that tend to have very specific structures – the matrix  $V$  is constructed in a very precise way, without any randomness. To finish this chapter, we discuss some possible future work that would combine techniques observed in [35, 212] to improve the ISF algorithm. They are based on the following definition, where here and in the rest of this chapter,  $\pi$  denotes a permutation of  $[1 : n]$ .

**Definition 4.5.8** (perfectly symmetric instances). A matrix  $V \in \mathbb{R}^{n \times p} = [v_1 \dots v_p]$  is said to be perfectly symmetric if for any  $i \in [1 : p]$  and any permutation  $\pi$  of  $[1 : n]$ , there exists an index  $j \in [1 : p]$  and a scalar  $\delta_{i,j,\pi}$  such that  $v_i^\pi = \delta_{i,j,\pi} v_j$  where  $(v_i^\pi)_k := (v_i)_{\pi(k)}$  for  $k \in [1 : n]$ .

An arrangement of hyperplanes governed by the columns of such a matrix  $V$  is said to be perfectly symmetric as well.  $\square$

**Definition 4.5.9** (symmetry group). Let  $V$  be a perfectly symmetric matrix. For  $i \in [1 : p]$ , the symmetry group of  $i$  is defined by all the indices  $j \in [1 : p]$  (counting  $i$  itself with the identity permutation) described in definition 4.5.8.  $\square$

In other words, all dimensions are “symmetric” or “equivalent”. This occurs for PERM, RESONANCE or DEMICUBE instances for example, though the THRESHOLD instances have similar properties.

**Example 4.5.10** (PERM and RESONANCE instances). Consider the following data

$$\text{PERM-4} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & -1 & -1 \end{bmatrix},$$

$$\text{RESONANCE-4} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix},$$

which are perfectly symmetric by routine verification. For PERM-4, there are two symmetry groups:  $\{1, 2, 3, 4\}$  and  $\{5, 6, 7, 8, 9, 10\}$ . For RESONANCE-4, there are four symmetry groups:  $\{1, 2, 4, 8\}$ ,  $\{3, 5, 6, 9, 10, 12\}$ ,  $\{7, 11, 13, 14\}$ ,  $\{15\}$ .  $\square$

While it seems difficult to make a general rule, it seems that types of instances verifying definition 4.5.8 could be treated in a particular way. Let us decline possibilities for algorithms using directions and others using stem vectors.

### For stem vectors

This part directly quotes [212, section 9]. When one wants to compute the circuits of perfectly symmetric instances, one can use the symmetry groups of the arrangement, to avoid many computations of (equivalent) circuits.

Let us illustrate this on PERM-4. Consider the subset formed by columns 1, 2 and 5 : it is clearly a circuit, but by the symmetry group, for any permutation  $\pi$ , so is the subset with  $j^1$ ,  $j^2$  and  $j^5$  (the indices given for 1, 2 and 5 by definition 4.5.8). For instance, with  $\pi = [4; 2; 3; 1]$ , i.e., the swap of coordinates 1 and 4, one has

$$e_1^\pi = e_4, e_2^\pi = e_2, (e_1 - e_2)^\pi = -(e_2 - e_4) = -v_9,$$

meaning that  $\{2, 4, 9\}$  is a circuit.

Once the circuits are obtained, in the tree algorithm, one could thus verify if a subvector of  $s$  is a stem vector but in a much shorter list that would take into account the symmetries – for instance, one could keep in memory the structure  $\{e_i, e_j, e_i - e_j\}$  for all those circuits instead of the  $\binom{n}{2}$  such circuits.

### For witness points

For algorithms based on the  $S$ -tree and linear optimization, such instances may have exploitable structure in  $\mathbb{R}^n$ . For instance, the permutohedron instances divide the space in  $n!$

“identical” regions of the form (see section 4.5.1)

$$R_\pi := \{x \in \mathbb{R}^n : x_{\pi(1)} > \cdots > x_{\pi(n)}\}.$$

Thus, one could compute a partial  $\mathcal{S}$ -tree for one region  $R_\pi$  and obtain the other chambers by permutation. Admittedly, the structure must be analyzed for each instance type, but its general idea may be used for other instance types.

For instance, for the RESONANCE instances, there is some symmetry in the orthants (in the classical sense,  $\mathbb{R}_+^n$  and the  $2^n - 1$  remaining ones). It is clear that  $\mathbb{R}_{++}^n$  is crossed by none of the hyperplanes, so it is a chamber (and  $\mathbb{R}_{--}^n = -\mathbb{R}_{++}^n$  as well).

Then, consider the neighboring orthants  $\{x \in \mathbb{R}^n : x_i < 0, x_{[1:n] \setminus \{i\}} > 0\}$  for  $i \in [1 : n]$ . Their decompositions by the hyperplanes are equivalent because the dimensions can be swapped. The same reasoning can be applied for orthants with 2 negative coordinates, then 3 negative coordinates and so on. By symmetry, one only needs to go to  $\lfloor \frac{n}{2} \rfloor$ , meaning that instead of  $2^n$  ( $2^{n-1}$  by symmetry of linear instances) orthants, only  $\lfloor \frac{n}{2} \rfloor$  need to be considered.

Naturally, the techniques evoked in this section are rather niche and specialized to particular instances. Moreover, they would require advanced machinery to be implemented – see [212, 35].



# Chapter 5

## Primal and dual approaches for the chamber enumeration of hyperplane arrangements

This chapter is composed of an article in preparation (initially submitted to *SIAM Journal on Discrete Mathematics* [79]). It describes the extension of the chapter 3 to noncentered arrangements.

We discuss properties such as the characterization of the symmetric chambers, the stem vectors (and the circuits, linked to matroids [151, 191, 141]), the notions of general position, of bounds on the number of chambers...

Another part of the chapter presents the so-called “compact” algorithms, which symmetrize the  $\mathcal{S}$ -tree to improve efficiency. These notions are implemented and benchmarked numerically.

This chapter goes along with chapter A, which contains details such as proofs or additional comments. In particular, we detail: a few properties of section 5.3, the proof of connectivity – as it is a more general case than in proposition 3.4.5, properties on the tested instances as well as additional results on the Julia code.

Furthermore, for harmonization purposes with the rest, the references are unified with those of the thesis, and are thus not added at the end of the article (some dates of references such as literature classics may differ from the published version). Similarly, page layout, fonts and font sizes are different.

Note: The Université de Sherbrooke asks that, for inserted articles, the contribution of the doctoral student is precised. In many publications in mathematics, the work is done collaboratively, with frequent discussions to coordinate contributions and viewpoints. This submitted paper was written jointly through a Git repository, so that all authors could edit and contribute. The Julia code providing the results at the end was written by myself.

# Primal and dual approaches for the chamber enumeration of real hyperplane arrangements

Jean-Pierre Dussault<sup>1</sup>, Jean Charles Gilbert<sup>2</sup> and Baptiste Plaquevent-Jourdain<sup>3</sup>

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 approaches, 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  $H := \{x \in \mathbb{R}^n : v^\top x = \tau\}$ , where  $v \in \mathbb{R}^n$ ,  $\tau \in \mathbb{R}$  and  $v^\top x = \sum_{i=1}^n v_i x_i$  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  $v_1, \dots, v_p \in \mathbb{R}^n$  and  $\tau_1, \dots, \tau_p \in \mathbb{R}$ , consider the collection of hyperplanes

---

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

<sup>2</sup>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

<sup>3</sup>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

$H_i := \{x \in \mathbb{R}^n : v_i^\top x = \tau_i\}$  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 properties 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  $r \geq 0$ ). For two integers  $n_1 \leq n_2$ ,  $[n_1 : n_2] := \{n_1, \dots, n_2\}$  is the set of the integers between  $n_1$  and  $n_2$ . We denote by  $\mathbb{R}_+^n := \{x \in \mathbb{R}^n : x \geq 0\}$  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  $u \cdot v \in \mathbb{R}^n$ , whose  $i$ th component is  $u_i v_i$ . The sign function  $\text{sgn} : \mathbb{R} \rightarrow \mathbb{R}$  is defined by  $\text{sgn}(t) = +1$  if  $t > 0$ ,  $\text{sgn}(t) = -1$  if  $t < 0$  and  $\text{sgn}(0) = 0$ . The sign of a vector  $x$  or a matrix  $M$  is defined componentwise:  $\text{sgn}(x)_i = \text{sgn}(x_i)$  and  $[\text{sgn}(M)]_{i,j} = \text{sgn}(M_{i,j})$  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  $\text{rank}(A) := \dim \mathcal{R}(A)$  and its nullity by  $\text{null}(A) := \dim \mathcal{N}(A) = n - \text{rank}(A)$  thanks to the rank-nullity theorem. The  $i$ th row (resp. column) of  $A$  is denoted by  $A_{i,:}$  (resp.  $A_{:,i}$ ). Transposition operates after a row and/or column selection:  $A_{i,:}^\top$  is a short notation for  $(A_{i,:})^\top$  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  $u \in \mathbb{R}^n$ ,  $\text{Diag}(u) \in \mathbb{R}^{n \times n}$  is the square diagonal matrix with  $\text{Diag}(u)_{i,i} = u_i$ . The orthogonal of a subspace  $Z \subseteq \mathbb{R}^n$  is denoted by  $Z^\perp := \{x \in \mathbb{R}^n : x^\top z = 0, \text{ for all } z \in Z\}$ .

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  $a \in \mathbb{R}^m$ ,

$$\exists x \in \mathbb{R}^n : Ax > a \iff \nexists \alpha \in \mathbb{R}_+^m \setminus \{0\} : A^\top \alpha = 0, a^\top \alpha \geq 0. \quad (5.1)$$

Gordan's theorem of the alternative [108, p. 1873] is recovered when  $a = 0$ :

$$\exists x \in \mathbb{R}^n : Ax > 0 \iff \nexists \alpha \in \mathbb{R}_+^m \setminus \{0\} : A^\top \alpha = 0. \quad (5.2)$$

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  $v_1, \dots, v_p$  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  $H := \{x \in \mathbb{R}^n : v^\top x = \tau\}$ , 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

$$H^- := \{x \in \mathbb{R}^n : v^\top x < \tau\} \quad \text{and} \quad H^+ := \{x \in \mathbb{R}^n : v^\top x > \tau\}.$$

If  $H$  is improper and  $\tau = 0$ , then  $H = \mathbb{R}^n$  and  $H^- = H^+ = \emptyset$ . If  $H$  is improper and  $\tau \neq 0$ , then  $H = \emptyset$  and  $H^+ = \mathbb{R}^n$  or  $\emptyset$ , while  $H^- = (H^+)^c$ .

A *hyperplane arrangement* is a collection of  $p \in \mathbb{N}^*$  hyperplanes  $H_i := \{x \in \mathbb{R}^n : v_i^\top x = \tau_i\}$ , for  $i \in [1 : p]$ , where  $v_1, \dots, v_p \in \mathbb{R}^n$  and  $\tau_1, \dots, \tau_p \in \mathbb{R}$ . 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  $\tau := [\tau_1; \dots; \tau_p] \in \mathbb{R}^{p \times 1}$ . 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  $\tau \in \mathcal{R}(V^\top)$  (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.

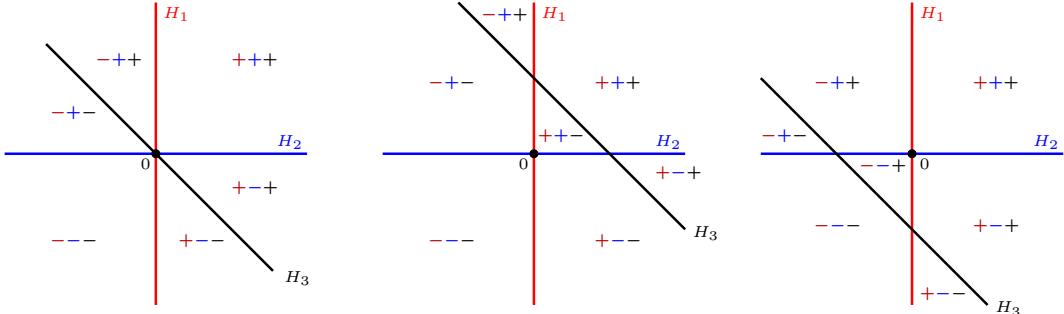


Figure 5.1: Arrangements in  $\mathbb{R}^2$  specified by the hyperplanes  $H_1 := \{x \in \mathbb{R}^2 : x_1 = 0\}$ ,  $H_2 := \{x \in \mathbb{R}^2 : x_2 = 0\}$ ,  $H_3(\text{left}) := \{x \in \mathbb{R}^2 : x_1 + x_2 = 0\}$ ,  $H_3(\text{middle}) := \{x \in \mathbb{R}^2 : x_1 + x_2 = 1\}$  and  $H_3(\text{right}) := \{x \in \mathbb{R}^2 : x_1 + x_2 = -1\}$ . The origin is contained in all the hyperplanes but in  $H_3(\text{middle})$  and  $H_3(\text{right})$ , 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

$$\mathfrak{B}([1 : p]) := \{(I_+, I_-) \in [1 : p]^2 : I_+ \cap I_- = \emptyset, I_+ \cup I_- = [1 : p]\}$$

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

$$C(I_+, I_-) := (\cap_{i \in I_+} H_i^+) \cap (\cap_{i \in I_-} H_i^-). \quad (5.3)$$

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

$$\mathfrak{C}(V, \tau) := \{(I_+, I_-) \in \mathfrak{B}([1 : p]) : C(I_+, I_-) \neq \emptyset\}. \quad (5.4)$$

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

$$\mathcal{S}(V, \tau) := \{s \in \{\pm 1\}^p : s \cdot (V^\top x - \tau) > 0 \text{ for some } x \in \mathbb{R}^n\}, \quad (5.5)$$

where “.” denotes the Hadamard product. A sign vector  $s \in \{\pm 1\}^p$  in  $\mathcal{S}(V, \tau)$  is said to be *feasible*, while it is 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

$$\phi : (I_+, I_-) \in \mathfrak{B}([1 : p]) \mapsto s \in \{\pm 1\}^p, \quad \text{where } s_i = \begin{cases} +1 & \text{if } i \in I_+ \\ -1 & \text{if } i \in I_-, \end{cases} \quad (5.6)$$

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  $\phi$  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

$$\begin{aligned} (I_+, I_-) \in \mathfrak{C}(V, \tau) &\iff \exists x \in \mathbb{R} : v_i^\top x > \tau_i \text{ for } i \in I_+ \text{ and } v_i^\top x < \tau_i \text{ for } i \in I_- \\ &\iff \exists x \in \mathbb{R} : s \cdot (V^\top x - \tau) > 0 \\ &\iff s \in \mathcal{S}(V, \tau). \end{aligned}$$

This proves the bijectivity of  $\phi : \mathfrak{C}(V, \tau) \rightarrow \mathcal{S}(V, \tau)$  and concludes the proof.  $\square$  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  $V = [v_1 \cdots v_p] \in \mathbb{R}^{n \times p}$ ,  $\tau = (\tau_1, \dots, \tau_p) \in \mathbb{R}^p$  and  $r := \text{rank}(V)$ . In the sequel, we consider the arrangement  $\mathcal{A}(V, \tau)$  formed by the  $p$  hyperplanes  $H_i := \{x \in \mathbb{R}^n : v_i^\top x = \tau_i\}$  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  $[V; \tau^\top]$ . 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  $x \mapsto x^\top v$  and  $x \mapsto x^\top \tilde{v}$  vanish simultaneously).

**Proposition 5.3.2** (parallel and identical hyperplanes). *Let  $H = \{x \in \mathbb{R}^n : v^\top x = \tau\}$  and  $\tilde{H} = \{x \in \mathbb{R}^n : \tilde{v}^\top x = \tilde{\tau}\}$  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  $(v, \tau)$  and  $(\tilde{v}, \tilde{\tau})$  are colinear in  $\mathbb{R}^n \times \mathbb{R}$ .*

The next proposition identifies some modifications of  $(V, \tau)$  that have no effect on the sign vector set  $\mathcal{S}(V, \tau)$ . For a matrix  $M$ , we denote by  $\text{sgn}(M)$  the matrix defined by  $[\text{sgn}(M)]_{i,j} = \text{sgn}(M_{i,j})$  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  $\mathcal{S}(VD, D\tau) = \text{sgn}(D)\mathcal{S}(V, \tau)$ .
- 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,  $r \leq n$  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

$$\mathcal{S}(V, 0) \text{ is symmetric.} \quad (5.7)$$

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)  $\tau \in \mathcal{R}(V^\top)$ ,
- (iii) the arrangement is centered.

*Proof.* [(i)  $\Rightarrow$  (ii)] One can decompose  $\tau$  as follows:

$$\tau = \tau^0 + V^\top \hat{x}, \quad (5.8a)$$

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  $I := \{i \in [1 : p] : \tau_i^0 \neq 0\}$  is nonempty. Define  $s \in \{\pm 1\}^p$  by

$$s_I := \text{sgn}(\tau_I^0), \quad (5.8b)$$

while  $s_{I^c}$  is defined below in order to get

$$s \notin \mathcal{S}(V, \tau) \quad \text{and} \quad -s \in \mathcal{S}(V, \tau).$$

These properties suffice to prove the implication “(i)  $\Rightarrow$  (ii)”. Set  $S_I := \text{Diag}(s_I)$ .

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

$$V_{:,I} S_I \alpha_I = 0 \quad \text{and} \quad (\tau_I^0)^\top S_I \alpha_I = \|\tau_I^0\|_2^2 \geq 0.$$

By Motzkin’s alternative (5.1) with  $A = S_I V_{:,I}^\top$  and  $a = S_I \tau_I^0$ , this is equivalent to

$$\nexists x \in \mathbb{R}^n : \quad S_I V_{:,I}^\top x > S_I \tau_I^0.$$

Hence, whatever  $s_{I^c} \in \{\pm 1\}^{I^c}$  is, there is no  $x \in \mathbb{R}^n$  such that  $s \cdot (V^\top x - \tau^0) > 0$ . Now, using (5.8a), we see that there is no  $x \in \mathbb{R}^n$  such that  $s \cdot (V^\top x - \tau) > 0$ , 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

$$-V_{:,I} S_I \alpha_I = 0 \quad \text{and} \quad -|\tau_I^0|^\top \alpha_I \geq 0$$

because the last inequality, with  $\alpha_I \geq 0$  and  $|\tau_I^0| > 0$ , implies that  $\alpha_I = 0$ . By Motzkin’s alternative (5.1) with  $A = -S_I V_{:,I}^\top$  and  $a = -|\tau_I^0| = -S_I \tau_I^0$ , this is equivalent to

$$\exists x \in \mathbb{R}^n : \quad -S_I V_{:,I}^\top x > -S_I \tau_I^0. \quad (5.8c)$$

Since the columns of  $V$  are nonzero, a small perturbation of  $x$  can maintain (5.8c) and ensures that the components of  $V_{:,I^c}^\top x$  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  $s_{I^c} := -\text{sgn}(V_{:,I^c}^\top x)$  and setting  $S_{I^c} := \text{Diag}(s_{I^c})$  leads to

$$-S_{I^c} V_{:,I^c}^\top x > 0 = -S_{I^c} \tau_{I^c}^0. \quad (5.8d)$$

Thanks to (5.8c) and (5.8d), there is an  $x \in \mathbb{R}^n$  such that  $-s \cdot (V^\top x - \tau^0) > 0$ . Now, using (5.8a), we see that there is an  $x \in \mathbb{R}^n$  such that  $-s \cdot (V^\top x - \tau) > 0$ , 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  $V^\top \hat{x} = \tau$ , 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  $V^\top \hat{x} = \tau$ . If  $s \in \mathcal{S}(V, \tau)$ , then  $s \cdot (V^\top x - \tau) > 0$  for some  $x \in \mathbb{R}^n$ , hence  $-s \cdot (V^\top (2\hat{x} - x) - \tau) = -s \cdot (V^\top (-x) + \tau) > 0$ , implying that  $-s \in \mathcal{S}(V, \tau)$ .  $\square$

It is short to prove that

$$\mathcal{S}(V, -\tau) = -\mathcal{S}(V, \tau). \quad (5.9)$$

Let us define the *symmetric* and *asymmetric* parts of  $\mathcal{S}(V, \tau)$  by

$$\mathcal{S}_s(V, \tau) := \mathcal{S}(V, \tau) \cap \mathcal{S}(V, -\tau) \quad \text{and} \quad \mathcal{S}_a(V, \tau) := \mathcal{S}(V, \tau) \setminus \mathcal{S}_s(V, \tau). \quad (5.10)$$

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 justifies the names given to  $\mathcal{S}_s(V, \tau)$  and  $\mathcal{S}_a(V, \tau)$ . One also observes that (see [80])

$$\mathcal{S}_s(V, -\tau) = -\mathcal{S}_s(V, \tau) = \mathcal{S}_s(V, \tau) \quad \text{and} \quad \mathcal{S}_a(V, -\tau) = -\mathcal{S}_a(V, \tau). \quad (5.11)$$

**Proposition 5.3.6** (symmetry in  $\mathcal{S}(V, \tau)$ ). 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  $s \cdot (V^T x) > 0$  for some  $x \in \mathbb{R}^n$ . Then,  $s \cdot (V^T(tx) - \tau) > 0$  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  $s \cdot (V^T x - \tau) > 0$  and  $-s \cdot (V^T \tilde{x} - \tau) > 0$ . After adding these inequalities side by side, one gets  $s \cdot (V^T(x - \tilde{x})) > 0$ , i.e.,  $s \in \mathcal{S}(V, 0)$ . [ $\supseteq$ ] Use  $\mathcal{S}(V, 0) \subseteq \mathcal{S}(V, \tau)$  (1a) for  $\tau$  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.  $\square$

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  $\mathcal{S}(V, \tau) \neq \emptyset$ :

$$2^r \leq |\mathcal{S}(V, \tau)| \leq 2^{\check{p}}, \quad (5.12)$$

where  $\check{p} := |\{i \in [1 : p] : V_{:,i} \neq 0\}|$ .

*Proof.* Let  $I := \{i \in [1 : p] : V_{:,i} = 0\}$ . By the assumption  $\mathcal{S}(V, \tau) \neq \emptyset$ ,  $\tau_I$  has no zero component and any  $s \in \mathcal{S}(V, \tau)$  verifies  $s_I := -\text{sgn}(\tau_I)$ . Therefore, one can only consider the components of the sign vectors that are not in  $I$ , which amounts to assuming that  $I = \emptyset$ . For the lower bound, one has  $\mathcal{S}(V, \tau) \supseteq \mathcal{S}(V, 0)$  by proposition 5.3.6(1) and  $|\mathcal{S}(V, 0)| \geq 2^r$  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  $2^{\check{p}}$ .  $\square$

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  $\tau$  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

$$C_\tau(s) := \{x \in \mathbb{R}^n : s \cdot (V^T x - \tau) > 0\},$$

which are in one to one correspondence with the sets denoted by  $C(I_1, I_2)$  in (5.3), thanks to the map  $\phi$  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

$$C_\tau(s)^\infty = \{d \in \mathbb{R}^n : s \cdot (V^\top d) \geq 0\}.$$

Its interior reads

$$\text{int } C_\tau(s)^\infty = \{d \in \mathbb{R}^n : s \cdot (V^\top d) > 0\} = C_0(s), \quad (5.13)$$

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)  $\text{int } C_\tau(s)^\infty \neq \emptyset$  for all  $s \in \mathcal{S}(V, \tau)$ ,
- (iii)  $C_0(s) \neq \emptyset$  for all  $s \in \mathcal{S}(V, \tau)$ ,
- (iv)  $C_{\tau'}(s) \neq \emptyset$  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.  $\square$

**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  $s^0, \dots, s^l \in S$  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  $S$ .  $\square$

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  $\phi$  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  $l := \sum_{i \in [1:p]} |\tilde{s}_i - s_i|/2 \leq p$  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

$$\mathcal{C}(V) := \{J \subseteq [1 : p] : J \neq \emptyset, \text{null}(V_{:,J}) = 1, \text{null}(V_{:,J_0}) = 0 \text{ for all } J_0 \subsetneq J\}, \quad (5.14)$$

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  $J := \{j_1, \dots, j_{|J|}\} \subseteq [1 : p]$  is a vector  $(\sigma_{j_1}, \dots, \sigma_{j_{|J|}})$  where the  $\sigma_j$ ’s are in  $\{-1, +1\}$ .

Note that an index set  $J \subseteq [1 : p]$  verifying  $\text{null}(V_{:,J}) = 1$  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  $\text{null}(V_{:,I}) = 1$  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

$$\begin{cases} J \in \mathcal{C}(V) \\ \sigma = \text{sgn}(\eta) \text{ for some } \eta \in \mathbb{R}^J \text{ verifying } \eta \in \mathcal{N}(V_{:,J}) \setminus \{0\} \text{ and } \tau_J^\top \eta \geqslant 0. \end{cases} \quad (5.15)$$

A stem vector is said to be *symmetric* if  $\tau_J^\top \eta = 0$  and *asymmetric* otherwise (these properties do not depend on the chosen vector  $\eta$ , as shown in remark 5.3.13(3) below). We denote

respectively by

$$\mathfrak{S}(V, \tau), \quad \mathfrak{S}_s(V, \tau) \quad \text{and} \quad \mathfrak{S}_a(V, \tau) := \mathfrak{S}(V, \tau) \setminus \mathfrak{S}_s(V, \tau)$$

the sets of stem vectors, symmetric stem vectors and asymmetric stem vectors of the arrangement  $\mathcal{A}(V, \tau)$ . We denote by  $\mathfrak{J} : \mathfrak{S}(V, \tau) \rightarrow \mathcal{C}(V)$  the map that associates with a stem vector  $\sigma \in \{\pm 1\}^J$  its circuit  $J := \mathfrak{J}(\sigma)$ .  $\square$

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  $[V; \tau^\top]$ ; the latter depend on  $\tau$ , 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),  $\text{null}(V_{:, J}) = 1$  and any  $\eta \in \mathcal{N}(V_{:, J}) \setminus \{0\}$  has no zero component (since  $\text{null}(V_{:, J_0}) = 0$  for all  $J_0 \subsetneq J$ ). Therefore,  $\text{sgn}(\eta) \in \{\pm 1\}^J$  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  $\tau_J^\top \eta = 0$  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  $\pm \text{sgn}(\eta_0)$  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  $\tau_J^\top \eta \neq 0$  for some  $\eta \in \mathcal{N}(V_{:, J})$  and, actually, for all  $\eta \in \mathcal{N}(V_{:, J}) \setminus \{0\}$  since  $\text{null}(V_{:, J}) = 1$ . In this case, there is a single asymmetric stem vector associated with  $J$ , namely  $\text{sgn}(\eta_+)$ , for some  $\eta_+ \in \mathcal{N}(V_{:, J})$  such that  $\tau_J^\top \eta_+ > 0$ .

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  $\tau_J^\top \eta > 0$ ).

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  $\sigma$  in  $\{\pm 1\}^J$ , so that  $-\sigma \in \{\pm 1\}^J$  (with the same  $J$ ). With this meaning given to  $-\sigma$ , one defines

$$-\mathfrak{S}(V, \tau) := \{-\sigma \in \{\pm 1\}^J : \sigma \in \mathfrak{S}(V, \tau) \text{ and } J := \mathfrak{J}(\sigma)\}. \quad (5.16)$$

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  $|J| \geq 2$ ,  $V_{:,J}$  has no zero column.  $\square$

It is easy to see that [80]

$$-\mathfrak{S}(V, \tau) = \mathfrak{S}(V, -\tau) \quad \text{and} \quad -\mathfrak{S}_a(V, \tau) = \mathfrak{S}_a(V, -\tau). \quad (5.17)$$

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 “ $\cup$ ” for the disjoint union of sets.

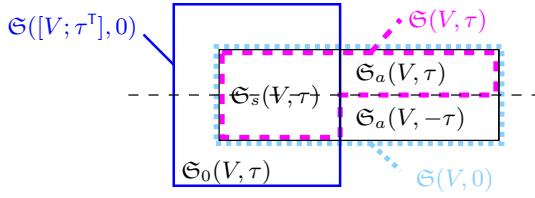


Figure 5.2: Symbolic representation of the sets  $\mathfrak{S}(V, \tau)$ ,  $\mathfrak{S}_s(V, \tau)$ ,  $\mathfrak{S}_a(V, \tau)$ ,  $\mathfrak{S}(V, 0)$ ,  $\mathfrak{S}_0(V, \tau)$  and  $\mathfrak{S}([V; \tau^T], 0)$ , respecting propositions 5.3.14, 5.3.21 and 5.3.23. The horizontal dashed line aims at representing the reflexion between a stem vector  $\sigma$  and its opposite  $-\sigma$ :  $\mathfrak{S}_s(V, \tau)$ ,  $\mathfrak{S}(V, 0)$ ,  $\mathfrak{S}_0(V, \tau)$  and  $\mathfrak{S}([V; \tau^T], 0)$  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  $-\mathfrak{S}([V; \tau^T], 0) = \mathfrak{S}([V; \tau^T], 0)$ . By propositions 5.3.15 and 5.3.21, the diagram simplifies when  $\tau \in \mathcal{R}(V^T)$ , since then  $\mathfrak{S}_a(V, \tau) = \mathfrak{S}_a(V, -\tau) = \mathfrak{S}_0(V, \tau) = \emptyset$  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)  $\mathfrak{S}_s(V, \tau) \cup \mathfrak{S}_a(V, \tau) \cup \mathfrak{S}_a(V, -\tau) = \mathfrak{S}(V, \tau) \cup \mathfrak{S}_a(V, -\tau) = \mathfrak{S}(V, 0)$ .

*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,  $\sigma = \text{sgn}(\eta) \in \{\pm 1\}^J$  for some  $J \in \mathcal{C}(V)$  and some  $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$  verifying  $\tau_J^T \eta \geq 0$  and, on the other hand,  $-\sigma = \text{sgn}(\tilde{\eta}) \in \{\pm 1\}^J$  (the same  $J$ , see remark 5.3.13(5)) for some  $\tilde{\eta} \in \mathcal{N}(V_{:,J}) \setminus \{0\}$  verifying  $\tau_J^T \tilde{\eta} \geq 0$ . Since  $\text{null}(V_{:,J}) = 1$  by (5.14),  $\tilde{\eta} = \alpha \eta$  for some  $\alpha \in \mathbb{R}^*$ . Then,  $-\sigma = \text{sgn}(\tilde{\eta}) = \text{sgn}(\alpha) \text{sgn}(\eta) = \text{sgn}(\alpha)\sigma$  shows that  $\text{sgn}(\alpha) = -1$ , so that  $0 \leq \tau_J^T \tilde{\eta} = \alpha(\tau_J^T \eta)$ . Hence  $\tau_J^T \eta \leq 0$ , so that  $\tau_J^T \eta = 0$  and  $\sigma$  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  $J \in \mathcal{C}(V)$ ,

$\sigma = \text{sgn}(\eta)$  for some  $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$  verifying  $\tau_J^\top \eta = 0$ . Then, clearly,  $-\sigma = \text{sgn}(-\eta)$  with  $-\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$  verifying  $\tau_J^\top (-\eta) = 0$ . 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  $\tau$ , which is quite clear since a stem vector  $\sigma := \text{sgn}(\eta) \in \mathfrak{S}(V, \tau)$  with  $J := \mathfrak{J}(\sigma)$  must satisfy one more property (namely  $\tau_J^\top \eta \geq 0$ ) than those in  $\mathfrak{S}(V, 0)$ .

$[\supseteq]$  Let  $\sigma \in \mathfrak{S}(V, 0)$  and  $J = \mathfrak{J}(\sigma)$ . Then,  $\sigma := \text{sgn}(\eta) \in \{\pm 1\}^J$  for some  $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ . We see that  $\sigma \in \mathfrak{S}(V, \tau)$  if  $\tau_J^\top \eta > 0$ ,  $\sigma \in \mathfrak{S}(V, -\tau)$  if  $\tau_J^\top \eta < 0$  and both sets  $\mathfrak{S}(V, \tau) \cap \mathfrak{S}(V, -\tau) = \mathfrak{S}_s(V, \tau)$  if  $\tau_J^\top \eta = 0$ .

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  $\mathfrak{S}(V, \tau) = \mathfrak{S}_s(V, \tau) \cup \mathfrak{S}_a(V, \tau)$  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.  $\square$

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  $\bar{\eta} \in \mathbb{R}^p$  the extended vector associated with  $\eta$  that is defined by  $\bar{\eta}_j = \eta_j$  for  $j \in J$  and  $\bar{\eta}_j = 0$  for  $j \notin J$ .

$[(i) \Rightarrow (ii)]$  If the arrangement is centered, one has  $\tau \in \mathcal{R}(V^\top) = \mathcal{N}(V)^\perp$  by proposition 5.3.5. A stem vector is of the form  $\text{sgn}(\eta)$  with  $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$  and  $\tau_J^\top \eta \geq 0$  for some  $J \in \mathcal{C}(V)$ . Then, the extended vector  $\bar{\eta}$  is in  $\mathcal{N}(V)$ , so that  $\tau_J^\top \eta = \tau^\top \bar{\eta} = 0$  (since  $\tau \in \mathcal{N}(V)^\perp$  and  $\bar{\eta} \in \mathcal{N}(V)$ ), showing that the stem vector is symmetric.

$[(ii) \Rightarrow (i)]$  If all the stem vectors are symmetric, then  $\tau_J^\top \eta = 0$  for all  $J \in \mathcal{C}(V)$  and  $\eta \in \mathcal{N}(V_{:,J})$ . If  $\bar{\eta}$  extends such an  $\eta$  and if we show that these  $\bar{\eta}$ 's generate  $\mathcal{N}(V)$ , we will have  $\tau \in \mathcal{N}(V)^\perp = \mathcal{R}(V^\top)$ , implying that the arrangement is centered (proposition 5.3.5).

Let  $r := \text{rank}(V)$ , 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  $\bar{\eta}$  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  $\bar{\eta}_i$  extending  $\eta_i \in \mathcal{N}(V_{:,J_i})$ ,  $i \in [1 : p]$ , are linearly independent in  $\mathbb{R}^p$ .  $\square$

The previous proposition can be rephrased in many ways, in particular as the following equivalence

$$\tau \in \mathcal{R}(V^\top) \iff \mathfrak{S}(V, \tau) = \mathfrak{S}(V, 0). \quad (5.18)$$

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  $s \in \{\pm 1\}^p$ ,*

$$s \in \mathcal{S}(V, \tau)^c \iff s_J \in \mathfrak{S}(V, \tau) \text{ for some } J \subseteq [1 : p]. \quad (5.19)$$

*Proof.*  $[ \Rightarrow ]$  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:

$$\{x \in \mathbb{R}^n : s_j(v_j^\top x - \tau_j) > 0 \text{ for all } j \in J\} = \emptyset, \quad (5.20a)$$

$$\forall J_0 \subsetneq J, \{x \in \mathbb{R}^n : s_j(v_j^\top x - \tau_j) > 0 \text{ for all } j \in J_0\} \neq \emptyset. \quad (5.20b)$$

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  $\{x \in \mathbb{R}^n : s_j(v_j^\top x - \tau_j) > 0 \text{ for all } j \in J \setminus \{j_0\}\} \neq \emptyset$  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),  $J \neq \emptyset$ . Next, by Motzkin's alternative (5.1) with  $A := \text{Diag}(s_J)V_{:,J}^\top$  and  $a := s_J \cdot \tau_J$ , (5.20a) and (5.20b) read

$$\exists \alpha \in \mathbb{R}_+^J \setminus \{0\} \text{ such that } V_{:,J}(s_J \cdot \alpha) = 0, \tau_J^\top(s_J \cdot \alpha) \geq 0, \quad (5.20c)$$

$$\forall J_0 \subsetneq J, \nexists \alpha' \in \mathbb{R}_+^{J_0} \setminus \{0\} \text{ such that } V_{:,J_0}(s_{J_0} \cdot \alpha') = 0, \tau_{J_0}^\top(s_{J_0} \cdot \alpha') \geq 0. \quad (5.20d)$$

From these properties, one deduces that  $\alpha > 0$  ( $\alpha \geq 0$  by (5.20c) and  $\alpha$  has no zero component since otherwise (5.20d) would not hold) and that  $\text{null}(V_{:,J}) \geq 1$  ( $s_J \cdot \alpha \in \mathcal{N}(V_{:,J}) \setminus \{0\}$ ). To show that  $\text{null}(V_{:,J}) = 1$ , we proceed by contradiction. Suppose that there is a nonzero  $\alpha'' \in \mathbb{R}^J$  that is not colinear with  $\alpha$  and that verifies  $V_{:,J}(s_J \cdot \alpha'') = 0$ . Since  $\alpha$  and  $\alpha''$  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  $r_1 := \max\{r \in \mathbb{R} : r\alpha \leq \alpha''\} < r_2 := \min\{r \in \mathbb{R} : \alpha'' \leq r\alpha\}$ ). One can also assume that  $\tau_J^\top(s_J \cdot \beta) \geq 0$  (otherwise replace  $\beta$  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  $\beta$  has positive components and the definition of  $t$ , we see that  $\emptyset \neq J_0 \subsetneq J$ . Let us introduce  $\alpha' := \alpha + t\beta \geq 0$ , which verifies  $\alpha'_j > 0$  for  $j \in J_0 \neq \emptyset$  and  $\alpha'_j = 0$  for  $j \in J \setminus J_0 \neq \emptyset$ . Therefore,

$$\begin{aligned} V_{:,J_0}(s_{J_0} \cdot \alpha'_{J_0}) &= V_{:,J}(s_J \cdot \alpha') \quad [\alpha'_{J \setminus J_0} = 0] \\ &= V_{:,J}(s_J \cdot \alpha) + t V_{:,J}(s_J \cdot \beta) \quad [\alpha' := \alpha + t\beta] \\ &= t V_{:,J}(s_J \cdot \alpha'') - rt V_{:,J}(s_J \cdot \alpha) \quad [V_{:,J}(s_J \cdot \alpha) = 0, \beta = \alpha'' - r\alpha] \\ &= 0 \quad [V_{:,J}(s_J \cdot \alpha'') = V_{:,J}(s_J \cdot \alpha) = 0] \end{aligned}$$

and

$$\begin{aligned} \tau_{J_0}^\top(s_{J_0} \cdot \alpha'_{J_0}) &= \tau_J^\top(s_J \cdot \alpha') \quad [\alpha'_{J \setminus J_0} = 0] \\ &= \tau_J^\top(s_J \cdot \alpha) + t \tau_J^\top(s_J \cdot \beta) \quad [\alpha' := \alpha + t\beta] \\ &\geq 0 \quad [\tau_J^\top(s_J \cdot \alpha) \geq 0, \tau_J^\top(s_J \cdot \beta) \geq 0, t > 0]. \end{aligned}$$

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  $\beta$  would be two linearly independent vectors in  $\mathcal{N}(V_{:,J})$  (since  $s_J \cdot \alpha$  has no zero component), contradicting  $\text{null}(V_{:,J}) = 1$ .

Now, since  $s_J = \text{sgn}(s_J \cdot \alpha)$ , since  $s_J \cdot \alpha \in \mathcal{N}(V_{:,J})$  and  $\tau_J^\top(s_J \cdot \alpha) \geq 0$  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  $s_J := \text{sgn}(\eta) \in \{\pm 1\}^J$  for some  $J \in \mathcal{C}$  and some  $\eta \in \mathbb{R}^J$  satisfying  $V_{:,J}\eta = 0$  and  $\tau_J^\top\eta \geq 0$ . Then,  $\alpha := s_J \cdot \eta = |\eta|$  is in  $\mathbb{R}_+^J \setminus \{0\}$  and verifies  $V_{:,J}(s_J \cdot \alpha) = 0$  and  $\tau_J^\top(s_J \cdot \alpha) \geq 0$ . By Motzkin's alternative (5.1), there is no  $x \in \mathbb{R}^n$  such that  $s_J \cdot (V_{:,J}^\top x - \tau_J) > 0$ . Hence, there is certainly no  $x \in \mathbb{R}^n$  such that  $s \cdot (V^\top x - \tau) > 0$ . This means that  $s \in \mathcal{S}(V, \tau)^c$ .  $\square$

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  $V := \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$  and  $\tau^\top := [0 \ 0 \ 1 \ 2]$  (add one hyperplane perpendicular to  $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$  in figure 5.1(middle)). More is said on this situation in the equivalence (5.22) below.  $\square$

### 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  $s \cdot (V^\top x - \tau) > 0$  or, equivalently, whether there is a pair  $(x, \xi) \in \mathbb{R}^n \times \mathbb{R}$  such that

$$s \cdot ([V; \tau^\top]^\top [x; \xi]) > 0 \quad \text{and} \quad \xi = -1.$$

The first condition above reads  $s \in \mathcal{S}([V; \tau^\top], 0)$  and refers to the linear arrangement in  $\mathbb{R}^{n+1}$  governed by the *augmented matrix*  $[V; \tau^\top]$ . This presentation of the problem shows that there must be links between the following sign vector sets and between the following stem vector sets

$$\mathcal{S}(V, 0), \quad \mathcal{S}(V, \tau) \quad \text{and} \quad \mathcal{S}([V; \tau^\top], 0), \tag{5.21a}$$

$$\mathfrak{S}(V, 0), \quad \mathfrak{S}(V, \tau) \quad \text{and} \quad \mathfrak{S}([V; \tau^\top], 0). \tag{5.21b}$$

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  $[V; \tau^\top]$  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  $[V; \tau^\top]$ , 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.

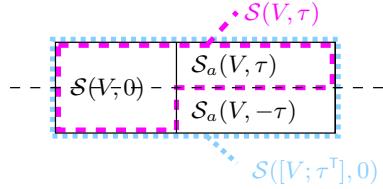


Figure 5.3: Symbolic representation of the sets  $\mathcal{S}(V, 0)$ ,  $\mathcal{S}(V, \tau)$ ,  $\mathcal{S}_a(V, \tau)$  and  $\mathcal{S}([V; \tau^T], 0)$ , respecting (5.9), (5.10), (5.11) and propositions 5.3.6 and 5.3.18. The horizontal dashed line aims at representing the reflection between a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

**Proposition 5.3.18** (properties with  $\mathcal{S}([V; \tau^T], 0)$ ). *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)  $\mathcal{S}(V, \tau) \cap \mathcal{S}(V, -\tau) = \mathcal{S}(V, 0) \subseteq \mathcal{S}(V, \tau) \subseteq \mathcal{S}([V; \tau^T], 0) = \mathcal{S}(V, \tau) \cup \mathcal{S}(V, -\tau)$ .
- 2)  $\mathcal{S}(V, \tau)^c \cup \mathcal{S}(V, -\tau)^c = \mathcal{S}(V, 0)^c \supseteq \mathcal{S}(V, \tau)^c \supseteq \mathcal{S}([V; \tau^T], 0)^c = \mathcal{S}(V, \tau)^c \cap \mathcal{S}(V, -\tau)^c$ .
- 3)  $\mathcal{S}(V, 0) \cup \mathcal{S}_a(V, \tau) \cup \mathcal{S}_a(V, -\tau) = \mathcal{S}([V; \tau^T], 0)$ .
- 4)  $2|\mathcal{S}(V, \tau)| = |\mathcal{S}(V, 0)| + |\mathcal{S}([V; \tau^T], 0)|$ .
- 5)  $2|\mathcal{S}(V, \tau)^c| = |\mathcal{S}(V, 0)^c| + |\mathcal{S}([V; \tau^T], 0)^c|$ .

*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  $s \cdot (V^T x - \tau) > 0$  for some  $x \in \mathbb{R}^n$  or  $s \cdot ([V; \tau^T]^T [x; -1]) > 0$ , implying that  $s \in \mathcal{S}([V; \tau^T], 0)$ .

Consider now the last equality.  $[\subseteq]$  Let  $s \in \mathcal{S}([V; \tau^T], 0)$ , so that  $s \cdot (V^T x + \tau \xi) > 0$  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,  $\mathcal{S}(V, \tau) \subseteq \mathcal{S}([V; \tau^T], 0)$  and  $\mathcal{S}(V, -\tau) = -\mathcal{S}(V, \tau) \subseteq -\mathcal{S}([V; \tau^T], 0) = \mathcal{S}([V; \tau^T], 0)$ .

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  $\mathcal{S}(V, 0) \cap \mathcal{S}_a(V, \pm \tau) = \emptyset$ . Use the same arguments

to get that

$$\begin{aligned}\mathcal{S}_a(V, \tau) \cap \mathcal{S}_a(V, -\tau) &= [\mathcal{S}(V, \tau) \cap \mathcal{S}(V, 0)^c] \cap [\mathcal{S}(V, -\tau) \cap \mathcal{S}(V, 0)^c] \\ &= [\mathcal{S}(V, \tau) \cap \mathcal{S}(V, -\tau)] \cap \mathcal{S}(V, 0)^c \\ &= \mathcal{S}(V, 0) \cap \mathcal{S}(V, 0)^c = \emptyset\end{aligned}$$

Consider now the identity:

$$\begin{aligned}\mathcal{S}(V, 0) \cup \mathcal{S}_a(V, \tau) \cup \mathcal{S}_a(V, -\tau) &= \mathcal{S}(V, 0) \cup [\mathcal{S}(V, \tau) \cap \mathcal{S}(V, 0)^c] \cup [\mathcal{S}(V, -\tau) \cap \mathcal{S}(V, 0)^c] \\ &= \mathcal{S}(V, 0) \cup \mathcal{S}(V, \tau) \cup \mathcal{S}(V, -\tau) \\ &= \mathcal{S}([V; \tau^T], 0) \quad [\text{point 1}].\end{aligned}$$

4) The identity results from

$$\begin{aligned}|\mathcal{S}([V; \tau^T], 0)| &= |\mathcal{S}(V, \tau)| + |\mathcal{S}(V, -\tau)| - |\mathcal{S}(V, \tau) \cap \mathcal{S}(V, -\tau)| \quad [\text{point 1}] \\ &= 2|\mathcal{S}(V, \tau)| - |\mathcal{S}(V, 0)| \quad [(5.9), (5.10), \text{proposition 5.3.6(2)}].\end{aligned}$$

5) Each set in point 4 is a part of  $\{\pm 1\}^p$  of cardinality  $2^p$ . Hence, the identity in point 4 gives

$$2(2^p - |\mathcal{S}(V, \tau)^c|) = (2^p - |\mathcal{S}(V, 0)^c|) + (2^p - |\mathcal{S}([V; \tau^T], 0)^c|).$$

Point 5 follows after subtracting  $2^{p+1}$  from both sides.  $\square$

**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  $\mathcal{A}([V; \tau^T], 0)$ .

2) As a corollary of proposition 5.3.18(2), one has (see remark 5.3.17):

$$\pm s \in \mathcal{S}(V, \tau)^c \iff s_J \in \mathfrak{S}([V; \tau^T], 0) \text{ for some } J \subseteq [1 : p]. \quad (5.22)$$

Indeed  $\pm s \in \mathcal{S}(V, \tau)^c$  if and only if  $s \in \mathcal{S}(V, \tau)^c \cap \mathcal{S}(V, -\tau)^c = \mathcal{S}([V; \tau^T], 0)^c$  (proposition 5.3.18(2)), which is equivalent to  $s_J \in \mathfrak{S}([V; \tau^T], 0)$  for some  $J \in [1 : p]$  (proposition 5.3.16). Note that  $\mathfrak{S}([V; \tau^T], 0)$  is symmetric, so that the properties in (5.22) are also equivalent to the fact that  $s_J \in -\mathfrak{S}([V; \tau^T], 0)$  for some  $J \subseteq [1 : p]$ .  $\square$

## 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  $[V; \tau^T]$ . Recall from proposition 5.3.5 that an arrangement is centered if and only if  $\tau \in \mathcal{R}(V^T)$ . Note also that

$$\text{rank}([V; \tau^T]) = \begin{cases} \text{rank}(V) & \text{if } \tau \in \mathcal{R}(V^T) \\ \text{rank}(V) + 1 & \text{otherwise,} \end{cases} \quad (5.23a)$$

$$\text{null}([V; \tau^T]) = \begin{cases} \text{null}(V) & \text{if } \tau \in \mathcal{R}(V^T) \\ \text{null}(V) - 1 & \text{otherwise.} \end{cases} \quad (5.23b)$$

The formula of  $\text{rank}([V; \tau^T])$  is clear and the one of  $\text{null}([V; \tau^T])$  can be deduced from (5.23a) by the rank-nullity theorem.

**Proposition 5.3.20** (circuits of  $V$  and  $[V; \tau^T]$ ). *Let  $V \in \mathbb{R}^{n \times p}$  and  $\tau \in \mathbb{R}^p$ . Then, the following properties are equivalent:*

- (i)  $\mathcal{C}(V) = \mathcal{C}([V; \tau^T])$ ,
- (ii)  $\mathcal{C}(V) \subseteq \mathcal{C}([V; \tau^T])$ ,
- (iii)  $\tau \in \mathcal{R}(V^T)$ , meaning that the arrangement  $\mathcal{A}(V, \tau)$  is centered.

*Proof.* [(i)  $\Rightarrow$  (ii)] Clear.

[(ii)  $\Rightarrow$  (iii)] Let  $J \in \mathcal{C}(V)$ . Then,  $\text{null}(V_{:,J}) = 1$  by (5.14). By assumption,  $J \in \mathcal{C}([V; \tau^T])$ , so that  $\text{null}([V; \tau^T]_{:,J}) = 1$ , as well. By (5.23b),  $\tau_J \in \mathcal{R}(V_{:,J}^T) = \mathcal{N}(V_{:,J})^\perp$ . 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  $\tau \in \mathcal{R}(V^T)$ , one has  $\tau_J \in \mathcal{R}(V_{:,J}^T)$  and  $\tau_{J_0} \in \mathcal{R}(V_{:,J_0}^T)$ , so that (5.23b) yields

$$\text{null}([V; \tau^T]_{:,J}) = \text{null}(V_{:,J}) \quad \text{and} \quad \text{null}([V; \tau^T]_{:,J_0}) = \text{null}(V_{:,J_0}).$$

It follows that  $\text{null}([V; \tau^T]_{:,J}) = 1$  and  $\text{null}([V; \tau^T]_{:,J_0}) = 0$  for all  $J_0 \subsetneq J$  if and only if  $\text{null}(V_{:,J}) = 1$  and  $\text{null}(V_{:,J_0}) = 0$  for all  $J_0 \subsetneq J$ . In other words,  $J \in \mathcal{C}([V; \tau^T])$  if and only if  $J \in \mathcal{C}(V)$ . We have shown that  $\mathcal{C}([V; \tau^T]) = \mathcal{C}(V)$ .  $\square$

The implication (ii)  $\Rightarrow$  (i) of lemma 5.3.20 is not based on the fact that one would always have  $\mathcal{C}(V) \supseteq \mathcal{C}([V; \tau^T])$ , which is not true. As a counter-example, take  $V = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$  and  $\tau^T = [0 \ 0 \ 1 \ 2]$ , in which case one has  $\mathcal{C}(V) = \{\{3, 4\}, \{1, 2, 3\}, \{1, 2, 4\}\}$ , while  $\mathcal{C}([V; \tau^T]) = \{\{1, 2, 3, 4\}\}$ . Actually, the property  $\mathcal{C}(V) \supseteq \mathcal{C}([V; \tau^T])$ , which is therefore weaker than those in lemma 5.3.20, has various other equivalent interesting formulations, including  $\mathfrak{S}_s(V, \tau) = \mathfrak{S}([V; \tau^T], 0)$ , 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  $\mathcal{A}(V, \tau)$  and  $\mathcal{A}([V; \tau^T], 0)$ ). *For any  $V$  and  $\tau$ ,*

$$\mathfrak{S}_a(V, \tau) \cap \mathfrak{S}([V; \tau^T], 0) = \emptyset \quad \text{and} \quad \mathfrak{S}_s(V, \tau) \subseteq \mathfrak{S}([V; \tau^T], 0), \quad (5.24)$$

*with equality in the last inclusion if the arrangement is centered. More precisely, the following properties are equivalent:*

- (i)  $\mathfrak{S}_s(V, \tau) = \mathfrak{S}([V; \tau^T], 0)$ ,
- (ii)  $\mathfrak{S}_s(V, \tau) \supseteq \mathfrak{S}([V; \tau^T], 0)$ ,
- (iii)  $\mathcal{C}(V) \supseteq \mathcal{C}([V; \tau^T])$ ,

(iv)  $\tau_J \in \mathcal{R}(V_{:,J}^T)$ , for all  $J \in \mathcal{C}([V; \tau^T])$ .

*Proof.* 1) [(5.24)<sub>1</sub>] Let  $\sigma \in \mathfrak{S}([V; \tau^T], 0)$  and  $J := \mathfrak{J}(\sigma)$ . Then,  $\sigma = \text{sgn}(\eta)$  for some  $\eta \in \mathcal{N}([V; \tau^T]_{:,J}) \setminus \{0\}$ . This  $\eta$  verifies  $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$  and  $\tau_J^T \eta = 0$ . It follows that  $\sigma \notin \mathfrak{S}_s(V, \tau)$ , 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  $\eta$ .

[(5.24)<sub>2</sub>] Let  $\sigma \in \mathfrak{S}_s(V, \tau)$  and  $J := \mathfrak{J}(\sigma)$ . Then,  $J \in \mathcal{C}(V)$  and  $\sigma = \text{sgn}(\eta)$  for some  $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$  verifying  $\tau_J^T \eta = 0$ . It follows that  $\eta \in \mathcal{N}([V; \tau^T]_{:,J}) \setminus \{0\}$ . To show that  $\sigma \in \mathfrak{S}([V; \tau^T], 0)$ , one still has to verify that  $J \in \mathcal{C}([V; \tau^T])$  (definition 5.14). First,  $J \neq \emptyset$ , since  $J \in \mathcal{C}(V)$ . Next,  $\text{null}([V; \tau^T]_{:,J}) = 1$ , since  $\eta \in \mathcal{N}([V; \tau^T]_{:,J}) \setminus \{0\}$  and  $\text{null}([V; \tau^T]_{:,J}) \leq \text{null}(V_{:,J}) = 1$  (because  $J \in \mathcal{C}(V)$ ). Finally, for all  $J_0 \subsetneq J$ , one has  $\text{null}([V; \tau^T]_{:,J_0}) = 0$ , since  $\text{null}([V; \tau^T]_{:,J_0}) \leq \text{null}(V_{:,J_0}) = 0$  (because  $J \in \mathcal{C}(V)$ ).

2) Suppose now that the arrangement  $\mathcal{A}(V, \tau)$  is centered (or  $\tau \in \mathcal{R}(V^T)$ ) and let us show that  $\mathfrak{S}_s(V, \tau) \supseteq \mathfrak{S}([V; \tau^T], 0)$  (this could also be viewed as a consequence of the implication (iv)  $\Rightarrow$  (i) proved below). Let  $\sigma \in \mathfrak{S}([V; \tau^T], 0)$  and  $J := \mathfrak{J}(\sigma)$ . Then,  $J \in \mathcal{C}([V; \tau^T])$  and  $\sigma = \text{sgn}(\eta)$  for some  $\eta \in \mathcal{N}([V; \tau^T]_{:,J}) \setminus \{0\}$ . Then,  $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$  and  $\tau_J^T \eta = 0$ . 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  $J \in \mathcal{C}([V; \tau^T])$ . By remark 5.3.13(3.a), there is a stem vector  $\sigma \in \{\pm 1\}^J$  that is in  $\mathfrak{S}([V; \tau^T], 0)$ , hence in  $\mathfrak{S}_s(V, \tau)$  by (ii). This latter fact implies that  $J \in \mathcal{C}(V)$ .

[(iii)  $\Rightarrow$  (iv)] Let  $J \in \mathcal{C}([V; \tau^T])$ . By (iii),  $J \in \mathcal{C}(V)$ , so that  $\text{null}([V; \tau^T]_{:,J}) = \text{null}(V_{:,J})$  (these nullities = 1). Then, (5.23b) implies that  $\tau_J \in \mathcal{R}(V_{:,J}^T)$ .

[(iv)  $\Rightarrow$  (ii)] Let  $\sigma = \text{sgn}(\eta) \in \mathfrak{S}([V; \tau^T], 0)$  and  $J := \mathfrak{J}(\sigma)$ . Since  $\eta \in \mathcal{N}([V; \tau^T]_{:,J}) \setminus \{0\} \subseteq \mathcal{N}(V_{:,J}) \setminus \{0\}$  and  $\tau^T \eta = 0$ , it suffices to show that  $J \in \mathcal{C}(V)$ . This property follows from  $J \in \mathcal{C}([V; \tau^T]_{:,J})$ , since  $J \in \mathcal{C}([V; \tau^T])$ , from  $\mathcal{C}([V; \tau^T]_{:,J}) = \mathcal{C}(V_{:,J})$ , by  $\tau_J \in \mathcal{R}(V_{:,J}^T)$  and the implication (iii)  $\Rightarrow$  (i) of proposition 5.3.20, and from  $\mathcal{C}(V_{:,J}) \subseteq \mathcal{C}(V)$ .  $\square$

**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  $V = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$  and  $\tau^T = [0 \ 0 \ 0 \ 1]$ . One has

$$\mathcal{C}(V) = \{\{3, 4\}, \{1, 2, 3\}, \{1, 2, 4\}\} \quad \text{and} \quad \mathcal{C}([V; \tau^T]) = \{\{1, 2, 3\}\}.$$

We see that  $\mathcal{C}([V; \tau^T])$  is included in  $\mathcal{C}(V)$  but is not equal to it. Note also that  $\tau \notin \mathcal{R}(V^T)$ ,

$$\begin{aligned}\mathfrak{S}_s(V, \tau) &= \mathfrak{S}([V; \tau^T], 0) = \left\{ \pm \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} \in \{\pm 1\}^{\{1,2,3\}} \right\} \quad \text{and} \\ \mathfrak{S}_a(V, \tau) &= \left\{ \begin{bmatrix} -1 \\ 1 \end{bmatrix} \in \{\pm 1\}^{\{3,4\}}, \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix} \in \{\pm 1\}^{\{1,2,4\}} \right\}.\end{aligned}$$

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  $\tau_K \in \mathcal{R}(V_{:,K}^T)$ , for  $K := \cup\{J \in \mathcal{C}([V; \tau^T])\}$ . To see this, take  $V = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \end{bmatrix}$  and  $\tau^T = [0 \ 0 \ 1 \ 0 \ 0 \ 1]$ . One has  $\mathcal{C}([V; \tau^T]) = \{\{1, 4\}, \{2, 5\}, \{3, 6\}\}$ ,  $\tau_J \in \mathcal{R}(V_{:,J}^T)$  for all  $J \in \mathcal{C}([V; \tau^T])$  and  $K = [1 : 6]$ , but  $\tau \notin \mathcal{R}(V^T)$ .  $\square$

Recall (5.3.21)<sub>2</sub>. In the algorithm of section 5.6.3, it will be interesting to partition  $\mathfrak{S}([V; \tau^T], 0)$  in  $\mathfrak{S}_s(V, \tau)$  and  $\mathfrak{S}_0(V, \tau)$ , where

$$\mathfrak{S}_0(V, \tau) := \mathfrak{S}([V; \tau^T], 0) \setminus \mathfrak{S}_s(V, \tau).$$

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  $\sigma \in \mathfrak{S}([V; \tau^T], 0)$  and  $J = \mathfrak{J}(\sigma)$ . Then,*

$$\sigma \in \mathfrak{S}_0(V, \tau) \iff \begin{cases} J \in \mathcal{C}([V; \tau^T]) \setminus \mathcal{C}(V) \\ \sigma = \text{sgn}(\eta) \text{ for some } \eta \in \mathcal{N}([V; \tau^T]_{:,J}) \setminus \{0\}. \end{cases} \quad (5.25a)$$

$$\sigma \in \mathfrak{S}_s(V, \tau) \iff J \in \mathcal{C}(V) \iff \text{null}(V_{:,J}) = 1 \iff \tau_J \in \mathcal{R}(V_{:,J}^T). \quad (5.25b)$$

### 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])

$$|\mathcal{S}(V, 0)| = \sum_{J \subseteq [1:p]} (-1)^{\text{null}(V_{:,J})}, \quad (5.26)$$

where the term in the right-hand side corresponding to  $J = \emptyset$  is 1 (one takes the convention that  $\text{null}(V_{:,\emptyset}) = 0$ ).

**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,*

$$|\mathcal{S}(V, \tau)| = \sum_{\substack{J \subseteq [1:p] \\ \tau_J \in \mathcal{R}(V_{:,J}^\top)}} (-1)^{\text{null}(V_{:,J})}, \quad (5.27a)$$

$$|\mathcal{S}_a(V, \tau)| = \sum_{\substack{J \subseteq [1:p] \\ \tau_J \notin \mathcal{R}(V_{:,J}^\top)}} (-1)^{\text{null}(V_{:,J})-1}, \quad (5.27b)$$

where the term in the right-hand side of (5.27a) corresponding to  $J = \emptyset$  is 1 (it is considered that  $\tau_J \in \mathcal{R}(V_{:,J}^\top)$  for  $J = \emptyset$ ).

*Proof.* Using proposition 5.3.18(4), Winder's formula (5.26) and (5.23b), one gets

$$\begin{aligned} 2|\mathcal{S}(V, \tau)| &= |\mathcal{S}(V, 0)| + |\mathcal{S}([V; \tau^\top], 0)| \\ &= \sum_{J \subseteq [1:p]} (-1)^{\text{null}(V_{:,J})} + \sum_{J \subseteq [1:p]} (-1)^{\text{null}([V; \tau^\top]_{:,J})} \\ &= 2 \sum_{\substack{J \subseteq [1:p] \\ \tau_J \in \mathcal{R}(V_{:,J}^\top)}} (-1)^{\text{null}(V_{:,J})} + \sum_{\substack{J \subseteq [1:p] \\ \tau_J \notin \mathcal{R}(V_{:,J}^\top)}} [(-1)^{\text{null}(V_{:,J})} + (-1)^{\text{null}(V_{:,J})-1}] \\ &= 2 \sum_{\substack{J \subseteq [1:p] \\ \tau_J \in \mathcal{R}(V_{:,J}^\top)}} (-1)^{\text{null}(V_{:,J})}, \end{aligned}$$

since  $(-1)^{\text{null}(V_{:,J})} + (-1)^{\text{null}(V_{:,J})-1} = 0$  (note that  $\text{null}(V_{:,J}) > 0$  if  $\tau_J \notin \mathcal{R}(V_{:,J}^\top)$ ). Formula (5.27a) follows. Formula (5.27b) of  $|\mathcal{S}_a(V, \tau)|$  comes from (5.10)<sub>2</sub>,  $\mathcal{S}_s(V, \tau) = \mathcal{S}(V, 0)$  and proposition 5.3.18(1), which implies that  $|\mathcal{S}_a(V, \tau)| = |\mathcal{S}(V, \tau)| - |\mathcal{S}(V, 0)|$ .  $\square$

Note that one recovers (5.26) from (5.27a) when the arrangement is centered (i.e.,  $\tau \in \mathcal{R}(V^\top)$ ). By proposition 5.3.18(1),  $\mathcal{S}(V, 0) \subseteq \mathcal{S}(V, \tau)$ , so that  $|\mathcal{S}(V, 0)| \leq |\mathcal{S}(V, \tau)|$ , 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])

$$|\mathcal{S}(V, 0)| \leq 2 \sum_{i \in [0:r-1]} \binom{p-1}{i}. \quad (5.28)$$

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

$$\begin{aligned}\forall I \subseteq [1 : p] : \quad & \dim(\cap_{i \in I} H_i^0) = n - \min(|I|, r), \\ \forall I \subseteq [1 : p] : \quad & \text{rank}(V_{:, I}) = \min(|I|, r),\end{aligned}$$

where  $H_i^0 := \{x \in \mathbb{R}^n : V_{:, i}^T x = 0\}$  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  $\dim(\cap_{i \in I} H_i^0) = n - \text{rank}(V_{:, I})$ . Observe that the inequality  $\text{rank}(V_{:, I}) \leq \min(|I|, r)$  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  $p \leq n$ . 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  $\text{rank}(V_{:, I}) = \min(|I|, r)$  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  $\tau \in \mathbb{R}^p$ . For the proper affine arrangement  $\mathcal{A}(V, \tau)$ , the following two conditions are equivalent:*

- (i)  $|\mathcal{S}([V; \tau^T], 0)| = 2 \sum_{i \in [0 : r]} \binom{p-1}{i}$ ,
- (ii)  $\forall I \subseteq [1 : p] : \text{rank}([V; \tau^T]_{:, I}) = \min(|I|, r + 1)$ .

**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  $H_i := \{x \in \mathbb{R}^n : V_{:, i}^T x = \tau_i\}$  for  $i \in [1 : p]$ . Then, the following*

properties are equivalent:

$$\forall I \subseteq [1 : p] : \begin{cases} \cap_{i \in I} H_i \neq \emptyset \text{ and } \dim(\cap_{i \in I} H_i) = n - |I| & \text{if } |I| \leq r \\ \cap_{i \in I} H_i = \emptyset & \text{if } |I| \geq r + 1, \end{cases} \quad (5.29a)$$

$$\forall I \subseteq [1 : p] : \begin{cases} \text{rank}(V_{:,I}) = |I| & \text{if } |I| \leq r \\ \text{rank}([V; \tau^T]_{:,I}) = r + 1 & \text{if } |I| \geq r + 1, \end{cases} \quad (5.29b)$$

$$\forall I \subseteq [1 : p] : \begin{cases} \text{rank}(V_{:,I}) = \min(|I|, r) \\ \text{rank}([V; \tau^T]_{:,I}) = \min(|I|, r + 1). \end{cases} \quad (5.29c)$$

**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.  $\square$

**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  $V = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$  and  $\tau^T = [0 \ 0 \ 1]$ , 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)<sub>2</sub>).

4) Affine general position usually holds when  $V$  of rank  $r$  and  $\tau$  are randomly generated.  $\square$

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)<sub>1</sub>]). The right-hand side of (5.30) is sequence A008949 in [185].

**Proposition 5.3.31** (bound on  $|\mathcal{S}(V, \tau)|$ ). *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,*

$$|\mathcal{S}(V, \tau)| \leq \sum_{i \in [0:r]} \binom{p}{i}, \quad (5.30)$$

*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  $\text{rank}([V; \tau^T]) \leq r + 1$ . Then, by (5.28), one has

$$|\mathcal{S}(V, 0)| \leq 2 \sum_{i \in [0:r-1]} \binom{p-1}{i}, \quad (5.31a)$$

$$|\mathcal{S}([V; \tau^T], 0)| \leq 2 \sum_{i \in [0:r]} \binom{p-1}{i}. \quad (5.31b)$$

Using these estimates in proposition 5.3.18(4) provides

$$\begin{aligned} |\mathcal{S}(V, \tau)| &= \frac{1}{2} |\mathcal{S}(V, 0)| + \frac{1}{2} |\mathcal{S}([V; \tau^T], 0)| \\ &\leq \underbrace{\binom{p-1}{0}}_{\binom{p}{0}} + \underbrace{\binom{p-1}{1}}_{\binom{p}{1}} + \cdots + \underbrace{\binom{p-1}{r-1}}_{\binom{p}{r-1}} \\ &\quad + \underbrace{\binom{p-1}{0}}_{\binom{p}{r}} + \cdots + \underbrace{\binom{p-1}{r-2}}_{\binom{p}{r-1}} + \underbrace{\binom{p-1}{r-1}}_{\binom{p}{r}} + \underbrace{\binom{p-1}{r}}_{\binom{p}{r}} \quad (5.31c) \\ &= \sum_{i \in [0:r]} \binom{p}{i}, \end{aligned}$$

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).  $\square$

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 than the bound (5.30) on  $|\mathcal{S}(V, \tau)|$ , unless  $r = p$ , in which case the affine arrangement is centered (necessarily  $\tau \in \mathcal{R}(V^T)$ ) 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  $\phi$ , 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:

$$\forall j \in [1:p] : V_{:,j} \neq 0. \quad (5.32)$$

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  $s \cdot (V^\top x - \tau) > 0$  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

$$\mathcal{S}_k \equiv \mathcal{S}_k(V, \tau) := \mathcal{S}(V_{\cdot, [1:k]}, \tau_{[1:k]}) \quad \text{and} \quad \mathcal{S}_k^c := \{\pm 1\}^k \setminus \mathcal{S}_k. \quad (5.33)$$

Hence,  $\mathcal{S}_k$  is the sign vector set of the arrangement associated with the matrix  $V_{\cdot, [1:k]} \in \mathbb{R}^{n \times k}$  and the vector  $\tau_{[1:k]} \in \mathbb{R}^k$ . Let us denote by  $v_i$  the  $i$ th column of  $V$ , by  $\tau_i$  the  $i$ th component of  $\tau$  and by  $H_i := \{x \in \mathbb{R}^n : v_i^\top x = \tau_i\}$  the  $i$ th 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  $(+1)(v_1^\top x_+ - \tau_1) > 0$  and  $(-1)(v_1^\top x_- - \tau_1) > 0$  are satisfied by the following two witness points, located on either side of the hyperplane  $H_1$ :

$$x_+ := (\tau_1 + 1)v_1/\|v_1\|^2 \quad \text{and} \quad x_- := (\tau_1 - 1)v_1/\|v_1\|^2. \quad (5.34)$$

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,

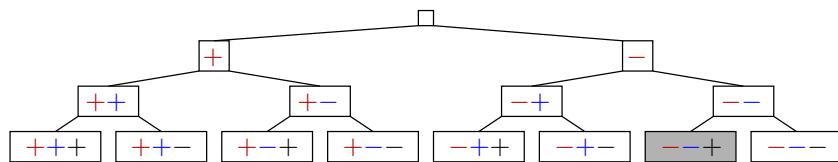


Figure 5.4:  $\mathcal{S}$ -tree of the arrangement in the middle pane of figure 5.1. The gray node is actually absent from the tree, since there is no chamber associated with  $s = (-1, -1, +1)$  (no  $x$  such that  $s \cdot (V^\top x - \tau) > 0$ ).

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.,  $\tau \in \mathcal{R}(V^\top)$ ),  $\mathcal{S}(V, \tau)$  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  $(V, \tau)$  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)<sub>1</sub>
2. **P\_STREE\_REC**(−1,  $x_-$ ) //  $x_-$  given by (5.34)<sub>2</sub>

**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** ( $v_{k+1}^\top x = \tau_{k+1}$ )
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.  $s_{k+1} := \text{sgn}(v_{k+1}^\top x - \tau_{k+1})$  //  $(s, s_{k+1}) \in \mathcal{S}_{k+1}$
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  $s_i(v_i^\top x - \tau_i) > 0$  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  $v_{k+1}^\top x = \tau_{k+1}$  (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  $\varepsilon > 0$ ,  $x_\pm^\varepsilon := x \pm \varepsilon v_{k+1}$  satisfies  $s_i(v_i^\top x_\pm^\varepsilon - \tau_i) > 0$ , for all  $i \in [1 : k]$  and  $\pm(v_{k+1}^\top x_\pm^\varepsilon - \tau_{k+1}) > 0$ . 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  $v_{k+1}^\top x \neq \tau_{k+1}$ , so that  $s_{k+1} := \text{sgn}(v_{k+1}^\top x - \tau_{k+1}) \in \{\pm 1\}$  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):

$$\begin{cases} s_i(v_i^\top \tilde{x} - \tau_i) > 0, & \text{for } i \in [1 : k] \\ -s_{k+1}(v_{k+1}^\top \tilde{x} - \tau_{k+1}) > 0. \end{cases} \quad (5.35)$$

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

$$\begin{aligned} \min_{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R}} \quad & \alpha \\ \text{s.t.} \quad & s_i(v_i^\top x - \tau_i) + \alpha \geq 0, \quad \text{for } i \in [1 : k] \\ & -s_{k+1}(v_{k+1}^\top x - \tau_{k+1}) + \alpha \geq 0 \\ & \alpha \geq -1. \end{aligned} \quad (5.36)$$

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  $(s, -\text{sgn}(\tau_{k+1}))$ .  $\square$

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,  $P_{k+1} : \mathbb{R}^n \rightarrow H_{k+1} - H_{k+1}$  denotes the orthogonal projector on the subspace  $v_{k+1}^\perp$  that is parallel to the affine space  $H_{k+1} := \{x \in \mathbb{R}^n : v_{k+1}^\top x = \tau_{k+1}\}$ ; while  $\hat{x}_{k+1} := \tau_{k+1}v_{k+1}/\|v_{k+1}\|^2$  is the unique point in  $\mathcal{N}(P_{k+1}) \cap H_{k+1}$ . We also denote by  $P_{k+1}$  the transformation matrix of the projector, so that  $P_{k+1} V_{:, [1:k]}$  can be viewed as the product of two matrices (its  $j$ th column is  $P_{k+1} v_j$ , for  $j \in [1 : k]$ ). Note that in (5.37b),  $\tilde{V} := P_{k+1} V_{:, [1:k]}$  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  $\mathcal{N}(P_{k+1}) \cap H_{k+1}$ . Then,*

$$(s, +1) \text{ and } (s, -1) \in \mathcal{S}_{k+1} \iff \exists x \in \mathbb{R}^n : s_i(v_i^\top x - \tau_i) > 0, \text{ for } i \in [1 : k], \text{ and } v_{k+1}^\top x - \tau_{k+1} = 0 \quad (5.37a)$$

$$\iff s \in \mathcal{S}(P_{k+1} V_{:, [1:k]}, \tau_{[1:k]} - V_{:, [1:k]}^\top \hat{x}_{k+1}). \quad (5.37b)$$

*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):

$$(s, +1) \text{ and } (s, -1) \in \mathcal{S}(V_{k+1}, \tau_{[1:k+1]}) \iff \begin{cases} \exists x_+ \in \mathbb{R}^n : s_i(v_i^\top x_+ - \tau_i) > 0, \text{ for } i \in [1 : k], \\ \text{and } +(v_{k+1}^\top x_+ - \tau_{k+1}) > 0 \\ \exists x_- \in \mathbb{R}^n : s_i(v_i^\top x_- - \tau_i) > 0, \text{ for } i \in [1 : k], \\ \text{and } -(v_{k+1}^\top x_- - \tau_{k+1}) > 0 \end{cases} \quad (5.38a)$$

$$\iff \exists x \in \mathbb{R}^n : s_i(v_i^\top x - \tau_i) > 0, \text{ for } i \in [1 : k], \text{ and } v_{k+1}^\top x - \tau_{k+1} = 0 \quad (5.38a)$$

$$\iff \exists x \in \mathbb{R}^n : s_i([P_{k+1} v_i]^\top x - [\tau_i - v_i^\top \hat{x}_{k+1}]) > 0, \text{ for } i \in [1 : k] \quad (5.38b)$$

$$\iff s \in \mathcal{S}(P_{k+1} V_k, \tau_{[1:k]} - V_k^\top \hat{x}_{k+1}).$$

The equivalence in (5.38a) is shown as follows.

[ $\Rightarrow$ ] Define  $t_- := +(v_{k+1}^\top x_+ - \tau_{k+1}) > 0$ ,  $t_+ := -(v_{k+1}^\top x_- - \tau_{k+1}) > 0$ ,  $\alpha_- := t_-/(t_- + t_+) \in (0, 1)$  and  $\alpha_+ := t_+/(t_- + t_+) \in (0, 1)$ . Then,  $x = \alpha_+ x_+ + \alpha_- x_-$  is appropriate since

$$\text{for } i \in [1 : k] : s_i(v_i^\top x - \tau_i) = \alpha_+ s_i[v_i^\top x_+ - \tau_i] + \alpha_- s_i[v_i^\top x_- - \tau_i] > 0, \\ v_{k+1}^\top x - \tau_{k+1} = \alpha_+[v_{k+1}^\top x_+ - \tau_{k+1}] + \alpha_-[v_{k+1}^\top x_- - \tau_{k+1}] = \alpha_+ t_- - \alpha_- t_+ = 0.$$

[ $\Leftarrow$ ] Take  $x_\pm = x \pm \varepsilon v_{k+1}$  for a sufficiently small  $\varepsilon > 0$ .

The equivalence in (5.38b) is shown as follows.

[ $\Rightarrow$ ] The point  $x$  in (5.38a) satisfies  $x \in H_{k+1}$ , so that  $x = P_{k+1} x + \hat{x}_{k+1}$ , since  $x - P_{k+1} x \in \mathcal{N}(P_{k+1}) \cap H_{k+1} = \{\hat{x}_{k+1}\}$ . Furthermore, by (5.38a), one has for  $i \in [1 : k]$ :

$k]$ :

$$0 < s_i(v_i^\top x - \tau_i) = s_i(v_i^\top [P_{k+1}x + \hat{x}_{k+1}] - \tau_i) = s_i([P_{k+1}v_i]^\top x - [\tau_i - v_i^\top \hat{x}_{k+1}]),$$

where we have used  $P_{k+1}^\top = P_{k+1}$  ( $P_{k+1}$  is an orthogonal projector).

[ $\Leftarrow$ ] Take  $x := P_{k+1}x_0 + \hat{x}_{k+1} \in H_{k+1}$  in (5.38a), where  $x_0$  is the  $x$  given by (5.38b).  $\square$

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  $\text{vect}\{v_1, \dots, v_k\}$  the vector space spanned by the vectors  $v_1, \dots, v_k$ .

**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,  $|\mathcal{S}_{k+1}^c| \geq 2|\mathcal{S}_k^c|$ .*
- 2) *If  $v_{k+1} \notin \text{vect}\{v_1, \dots, v_k\}$ , 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  $v_{k+1} \in \text{vect}\{v_1, \dots, v_k\}$ ,  $V_{:, [1:k+1]}$  has no zero column,  $H_{k+1} \neq H_i$  for  $i \in [1 : k]$ , and  $r_k := \dim \text{vect}\{v_1, \dots, v_k\}$ , then  $|\mathcal{S}_{k+1}| \geq |\mathcal{S}_k| + 2^{r_k-1}$ .*

*Proof.* 1) If  $s \in \mathcal{S}_k^c$ , there is no  $x \in \mathbb{R}^n$  such that  $s_i(v_i^\top x - \tau_i) > 0$  for  $i \in [1 : k]$ . Therefore, there is certainly no  $x \in \mathbb{R}^n$  satisfying  $s_i(v_i^\top x - \tau_i) > 0$  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  $|\mathcal{S}_{k+1}^c| \geq 2|\mathcal{S}_k^c|$ .

2) Let  $Q$  be the orthogonal projector on  $\text{vect}\{v_1, \dots, v_k\}^\perp$  for the Euclidean scalar product. By assumption,  $Qv_{k+1} \neq 0$ . Let  $s \in \mathcal{S}_k$ , so that there is an  $x \in \mathbb{R}^n$  such that  $s_i(v_i^\top x - \tau_i) > 0$  for  $i \in [1 : k]$ . For any  $t \in \mathbb{R}$  and  $i \in [1 : k]$ , the points  $x_\pm := x \pm t Qv_{k+1}$  verify  $s_i(v_i^\top x_\pm - \tau_i) = s_i(v_i^\top x - \tau_i) > 0$  (because  $v_i^\top Qv_{k+1} = 0$ ). In addition, for  $t > 0$  sufficiently large, one has  $\pm(v_{k+1}^\top x_\pm - \tau_{k+1}) = \pm(v_{k+1}^\top x - \tau_{k+1}) + t\|Qv_{k+1}\|^2 > 0$  (because  $Q^2 = Q$  and  $Q^\top = Q$ ). We have shown that both  $(s, +1)$  and  $(s, -1)$  are in  $\mathcal{S}_{k+1}$ . Therefore,  $|\mathcal{S}_{k+1}| \geq 2|\mathcal{S}_k|$ .

Now,  $|\mathcal{S}_k| + |\mathcal{S}_k^c| = 2^k$ ,  $|\mathcal{S}_{k+1}| + |\mathcal{S}_{k+1}^c| = 2^{k+1}$  and  $|\mathcal{S}_{k+1}^c| \geq 2|\mathcal{S}_k^c|$  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  $\text{null}(P_{k+1}) = 1$  (since  $\mathcal{N}(P_{k+1}) = \mathbb{R}v_{k+1}$ ) and  $\text{rank}(V_{:, [1:k]}) = r_k$  (by definition). Then,  $\text{rank}(P_{k+1}V_{:, [1:k]}) \geq r_k - 1$ . To apply (5.12) to the arrangement  $\mathcal{A}(P_{k+1}V_{:, [1:k]}, \tau_{[1:k]} - V_{:, [1:k]}^\top \hat{x}_{k+1})$ , one must show that  $v_i^\top \hat{x}_{k+1} \neq \tau_i$  when  $i \in [1 : k]$  and  $P_{k+1}v_i = 0$  (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  $v_i^\top \hat{x}_{k+1} = \tau_i$  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),  $|\mathcal{S}(P_{k+1}V_{:, [1:k]}, \tau_{[1:k]} - V_{:, [1:k]}^\top \hat{x}_{k+1})| \geq 2^{r_k-1}$ .

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  $|\mathcal{S}_{k+1}| \geq |\mathcal{S}_k| + 2^{r_k-1}$ .  $\square$

**Corollary 5.4.6** (lower bound of  $|\mathcal{S}(V, \tau)|$ ). *For a matrix  $V$  of rank  $r$  without zero column and a vector  $\tau$  such that all the hyperplanes  $H_i$  are different, one has*

$$2^r + 2^{r-1}(p - r) \leq |\mathcal{S}(V, \tau)|. \quad (5.39)$$

*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  $k \in [r : p - 1]$ ,  $\dim \text{vect}\{v_1, \dots, v_k\} = r$ , so that proposition 5.4.5(3) implies that  $|\mathcal{S}_{k+1}| \geq |\mathcal{S}_k| + 2^{r-1}$ . By induction, one gets (5.39).  $\square$

### 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  $r := \text{rank}(V)$  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  $v_1, \dots, v_r$  (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  $V_r^\top \hat{x} = \tau_{[1:r]}$ , by  $\hat{x} := V_r(V_r^\top V_r)^{-1} \tau_{[1:r]}$ . Next, for any  $s \in \{\pm 1\}^r$ , one computes

$d_s := Q_r R_r^{-\top} s \in \mathbb{R}^n$ . Let us show that  $x_s := \hat{x} + d_s$  is a witness point of the considered  $s$ . One has

$$V_r^\top x_s - \tau_{[1:r]} = V_r^\top [V_r(V_r^\top V_r)^{-1} \tau_{[1:r]} + Q_r R_r^{-\top} s] - \tau_{[1:r]} = (Q_r R_r)^\top Q_r R_r^{-\top} s = s.$$

Therefore,  $s \cdot (V_r^\top x_s - \tau_{[1:r]}) = s \cdot s = e > 0$ , 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.,  $v_{k+1}^\top x = \tau_{k+1}$ , 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  $s \cdot (V_{:, [1:k]}^\top x - \tau_{[1:k]}) > 0$ . Suppose that  $v_{k+1} \neq 0$  and*

$$\underbrace{\max_{s_i v_i^\top v_{k+1} > 0} \frac{\tau_i - v_i^\top x}{v_i^\top v_{k+1}}}_{=: t_{\min}} < \underbrace{\frac{\tau_{k+1} - v_{k+1}^\top x}{\|v_{k+1}\|^2}}_{=: t_0} < \underbrace{\min_{s_i v_i^\top v_{k+1} < 0} \frac{\tau_i - v_i^\top x}{v_i^\top v_{k+1}}}_{=: t_{\max}}. \quad (5.40)$$

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  $v_{k+1}^\top x = \tau_{k+1}$  ( $t_0 = 0$ ), that is, when  $x \in H_{k+1}$ . One has, for  $i \in [1 : k]$ ,

$$s_i(v_i^\top(x + tv_{k+1}) - \tau_i) > 0 \iff \begin{cases} t < \frac{s_i(\tau_i - v_i^\top x)}{s_i v_i^\top v_{k+1}} & \text{if } s_i v_i^\top v_{k+1} < 0 \\ t \in \mathbb{R} & \text{if } s_i v_i^\top v_{k+1} = 0 \\ t > \frac{s_i(\tau_i - v_i^\top x)}{s_i v_i^\top v_{k+1}} & \text{if } s_i v_i^\top v_{k+1} > 0. \end{cases}$$

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

$$\begin{aligned} t > t_0 := \frac{\tau_{k+1} - v_{k+1}^\top x}{\|v_{k+1}\|^2} &\iff v_{k+1}^\top(x + tv_{k+1}) - \tau_{k+1} > 0, \\ t < t_0 := \frac{\tau_{k+1} - v_{k+1}^\top x}{\|v_{k+1}\|^2} &\iff v_{k+1}^\top(x + tv_{k+1}) - \tau_{k+1} < 0. \end{aligned}$$

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.  $(s, -1)$ ).  $\square$

## 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  $(v_i, \tau_i)$ , 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  $T_s := \{i_1^s, \dots, i_k^s\}$  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  $|v_i^\top x - \tau_i| / \|v_i; \tau_i\|$  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.  $\mathfrak{S}_s = \emptyset$  and  $\mathfrak{S}_a = \emptyset$
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  $(\tau_J^\top \eta_J = 0)$
6.  $\mathfrak{S}_s := \mathfrak{S}_s \cup \{\text{sgn}(\eta_J)\} \cup \{-\text{sgn}(\eta_J)\}$
7. ELSE
8.  $\mathfrak{S}_a := \mathfrak{S}_a \cup \{\text{sgn}(\tau_J^\top \eta_J) \text{ sgn}(\eta_J)\}$
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  $\sigma$  such that  $\mathfrak{J}(\sigma)$  is contained in the set formed of  $I_0$  and indices larger than  $\max(I_0)$ .
- 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,  $\text{null}(V_{:,I}) = 1$  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  $\text{null}(V_{:,I'}) = 1$  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,  $\eta_I$  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  $\pm \text{sgn } \eta_J$  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  $\tau^T = [1, 0, 1]$ , 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.  $Sc = \emptyset$  // 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.    $Sc = Sc \cup STEM\_TO\_INFEAS\_SIGN\_VECTORS(\sigma, p)$
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  $(s, s_{k+1}) \in \mathcal{S}_{k+1}^c$  and any  $\tilde{s} \in \{\pm 1\}^p$ , extending  $(s, s_{k+1})$  by  $\pm 1$ , 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( $\emptyset, \mathfrak{S}_s \cup \mathfrak{S}_a$ )

**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

$$\begin{aligned} \max_{(\lambda, \mu) \in \mathbb{R}^{k+1} \times \mathbb{R}} \quad & \sum_{i \in [1:k]} \lambda_i s_i \tau_i - \lambda_{k+1} s_{k+1} \tau_{k+1} - \mu \\ \text{s.t.} \quad & \lambda \geq 0, \mu \geq 0 \\ & \sum_{i \in [1:k]} \lambda_i s_i v_i = \lambda_{k+1} s_{k+1} v_{k+1} \\ & \sum_{i \in [1:k+1]} \lambda_i + \mu = 1, \end{aligned} \tag{5.41}$$

where  $\lambda \in \mathbb{R}^{k+1}$  is the dual variable associated with the first  $k+1$  constraints of (3.43) and  $\mu$  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  $\text{val}(3.43)$  (resp.  $\text{val}(5.41)$ ) 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  $\text{val}(3.43) = \text{val}(5.41)$ .

**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  $\text{val}(3.43) \geq 0$ ,  $\lambda_{k+1} > 0$  and  $\mu = 0$ .*
- 3) *If, in addition,  $(\lambda, \mu)$  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  $\text{val}(3.43) = 0$ ,  $\pm(s, -s_{k+1})_J$  are the two symmetric stem vectors associated with  $J$ ,*
  - *if  $\text{val}(3.43) > 0$ ,  $(s, -s_{k+1})_J$  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  $(\lambda, \mu)$ .

2) Suppose that  $s \in \mathcal{S}_k$  and that  $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}$ . Let  $(x, \alpha)$  be a solution to (3.43) ( $\alpha = \text{val}(3.43)$  is uniquely determined). Let us show that

$$\lambda_{k+1} > 0 \quad \text{and} \quad \mu = 0. \tag{5.42a}$$

The optimal multiplier  $\mu$  is associated with the constraint  $\alpha \geq -1$  of the optimization problem (3.43), which is inactive ( $\alpha \geq 0$  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  $0 \leq \alpha = \text{val}(3.43) = \text{val}(5.41) = \sum_{i \in [1:k]} \lambda_i s_i \tau_i$ , 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  $s_i(v_i^\top x - \tau_i) > 0$ , 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]

$$\left\{ \begin{pmatrix} s_i v_i \\ 1 \end{pmatrix}_{i \in I}, \begin{pmatrix} -s_{k+1} v_{k+1} \\ 1 \end{pmatrix} \right\} \text{ are linearly independent,} \quad (5.42b)$$

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

$$\{s_i v_i\}_{i \in I} \text{ are linearly independent.} \quad (5.42c)$$

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  $\alpha_i$ '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

$$\text{null}(V_{:,J} D) = 1.$$

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  $\tau_J^\top \eta = \text{val}(3.43)$ , we see that the number of stem vectors associated with  $J$  is governed by  $\text{val}(3.43)$ , as described in remark 5.3.13(3). In addition,  $\text{sgn}(\eta) = (s, -s_{k+1})_J$ , because  $\lambda_J > 0$ , showing that  $(s, -s_{k+1})_J$  is a stem vector.  $\square$

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.  $\mathfrak{S} = \emptyset$
2. PD\_STREE\_REC(+1,  $x_+$ ,  $\mathfrak{S}$ ) //  $x_+$  given by (5.34)<sub>1</sub>
3. PD\_STREE\_REC(-1,  $x_-$ ,  $\mathfrak{S}$ ) //  $x_-$  given by (5.34)<sub>2</sub>

**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 ( $v_{k+1}^\top x = \tau_{k+1}$ )
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.  $s_{k+1} := \text{sgn}(v_{k+1}^\top x - \tau_{k+1})$  //  $(s, s_{k+1}) \in \mathcal{S}_{k+1}$
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  $(s, -s_{k+1}) \in \mathcal{S}_{k+1}^c$ , 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  $\mathcal{S}_k(V, \tau)$ )

$$\mathcal{T}(V, \tau) := \bigcup_{k \in [1:p]} \mathcal{S}_k(V, \tau). \quad (5.43)$$

When the arrangement is not centered (equivalently,  $\tau \notin \mathcal{R}(V^\top)$ ), 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  $\mathcal{T}(V, 0) = \bigcup_{k \in [1:p]} \mathcal{S}_k(V, 0)$ . 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  $\tau$  by

$$\mathcal{A}_k(V, \tau) := \mathcal{A}(V_{:, [1:k]}, \tau_{[1:k]}).$$

By proposition 5.3.18, we have that

$$\mathcal{S}_k(V, 0) \subseteq \mathcal{S}_k(V, \tau) \subseteq \mathcal{S}_k([V; \tau^\top], 0) \quad (5.44a)$$

$$\mathcal{S}_k([V; \tau^\top], 0) \setminus \mathcal{S}_k(V, 0) = \mathcal{S}_{a,k}(V, \tau) \cup \mathcal{S}_{a,k}(V, -\tau). \quad (5.44b)$$

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  $\mathcal{T}(V, 0) \cup \mathcal{T}_a(V, \tau)$ . Taking the intersection with  $\mathcal{T}^+(V, \tau)$  and  $\mathcal{T}^-(V, \tau)$  partitions  $\mathcal{T}(V, \tau)$  into four sets:

$$\mathcal{T}(V, 0) \cap \mathcal{T}^+(V, \tau), \mathcal{T}_a(V, \tau) \cap \mathcal{T}^+(V, \tau), \mathcal{T}(V, 0) \cap \mathcal{T}^-(V, \tau), \mathcal{T}_a(V, \tau) \cap \mathcal{T}^-(V, \tau). \quad (5.45)$$

Since

$$\mathcal{T}(V, 0) \cap \mathcal{T}^-(V, \tau) = -[\mathcal{T}(V, 0) \cap \mathcal{T}^+(V, \tau)], \quad (5.46)$$

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:

$$\mathcal{T}^+(V, \tau) \quad \text{and} \quad \mathcal{T}_a(V, \tau) \cap \mathcal{T}^-(V, \tau).$$

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

$$\mathcal{S}_k(V, 0), \quad \mathcal{S}_{a,k}(V, \tau) \quad \text{or} \quad \mathcal{S}_{a,k}(V, -\tau). \quad (5.47)$$

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  $\boxdot$  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  $\mathcal{T}^+(V, \tau)$ .

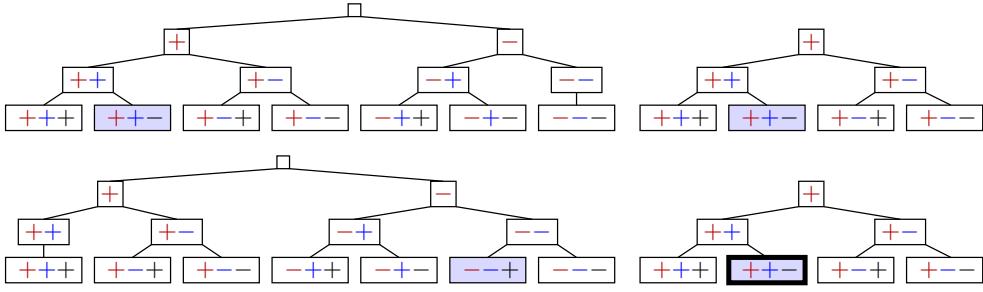


Figure 5.5: Standard  $\mathcal{S}$ -trees (left) and compact  $\mathcal{S}$ -trees (right) of the arrangements in the middle pane (above, compare with figure 5.4) and the right-hand side pane (below) of figure 5.1. The sign vectors in the white boxes are in  $\mathcal{T}(V, 0)$ , those in the blue/gray 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  $s \in \mathcal{S}_k(V, 0) \cup \mathcal{S}_{a,k}(V, \tau) \cup \mathcal{S}_{a,k}(V, -\tau)$  be a sign vector of the compact  $\mathcal{S}$ -tree. Set  $\mathcal{S}_k^\pm(V, \tau) := \mathcal{S}_k(V, \tau) \cap \mathcal{T}^\pm(V, \tau)$ .*

- 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  $\mathcal{S}_k^+(V, \tau) \cup (-[\mathcal{S}_{a,k}(V, \tau) \cap \mathcal{T}^-(V, \tau)])$ .

### 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  $(s, x, \boxed{s})$ , where  $s \in \{\pm 1\}^k$  is the sign vector of the node,  $\boxed{s} \in \{-1, 0, +1\}$  is a flag specifying to which sign set  $s$  belongs and  $x$  is some witness point. More specifically,

$$\begin{cases} s \in \mathcal{S}_k(V, 0), x \text{ is a witness point for } s \text{ in } \mathcal{A}_k(V, 0) & \text{if } \boxed{s} = 0, \\ s \in \mathcal{S}_{a,k}(V, \tau), x \text{ is a witness point for } s \text{ in } \mathcal{A}_k(V, \tau) & \text{if } \boxed{s} = +1, \\ s \in \mathcal{S}_{a,k}(V, -\tau), x \text{ is a witness point for } s \text{ in } \mathcal{A}_k(V, -\tau) & \text{if } \boxed{s} = -1. \end{cases} \quad (5.48)$$

The flag  $\boxed{s}$  is used below as a scalar, hence  $\boxed{s}s$  is the vector whose  $i$ th component is  $\boxed{s}s_i$ . 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  $(s, x, \boxed{s})$  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.,  $\boxed{s} = 0$ .

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,  $v_{k+1}^T x \neq 0$ , implying that  $(s, s_{k+1}) \in \mathcal{S}_{k+1}(V, 0)$  for  $s_{k+1} := \text{sgn}(v_{k+1}^T x)$  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  $\mathcal{S}_{k+1}([V; \tau^T], 0)^c$  (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  $(x_0, \alpha_0)$  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  $(s, -s_{k+1}) \in \mathcal{S}_{k+1}([V; \tau^T], 0)$  by solving the following LOP, which is similar to (3.43) with  $\tau = 0$ , but for the arrangement  $\mathcal{A}([V; \tau^T], 0)$  instead of  $\mathcal{A}(V, 0)$ :

$$\begin{aligned} \min_{(x, \xi, \alpha) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}} \quad & \alpha \\ \text{s.t.} \quad & s_i(v_i^T x + \tau_i \xi) + \alpha \geq 0, \quad \text{for } i \in [1 : k] \\ & -s_{k+1}(v_{k+1}^T x + \tau_{k+1} \xi) + \alpha \geq 0 \\ & \alpha \geq -1. \end{aligned} \tag{5.49}$$

Denote by  $(x_1, \xi_1, \alpha_1)$  a solution to this problem.

1.2.1. If  $\alpha_1 \geq 0$ , then  $(s, -s_{k+1}) \notin \mathcal{S}_{k+1}([V; \tau^T], 0)$ , 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  $(s, -s_{k+1}) \in \mathcal{S}_{k+1}([V; \tau^T], 0) \setminus \mathcal{S}_{k+1}(V, 0) = \mathcal{S}_{a,k+1}(V, \tau) \cup \mathcal{S}_{a,k+1}(V, -\tau)$  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  $x_1/\xi_1$  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.,  $\boxed{s} = +1$ .

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,  $v_{k+1}^\top x \neq \tau_{k+1}$ , implying that  $(s, s_{k+1}) \in \mathcal{S}_{k+1}(V, \tau)$  for  $s_{k+1} := \text{sgn}(v_{k+1}^\top x - \tau_{k+1})$  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  $(x_2, \alpha_2)$  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.,  $\boxed{s} = -1$ , 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  $\boxed{s}$  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,  $\boxed{s}$ )`).**

It is assumed that  $s \in \{\pm 1\}^p$  and that  $\boxed{s} \in \{-1, 0, +1\}$ .

1. `IF` ( $\boxed{s} = 0$ )
2.   `OUTPUT`  $\pm s$       //  $s \in \mathcal{S}(V, 0)$
3. `ELSE`
4.   `OUTPUT`  $\boxed{s} s$       //  $s \in \mathcal{S}_a(V, \boxed{s} \tau)$
5. `ENDIF`

**Algorithm 5.6.3 (`C_P_TWO_CHILDREN(s, x,  $\boxed{s}$ )`).**

It is assumed that  $s \in \{\pm 1\}^k$  and that  $(s, x, \boxed{s})$  satisfies (5.48).

1. `IF` ( $v_{k+1}^\top x \simeq \boxed{s} \tau_{k+1}$ )      // two easy children in  $\mathcal{A}_{k+1}(V, \boxed{s} \tau)$
2.   `C_P_STREE_REC`(( $s, +1$ ),  $x + t_+ v_{k+1}, \boxed{s}$ ) for some  $t_+ \in (t_0, t_{\max})$
3.   `C_P_STREE_REC`(( $s, -1$ ),  $x + t_- v_{k+1}, \boxed{s}$ ) 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  $\mathbf{s} = \pm 1$  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( $s, x, \mathbf{s}$ )).**

It is assumed that  $s \in \{\pm 1\}^k$  and that  $(s, x, \mathbf{s})$  satisfies (5.48).

1. IF ( $k = p$ ) //  $s$  is a leaf of the compact  $\mathcal{S}$ -tree
2. OUTPUT\_S( $s, \mathbf{s}$ )
3. RETURN
4. ENDIF
5. IF (C\_P\_TWO\_CHILDREN( $s, x, \mathbf{s}$ )) // two easy children in  $\mathcal{A}_{k+1}(V, \mathbf{s}\tau)$
6. RETURN
7. ENDIF
8.  $s_{k+1} := \text{sgn}(v_{k+1}^\top x - \mathbf{s}\tau_{k+1})$  //  $(s, s_{k+1}) \in \mathcal{S}_{k+1}(V, \mathbf{s}\tau)$
9. c\_p\_STREE\_REC( $(s, s_{k+1}), x, \mathbf{s}$ )
10. IF ( $\mathbf{s} = 0$ ) //  $s \in \mathcal{S}_k(V, 0)$  with  $x$  as witness point in  $\mathcal{A}_k(V, 0)$
11. Solve (3.43) with  $\tau = 0$ ; let  $(x_0, \alpha_0)$  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  $(x_1, \xi_1, \alpha_1)$  be a solution // here  $\xi_1 \neq 0$
16. IF ( $\alpha_1 < 0$ ) //  $(s, -s_{k+1}) \in \mathcal{S}_{a,k+1}(V, -\text{sgn}(\xi_1)\tau)$
17. c\_p\_STREE\_REC( $(s, -s_{k+1}), x_1/|\xi_1|, -\text{sgn}(\xi_1)$ )
18. ENDIF
19. ENDIF
20. RETURN
21. ENDIF // here  $\mathbf{s} \in \{\pm 1\}$ , hence  $s \in \mathcal{S}_{a,k}(V, \mathbf{s}\tau)$
22. Solve (3.43) with  $\tau \curvearrowright \mathbf{s}\tau$ ; let  $(x_2, \alpha_2)$  be a solution
23. IF ( $\alpha_2 < 0$ ) //  $(s, -s_{k+1}) \in \mathcal{S}_{a,k+1}(V, \mathbf{s}\tau)$
24. c\_p\_STREE\_REC( $(s, -s_{k+1}), x_2, \mathbf{s}$ )
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  $\leq 1$ ).

### 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.

---

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	$\mathcal{S}_{k+1}([V; \tau^T], 0)$	$\mathfrak{S}([V; \tau^T], 0)$	$\tilde{\mathfrak{S}}_s \cup \tilde{\mathfrak{S}}_0$
22..25	30..37	$\mathcal{S}_{a,k+1}(V, \mathfrak{s}\tau)$	$\mathfrak{S}(V, \mathfrak{s}\tau)$	$\tilde{\mathfrak{S}}_s \cup (\mathfrak{s}\tilde{\mathfrak{S}}_a)$

---

Table 5.1: Corresponding lines in algorithms 5.6.5 and 5.6.8.

- 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  $(s, -s_{k+1})$  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, [S])`).

It is assumed that  $s \in \{\pm 1\}^k$  and  $(s, x, [S])$  satisfies (5.48).

1. IF  $(v_{k+1}^\top x \simeq [S] \tau_{k+1})$  // two easy children in  $\mathcal{A}_{k+1}(V, [S] \tau)$
2.    `c_PD_STREE_REC` $((s, +1), x + t_+ v_{k+1}, [S])$  for some  $t_+ \in (t_0, t_{\max})$
3.    `c_PD_STREE_REC` $((s, -1), x + t_- v_{k+1}, [S])$  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.  $\tilde{\mathfrak{S}}_s = \emptyset, \tilde{\mathfrak{S}}_a = \emptyset, \tilde{\mathfrak{S}}_0 = \emptyset$  // 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, [S],  $\tilde{\mathfrak{S}}_s, \tilde{\mathfrak{S}}_a, \tilde{\mathfrak{S}}_0$ )`).

It is assumed that  $s \in \{\pm 1\}^k$ , that  $(x, s, [S])$  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` $(s, [S])$
3.    RETURN
4. ENDIF

```
5. IF (C_PD_TWO_CHILDREN(s, x,  $\underline{s}$ ))      // two easy children in  $\mathcal{A}_{k+1}(V, \underline{s}\tau)$ 
6.   RETURN
7. ENDIF
8.  $s_{k+1} := \text{sgn}(v_{k+1}^\top x - \underline{s}\tau_{k+1})$       //  $(s, s_{k+1}) \in \mathcal{S}_{k+1}(V, \underline{s}\tau)$ 
9. C_PD_STREE_REC(( $s, s_{k+1}$ ),  $x, \underline{s}, \tilde{\mathfrak{S}}_s, \tilde{\mathfrak{S}}_a, \tilde{\mathfrak{S}}_0$ )
10. IF ( $\underline{s} = 0$ )      //  $s \in \mathcal{S}_k(V, 0)$  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  $(x_0, \alpha_0)$  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  $\in \mathcal{S}_{k+1}([V; \tau^\top], 0)$ 
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  $(x_1, \xi_1, \alpha_1)$  be a solution      // here  $\xi_1 \neq 0$ 
22.   IF ( $\alpha_1 < 0$ )      //  $(s, -s_{k+1}) \in \mathcal{S}_{a,k+1}(V, -\text{sgn}(\xi_1)\tau)$ 
23.     C_PD_STREE_REC(( $s, -s_{k+1}$ ),  $x_1/|\xi_1|, -\text{sgn}(\xi_1), \tilde{\mathfrak{S}}_s, \tilde{\mathfrak{S}}_a, \tilde{\mathfrak{S}}_0$ )
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  $\underline{s} \in \{\pm 1\}$ , hence  $s \in \mathcal{S}_{a,k}(V, \underline{s}\tau)$ 
30. IF (( $s, -s_{k+1}$ ) does not cover a stem vector of  $\tilde{\mathfrak{S}}_s \cup (\underline{s}\tilde{\mathfrak{S}}_a)$ )
31.   Solve (3.43) with  $\tau \curvearrowright \underline{s}\tau$ ; let  $(x_2, \alpha_2)$  be a solution
32.   IF ( $\alpha_2 < 0$ )
33.     C_PD_STREE_REC(( $s, -s_{k+1}$ ),  $x_2, \underline{s}, \tilde{\mathfrak{S}}_s, \tilde{\mathfrak{S}}_a, \tilde{\mathfrak{S}}_0$ )
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  $r := \text{rank}(V) = n$ . 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  $\tau$  are randomly generated integers in  $[-20, +20]$  [208, 77].
- **PERM-N:** This problem refers to the hyperplane arrangements that are called *permutohedron* 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  $k$ th basis vector of  $\mathbb{R}^n$ ). The vector  $\tau$  is defined by  $\tau_i = 1$  for  $i \in [1 : n]$  and  $\tau_i = 0$  for  $i \in [n+1 : p]$ . Since  $(1, \dots, 1)$  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  $\tau \in [0, 1]$ . Then, the remaining columns of  $[V; \tau^\top]$  can either be random with probability  $1 - \tau$  or randomly generated linear combinations in  $[-4 : 4]$  of the previous vectors. One recovers problem **RAND-N-P** when  $\tau = 0$ .

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		Stem vectors		Chambers
	$ \mathcal{C}(V) $	$ \mathfrak{S}_s(V, \tau) /2$	$ \mathfrak{S}_a(V, \tau) $	$ \mathfrak{S}([V; \tau^\top], 0) /2$	$ \mathcal{S}(V, \tau) $
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  $\text{rank}(V) = n$ .

**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

Problem	$n$	$p$	Circuits	Stem vectors		Stem vectors		Chambers	
			of $V$	of $\mathcal{A}(V, \tau)$	of $\mathcal{A}([V; \tau^T], 0)$	$ \mathfrak{S}_s /2$	$ \mathfrak{S}_a $	$ \mathfrak{S} /2$	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  $\mathcal{A}([V; \tau^T], 0)$  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).

- $|\mathfrak{S}_s(V, \tau)|/2 + |\mathfrak{S}_a(V, \tau)|$  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  $|\mathfrak{S}([V; \tau^T], 0)|/2$  of circuits of  $[V; \tau^T]$  (7th column, see [77, after definition 3.9]) that reaches its maximum (8th column), see remark 5.3.13(6); and by a number  $|\mathcal{S}(V, \tau)|$  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  $\mathcal{A}([V; \tau^T], 0)$  (7th column) is also the number  $|\mathcal{C}([V; \tau^T])|$  of circuits of  $[V; \tau^T]$  (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  $\mathcal{C}(V) \subseteq \mathcal{C}([V; \tau^T])$  nor  $\mathcal{C}([V; \tau^T]) \subseteq \mathcal{C}(V)$  must hold.  $\square$

### 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 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.

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).

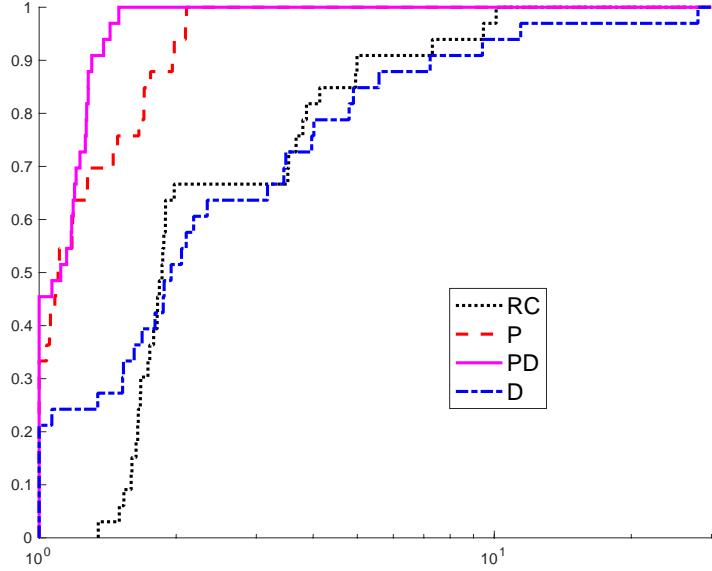


Figure 5.6: Performance profiles of the RC, P, PD and D algorithms, for the computing time.

### 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 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.

## 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

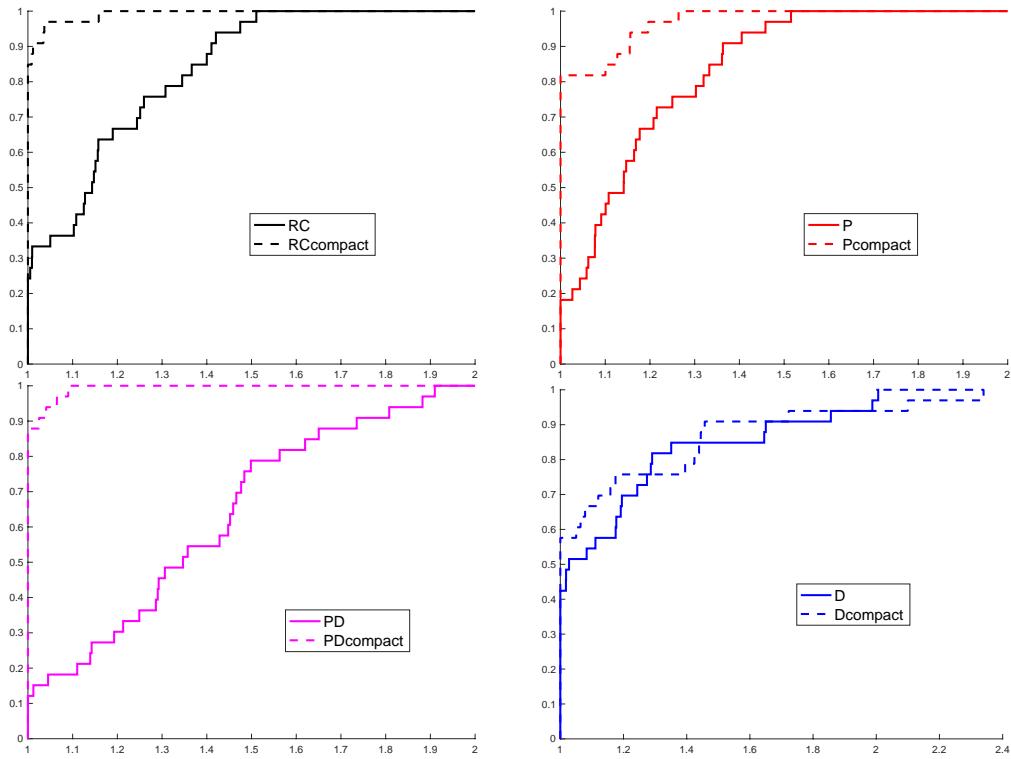


Figure 5.7: Performance profiles of the RC vs RC/C, P vs P/C, PD vs PD/C and D vs D/C algorithms, for the computing time. The dashed lines refer to the compact versions of the algorithms.

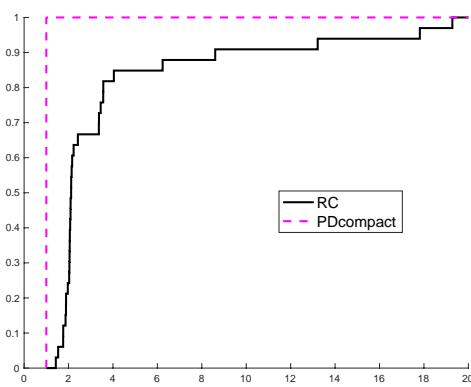


Figure 5.8: Performance profiles of the RC vs PD/C solvers, for the computing time.

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].

Problems	RC	P		PD		D	
	time	time	ratio	time	ratio	time	ratio
RAND-2-8	0.09	0.03	<b>3.76</b>	0.03	<b>3.51</b>	<b>0.02</b>	<b>4.13</b>
RAND-4-8	0.12	0.07	<b>1.71</b>	0.08	<b>1.47</b>	<b>0.06</b>	<b>1.86</b>
RAND-4-9	0.20	0.12	<b>1.67</b>	0.13	<b>1.56</b>	<b>0.10</b>	<b>1.99</b>
RAND-5-10	0.48	0.27	<b>1.79</b>	0.31	<b>1.58</b>	<b>0.26</b>	<b>1.89</b>
RAND-4-11	0.56	0.35	<b>1.58</b>	0.38	<b>1.47</b>	<b>0.30</b>	<b>1.88</b>
RAND-6-12	1.89	1.18	<b>1.59</b>	1.34	<b>1.41</b>	<b>1.09</b>	<b>1.73</b>
RAND-5-13	2.32	1.37	<b>1.69</b>	1.45	<b>1.60</b>	<b>1.30</b>	<b>1.79</b>
RAND-7-14	6.96	<b>4.37</b>	<b>1.59</b>	5.14	<b>1.35</b>	4.66	1.50
RAND-7-15	12.40	<b>7.51</b>	<b>1.65</b>	8.92	<b>1.39</b>	10.10	1.22
RAND-8-16	29.20	<b>17.50</b>	<b>1.67</b>	21.10	<b>1.38</b>	28.30	1.03
RAND-9-17	63.00	<b>39.40</b>	<b>1.60</b>	50.50	<b>1.25</b>	70.80	0.89
SRAND-8-20-2	33.30	9.70	<b>3.43</b>	<b>6.67</b>	<b>5.00</b>	23.00	1.45
SRAND-8-20-4	199.00	<b>105.00</b>	<b>1.90</b>	137.00	<b>1.45</b>	989.00	0.20
SRAND-8-20-6	238.00	<b>131.00</b>	<b>1.81</b>	196.00	<b>1.21</b>	947.00	0.25
2D-4-20	2.12	0.89	<b>2.38</b>	<b>0.60</b>	<b>3.53</b>	1.01	2.11
2D-5-20	3.50	1.58	<b>2.22</b>	<b>0.95</b>	<b>3.67</b>	1.86	1.88
2D-6-20	5.84	2.82	<b>2.07</b>	<b>1.66</b>	<b>3.53</b>	3.11	1.88
2D-7-20	9.86	4.48	<b>2.20</b>	<b>2.55</b>	<b>3.86</b>	5.24	1.88
2D-8-20	16.40	7.35	<b>2.23</b>	<b>4.32</b>	<b>3.80</b>	8.13	2.02
PERM-5	1.17	0.46	<b>2.53</b>	<b>0.24</b>	<b>4.96</b>	0.82	1.42
PERM-6	10.60	3.04	<b>3.50</b>	<b>1.45</b>	<b>7.33</b>	7.11	1.50
PERM-7	106.00	23.60	<b>4.49</b>	<b>11.20</b>	<b>9.50</b>	128.00	0.83
PERM-8	1070.00	210.00	<b>5.10</b>	<b>106.00</b>	<b>10.14</b>	2970.00	0.36
RATIO-3-20-0.7	2.53	1.54	<b>1.65</b>	<b>1.39</b>	<b>1.82</b>	2.12	1.19
RATIO-3-20-0.9	2.59	1.80	<b>1.44</b>	<b>1.41</b>	<b>1.84</b>	2.16	1.20
RATIO-4-20-0.7	11.10	6.17	<b>1.80</b>	<b>5.94</b>	<b>1.87</b>	12.50	0.89
RATIO-4-20-0.9	7.07	5.53	<b>1.28</b>	<b>4.33</b>	<b>1.63</b>	9.46	0.75
RATIO-5-20-0.7	21.00	<b>12.00</b>	<b>1.75</b>	12.80	<b>1.64</b>	38.10	0.55
RATIO-5-20-0.9	16.00	11.30	<b>1.42</b>	<b>9.57</b>	<b>1.67</b>	22.40	0.71
RATIO-6-20-0.7	75.90	<b>46.10</b>	<b>1.65</b>	58.50	<b>1.30</b>	183.00	0.42
RATIO-6-20-0.9	65.40	<b>43.60</b>	<b>1.50</b>	50.10	<b>1.30</b>	175.00	0.37
RATIO-7-20-0.7	147.00	<b>109.00</b>	<b>1.36</b>	151.00	<b>0.98</b>	523.00	0.28
RATIO-7-20-0.9	148.00	<b>96.40</b>	<b>1.54</b>	138.00	<b>1.07</b>	538.00	0.28
Mean			2.11		<b>2.76</b>		1.28
Median			<b>1.71</b>		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			P/C			PD/C			D/C		
	time	ratio	ratio	time	ratio	ratio	time	ratio	ratio	time	ratio	ratio
RAND-2-8	0.08	1.25	<b>1.25</b>	0.03	0.89	<b>3.33</b>	0.03	0.96	<b>3.37</b>	<b>0.03</b>	0.85	<b>3.52</b>
RAND-4-8	0.08	1.37	<b>1.37</b>	0.05	1.33	<b>2.28</b>	0.05	1.45	<b>2.14</b>	<b>0.05</b>	1.24	<b>2.32</b>
RAND-4-9	0.16	1.24	<b>1.24</b>	0.10	1.21	<b>2.02</b>	0.11	1.21	<b>1.89</b>	<b>0.09</b>	1.11	<b>2.20</b>
RAND-5-10	0.35	1.40	<b>1.40</b>	<b>0.22</b>	1.25	<b>2.24</b>	0.24	1.29	<b>2.05</b>	0.22	1.18	2.23
RAND-4-11	0.49	1.15	<b>1.15</b>	0.30	1.16	<b>1.84</b>	0.30	1.29	<b>1.89</b>	<b>0.29</b>	1.03	<b>1.93</b>
RAND-6-12	1.34	1.41	<b>1.41</b>	<b>0.87</b>	1.36	<b>2.18</b>	0.90	1.48	<b>2.09</b>	0.92	1.19	2.07
RAND-5-13	1.95	1.19	<b>1.19</b>	<b>1.20</b>	1.14	<b>1.93</b>	1.11	1.31	<b>2.09</b>	1.20	1.08	1.93
RAND-7-14	4.90	1.42	<b>1.42</b>	<b>3.11</b>	1.41	<b>2.24</b>	3.43	1.50	<b>2.03</b>	3.90	1.19	1.78
RAND-7-15	9.22	1.34	<b>1.34</b>	<b>5.69</b>	1.32	<b>2.18</b>	6.04	1.48	<b>2.05</b>	8.58	1.18	1.45
RAND-8-16	19.80	1.47	<b>1.47</b>	<b>12.00</b>	1.46	<b>2.43</b>	13.50	1.56	<b>2.16</b>	17.20	1.65	1.70
RAND-9-17	41.70	1.51	<b>1.51</b>	<b>26.00</b>	1.52	<b>2.42</b>	30.60	1.65	<b>2.06</b>	42.90	1.65	<b>1.47</b>
SRAND-8-20-2	30.20	1.10	<b>1.10</b>	9.45	1.03	<b>3.52</b>	<b>5.34</b>	1.25	<b>6.24</b>	22.60	1.02	<b>1.47</b>
SRAND-8-20-4	158.00	1.26	<b>1.26</b>	<b>80.60</b>	1.30	<b>2.47</b>	93.90	1.46	<b>2.12</b>	493.00	2.01	0.40
SRAND-8-20-6	182.00	1.31	<b>1.31</b>	<b>96.10</b>	1.36	<b>2.48</b>	121.00	1.62	<b>1.97</b>	701.00	1.35	0.34
2D-4-20	2.10	1.01	<b>1.01</b>	1.03	0.87	<b>2.06</b>	<b>0.62</b>	0.98	<b>3.45</b>	1.74	0.58	1.22
2D-5-20	3.54	0.99	<b>0.99</b>	1.89	0.84	<b>1.85</b>	<b>1.04</b>	0.92	<b>3.37</b>	2.71	0.69	1.29
2D-6-20	5.89	0.99	<b>0.99</b>	3.26	0.87	<b>1.79</b>	<b>1.64</b>	1.01	<b>3.56</b>	4.43	0.70	1.32
2D-7-20	9.81	1.01	<b>1.01</b>	4.93	0.91	<b>2.00</b>	<b>2.44</b>	1.05	<b>4.04</b>	7.31	0.72	1.35
2D-8-20	16.40	1.00	<b>1.00</b>	9.29	0.79	<b>1.77</b>	<b>4.60</b>	0.94	<b>3.57</b>	11.70	0.69	1.40
PERM-5	1.04	1.12	<b>1.12</b>	0.42	1.11	<b>2.80</b>	<b>0.14</b>	1.74	<b>8.60</b>	0.64	1.29	1.83
PERM-6	9.27	1.14	<b>1.14</b>	2.82	1.08	<b>3.76</b>	<b>0.80</b>	1.81	<b>13.22</b>	5.58	1.27	1.90
PERM-7	94.00	1.13	<b>1.13</b>	22.30	1.06	<b>4.75</b>	<b>5.95</b>	1.88	<b>17.82</b>	64.40	1.99	1.65
PERM-8	1070.00	1.00	<b>1.00</b>	195.00	1.08	<b>5.49</b>	<b>55.50</b>	1.91	<b>19.28</b>	1600.00	1.86	0.67
RATIO-3-20-0.7	2.53	1.00	<b>1.00</b>	1.45	1.06	<b>1.74</b>	<b>1.22</b>	1.14	<b>2.07</b>	4.96	0.43	0.51
RATIO-3-20-0.9	2.68	0.97	<b>0.97</b>	1.65	1.09	<b>1.57</b>	<b>1.27</b>	1.11	<b>2.04</b>	3.12	0.69	0.83
RATIO-4-20-0.7	11.00	1.01	<b>1.01</b>	5.73	1.08	<b>1.94</b>	<b>4.98</b>	1.19	<b>2.23</b>	13.30	0.94	0.83
RATIO-4-20-0.9	8.19	0.86	<b>0.86</b>	5.30	1.04	<b>1.33</b>	<b>3.79</b>	1.14	<b>1.87</b>	7.33	1.29	0.96
RATIO-5-20-0.7	20.00	1.05	<b>1.05</b>	10.90	1.10	<b>1.93</b>	<b>9.92</b>	1.29	<b>2.12</b>	80.00	0.48	0.26
RATIO-5-20-0.9	13.90	1.15	<b>1.15</b>	9.67	1.17	<b>1.65</b>	<b>6.61</b>	1.45	<b>2.42</b>	22.00	1.02	0.73
RATIO-6-20-0.7	68.50	1.11	<b>1.11</b>	<b>40.20</b>	1.15	<b>1.89</b>	43.10	1.36	<b>1.76</b>	212.00	0.86	0.36
RATIO-6-20-0.9	67.80	0.96	<b>0.96</b>	38.20	1.14	<b>1.71</b>	<b>37.20</b>	1.35	<b>1.76</b>	196.00	0.89	0.33
RATIO-7-20-0.7	127.00	1.16	<b>1.16</b>	<b>89.70</b>	1.22	<b>1.64</b>	103.00	1.47	<b>1.43</b>	564.00	0.93	0.26
RATIO-7-20-0.9	128.00	1.16	<b>1.16</b>	<b>81.90</b>	1.18	<b>1.81</b>	96.60	1.43	<b>1.53</b>	565.00	0.95	0.26
Mean		1.16	<b>1.16</b>		1.14	<b>2.33</b>		1.35	<b>3.95</b>		1.09	<b>1.30</b>
Median		1.14	<b>1.14</b>		1.14	<b>2.02</b>		1.35	<b>2.12</b>		1.03	<b>1.35</b>

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).

## Funding

The first author's research was partially supported by NSERC grant OGP0005491. The third author was partially supported by a Mitacs-Inria grant.

# Chapter 6

## Levenberg-Marquardt least-squares globalization of the PNM algorithm

After this long and detailed focus on hyperplane arrangements, this chapter details some ongoing work on the initial motivation of the thesis, the globalization of the Polyhedral Newton-Min (PNM) algorithm from [72], to appear in *Mathematical Programming*. It is split into two parts: section 6.1 presents a possible way to modify the PNM algorithm, while section 6.2 discusses some convergence properties of the corresponding modified algorithm.

Let us recall that we consider the following general nonlinear complementarity problem

$$0 \leq F(x) \perp G(x) \geq 0, \quad (6.1)$$

where  $F$  and  $G$  are smooth functions from  $\mathbb{R}^n$  to  $\mathbb{R}^n$ , and possibly its affine version

$$0 \leq Ax + a \perp Bx + b \geq 0.$$

Using the reformulation via the minimum C-function, this leads to the nonsmooth system and the minimization of its associated merit function

$$H(x) := \min(F(x), G(x)) = 0 \quad \text{and} \quad \min \theta(x) := \frac{1}{2} \|H(x)\|^2. \quad (6.2)$$

The following index sets play a particularly relevant role:

$$\begin{aligned} \mathcal{E}(x) &:= \{i \in [1 : n] : F_i(x) = G_i(x)\}, \\ \mathcal{F}(x) &:= \{i \in [1 : n] : F_i(x) < G_i(x)\}, \\ \mathcal{G}(x) &:= \{i \in [1 : n] : F_i(x) > G_i(x)\}. \end{aligned} \quad (6.3)$$

The resulting system and minimization problem (6.2) are piecewise smooth. As it was mentioned in section 2.3.3 about the Newton-min algorithm, at an iterate one may face the question of “which piece to choose?”. Choosing an appropriate piece and obtaining a descent direction is not necessarily easy and, as we shall see, even detecting whether the iterate is

stationary is in general co-NP-complete. This is shown by a geometric reformulation partly related to chapter 3.

Nonetheless, under assumptions that make this difficulty tractable (injectivity of an involved submatrix for instance), an algorithm based on a Levenberg-Marquardt curve tangent to an element of  $\partial\theta(x)$  at each iteration is proposed, which benefits from the usual properties of Levenberg-Marquardt algorithms.

## 6.1 Modifying the Polyhedral Newton-Min algorithm

### 6.1.1 Presentation of the method

Let us first recall the Newton-min algorithm. It solves the nonsmooth equation  $H(x) = 0$  by solving a sequence of Newton-type steps where the indices in  $\mathcal{E}(x)$ , which make the system nonsmooth, are arbitrarily assigned to  $\mathcal{F}(x)$  or  $\mathcal{G}(x)$ .

**Algorithm 6.1.1** (NEWTON-MIN). Consider a starting point  $x^0 \in \mathbb{R}^n$ .

1. *Stopping criterion.* If  $H(x^k) = 0$ , stop.
2. *Index decomposition.* Define the following index sets:

$$\mathcal{E}(x^k), \quad \mathcal{F}(x^k), \quad \mathcal{G}(x^k).$$

Let  $\tilde{\mathcal{F}}(x^k), \tilde{\mathcal{G}}(x^k)$  be a partition of  $[1 : n]$  such that  $\tilde{\mathcal{F}}(x^k) \supseteq \mathcal{F}(x^k)$  and  $\tilde{\mathcal{G}}(x^k) \supseteq \mathcal{G}(x^k)$ . Solve the linear system for the variable  $d$ .

$$\begin{cases} F(x^k)_{\tilde{\mathcal{F}}(x^k)} + F'(x^k)_{\tilde{\mathcal{F}}(x^k)}d = 0, \\ G(x^k)_{\tilde{\mathcal{G}}(x^k)} + G'(x^k)_{\tilde{\mathcal{G}}(x^k)}d = 0. \end{cases} \quad (6.4)$$

3. *Update.* Set  $x^{k+1} = x^k + d^k$  where  $d^k$  is a solution of (6.4),

The algorithm assumes that the system (6.4) has a solution. This is guaranteed near a point satisfying a certain regularity solution.

As presented in section 2.3.4, a variant of this algorithm solves a system involving inequalities instead of only equalities, to ensure that the obtained direction is a descent direction for the merit function  $\theta$  [72]. Define the following partition of  $\mathcal{E}(x)$ :

$$\begin{aligned} \mathcal{E}^{0+}(x) &:= \{i \in \mathcal{E}(x) : F_i(x) = G_i(x) \geq 0\}, \\ \mathcal{E}^-(x) &:= \{i \in \mathcal{E}(x) : F_i(x) = G_i(x) < 0\}. \end{aligned} \quad (6.5)$$

Then, consider a bipartition of  $\mathcal{E}^{0+}(x)$  into  $\mathcal{E}_{\mathcal{F}}^{0+}(x) \cup \mathcal{E}_{\mathcal{G}}^{0+}(x)$ . The proposed idea is to keep one equality for indices of  $\mathcal{E}^{0+}(x)$  and to introduce two inequalities for those in  $\mathcal{E}^-(x)$ .

Thus, one solves

$$\begin{cases} F_i(x) + F'_i(x)d = 0 & \text{if } i \in \mathcal{F}(x) \cup \mathcal{E}_\mathcal{F}^{0+}(x), \\ G_i(x) + G'_i(x)d = 0 & \text{if } i \in \mathcal{G}(x) \cup \mathcal{E}_\mathcal{G}^{0+}(x), \\ F_i(x) + F'_i(x)d \geq 0 & \text{if } i \in \mathcal{E}^-(x), \\ G_i(x) + G'_i(x)d \geq 0 & \text{if } i \in \mathcal{E}^-(x). \end{cases} \quad (6.6)$$

Observe that this system has  $n - |\mathcal{E}^-(x)|$  equalities and  $2|\mathcal{E}^-(x)|$  inequalities. We show that, for a direction  $d$  solving this system, one has  $\theta'(x; d) \leq -2\theta(x)$ ; recall that, for the Newton method on a smooth  $H$ , one has  $\nabla\theta(x)^\top d = -2\theta(x)$ . Indeed, using that  $\theta'(x; d) = H(x)^\top H'(x; d)$ , one gets<sup>1</sup>

$$\begin{aligned} \theta'(x; d) &= \sum_{i \in \mathcal{F}(x)} F_i(x)F'_i(x)d + \sum_{i \in \mathcal{G}(x)} G_i(x)G'_i(x)d + \sum_{i \in \mathcal{E}(x)} H_i(x) \min(F'_i(x)d, G'_i(x)d) \\ &= -\|F_{\mathcal{F}(x)}(x)\|^2 - \|G_{\mathcal{G}(x)}(x)\|^2 - \|H_{\mathcal{E}(x)}(x)\|^2 \\ &\quad + H_{\mathcal{E}(x)}(x)^\top \min(F_{\mathcal{E}(x)}(x) + F'_{\mathcal{E}(x)}(x)d, G_{\mathcal{E}(x)}(x) + G'_{\mathcal{E}(x)}(x)d) \\ &= -2\theta(x) + H_{\mathcal{E}(x)}(x)^\top \min(F_{\mathcal{E}(x)}(x) + F'_{\mathcal{E}(x)}(x)d, G_{\mathcal{E}(x)}(x) + G'_{\mathcal{E}(x)}(x)d). \end{aligned}$$

Let us examine the last line term by term. For  $i \in \mathcal{E}^{0+}(x)$ , either  $F_{\mathcal{E}(x)}(x) + F'_{\mathcal{E}(x)}(x)d = 0$  or  $G_{\mathcal{E}(x)}(x) + G'_{\mathcal{E}(x)}(x)d = 0$ , implying the minimum of these quantities is  $\leq 0$ . Then, the multiplication by  $H_i(x) \geq 0$  implies the product remains nonpositive. For  $i \in \mathcal{E}^-(x)$ , both arguments of the minimum are  $\geq 0$ , so their minimum is  $\geq 0$ ; now, it is multiplied by a negative term. In the end, all involved quantities are nonpositive, implying  $\theta'(x; d) \leq -2\theta(x)$ .

Let us observe that  $\theta'(x; d)$  is convex (in  $d$ ) when  $\mathcal{E}^{0+}(x) = \emptyset$  and concave when  $\mathcal{E}^-(x) = \emptyset$  (in  $d$ ). Indeed, using that the maximum / minimum of linear functions is convex / concave:<sup>2</sup>

$$\begin{aligned} \text{for } i \in \mathcal{E}^-(x) : \quad H_i(x) \min(F'_i(x)d, G'_i(x)d) &= \max(H_i(x)F'_i(x)d, H_i(x)G'_i(x)d), \\ \text{for } i \in \mathcal{E}^{0+}(x) : \quad H_i(x) \min(F'_i(x)d, G'_i(x)d) &= \min(H_i(x)F'_i(x)d, H_i(x)G'_i(x)d). \end{aligned}$$

In [72], other variants of the system (6.6) are also considered, in particular extending the set  $\mathcal{E}^-(x)$  to  $\mathcal{E}_\tau^-$  for some  $\tau > 0$ , the indices such that

$$\mathcal{E}_\tau^- := \{i \in [1 : n] : F_i(x) < 0, G_i(x) < 0, |F_i(x) - G_i(x)| < \tau\}.$$

This tolerance around the “negative kinks” ( $\mathcal{E}^-(x)$ , in the limit  $\tau \rightarrow 0$ ) is also highly relevant from a computational point of view – though this could concern, outside the PNM algorithm, the “nonnegative kinks” as well; we return to this later in the chapter. Here is a simple version of the polyhedral algorithm.

---

<sup>1</sup>Use, for  $i \in \mathcal{E}(x)$ ,  $F_i(x) = H_i(x) = G_i(x)$  and  $H_i(x) \min(F'_i(x)d, G'_i(x)d) = -H_i(x)^2 + H_i(x) \min(F_i(x) + F'_i(x)d, G_i(x) + G'_i(x)d)$ .

<sup>2</sup>Technically, there can be indices for which  $F_i(x) = 0 = G_i(x)$  (which were arbitrarily put with those such that  $F_i(x) = G_i(x) > 0$ ).

**Algorithm 6.1.2** (Algorithm PNM [72]). Let  $\tau \in (0, \infty]$  be the kink tolerance constant,  $\omega \in (0, 1/2)$  and  $\beta \in (0, 1)$  be the two constants used in the linesearch of step 4 below. The next iterate  $x_+$  is computed from iterate  $x$  as follows.

1. *Stopping criterion.* If  $\theta(x) = 0$ , stop ( $x$  is a solution).
2. *Index sets.* Choose some partition  $(\mathcal{E}_{\mathcal{F}}^{0+}(x), \mathcal{E}_{\mathcal{G}}^{0+}(x))$  of  $\mathcal{E}^{0+}(x)$  and compute the index sets  $\mathcal{E}_{\tau}^{-}, \mathcal{F}(x) \setminus \mathcal{E}_{\tau}^{-}$  and  $\mathcal{G}(x) \setminus \mathcal{E}_{\tau}^{-}$ .
3. *Direction.* Compute a direction  $d$  as a solution to the following problem

$$\min\{||d|| : d \text{ satisfies (6.6)}\} \quad (6.7)$$

by using the index sets of point 2.

4. *Stepsize.* Set  $\alpha := \beta^i$  where  $i$  is the smallest nonnegative integer such that

$$\theta(x + \alpha d) \leqslant (1 - 2\omega\alpha)\theta(x). \quad (6.8)$$

5. *New iterate.* Set  $x_+ = x + \alpha d$ .

### 6.1.2 Levenberg-Marquardt least-squares variant

The main issue with the system (6.6) is that the defined polyhedron may be empty. In [72], this emptiness is avoided by technical regularity assumptions. Let us slightly rewrite (6.6) with the variables  $\gamma_i \in \{0, 1\}$  for  $i \in \mathcal{E}^{0+}(x)$  and  $\bar{\gamma}_i := 1 - \gamma_i$ , to have  $\gamma_i = +1 \Leftrightarrow i \in \mathcal{E}_{\mathcal{F}}^{0+}(x)$  and  $\gamma_i = 0 \Leftrightarrow i \in \mathcal{E}_{\mathcal{G}}^{0+}(x)$ :

$$\left\{ \begin{array}{ll} F_i(x) + F'_i(x)d = 0 & \text{if } i \in \mathcal{F}(x), \\ G_i(x) + G'_i(x)d = 0 & \text{if } i \in \mathcal{G}(x), \\ \gamma_i(F_i(x) + F'_i(x)d) + \bar{\gamma}_i(G_i(x) + G'_i(x)d) = 0 & \text{if } i \in \mathcal{E}^{0+}(x), \\ F_i(x) + F'_i(x)d \geqslant 0 & \text{if } i \in \mathcal{E}^{-}(x), \\ G_i(x) + G'_i(x)d \geqslant 0 & \text{if } i \in \mathcal{E}^{-}(x). \end{array} \right.$$

Furthermore, observe that the indices of  $\mathcal{E}^{-}(x)$  are duplicated: in a least-squares formulation, they may get additional weight compared to the other indices. We have thus tried to introduce a convex ponderation on the linearizations of  $F$  and  $G$  for these indices, so that all indices have total weight 1. This modelling shall play a specific role later. Let

$$\Gamma = \text{Diag}(\gamma) \in \mathbb{R}^{\mathcal{E}(x) \times \mathcal{E}(x)}, \quad \bar{\Gamma} = I - \text{Diag}(\Gamma) \in \mathbb{R}^{\mathcal{E}(x) \times \mathcal{E}(x)}.$$

---

<sup>3</sup>

<sup>3</sup>Below, we also denote  $\Gamma_{\mathcal{E}^{0+}(x)} := \Gamma_{\mathcal{E}^{0+}(x), \mathcal{E}^{0+}(x)}$  and  $\Gamma_{\mathcal{E}^{-}(x)} := \Gamma_{\mathcal{E}^{-}(x), \mathcal{E}^{-}(x)}$ , which is a slight notational abuse for a diagonal matrix.

Thus, the weighted least-squares version of (6.6) becomes

$$\min_{d \in \mathbb{R}^n} \frac{1}{2} \|w(x, d)\|_2^2, \quad \text{where } w(x, d) := \begin{bmatrix} F_{\mathcal{F}(x)}(x) + F'_{\mathcal{F}(x)}(x)d \\ G_{\mathcal{G}(x)}(x) + G'_{\mathcal{G}(x)}(x)d \\ (\Gamma_{\mathcal{E}^{0+}(x)})^{1/2}[F_{\mathcal{E}^{0+}(x)}(x) + F'_{\mathcal{E}^{0+}(x)}(x)d] \\ (\bar{\Gamma}_{\mathcal{E}^{0+}(x)})^{1/2}[G_{\mathcal{E}^{0+}(x)}(x) + G'_{\mathcal{E}^{0+}(x)}(x)d] \\ -(\Gamma_{\mathcal{E}^-(x)})^{1/2}[F_{\mathcal{E}^-(x)}(x) + F'_{\mathcal{E}^-(x)}(x)d]^- \\ -(\bar{\Gamma}_{\mathcal{E}^-(x)})^{1/2}[G_{\mathcal{E}^-(x)}(x) + G'_{\mathcal{E}^-(x)}(x)d]^- \end{bmatrix}, \quad (6.9)$$

which deserves some comments. First,  $\|w\|^2$  is smooth (but clearly not  $w$  in general) since  $((\cdot)^-)^2$  is. The weights  $\Gamma$  and  $\bar{\Gamma}$  apply to  $\|w\|^2$ , which is why they appear with square roots in the expression of  $w$ . For indices in  $\mathcal{E}^-(x)$ , the minus sign and the  $[\cdot]^-$  come from the inequalities in the previous system, where only negative quantities in  $\mathcal{E}^-(x)$  must be penalized.

For some  $\lambda \in \mathbb{R}_+$  and a positive definite matrix  $S$  (often,  $S = I$ ), a Levenberg-Marquardt variant reads

$$\min_{d \in \mathbb{R}^n} \left( \varphi_x(d) := \frac{1}{2} (\|w(x, d)\|_2^2 + \lambda d^\top S d) \right). \quad (6.10)$$

For some given (iterate)  $x$ , the parameter  $\lambda$  defines a certain curve of solutions  $d(\lambda)$  ( $S$  is fixed at least during an iteration) and varies to obtain suitable descent properties. One important aspect, which is what we discuss from now on, is the *tangency* between  $\theta$  and  $\varphi_x$ , i.e., how do the model  $\varphi_x$  and the real function  $\theta$  fit at first order. This tangency is modelled by the gradient of the model  $\varphi_x$  at  $d = 0$ .<sup>4</sup>

This closeness is heavily related to the values of  $\gamma$ , which correspond to which piece is considered. Observe that the splitting of indices in algorithm 6.1.1 or the partition of  $\mathcal{E}^{0+}(x)$  into  $\mathcal{E}_{\mathcal{F}}^{0+}(x)$  and  $\mathcal{E}_{\mathcal{G}}^{0+}(x)$  in (6.6) are precursory aspects of the convex weights  $\gamma$ .

The gradient of  $\varphi_x$  in  $d = 0$  is denoted below by  $g := g(\gamma) = g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)})$  (to avoid heavier notation, we do not always specify  $g$ 's dependency on  $\gamma$ ). It plays a central role and its expression is given below. In this expression,  $x$  is fixed and, for simplicity of the notation,  $\gamma$  is taken in  $[0, 1]^{\mathcal{E}(x)}$  instead of  $\{0, 1\}^{\mathcal{E}^{0+}(x)} \times [0, 1]^{\mathcal{E}^-(x)}$ . We shall see below that the values  $\{0, 1\}^{\mathcal{E}^{0+}(x)}$  do, however, play a preeminent role among  $[0, 1]^{\mathcal{E}^{0+}(x)}$ .

$$\begin{aligned} g &:= F'_{\mathcal{F}(x)}(x)^\top F_{\mathcal{F}(x)}(x) + G'_{\mathcal{G}(x)}(x)^\top G_{\mathcal{G}(x)}(x) + [F'_{\mathcal{E}(x)}(x)^\top \Gamma + G'_{\mathcal{E}(x)}(x)^\top \bar{\Gamma}] H_{\mathcal{E}(x)}(x) \\ &= g_0(x) + [F'_{\mathcal{E}^{0+}(x)}(x) - G'_{\mathcal{E}^{0+}(x)}(x)]^\top \text{Diag}(H_{\mathcal{E}^{0+}(x)}(x)) \gamma_{\mathcal{E}^{0+}(x)} \\ &\quad + [F'_{\mathcal{E}^-(x)}(x) - G'_{\mathcal{E}^-(x)}(x)]^\top \text{Diag}(H_{\mathcal{E}^-(x)}(x)) \gamma_{\mathcal{E}^-(x)} \\ &= g_0(x) + \mathcal{M}_+ \gamma_{\mathcal{E}^{0+}(x)} + \mathcal{M}_- \gamma_{\mathcal{E}^-(x)} \end{aligned} \quad (6.11)$$

---

<sup>4</sup>For smooth least-squares, solving the normal equation always gives a descent direction unless the iterate is stationary.

with the intermediary variables

$$\begin{aligned} g_0(x) &:= F'_{\mathcal{F}(x)}(x)^T F_{\mathcal{F}(x)}(x) + G'_{\mathcal{G}(x)}(x)^T G_{\mathcal{G}(x)}(x) + G'_{\mathcal{E}(x)}(x)^T G_{\mathcal{E}(x)}(x) \\ \mathcal{M}_+ &:= (F'_{\mathcal{E}^{0+}(x)}(x) - G'_{\mathcal{E}^{0+}(x)}(x))^T \text{Diag}(H_{\mathcal{E}^{0+}(x)}(x)) \\ \mathcal{M}_- &:= (F'_{\mathcal{E}^-(x)}(x) - G'_{\mathcal{E}^-(x)}(x))^T \text{Diag}(H_{\mathcal{E}^-(x)}(x)) \end{aligned} \quad (6.12)$$

In fact, this expression of  $g$  corresponds, since the weights  $\gamma$  are arbitrary, to elements in  $H(x)^T \partial_{\times} H(x)$ . Indeed, recall that (see (3.10))

$$\begin{aligned} J_0 &= [F'_{\mathcal{F}(x)}(x); G'_{\mathcal{G}(x)}(x)], \\ \partial_B H(x) &\subseteq \partial_B^{\times} H(x) := \{J \in \mathbb{R}^{n \times n} : J_{\mathcal{F}(x) \cup \mathcal{G}(x),:} = J_0, J_{i,:} \in \{F'_i(x), G'_i(x)\}, i \in \mathcal{E}(x)\}, \\ \partial_C H(x) &\subseteq \text{conv}(\partial_B^{\times} H(x)) := \partial_{\times} H(x), \\ \partial_{\times} H(x) &= \{J \in \mathbb{R}^{n \times n} : J_{\mathcal{F}(x) \cup \mathcal{G}(x),:} = J_0, J_{i,:} = \gamma_i F'_i(x) + \bar{\gamma}_i G'_i(x), i \in \mathcal{E}(x), \gamma_i \in [0, 1]\}. \end{aligned}$$

Let us explain these relations, focusing on the indices of  $\mathcal{E}(x)$ :

- $\partial_B H(x)$  corresponds to *some* values  $\gamma \in \{0, 1\}^{\mathcal{E}(x)}$  (see chapters 3 and section 4.3);
- $\partial_B^{\times} H$  corresponds to *all* values of  $\gamma \in \{0, 1\}^{\mathcal{E}(x)}$  (it is the superset of the B-differential of  $H$ );
- $\partial_C H(x)$  corresponds to *some* values of  $\gamma \in [0, 1]^{\mathcal{E}(x)}$ ;
- $\partial_{\times} H(x)$  corresponds to *all* values of  $\gamma \in [0, 1]^{\mathcal{E}(x)}$ , which is what is considered with arbitrary values of  $\gamma$ .

In particular,  $\partial_C H(x)$  is a polytope (convex hull of a finite set). Then, one gets

$$\begin{aligned} \partial_{\times} H(x)^T H(x) &= F'_{\mathcal{F}(x)}(x)^T F_{\mathcal{F}(x)}(x) + G'_{\mathcal{G}(x)}(x)^T G_{\mathcal{G}(x)}(x) \\ &\quad + \{[\Gamma F'_{\mathcal{E}(x)}(x) + \bar{\Gamma} G'_{\mathcal{E}(x)}(x)] : \gamma \in [0, 1]^{\mathcal{E}(x)}\}^T H_{\mathcal{E}(x)}(x). \end{aligned}$$

Now that the role of the weights  $\gamma$  is partly explained, let us discuss the relevance of their values. It is mainly conveyed by the quantity  $\theta'(x; -g)$ , which ideally should be negative: this means the opposite of the gradient of the quadratic model  $\varphi_x$  is a descent direction for  $\theta$  as well.<sup>5</sup>

For arbitrary  $\gamma$ , there is no guarantee that  $\theta'(x; -g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)})) < 0$ , since  $\varphi_x$  and  $\theta$  may be too different. This issue is due to nonsmoothness – similar phenomenon may occur in the Newton-min algorithm 2.3.29 when the indices are not properly split, see [20,

---

<sup>5</sup>For simplicity, we present with  $S = I$ ; otherwise, one would have to look at  $\theta'(x; -S^{-1}g)$ .

example 5.8] for instance. We now express  $\theta'(x, -g)$  to further analyze its properties:

$$\begin{aligned}
 \theta'(x; -g) &= F_{\mathcal{F}(x)}(x)^\top F'_{\mathcal{F}(x)}(x)(-g) + G_{\mathcal{G}(x)}(x)^\top G'_{\mathcal{G}(x)}(x)(-g) \\
 &\quad + H_{\mathcal{E}(x)}(x)^\top \min(F'_{\mathcal{E}(x)}(x)(-g), G'_{\mathcal{E}(x)}(x)(-g)) \\
 &= [F_{\mathcal{F}(x)}(x)^\top F'_{\mathcal{F}(x)}(x) + G_{\mathcal{G}(x)}(x)^\top G'_{\mathcal{G}(x)}(x) \\
 &\quad \pm H_{\mathcal{E}(x)}(x)^\top (\Gamma F'_{\mathcal{E}(x)}(x) + \bar{\Gamma} G'_{\mathcal{E}(x)}(x))](-g) \\
 &\quad + H_{\mathcal{E}(x)}(x)^\top \min[F'_{\mathcal{E}(x)}(x)(-g), G'_{\mathcal{E}(x)}(x)(-g)] \\
 &= -\|g\|^2 - [\Gamma F'_{\mathcal{E}(x)}(x) + \bar{\Gamma} G'_{\mathcal{E}(x)}(x)](-g) \\
 &\quad + H_{\mathcal{E}(x)}(x)^\top \min[F'_{\mathcal{E}(x)}(x)(-g), G'_{\mathcal{E}(x)}(x)(-g)] \\
 &= -\|g\|^2 + H_{\mathcal{E}(x)}(x)^\top [\min(F'_{\mathcal{E}(x)}(x)(-g), G'_{\mathcal{E}(x)}(x)(-g)) \\
 &\quad - (\Gamma F'_{\mathcal{E}(x)}(x) + \bar{\Gamma} G'_{\mathcal{E}(x)}(x))(-g)]
 \end{aligned} \tag{6.13}$$

In the last expression, observe the term  $-\|g\|^2$  is similar to the smooth setting, whereas the remaining term corresponds to the nonsmoothness. If  $\mathcal{E}(x) = \emptyset$ , i.e., the system is smooth at  $x$ , this complicated term does not intervene. We split this last term into the  $\mathcal{E}^{0+}(x)$  and the  $\mathcal{E}^-(x)$  parts.

$$\begin{aligned}
 \theta'(x; -g) &= -\|g\|^2 + H_{\mathcal{E}^{0+}(x)}(x)^\top [\min(F'_{\mathcal{E}^{0+}(x)}(x)(-g), G'_{\mathcal{E}^{0+}(x)}(x)(-g)) \\
 &\quad - (\Gamma_{\mathcal{E}^{0+}(x)} F'_{\mathcal{E}^{0+}(x)}(x) + \bar{\Gamma}_{\mathcal{E}^{0+}(x)} G'_{\mathcal{E}^{0+}(x)}(x))(-g)] \\
 &\quad + H_{\mathcal{E}^-(x)}(x)^\top [\min(F'_{\mathcal{E}^-}(x)(-g), G'_{\mathcal{E}^-}(x)(-g)) \\
 &\quad - (\Gamma_{\mathcal{E}^-} F'_{\mathcal{E}^-}(x) + \bar{\Gamma}_{\mathcal{E}^-} G'_{\mathcal{E}^-}(x))(-g)]
 \end{aligned} \tag{6.14}$$

Then, observe the components of the terms into brackets are of the form  $\min(a, b) - \gamma a - \bar{\gamma} b$ , i.e., a minimum of two values minus a convex combination of these values. Thus, they are clearly nonpositive<sup>6</sup>. Then, the one with indices in  $\mathcal{E}^{0+}(x)$  is multiplied by  $H_{\mathcal{E}^{0+}(x)} \geq 0$ , lines 1-2 are only comprised of nonpositive terms, whereas the remaining terms (indices in  $\mathcal{E}^-(x)$ , lines 3-4) are nonnegative. Let us show that this nonnegative term can make  $\theta'(x; -g) \geq 0$ .

**Example 6.1.3** (ascent along  $-g$  (1)). Let  $\delta \in \mathbb{R}_*^+$  and  $n = 2$ , and define

$$\begin{aligned}
 F_1(x) &= x_1 - 2, & G_1(x) &= -2x_1 + 1, \\
 F_2(x) &= x_2 - 1 - \delta, & G_2(x) &= 2x_2 - 2 - \delta,
 \end{aligned} \tag{6.15}$$

and let  $x = (1, 1)$ . Clearly, one has

$$F_1(x) = -1 = G_1(x), \quad F_2(x) = -\delta = G_2(x), \quad \mathcal{E}^-(x) = \{1, 2\}.$$

With the convex weights framework,  $g(\gamma)$  reads

$$\begin{aligned}
 g(\gamma) &= (\gamma_1 \nabla F_1(x) + \bar{\gamma}_1 \nabla G_1(x))H_1(x) + (\gamma_2 \nabla F_2(x) + \bar{\gamma}_2 \nabla G_2(x))H_2(x) \\
 &= (\gamma_1 - 2\bar{\gamma}_1) \times (-1)e_1 + (\gamma_2 + 2\bar{\gamma}_2) \times (-\delta)e_2 \\
 &= (2\bar{\gamma}_1 - \gamma_1)e_1 - \delta(\gamma_2 + 2\bar{\gamma}_2)e_2.
 \end{aligned}$$

<sup>6</sup>Indeed,  $\min(a, b) - \gamma a - \bar{\gamma} b = \min(a - \gamma a - \bar{\gamma} b, b - \gamma a - \bar{\gamma} b) = \min(\bar{\gamma}(a - b), \gamma(b - a)) \leq 0$ .

Now, consider  $\gamma = (1/2, 1/2)$ , i.e.,  $g(1/2, 1/2) = (e_1 - 3\delta e_2)/2$ , one has

$$\begin{aligned}\theta'(x; -g) &= (-1) \min(e_1^\top(-g), -2e_1^\top(-g)) + (-\delta) \min(e_2^\top(-g), 2e_2^\top(-g)) \\ &= -\min(-1/2, 1) - \delta \min(3\delta/2, 6\delta/2) \\ &= \frac{1}{2} - \frac{3}{2}\delta^2\end{aligned}$$

which is positive for  $\delta$  small enough. This is illustrated in figure 6.1.

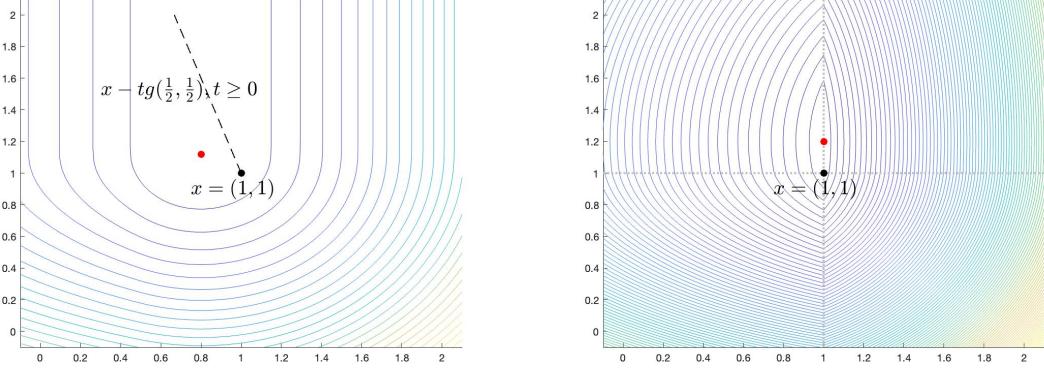


Figure 6.1: Left: level sets of  $\varphi_x$  with the choice of  $\gamma = (1/2, 1/2)$ . Right: level sets of  $\theta$ ; the dotted lines are the kinks ( $\theta$  is not smooth). The red dot indicates a local minimum. The level sets reveal too much difference between  $\theta$  and  $\varphi_x$ , so the direction given by  $\varphi_x$  increases  $\theta$ .

Consider  $\gamma_{\mathcal{E}^-(x)} = (2/3, 1)$  (that will be explained later). One gets, see figure 6.2,

$$\begin{aligned}g(\gamma) &= g(2/3, 1) = [2(1 - 2/3) - 2/3]e_1 - \delta[1 + 2(1 - 1)]e_2 = -\delta e_2 \\ \theta'(x; -g) &= (-1) \min(-e_1^\top g, +2e_1^\top g) - \delta \min(-e_2^\top g, -2e_2^\top g) = -\delta \min(\delta_2, 2\delta_2) = -\delta^2\end{aligned}$$

which confirms the decrease of  $\theta$  with these weights.  $\square$

This phenomenon may also be observed for LCP( $M, q$ ) with a **P**-matrix.

**Example 6.1.4** (ascent along  $-g$  (2)). Consider the following LCP data  $(P, q)$ , with a **P**-matrix  $P$ , and the point  $\hat{x}$ :

$$P = \begin{pmatrix} 1/2 & 1/2 \\ -5 & 1 \end{pmatrix}, \quad q = \begin{pmatrix} 0 \\ -1/10 \end{pmatrix} \quad \text{and} \quad \hat{x} = \begin{pmatrix} -1/50 \\ -1/50 \end{pmatrix}.$$

At  $\hat{x}$ ,  $\mathcal{E}(\hat{x}) = \{1, 2\}$ ; the green oblique arrow is the gradient  $g(\gamma)$  for  $\gamma = (0, 0)$  (its opposite), which is therefore the gradient at  $\hat{x}$  of  $\theta$  in the south-east part of the figure, where it reads  $x \mapsto \frac{1}{2}\|Px + q\|_2^2$ . Hence

$$g = P^\top(P\hat{x} + q) = P^\top\hat{x} = \begin{pmatrix} 9/100 \\ -3/100 \end{pmatrix}.$$

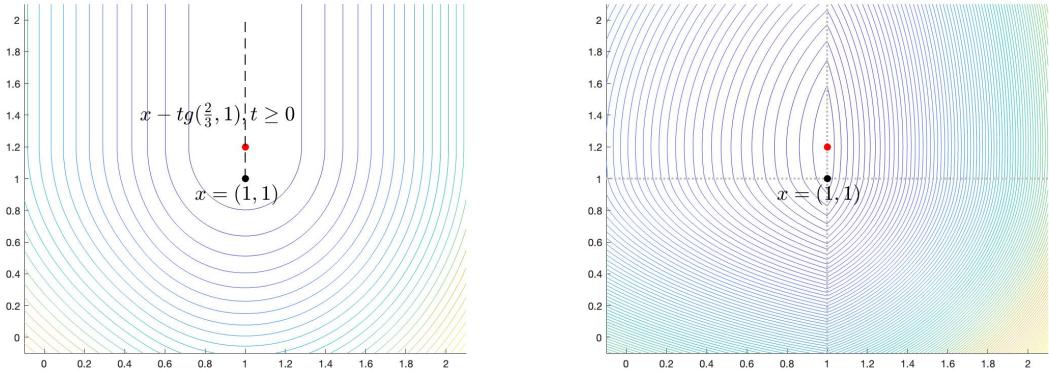


Figure 6.2: Left: level sets of  $\varphi_x$  with the choice of  $\gamma = (2/3, 1)$ . Right: level sets of  $\theta$ ; the dotted lines are the kinks ( $\theta$  is not smooth). The red dot indicates a local minimum. The level sets of  $\varphi_x$  are (at least locally) close enough to those of  $\theta$  so a descent direction of  $\varphi_x$  decreases  $\theta$ .

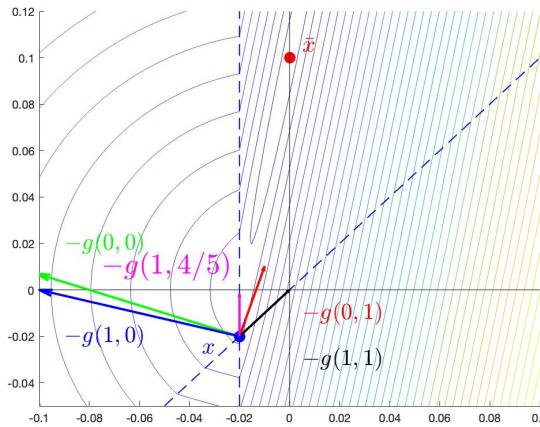


Figure 6.3: Illustration of example 6.1.4. The level curves of  $\theta$  ( $\sqrt{\theta}$  for visibility) are drawn in color, the blue dashed lines are the kinks of nondifferentiability of  $H$  (they are defined by  $x_i = (Px + q)_i$  for  $i \in \{1, 2\}$ ), the red point above is the unique solution  $\bar{x} = (0, 1/10)$  to the problem and the blue point is the current point. The arrows in green, blue, red and black correspond to four possible  $-g$  for extremal choices of  $\gamma$ , the one in magenta to a descent direction.

The picture clearly shows that  $\theta$  increases at  $\hat{x}$  along  $-g$  and this is confirmed by calculation: since for  $t > 0$ ,  $\hat{x} - tg \leq P(\hat{x} - tg) + q$  and it follows that

$$\theta(\hat{x} - tg) = \frac{1}{2} \|\hat{x} - tg\|_2^2 = \frac{1}{2} \|\hat{x}\|_2^2 - t\hat{x}^\top g + \frac{t^2}{2} \|g\|_2^2 > \theta(\hat{x}),$$

since  $-\hat{x}^\top g = 6/5000 > 0$ .

Nevertheless, for  $\gamma = (1, 4/5)$ ,  $\theta$  decreases. Indeed, the value of  $g(1, 4/5)$  is

$$g = \Gamma \hat{x} + P^T \bar{\Gamma} \hat{x} = \begin{pmatrix} -1/50 \\ -4/250 \end{pmatrix} + \begin{pmatrix} 1/2 & -5 \\ 1/2 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ -1/250 \end{pmatrix} = \begin{pmatrix} 0 \\ -1/50 \end{pmatrix}.$$

This one is represented by the magenta vertical arrow in the picture. We observe that, with this last  $g$ ,  $-g$  is a descent direction of  $\theta$  at  $\hat{x}$ , since it forces the decrease of  $\theta$ , which reads  $x \mapsto \frac{1}{2}\|x\|_2^2$  along that direction.  $\square$

**Remark 6.1.5** (interaction between  $\gamma_{\mathcal{E}^-(x)}$  and  $\gamma_{\mathcal{E}^{0+}(x)}$ ). We have seen that the indices in  $\mathcal{E}^-(x)$  have a different impact than those of  $\mathcal{E}^{0+}(x)$ . However, in (6.14), the  $\mathcal{E}^-(x)$  part is quadratic in  $\gamma$  (recall that  $x$  is fixed). Indeed, by (6.11) and (6.12),  $g$  is affine in  $\gamma$  thus the quantities inside brackets are quadratic in  $\gamma$ . Thus, under this form, there is no immediate value of  $\gamma_{\mathcal{E}^-(x)}$  and  $\gamma_{\mathcal{E}^{0+}(x)}$  that cancels this term. Observe that if  $\mathcal{E}^-(x) = \emptyset$ , any choice of  $\gamma_{\mathcal{E}^{0+}(x)}$  is suitable.  $\square$

The following lemma presents a central piece of what follows: the vanishing of the last term in (6.14). Its proof can be made rather simple but requires some notation defined in the remainder.<sup>7</sup>

**Lemma 6.1.6** (cancelling the term of  $\mathcal{E}^-(x)$ ). *Let  $\gamma_{\mathcal{E}^{0+}(x)} \in [0, 1]^{\mathcal{E}^{0+}(x)}$ , then there exists a  $\gamma_{\mathcal{E}^-(x)} \in [0, 1]^{\mathcal{E}^-(x)}$  such that the term lines 3-4 of (6.14) vanishes, i.e.,*

$$\begin{aligned} H_{\mathcal{E}^-(x)}(x)^T [\min(F'_{\mathcal{E}^-(x)}(x)(-g), G'_{\mathcal{E}^-(x)}(x)(-g)) \\ - (\Gamma_{\mathcal{E}^-(x)} F'_{\mathcal{E}^-(x)}(x) + \bar{\Gamma}_{\mathcal{E}^-(x)} G'_{\mathcal{E}^-(x)}(x))(-g)] = 0 \\ \Leftrightarrow \min(F'_{\mathcal{E}^-(x)}(x)(-g), G'_{\mathcal{E}^-(x)}(x)(-g)) = -(\Gamma_{\mathcal{E}^-(x)} F'_{\mathcal{E}^-(x)}(x) + \bar{\Gamma}_{\mathcal{E}^-(x)} G'_{\mathcal{E}^-(x)}(x))g. \end{aligned}$$

$\square$

The equivalence between the two lines comes from the fact that, for  $u < 0$  and  $v \leq 0$ ,  $u^T v = 0 \Leftrightarrow v = 0$ . The proof of this lemma was greatly simplified after the detour of chapter 3, which eventually brought to light that such value of  $\gamma_{\mathcal{E}^-(x)}$  is essentially obtained by an orthogonal projection, i.e., a rather simple operation. The difficulty is to identify the point and the set it is projected on.

In the following remark,  $V := (G'_{\mathcal{E}(x)}(x) - F'_{\mathcal{E}(x)}(x))^T$  denotes the differences of the partial Jacobian matrices, which is related to the B-differential of  $H$  (see chapter 3).

**Remark 6.1.7** (reformulation of the term of  $\mathcal{E}^-(x)$ ). In lemma 6.1.6, the equation can be reformulated as follows, by successively isolating the term  $G'_{\mathcal{E}^-(x)}(x)d$ , using the definition

---

<sup>7</sup>Note that one can have  $\theta'(x; -g) \leq 0$  even if this term is positive, if the others are sufficiently negative.

of  $V$  and multiplying by  $\text{Diag}(H_{\mathcal{E}^-(x)}(x))$ .

$$\begin{aligned}
 \Gamma_{\mathcal{E}^-(x)} F'_{\mathcal{E}^-(x)}(x)(-g) + \bar{\Gamma}_{\mathcal{E}^-(x)} G'_{\mathcal{E}^-(x)}(x)(-g) &= \min(F'_{\mathcal{E}^-(x)}(x)(-g), G'_{\mathcal{E}^-(x)}(x)(-g)) \\
 \Gamma_{\mathcal{E}^-(x)} [F'_{\mathcal{E}^-(x)}(x) - G'_{\mathcal{E}^-(x)}(x)] + G'_{\mathcal{E}^-(x)}(x)(-g) &= \min((F'_{\mathcal{E}^-(x)}(x) - G'_{\mathcal{E}^-(x)}(x))(-g), 0) \\
 &\quad + G'_{\mathcal{E}^-(x)}(x)(-g) \\
 \Gamma_{\mathcal{E}^-(x)} \text{Diag}(H_{\mathcal{E}^-(x)}(x))[-V_{:, \mathcal{E}^-(x)}^\top](-g) &= \text{Diag}(H_{\mathcal{E}^-(x)}(x)) \min(V_{:, \mathcal{E}^-(x)}^\top g, 0) \\
 \Gamma_{\mathcal{E}^-(x)} \mathcal{M}_-^\top(-g) &= \max(\mathcal{M}_-^\top(-g), 0)
 \end{aligned} \tag{6.16}$$

Using  $\tilde{g} = \mathcal{M}_-^\top(-g)$ , the last expression also reads, for all  $i \in \mathcal{E}^-(x)$ ,  $\gamma_i \tilde{g}_i = \max(\tilde{g}_i, 0)$ .

Now, let us define a few intermediary variables. In particular, we map  $\gamma_{\mathcal{E}^{0+}(x)}$  and  $\gamma_{\mathcal{E}^-(x)}$  who belong in  $[0, 1]$  to  $\eta$  and  $\zeta$  who belong in  $[-1, +1]$ .

**Rule 6.1.8** (variables correspondence). In what follows, we use the following quantities:

$$\begin{aligned}
 X &= \frac{1}{2}\mathcal{M}_+, \quad Y = -\frac{1}{2}\mathcal{M}_-, \quad \bar{x} - \bar{y} := g_0(x) + \frac{\mathcal{M}_+}{2}e + \frac{\mathcal{M}_-}{2}e \\
 \eta &= 2\gamma_{\mathcal{E}^{0+}(x)} - e, \quad \zeta = 2\gamma_{\mathcal{E}^-(x)} - e \\
 g &= g_0(x) + \mathcal{M}_+ \gamma_{\mathcal{E}^{0+}(x)} + \mathcal{M}_- \gamma_{\mathcal{E}^-(x)} \\
 &= \frac{\mathcal{M}_+}{2}\eta + \frac{\mathcal{M}_-}{2}\zeta + \bar{x} - \bar{y} = \bar{x} - \bar{y} + X\eta - Y\zeta := g(\eta, \zeta)
 \end{aligned} \tag{6.17}$$

Observe that  $\bar{x}$  and  $\bar{y}$  are not explicitly defined. In fact, in what we need below, we only need their difference  $\bar{x} - \bar{y}$ .<sup>8</sup> These variables allow for the following reformulation.

**Remark 6.1.9** (reformulation of  $\theta'(x; -g)$ ). One may express  $\theta'(x; -g)$  in (6.14) as:

$$\begin{aligned}
 \theta'(x; (-g)) &= -||g||^2 - \gamma_{\mathcal{E}^{0+}(x)}^\top [\text{Diag}(H_{\mathcal{E}^{0+}(x)})(F'_{\mathcal{E}^{0+}(x)}(x) - G'_{\mathcal{E}^{0+}(x)}(x))](-g) \\
 &\quad - H_{\mathcal{E}^{0+}(x)}^\top \max((G'_{\mathcal{E}^{0+}(x)}(x) - F'_{\mathcal{E}^{0+}(x)}(x))(-g), 0) \\
 &\quad - \gamma_{\mathcal{E}^-(x)}^\top [\text{Diag}(H_{\mathcal{E}^-(x)})(F'_{\mathcal{E}^-(x)}(x) - G'_{\mathcal{E}^-(x)}(x))](-g) \\
 &\quad - H_{\mathcal{E}^-(x)}^\top \max((G'_{\mathcal{E}^-(x)}(x) - F'_{\mathcal{E}^-(x)}(x))(-g), 0) \\
 &= -||g||^2 - \gamma_{\mathcal{E}^{0+}(x)}^\top \mathcal{M}_+^\top(-g) - e^\top \max(-\mathcal{M}_+^\top(-g), 0) \\
 &\quad - \gamma_{\mathcal{E}^-(x)}^\top \mathcal{M}_-^\top(-g) - e^\top \min(-\mathcal{M}_-^\top(-g), 0) \\
 &= -||g||^2 - \eta^\top X^\top(-g) - e^\top X^\top(-g) - e^\top \max(-2X^\top(-g), 0) \\
 &\quad + \zeta^\top Y^\top(-g) + e^\top Y^\top(-g) - e^\top \min(2Y^\top(-g), 0) \\
 &= -||g||^2 - \eta^\top X^\top(-g) + \zeta^\top Y^\top(-g) \\
 &\quad - e^\top \max(X^\top(-g), -X^\top(-g)) - e^\top \min(Y^\top(-g), -Y^\top(-g)) \\
 &= -||g||^2 - \eta^\top X^\top(-g) + \zeta^\top Y^\top(-g) - ||X^\top(-g)||_1 + ||Y^\top(-g)||_1
 \end{aligned}$$

In particular, it confirms that the term with indices in  $\mathcal{E}^{0+}(x)$  is nonpositive and the one with indices in  $\mathcal{E}^-(x)$  is nonnegative.  $\square$

<sup>8</sup>Like in chapter 3 where the difference of partial Jacobians intervene.

**Proposition 6.1.10** (lemma as a projection). Let  $\gamma_{\mathcal{E}^0(x)} \in [0, 1]^{\mathcal{E}^0(x)}$ ,  $\eta = 2\gamma_{\mathcal{E}^0(x)} - e$ . Let  $Z_y = Y[-1, +1]^{\mathcal{E}^-(x)}$  (zonotope generated by  $Y$ , see section B.2). Let  $\eta \in [-1, +1]^{\mathcal{E}^0(x)}$ ,  $z = P_{Z_y}(\bar{x} - \bar{y} + X\eta)$  and  $\zeta^* \in [-1, +1]^{\mathcal{E}^-(x)}$  such that  $z = Y\zeta^*$ .

Then  $\gamma_{\mathcal{E}^-(x)} = (1 + \zeta^*)/2$  is a value verifying lemma 6.1.6 for  $\gamma_{\mathcal{E}^0(x)}$ .  $\square$

*Proof.* Let  $z_0 = \bar{x} - \bar{y} + X\eta$  and consider  $P_{Z_y}(\bar{x} - \bar{y} + X\eta)$  (see remark B.2.7 and proposition B.2.8). Since the set  $Z_y$  is closed convex (affine transformation of a compact convex), the projection is well-defined. This problems reads

$$\min_{z \in Z_y} \frac{1}{2} \|z - z_0\|^2 = \min_{\zeta \in [-1, +1]^{\mathcal{E}^-(x)}} \frac{1}{2} \|Y\zeta - z_0\|^2$$

which clearly has qualified constraints (affine and set is nonempty). Thus, the KKT conditions read as follows

$$\text{KKT} \quad \begin{cases} Y^\top(Y\zeta - z_0) - \mu + \nu = 0, \\ 0 \leq \mu \perp (-e - \zeta) \leq 0, \\ 0 \leq \nu \perp (\zeta - e) \leq 0. \end{cases}$$

By the complementarity conditions, one gets  $\zeta_i = -1 \Rightarrow \nu_i = 0$  and  $\zeta_i = +1 \Rightarrow \mu_i = 0$ . Moreover, if  $\zeta_i \in (-1, +1)$ ,  $\mu_i = 0 = \nu_i = (Y^\top(Y\zeta - z_0))_i$ . The KKT system becomes (denoting by  $y_i$  the columns of  $Y$ )

$$\begin{cases} \zeta_i = +1, & y_i^\top(Y\zeta - z_0) = -\nu_i \leq 0, \\ \zeta_i = -1, & y_i^\top(Y\zeta - z_0) = +\mu_i \geq 0, \\ \zeta_i \in (-1, +1), & y_i^\top(Y\zeta - z_0) = 0, \end{cases} \Leftrightarrow \begin{cases} \zeta_i \in \{-1, +1\}, & \zeta_i y_i^\top(z_0 - Y\zeta) \geq 0, \\ \zeta_i \in (-1, +1), & y_i^\top(z_0 - Y\zeta) = 0. \end{cases}$$

Now, recall that  $z_0 = \bar{x} - \bar{y} + X\eta$ , so  $z_0 - Y\zeta = g(\gamma_{\mathcal{E}^0(x)}, \gamma_{\mathcal{E}^-(x)})$ . Furthermore, in remark 6.1.7, since  $Y = -\mathcal{M}_-/2$ , one can rewrite (6.16) as

$$\begin{aligned} \Gamma_{\mathcal{E}^-(x)} \mathcal{M}_-^\top(-g) &= \max(\mathcal{M}_-^\top(-g), 0) \\ \Leftrightarrow \frac{1}{2} (\zeta + e) \cdot (-2Y)^\top(-g) &= \max((-2Y)^\top(-g), 0) \\ \Leftrightarrow (\zeta + e) \cdot Y^\top g &= \max(2Y^\top g, 0) \\ \Leftrightarrow \zeta \cdot Y^\top g &= \max(Y^\top g, -Y^\top g) \\ \Leftrightarrow \zeta \cdot Y^\top g &= |Y^\top g| \\ \Leftrightarrow \forall i \in \mathcal{E}^-(x), \zeta_i y_i^\top g &= |y_i^\top g| \end{aligned}$$

which is the same conditions as the KKT system.  $\square$

Let us mention a relevant point: while the projection  $z$  is unique since it is well-defined, the set  $\{\zeta \in [-1, +1]^{\mathcal{E}^-(x)} : Y\zeta = z\}$  may not be a singleton. However, any of such  $\zeta$  produces the same value of  $g(\gamma_{\mathcal{E}^0(x)}, \gamma_{\mathcal{E}^-(x)})$ . Indeed, recall that  $g = \bar{x} - \bar{y} + X\eta - Y\zeta = \bar{x} - \bar{y} + X\eta - z$  which is independent of the  $\zeta$  chosen.

**Counter-example 6.1.11** (nonuniqueness of  $\zeta$ ). Consider the data  $Y = [e_1 \ e_2 \ e_1 + e_2]$  in dimension 3 and project  $e_3$  onto  $Y[-1, +1]^3$ . The KKT system is solved by  $\zeta = (t, t, -t)$  for any  $t \in [-1, +1]$  and  $\mu = 0 = \nu$ . Indeed,  $Y^T(Y[t; t; -t] - e_3) = Y^T(0 - e_3) = 0$ .  $\square$

We summarize the previous properties in the following discussion and definition. Let  $X$ ,  $Y$  and  $\bar{x} - \bar{y}$  be as described by rule 6.1.8. For any  $\gamma_{\mathcal{E}^0(x)} \in [0, 1]^{\mathcal{E}^0(x)}$  and its equivalent  $\eta = 2\gamma_{\mathcal{E}^0(x)} - e \in [-1, +1]^{\mathcal{E}^0(x)}$ , define  $z := P_{Y[-1, +1]^{\mathcal{E}^-(x)}}(\bar{x} - \bar{y} + X\eta)$  and  $\zeta \in [-1, +1]^{\mathcal{E}^-(x)}$  such that  $z = Y\zeta$  and  $\gamma_{\mathcal{E}^-(x)} = (\zeta + e)/2$ . Then one has, by lemma 6.1.6,

$$\min(F'_{\mathcal{E}^-(x)}(x)(-g), G'_{\mathcal{E}^-(x)}(x)(-g)) = (\Gamma_{\mathcal{E}^-(x)}F'_{\mathcal{E}^-(x)}(x) + \bar{\Gamma}_{\mathcal{E}^-(x)}G'_{\mathcal{E}^-(x)}(x))(-g), \\ \theta'(x; -g) \leq -\|g\|^2,$$

which means that  $-g$  is a descent direction of  $\theta$  if it is nonzero.

**Definition 6.1.12** (choosing  $\gamma_{\mathcal{E}^-(x)}$  for a descent property). Let  $\gamma_{\mathcal{E}^0(x)} \in [0, 1]^{\mathcal{E}^0(x)}$ ,  $\eta = 2\gamma_{\mathcal{E}^0(x)} - e$  and  $z = P_{Y[-1, +1]^{\mathcal{E}^-(x)}}(\bar{x} - \bar{y} + X\eta) = Y\zeta$ .

In what follows, we define  $\mathbb{G}(\gamma_{\mathcal{E}^0(x)}) \subseteq [0, 1]^{\mathcal{E}^-(x)}$  as the subset  $\mathbb{G}(\gamma_{\mathcal{E}^0(x)}) := \{\gamma_{\mathcal{E}^0(x)} \in [0, 1]^{\mathcal{E}^0(x)} : Y(2\gamma_{\mathcal{E}^-(x)} - e) = Y\zeta\}$ . Next, we consider a particular value of this set, defined and denoted by <sup>9</sup>

$$\gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^0(x)}) := \operatorname{argmin} \frac{1}{2}\|\gamma_{\mathcal{E}^-(x)} - e/2\|^2 \\ \text{s.t. } \gamma_{\mathcal{E}^-(x)} \in \mathbb{G}(\gamma_{\mathcal{E}^0(x)}) \Leftrightarrow \begin{cases} \gamma_{\mathcal{E}^-(x)} \in [0, 1]^{\mathcal{E}^-(x)} \\ Y(2\gamma_{\mathcal{E}^-(x)} - e) = Y\zeta. \end{cases} \quad \square$$

Observe that  $\gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^0(x)})$  is well-defined by virtue of being the projection of  $e/2$  on a nonempty (since  $(\zeta + e)/2$  belongs to it) set defined by affine equalities and inequalities. We chose the element closest to the center of the hypercube ( $\gamma_{\mathcal{E}^-(x)} = e/2$  or equivalently  $\zeta = 0$  is the center of the hypercube), but one could have chosen another convention.

Let us conclude this section by a few remarks and illustrations on the previous properties.

- For a given  $\gamma_{\mathcal{E}^0(x)}$ ,  $\mathbb{G}(\gamma_{\mathcal{E}^0(x)})$  is closed convex compact.
- The set of  $\gamma_{\mathcal{E}^0(x)}$  such that  $g(\gamma_{\mathcal{E}^0(x)}, \gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^0(x)})) = 0$  may not be of measure 0 in  $[0, 1]^{\mathcal{E}^0(x)}$ .

Now, consider example 6.1.3: by simple computations, one has

$$X = \emptyset, Y = \frac{1}{2} \begin{bmatrix} 3 & 0 \\ 0 & -\delta \end{bmatrix}, \bar{x} - \bar{y} = \begin{bmatrix} 2 \\ -2\delta \end{bmatrix} - \frac{1}{2} \begin{bmatrix} 3 & 0 \\ 0 & -\delta \end{bmatrix} e = \begin{bmatrix} 1/2 \\ -\delta \end{bmatrix}.$$

The projection of the point  $(1/2, -\delta)$  on  $Y[-1, +1]^2 = [-3/2, 3/2] \times [-\delta/2, +\delta/2]$  is the point  $(1/2, -\delta/2)$  corresponding to  $\zeta = (1/3, 1)$  and  $\gamma_{\mathcal{E}^-(x)} = (2/3, 1)$ .

<sup>9</sup>The last constraint also reads  $\bar{x} - \bar{y} + X\eta - Y(2\gamma_{\mathcal{E}^-(x)} - e) = g$ .

Similarly, in example 6.1.4, by simple computation, one has

$$X = \emptyset, Y = \frac{1}{200} \begin{bmatrix} 1 & 10 \\ 1 & 0 \end{bmatrix}, \bar{x} - \bar{y} = \frac{1}{100} \begin{bmatrix} 9 \\ -3 \end{bmatrix} - \frac{1}{200} \begin{bmatrix} 11 \\ 1 \end{bmatrix} = \frac{1}{200} \begin{bmatrix} 7 \\ -7 \end{bmatrix}.$$

The projection of the point  $(7, -7)/200$  on  $Y[-1, +1]^2$  is the point  $(7, -1)/200$  corresponding to  $\zeta = (1, 3/5)$  and  $\gamma_{\mathcal{E}^0+(x)} = (1, 4/5)$ .

### 6.1.3 Choice of the weights and stationarity

#### Necessary and sufficient condition of stationarity

The framework of the previous section introduces convex weights for the indices in  $\mathcal{E}(x)$ , and proposes a way to choose a part of the values,  $\gamma_{\mathcal{E}^-(x)}$ , to ensure the algorithm finds a descent direction. However, the choice of  $\gamma_{\mathcal{E}^0+(x)}$  remains open. One relevant obstacle to this is the following observation. Recall that  $\mathbb{G}$  is introduced in definition 6.1.12.

**Remark 6.1.13** (case  $g = 0$ ). It may happen that for a given  $\gamma_{\mathcal{E}^0+(x)}$ , for any  $\gamma_{\mathcal{E}^-(x)} \in \mathbb{G}(\gamma_{\mathcal{E}^0+(x)})$ ,  $g(\gamma_{\mathcal{E}^0+(x)}, \gamma_{\mathcal{E}^-(x)}) = 0$ .  $\square$

This is actually relevant for the next proposition, which characterizes the Dini/strong stationarity of  $\theta$  (definition 2.3.39) at a given point  $x$ .

**Proposition 6.1.14** (NSC of  $\theta$ -stationarity). *The following properties are equivalent:*

- (i)  $x$  is a Dini/strong stationary point of  $\theta$ , i.e.  $\forall h \in \mathbb{R}^n, \theta'(x; h) \geq 0$ ,
- (ii)  $(\mathcal{C}_\theta^\square)$  for any  $\gamma_{\mathcal{E}^0+(x)} \in [0, 1]^{\mathcal{E}^0+(x)}$  and  $\gamma_{\mathcal{E}^-(x)} \in \mathbb{G}(\gamma_{\mathcal{E}^0+(x)})$ , one has  $g = 0$ ,
- (iii)  $(\mathcal{C}_\theta^\{ \})$  for any  $\gamma_{\mathcal{E}^0+(x)} \in \{0, 1\}^{\mathcal{E}^0+(x)}$  and  $\gamma_{\mathcal{E}^-(x)} \in \mathbb{G}(\gamma_{\mathcal{E}^0+(x)})$ , one has  $g = 0$ .

In what follows, we use  $(\mathcal{C}_\theta)$  to refer to  $(\mathcal{C}_\theta^\square)$  or  $(\mathcal{C}_\theta^\{ \})$ . It is clear that point (iii) is a subcase of (ii), and it corresponds to a partition of  $\mathcal{E}^0+(x)$ , as it is done for instance in the Newton-min or PNM algorithms. Observe that even in this case, the corresponding  $\gamma_{\mathcal{E}^-(x)}$  given by definition 6.1.12 are not necessarily in  $\{0, 1\}^{\mathcal{E}^-(x)}$ .

*Proof.* [(i)  $\Rightarrow$  (ii)] By contrapositive, if (ii) is not verified, there exists a certain  $\gamma_{\mathcal{E}^0+(x)}$  such that  $g(\gamma_{\mathcal{E}^0+(x)}, \gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^0+(x)})) \neq 0$  where  $\gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^0+(x)}) \in \mathbb{G}(\gamma_{\mathcal{E}^0+(x)})$  (definition 6.1.12). Thus, it is a nonzero strict descent direction, meaning  $\theta$  is not Dini-stationary at  $x$ .

[(ii)  $\Rightarrow$  (iii)] Clear since (iii) considers only some cases of (ii).

[(iii)  $\Rightarrow$  (i)] Let  $h \in \mathbb{R}^n$  be arbitrary and let us show that  $\theta'(x; h) \geq 0$ . Consider the following particular value of  $\gamma_{\mathcal{E}^0+(x)}$  defined as follows:

$$\begin{cases} \mathcal{E}_{\mathcal{F}}^{0+}(x) := \{i \in \mathcal{E}^{0+}(x) : F'_i(x)h \leq G'_i(x)h\}, \\ \mathcal{E}_{\mathcal{G}}^{0+}(x) := \{i \in \mathcal{E}^{0+}(x) : F'_i(x)h > G'_i(x)h\}, \end{cases} \quad \gamma_i = \begin{cases} 1 & i \in \mathcal{E}_{\mathcal{F}}^{0+}(x), \\ 0 & i \in \mathcal{E}_{\mathcal{G}}^{0+}(x). \end{cases}$$

Clearly  $\gamma_{\mathcal{E}^{0+}(x)} \in \{0, 1\}^{\mathcal{E}^{0+}(x)}$ . Now, consider the associated  $\gamma_{\mathcal{E}^{-}(x)}(\gamma_{\mathcal{E}^{0+}(x)})$  of definition 6.1.12. By (iii),  $g = 0$  and (6.14) becomes

$$\begin{aligned} \theta'(x; h) &= (F_{\mathcal{F}(x)}(x)^T F'_{\mathcal{F}(x)}(x) + G_{\mathcal{G}(x)}(x)^T G'_{\mathcal{G}(x)}(x))h \\ &\quad + H_{\mathcal{E}(x)}(x)^T \min(F'_{\mathcal{E}(x)}(x)h, G'_{\mathcal{E}(x)}(x)h) \\ [(6.11)] \quad &= (g - [F'_{\mathcal{E}(x)}(x)^T \Gamma + G'_{\mathcal{E}(x)}(x)^T \bar{\Gamma}] H_{\mathcal{E}(x)}(x))^T h \\ &\quad + H_{\mathcal{E}(x)}(x)^T \min(F'_{\mathcal{E}(x)}(x)h, G'_{\mathcal{E}(x)}(x)h) \\ [g = 0] \quad &= H_{\mathcal{E}(x)}(x)^T [\min(F'_{\mathcal{E}(x)}(x)h, G'_{\mathcal{E}(x)}(x)h) - (\Gamma F'_{\mathcal{E}(x)}(x) + \bar{\Gamma} G'_{\mathcal{E}(x)}(x))h] \end{aligned}$$

Now, let us split the indices into  $\mathcal{E}_{\mathcal{F}}^{0+}(x)$  and  $\mathcal{E}_{\mathcal{G}}^{0+}(x)$  and  $\mathcal{E}^-(x)$ :

- if  $i \in \mathcal{E}_{\mathcal{F}}^{0+}(x)$ , one has  $\min(F'_i(x)h, G'_i(x)h) = F'_i(x)h$ ,  $\gamma_i = 1$  and  $\bar{\gamma}_i = 0$ , so the quantity inside the brackets vanishes;
- if  $i \in \mathcal{E}_{\mathcal{G}}^{0+}(x)$ , one has  $\min(F'_i(x)h, G'_i(x)h) = G'_i(x)h$ ,  $\gamma_i = 0$  and  $\bar{\gamma}_i = 1$ , so the quantity inside the brackets vanishes;
- if  $i \in \mathcal{E}^-(x)$ , the quantity inside brackets is nonpositive since it is of the form  $\min(a, b) - \gamma a - \bar{\gamma} b$ , thus when multiplied by  $H_i(x) < 0$ , one gets a nonnegative quantity.

Summing over all the indices,  $\theta'(x; h) \geq 0$  for all  $h$ , meaning  $\theta$  is Dini-stationary at  $x$ .  $\square$

The main consequence of this proposition is to state, in terms of the convex weights framework described, that either the point is Dini-stationary or there is a descent direction.

### Comments on the complexity of $(\mathcal{C}_\theta)$

While  $(\mathcal{C}_\theta)$  may appear as a handy stopping criterion (either the algorithm stops or is guaranteed to progress), its verification may not be feasible in polynomial time. We show this in two simple steps: the reformulation of points (ii) and (iii) of proposition 6.1.14 as a geometric problem, and the use of a pair of references that deal with this specific geometric problem and show it has a combinatorial nature.

**Proposition 6.1.15** (polytopic formulation). *The following properties are equivalent:*

- (i)  $x$  is a Dini stationary point of  $\theta$ ,
- (ii)  $\forall \gamma = (\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)})$  with  $\gamma_{\mathcal{E}^-(x)} \in \mathbb{G}(\gamma_{\mathcal{E}^{0+}(x)})$ ,  $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}) = 0$ ,
- (iii)  $\forall \gamma_{\mathcal{E}^{0+}(x)} \in [0, 1]^{\mathcal{E}^{0+}(x)}$ ,  $\exists \gamma_{\mathcal{E}^-(x)} \in [0, 1]^{\mathcal{E}^-(x)}$  such that  $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}) = 0$ ,
- (iv)  $\mathcal{M}_+[0, 1]^{\mathcal{E}^{0+}(x)} \subseteq -g_0(x) - \mathcal{M}_-[0, 1]^{\mathcal{E}^-(x)}$ .
- (v)  $\bar{x} + X[-1, +1]^{\mathcal{E}^{0+}(x)} \subseteq \bar{y} + Y[-1, +1]^{\mathcal{E}^-(x)}$ .

*Proof.* [(i)  $\Leftrightarrow$  (ii)] Clear since (ii) is a reformulation of point (ii) of proposition 6.1.14.

$[(ii) \Leftrightarrow (iii)]$  The  $\Rightarrow$  sense is clear; the  $\Leftarrow$  sense comes from the fact that values  $\gamma_{\mathcal{E}^-(x)}$  such that  $g = 0$  are solutions.

$[(iii) \Leftrightarrow (iv)]$  Recall (6.11), where  $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}) = g_0(x) + \mathcal{M}_+ \gamma_{\mathcal{E}^{0+}(x)} + \mathcal{M}_- \gamma_{\mathcal{E}^-(x)}$ . Thus,  $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}) = 0$  reads

$$g_0(x) + \mathcal{M}_+ \gamma_{\mathcal{E}^{0+}(x)} + \mathcal{M}_- \gamma_{\mathcal{E}^-(x)} = 0 \Leftrightarrow \mathcal{M}_+ \gamma_{\mathcal{E}^{0+}(x)} = -g_0(x) - \mathcal{M}_- \gamma_{\mathcal{E}^-(x)}.$$

Then, the equivalence stems from the inclusion in (iv).

$[(iv) \Leftrightarrow (v)]$  Using rule 6.1.8, one has the following equivalent inclusions

$$\begin{aligned} \mathcal{M}_+[0, 1]^{\mathcal{E}^{0+}(x)} &\subseteq -g_0(x) - \mathcal{M}_-[0, 1]^{\mathcal{E}^-(x)} \\ [\mathcal{M}_+ = 2X] \quad 2X[0, 1]^{\mathcal{E}^{0+}(x)} &\subseteq -g_0(x) + 2Y[0, 1]^{\mathcal{E}^-(x)} \quad [\mathcal{M}_- = -2Y] \\ X([-1, +1]^{\mathcal{E}^{0+}(x)} + e) &\subseteq -g_0(x) + Y([-1, +1]^{\mathcal{E}^-(x)} + e) \\ X[-1, +1]^{\mathcal{E}^{0+}(x)} &\subseteq -g_0(x) - Xe + Ye + Y[-1, +1]^{\mathcal{E}^-(x)} \\ X[-1, +1]^{\mathcal{E}^{0+}(x)} &\subseteq -g_0(x) - \frac{\mathcal{M}_+}{2}e - \frac{\mathcal{M}_-}{2}e + Y[-1, +1]^{\mathcal{E}^-(x)} \\ \bar{x} + X[-1, +1]^{\mathcal{E}^{0+}(x)} &\subseteq \bar{y} + Y[-1, +1]^{\mathcal{E}^-(x)} \quad [g_0(x) + \frac{\mathcal{M}_+}{2}e + \frac{\mathcal{M}_-}{2}e = \bar{x} - \bar{y}] \end{aligned}$$

which concludes the proof.  $\square$

In the last line, we put  $\bar{x}$  and  $\bar{y}$  on different sides to comply with the formalism of a relevant paper treating (partially) this question (see appendix C). This clearly does not change whether the inclusion holds or not.

In particular, points (iv) and (v) show that assessing stationarity is equivalent to a polytope inclusion problem. This is where issues arise: the complexity of polytope inclusion is mostly determined by how the polytopes are described. Recall that polytopes are generally defined either by a list of vertices, the  $V$ -formulation, or an intersection of halfspaces, the  $H$ -formulation. For instance, determining if a  $H$ -polytope is included in a  $V$ -polytope is in general not solvable in polynomial time while the other cases are. This is highly related to the fact that swapping from a  $V$ -formulation to a  $H$ -formulation is (in general) not feasible in polynomial time (for references, see for instance [81, 96, 13, 34, 263]).

The polytopes in points (iv) and (v) are neither in  $H$  nor  $V$  form, but are *zonotopes* (for additional information on these polytopes, see section B.2 and for instance references [167, 263]). Since we are mainly interested in zonotope inclusion, we mention two relatively recent papers dealing with this topic. The first one [223] discusses the algorithmical resolution of various polytope inclusion problems and in particular for zonotopes, proposing a simple method which may ensure the inclusion holds. It is based on solving a linear optimization problem which computes a scalar value, some sort of dilation factor.

When the optimal value of the linear problem is  $\leq 1$ , the inclusion holds, i.e., the first zonotope  $\bar{x} + X[-1, +1]^{\mathcal{E}^{0+}(x)}$  is contained in the second one  $\bar{y} + Y[-1, +1]^{\mathcal{E}^-(x)}$ . However,

when it is  $> 1$ , it may still be possible that inclusion holds – see [223, example 2]. This is linked with the property given in the second paper [149], where the authors show that, in general, zonotope containment is co-NP-complete.

**Theorem 6.1.16** (corollary 4 of [149]). *Zonotope containment is co-NP-complete.*  $\square$

**Remark 6.1.17** ('discrete' versus 'continuous'). In propositions 6.1.14 and 6.1.15, one has the equivalence between  $[0, 1]^{\mathcal{E}^0(x)}$  et  $\{0, 1\}^{\mathcal{E}^0(x)}$  since checking if a convex polytope  $P_1$  ( $[0, 1]$ ) is contained in polytope  $P_2$  is equivalent to checking if the vertices of  $P_1$  ( $\{0, 1\}$ ) are contained in  $P_2$ . Furthermore, it is clear that the vertices of the zonotope  $V[0, 1]^m$  for  $V \in \mathbb{R}^{n \times m}$  are contained in  $V\{0, 1\}^m$ .  $\square$

Let us mention a specific case, though restrictive, that reduces the zonotope inclusion to a simple linear optimization problem.

**Proposition 6.1.18** (injectivity in  $\mathcal{E}^-(x)$ ). *If the matrix  $(G'(x) - F'(x))_{\mathcal{E}^-(x), :}$  is surjective, or equivalently the matrices  $Y$  and  $\mathcal{M}_-$  are injective, then the inclusion can be solved in polynomial time, and a  $\gamma_{\mathcal{E}^0(x)}$  such that  $g(\gamma_{\mathcal{E}^0(x)}, \gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^0(x)})) \neq 0$  can be obtained when the inclusion does not hold.*  $\square$

The proof is given in the appendices, in proposition C.0.7 (particular case of [223, thm 3, corollary 2]). In short, when  $Y$  is injective, it has a left-inverse, so the second zonotope is similar to a hypercube and the parametrization  $\zeta$  is unique (compare with example 6.1.11). This particular case usually cannot happen for zonotopes, since the matrix  $Y$  has more columns than rows. It may happen here since  $Y \in \mathbb{R}^{n \times |\mathcal{E}^-(x)|}$  and clearly  $|\mathcal{E}^-(x)| \leq n$  since  $\mathcal{E}^-(x) \subseteq [1 : n]$ .

In [149], a few algorithms, essentially relying on enumeration, are proposed. The appendix C gives further details about the question of zonotope inclusion and the simple (but inexact) algorithm proposed in [223], and presents a way to include the enumeration in a single linear optimization problem involving binary variables<sup>10</sup>.

The exposed difficulty of finding a descent direction / guaranteeing stationarity in the nonsmooth nonconvex case is discussed for instance in [19]: "Accordingly, there are very few methods that are guaranteed to converge to stationary points in the case of a non-smooth and nonconvex objective function." (see the paragraph between pp. 57 and 58). The method they propose, though not designed for complementarity problems, deals with a nonsmooth nonconvex constrained problem. In particular, their algorithm computes a "positive spanning set" of the feasible directions, which may have exponential complexity. The remedy they propose is to use a random selection instead of the whole computation.

In our setting, we could imagine selecting randomly values  $\gamma_{\mathcal{E}^0(x)} \in \{0, 1\}^{\mathcal{E}^0(x)}$  until one returns a nonzero  $g$  (if an approximate stopping criterion does not hold). Indeed, since

---

<sup>10</sup>Which may be easier to setup than the explicit combinatorial enumeration and can be solved by efficient optimization software such as GUROBI.

one wants to find a descent direction, it “suffices” to find one to begin the iteration starting at point  $x$ .

In addition, let us underline the following observation: proposition 6.1.14 and theorem 6.1.16 indicate that even the *verification* of Dini-stationarity of  $\theta$  is in general not polynomial. Thus, *obtaining* such a point seems to be, without strong assumptions, unrealistic. Some conditions for LCPs are proposed in section 6.1.5.

#### 6.1.4 Choice of the weights and differentials

The previous parts have exhibited the link between finding descent directions of  $\theta$  and suitable treatment of the indices in  $\mathcal{E}(x)$  as in the Newton-min algorithm for instance. The following observation seems somewhat related to these questions.

**Proposition 6.1.19** (projection of zero on the subdifferential). *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a locally Lipschitz function and let  $\partial f(x)$  be its (Clarke) differential at  $x$ . Define  $g := P_{\partial f}(0)$ <sup>11</sup>, then  $-g$  is a descent direction of  $f$  at  $x$ .*

*Proof.* By the definition of the orthogonal projection on a convex set, one has the following inequality:

$$\forall \xi \in \partial f(x), (\xi - g, g - 0) \geq 0.$$

This also reads  $(\xi, g) \geq \|g\|^2$  or  $(\xi, -g) \leq -\|g\|^2$  which is negative if  $g \neq 0$ . However,  $g = 0$  means 0 is in the differential, so  $f$  is stationary in a certain sense at  $x$ . Then, recall that  $f' \leq f^\circ$  (see the remark below definition 2.3.7), one has

$$f'(x; -g) \leq f^\circ(x; -g) = \max\{(\xi, -g) : \xi \in \partial f(x)\} \leq -\|g\|^2 \leq 0$$

where the second equality comes from definition 2.3.9. Clearly,  $-g$  is a strict descent direction if it is nonzero.  $\square$

This proposition, inspired from the convex case, can in particular be used for  $\theta$ . Naturally, it requires the knowledge of  $\partial\theta = \partial H^\top H$ , which is unlikely to be easy to compute.<sup>12</sup>

Moreover, when the returned  $g$  equals 0, i.e.,  $0 \in \partial\theta(x)$ , there may still be descent directions – the point is stationary but not Dini-stationary (definitions 2.3.38 and 2.3.39). Such situation is for instance illustrated in the following example, which shall return later.

The inadequacy of proposition 6.1.19 may come from the following observation. In the proof, the inequality  $f' \leq f^\circ$  is used. However, as it was remarked in section 2.3.2, this inequality may be strict and rather imprecise when the function is defined by a componentwise minimum. Said differently, this proposition cannot go beyond Clarke stationary

<sup>11</sup>Clarke's differential is a closed convex set, see definition 2.3.9 and [51, §2].

<sup>12</sup>No formula for  $\partial H$ , using  $\text{conv}(\partial_B H)$  requires the nontrivial characterization of  $\partial_B H$  (chapters 3-4).

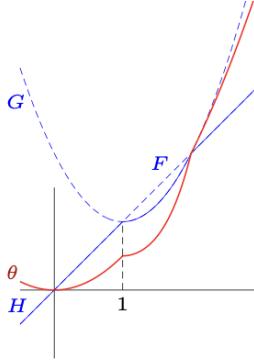


Figure 6.4: Illustration of nonDini stationarity. One has  $F(x) = x$ ,  $G(x) = 1 + (x - 1)^2$ , so the problem has a solution at  $x = 0$ . At  $x = 1$ , neither  $H$  nor  $\theta$  are differentiable. Since  $G'(1) = 0$ , by taking a sequence  $1 + t_k \rightarrow 1$ , one gets that  $0 \in \partial\theta(1)$ , but  $x = 1$  is clearly not strongly stationary. Such point is sometimes called a “concave kink”, which, as we shall see, may cause some difficulties to algorithms.

points since such points return  $g = 0$ . At  $x = 1$  in figure 6.4,  $g(\gamma) = \gamma \times 1 + \bar{\gamma} \times 0$  which is nonzero for  $\gamma > 0$ , in coherence with proposition 6.1.14.

We conclude this section by the following property and discussion. Let us summarize the steps of this section so far:

- replace the polyhedral system and the nonvacuity assumptions by least-squares;
- introduce convex weights to balance the contributions of all indices;
- link these weights with the differential of  $\theta$ ;
- give a condition for some of the weights to get a descent direction;
- deduce a NSC of Dini-stationarity with explicit complexity.

Before presenting an algorithm based on a Levenberg-Marquardt regularization of the least-squares, we mention a last property. Since its proof is rather long and technical, we postpone it to appendix D.

**Proposition 6.1.20** ( $g \in \partial\theta(x)$  with suitable  $\gamma_{\mathcal{E}^0+(x)}$ ). *Suppose  $\theta$  is not Dini-stationary at  $x$ . Then, there exists  $\gamma_{\mathcal{E}^0+(x)}$  such that  $g(\gamma_{\mathcal{E}^0+(x)}, \gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^0+(x)})) \in \partial\theta(x)$ . Furthermore, such a  $\gamma_{\mathcal{E}^0+(x)}$  may be found by solving the following equivalent problems ( $\gamma_{\mathcal{E}^0+(x)} = (\eta + e)/2$ )*

$$\begin{aligned} & \max_{\eta \in [-1, +1]^{\mathcal{E}^0+(x)}} \min_{\zeta \in [-1, +1]^{\mathcal{E}^-(x)}} \frac{1}{2} \|g(\eta, \zeta)\|^2 \\ \Leftrightarrow & \max_{\eta \in [-1, +1]^{\mathcal{E}^0+(x)}} \frac{1}{2} \text{dist}(\bar{x} - \bar{y} + X\eta, Y[-1, +1]^{\mathcal{E}^-(x)})^2 \quad (6.18) \\ \Leftrightarrow & \max_{\eta \in [-1, +1]^{\mathcal{E}^0+(x)}} \frac{1}{2} \|\bar{x} - \bar{y} + X\eta - P_{Y[-1, +1]^{\mathcal{E}^-(x)}}(\bar{x} - \bar{y} + X\eta)\|^2 \end{aligned}$$

*Strict local maxima also imply  $g(\gamma_{\mathcal{E}^0+(x)}, \gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^0+(x)})) \in \partial\theta(x)$ .*

□

Recall that the distance to a convex set is a convex function. Thus, the second problem maximizes a convex quadratic function over a polyhedron, so the maximum is attained at one of the vertices. Observe that this problem is a more demanding version of the previous requirement of finding  $\gamma_{\mathcal{E}^0(x)}$  such that  $g(\gamma_{\mathcal{E}^0(x)}, \gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^0(x)})) \neq 0$ , which makes sense since we added the requirement of  $g \in \partial\theta(x)$  (in addition to  $g \neq 0$ ).

Let us point out what is the relevance of this convoluted proposition. The Levenberg-Marquardt methods allows one to prove, under reasonable assumptions, that the tangent directions it produces tend to zero (theorem 6.2.4 below). In the smooth setting, it is sufficient to ensure the limit point is stationary. In the nonsmooth setting, this is where the proposition intervenes: if one can ensure  $g_k \in \partial\theta(x_k)$  for all  $k$  and  $x$  is an accumulation point of  $\{x_k\}$ , by the properties of  $\partial$  (see definition 2.3.9 and [51, proposition 2.1.5b p. 29]),  $0 \in \partial\theta(x)$ .

### 6.1.5 Discussion of regularity of solutions

This section aims at discussing, with the framework developed in the previous sections, regularity conditions to ensure that a point  $x$  is a solution to the problem. This can be seen as a strengthening of proposition 6.1.14, since solutions are particular Dini-stationary points. While it may be possible to extend what is developed below for the nonlinear counterparts, we only consider the basic LCP

$$0 \leq x \perp Mx + q \geq 0 \tag{6.19}$$

where we use the following notation

$$A := \mathcal{F}(x), \quad E := \mathcal{E}(x) \quad \text{and} \quad I := \mathcal{G}(x), \tag{6.20}$$

where we have assumed that  $F(x) \equiv x$  and  $G(x) \equiv Mx + q$ . In what follows, to avoid confusion with the index set  $I$ , the identity matrix is denoted by  $\text{Id}$ .

To start, we recall a proposition from [58, proposition 5.8.4], which identifies regularity conditions that a  $\theta$ -stationary point  $x$  must satisfy to be a solution to the problem. The regularity condition at  $x$  is the following (recall that  $M \in \mathbf{Q}$  if the LCP has a solution for any  $q$ , see definition 2.2.2):

$$\begin{cases} M_{I,I} \text{ is nonsingular,} \\ \text{the Schur complement } M_{E,E} - M_{E,I}M_{I,I}^{-1}M_{I,E} \in \mathbf{Q}. \end{cases} \tag{6.21}$$

By (6.20), the index sets  $I$  and  $E$  used in this regularity condition depends on  $x$ . These conditions are satisfied independently of  $x$  when  $M \in \mathbf{P}$ , since a  $\mathbf{P}$ -matrix has positive principal minors and has Schur complements that are in  $\mathbf{P}$  [192], hence in  $\mathbf{Q}$ .

**Proposition 6.1.21** ( $\theta$ -stationarity for LCP [58]). *If  $x$  is a stationary point of  $\theta$  verifying the regularity condition (6.21), then  $x$  is a solution to the LCP. In particular, if  $M \in \mathbf{P}$ , the unique stationary point of  $\theta$  is the solution to the LCP.*  $\square$

Therefore, for an LCP with a  $\mathbf{P}$ -matrix,  $\theta$ -stationary point and solution are two identical concepts.

Let us now see whether one can link the Dini-stationary of points or solutions to the LCP in terms of zero gradient of the piecewise quadratic models  $\varphi_x$  in (6.10), depending on the choice of  $\Gamma \in [0, \text{Id}]$ . With the notation (6.20), the zero gradient of the piecewise quadratic models  $\varphi_x$  in (6.10) is expressed as follows (recall that  $x_E = (Mx + q)_E$ )

$$I_{A,:}^\top x_A + M_{I,:}^\top (Mx + q)_I + I_{E,:}^\top \Gamma x_E + M_{E,:}^\top \bar{\Gamma} (Mx + q)_E = 0. \quad (6.22)$$

As evoked in the introduction of this section, one must strengthen the hypotheses to obtain a characterization of solutions, since they are particular Dini-stationary points. We have identified two such regularity conditions. The first one is the following (recall definition 2.2.3:  $M \in \mathbf{ND}$ , or is *nondegenerate*, if its principal submatrices are nonsingular):

$$\begin{cases} M_{I,I} \text{ is nonsingular,} \\ \text{the Schur complement } M_{E,E} - M_{E,I} M_{I,I}^{-1} M_{I,E} \in \mathbf{ND}. \end{cases} \quad (6.23)$$

These conditions are satisfied when  $M \in \mathbf{P}$  for similar reasons as those quoted for saying that (6.21) is verified for a  $\mathbf{P}$ -matrix.

The regularity assumption (6.23) used in the following proposition 6.1.14 has similarities with and is weaker than the notion of *R-regularity* defined by Facchinei and Soares [87, definition 2.1]. It is defined at a solution to a nonlinear CP but, extended to a nonsolution to an LCP, it would read (recall that  $M \in \mathbf{P}$  if the LCP (6.19) has one and only one solution for any  $q$ ):

$$\begin{cases} M_{I,I} \text{ is nonsingular,} \\ \text{the Schur complement } M_{E,E} - M_{E,I} M_{I,I}^{-1} M_{I,E} \in \mathbf{P}. \end{cases} \quad (6.24)$$

This last notion is also called *strong regularity* in [58, definition 5.8.3] and is related to the strong regularity of Robinson [220] (see also [196]). Again, (6.24) is satisfied, independently of  $x$ , if  $M \in \mathbf{P}$ . In the remainder of this part, we use extensively  $\Gamma = \text{Diag}(\gamma) \in [0, \text{Id}_{\mathcal{E}(x)}]$  since it intervenes in many computations.

**Proposition 6.1.22** (NSC of  $\theta$ -stationarity for LCP - I). *For  $x \in \mathbb{R}^n$ , consider the following properties:*

- (i)  *$x$  is a solution to the LCP (6.19),*
- (ii) *for any  $\Gamma \in [0, \text{Id}]$ , the identity (6.22) holds,*
- (iii) *for some  $\Gamma \in \text{ext}[0, \text{Id}]$ , the identity (6.22) holds.*

*Then, (i)  $\Rightarrow$  (ii)  $\Rightarrow$  (iii). One also has the implication (iii)  $\Rightarrow$  (i), provided the regularity assumption (6.23) holds at  $x$ .*

*Proof.* [(i)  $\Rightarrow$  (ii)] Since  $x$  is a solution, one has  $F_{\mathcal{F}(x)}(x) = x_{\mathcal{F}(x)} = x_A = 0$ ,  $G_{\mathcal{G}(x)}(x) = (Mx + q)_{\mathcal{G}(x)} = (Mx + q)_I = 0$  and  $H_{\mathcal{E}(x)} = x_E = 0 = (Mx + q)_E$ , so clearly  $g = 0$  whatever  $\Gamma$  is.

$[(ii) \Rightarrow (iii)]$  Clear.

$[(iii) \Rightarrow (i)]$  Suppose that (6.23) holds at  $x$ . The system (6.22) also reads

$$\begin{cases} x_A + M_{I,A}^T(Mx + q)_I + M_{E,A}^T(\text{Id} - \Gamma)(Mx + q)_E = 0 \\ M_{I,E}^T(Mx + q)_I + \Gamma x_E + M_{E,E}^T(\text{Id} - \Gamma)(Mx + q)_E = 0 \\ M_{I,I}^T(Mx + q)_I + M_{E,I}^T(\text{Id} - \Gamma)(Mx + q)_E = 0. \end{cases} \quad (6.25a)$$

Since  $M_{I,I}$  is nonsingular, the last equation gives

$$(Mx + q)_I = -M_{I,I}^{-1}M_{E,I}^T(\text{Id} - \Gamma)(Mx + q)_E. \quad (6.25b)$$

After substitution in the second equation of (6.25a) and the use of  $(Mx + q)_E = x_E$ , this one becomes

$$(M_{E,E} - M_{E,I}M_{I,I}^{-1}M_{I,E})^T(\text{Id} - \Gamma)x_E + \Gamma x_E = 0. \quad (6.25c)$$

Since  $\Gamma \in \text{ext}[0, \text{Id}]$ , one has  $\Gamma^2 = \Gamma$ , so that  $(\text{Id} - \Gamma)\Gamma = 0$  and, after a left-multiplication of its two sides by  $(\text{Id} - \Gamma)$ , the previous equation becomes

$$\left( (\text{Id} - \Gamma) (M_{E,E} - M_{E,I}M_{I,I}^{-1}M_{I,E}) (\text{Id} - \Gamma) \right)^T x_E = 0.$$

By (6.23) and  $\Gamma \in \text{ext}[0, \text{Id}]$ , the principal submatrix  $(\text{Id} - \Gamma)(M_{E,E} - M_{E,I}M_{I,I}^{-1}M_{I,E})(\text{Id} - \Gamma)$  of  $M_{E,E} - M_{E,I}M_{I,I}^{-1}M_{I,E}$  is nonsingular, implying that

$$(\text{Id} - \Gamma)x_E = 0.$$

Consequently, by (6.25b), (6.25c) and the first equation in (6.25a), we get

$$(Mx + q)_I = 0, \quad \Gamma x_E = 0 \quad \text{and} \quad x_A = 0.$$

To show that  $x$  is a solution to (6.19), one still has to prove that  $x_I \geq 0$  and  $(Mx + q)_A \geq 0$ . This can be deduced from the definition of  $A$  and  $I$ , since one has

$$0 = x_A < (Mx + q)_A, \quad \text{and} \quad 0 = (Mx + q)_I < x_I. \quad \square$$

Interestingly enough, we see that propositions 6.1.14 and 6.1.22 give related but different results (of course the former is for the nonlinear CP, while the latter deals with the LCP):

- the main difference is that property (iii) of proposition 6.1.14 requires a condition for *all*  $\Gamma_{\mathcal{E}^{0+}(x)}$  in  $[0, \text{Id}_{\mathcal{E}^{0+}(x)}]$ , while property (iii) of proposition 6.1.22 requires a condition for a *single*  $\Gamma \in \text{ext}[0, \text{Id}]$ ,
- proposition 6.1.14 uses no regularity assumption, but only  $\Gamma_{\mathcal{E}^{0+}(x)}$  can be chosen arbitrarily in  $[0, \text{Id}_{\mathcal{E}^{0+}(x)}]$ , while  $\Gamma_{\mathcal{E}^{-(x)}}$  is determined definition 6.1.12
- proposition 6.1.14 deals with Dini-stationary points, which in general are not solutions without regularity assumption.

The next result strengthens proposition 6.1.22, in the sense that  $\Gamma$  can now be chosen arbitrarily in  $[0, \text{Id}]$ , but the stronger regularity condition (6.24) is required instead of (6.23).

**Proposition 6.1.23** (NSC of  $\theta$ -stationarity for LCP - II). *For  $x \in \mathbb{R}^n$ , consider the following properties:*

- (i)  *$x$  is a solution to the LCP (6.19),*
- (ii) *for any  $\Gamma \in [0, \text{Id}]$ , the identity (6.22) holds,*
- (iii) *for some  $\Gamma \in [0, \text{Id}]$ , the identity (6.22) holds.*

*Then, (i)  $\Rightarrow$  (ii)  $\Rightarrow$  (iii). One also has the implication (iii)  $\Rightarrow$  (i), provided (6.24) holds.*

*Proof.* The proof of the implication (iii)  $\Rightarrow$  (i) is identical to the one of proposition 6.1.22, up to (6.25c), which is rewritten here after using the identity  $(Mx + q)_E = x_E$  (a consequence of the definition of  $E$ ):

$$\left( (\text{Id} - \Gamma) (M_{E,E} - M_{E,I} M_{I,I}^{-1} M_{I,E}) + \Gamma \right)^T x_E = 0.$$

By (6.24),  $M_{E,E} - M_{E,I} M_{I,I}^{-1} M_{I,E} \in \mathbf{P}$ . Now, any rowwise convex combination of a  $\mathbf{P}$ -matrix and  $\text{Id}$  is nonsingular [3, lemma 2.1]. Therefore, the matrix of the previous linear system is nonsingular. It results that  $x_E = 0$ . The rest of the proof is similar to the one of proposition 6.1.22.  $\square$

The R-regularity condition (6.24) depends on the considered point but we have said that it can be made uniform in  $x \in \mathbb{R}^n$  by assuming that  $M \in \mathbf{P}$ .

Let us clarify one meaning of the proposition, by focusing on the contrapositive of the implication (iii)  $\Rightarrow$  (i). This contrapositive tells us that when  $x$  is not a solution to the LCP (6.19), then  $g$  given by (6.22) is nonzero whatever  $\Gamma$  is in  $[0, \text{Id}]$ . We want to stress that this claim does not imply that, for any  $\Gamma \in [0, \text{Id}]$ ,  $-g$  is a descent direction of  $\theta$  at  $x$  since a phenomenon similar to the one encountered with a nonsmooth convex function can occur ( $g$  is a subgradient, but  $-g$  is not a descent direction [126, end of § VIII.1.1]), see example 6.1.4. To ensure  $-g$  is a descent direction, one possibility is to chose  $\gamma_{\varepsilon-(x)}$  as detailed in definition 6.1.12.

## 6.2 A considered algorithm

### 6.2.1 The method and its properties

This section details the properties one may obtain for the globalization of the PNM algorithm [72] regularized by a Levenberg-Marquardt approach. In this sequence, we do not assume  $S = I$  ( $S_k = I$ ), though  $S$  and  $S_k$  do not play a major role (in the theory). The

properties of the previous part all holds (essentially  $F'$  and  $G'$  are post-multiplied by  $S^{-1/2}$  which is well-defined).

Without regularity assumptions, the algorithm cannot do better than finding a (Clarke) stationary point of the least-squares function  $\theta$ ; hence it stops at such a point. If the current iterate  $x_k$  is not a stationary point of  $\theta$ , the algorithm determines a local model of  $\theta$ , of the form given by formula (6.10), namely

$$\varphi_k : d \in \mathbb{R}^n \mapsto \varphi_k(d) := \frac{1}{2} (\|w_k(d)\|_2^2 + \lambda d^\top S_k d).$$

We have denoted by  $w_k$  the function  $w(x_k, \cdot)$  which associates with  $d \in \mathbb{R}^n$  a vector  $w_k(d) \in \mathbb{R}^{n+|\mathcal{E}(x_k)|}$ , whose formula is the one of (6.9) with  $x \equiv x_k$

$$w_k(d) = \begin{bmatrix} F_{\mathcal{F}(x_k)}(x_k) + F'_{\mathcal{F}(x_k)}(x_k)d \\ G_{\mathcal{G}(x_k)}(x_k) + G'_{\mathcal{G}(x_k)}(x_k)d \\ (\Gamma_{\mathcal{E}^{0+}(x_k)})^{1/2} [H_{\mathcal{E}^{0+}(x_k)}(x_k) + F'_{\mathcal{E}^{0+}(x_k)}(x_k)d] \\ (\bar{\Gamma}_{\mathcal{E}^{0+}(x_k)})^{1/2} [H_{\mathcal{E}^{0+}(x_k)}(x_k) + G'_{\mathcal{E}^{0+}(x_k)}(x_k)d] \\ -(\Gamma_{\mathcal{E}^{-(x_k)}})^{1/2} [H_{\mathcal{E}^{-(x_k)}}(x_k) + F'_{\mathcal{E}^{-(x_k)}}(x_k)d]^- \\ -(\bar{\Gamma}_{\mathcal{E}^{-(x_k)}})^{1/2} [H_{\mathcal{E}^{-(x_k)}}(x_k) + G'_{\mathcal{E}^{-(x_k)}}(x_k)d]^- \end{bmatrix}$$

Note that, since  $F_i(x_k) = G_i(x_k)$  for  $i \in \mathcal{E}(x_k)$  and  $\gamma_k + \bar{\gamma}_k = 1$ , one has

$$\varphi_k(0) = \theta(x_k). \quad (6.27)$$

This model requires to define the weighting vector  $\gamma_k \in [0, 1]^{|\mathcal{E}(x_k)|}$  and this is done so that the gradient of  $\|w_k\|^2/2$  at zero does not vanish (as we have seen it in section 6.1.4, this is possible when  $x_k$  is not a Dini-stationary point of  $\theta$ ). The model also requires to define a matrix  $S_k \succ 0$  and a parameter  $\lambda_k$  known at the beginning of the iteration; some details are given in theorem 6.2.4 below.

Recall that  $\varphi_k$  is differentiable, so that a minimizer  $d_k$  of  $\varphi_k$  with  $\lambda = \lambda_k$  verifies  $\nabla \varphi_k(d_k) = 0$  or

$$\begin{aligned} & \left( F'_{\mathcal{F}(x_k)}(x_k)^\top F'_{\mathcal{F}(x_k)}(x_k) + G'_{\mathcal{G}(x_k)}(x_k)^\top G'_{\mathcal{G}(x_k)}(x_k) \right. \\ & + F'_{\mathcal{E}^{0+}(x_k)}(x_k)^\top (\Gamma_k)_{\mathcal{E}^{0+}(x_k)} F'_{\mathcal{E}^{0+}(x_k)}(x_k) + G'_{\mathcal{E}^{0+}(x_k)}(x_k)^\top (\bar{\Gamma}_k)_{\mathcal{E}^{0+}(x_k)} G'_{\mathcal{E}^{0+}(x_k)}(x_k) \Big) d_k \\ & - F'_{\mathcal{E}^{-(x_k)}}(x_k)^\top (\Gamma_k)_{\mathcal{E}^{-(x_k)}} \left( F'_{\mathcal{E}^{-(x_k)}}(x_k) d_k \right)^- \\ & - G'_{\mathcal{E}^{-(x_k)}}(x_k)^\top (\bar{\Gamma}_k)_{\mathcal{E}^{-(x_k)}} \left( G'_{\mathcal{E}^{-(x_k)}}(x_k) d_k \right)^- + \lambda_k S_k d_k = -g_k, \end{aligned} \quad (6.28)$$

where  $\Gamma_k := \text{Diag}(\gamma_k)$  and  $\bar{\Gamma}_k = \text{Id} - \Gamma_k$  and  $g_k := \nabla \varphi_k(0)$  is given by (see (6.11) in  $x \equiv x_k$ ):

$$\begin{aligned} g_k &= F'_{\mathcal{F}(x_k)}(x_k)^\top F_{\mathcal{F}(x_k)}(x_k) + G'_{\mathcal{G}(x_k)}(x_k)^\top G_{\mathcal{G}(x_k)}(x_k) \\ &+ [F'_{\mathcal{E}(x_k)}(x_k)^\top \Gamma_k + G'_{\mathcal{E}(x_k)}(x_k)^\top \bar{\Gamma}_k] H_{\mathcal{E}(x_k)}(x_k). \end{aligned} \quad (6.29)$$

Once the weights are chosen, with the penalty parameter  $\lambda_k \geq 0$ , the algorithm minimizes the model  $\varphi_k$

$$\min_{d \in \mathbb{R}^n} \varphi_k(d). \quad (6.30)$$

The computed solution depends on  $\lambda_k$  and is denoted  $d_k(\lambda_k)$  (it is uniquely determined if  $\lambda_k > 0$ ). The penalty parameter  $\lambda_k$  is (possibly) modified in order to get a sufficient decrease of  $\theta$ , in the sense that<sup>13</sup>

$$\theta(x_k + d_k(\lambda_k)) \leq \theta(x_k) + \eta_1 g_k^\top d_k(\lambda_k), \quad (6.31)$$

when this inequality is satisfied, the next iterate is set to  $x_{k+1} := x_k + d_k(\lambda_k)$ .

We now give a precise and formal description of the proposed LM-type algorithm for solving (6.1), on which the results given below are based. In the description of the algorithm, the term “constant” refers to a quantity that does not depend on the iteration.

**Algorithm 6.2.1** (LM-like algorithm). The algorithm uses the following constants:  $0 < \sigma_1 < 1 < \sigma_2$  for the update of  $\lambda_k$  and  $0 < \eta_1 < \eta_2 < 1$  as satisfaction thresholds for the decrease of  $\theta$ . One assume that  $S_k \succ 0$ . One iteration of the algorithm, computing  $(x_{k+1}, \lambda_{k+1}, S_{k+1})$  from  $(x_k, \lambda_k, S_k)$ , proceeds as follows.

1. *Stopping test and weights.* Stop if  $x_k$  is a stationary point of  $\theta$ . Otherwise, set  $\Gamma_k \in [0, \text{Id}]$  such that the gradient  $g_k$  given by (6.29) is nonzero.
2. *Displacement.* Set  $\lambda_{k,0} := \lambda_k$  and repeat the following operations indexed by  $i \in \mathbb{N}$ , up to satisfaction (6.31).
  - 2.1. Compute a solution  $d_{k,i}$  to (6.30).
  - 2.2. If

$$\theta(x_k + d_{k,i}) \leq \theta(x_k) + \eta_1 g_k^\top d_{k,i}, \quad (6.32)$$

holds, exit the present loop with  $d_k := d_{k,i}$  (go to step 3), otherwise  $\lambda_{k,i+1} = \sigma_2 \lambda_{k,i}$  and repeat the loop (go to step 2.1).

3. *New penalty parameter.* If

$$\theta(x_k + g_k) \leq \theta(x_k) + \eta_2 g_k^\top d_k, \quad (6.33)$$

then  $\lambda_{k+1} := \sigma_1 \lambda_{k,i}$ , else  $\lambda_{k+1} := \lambda_{k,i}$ .

4. *New iterate.* Update  $x_{k+1} := x_k + d_k$ .
5. *New scaling matrix.* Choose  $S_{k+1} \succ 0$ .

**Remarks 6.2.2.** 1) It is assumed that the stationarity test can be solved at each iteration; this can be achieved by enumeration if  $|\mathcal{E}(x_k)|$  is small or under the hypotheses of proposition 6.1.18 for instance.

---

<sup>13</sup>Using  $\theta(x_k + d_k(\lambda_k)) \leq \theta(x_k) + \eta_1 \theta'(x_k; d_k(\lambda_k))$  was considered as well, but without major differences arising.

- 2) The computing cost of this algorithm is mainly linked to the number of optimization problems (6.30) to solve, possibly several at each iteration.

An algorithm with quadratic subproblems at each iteration can for instance be found in [86, §9.2, algorithm 9.2.2]. Similar considerations appear in [17], where a differentiable strongly convex piecewise quadratic function must be minimized without constraints, and for which a specific semismooth Newton with exact linesearch can be used. Point 4 of the next proposition has the consequence that the loop in step 2 of algorithm 6.2.1 is processed a finite number of times at each iteration. The next proposition adapts usual Levenberg-Marquardt properties to the current setting.

**Proposition 6.2.3** (sufficient decrease). *Suppose that the gradient given by (6.29)  $g_k \neq 0$ , that  $S_k \succ 0$  and that  $\eta_1 < 1$ . Denote by  $d_k(\lambda)$  a solution to (6.30). Then,*

- 1)  $d_k(\lambda) \neq 0$  for all  $\lambda \geq 0$ ,
- 2)  $d_k(\lambda) \rightarrow 0$  when  $\lambda \rightarrow +\infty$ ,
- 3)  $d_k(\lambda)/\|d_k(\lambda)\| \rightarrow -S_k^{-1}g_k/\|S_k^{-1}g_k\|$  when  $\lambda \rightarrow \infty$ ,
- 4) the sufficient decrease condition (6.31) holds for all  $\lambda$  sufficiently large.

*Proof.* 1) We proceed by contradiction. If  $d_k(\lambda) = 0$ , one would have  $\nabla\varphi_k(0) = 0$ , since  $d_k(\lambda)$  solves problem (6.30) and  $\varphi_k$  is differentiable. Equivalently, by the definition of  $g_k$  in (6.29), one would have  $g_k = 0$ , which has been supposed not to occur by assumption. Therefore  $d_k(\lambda) \neq 0$ .

2) The convergence of  $d_k(\lambda)$  to zero is due to the fact that  $d_k(\lambda)$  minimizes  $\varphi_k(\cdot)$ , so that

$$0 \leq \frac{\lambda}{2} d_k(\lambda)^T S_k d_k(\lambda) \leq \varphi_k(d_k(\lambda)) \leq \varphi_k(0) = \theta(x_k),$$

by (6.27). Dividing by  $\lambda$  and taking the limit when  $\lambda \rightarrow +\infty$  show that  $d_k(\lambda)^T S_k d_k(\lambda) \rightarrow 0$ . Since  $S_k \succ 0$ , it results that  $d_k(\lambda) \rightarrow 0$ .

3) Since  $d_k(\lambda)$  minimizes  $\varphi_k(\cdot)$  and  $\varphi_k$  is differentiable, one has  $\nabla\varphi_k(d_k(\lambda)) = 0$ . Then, taking the limit when  $\lambda \rightarrow +\infty$  in the equation  $\nabla\varphi_k(d_k(\lambda)) = 0$  and using point 2 yields

$$\lambda S_k d_k(\lambda) \rightarrow -g_k \quad \text{or equivalently} \quad \lambda d_k(\lambda) \rightarrow -S_k^{-1}g_k,$$

where  $g_k$  is defined by (6.29). Now, since  $\lambda d_k(\lambda) \neq 0$  by point 1, the claim in point 3 follows.

4) We proceed by contradiction. Suppose that (6.31) does not hold for a sequence of  $\lambda \rightarrow +\infty$ . Then, for these  $\lambda \rightarrow +\infty$ , one has

$$\frac{\theta(x_k + d_k(\lambda)) - \theta(x_k) - g_k^T d_k(\lambda)}{\|d_k(\lambda)\|} > (1 - \eta_1) \frac{-g_k^T d_k(\lambda)}{\|d_k(\lambda)\|}.$$

By point 3, the right-hand side tends to  $(1 - \eta_1)g_k^T S_k^{-1}g_k/\|S_k^{-1}g_k\|$ , which is positive by the positive definiteness of  $S_k$ ,  $g_k \neq 0$  and  $\eta_1 < 1$ . We claim that the left-hand side tends to a nonpositive number, which will yield the expected contradiction.

Recall that  $\theta$  is directionally differentiable. Moreover,  $\theta$  is also Lipschitz continuous, so that  $\theta'(x; \cdot)$  is also Lipschitz continuous and  $\theta$  is also directionally differentiable in the sense of Hadamard (see for example [230] and the references therein), meaning that

$$\theta'(x; d) = \lim_{\substack{t \downarrow 0 \\ d' \rightarrow d}} \frac{\theta(x + td') - \theta(x)}{t}.$$

Therefore, setting  $t := \|d_k(\lambda)\|$ , with  $t \rightarrow 0$  when  $\lambda \rightarrow \infty$ , and  $d' := d_k(\lambda)/\|d_k(\lambda)\|$ , which tends to  $-S_k^{-1}g_k/\|S_k^{-1}g_k\|$  by point 3, one deduces from the previous exposed formula that

$$\lim_{\lambda \rightarrow +\infty} \frac{\theta(x_k + d_k(\lambda)) - \theta(x_k)}{\|d_k(\lambda)\|} = \theta'(x_k; -S_k^{-1}g_k/\|S_k^{-1}g_k\|).$$

Therefore,

$$\lim_{\lambda \rightarrow \infty} \frac{\theta(x_k + d_k(\lambda)) - \theta(x_k) - g_k^T d_k(\lambda)}{\|d_k(\lambda)\|} = \frac{\theta'(x_k; -S_k^{-1}g_k) + g_k S_k^{-1} g_k}{\|S_k^{-1}g_k\|}.$$

By definition 6.1.12, the right-hand side is nonpositive.  $\square$

The following result requires the boundedness of some quantities generated by the algorithm, the product  $\lambda S$ , as well as  $F'$  and  $G'$ , the hypothesis on  $F$  and  $G$  is rather innocuous. It also assumes that the algorithm generates a sequence, so that it never finds a point  $x_k$  such that  $g_k = 0$ , which is a favorable case since then the algorithm stops after a finite number of iterations.

**Theorem 6.2.4** (convergence of the LM algorithm). *Let  $\{(x_k, \lambda_k)\}$  be a sequence generated by algorithm 6.2.1. Define  $g_k$  by (6.29). Then,*

- 1)  $\{\theta(x_k)\}$  converges,
- 2) for any subsequence  $\mathcal{K}$  of  $\mathbb{N}$  such that  $\{(F'(x_k), G'(x_k), \lambda_k S_k)\}_{k \in \mathcal{K}}$  is bounded, one has that  $\{g_k\}_{k \in \mathcal{K}} \rightarrow 0$  when  $k \rightarrow +\infty$  in  $\mathcal{K}$ .

*Proof.* 1) The convergence of the sequence  $\{\theta(x_k)\}$  follows from its decrease and its boundedness from below (by zero).

2) From the sufficient decrease condition (6.31) and the convergence of  $\theta(x_k)$ , it follows that

$$g_k^T d_k \rightarrow 0, \quad \text{when } k \rightarrow \infty. \quad (6.34)$$

It is now a question of showing that the convergence in (6.34) is due to a gradient  $g_k$  that tends to zero and not to the convergence of  $d_k$  to zero (a fact that is probably true as well). Taking the scalar product of both sides of (6.28) with  $d_k$  yields

$$\begin{aligned} -g_k^T d_k &= \|F'_{\mathcal{F}(x_k)}(x_k)d_k\|^2 + \|G'_{\mathcal{G}(x_k)}(x_k)d_k\|^2 \\ &\quad + \|(\Gamma_k)_{\mathcal{E}^{0+}(x_k)}^{1/2} F'_{\mathcal{E}^{0+}(x_k)}(x_k)d_k\|^2 + \|(\bar{\Gamma}_k)_{\mathcal{E}^{0+}(x_k)}^{1/2} G'_{\mathcal{E}^{0+}(x_k)}(x_k)d_k\|^2 \\ &\quad + \|(\Gamma_k)_{\mathcal{E}^{-(x_k)}}^{1/2} [F'_{\mathcal{E}^{-(x_k)}}(x_k)d_k]^- \|^2 + \|(\bar{\Gamma}_k)_{\mathcal{E}^{-(x_k)}}^{1/2} [G'_{\mathcal{E}^{-(x_k)}}(x_k)d_k]^- \|^2 \\ &\quad + \lambda_k S_k d_k. \end{aligned}$$

One deduces now from (6.34) and the positivity of  $\lambda_k$  that when  $k \rightarrow \infty$ :

$$\begin{aligned} F'_{\mathcal{F}(x_k)}(x_k)d_k &\rightarrow 0, & G'_{\mathcal{G}(x_k)}(x_k)d_k &\rightarrow 0, \\ (\Gamma_k)_{\mathcal{E}^{0+}(x_k)}^{1/2} F'_{\mathcal{E}^{0+}(x_k)}(x_k)d_k &\rightarrow 0, & (\bar{\Gamma}_k)_{\mathcal{E}^{0+}(x_k)}^{1/2} G'_{\mathcal{E}^{0+}(x_k)}(x_k)d_k &\rightarrow 0, \\ (\Gamma_k)_{\mathcal{E}^-(x_k)}^{1/2} [F'_{\mathcal{E}^-(x_k)}(x_k)d_k]^- &\rightarrow 0, & (\bar{\Gamma}_k)_{\mathcal{E}^-(x_k)}^{1/2} [G'_{\mathcal{E}^-(x_k)}(x_k)d_k]^- &\rightarrow 0, \\ \lambda_k^{1/2} S_k^{1/2} d_k &\rightarrow 0. \end{aligned}$$

Now, observing that  $\{\Gamma_k, \bar{\Gamma}_k\}$  is bounded, we see that, if the sequence  $\{(F'(x_k), G'(x_k), \lambda_k S_k)\}_{k \in \mathcal{K}}$  is bounded, one has when  $k \rightarrow \infty$  in  $\mathcal{K}$ :

$$\begin{aligned} F'_{\mathcal{F}(x_k)}(x_k)^\top F'_{\mathcal{F}(x_k)}(x_k)d_k &\rightarrow 0, \\ G'_{\mathcal{G}(x_k)}(x_k)^\top G'_{\mathcal{G}(x_k)}(x_k)d_k &\rightarrow 0, \\ F'_{\mathcal{E}^{0+}(x_k)}(x_k)^\top (\Gamma_k)_{\mathcal{E}^{0+}(x_k)} F'_{\mathcal{E}^{0+}(x_k)}(x_k)d_k &\rightarrow 0, \\ G'_{\mathcal{E}^{0+}(x_k)}(x_k)^\top (\bar{\Gamma}_k)_{\mathcal{E}^{0+}(x_k)} G'_{\mathcal{E}^{0+}(x_k)}(x_k)d_k &\rightarrow 0, \\ F'_{\mathcal{E}^-(x_k)}(x_k)^\top (\Gamma_k)_{\mathcal{E}^-(x_k)} [F'_{\mathcal{E}^-(x_k)}(x_k)d_k]^- &\rightarrow 0, \\ G'_{\mathcal{E}^-(x_k)}(x_k)^\top (\bar{\Gamma}_k)_{\mathcal{E}^-(x_k)} [G'_{\mathcal{E}^-(x_k)}(x_k)d_k]^- &\rightarrow 0, \\ \lambda_k S_k d_k &\rightarrow 0. \end{aligned}$$

The definition (6.28) of the iteration now implies that  $g_k \rightarrow 0$  when  $k \rightarrow \infty$  in  $\mathcal{K}$ .  $\square$

Observe that without additional work, the algorithm may be blocked at points that are not necessarily satisfactory. The hypothesis on  $\lambda$  being bounded is relatively strong, it supposes that the Levenberg-Marquardt part does not make too small displacements. Obtaining a stronger version of the theorem without such hypothesis is an interesting perspective.

**Counter-example 6.2.5** (accumulation point is not Dini-stationary). Consider figure 6.4 and the associated problem, where  $F(x) \equiv x$ ,  $G(x) \equiv 1 + (x - 1)^2$ . If  $\lambda S$  is large enough, i.e., the directions are small enough, one can justify the algorithm never reaches the set  $[0, 1]$ . Thus, the framework developed with the weights  $\gamma$  never intervenes.  $\square$

Let us mention that, by proposition 6.1.20, the accumulation point is (Clarke) stationary.

## 6.2.2 Modifications and potential improvements

Now that the algorithm and its properties have been studied, we focus on possible improvements that could be brought to it and are interesting extensions for future work.

### Choice of the weights

Let us recall that despite the stationarity detection having nonpolynomial complexity in the general case (propositions 6.1.14 and 6.1.15 alongside theorem 6.1.16), in the setting of

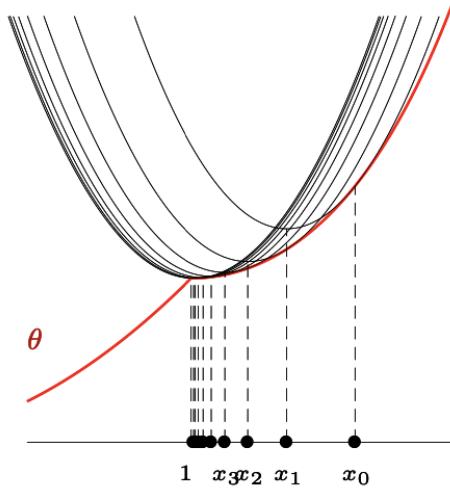


Figure 6.5: The curves above  $\theta$  are the quadratic models  $\varphi_{x_k}$ . While there may be fast convergence to  $x = 1$ , it may never be reached ( $x_k > 1$ ) so  $\forall k, \mathcal{E}(x_k) = \emptyset$ .

this chapter it may often be relatively easy to solve, for instance in the following cases (see appendix C for the details):

- the lines of  $G'_{\mathcal{E}^-(x)}(x) - F'_{\mathcal{E}^-(x)}(x)$  are independent,
- $\mathcal{E}^-(x) = \emptyset$  (no projection),
- $\mathcal{E}^{0+}(x) = \emptyset$  (just one point to project),
- $\mathcal{R}([(G'_{\mathcal{E}^{0+}(x)}(x) - F'_{\mathcal{E}^{0+}(x)}(x))^T \bar{x} - \bar{y}]) \not\subseteq \mathcal{R}((G'_{\mathcal{E}^{0+}(x)}(x) - F'_{\mathcal{E}^{0+}(x)}(x))^T)$  (in short, one dimension of the first zonotope cannot be generated by the second, so it is clearly not contained by it).

All these cases can be verified and solved easily. In the general case, especially “far” from the solution, i.e., in the early stages of the algorithm, one could think about adding some randomization aspect to the choice of the weights (as it is done in [19]). For instance, a few arbitrarily chosen values of weights could be tested to see if a descent direction can be obtained. One could also think at changing only weights corresponding to indices  $i$  such that  $i$  is not in the same set index set.

### Alternative to the choice of the weights

To avoid the complicated and costly iteration of Levenberg-Marquardt, one possibility is to use hybridization schemes, which, as described in section 2.3.3, consist in using a simply computed direction instead of the complete iteration. In our case, one could use a simpler

expression the model  $\varphi_x$  from (6.10) by using

$$w_k(d) = \begin{bmatrix} F_{\mathcal{F}(x_k)}(x_k) + F'_{\mathcal{F}(x_k)}(x_k)d \\ G_{\mathcal{G}(x_k)}(x_k) + G'_{\mathcal{G}(x_k)}(x_k)d \\ \Gamma^{1/2}[F_{\mathcal{E}(x_k)}(x_k) + F'_{\mathcal{E}(x_k)}(x_k)d] \\ \bar{\Gamma}^{1/2}[G_{\mathcal{E}(x_k)}(x_k) + G'_{\mathcal{E}(x_k)}(x_k)d] \end{bmatrix}.$$

One could also use a possible Newton-min direction (recall that this is equivalent to choosing  $\gamma \in \{0, 1\}^{\mathcal{E}(x)}$ ), which has very low cost. Several methods evoked in section 2.3.3 use this trick with satisfying success.

In the aforementioned algorithms, this would mean replacing step 3 of the PNM algorithm 6.1.2 by the following step.

**Algorithm 6.2.6** (Hybrid PNM [72, algorithm 3.8]). 3.1 For some partition  $\tilde{\mathcal{F}}(x)$ ,  $\tilde{\mathcal{G}}(x)$  of  $[1 : n]$  that satisfies  $\tilde{\mathcal{F}}(x) \supseteq \mathcal{F}(x)$  and  $\tilde{\mathcal{G}}(x) \supseteq \mathcal{G}(x)$ , compute a Newton-min direction  $d^{\text{NM}}$  as a solution to

$$\begin{cases} F_i(x) + F'_i(x)d = 0 & \text{if } i \in \tilde{\mathcal{F}}(x), \\ G_i(x) + G'_i(x)d = 0 & \text{if } i \in \tilde{\mathcal{G}}(x). \end{cases}$$

3.2 *Decrease test.* If (6.8) is verified with  $d$  and  $\alpha = 1$ , go to step 5. Otherwise, compute  $d$  as given by (6.7).

Similarly, we replace steps 1 and 2 of the LM algorithm 6.2.1 by the following step.

**Algorithm 6.2.7** (Hybrid LM). 1.1 For some partition  $\tilde{\mathcal{F}}(x)$ ,  $\tilde{\mathcal{G}}(x)$  of  $[1 : n]$  that satisfies  $\tilde{\mathcal{F}}(x) \supseteq \mathcal{F}(x)$  and  $\tilde{\mathcal{G}}(x) \supseteq \mathcal{G}(x)$ , compute a Newton-min direction  $d^{\text{NM}}$  as a solution to

$$\begin{cases} F_i(x) + F'_i(x)d = 0 & \text{if } i \in \tilde{\mathcal{F}}(x), \\ G_i(x) + G'_i(x)d = 0 & \text{if } i \in \tilde{\mathcal{G}}(x). \end{cases}$$

1.2 *Decrease test.* If (6.32) is verified with  $d$ , go to step 4. Otherwise, compute  $d$  as given by steps 1 and 2 of algorithm 6.2.1.

## Tolerance and numerical precision

As it was discussed in counterexample 6.2.5, such “concave kinks” that  $\theta$  may have are issues not necessarily dealt with by the algorithm. Furthermore, especially for the numerical aspect, defining the set  $\mathcal{E}$  by  $F_i(x) = G_i(x)$  is not satisfying. This justifies the need to introduce a tolerance  $\tau > 0$  such that the equality indices are defined by  $|F_i(x) - G_i(x)| < \tau$ . This tolerance is already presented in [72].

With such tolerance in mind, counterexample 6.2.5 is unlikely to be problematic: for  $k$  large enough,  $x_k$  shall be close enough to 1, meaning that instead of having the index set be  $\mathcal{G}$ , it will be  $\mathcal{E}(x)^\tau$ : the other piece (corresponding to  $x < 1$ ) will be considered and, for  $\gamma$

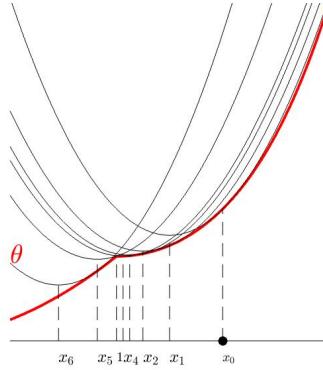


Figure 6.6: Illustration of a few iterates, for some  $\tau > 0$ , of algorithm 6.2.1 using  $\mathcal{E}(x) := \{i \in [1 : n] : |F_i(x) - G_i(x)| < \tau\}$ .

not too close to 0,  $\gamma F'(x_k) + (1 - \gamma)G'(x_k) = \gamma + (1 - \gamma) \times 2(x_k - 1)$  will be large enough to pass “over” the kink at 1, as schematized in the following picture.

To conclude, the work in progress around this algorithm could clearly benefit from additional tuning techniques, some of which were discussed in chapter 2. In particular, studying and observing the impact of the choice of the weights seems to be an interesting track of what was presented.

## Historic

Let us finish by a brief comment on lemma 6.1.6. We initially found a rather involved and convoluted reasoning to prove the existence (but not really in a constructive way) of the “solution”  $\gamma_{\mathcal{E}^-(x)}(\gamma_{\mathcal{E}^{0+}(x)})$ . However, the understanding of the structure of the minimum function (see chapter 3) seems to provide a simpler (and somewhat constructive) approach.



# Conclusion and perspectives

In this thesis, we have discussed various elements surrounding the use of the minimum C-function for complementarity problems. Local and global methods arising from the non-smooth system induced by this C-function may lead to strong geometric considerations, which are a result of the intrinsic structure of the minimum.

This was initiated in chapter 6, where the research of a descent direction (or the validation of stationarity) was shown to be a NP-complete problem, due to the combinatorial nature of the minimum. The associated algorithm, while it technically ensures one reaches a Clarke-stationary point, still needs to be tested on classic problems.

To justify these properties, the geometry of zonotopes, particular symmetric polytopes, was essential. This is highly related to chapter 3 ([77]), where it was shown that some elements of the Bouligand differential of the minimum C-function are related to vertices of specific zonotopes and thus to hyperplane arrangements.

For that particular problem, we proposed some improvements on a state of the art algorithm, the most promising ones being based on interesting heuristics such as the use of duality, via the circuits of the underlying matroid. These improvements were further used in chapter 5, for general arrangements – not ones necessarily centered ones related to the application of the minimum on a CP. Further testing of these heuristics on the computation of the full arrangement, as well as their combination with other methods based on combinatorial symmetries, would surely lead to interesting observations.



# Appendix A

## Detailed information on affine hyperplane arrangements: theory, numerics and complements

This appendix completes chapter 5, by detailing some proofs and giving some details, as it was done in section 4.5, on some numerical values observed.

### A.1 Details on properties of chapter 5

$$(5.9) \quad \mathcal{S}(V, -\tau) = -\mathcal{S}(V, \tau).$$

Indeed, if  $s \in \mathcal{S}(V, -\tau)$ ,  $s \cdot (V^\top x + \tau) > 0$  for some  $x \in \mathbb{R}^n$ . Therefore,  $-s \cdot (V^\top (-x) - \tau) > 0$ , showing that  $-s \in \mathcal{S}(V, \tau)$  or  $s \in -\mathcal{S}(V, \tau)$ . We have shown the inclusion  $\mathcal{S}(V, -\tau) \subseteq -\mathcal{S}(V, \tau)$ . By changing  $\tau$  into  $-\tau$ , one gets  $\mathcal{S}(V, \tau) \subseteq -\mathcal{S}(V, -\tau) \subseteq \mathcal{S}(V, \tau)$ , hence (5.9).

$$(5.11) \quad \mathcal{S}_s(V, -\tau) = -\mathcal{S}_s(V, \tau) = \mathcal{S}_s(V, \tau) \quad \text{and} \quad \mathcal{S}_a(V, -\tau) = -\mathcal{S}_a(V, \tau).$$

Indeed,

$$\begin{aligned} \mathcal{S}_s(V, -\tau) &= \mathcal{S}(V, -\tau) \cap \mathcal{S}(V, \tau) = \mathcal{S}_s(V, \tau), \\ -\mathcal{S}_s(V, \tau) &= [-\mathcal{S}(V, \tau)] \cap [-\mathcal{S}(V, -\tau)] = \mathcal{S}(V, -\tau) \cap \mathcal{S}(V, \tau) = \mathcal{S}_s(V, \tau), \\ \mathcal{S}_a(V, -\tau) &= \mathcal{S}(V, -\tau) \setminus \mathcal{S}_s(V, -\tau) = [-\mathcal{S}(V, \tau)] \setminus [-\mathcal{S}_s(V, \tau)] = -\mathcal{S}_a(V, \tau). \end{aligned}$$

**Proposition A.1.1** (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  $l := \sum_{i \in [1:p]} |\tilde{s}_i - s_i|/2 \leq p$  and there is no path in  $\mathcal{S}(V, \tau)$  joining  $s$  and  $\tilde{s}$  of smaller length.*

---

*Proof.* The fact that any path joining  $s$  and  $\tilde{s}$  in  $\mathcal{S} \equiv \mathcal{S}(V, \tau)$  is of length  $\geq l$  is due to the fact that  $s$  and  $\tilde{s}$  have  $l$  different components and that two adjacent sign vectors differ by a single component.

[ $\Rightarrow$ ] We prove the contrapositive. Suppose that, for some  $i \neq j$  in  $[1 : p]$ , the hyperplanes  $H_i$  and  $H_j$  are identical. Then, by proposition 5.3.2(2), the nonzero pairs  $(v_i, \tau_i)$  and  $(v_j, \tau_j)$  are colinear in  $\mathbb{R}^n \times \mathbb{R}$ :  $(v_j, \tau_j) = \alpha(v_i, \tau_i)$ , for some  $\alpha \in \mathbb{R}^*$ . Assume that  $\alpha > 0$  (resp.  $\alpha < 0$ ). For any  $\tilde{s} \in \mathcal{S}$ , there is an  $\tilde{x} \in \mathbb{R}^n$  such that  $\tilde{s} \cdot (V^\top \tilde{x} - \tau) > 0$ , which implies that one must have  $\tilde{s}_i = \tilde{s}_j$  (resp.  $\tilde{s}_i = -\tilde{s}_j$ ). Take any  $s \in \mathcal{S}(V, 0)$  ( $\neq \emptyset$  by (5.12)), so that  $-s \in \mathcal{S}(V, 0)$  by the symmetry of  $\mathcal{S}(V, 0)$ , asserted in (5.7). We claim that one cannot find a path in  $\mathcal{S}$  joining  $s \in \mathcal{S}$  and  $-s \in \mathcal{S}$ . Indeed, all the components of  $s$  must change their sign. Now, the components  $i$  and  $j$  of any sign vector  $\tilde{s}$  on such a path are necessarily equal (resp. opposite), so that they would change simultaneously, while adjacency imposes to change only a single sign between two consecutive sign vectors of a path.

[ $\Leftarrow$ ] Let  $s$  and  $\tilde{s} \in \mathcal{S}$ , which are assumed to be distinct (otherwise the result is straightforward). One has to show that there is a path of the given length  $l$  in  $\mathcal{S}$  joining  $s$  to  $\tilde{s}$ . By the definition of  $\mathcal{S}$ , one can find  $x$  and  $\tilde{x}' \in \mathbb{R}^n$  such that

$$s \cdot (V^\top x - \tau) > 0 \quad \text{and} \quad \tilde{s} \cdot (V^\top \tilde{x}' - \tau) > 0.$$

The desired path in  $\mathcal{S}$  is determined by the sign vectors of the chambers, given by  $\phi$  in (5.6), that are visited along the segment joining  $x$  to a small modification  $\tilde{x}$  of  $\tilde{x}'$ . The modification is introduced so that  $\tilde{x}$  of  $\tilde{x}'$  belong to the same chamber and the segment joining them does not cross two or more hyperplanes simultaneously.

Here is how the modification  $\tilde{x}$  of  $\tilde{x}'$  is obtained. Let  $d' := \tilde{x}' - x$ , which is nonzero, since  $s \neq \tilde{s}$ . By proposition 5.3.2(2), since the hyperplanes of the arrangement are all different, the vectors  $\{(v_i, \tau_i) \in \mathbb{R}^n \times \mathbb{R} : i \in [1 : p]\}$  are not colinear, so that the vectors  $\{v_i/(v_i^\top x - \tau_i) \in \mathbb{R}^n : i \in [1 : p]\}$  are distinct<sup>1</sup>. Now, for a  $d$  arbitrary close to  $d'$ , hence for an

$$\tilde{x} := x + d \tag{A.1a}$$

arbitrary close to  $\tilde{x}'$ , one can guarantee the inequality  $\tilde{s} \cdot (V^\top \tilde{x} - \tau) > 0$  and, by lemma 3.2.6,

$$|\{(v_i^\top d)/(v_i^\top x - \tau_i) : i \in [1 : p]\}| = p. \tag{A.1b}$$

Since, when applying lemma 3.2.6, one could have added  $v_0 = 0$  to the list of distinct nonzero vectors  $v_i/(v_i^\top x - \tau_i)$ 's, one can also guarantee that

$$v_i^\top d \neq 0, \quad \forall i \in [1 : p]. \tag{A.1c}$$

---

<sup>1</sup>Note first that  $v_i^\top x \neq \tau_i$  for all  $i \in [1 : p]$ , since  $x$  does not belong to a hyperplane, so that the vectors  $v_i/(v_i^\top x - \tau_i)$  are well defined. Now, if one would have  $v_i/(v_i^\top x - \tau_i) = v_j/(v_j^\top x - \tau_j)$  for some  $i \neq j$ , it would follow that  $v_i = \alpha v_j$ , for  $\alpha := (v_i^\top x - \tau_i)/(v_j^\top x - \tau_j)$ . Then,  $v_i^\top x - \tau_i = \alpha(v_j^\top x - \tau_j)$  or  $\tau_i = \alpha\tau_j$ . This would imply that  $(v_i, \tau_i)$  and  $(v_j, \tau_j)$  would be colinear in  $\mathbb{R}^n \times \mathbb{R}$ .

To determine the sign vectors of the chambers that are visited along the path  $t \mapsto x + td$  in  $\mathbb{R}^n$ , we first determine the  $t_i$ 's at which this path encounters a hyperplane. By (A.1c), one can define  $t_i := -(v_i^\top x - \tau_i)/(v_i^\top d)$ , for  $i \in [1 : p]$ , which are  $p$  distinct values by (A.1b). It results that the following equivalent expressions hold for all  $i \in [1 : p]$ :  $(v_i^\top x - \tau_i) + t_i(v_i^\top d) = 0$  or, using (A.1a),

$$(1 - t_i)(v_i^\top x - \tau_i) + t_i(v_i^\top \tilde{x} - \tau_i) = 0 \quad \text{or} \quad v_i^\top[(1 - t_i)x + t_i\tilde{x}] - \tau_i = 0. \quad (\text{A.1d})$$

By the last expression in (A.1d), the point  $z^i := (1 - t_i)x + t_i\tilde{x} = x + t_id$  belongs to the  $i$ th hyperplane, as announced. Now, by the first expression in (A.1d),  $t_i \in (0, 1)$  (i.e.,  $z^i$  is in the relative interior of  $[x, \tilde{x}]$ ) if and only if  $(v_i^\top x - \tau_i)$  and  $(v_i^\top \tilde{x} - \tau_i)$  have opposite signs, which also reads  $s_i \tilde{s}_i = -1$ . Therefore, the number of  $t_i$ 's in  $(0, 1)$  is equal to  $l = \sum_{i \in [1:p]} |\tilde{s}_i - s_i|/2 \leq p$ . Let us denote them by

$$0 < t_{i_1} < \dots < t_{i_l} < 1.$$

By definition of the  $t_i$ 's, for  $t \in (t_{i_j}, t_{i_{j+1}})$ , the sign vector  $s^{ij} := \text{sgn}(V^\top[(1 - t)x + t\tilde{x}] - \tau)$  is constant, which also reads

$$s^{ij} \cdot (V^\top[(1 - t)x + t\tilde{x}] - \tau) > 0, \quad \text{for } t \in (t_{i_j}, t_{i_{j+1}}).$$

Furthermore, when  $t \in (0, 1)$  crosses a  $t_{i_j} \in (0, 1)$ , a single component of  $V^\top[(1 - t)x + t\tilde{x}] - \tau$  changes its sign. Therefore, we have defined a path of length  $l \leq p$  in  $\mathcal{S}$ , namely  $s^{i_0} = s, s^{i_1}, \dots, s^{i_l} = \tilde{s}$ , joining  $s$  to  $\tilde{s}$ . This proves the implication.  $\square$

$$(5.17) \quad -\mathfrak{S}(V, \tau) = \mathfrak{S}(V, -\tau) \quad \text{and} \quad -\mathfrak{S}_a(V, \tau) = \mathfrak{S}_a(V, -\tau).$$

*Proof.* Indeed, let  $\sigma \in \mathfrak{S}(V, \tau)$  with  $J = \mathfrak{J}(\sigma)$ . Then,  $J \in \mathcal{C}(V)$  and  $\sigma = \text{sgn}(\eta)$  for some  $\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$  satisfying  $\tau_J^\top \eta \geq 0$ . Since  $J \in \mathcal{C}(V)$  and  $-\sigma = \text{sgn}(-\eta)$  for some  $-\eta \in \mathcal{N}(V_{:,J}) \setminus \{0\}$  satisfying  $(-\tau_J)^\top (-\eta) \geq 0$ , it follows that  $-\sigma \in \mathfrak{S}(V, -\tau)$ . We have shown that  $-\mathfrak{S}(V, \tau) \subseteq \mathfrak{S}(V, -\tau)$ . Changing  $\tau$  into  $-\tau$  and using (5.16) yield  $\mathfrak{S}(V, -\tau) \subseteq -\mathfrak{S}(V, \tau)$ . The desired identity  $-\mathfrak{S}(V, \tau) = \mathfrak{S}(V, -\tau)$  follows.

One proceeds similarly to show that  $-\mathfrak{S}_s(V, \tau) = \mathfrak{S}_s(V, -\tau)$ , by changing the inequalities “ $\geq$ ” into equalities (actually, proposition 5.3.14(1) below will show that  $-\mathfrak{S}_s(V, \tau) = \mathfrak{S}_s(V, \tau)$ ). Now, the last formula in (5.17) is deduced directly from the definition  $\mathfrak{S}_a(V, \tau) := \mathfrak{S}(V, \tau) \setminus \mathfrak{S}_s(V, \tau)$ .  $\square$

The coning, related to section 5.3.4, may be viewed as follows. If one uses the matrix  $\mathcal{V}_t := [V \ 0; \tau^\top \ t]$  for some scalar  $t$ , one has

$$\mathcal{S}(\mathcal{V}_t) = [\mathcal{S}(V, \tau) \times \{\text{sgn}(-t)\}] \cup [-\mathcal{S}(V, \tau) \times \{\text{sgn}(+t)\}],$$

and clearly taking half of this linear arrangement yields  $\mathcal{S}(V, \tau)$  (up to  $\text{sgn}(-t)$ ). Very similar observations hold for the stem vectors.

---

**Remark A.1.2** ( $D_2$  is a combinatorial heuristic). The strategy discussed in sections 3.5.2. $D_2$  and 5.5.2 (the primal-dual method) has some similarities with heuristics developed for other combinatorial problems such as the SAT problem with constraint learning.  $\square$

Indeed,  $D_2$  can be understood as follows: when arriving at an infeasible leaf of the tree, a stem vector is obtained. A stem vector, by definition, is an infeasible combination of signs of minimal size. Therefore, learning this minimal infeasibility means it and can be reused in other subtrees to prevent further exploration at a low cost. The same notion is used in constraint learning for the clause satisfiability problem SAT, see for instance [201].

## A.2 Instance values

The instance types are defined in section 5.7.1. The following values are to be expected from the **RAND**, **2D** and **PERM** instances. For the **SRAND** and **RATIO** instances, one cannot find precise formulas.

Problems	chambers			circuits / stem vectors		
	$ \mathcal{S}(V, \tau) $	$ \mathfrak{S}_s(V, \tau) /2$	$ \mathfrak{S}_a(V, \tau) $	$ \mathfrak{S}([V; \tau^T], 0) /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	$(n+1)!$	$\sum_{i=3}^{n+1} \frac{i!}{2^i} \binom{n+1}{i}$	0	$\sum_{i=3}^{n+1} \frac{i!}{2^i} \binom{n+1}{i}$		

Table A.1: Known cardinalities for some of the instances. The values for the **RAND-N-P** and **2D-N-P** problems are obtained via affine general position, thus propositions 5.3.31, 3.4.6 and remark 5.3.13 6). Recall that the symmetric stem vectors are counted in pairs (thus the factor 1/2) – the number of circuits does *not* take this factor into account.

Let us justify these values. For the **RAND** instances, since they are randomly generated general position holds, so one only has to use the upper bounds. For the **2D** instances, one may proceed as follows. By independence of the first  $n-2$  vectors, one can decompose the sign vectors (of size  $p$ ):

$$\begin{aligned}
s \in \mathcal{S}(V, \tau) &\iff \exists x \text{ s.t. } \begin{cases} s_i \begin{bmatrix} 0 & (v_i)_{[3:n]}^T \\ (v_i)_{[1:2]}^T & 0 \end{bmatrix} x > s_i \tau_i, & \forall i \in [1 : n-2] \\ s_i \begin{bmatrix} 0 & (v_i)_{[3:n]}^T \\ (v_i)_{[1:2]}^T & 0 \end{bmatrix} x > s_i \tau_i, & \forall i \in [n-1 : p] \end{cases} \\
&\iff \begin{cases} \exists x_{[3:n]} \text{ s.t. } s_i(v_i)_{[3:n]}^T x_{[3:n]} > s_i \tau_i, & \forall i \in [1 : n-2] \\ \exists x_{[1:2]} \text{ s.t. } s_i(v_i)_{[1:2]}^T x_{[1:2]} > s_i \tau_i, & \forall i \in [n-1 : p] \end{cases} \\
&\iff \begin{cases} s_{[1:n-2]} \in \mathcal{S}(V_{[3:n],[1:n-2]}, \tau_{[1:n-2]}) \\ s_{[n-1:p]} \in \mathcal{S}(V_{[1:2],[n-1:p]}, \tau_{[n-1:p]}). \end{cases}
\end{aligned}$$

Therefore, one has:

$$\begin{aligned} |\mathcal{S}(V_{:, [n-1:p]}, \tau_{[n-1:p]})| &= \sum_{i=0}^2 \binom{n-p+2}{i}, \\ \mathcal{C}(V_{:, [n-1:p]}, \tau_{[n-1:p]}) &= \{(i, j, k) \in [n-1:p]^3 : i, j, k \text{ different}\}, \\ \mathcal{C}([V; \tau^T]_{:, [n-1:p]}, 0) &= \{(i, j, k, l) \in [n-1:p]^4 : i, j, k, l \text{ different}\}. \end{aligned}$$

Now, since the remaining vectors (indices in  $[1 : n-2]$ ) are independent from the others and between themselves, circuits cannot contain any of those indices and

$$\begin{aligned} \mathcal{S}(V, \tau) &= \{-1, +1\}^{[1:n-2]} \times \mathcal{S}_{[n-1:p]}(V, \tau), \quad |\mathcal{S}(V, \tau)| = 2^{n-2} \times \sum_{i=0}^2 \binom{n-p+2}{i}, \\ \mathcal{C}(V, \tau) &= \mathcal{C}_{[n-1:p]}(V, \tau), \quad \mathcal{C}([V; \tau^T], 0) = \mathcal{C}_{[n-1:p]}([V; \tau^T], 0). \end{aligned}$$

For the PERM instances, the reasoning is identical to the one from section 4.5.1. Let

$$H_i := \{x : x_i = 1\} \text{ for } 1 \leq i \leq n, \quad H_{ij} = \{x : x_i - x_j = 0\} \text{ for } 1 \leq i < j \leq n. \quad (\text{A.2})$$

Then, using the fact that

$$x \in \mathbb{R}^n \setminus (\cup_{i,j} H_{ij}) \iff (x_1, \dots, x_n) \text{ are all different},$$

the  $H_{ij}$  hyperplanes split the space into  $n!$  regions of the form  $x_{\sigma(1)} > \dots > x_{\sigma(n)}$ , one for each of the permutations  $\sigma$  of  $[1 : n]$ . For  $\sigma$  fixed, one can have the following configurations:

$$\begin{array}{ll} x_{\sigma(i)} > 1 \forall i \in [1 : n], & x_{\sigma(1)} < 1, x_{\sigma(i)} > 1 \forall i > 1, \\ \dots & \dots \\ x_{\sigma(i)} < 1, x_{\sigma(n)} > 1 \forall i < n, & x_{\sigma(i)} < 1 \forall i \in [1 : n]. \end{array}$$

any other combination is of the form

$$\{x_{\sigma(1)} < 1, \dots, x_{\sigma(i^*)} > 1, \dots, x_{\sigma(j^*)} < 1, \dots, x_{\sigma(n)} > 1\}$$

which does not respect the definition of  $\sigma$ .

For the stem vectors, recall first that the circuits are independent from  $\tau$ , thus we know by proposition 4.5.2 the number of circuits and their structure. Let us justify the corresponding stem vectors are all symmetric.

For a given circuit, if it is only composed of vectors  $e_i - e_j$ , since  $\tau_{ij} = 0$ , the resulting stem vectors are symmetric. Otherwise, since there are only two vectors  $e_i$  and  $e_j$  in the circuit, and their weights are opposite, the resulting stem vectors are symmetric.

For the augmented matrix, by proposition 5.3.20, since the arrangement is centered  $\mathcal{C}(V) = \mathcal{C}([V; \tau^T])$ , meaning that  $\mathfrak{S}_0(V, \tau) = \emptyset$ .

For the 2D instances, there are some slight irregularities in table 5.3. For  $n = 4$  and  $n = 6$ , there is actually (exactly) one symmetric stem vector; one can show this reduces

---

the number of the subarrangement with indices in  $[n - 1 : p]$  by exactly one, thus, after multiplication by  $2^{n-2}$ , by  $2^{n-2}$ . For  $n = 7$ , this is slightly different: there is an asymmetric stem vector of size 2 (say  $\{i, j\}$ , so no  $\{i, j, k\}$  for  $k \in [n - 1 : p] \setminus \{i, j\}$  are circuits), meaning two hyperplanes are parallel; while this also reduces the number of chambers by exactly  $2^{n-2}$ , this reduces the number of stem vectors.

For the RATIO instances, there may be “bad conditioning” due to the way the instances are generated: since vectors are linear combinations of the previous ones, some may have large coordinates. This sometimes resulted in 1 or 2 less chambers detected by some algorithms.<sup>2</sup>

## A.3 Algorithmic behaviors

### A.3.1 Primal heuristics

#### Heuristic B

First, we consider modification B, which avoids using linear optimization and/or stem vectors whenever the current point is “near” the newly added hyperplane. While it is unlikely to have a huge impact for instances with important randomness, for instances such as PERM it may play an important role.

Indeed, recall that after modification A the subtrees start with  $n$  signs, with indices corresponding to the part of  $V$  that is the identity. Therefore, the corresponding witness point is gonna be of the form  $x^s = s/\sqrt{n}$  (assuming it is normed). Then, many of the remaining hyperplanes, which are the  $(e_i - e_j)^\perp$ , shall contain  $x^s$ . This is explicitly detailed below.

#### Heuristic C

Now, we discuss modification C, which, at sign vector  $s = (s_{i_1}, \dots, s_{i_k}) \in \{\pm 1\}^k$ , suggests the next hyperplane with index in  $[1 : p] \setminus \{i_1, \dots, i_k\}$  which ideally makes  $s$  have only one descendant. Let us start with the 2D instances as an illustration.

The matrices  $V$  involved have a specific shape, with two independent subarrangements. Therefore, after modification A, the hyperplanes of the subarrangement in dimension  $n - 2$  are already treated. The remaining hyperplanes form an arrangement of dimension 2, with the first two hyperplanes already taken into account by the QR factorization. When adding the remaining hyperplanes, it is likely that hyperplanes leading to only one child instead

---

<sup>2</sup>Even by norming the data  $V$  and  $\tau$ , depending on which norm was used, sometimes a chamber was detected and sometimes not.

of two are found and added first, see figure A.1.

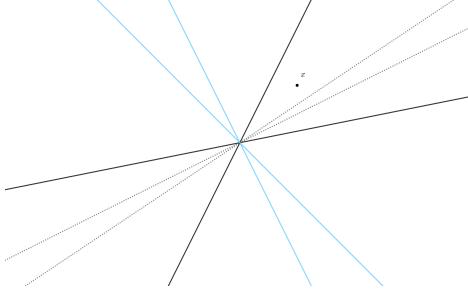


Figure A.1: The black lines represent the two hyperplanes already considered,  $x$  a point of the current region. It is simple to add first the blue hyperplanes, that lead to only one child, then add the dotted hyperplanes rather than doing the opposite. While the figure is shown for a central arrangement, the principle remains the same for an affine arrangement.

We continue with the linear **PERM** instances: recall that for any  $s \in \{\pm 1\}^n$ , the subtree starting with root  $s$  has witness point  $x^s = s/\sqrt{n}$ . This simple expression, combined with the form of  $V$ , renders very easy to choose hyperplanes that result in only one child (depending on  $s$ ). Consider for instance that  $s_1 = +1, s_2 = -1$ , then hyperplane  $(e_1 - e_2)^\perp$  can only have sign +1 in the subtree of  $s$ .

Among the first  $n$  indices, denote  $I_s^+ = \{i \in [1 : n], s_i = +1\}$  and  $I_s^- = \{i \in [1 : n], s_i = -1\}$ . Let  $i^+ \in I_s^+, i^- \in I_s^-$ , if  $i^+ < i^-$  then the sign of hyperplane  $(e_{i^+} - e_{i^-})^\perp$  must be +1 to get a feasible sign vector, otherwise one would have  $x_{i^+} > 0, x_{i^-} < 0$  and  $x_{i^+} - x_{i^-} < 0$ , which is an infeasible (sub)system. If  $i^+ > i^-$ , the same is true with a change of sign (since the hyperplane has expression  $(e_{i^-} - e_{i^+})^\perp$ ). Therefore, in each node  $s \in \{\pm 1\}^n$ , each of the hyperplane of this form implies only one descendant: among the remaining  $n(n-1)/2$  hyperplanes,  $|I_s^+||I_s^-|$  imply only one child and can easily be detected. With the above reasoning, there are

$$\sum_{s \in \{\pm 1\}^n} |I_s^+||I_s^-| = \sum_{i=|I_s^+| \in [0, n]} \binom{n}{i} i(n-i) = n(n-1)2^{n-2}$$

easy hyperplanes to add, which is half the total number of hyperplanes since there are  $2^n$  starting nodes and each has  $n(n-1)/2$  hyperplanes to add, that is  $2^n n(n-1)/2$  hyperplanes to add in total.<sup>3</sup> This may be an explanation of the good improvement ratios observed for these instances in table A.5.

Furthermore, for the affine **PERM** instances, the starting points are  $e + s/\sqrt{n}$ , but for any pair  $i \neq j \in [1 : n]^2$ ,  $\langle e_i - e_j, e + s/\sqrt{n} \rangle = \langle e_i - e_j, s/\sqrt{n} \rangle$  so the reasoning stays true.

For the **CROSSPOLYTOPE** instances, both their stem vectors and their chambers are analytically known, see section 4.5.2. In particular, the stem vectors are of the form  $s_i = s_{i+n} \neq$

<sup>3</sup>Even taking the symmetry into account, both numbers are divided by two so still half the hyperplanes can theoretically be added directly with only one descendant.

---

$s_j = s_{j+n}$  for  $i \neq j \in [1 : n]$  and all other components equal zero. Since the matrix  $V$  is of rank  $n + 1$ , the starting nodes are of size  $n + 1$ . For each of the pairs  $(e_0 + e_i, e_0 - e_i)$ , one and only one of the vectors is among the starting  $n + 1$  ones, except for one index, say  $i^*$ , for which both vectors of the pair are in the starting  $n + 1$ <sup>4</sup>. Suppose first that  $s_{i^*} \neq s_{i^*+n}$ , for instance  $s_{i^*} = +1, s_{i^*+n} = -1$ , the two corresponding equations read

$$x_0 + x_{i^*} > 0 \quad \text{and} \quad -x_0 + x_{i^*} > 0,$$

which are verified for  $x_{i^*}$  positive large enough. If  $s_{i^*} = -1, s_{i^*+n} = +1$ , the reversed inequalities are verified for  $x_{i^*}$  negative large enough. In both cases, if  $s_{i^*} \neq s_{i^*+n}$ , both constraints  $i^*$  and  $i^* + n$  can be satisfied by the choice of  $x_{i^*}$ . If  $s_{i^*} = +1 = s_{i^*+n}$ , the two corresponding equations are

$$x_0 + x_{i^*} > 0, \quad x_0 - x_{i^*} > 0 \quad \Rightarrow \quad x_0 > 0$$

Therefore, for any of the remaining  $n - 1$  hyperplanes, one cannot have  $s_j = -1 = s_{j+n}$ , since with a similar argument this would imply  $x_0 < 0$ . Equivalently, the sign vector would cover a stem vector. (If  $s_{i^*} = -1 = s_{i^*+n}$ , the same argument holds with reversed signs.)

Let us count how many hyperplanes with only one descendant can be added. Among the  $2^{n-1}$  starting nodes having  $s_{i^*} = +1 = s_{i^*+n}$ , one can add, denoting  $s'$  the signs associated with the starting vectors of index different from  $i^*$  and  $i^* + n$ ,

$$\sum_{s' \in \{\pm 1\}^{n-1}} |\{j \in [1 : n] \setminus \{i^*\} : s'_j = -1 \text{ or } s'_{j+n} = -1\}| = \sum_{k=0}^{n-1} \binom{n-1}{k} k = (n-1)2^{n-2}$$

hyperplanes who lead to only one child. Therefore, there are  $(n-1)2^{n-1}$  hyperplanes that can be added while maintaining only one child. Compared to the  $(n-1)2^{n+1}$  number of hyperplanes, this means a quarter of the total hyperplanes are added with only one child every time.

Now, in the nodes such that  $s_{i^*} \neq s_{i^*+n}$ , their inequalities can be ignored by choosing  $x_{i^*}$  accordingly. Suppose the next index added is  $j^*$  or  $j^* + n$  such that  $s_{j^*} = s_{j^*+n}$ , then one can reiterate on the remaining indices. If  $s_{j^*} \neq s_{j^*+n}$ , then one continues until all signs are added or until the first equality between some signs of same index modulo  $n$ . However, in these nodes, the first hyperplane added leads to two descendants (necessarily, no stem vector can be covered). Thus, while in *these* descendants it may be easy to find some hyperplanes leading to only one descendant, the counting becomes tedious.<sup>5</sup>

### A.3.2 Dual heuristics

In this section, since we consider the linear instances, by “stem vector” we mean a pair  $\pm \text{sgn}(\eta)$  of two opposite stem vectors. Thus, we count the number of circuits (see also the

---

<sup>4</sup>Otherwise, if there are two or more of such pairs, the vectors cannot span  $\mathbb{R}^{n+1}$ .

<sup>5</sup>We conjecture, taking this observation into account, that  $n2^n - 2^{n+1} + 2 = \sum_{i=0}^{n-1} i2^i$  out of the  $2^{n+1}(n-1)$  hyperplanes can be added directly. This amounts to approximately half the total hyperplanes.

factors 1/2 in chapter 5).

## General position and high randomness

Let us discuss an important observation: random instances are in general position, which means maximal number of chambers (proposition 5.3.31) *and* maximal number of circuits (remark 5.3.13 6)). This may seem contradictory, since circuits generate  $\mathcal{S}^c$ , so one may think at first that many circuits mean less sign vectors in  $\mathcal{S}$ . The explanation is the following: by general position, the circuits are *all* the subsets of size  $r + 1$ . However, since this is the maximal size of circuits ( $j$  vectors with  $j \geq r + 2$  have nullity  $j - r > 1$ ), all those circuits generate “less” sign vectors in  $\mathcal{S}^c$ .

This observation is interesting for two aspects. First, why the fully dual algorithm “takes double punishment”: in addition to covering tests not necessarily being much more efficient compared to linear optimization, there are lots of stem vectors and many covering tests to perform. Note that instances with a lot of randomness like the RATIO or SRAND with many components (Q not small compared to N) shall suffer from the same observation. Then, this also hinders, though in a much smaller way, the primal-dual method: the acquired stem vectors have limited impact.

## Relevance of the primal-dual method

In the light of the previous paragraph, the primal-dual approach is justified for instances that do not have too much stem vectors (i.e., the instance is not too random). This is where instances with a certain (combinatorial) structure are interesting: since the number of hyperplanes may increase fast, computing all stem vectors, without the techniques of [212] (and [35]), takes too much time.

However, such instances tend to have small stem vectors – not necessarily only small ones, as we have seen for the PERM instances in section 4.5.1. The small size of the stem vectors make them particularly useful since they may prune more easily branches in the other parts of the tree.

Indeed, for the PERM instances, as detailed in section 4.5.1, there are  $(k - 1)! \binom{n+1}{k}/2$  stem vectors of size  $k$ , and approximately half have their first  $n$  components to be zero. Thus, if those are found, they are independent from the starting node (corresponding to hyperplanes  $e_i^\perp$  for  $i \in [1 : n]$ ), meaning they may be reused in all other starting nodes; however, even the other type of circuits, i.e., involving indices in  $[1 : n]$ , may be used in many other subbranches since they have only two nonzero components among these indices.

Furthermore, consider the CROSPOLYTOPE instances. We illustrate this phenomenon on CROSPOLYTOPE-9, and its  $\binom{n}{2} = \binom{9}{2} = 36$  circuits. Among the  $2^{n+1-1} = 2^9 = 512$  starting

---

nodes, all stem vectors were found after the first 130 nodes, so after the first 25% of starting nodes, and 33/36 were found after the first 7% of starting nodes. Thus, the primal-dual algorithm may benefit here from the knowledge of (nearly) all stem vectors without the explicit computation of all of them from the start.

Finally, in the 2D instances, recall that the stem vectors of the 2D instances are the  $\binom{p+2-n}{3}$  vectors with nonzero coordinates on exactly three of the last  $p + 2 - n$  indices. Any learnt stem vector that has zero components in indices  $n - 1$  and  $n$  can be used to prune the tree in each subsequent starting node of size  $r$ . Therefore, any stem vector learnt can be used a lot, and since it is of size 3, used “early” in the tree. Moreover, the percentage of successful covering checks is around 95%, meaning that despite the small numbers of stem vectors they suffice to detect when the tree should be stopped.

### A.3.3 Analysis of the compact algorithm

#### Theory

To conclude this section, let us explain what we can expect from the compaction technique presented in section 5.6. To simplify the presentation, we analyze the (fully) primal compact version, algorithm 5.6.4-5.6.5, and the (fully) dual compact version, that was not presented.

This is done in two steps. The first exposes the situations where the compact algorithms (purely primal and purely dual algorithms without additional heuristics) solve less subproblems, and the second quantifies these situations.

**Remark A.3.1** (imprecision of the dual estimation). In the (fully) dual algorithm, at a given node  $s$ , the following situations may occur. The first descendant covers a stem vector thus the second descendant is always feasible (only one check) or it does not and the second descendant requires a second check. Thus, when not all descendants are feasible, one cannot know (a priori) the number of checks.

**Proposition A.3.2** (difference in numbers of subproblems). *The primal and dual  $\mathcal{S}$ -tree algorithms, at a given node  $s \in \mathcal{S}_k$  and possibly its opposite  $-s$ , solve an amount of subproblems given by the following table.*

*In particular, for the primal algorithm<sup>6</sup>, the only situations where less subproblems are solved are symmetric nodes with both having two descendants, i.e., symmetric nodes with symmetric descendants (so 4 total descendants).*

The proof briefly justify all cases, evoking all 10 columns after the first.

*Proof.*

---

<sup>6</sup>Primal algorithm without heuristics = RC algorithm.

	LOPs					covering checks				
	$\pm s \in \mathcal{S}_k$		$(-)s \in \mathcal{S}_k$			$\pm s \in \mathcal{S}_k$		$(-)s \in \mathcal{S}_k$		
	4	3	2	2	1	4	3	2	2	1
number of descendants	4	3	2	2	1	4	3	2	2	1
classic algorithm	2	2	2	1	1	4	3 or 4	3	1 or 2	1 or 2
compact algorithm	1	2	2	1	1	2	1 or 2	1 or 2	1 or 2	1 or 2

Table A.2: Number of subproblems solved, depending on the state of the current node. Moreover,  $\pm s \in \mathcal{S}_k \iff \underline{s} = 0$  and  $(-)s \in \mathcal{S}_k \iff \underline{s} \neq 0$ .

[4 descendants, primal] [classic] The classic algorithms gets two of the four descendants then requires two LOPs (one for  $s$ , one for  $-s$ , both concluding the sign vector is feasible). [compact] The compact algorithm gets two easy descendants and, by symmetry, the LOP will conclude  $\pm(s, -s_{k+1})$  are feasible.

[3 descendants, primal] [classic] Similarly, two LOPs are required (but they return opposite conclusions). [compact] Similarly, two easy descendants, but the LOP will fail and require a new LOP (in  $\mathbb{R}^{n+1}$ ) that will be feasible ( $\underline{s} \neq 0$ ).

[2 descendants,  $s$  and  $-s$ , primal] [classic] Similarly, two LOPs are required (they both conclude the sign vector is infeasible). [compact] Similarly, after the two easy descendants (since  $\underline{s} = 0$  here) the first LOP tells the descendant is infeasible, as does the second (in  $\mathbb{R}^{n+1}$ ).

[2 descendants, just  $s$ , primal] [classic] Only one node so one LOP (feasible descendant). [compact] Here,  $\underline{s} \neq 0$ , so the algorithm uses one LOP (feasible descendant).

[1 descendant, primal] [classic] Only one node so one LOP (infeasible sign vector). [compact] Here,  $\underline{s} \neq 0$ , so the algorithm uses one LOP (infeasible descendant).

[4 descendants, dual] [classic] For both  $s$  and  $-s$ , no stem vector is covered so both require two checks (4 total). [compact] At  $s$ , no stem vector is covered so two checks are required.

[3 descendants, dual] [classic] Without loss of generality, assume  $s$  has two descendants and  $-s$  one:  $s$  requires two covering checks and  $-s$  one or two (see remark A.3.1). [compact] Similarly, depending on the order the potential descendants are considered, one needs one covering check (the first check returns some covered stem vector) or two (the first check returns no covered one).

[2 descendants,  $s$  and  $-s$ , dual] Actually, by symmetry one can show that  $s$  and  $-s$  have opposite descendants. If they have descendants  $(s, +1)$  and  $(-s, +1)$ , this contradicts the fact the region of  $s$  and  $-s$  have opposite asymptotic cones. [classic] Thus, there are three covering checks performed (since the descendants are opposed so one requires two tests, the other one). [compact] There is one covering check if the first potential descendant is

---

infeasible (even in  $\mathbb{R}^{n+1}$ ) and two if this one is tested after.

[2 descendants, just  $s$ , dual] [classic] Similarly, one or two covering checks depending on the order. [compact] Same as the classic case.

[1 descendant, dual] [classic] Similarly, one or two covering checks depending on the order. [compact] Same as the classic case.  $\square$

**Remark A.3.3** (precision on the covering checks). In the dual variant, when  $\pm s \in \mathcal{S}_k$  ( $\mathbb{S} = 0$ ), there are less covering checks but they may require more stem vectors since they use those of  $([V; \tau^T], 0)$  and  $(V, \tau)$ . Once there is only  $s$  or  $-s$  ( $\mathbb{S} \neq 0$ ), the stem vectors considered by the compact algorithm are only the ones of  $[V; \tau^T]$ .  $\square$

We now move onto the second part, where we give a special meaning to the part of table A.2 with 4 descendants.

**Definition A.3.4** (complete direct lineage). Let  $\mathcal{A}(V, \tau)$  for  $V \in \mathbb{R}^{n \times p}$  and  $\tau \in \mathbb{R}^p$  be an arrangement. Consider the computation of its chambers by the  $\mathcal{S}$ -tree algorithm. For  $k \in [1 : p - 1]$ ,  $s \in \mathcal{S}_k$  is said to have complete direct lineage if  $-s \in \mathcal{S}_k$ ,  $s$  and  $-s$  both have two descendants, i.e.,  $(s, +1)$ ,  $(s, -1)$ ,  $(-s, +1)$ ,  $(-s, -1)$  all belong to  $\mathcal{S}_{k+1}$ .

Naturally,  $s$  has complete direct lineage  $\iff -s$  has complete direct lineage.  $\square$

Observe that, a priori, complete direct lineage (CDL) depends on the ordering of the sign vectors considered. Furthermore, according to proposition A.3.2, one less subproblem is solved by the compact primal algorithm at each pair of opposite sign vectors with CDL. By convention, the leaves of the tree are considered to have no descendants.

**Proposition A.3.5** (pairs with 4 descendants and  $\mathcal{S}_s$ ). *The number of sign vectors with CDL in the  $\mathcal{S}$ -tree equals  $|\mathcal{S}_s| - 2$ .*<sup>7</sup>

*Proof.* The proof proceeds by induction of  $p$ , the number of vectors and the depth of the tree. [Initialization] Consider the case  $p = 1$ . One has  $\mathcal{S} = \{+1, -1\}$  (assuming  $V \neq 0_n$ ), and the number of symmetric pairs is  $0 = 2 - 2$ .<sup>8</sup>

[Heredity] Suppose this is true for any arrangement with  $p$  vectors and consider a  $p + 1$ -th vector. Consider a sign vector  $s \in \mathcal{S}_p$ . If it is asymmetric, i.e.,  $-s \notin \mathcal{S}_p$ , then the descendants of  $s$  are necessarily asymmetric: neither the number of symmetric sign vectors with CDL nor  $|\mathcal{S}_s|$  increases.

If  $s$  and  $-s$  are in  $\mathcal{S}_p$ , and they have two or three descendants total, only two of them are symmetric (see the proof of proposition A.3.2): neither the number of sign vectors with

---

<sup>7</sup>This  $-2$  is somewhat curious, though it seems to be an intrinsic phenomenon. One could consider the empty root to be both “ $+\emptyset$  and  $-\emptyset$ ” with their descendants  $\{(\emptyset, 1), (\emptyset, -1), (-\emptyset, 1), (-\emptyset, -1)\}$ , one gets  $|\mathcal{S}_s|$ .

<sup>8</sup>If one considers the root as 2, one has  $2 = |\mathcal{S}_s| = 2$ .

CDL nor  $|\mathcal{S}_s|$  increases ( $s$  and  $-s$  in  $\mathcal{S}_p$  to  $(s, +1)$  and  $(-s, -1)$  or  $(s, -1)$  and  $(-s, +1)$  in  $\mathcal{S}_p$ ).

The last case is the case of CDL, i.e.,  $\pm s \in \mathcal{S}_p$  and all four descendants are in  $\mathcal{S}_{p+1}$ . It is clear that these two sign vectors increase by two  $|\mathcal{S}_s|$  since there was  $s$  and  $-s$  in  $\mathcal{S}_p$  and now  $(s, +1), (-s, -1), (s, -1), (-s, +1)$  in  $\mathcal{S}_{p+1}$ . Finally, there are two more sign vectors with complete lineage,  $s$  and  $-s$ . This completes the proof.  $\square$

**Corollary A.3.6** (CDL and  $\mathcal{S}_s$ ). The number of (partial) sign vectors in the  $\mathcal{S}$ -tree with complete direct lineage is equal to  $|\mathcal{S}_s| - 2$ . Since one less problem is solved at each opposite pair of such sign vectors, the compact primal algorithm solves  $|\mathcal{S}_s|/2 - 1$  less subproblems.<sup>9</sup>

Observe that only LOPs with *feasible* sign vectors are skipped by the compact algorithm. Before mentioning an interesting corollary, we have (“(c)” stands for compact)

$$\#LOP(c) = \#LOP - \frac{|S_s(V, \tau)|}{2}, \quad \frac{\#LOP}{\#LOP(c)} = 1 + \frac{|S_s(V, \tau)|}{2\#LOP(c)}. \quad (\text{A.3})$$

Recall that  $S_s(V, \tau) = S(V, 0)$ , though we shall use the first expression to distinguish more easily between affine and linear arrangements. However, since the numerical experiments are conducted with modification A, i.e., starting with  $2^n$  subtrees coming from a complete subarrangement, the formula takes the modified form:

$$\#LOP(c) = \#LOP - \frac{|S_s(V, \tau)| - 2^n}{2}, \quad \frac{\#LOP}{\#LOP(c)} = 1 + \frac{|S_s(V, \tau)| - 2^n}{2\#LOP(c)}. \quad (\text{A.4})$$

**Corollary A.3.7** (CDL independent of vector ordering). *Since  $|\mathcal{S}_s|$  is independent of the vector ordering, so is the amount of nodes with CDL.*

## Numerics

Now, we observe the relevance of the compact algorithms introduced in section 5.6 and displayed in table 5.5. The first column of times represents the symmetrized algorithm with modification A, which is very close to the symmetrized RC algorithm (modification A has little effect on the computation times). Let us propose some main observations.

- For the **RAND** instances, the compact algorithms are always better – the ratios seem to increase by 0.5-0.6 on average.
- For the **SRAND** instances, the compact algorithms are slightly better.
- For the **2D** instances, the compact algorithms perform worse.

---

<sup>9</sup>The  $-1$  term could be cancelled if, by convention, one would count 1 LOP to get the first two sign vectors  $S_1 = \{-1, +1\}$ .

- 
- For the **PERM** instances, the compact algorithms show rather convincing improvement – this is to be expected since the instances are centered.
  - For the **RATIO** instances, the compact algorithms are more efficient when they are primal but worse when they compute the full list of stem vectors.

One possible explanation concerning the **2D** instances is the following. Consider the subarrangement defined by indices  $[n - 1 : p]$  and let  $q = p - n + 2$ . It has  $\sum_{i=0}^n \binom{q}{i}$  chambers and the symmetric part is  $2 \sum_{i=0}^{n-1} \binom{q-1}{i} = \sum_{i=0}^n \binom{q}{i} - \binom{p-1}{n}$ . However, since this term is multiplied by  $2^{n-2}$ , it results that the **2D** instances have a much larger proportion of asymmetric chambers compared to the other instances, see table A.3. Since the goal of compact algorithm is to avoid symmetric computations, it is logical that its performance are worsened on less symmetric instances.

instance type	RAND	SRAND	2D	PERM	RATIO
approximate $\frac{ \mathcal{S}(V,0) }{ \mathcal{S}(V,\tau) }$ (%)	> 80	> 70	25	100	50

Table A.3: Approximate proportions of symmetric chambers. The **2D** instances have particularly low proportion of symmetric chambers.

**Subproblems** In the next table, we mention the instances tested (which values of  $n$  and  $p$ ). The RC stands for the original algorithm [208] while (c) stands for “compact”.

Let us detail what happens for the **PERM** instances, which are a bit shifted to the right in figures A.2b and A.2d. Since the coefficients of the vectors are in  $\{-1, 0, +1\}$ , and each  $(v_i, \tau_i)$  has only two nonzero components, say for instance  $x = 0$ , belongs to most of the hyperplanes, so many linear optimization problems are skipped since there are clearly two descendants. If this wasn’t done, the number of feasible LOPs and feasible LOP(c) numbers should be approximately doubled: the coordinates in the  $y$ -axis would thus remain at the same value while those in the  $x$ -axis would be halved, meaning the values would be much closer to the others and to the line.

## A.4 Linear instances and other topics

While the linear instances are not the main topic of this chapter, it did seem relevant to showcase the behaviors of the various algorithms. Results are displayed in table A.5. Overall, for the instance types described above, we observe similar improvement ratios. For the combinatorial instances presented in the last part of section 3.5.2.A, we also obtain encouraging results, especially for the primal-dual algorithm.

In what follows, we recall that D1 means using a few stem vectors computed at the start, D2 is the primal-dual method, D3 is the intermediate method using all stem vectors

APPENDICES

---

Problems	$ S(V, \tau) $	$ S_s(V, \tau) $	Feas.(RC)	Feas.	Feas.(c)	Feas. - Feas.(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: Relevant values for the affine instances. Columns 2 and 3 represent the cardinalities of the sign vector sets, columns 4-5-6 the number of feasible problems solved. Column 7 is the difference of the two previous ones. The last column represents the second term in the right-hand side of (A.4). This table is illustrated below, in figure A.2. The \* represent the irregularities mentioned above (no perfect general position in the subarrangement).

and directions to get a free descendant, and D4 is the fully dual method.

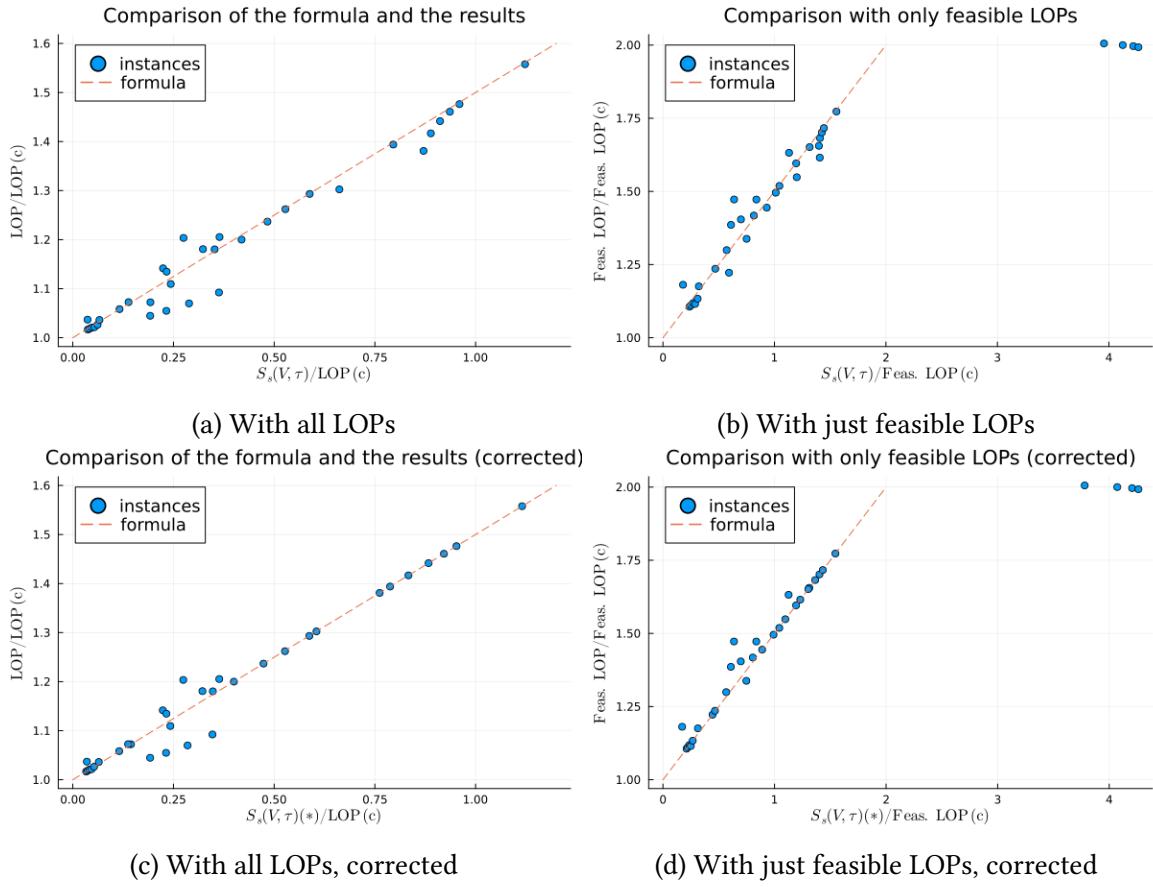


Figure A.2: Illustration of (A.4) (and (A.3)). The correction, i.e., the difference between (A.3) and (A.4), is denoted by  $(*)$  and is added to the bottom pictures; for the instances on the top-right of the right graphs, it brings the points on the line corresponding to the formula, given by  $y = 1 + x/2$ . The (c) denotes the number of LOPs of the compact variant. for the pictures on the right, four points, corresponding to the PERM instances, are shifted to the right. A possible explanation is proposed.

#### A.4.1 Circuits computation

Computing the circuits can be done with algorithm 5.5.1. When the arrangement is linear, the algorithm is essentially the same, except one always get  $\tau_j^T \eta = 0$  since  $\tau = 0$ . For the compact algorithm, since the circuits of  $V$  and  $[V; \tau^T]$  are needed, there is a slight change to verify, when a circuit of  $V$  is found, if it is already a circuit of  $[V; \tau^T]$ : if yes, the recursion stops, otherwise it continues. See section 5.3 and in particular figure 5.2 and subsection 5.3.4.

In most instances, obtaining the circuits takes a small fraction of the total time – the major part is devoted to the computations using the stem vectors. However, for larger instances, when  $p$  becomes larger than 30, such as in PERM-8 or the larger instances from [35], the algorithm can be very slow (does not finish for RESONANCE-6). Specialized methods [35, 212] focusing on such instances, use different approaches since algorithm 5.5.1 may not

Problem	Computation times (in sec) and improvement ratios						
	RC	P	PD	D			
RAND-8-15-7	6.04	<b>3.36</b>	<b>1.80</b>	3.74	<b>1.62</b>	4.53	1.33
RAND-9-16-8	14.3	<b>8.34</b>	<b>1.72</b>	9.54	<b>1.50</b>	13.6	1.05
RAND-10-17-9	32.7	<b>19.7</b>	<b>1.66</b>	23.3	<b>1.40</b>	33.1	0.987
SRAND-8-20-2	15.7	3.96	<b>3.97</b>	<b>2.90</b>	<b>5.42</b>	9.74	1.62
SRAND-8-20-4	79.6	<b>36.7</b>	<b>2.17</b>	42.7	<b>1.86</b>	204	0.389
SRAND-8-20-6	100	<b>50.2</b>	<b>1.99</b>	67.0	<b>1.49</b>	42	0.238
2D-6-20	1.10	0.551	<b>1.99</b>	<b>0.20</b>	<b>5.49</b>	0.566	1.94
2D-7-20	1.93	1.34	<b>1.44</b>	<b>0.444</b>	<b>4.36</b>	1.01	1.92
2D-8-20	3.42	1.72	<b>1.99</b>	<b>0.580</b>	<b>5.90</b>	1.73	1.98
PERM-6	5.11	1.21	<b>4.22</b>	<b>0.448</b>	<b>11.4</b>	3.55	1.44
PERM-7	50.7	10.2	<b>4.97</b>	<b>3.92</b>	<b>13.0</b>	64.6	0.786
PERM-8	551	99.7	<b>5.53</b>	<b>41.7</b>	<b>13.2</b>	1520	0.362
RATIO-5-20-0.7	8.43	4.50	<b>1.87</b>	<b>3.76</b>	<b>2.24</b>	12.2	0.691
RATIO-6-20-0.7	20.7	11.6	<b>1.79</b>	<b>10.8</b>	<b>1.91</b>	40.1	0.516
RATIO-7-20-0.7	53.0	<b>31.7</b>	<b>1.67</b>	33.6	<b>1.58</b>	165	0.321
THRESHOLD-3	0.0464	0.0245	<b>1.89</b>	0.0260	<b>1.78</b>	<b>0.0233</b>	<b>1.99</b>
THRESHOLD-4	1.67	0.784	<b>2.13</b>	<b>0.557</b>	<b>3.00</b>	0.975	1.71
THRESHOLD-5	167	50.2	<b>3.33</b>	<b>35.8</b>	<b>4.67</b>	1300	0.129
RESONANCE-4	0.367	0.103	<b>3.55</b>	<b>0.0517</b>	<b>7.10</b>	0.221	1.66
RESONANCE-5	22.5	4.28	<b>5.27</b>	<b>2.11</b>	<b>10.7</b>	70.5	0.320
RESONANCE-6	4240	<b>565</b>	<b>7.50</b>	684	<b>6.19</b>	*	*
CROSSPOLYTOPE-10	62.7	11.	<b>5.73</b>	<b>7.39</b>	<b>8.49</b>	38.6	1.62
CROSSPOLYTOPE-11	193	30.1	<b>6.43</b>	<b>21.9</b>	<b>8.84</b>	123	1.57
CROSSPOLYTOPE-12	595	83.0	<b>7.16</b>	<b>65.3</b>	<b>9.11</b>	369	1.61
DEMICUBE-4	0.0546	0.0137	<b>3.97</b>	<b>0.00854</b>	<b>6.39</b>	0.0299	1.83
DEMICUBE-5	2.79	1.09	<b>2.55</b>	<b>0.718</b>	<b>3.89</b>	1.58	1.77
DEMICUBE-6	434	119	<b>3.65</b>	<b>90.1</b>	<b>4.82</b>	3210	0.135

Table A.5: Computation times in black and ratio time(RC) / time(A) in blue for the linear instances and the different algorithms; bold ratios are the best ones. For algorithm D, the stem vectors and the covering tests are computed in slightly faster ways described in sections A.4.1 and A.4.2.

be useable. They use the combinatorial symmetries of the instances, through symmetry groups, to compute much less circuits while getting all the information through the symmetries. For instance, in the PERM instances, when the circuit of vectors  $e_1, e_2$  and  $(e_1 - e_2)$  is found, by symmetry one could add all the circuits formed by the triples  $e_i, e_j, (e_i - e_j)$  for all  $i < j$ .

This section presents a simple refinement inspired from TOPCOM [214, 212] that can be applied for any arrangement. Instead of computing a null space every time a new index is added, therefore computing a null space at every level of the tree, one can compute the

---

echelon form of the matrix  $V_{:, [1:k]}$  (or  $[V; \tau^T]_{:, [1:k]}$ ). This means the null space computation is replaced by the update of the echelon form, which costs less. However, since there is an additional matrix kept in memory, and the cost of a null space is already low, it is not necessarily faster to compute stem vectors this way. Indeed, if the recursion depth in algorithm 5.5.2 is low, it may be better to do a few null space computations rather than keeping a variable in memory. The comparisons are illustrated in tables A.6 and A.7 on the affine instances.

Let us comment the improvement ratios brought by the implementation of the echelon form, starting with table A.6 and the normal variants. It is reasonably efficient for the **RAND** and **RATIO** instances and seems to get better when the dimension increases. This is coherent since the stem vectors have bigger sizes, therefore the recursion in algorithm 5.5.2 has more levels. Conversely, when the recursion is shallow as in the 2D instances, the computing a null space at each level can be better. The **PERM** instances do not benefit much from computing the echelon form, whereas the **SRAND** instances show some medium ratios.

For the compact case in table A.7, the improvement is slightly worse, which is coherent since both times are increased, so the ratio becomes lower. A simple empirical rule can be the following: when the QR factorization is computed, if the rank is very small or the matrix  $R$  has many (close to) zero blocks as in the 2D instances, this feature is not used.

#### A.4.2 Recursive covering test

**Remark A.4.1** (recursive covering). The covering tests can be performed recursively. Indeed, consider that node  $s \in \mathcal{S}_k$  has been verified by a covering test: the product  $M_{:, I^s s}$  was computed, where  $M$  represents a matrix of stem vectors (of any type) and  $I^s$  contains the hyperplanes already treated in  $s$ . Let  $s_+ = (s, +1)$  and  $s_- = (s, -1)$  be the descendants of  $s$  and  $i_{k+1}$  the index of the added hyperplane. The covering test computes  $M_{:, I^{s+} s_+} = M_{:, I^s s} + M_{:, i_{k+1}}$  and  $M_{:, I^{s-} s_-} = M_{:, I^s s} - M_{:, i_{k+1}}$ . Therefore, the matrix-vector product can be incrementally computed, and at each node there is only an addition/subtraction of vectors. It is illustrated in table A.8.  $\square$

It is however not obvious this necessarily reduces the total computation time. Indeed, the matrix-vector product of the covering test introduced is done in the dimension equal to the number of signs, the other being the number of stem vectors which can be large. The recursive computation proposed here still has to add (or subtract) two vectors whose length is the number of stem vectors. In the numerical comparisons below, variant D1 is not shown since there are very few stem vectors. Variant D2 is also not shown for the following reason: since the number of stem vectors increases, when the algorithm goes back to an unfinished subtree one must adjust the size of the vector keeping the product in memory. Though it can be implemented, it is indeed inefficient therefore not shown here. This recursive method is thus only compared for variants D3 and D4. Before discussing the

Problem	noncompact version						
	$ \mathfrak{S}_s(V, \tau) $	duplicates	$ \mathfrak{S}_a(V, \tau) $	duplicates	time	echelon	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
Mean							1.34
Median							1.22

Table A.6: Computation times of the stem vectors in the regular variants. The second and fourth columns represent the numbers of stem vectors, the third and fifth columns the number of duplicates. The three remaining columns indicate the time of the initial computation, the computation time with echelon form and their ratio: if over 1, it means the echelon form is faster.

comparisons, table A.8 presents an illustration of this recursive computation.

Problem	compact version						
	$ \mathfrak{S}_s(V, \tau) $	duplicates	$ \mathfrak{S}_a(V, \tau) $	duplicates	time	echelon	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
Mean							1.23
Median							1.16

Table A.7: Computation times of the stem vectors in the compact variants. The second and fourth columns represent the numbers of stem vectors, the third and fifth columns the duplicates. The three remaining columns indicate the time of the initial computation, the computation time with echelon form and their ratio: if over 1, it means the echelon form is faster.

In tables A.9 and A.10, we compare the total times of the D3 and D4 variants using the recursive covering tests or not. The tests were conducted on some of the linear instances,

index	sign	current vector								$\mathfrak{S}$							
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 of the recursive implementation of the covering test. There are  $p = 5$  vectors in  $\mathbb{R}^n$ , and the matrix of stem vectors is given on the right in transpose form in the right half. For instance the first column means  $[v_1 \ v_2 \ -v_3]$  is of nullity one in  $\mathbb{R}_+^{\{1,2,3\}}$ . On the left, on the line with index = 1 and sign = +, the current vector is  $+M_{:,1}$  (the first line of the transposed matrix of stem vectors). On the following line, since the sign is also +, the second line of the matrices of stem vectors is added. On line with index = 3 and sign = -, the current vector is thus the first line of  $\mathfrak{S}$  plus the second minus the third. In particular, coordinate 1 of the current vector equals 3, which is the size of the first stem vector (first column of  $\mathfrak{S}$ ), so the covering test stops the recursion.

which is why there is only one type of stem vectors.

One may observe that the efficiency of this method is mostly dependent on the cardinality of  $\mathfrak{S}$ , which is to be expected. However, the nature of the arrangement also matters: very close improvements are obtained for PERM-8 and RATIO-7-20-0.7 but the number of covering tests done vary. A possible empirical rule is chosen as follows: if there are many vectors ( $p$  large), or if the ratio of the number of stem vectors divided by the maximal number of stem vectors  $\binom{p}{r+1}$  is high, then this method is used.

## A final commentary

Another interesting aspect would be to study the behavior of the algorithms for combinatorial *affine* instances. One could consider adding perturbations ( $\tau \neq 0$ ) to the threshold, resonance, or demicube instances, or studying the arrangements considered for instance in [203].

For D3	basic version		recursive version		$ \mathcal{C} $	ratios	
	Name	total time	cover time	total time	cover time	total	cover
RAND-4-8-2	0.00313	0.00184	0.00901	0.00548	56	0.35	<b>0.34</b>
RAND-7-8-4	0.0224	0.00507	0.0460	0.0199	56	0.49	<b>0.25</b>
RAND-7-9-4	0.0343	0.00823	0.0749	0.0349	126	0.46	<b>0.24</b>
RAND-7-10-5	0.105	0.0233	0.207	0.0878	210	0.50	<b>0.27</b>
RAND-7-11-4	0.0849	0.0236	0.173	0.0793	462	0.49	<b>0.30</b>
RAND-7-12-6	0.513	0.112	0.894	0.357	792	0.57	<b>0.31</b>
RAND-7-13-5	0.471	0.125	0.805	0.336	1716	0.59	<b>0.37</b>
RAND-7-14-7	2.42	0.662	3.84	1.49	3003	0.63	<b>0.44</b>
RAND-8-15-7	4.30	1.44	6.12	2.43	6435	0.70	<b>0.59</b>
RAND-9-16-8	13.4	5.54	16.2	6.44	11440	0.82	<b>0.86</b>
RAND-10-17-9	40.4	19.8	41.7	17.4	19448	0.97	<b>1.13</b>
2D-4-20	0.0696	0.0397	0.179	0.109	680	0.39	<b>0.36</b>
2D-5-20	0.132	0.0660	0.347	0.195	680	0.38	<b>0.34</b>
2D-6-20	0.270	0.147	0.667	0.391	560	0.41	<b>0.38</b>
2D-7-20	0.621	0.332	1.52	0.889	455	0.41	<b>0.37</b>
2D-8-20	0.795	0.391	2.27	1.37	364	0.35	<b>0.29</b>
SRAND-8-20-2	4.14	0.973	9.35	4.22	540	0.44	<b>0.23</b>
SRAND-8-20-4	202	159	141	85.7	84390	1.44	<b>1.85</b>
SRAND-8-20-6	466	399	258	182	160074	1.81	<b>2.19</b>
PERM-5	0.0853	0.0369	0.307	0.176	197	0.28	<b>0.21</b>
PERM-6	1.03	0.336	2.56	1.29	1172	0.40	<b>0.26</b>
PERM-7	21.2	7.60	28.8	11.1	8018	0.74	<b>0.69</b>
PERM-8	992	645	646	257	62814	1.53	<b>2.51</b>
RATIO-3-20-0.7	0.227	0.119	0.364	0.197	3486	0.62	<b>0.60</b>
RATIO-3-20-0.9	0.141	0.0776	0.305	0.174	1332	0.46	<b>0.45</b>
RATIO-4-20-0.7	2.23	1.44	2.21	1.34	15138	1.01	<b>1.27</b>
RATIO-4-20-0.9	1.85	1.18	1.96	1.01	14052	0.94	<b>1.17</b>
RATIO-5-20-0.7	11.4	8.57	8.64	4.83	34556	1.33	<b>1.77</b>
RATIO-5-20-0.9	10.4	7.52	7.92	4.38	31334	1.31	<b>1.72</b>
RATIO-6-20-0.7	48.9	39.1	28.9	17.0	56184	1.70	<b>2.30</b>
RATIO-6-20-0.9	52.2	42.2	33.9	21.1	36970	1.54	<b>2</b>
RATIO-7-20-0.7	232	201	117	80.3	112576	1.98	<b>2.51</b>
RATIO-7-20-0.9	131	108	73.0	45.4	74970	1.79	<b>2.38</b>

Table A.9: Run times for the linear instances with option D3. Columns 2-3 represent the total and covering times with the full matrix-vector product in the test. Columns 4-5 represent the total and covering times done recursively. Column 6 gives the number of stem vectors. Columns 7-8 show the ratios for the total times and the covering times.

For D4	basic version		recursive version		C	ratios	
	Name	total time	cover time	total time	cover time		
RAND-4-8-2	0.00803	0.00441	0.014	0.00837	56	0.57	<b>0.53</b>
RAND-7-8-4	0.0279	0.0171	0.0453	0.0277	56	0.62	<b>0.62</b>
RAND-7-9-4	0.0516	0.0320	0.0820	0.0504	126	0.63	<b>0.63</b>
RAND-7-10-5	0.122	0.0754	0.203	0.122	210	0.60	<b>0.62</b>
RAND-7-11-4	0.132	0.0861	0.206	0.129	462	0.64	<b>0.67</b>
RAND-7-12-6	0.565	0.359	0.885	0.552	792	0.64	<b>0.65</b>
RAND-7-13-5	0.533	0.346	0.869	0.551	1716	0.61	<b>0.63</b>
RAND-7-14-7	2.42	1.68	3.31	2.14	3003	0.73	<b>0.79</b>
RAND-8-15-7	4.57	3.4	5.52	3.68	6435	0.83	<b>0.92</b>
RAND-9-16-8	15.6	12.6	14.1	9.63	11440	1.10	<b>1.30</b>
RAND-10-17-9	49.1	41.9	35.3	25.1	19448	1.39	<b>1.67</b>
2D-20-4	0.173	0.106	0.303	0.190	680	0.57	<b>0.56</b>
2D-20-5	0.359	0.214	0.623	0.391	680	0.58	<b>0.55</b>
2D-20-6	0.652	0.40	1.23	0.779	560	0.53	<b>0.51</b>
2D-20-7	1.11	0.658	2.17	1.41	455	0.51	<b>0.47</b>
2D-20-8	1.92	1.14	4.03	2.60	364	0.48	<b>0.44</b>
SRAND-8-20-2	10.5	6.30	19.3	12.5	540	0.54	<b>0.50</b>
SRAND-8-20-4	371	350	202	172	84390	1.83	<b>2.03</b>
SRAND-8-20-6	890	863	404	364	160074	2.21	<b>2.37</b>
PERM-5	0.375	0.227	0.691	0.431	197	0.54	<b>0.53</b>
PERM-6	3.81	2.27	6.98	4.34	1172	0.55	<b>0.52</b>
PERM-7	63.7	42.6	65.8	37.0	8018	0.97	<b>1.15</b>
PERM-8	3518	3081	1593	1058	62814	2.21	<b>2.90</b>
RATIO-3-20-0.7	0.364	0.253	0.532	0.339	3486	0.68	<b>0.69</b>
RATIO-3-20-0.9	0.177	0.111	0.318	0.196	1332	0.56	<b>0.57</b>
RATIO-4-20-0.7	3.26	2.70	3.01	2.06	15138	1.08	<b>1.31</b>
RATIO-4-20-0.9	2.87	2.34	3.14	2.16	14052	0.91	<b>1.08</b>
RATIO-5-20-0.7	18.6	16.7	12.6	9.35	34556	1.48	<b>1.78</b>
RATIO-5-20-0.9	15.1	13.3	10.6	7.77	31334	1.42	<b>1.71</b>
RATIO-6-20-0.7	75.7	70.2	40.3	32.1	56184	1.88	<b>2.18</b>
RATIO-6-20-0.9	92.9	86.9	48.9	40.1	63970	1.90	<b>2.17</b>
RATIO-7-20-0.7	374	360	160	140	112576	2.34	<b>2.58</b>
RATIO-7-20-0.9	179	169	87.4	73.0	74970	2.05	<b>2.32</b>

Table A.10: Run times for the linear instances with option D4. Columns 2-3 represent the total and covering times with the full matrix-vector product in the test. Columns 4-5 represent the total and covering times done recursively. Column 6 gives the number of stem vectors. Columns 7-8 show the ratios for the total times and the covering times.



# Appendix B

## Geometric elements on polytopes

This appendix aims at describing a few useful geometrical properties about polytopes, as well as a few additional ones concerning zonotopes. Most of them are, up to notation and convention, well-known and basic properties.

In this section, we use “faces” to describe the faces of all dimensions of a polytope. To emphasize, we shall explicitly use “of maximal dimension” to describe what is often called “facets” in combinatorial geometry.

### B.1 Polytopes and their face(t)s

In what follows, we use the “H-representation” of (convex) polytopes, i.e., their representation as a finite intersection of half-spaces and hyperplanes of the form  $P = \{x \in \mathbb{R}^n : Ax \leqslant a, Bx = b\}$ . The first lemma describes the relative interior of such polyhedrons.

**Lemma B.1.1** (relative interior of polyhedrons). *Let  $P = \{x \in \mathbb{R}^n : Ax \leqslant a, Bx = b\}$  for some  $A \in \mathbb{R}^{m \times n}$  and  $a \in \mathbb{R}^m$ . Assume that for every  $i \in [1 : m]$ , there exists some  $x^i \in P$  such that  $(Ax^i - a)_i < 0$ , i.e., none of the inequalities is actually an equality. Then  $\text{ri}(P) = \{x \in \mathbb{R}^n : Ax < a, Bx = b\}$ .*

*Proof.* [ $\subseteq$ ] Let  $x \in \text{ri}(P)$ , clearly  $x \in P$  so  $Bx = b$ . Suppose there exists some  $i$  such that  $(Ax - a)_i = 0$ . Then, using the  $x^i \in P$  such that  $(Ax^i - a)_i < 0$ , we know that there exists some  $\delta^i > 0$  such that  $x + \delta^i(x - x^i) \in P$ , by definition of the relative interior. However, looking at index  $i$ , we get

$$\begin{aligned} (A(x + \delta^i(x - x^i)) - a)_i &= (Ax - a + \delta^i(Ax - Ax^i - a + a))_i \\ &= (Ax - a + \delta^i((Ax - a) - (Ax^i - a)))_i \\ &= 0 + \delta^i(0 - (Ax^i - a)_i) > 0, \end{aligned}$$

which is a contradiction with the definition of  $P$ .

---

[ $\supseteq$ ] Let  $x$  such that  $Ax < a$  and  $Bx = b$ ,  $x_0 \in P$ , we show there exists some  $\delta > 0$  such that  $x + \delta(x - x_0)$  belongs to  $P$ . Indeed,  $B(x + \delta(x - x_0)) = b + \delta(b - b) = b$ . Moreover,

$$A(x + \delta(x - x_0)) - a = (Ax - a) + \delta(Ax - a - (Ax_0 - a))$$

which is  $< 0$  for  $\delta$  small enough since  $(Ax - a) < 0$ .  $\square$

The next lemma gives an expression of faces, whose definition is recalled, that is more easily useable.

**Definition B.1.2** (face of a polyhedron). A face  $F$  of a polytope  $P$  is a subset of  $P$  such that for all  $x, y \in P$  and all  $t \in (0, 1)$ ,  $(1-t)x + ty \in F \Rightarrow x, y \in F$ .  $\square$

**Lemma B.1.3** (faces and subsets of indices). *Let  $P = \{x \in \mathbb{R}^n : Ax \leq a, Bx = b\}$  be a convex polyhedron,  $F$  is a face of  $P$  if and only if there exists a subset of indices  $I$  such that  $F = \{x \in P : (Ax - a)_I = 0\}$ .*

*Proof.* If  $F = \{x \in P : (Ax - a)_I = 0\}$ , let  $x_1, x_2 \in P$  such that  $(x_1 + x_2)/2 \in F$ . This reads  $(A(x_1 + x_2)/2 - a)_I = 0$ , but since  $(Ax_1 - a)_I \leq 0$  and  $(Ax_2 - a)_I \leq 0$ , both quantities must be zero, meaning  $x_1$  and  $x_2$  are in  $F$ .

Conversely, since  $F$  is nonempty convex take  $x_0$  in its relative interior, and define  $I = \{i : (Ax_0)_i = a_i\}$ . Let  $x \in F \subseteq P$ , by definition of  $\text{ri}(F)$ ,  $(1-t)x_0 + tx \in F$  for some  $t > 1$ . Now,

$$a_I \geq (A((1-t)x_0 + tx))_I = (1-t)(Ax_0)_I + t(Ax)_I \iff 0 \geq (1-t)(Ax - a)_I + t(Ax_0 - a)_I,$$

where  $(Ax_0 - a)_I = 0$  and  $1 - t < 0$ . Therefore, one must have  $(Ax - a)_I = 0$  since  $(Ax - a)_I \leq 0$ . Similarly, if  $x \in P$  verifies  $(Ax - a)_I = 0$ , let us show it belongs to  $F$ . Since  $F$  is a face, it suffices to show  $x_t = (1-t)x_0 + tx \in P$  for some small negative  $t$ , because then  $x_0$  is a convex combination of  $x_t$  and  $x$  (which clearly belongs to  $F$ , which implies  $x_t$  and  $x$  are in  $F$ ). Clearly,  $(Ax_t - a)_I = (1-t)(Ax_0 - a)_I + t(Ax - a)_I = (1-t)0 + t0$ . On the remaining indices, since  $(Ax_0 - a)_i < 0$  for  $i$  close enough to zero one still has  $(Ax_t - a)_i < 0$ .  $\square$

Lemmas B.1.1 and B.1.4, which follows, are illustrated in figure B.1.

**Lemma B.1.4** (decomposition into relative interior of faces). *Let  $P := \{x \in \mathbb{R}^n : Ax \leq a, Bx = b\}$  be a polyhedron, one has*

$$P = \bigcup_{F \in \text{faces}(P)} \text{ri}(F),$$

where the union is disjoint and contains  $P$  itself as a face.

*Proof.* First, let us justify the union is disjoint. Let  $F^1$  and  $F^2$  be two faces, using lemma B.1.3 denote by  $I_1$  and  $I_2$  the corresponding index subsets, i.e.,

$$\begin{aligned} F^1 &= \{x \in \mathbb{R}^n : Bx = b, (Ax - a)_{I_1} = 0, (Ax - a)_{I_1^c} \leq 0\} \\ F^2 &= \{x \in \mathbb{R}^n : Bx = b, (Ax - a)_{I_2} = 0, (Ax - a)_{I_2^c} \leq 0\}. \end{aligned}$$

Then, using lemma B.1.1, one has

$$\begin{aligned} \text{ri}(F^1) &= \{x \in \mathbb{R}^n : Bx = b, (Ax - a)_{I_1} = 0, (Ax - a)_{I_1^c} < 0\} \\ \text{ri}(F^2) &= \{x \in \mathbb{R}^n : Bx = b, (Ax - a)_{I_2} = 0, (Ax - a)_{I_2^c} < 0\}. \end{aligned}$$

Now, suppose there exists some  $x \in \text{ri}(F^1) \cap \text{ri}(F^2)$ . Then, for indices in  $I_1 \cap I_2^c$  and  $I_2 \cap I_1^c$ , one must have  $(Ax - a)_i < 0$  and  $(Ax - a)_i = 0$ , which is a contradiction unless  $I_1 = I_2$ , i.e.,  $F^1 = F^2$ .

[ $\subseteq$ ] Let  $x \in P$  and set  $I_0^x := \{i : (Ax - a)_i = 0\}$  as well as  $I_*^x := \{i : (Ax - a)_i < 0\}$ . Let  $F^x$  be the face defined by the subset  $I_0^x$ , one has  $x \in \text{ri}(F^x)$ .

[ $\supseteq$ ] Clear since  $x \in \text{ri}(F) \subseteq F \subseteq P$ . □

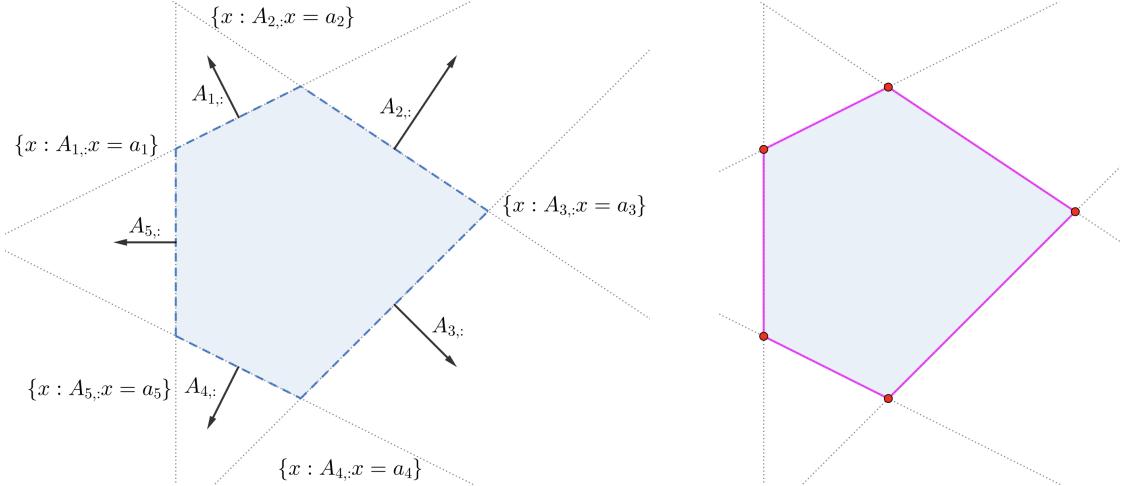


Figure B.1: Illustration of lemmas B.1.1 (left) and B.1.4 (right). On the left, we see that the relative interior is obtained by removing the parts of  $P$  where the equalities  $A_{:,i}x = a_i$  hold. However, consider the same polytope in dimension 3 (thus with empty interior), defined by the additional inequalities  $e_3^\top x \leq 0$  and  $-e_3^\top x \leq 0$ , one cannot take the strict inequalities since it would result in an empty set. This is because these two inequalities actually form an equality ( $e_3^\top x = 0$ ). On the right, we see that the relative interior in blue corresponds to the relative interior of  $P$  (seen as a face of itself), whereas the boundary is composed of the relative interiors of the edges in magenta, then the vertices in red.

The next lemmas define and describe properties of *normal vectors*, called “normals”, to  $\text{face}(t)$ s.

---

**Lemma B.1.5** (faces and “normals”). Let  $P$  be a convex polyhedron, let  $F \subseteq P$ ,  $F$  is a face (of any dimension) if and only if there exists some  $c \in \mathbb{R}^n$  such that  $F = \operatorname{argmin}\{c^\top x : x \in P\}$ . In particular  $F = P \cap H$  where  $H = c^\perp + x_F$  for any  $x_F \in F$ .

*Proof.* If there exists such a  $c$ , for any  $x_1, x_2 \in P$ , one has  $c^\top x \geq c^\top x_1$  and  $c^\top x \geq c^\top x_2$  for any  $x \in P$ . When  $c^\top(x_1 + x_2)/2 = \alpha := \min\{c^\top x : x \in P\}$ , one must have  $c^\top x_1 = \alpha = c^\top x_2$ , so  $x_1, x_2 \in F$ .

Conversely, let  $F = \{x \in P, (Ax)_I = a_I\}$  for some index subset  $I$ , using lemma B.1.5. One must find a  $c$  such that  $F = \operatorname{argmin}\{c^\top x : x \in P\}$ . The optimality conditions read  $c = -A^\top \mu$ ,  $\mu^\top(Ax - a) = 0$ . For  $I^c$ , since  $Ax < a$ , one has  $\mu_{I^c} = 0$ . With  $\mu_I > 0$  let  $\alpha = -a^\top \mu$ . For  $x \in P$ , one has  $c^\top x = -\mu^\top Ax \geq -\mu_I^\top(Ax)_I = \alpha$ . For  $x \in F$ ,  $c^\top x = \alpha$  so  $F \subseteq \operatorname{argmin}\{c^\top x : x \in P\}$ . Reciprocally, for some  $x' \in \operatorname{argmin}\{c^\top x : x \in P\}$ , for all  $x \in P$ , one has  $c^\top x \geq c^\top x'$  which reads  $0 \geq \mu^\top A(x - x')$ . For  $x \in F$ , one has  $0 \geq \mu^\top(a - Ax')$ , and  $\mu_I > 0$  implies  $(Ax')_I = a_I$ .

Now, let  $\alpha := \min\{c^\top x : x \in P\}$ ,  $\operatorname{argmin}\{c^\top x : x \in P\} = P \cap \{x : c^\top x = \alpha\} = P \cap H$ .  $\square$

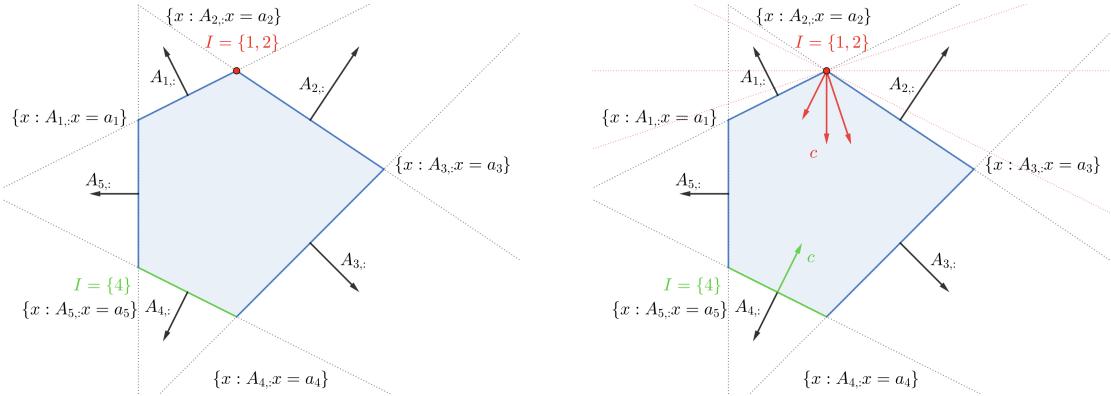


Figure B.2: Illustration of lemmas B.1.3 (left) and B.1.5 (right). On the left, one can observe that the green face corresponds to a set  $I = \{4\}$  of size 1, whereas the vertex in red corresponds to a set  $I = \{1, 2\}$  of size 2. On the right, the same green face has a unique (up to multiplicative constant) normal  $c$  since the face is of maximal dimension  $n - 1$  whereas the vertex in red has multiple noncolinear possible normals  $c$  (see remark B.1.6).

**Remark B.1.6** (faces of full dimension / other faces). Let  $P \subseteq \mathbb{R}^n$  be a polyhedron, if a face  $F$  is of dimension  $n - 1$  then  $\operatorname{aff}(F)$  is a hyperplane and  $c$  is unique up to a multiplicative positive constant. This comes from the fact  $I = \{i\}$  for some  $i$  (or there is some redundancy in the constraints). Whenever the face is of smaller dimension,  $I$  is not reduced to a single index, and since any  $\mu_I > 0$  in the proof of lemma B.1.5 is convenient, there is no uniqueness.  $\square$

Observe that in the previous lemma, one could have used  $\operatorname{argmax}$  instead. This essentially add a minus sign to the normal(s).

**Proposition B.1.7** (outward-pointing normal). *Let  $P \subseteq \mathbb{R}^n$  be a convex polyhedron,  $L = \text{aff}(P) - \text{aff}(P)$  be the subspace parallel to its affine hull,  $F$  be a face of  $P$  and  $L_F = \text{aff}(F) - \text{aff}(F)$ . There exists  $c \in L$  such that  $\|c\| = 1$ ,  $c \in L_F^\perp$  and  $c$  is pointing outwards, i.e.,  $x_F + tc \notin P$  for  $x_F \in F$  and  $t > 0$ .*

*Proof.* Let  $d$  be the vector obtained in lemma B.1.5. Let  $d = c_L + c_\perp$  be its decomposition onto  $L + L^\perp$ . Since  $P \subseteq x_P + L$  for some  $x_P \in P$ , one has

$$\begin{aligned} F = \operatorname{argmin}\{d^\top x : x \in P\} &= \operatorname{argmin}\{c_L^\top x + c_\perp^\top x : x \in P\} \\ &= \operatorname{argmin}\{c_L^\top x + c_\perp^\top x_P : x \in P\} = \operatorname{argmin}\{c_L^\top x : x \in P\} \end{aligned}$$

If  $c_L = 0$ , then  $F = P$ . Otherwise, let  $c := -c_L/\|c_L\|$ . Clearly  $\|c\| = 1$  and  $c \in L$ . Let  $v \in L_F$ , one has  $v = v_1 - v_2$  for some  $v_1$  and  $v_2$  in  $\text{aff}(F)$ . Then, write  $v_i = (1-t_i)x_i + t_i x'_i$  for some  $x_i$  and  $x'_i$  in  $F$  (since it is convex) for  $i \in \{1, 2\}$ . One has thus

$$c^\top v = c^\top((1-t_1)x_1 + t_1 x'_1 - (1-t_2)x_2 - t_2 x'_2) = -\frac{1}{\|c_L\|}[(1-t_1)\alpha + t_1\alpha - (1-t_2)\alpha - t_2\alpha] = 0$$

where  $\alpha$  is the optimal value from the definition of  $F$ . Now, we show that  $c$  is pointing outwards. Let  $x_F \in F$  and  $t > 0$ ,  $c_L^\top(x^F + tc) = \alpha + tc_L^\top(-c_L/\|c_L\|) = \alpha - t\|c_L\| = \alpha - t < \alpha$ , meaning  $x^F + tc \notin P$ .  $\square$

This property essentially takes the opposite, projects on  $L$  then norms the normal(s) described in lemma B.1.5. The elements discussed in this section are related to the notion of normal fan [263, p. 193], which is the decomposition of  $\mathbb{R}^n$  into the (interiors of the) normal cones to the faces.

## B.2 Specific properties of zonotopes

The next properties focus on zonotopes, which are particular polytopes. After recalling the definition of a zonotope, we focus on their face(t)s, with a property illustrated in figure B.3. Its proof is taken from [167].

**Definition B.2.1** (zonotopes). Let  $V \in \mathbb{R}^{n \times m}$  and  $\bar{z} \in \mathbb{R}^n$ , the zonotope  $Z(V, \bar{z})$  is a convex polytope defined by

$$Z(V, \bar{z}) := V[-1, +1]^m + \bar{z}.$$

In particular, it is compact, centrally symmetric around  $\bar{z}$  ( $z + \bar{z} \in Z(V, \bar{z}) \iff \bar{z} - z \in Z(V, \bar{z})$ ). Often, it is considered that  $\bar{z} = 0$ .  $\square$

**Proposition B.2.2** ( $k$ -faces of a zonotope). *Let  $V \in \mathbb{R}^{n \times m}$  and  $Z = V[-1, +1]^m$ ,  $V = [v_1 \dots v_m]$ , the faces of  $Z$  can be expressed as  $F = V_{:, I^F}[-1, +1]^{I^F} + V_{:, I^*} \kappa$  for some  $\kappa \in \{-1, +1\}^{I^*}$  where  $I^F$  are the indices of the generators of the face and  $I^* = [1 : m] \setminus I^F$ . In particular, the  $k$ -faces (faces of dimension  $k$ ) of a zonotope are zonotopes themselves.*

*Proof.* Since  $Z$  is a convex polytope, for a given face  $F$  let  $c$  be a normal given by proposition B.1.7,  $H = c^\top + z^F$  for some  $z^F \in F$ , i.e.,  $H = \{x \in \mathbb{R}^n : c^\top x = c^\top z^F\}$ . In particular, one has  $F = Z \cap H$  and  $Z \subseteq H^-$ . Let  $H_0 = H - H$  and  $I^F := \{i \in [1 : m] : v_i \in H_0\}$ . Define  $\kappa_i$  such that  $\kappa_i v_i \in \text{int } H_0^+$  for  $i \in I^* := [1 : m] \setminus I^F$ . For  $z \in F = Z \cap H$ ,  $c^\top z = \alpha$  reads

$$\alpha = c^\top z = c^\top \sum_{i \in I^*} t_i v_i = c^\top \sum_{i \in I^*} t_i v_i \underset{c^\top t_i v_i \leq c^\top \kappa_i v_i}{\leq} c^\top \sum_{i \in I^*} v_i \kappa_i = c^\top V_{:,I^*} \kappa$$

However,  $V_{:,I^*} \kappa \in Z$  so  $c^\top V_{:,I^*} \kappa \leq \alpha$ . This means  $t_i = \kappa_i$ , so  $z$  has the desired form. Reciprocally, a point  $z \in V_{:,I^F}[-1, +1]^{I^F} + V_{:,I^*} \kappa$  is clearly in  $Z$ . Moreover,  $c^\top z = c^\top V_{:,I^*} \kappa = \alpha$  using the definition of  $\kappa$ .  $\square$

If the zonotope is not centered, i.e.,  $Z = \bar{z} + V[-1, +1]^m$ , the property holds up to a translation of  $\bar{z}$ . This is true for the next proposition, which discusses some technical property of the “normals” of a face (the vectors  $c$  in the previous propositions).

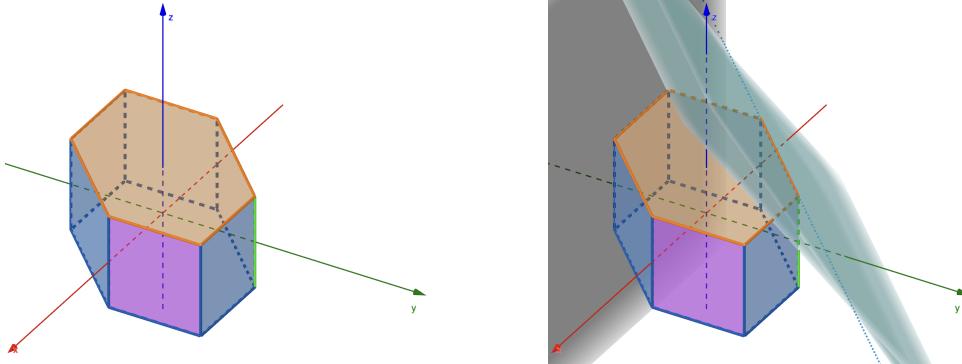


Figure B.3: Example of a simple zonotope with  $V = [e_1 \ e_2 \ e_1 + e_2 \ e_3]$ . The upper face in orange is a face of dimension two, generated by  $e_1, e_2, e_1 + e_2$ , with  $I^* = \{4\}$ . The face in purple at the front is generated by  $e_2$  and  $e_3$ , with  $I^* = \{1, 3\}$ . The face in green on the right, which is an edge, is generated by  $e_3$  with  $I^* = \{1, 2, 3\}$ . All the vertices are also faces with no generators. On the right, some hyperplanes corresponding to normals to faces were added. The full dimensional faces have only one hyperplane but the edges have multiple (since the dimension is 3).

The following property may be new, but is it admittedly rather niche and designed for a specific purpose. It refers to a normal vector given by proposition B.1.7 which is unique (up to a constant factor) only for faces of maximal dimension (dimension  $n - 1$  in a space of dimension  $n$ ). An illustration is proposed in figure B.4.

**Proposition B.2.3** (property of the specific normal). *Let  $V \in \mathbb{R}^{n \times m}$ ,  $Z = V[-1, +1]^m$  be a zonotope,  $F$  be a face of  $Z$ ,  $I^F$  be the set of generators of  $F$ ,  $I^*$  and  $\kappa \in \{\pm 1\}^{I^*}$  such that the center of  $F$  is  $V_{:,I^*} \kappa$  according to proposition B.2.2. Let  $c$  be a “normal” to  $F$  given by proposition B.1.7, then*

$$\kappa \cdot V_{:,I^*}^\top c > 0, \quad \text{i.e.,} \quad \forall i \in I^*, \kappa_i v_i^\top c > 0.$$

*Proof.* Observe that for  $i \in I^*$ ,  $\kappa_i v_i^\top c$  cannot be zero since it implies  $v_i^\top c = 0$  thus  $\text{Vect}(v_i) \subseteq c^\perp = L$  (in  $\mathcal{R}(V)$ ) meaning  $i \in I^F$ . Suppose by contradiction that there exists a partition  $(I_-^*, I_+^*)$  of  $I^*$  such that  $\kappa_{I_-^*} \cdot V_{:,I_-^*}^\top c < 0$  and  $\kappa_{I_+^*} \cdot V_{:,I_+^*}^\top c > 0$  with  $I_-^*$  nonempty (otherwise the result is shown). Then, the point

$$z_* := V_{:,I_+^*} \kappa_{I_+^*} - V_{:,I_-^*} \kappa_{I_-^*} = V_{:,I_+^*} \kappa_{I_+^*} + V_{:,I_-^*} \kappa_{I_-^*} - 2V_{:,I_-^*} \kappa_{I_-^*} = z^F - 2V_{:,I_-^*} \kappa_{I_-^*}$$

is a point of  $Z$  since it reads  $z_* = V\eta$  for some  $\eta \in [-1, +1]^m$ . However,  $c^\top z_* = c^\top z^F - 2 \sum_{i \in I_-^*} \kappa_i v_i^\top c > c^\top z^F$ , which contradicts the definition of  $c$ . Therefore,  $I_-^*$  is empty, which concludes the proof.  $\square$

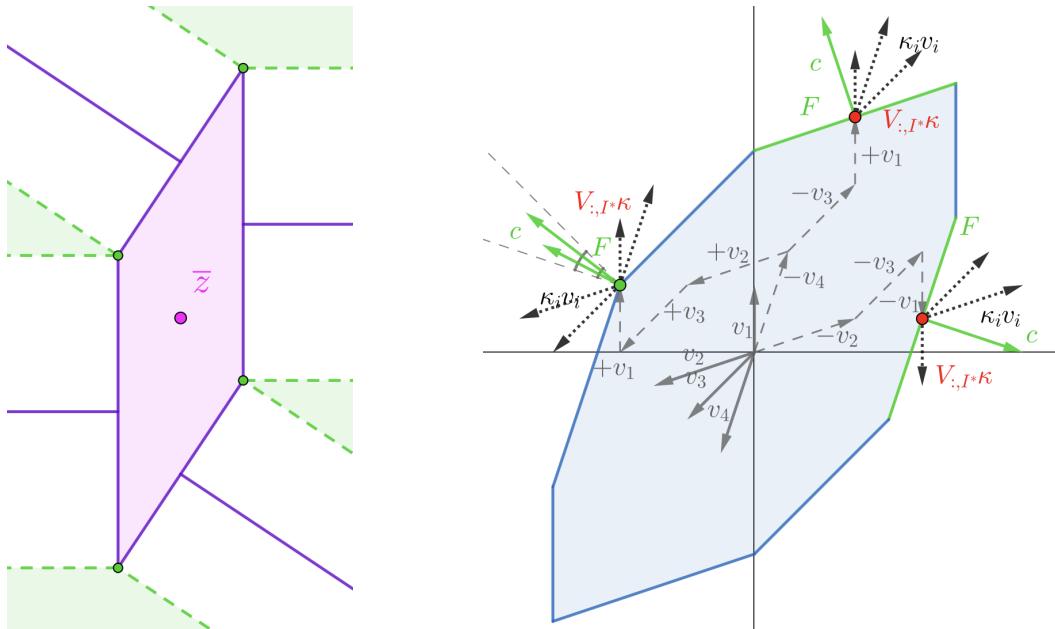


Figure B.4: Left: normal fan of a zonotope. Light green: vertices and their normal cones (dashed boundaries since the boundaries are the normals of the edges); purple: edges and their normal cones. Right: illustration of proposition B.2.3. Green sets: some of the faces considered; green arrows: their normals. Red points: the centers of the faces (equal to the faces for the vertices). Grey: vectors generating the zonotope. Black: the dotted arrows are the  $V_{:,i} \kappa_i$  for  $i \in I^*$ . We translated to the centers of the faces to show more easily the normals  $c$  in green from proposition B.1.7 make a positive scalar product with the black dotted arrows.

**Example B.2.4** (proposition B.2.3 for arbitrary faces). For the zonotope of figure B.3, for the upper face in orange, its normal is  $c = e_3$ , unique up to a positive multiplicative constant since the face is of maximal dimension. The same is true for the purple face, with normal  $c = e_1$ . For the green edge however, any  $c$  such that  $c_3 = 0$ ,  $c_1 < 0$  and  $c_2 + c_1 > 0$  will verify the required properties, since one has

$$F_{\text{green}} = [0; 2; 0] + e_3[-1, 1] = -e_1 + e_2 + (e_1 + e_2) + e_3[-1, 1].$$

Therefore,  $-e_1^\top c > 0$ ,  $(e_1 + e_2)^\top c > 0$  and  $e_2^\top c > 0$ .  $\square$

---

**Remark B.2.5** (proposition B.2.3 for vertices). When the considered face  $F$  is a vertex, any witness point  $d$  of the vertex is a suitable normal. This comes from the fact the vertices are faces with no generators, therefore  $I^* = [1 : m]$ . Moreover, the normals can be taken in the interior of a cone which is the set of directions  $d$  verifying the vertex, since this reads  $\kappa_i v_i^\top c > 0$ , or equivalently “ $s_i v_i^\top d > 0$ ” (chapter 3).  $\square$

Let us finish by observing a rather natural property of zonotopes.

**Proposition B.2.6** (relative interior of zonotopes). *Let  $Z = \bar{z} + V[-1, +1]^m$  for some  $\bar{z} \in \mathbb{R}^n$  and  $V \in \mathbb{R}^{n \times m}$ . A point  $z$  is in the relative interior of  $Z$  if and only if it reads  $z = \bar{z} + V\kappa$ ,  $\kappa \in (-1, +1)^m$ .*

*Proof.* One may use a direct proof using the definition of the relative interior, but using the commutation with affine transformations ([105, proposition 2.17 3)<sub>1</sub>]), one gets

$$\text{ri}(c + V[-1, +1]^m) = c + V\text{ri}([-1, +1]^m) = c + V(-1, +1)^m.$$

$\square$

In particular, the previous proposition can be used on the faces of a zonotope, since they are zonotopes themselves. We finish this section by presenting what the projection on a zonotope looks like. The goal of this explanation comes from chapter 6.

**Remark B.2.7** (projection on a zonotope). Let  $c \in \mathbb{R}^n$ ,  $Z \in \mathbb{R}^{n \times m}$  and consider the zonotope  $\bar{z} + Z[-1, +1]^m$ . The projection of a point  $x \in \mathbb{R}^n$  on  $\bar{z} + Z[-1, +1]^m$  reads

$$\min_{z \in \bar{z} + Z[-1, +1]^m} \frac{1}{2} \|z - x\|^2 = \min_{\xi \in [-1, +1]^m} \frac{1}{2} \|\bar{z} - x + Z\xi\|^2$$

which is a linear least-squares problem with box constraints.

**Proposition B.2.8** (projection and normal). *Let  $Z = \bar{z} + V[-1, +1]^m$  for some  $\bar{z} \in \mathbb{R}^n$  and  $V \in \mathbb{R}^{n \times m}$ . Let  $p \in \mathbb{R}^n$ , let  $z = P_Z(p)$  be the projection of  $p$  on  $Z$ , then the projection of  $p - z$  on  $\mathcal{R}(V)$ ,  $P_{\mathcal{R}(V)}(p - z)$ , verifies the inequalities of propositions B.1.7 and B.2.3 with nonstrict inequalities.*

*Proof.* Since  $Z$  is a nonempty convex, the projection  $P_Z(p)$  is well-defined by the problem

$$\min_{z \in Z} \frac{\|z - p\|^2}{2} = \min_{\xi \in [-1, +1]^m} \frac{\|\bar{z} - p + Z\xi\|^2}{2}$$

which clearly has qualified constraints. The KKT system reads

$$\begin{cases} Z^\top(Z\xi + \bar{z} - p) + \lambda - \mu = 0, \\ \lambda^\top(\xi - 1) = 0, \\ \mu^\top(-1 - \xi) = 0. \end{cases}$$

Then, denote  $I_-^* := \{i \in [1 : m] : \xi_i = -1\}$  and  $I_+^* := \{i \in [1 : m] : \xi_i = +1\}$ . Using the complementarity conditions, one gets  $i \in I_-^* \Rightarrow \mu_i = 0$  and  $i \in I_+^* \Rightarrow \lambda_i = 0$ . The KKT system becomes then

$$\begin{cases} i \in I_+^*, & z_i^\top(p - \bar{z} - Z\xi) = \lambda_i, \\ i \in I_-^*, & z_i^\top(p - \bar{z} - Z\xi) = -\mu_i, \\ \xi_i \in (-1, +1), & z_i^\top(p - \bar{z} - Z\xi) = 0, \end{cases} \Leftrightarrow \begin{cases} i \in I_+^*, & z_i^\top(p - \bar{z} - Z\xi) \geq 0, \\ i \in I_-^*, & z_i^\top(p - \bar{z} - Z\xi) \leq 0, \\ \xi_i \in (-1, +1), & z_i^\top(p - \bar{z} - Z\xi) = 0. \end{cases}$$

Observe that, as described in counterexample B.2.9, there is not necessarily strict complementarity, i.e., one does not have  $i \in I_+^* \iff \lambda_i > 0$  and  $i \in I_-^* \iff \mu_i < 0$ . Then, the previous system reads

$$\begin{cases} \xi_i \in \{-1, +1\}, & \xi_i z_i^\top(p - \bar{z} - Z\xi) \geq 0, \\ \xi_i \in (-1, +1), & z_i^\top(p - \bar{z} - Z\xi) = 0. \end{cases}$$

Moreover, let  $c := p - \bar{z} - Z\xi$  be the direction from the projection to the projected point,  $c$  verifies the inequalities of the normal up to the strictness. To conclude, observe that dividing the vector by its norm does not change the inequalities. Moreover, the projection on  $\mathcal{R}(V)$  maintains the inequalities as well: let  $p - z = P_{\mathcal{R}(V)}(p - z) + [p - z - P_{\mathcal{R}(V)}(p - z)]$ , by definition of the orthogonal projection on a vector subspace, in the products defining the inequalities the second term disappears (since  $v_i \in \mathcal{R}(V)$  is orthogonal to it), so the only remaining term is  $P_{\mathcal{R}(V)}(p - z)$ . Finally, it is clearly pointing outwards.  $\square$

Observe that the projection on a zonotope is a constrained least-squares problem. The following counterexample illustrates the (possible) absence of strictness in the described equations. This will have some relevance in the later sections.

**Counter-example B.2.9** (no strict complementarity / normal). Consider the following data, there is no strict complementarity in the KKT system, i.e., the difference  $p - z = p - P_Z(p)$  is not a strict normal. It is illustrated in figure B.5.

$$p = [-2; 6], \quad V = \begin{bmatrix} 0 & 1 \\ 5/2 & 3/2 \end{bmatrix}, \quad \bar{z} = 0$$

It is clear that the projection of  $p$  on  $Z$  is  $z = [1; 4]$  and  $\xi = [1; 1]$ . However, the KKT system is

$$Z^\top(Z\xi + \bar{z} - z) = \begin{bmatrix} 0 & 5/2 \\ 1 & 3/2 \end{bmatrix} \left( \begin{bmatrix} 0 & 1 \\ 5/2 & 3/2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 0 - \begin{bmatrix} -2 \\ 6 \end{bmatrix} \right) = \begin{bmatrix} 0 & 5/2 \\ 1 & 3/2 \end{bmatrix} \begin{bmatrix} 1+2 \\ 4-6 \end{bmatrix} = \begin{bmatrix} -5 \\ 0 \end{bmatrix}$$

So  $\mu = 0$  and  $\lambda = [5; 0]$  solves the system but without strict complementarity.  $\square$

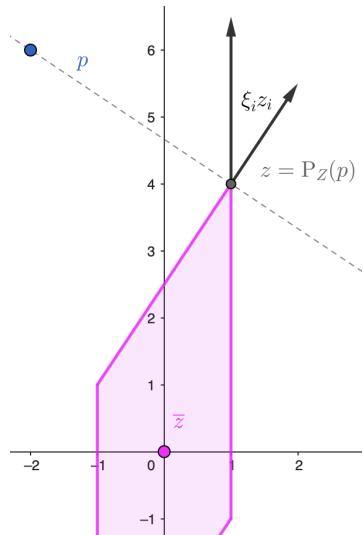


Figure B.5: Example of a point such that the direction point – projection does not verify strict complementarity. This occurs at points where the projection is not differentiable. (The distance itself is differentiable outside of the boundary of the (closed) convex).

# Appendix C

## Inclusion of zonotopes

This appendix details relevant notions about the zonotope inclusion problem. While this question was essentially answered in [223] and [149], where it is shown that determining if a zonotope is contained in another zonotope is, in general, co-NP-complete, we detail some related elements that may improve understanding of this problem. In particular, the approximate algorithm in [223] does not correspond ideally to what is required in chapter 6, since we mostly want to *disprove* inclusion – if possible without combinatorial enumeration.

This appendix details the work of [223]. In particular, let us mention that the considered framework is such that the dimension  $n$  of the space containing the zonotope is greater than the number of generators  $m$ . This rather unusual setting is such that the theorem 3 of [223], theorem C.0.1 below, may be a NSC (the article assumes  $m > n$  so it cannot occur).

**Theorem C.0.1** (theorem 3 of [223]). *Let  $Z_x = \bar{x} + X[-1, +1]^q$  and  $Z_y = \bar{y} + Y[-1, +1]^p$  for  $\bar{x}, \bar{y} \in \mathbb{R}^n$ ,  $X \in \mathbb{R}^{n \times q}$  and  $Y \in \mathbb{R}^{n \times p}$ . If there exist  $\Delta \in \mathbb{R}^{p \times q}$ ,  $\beta \in \mathbb{R}^p$  such that the following condition holds, then  $Z_x \subseteq Z_y$ .*

$$Y[\Delta \beta] = [X \bar{y} - \bar{x}], \|\Delta \beta\|_\infty \leq 1$$

□

In the theorem, the infinite norm of the matrix  $[\Delta \beta]$  is defined as the maximum of the 1-norm of the lines, i.e.,

$$\|\Delta \beta\|_\infty = \max_{i \in [1:p]} \|\Delta_{i,:} \beta_i\|_1$$

Two questions then arise. When the condition is not verified, how to verify if whether the inclusion holds or not. And in any case, how to obtain a point confirming there is no inclusion. Let us detail these properties. First, observe that, since the problem is co-NP-complete and clearly has a combinatorial nature, one could verify if every vertex (see also chapter 3 for the computation of the vertices) is contained in the second zonotope or not and stop at the first one not belonging to it.

The condition of theorem C.0.1 can be checked by solving the following optimization

---

problem

$$\begin{aligned} \min \quad & \|[\Delta \beta]\|_\infty \\ \text{s.t.} \quad & Y[\Delta \beta] = [X \bar{y} - \bar{x}] \end{aligned} \tag{C.1}$$

and checking if the optimal cost is  $\leq 1$ . When the optimum is  $> 1$ , the inclusion  $Z_x \subseteq Z_y$  may still hold: see example 2 given in [223]<sup>1</sup>. But it is not clear how to find a point in  $Z_x \setminus Z_y$  when the inclusion does not hold. Let us first consider the case where the linear constraints are not feasible.

**Proposition C.0.2** (infeasible constraints – 1). *If  $Y\beta = \bar{y} - \bar{x}$  is not feasible, then the point  $\bar{x} = \bar{x} + X0$  does not belong to  $Z_y = \bar{y} + Y[-1, +1]^p$ .*

*Proof.* If the constraint is not feasible, then  $\bar{y} - \bar{x} \notin \mathcal{R}(Y)$ . This also reads  $\bar{x} \notin \bar{y} + \mathcal{R}(Y)$ . Since  $\bar{y} + Y[-1, +1]^p \subseteq \bar{y} + \mathcal{R}(Y)$ , this means that  $\bar{x}$  does not belong to  $Z_y$ .  $\square$

**Proposition C.0.3** (infeasible constraints – 2). *If  $Y\Delta = X$  is not feasible, then one can find in polynomial time a point verifying the inclusion doesn't hold.*

*Proof.* If these constraints cannot be verified, then for some index  $j \in [1 : q]$ ,  $X_{:,j} \notin \mathcal{R}(Y)$  (this index can be found in polynomial time). Suppose that  $\bar{x} + X_{:,j}$  and  $\bar{x} - X_{:,j}$  both belong to  $Z_y$ . By convexity of  $Z_y$ , one has  $[\bar{x} - X_{:,j}, \bar{x} + X_{:,j}] \subseteq Z_y$ . This implies that  $\bar{x} + \text{vect}(X_{:,j}) \subseteq \bar{y} + \mathcal{R}(Y)$ . By using a dimensional argument, one gets that  $\text{vect}(X_{:,j}) \subseteq \mathcal{R}(Y)$ , which is a contradiction. Thus,  $\bar{x} + X_{:,j}$  and/or  $\bar{x} - X_{:,j}$  does not belong to  $Z_y$ .  $\square$

These two simple cases, in which the optimal value is  $+\infty$ , mean that either the centers are not “aligned” (proposition C.0.2) or  $Z_x$  extends in a dimension not generated in  $Y$  (proposition C.0.3). While they are essentially irrelevant in [223], since the zonotopes have full dimension, it is reasonable to consider, in the setting of chapter 6, that the hypotheses of these propositions may be true.

Let us mention a possible similarity in what follows: the part of the problem with  $\beta$  and  $\bar{y} - \bar{x}$  plays a slightly different role than the columns of  $X$  and  $\Delta$ , which may be reminiscent of sections 5.3.4 and 5.6 of chapter 5, where the affine part of the arrangement,  $\tau$ , has a role inbetween dimension  $n$  and dimension  $n + 1$ .

In what follows, we assume (C.1) is feasible, meaning, by the properties of linear optimization (see [105, 29]), that it has a finite optimal value (otherwise propositions C.0.2 and C.0.3 answer the question) and that its optimal value is  $\lambda^* > 1$  (otherwise there is inclusion).

Proposition C.0.4 below gives a possible meaning to the value of  $\lambda^*$  (slightly different from  $\lambda_{\text{Theorem3}}$  of [223] and closer to the zonotope norm used in [149]):  $\lambda^*$  serves as a guaranteed dilation factor for  $Z_y$  to contain  $Z_x$  (since  $X$  is generated by  $Y$ ). It is illustrated in example C.0.5.

---

<sup>1</sup>We conjecture it is impossible to have a counterexample in dimension 2, as the authors suggest.

**Proposition C.0.4** (dilation by  $\lambda^*$  finite). *One has  $\bar{x} + X[-1, +1]^q \subseteq \bar{y} + \lambda^*Y[-1, +1]^p$ .*

*Proof.* Using the constraints of the problem, one has

$$\begin{aligned}\bar{x} + X[-1, +1]^q &= \bar{y} - Y\beta + Y\Delta[-1, +1]^q = \bar{y} + Y[\Delta[-1, +1]^q - \beta] \\ &= \bar{y} + Y[\Delta \beta] \begin{bmatrix} [-1, +1]^q \\ -1 \end{bmatrix} \subseteq \bar{y} + \lambda^*Y[-1, +1]^p\end{aligned}$$

using the definition of  $\lambda^* = ||[\Delta \beta]||_\infty$ .  $\square$

In particular, the proof shows the aforementioned slight difference between the columns of  $X$  (indices  $[1 : q]$ ) and  $\bar{y} - \bar{x}$ .

**Example C.0.5** (the value of  $\lambda^*$ ). Consider the following data, where  $\lambda^* = 6$ :

$$(\bar{x}, \bar{y}, X, Y) = \left( \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \frac{1}{2} \begin{bmatrix} 0 & 1 & 1 \\ 2 & 2 & -2 \end{bmatrix}, \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right), [\Delta \beta] = \frac{1}{2} \begin{bmatrix} 2 & 3 & -1 & 2 \\ -2 & -1 & 3 & 6 \end{bmatrix} \quad \square$$

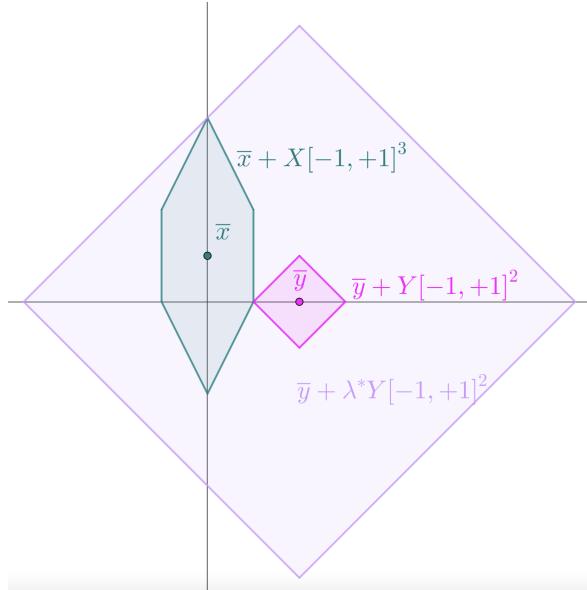


Figure C.1: In this particular example, the (unique) solution  $(\Delta, \beta)$  of the problem detailed in example C.0.5 yields  $\lambda^* = 6$ . We observe that when dilating  $Z_y$  by  $\lambda^*$ , one has  $Z_x \subseteq \bar{y} + \lambda^*Y[-1, +1]^2$ , but dilating by any  $\lambda < \lambda^*$ , the inclusion does not hold (the top point of the green area is not contained in the light purple area).

When  $Z_x \subseteq Z_y$  but  $\lambda^* > 1$ , i.e., theorem C.0.1 fails, the value of  $\lambda^*$  may be imprecise (it is some sort of overestimation) since  $Z_y = \bar{y} + Y[-1, +1]^p \subsetneq \bar{y} + \lambda^*Y[-1, +1]^p$ . Nonetheless, the theorem becomes a necessary and sufficient condition if  $Y$  has independent columns. This observation is not told in [223] since its setting assumes the dimension  $n$  is smaller than the numbers of generators  $m$ , which is the usual assumption for zonotopes. However,

---

this is not the case in our setting. Before detailing this case, let us derive the dual problem of (C.1). Using Lagrangian duality<sup>2</sup>, one gets

$$\begin{aligned} \max_{\Lambda} & \quad (\Lambda, [X \bar{y} - \bar{x}]) \\ \text{s.t.} & \quad \|Y^T \Lambda\|_1 \leq 1 \end{aligned} \tag{C.2}$$

where the involved norm, the 1-norm over matrices, is dual to the previous norm and defined as the sum of the infinite norm of the lines, i.e.,  $\|M\|_1 = \sum_{i=1}^p \|M_{i,:}\|_\infty$ . Moreover, as one can reformulate these problems as linear optimization problems, strong duality holds.

**Remark C.0.6** (1-norm of matrices with special form). Let  $M = ms^T$  for some  $m \in \mathbb{R}^n$  and  $s \in \{\pm 1\}^l$ , then  $\|M\|_1 = \sum_{i=1}^n \|m_i s^T\|_\infty = \sum_{i=1}^n |m_i| = \|m\|_1$ .  $\square$

**Proposition C.0.7** ( $Y$  is of full rank / injective). *When  $Y$  has independent columns, and  $\lambda^* > 1$ , one can obtain a point  $\bar{x} + X\eta \notin Z_y$  for some  $\eta \in [-1, +1]^q$ .*

*Proof.* If  $Y$  has independent columns, since  $Y \in \mathbb{R}^{n \times p}$  and  $n > p$ ,  $Y$  is injective and  $Y^T Y$  is nonsingular. Therefore, the only primal solution is

$$[\Delta \beta] = (Y^T Y)^{-1} Y^T [X \bar{y} - \bar{x}].$$

Let  $i^* \in [1 : p]$  be such that  $\lambda^* = \|[\Delta \beta]_{i,:}\|_1$  and  $z := Y(Y^T Y)^{-1} e_{i^*}$ , in particular  $z \in \mathcal{R}(Y)$  and  $Y^T z = e_{i^*}$ . We construct a dual variable of the form  $\Lambda = z s^T$  that attains the value  $\lambda^*$  for some well-chosen  $s \in \{\pm 1\}^{q+1}$ . First, observe that

$$\|Y^T \Lambda\|_1 = \|Y^T z s^T\|_1 = \|(Y^T z) s^T\|_1 = \|e_{i^*} s^T\|_1 = \|s^T\|_\infty = 1$$

where  $\|s^T\|_\infty$  is the infinity norm on  $\mathbb{R}^{q+1}$  (for a row vector). The dual cost reads

$$\begin{aligned} (\Lambda, [X \bar{y} - \bar{x}]) &= (z s^T, Y [\Delta \beta]) = (Y^T z s^T, [\Delta \beta]) \\ &= (e_{i^*} s^T, [\Delta \beta]) = (s, [\Delta \beta]_{i^*,:}) = \sum_{j=1}^q s_j \Delta_{i^*,j} + s_{q+1} \beta_{i^*} \end{aligned}$$

Now, take  $s_{q+1} = \text{sgn}(\beta_{i^*})$  and  $s_j = \text{sgn}(\Delta_{i^*,j})$  for  $j \in [1 : q]$  (if  $\beta_{i^*}$  or  $\Delta_{i^*,j} = 0$ , choose the corresponding sign arbitrarily). The dual cost becomes:

$$\sum_{j=1}^q s_j \Delta_{i^*,j} + s_{q+1} \beta_{i^*} = \sum_{j=1}^q |\Delta_{i^*,j}| + |\beta_{i^*}| = \lambda^*$$

---

<sup>2</sup>Indeed, let  $\tilde{\Delta} := [\Delta \beta]$ ,

$$\begin{aligned} \min_{\tilde{\Delta}} \max_{\Lambda} \|\tilde{\Delta}\|_\infty + (-Y \tilde{\Delta} + [X \bar{y} - \bar{x}], \Lambda) &\iff \max_{\Lambda} \min_{\tilde{\Delta}} (\Lambda, [X \bar{y} - \bar{x}]) + \|\tilde{\Delta}\|_\infty - (\tilde{\Delta}, Y^T \Lambda) \\ &\iff \max_{\Lambda} (\Lambda, [X \bar{y} - \bar{x}]), \text{s.t. } \|Y^T \Lambda\|_1 \leq 1 \end{aligned}$$

Finally, consider  $z^* = \bar{x} + X(-\text{sgn}(\beta_{i^*})s_{[1:q]})$  assuming  $\beta_{i^*} \neq 0$  for now. Let us show this point does not belong to  $Z_y$ . Using equations 5 and 6 from [149], one has

$$\bar{x} + Xs_{[1:q]} \in Z_y \iff \begin{array}{l} \min \quad ||\zeta||_\infty \\ \text{s.t.} \quad \bar{x} + Xs_{[1:q]} = \bar{y} + Y\zeta \end{array} \text{ has optimal value } \leq 1. \quad (\text{C.3})$$

Indeed, when the second condition occurs  $\bar{x} + Xs_{[1:q]}$  can be expressed as a point of  $Z_y$ . Using the primal constraints and the injectivity of  $Y$ , one obtains

$$\begin{aligned} & \bar{x} + X(-\text{sgn}(\beta_{i^*})s_{[1:q]}) = \bar{y} + Y\zeta \\ \iff & \bar{x} - \bar{y} + X(-\text{sgn}(\beta_{i^*})s_{[1:q]}) = Y\zeta \\ \iff & -Y(\beta + \text{sgn}(\beta_{i^*})\Delta s_{[1:q]}) = Y\zeta \\ \iff & -\beta - \text{sgn}(\beta_{i^*})\Delta s_{[1:q]} = \zeta \end{aligned}$$

where  $\zeta$  is uniquely determined. Then, one has:

$$\zeta_{i^*} = -\beta_{i^*} - \text{sgn}(\beta_{i^*}) \sum_{j=1}^q s_j \Delta_{i^*,j} = -\beta_{i^*} - \text{sgn}(\beta_{i^*}) \sum_{j=1}^q |\Delta_{i^*,j}| = -\text{sgn}(\beta_{i^*})\lambda^* \notin [-1, +1].$$

If  $\beta_{i^*} = 0$ , let  $z^* = \bar{x} + X(\pm s_{[1:q]})$ , one gets

$$\begin{aligned} & \bar{x} \pm Xs_{[1:q]} = \bar{y} + Y\zeta \\ \iff & \bar{x} - \bar{y} \pm Xs_{[1:q]} = Y\zeta \\ \iff & -Y\beta \pm Y\Delta s_{[1:q]} = Y\zeta \\ \iff & -\beta \pm \Delta s_{[1:q]} = \zeta \end{aligned}$$

and  $\zeta_{i^*} = 0 \pm \sum_{j=1}^q s_j \Delta_{i^*,j} = \pm \lambda^* \notin [-1, +1]$ .  $\square$

The case of  $Y$  of full rank occurs for instance in example C.0.5. Observe that the indices for which the sign  $s_j$  is chosen arbitrarily do not intervene in the final computations. The hypothesis on  $Y$  is a relatively strong assumption. However, when it does not hold, one may still obtain a dual variable of the same form. From there, a similar reasoning can be employed.

**Proposition C.0.8** (dual solution of the form  $zs^\top$ ). *Let  $\Lambda = zs^\top$  for  $z \in \mathbb{R}^n$  and  $s \in \{\pm 1\}^{q+1}$  be a dual solution, assuming  $\lambda^* > 1$  and  $s_{q+1} = -1$  (up to taking  $(-z, -s)$ ). Then  $Z_x \not\subseteq Z_y$ .*

*Proof.* Let us show that the point  $\bar{x} + Xs_{[1:q]}$  does not belong to  $Z_y$ . As in (C.3), the primal and dual problems are

$$(P) \quad \left\{ \begin{array}{l} \min \quad ||\zeta||_\infty \\ \text{s.t.} \quad \bar{x} + Xs_{[1:q]} = \bar{y} + Y\zeta \end{array} \right. \quad (D) \quad \left\{ \begin{array}{l} \max \quad w^\top Y(\Delta s_{[1:q]} - \beta) \\ \text{s.t.} \quad ||Y^\top w||_1 \leq 1 \end{array} \right. \quad (\text{C.4})$$

We show that, in the dual problem,  $z$  is a variable such that the dual cost is  $\lambda^* > 1$ . By (strong) duality, the primal cost is also  $\lambda^* > 1$ . Indeed, the evaluation of the dual cost in  $z$  reads

$$(Y^\top z)^\top (\Delta s_{[1:q]} - \beta) = (Y^\top z)^\top (\Delta s_{[1:q]} + \beta s_{q+1}) = (Y^\top z)^\top [\Delta \beta] s$$

---

whereas the dual cost of the problem with  $\Lambda$  reads

$$\lambda^* = (\Lambda, [X \bar{y} - \bar{x}]) = (Y^\top z, [\Delta \beta] s). \quad \square$$

In particular, this does not hold for the example 2 of [223]. The following table summarizes the cases treated so far.

case	$Z_x \subseteq Z_y$ ?	parameter recovery	cost
$\lambda^* \leq 1$	yes	nothing to do	LOP
$\text{rank}(Y) = p$	no	possible	LOP
dual solution $\Lambda = zs^\top$	no	given by $s$	LOP

Let us show now that one also has the reverse implication of proposition C.0.8, i.e., when there is no inclusion but the problems are feasible there exists a dual solution with colinear columns. The proof uses some properties of zonotopes (and polytopes) detailed in appendix B.

**Proposition C.0.9** (contrapositive of proposition C.0.8). *If  $Z_x \not\subseteq Z_y$  but the problems are feasible, there exists a dual solution of the form  $\Lambda = zs^\top$ .*

*Proof.* Since  $Z_x \not\subseteq Z_y$ , there exists some  $\eta \in \{\pm 1\}^q$  such that  $\bar{x} + X\eta \in \bar{y} + \lambda^* Y[-1, +1]^p$ , which reads

$$\bar{x} + X\eta = \bar{y} + \lambda^* Y\zeta \quad \text{for some } \zeta \in [-1, +1]^p.$$

Let us show that, for a suitable pair  $(z, s)$ , one can find a dual solution  $\Lambda = zs^\top$  for  $s_{[1:q]} = \eta$  and  $s_{q+1} = -1$  with a suitable  $z$ . First, observe that  $\bar{y} + \lambda^* Y\zeta$  belongs to a face of  $\bar{y} + \lambda^* Y[-1, +1]^p$ . Therefore, there exists a certain normal  $c$  as described in proposition B.1.7. Since  $c$  can be multiplied by a positive constant, assume that  $\|Y^\top c\|_1 = 0$  and let  $z = c$ . The dual constraint reads

$$\|Y^\top z s^\top\|_1 = \sum_i \|y_i^\top z s^\top\|_\infty = \sum_i |y_i^\top z| = \|Y^\top c\|_1 = 1.$$

Moreover, the dual cost reads

$$(zs^\top, [X \bar{y} - \bar{x}]) = (z, [X \bar{x} - \bar{y}][\eta; 1]) = (c, X\eta + \bar{x} - \bar{y}) = \lambda^*(Y^\top c, \zeta) = \lambda^* \sum_{i=1}^p \zeta_i y_i^\top c$$

and using the properties of  $c$  from proposition B.2.3, one has  $\zeta_i y_i^\top c \geq 0$ . If  $\zeta_i \in (-1, +1)$ ,  $y_i$  is a generator of the face and  $y_i^\top c = 0$ , so either  $y_i^\top c = 0$ , or  $\zeta_i \in \{-1, +1\}$  and thus the sum equals  $\|Y^\top c\| = 1$ , which concludes the proof.  $\square$

If  $\zeta_i \in (-1, +1)$ ,  $y_i$  is a generator of the face and  $y_i^\top c = 0$ , so either  $y_i^\top c = 0$ , or  $\zeta_i \in \{-1, +1\}$  and thus the sum equals  $\|Y^\top c\| = 1$ , which concludes the proof.

To conclude, let us summarize these various observations into a possible algorithm.

**Algorithm C.0.10** (solving the zonotope inclusion). Input:  $\bar{x}, \bar{y}, X, Y$ . Output: boolean answer to  $Z_x \subseteq Z_y$  and point  $z$  verifying the noninclusion if so. When relevant, one also returns  $\eta$  such that  $\bar{x} + X\eta \notin \bar{y} + Y[-1, +1]^p$ . The following cases are disjoint.

1. *Primal constraints infeasible.* If the constraint  $Y\beta = \bar{y} - \bar{x}$  is not feasible, there is no inclusion, return (`FALSE`,  $z = \bar{x} + X0$ ,  $\eta = 0$ ). If the constraint  $Y\Delta = X$  is not feasible, obtain an index  $j \in [1 : q]$  such that  $X_{:,j} \notin \mathcal{R}(Y)$ , verify whether  $\bar{x} \pm X_{:,j} = \bar{x} \pm Xe_j$  does not belong to  $\bar{y} + Y[-1, +1]^p$ , return (`FALSE`,  $z = \bar{x} \pm Xe_j$ ,  $\eta = e_j$ ).
2. *Primal check.* If (C.1) is feasible and  $\lambda^* \leq 1$ , then  $Z_x \subseteq Z_y$ , return (`TRUE`).
3. *Case of  $Y$  with independent columns.* If  $Y$  is of full rank, use proposition C.0.7 to obtain the  $\eta$  such that  $z = \bar{x} + X\eta \notin Z_y$ , return (`FALSE`,  $z$ ,  $\eta$ ).
4. *Dual check.* Solve the dual problem (C.2). If the dual variable has columns all equal up to the sign, thus  $\Lambda = zs^\top$ , let  $z$  and  $s$  such that  $s_{q+1} = -1$  and let the point  $z = \bar{x} + Xs_{[1:q]} \notin \bar{y} + Y[-1, +1]^p$  verify the inclusion does not hold, return (`FALSE`,  $z = \bar{x} + X\eta$ ,  $\eta$ )
5. *Dual problem with opposite/equal columns.* Solve the dual problem by imposing the columns to be equal or opposite. If such a solution has value  $\lambda^*$ ,  $Z_x \not\subseteq Z_y$ , let  $\eta := -\text{sgn}(s_{q+1})s_{[1:q]}$ ,  $\bar{x} + X\eta \notin \bar{y} + Y[-1, +1]^p$ , return (`FALSE`,  $z = \bar{x} + X\eta$ ,  $\eta$ )
6. *Dual problem has no solution with colinear columns.* Then the theorem “fails” and there is inclusion, return (`TRUE`).

In algorithm C.0.10, step 5 must either use nonlinear constraints of the form  $\Lambda_{:,j} = s_j\Lambda_{:,q+1}$  for instance or binary variables (among possible reformulations). In both cases, the problem is not polynomial anymore. Steps 4 and 5 differ in the sense that even when the inclusion does not hold, therefore there exists a dual solution of the form  $\Lambda = zs^\top$ , the linear solver may or may not obtain it<sup>3</sup>.

Algorithm C.0.10 may look complicated, the reason being its formulation tries to avoid the combinatorial problem if possible. One could verify every vertex of the zonotope and check if it belongs to the other zonotope.

---

<sup>3</sup>Solvers such as GUROBI may obtain it naturally even without imposing explicitly the constraint as in step 5.



# Appendix D

## Weights and element of Clarke's differential

The goal of this appendix is to give a proof of proposition 6.1.20. The proof turned out to be rather long and technical – there may be simpler paths using the properties of the Levenberg-Marquardt approach. Along the way, we discuss various related technicalities.

### D.1 A geometric property on sign vector sets

This brief section aims discussing a rather abstract result which explains some difficulties encountered later, in the main proof of this appendix. It discusses a property of independent subsets of vectors and the impact on the corresponding subarrangements. In what follows,  $X$  and  $Y$  are matrices with the same number of lines, and recall that  $[X \ Y]$  (resp.  $[X \ -Y]$ ) is the horizontal concatenations of  $X$  and  $Y$  (resp.  $X$  and  $-Y$ ). This proposition seems to generalize proposition 3.3.17. Since it is mainly used to explain some observations below, we did not put it into chapter 4.

**Proposition D.1.1**  $((X, Y), (X, -Y)$  and ranges). *Let  $X \in \mathbb{R}^{n \times q}$  and  $Y \in \mathbb{R}^{n \times p}$ , the following equivalence holds:*

$$\mathcal{R}(X) \cap \mathcal{R}(Y) = \{0\} \iff \mathcal{S}([X \ Y]) = \mathcal{S}([X \ -Y]). \quad (\text{D.1})$$

The meaning of the left-hand side is that the vectors of  $X$  and  $Y$  belong to subspaces with empty intersection, whereas the right-hand side means that the sign vector set is invariant if some part of the vectors are reversed (for what is needed below, it means the zonotopes corresponding to  $[X \ Y]$  and  $[X \ -Y]$  have vertices with the same sign vectors identifications, see below). Recall that “support” refers to a subset of indices such that a considered quantity has nonzero components on the indices of its support.

---

*Proof.*  $\Rightarrow$ ] Let us show the contrapositive by leveraging the stem vectors framework. If the right-hand side equality does not hold, there exists some sign vector in  $\mathcal{S}([X \ Y]) \setminus \mathcal{S}([X - Y])$  (since opposite vectors yield the same hyperplane, the hyperplanes in both arrangements are the same, so there is the same number of sign vectors in both, thus after assuming they are different, one cannot be included in the other). Let  $s$  be this sign vector,  $s \notin \mathcal{S}([X - Y])$  reads, with  $s = (s^x, s^y)$  the parts corresponding to the indices of  $X$  and those of  $Y$

$$\begin{aligned}\exists \alpha = (\alpha_x, \alpha_y) \in \mathbb{R}_+^{q+p} \setminus \{0\}, \sum (\alpha_x)_i s_i^x X_{:,i} + \sum (\alpha_y)_i s_i^y (-Y_{:,i}) &= 0 \\ &= X \operatorname{Diag}(s^x) \alpha_x - Y \operatorname{Diag}(s^y) \alpha_y = 0.\end{aligned}$$

Observe that one cannot have  $X \operatorname{Diag}(s^x) = 0$  since it would imply  $Y \operatorname{Diag}(s^y) \alpha_y = 0$ , meaning  $(0, \alpha_y) \neq 0$  would be a stem vector, but this contradicts the fact  $s \in \mathcal{S}([X \ Y])$ . Similarly,  $Y \operatorname{Diag}(s^y) \alpha_y = 0$  is impossible. Therefore,  $d := X \operatorname{Diag}(s^x) \alpha_x = Y \operatorname{Diag}(s^y) \alpha_y$  is nonzero, meaning we have found a nonzero  $d \in \mathcal{R}(X) \cap \mathcal{R}(Y)$ .

$\Leftarrow$ ] Suppose that both  $\mathcal{S}([X \ Y]) = \mathcal{S}([X - Y])$  and  $\mathcal{R}(X) \cap \mathcal{R}(Y) \neq \{0\}$  hold, and let us observe a contradiction. First, let us show a consequence of the assumption  $\mathcal{S}([X \ Y]) = \mathcal{S}([X - Y])$ . For any pair of subsets  $J_x \subseteq [1 : q]$  and  $J_y \subseteq [q + 1 : q + p]$ , one has  $\mathcal{S}([X_{:,J_x} \ Y_{:,J_y}]) = \mathcal{S}([X_{:,J_x} - Y_{:,J_y}])$ . Indeed, if this inequality does not hold, there exists a sign vector  $s$  belonging to one of the two sets and not the other (for the same reasons as in the other implication). Therefore, using the fact that (with a natural reordering of the indices)<sup>1</sup>

$$\begin{cases} \mathcal{S}([X \ Y]) & \subseteq \mathcal{S}([X_{:,J_x} \ Y_{:,J_y}]) \times \{\pm 1\}^{J_x^c \cup J_y^c} \\ \mathcal{S}([X - Y]) & \subseteq \mathcal{S}([X_{:,J_x} - Y_{:,J_y}]) \times \{\pm 1\}^{J_x^c \cup J_y^c} \end{cases},$$

and that every sign vector has at least one descendant, the sign vector  $s$  has a descendant not belonging to the other sign vector set with all the columns of  $X$  and  $(\pm)Y$ .

The negation of the left-hand side means there exists some nonzero  $d \in \mathbb{R}^n$  such that

$$\operatorname{Vect}(d) \subseteq \mathcal{R}(X) \cap \mathcal{R}(Y), \tag{D.2a}$$

which also reads

$$Xd^x = d = Yd^y, \quad \text{for some } d^x \in \mathbb{R}^q \text{ and } d^y \in \mathbb{R}^p. \tag{D.2b}$$

Moreover, let  $B_x \subseteq [1 : q]$  be a base of  $\mathcal{R}(X)$  and suppose, without loss of generality, that  $d_{B_x^c}^x = 0$ , i.e., the support of  $d^x$  is contained in  $B_x$ . Let us express (D.2b) as

$$Xd^x - Yd^y = 0, \quad \text{or} \quad X \operatorname{Diag}(\operatorname{sgn}(d^x))|d^x| - Y \operatorname{Diag}(\operatorname{sgn}(d^y))|d^y| = 0, \tag{D.2c}$$

where  $0 \times \operatorname{sgn}(0) = 0$ . Under this last form, one recognizes, by Gordan's alternative, that

$$(s^x, s^y) := (\operatorname{sgn}(d^x), \operatorname{sgn}(d^y)) \tag{D.2d}$$

---

<sup>1</sup>And using  $J_x^c = [1 : q] \setminus J_x$ ,  $J_y^c = [q + 1 : q + p] \setminus J_y$ .

is an infeasible sign subvector (taking the nonzero components) of  $[X_{:,B_x} - Y]$ . By the assumption of the right-hand side and its detailed consequence, it must also be an infeasible subvector of  $[X_{:,B_x} Y]$ . This means there exists some nonzero  $(\alpha_x, \alpha_y) \geq 0$  with support in  $B_x \times \text{supp}(d^y)$  such that

$$X \text{Diag}(s^x) \alpha_x + Y \text{Diag}(s^y) \alpha_y = 0. \quad (\text{D.2e})$$

Observe that one cannot have  $Y \text{Diag}(s^y) \alpha_y = 0$  since it would imply  $X \text{Diag}(s^x) \alpha_x = 0$ , meaning that  $s^x$  is infeasible which is not possible since it has support in  $B_x$ , i.e., corresponds to vectors  $X_{:,i}, i \in B_x$  are independent. Multiply (D.2e) by

$$\alpha_0 := \max \left\{ \frac{|d_j^y|}{(\alpha_y)_j}, (\alpha_y)_j \neq 0 \right\} > 0, \quad (\text{D.2f})$$

which is well-defined since one cannot have  $\alpha_y = 0$  and is nonzero by the support assumption. Then, sum  $\alpha_0$ (D.2e) and (D.2c) to get

$$\begin{aligned} & X \text{Diag}(s^x)(|d^x| + \alpha_0 \alpha_x) + Y \text{Diag}(s^y)(\alpha_0 \alpha_y - |d^y|) = 0, \\ \iff & X \text{Diag}(s^x)(|d^x| + \alpha_0 \alpha_x) - Y[-\text{Diag}(s^y) \text{Diag}(\text{sgn}(\alpha_0 \alpha_y - |d^y|))]|\alpha_0 \alpha_y - |d^y|| = 0 \end{aligned} \quad (\text{D.2g})$$

where the support of  $\alpha_0 \alpha_y - |d^y|$  is strictly included in the one of  $d^y$  by definition of  $\alpha_0$  and clearly  $|d^x| + \alpha_0 \alpha_x \geq 0$ . Then, one has that the pair  $(s^x, -\text{Diag}(s^y)\text{sgn}(\alpha_0 \alpha_y - |d^y|))$  is infeasible in  $[X_{:,B_x} - Y_{:,J_y^1}]$  with  $J_y^1$  the support of  $\alpha_0 \alpha_y - |d^y|$ . Then, it must also be infeasible in  $[X_{:,B_x} Y_{:,J_y^1}]$ . Using the same argument, there exists some nonzero  $(\alpha_x^1, \alpha_y^1) \geq 0$  with support in  $B_x \times J_y^1$ , still such that  $\alpha_y^1$  cannot be zero, and one can reuse the same argument.

To conclude, one can reduce from at least 1 the size of the support of  $J_y^l$  for each iteration  $l$ , meaning at some point one gets  $X \text{Diag}(s^x)[\dots] = 0$ , which is a contradiction.  $\square$

## D.2 Main proof

Now, we turn to the proof of proposition 6.1.20. Along the way, some related properties will be presented and analyzed, partly through the zonotope framework, which has the advantage of presenting a rather visual aspect.

The main idea is the following: first, choose a particular extremal  $\gamma_{\mathcal{E}^{0+}(x)}$ , then, by the projection of proposition 6.1.10, find a suitable  $\gamma_{\mathcal{E}^-(x)}$ . Then, that  $\gamma_{\mathcal{E}^-(x)}$  corresponds (by lemma B.1.4) to the relative interior of a certain face of the associated zonotope. This  $\gamma_{\mathcal{E}^-(x)}$  is naturally a convex combination of the vertices of the face, say the  $\gamma_{\mathcal{E}^-(x)}^{\text{vertex}}$ . Then, thanks to the properties of the “normals” to the face discussed in appendix C, one can justify that the couples  $(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}^{\text{vertex}})$  corresponds to a sign vector in  $\mathcal{S}$ , i.e., a Jacobian matrix in

---

$\partial_B H(x)$ . Then, by properties of convex combinations, the  $g$  belongs to the (C-)differential of  $\theta$ . Actually, we shall consider only the sign vectors of the linearization of  $F$  and  $G$ , i.e., the part of  $\partial_B H(x)$  governed by the matrix  $V$  (see chapters 3 and 4). This means the process is ignorant of some Jacobian matrices in  $\partial_B H(x)$ , thus in some points in  $\partial\theta(x)$ . However, the result is still true.

This roadmap is unfortunately hindered, besides its relative technicality, by what is called “degeneracies” in what follows. Easily visible thanks to the zonotope aspect, they actually show that for the values  $(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^{-}(x)}^{\text{vertex}})$ , the corresponding  $(\eta, \zeta) \in \{-1, +1\}^{\mathcal{E}(x)}$  do not necessarily all correspond to sign vectors of  $\mathcal{S}$ . They are related to the absence of strict inequalities discussed in appendix C around counterexample B.2.9. Let us summarize these observations.

- There is a significant difference between the elements of  $\partial_B H(x)$  and  $\partial\theta(x)$ , despite the relations between the two sets.
- Degeneracies may occur, i.e., the general proof is more technical.
- The method solving zonotope inclusion as discussed in appendix C is not the correct tool: it can return a value  $\gamma_{\mathcal{E}^{0+}(x)}(\eta)$  such that  $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^{-}(x)}(\gamma_{\mathcal{E}^{0+}(x)})) \notin \partial\theta(x)$  for any  $\gamma_{\mathcal{E}^{-}(x)} \in \mathbb{G}(\gamma_{\mathcal{E}^{0+}(x)})$ .
- For this method, one can correct the degeneracies to get an element in  $\partial\theta(x)$ , but it is unknown if it is still a descent direction since the projection is not respected.
- The furthest point of  $Z_x$  from  $Z_y$  (in Euclidean distance), which may not be the same as the one in the previous method, after applying the projection, corresponds to an element of  $\partial\theta(x)$ . The same is true for strict local maxima of the distance.

Let us recall some useful notation (equation (6.12) and rule 6.1.8)

$$\begin{aligned} g_0(x) &:= F'_{\mathcal{F}(x)}(x)^T F_{\mathcal{F}(x)}(x) + G'_{\mathcal{G}(x)}(x)^T G_{\mathcal{G}(x)}(x) + G'_{\mathcal{E}(x)}(x)^T G_{\mathcal{E}(x)}(x), \\ \mathcal{M}_+ &:= (F'_{\mathcal{E}^{0+}(x)}(x) - G'_{\mathcal{E}^{0+}(x)}(x))^T \text{Diag}(H_{\mathcal{E}^{0+}(x)}(x)), \\ \mathcal{M}_- &:= (F'_{\mathcal{E}^{-}(x)}(x) - G'_{\mathcal{E}^{-}(x)}(x))^T \text{Diag}(H_{\mathcal{E}^{-}(x)}(x)). \end{aligned} \quad (\text{D.3})$$

In the following rule,  $\lambda^*$  and  $\bar{\cdot}$  refer to the optimal values returned by the algorithm C.0.10 of appendix C (when  $\lambda^*$  is finite).

**Rule D.2.1** (variables correspondence). In what follows, we use the following quantities:

$$\begin{aligned} X &= \frac{1}{2}\mathcal{M}_+ & Y &= -\frac{1}{2}\mathcal{M}_- & \bar{x} - \bar{y} &= g_1 := g_0(x) + \frac{\mathcal{M}_+}{2}e + \frac{\mathcal{M}_-}{2}e \\ \eta &= 2\gamma_{\mathcal{E}^{0+}(x)} - e & \zeta &= 2\gamma_{\mathcal{E}^{-}(x)}\gamma_{\mathcal{E}^{-}(x)} - e & \bar{\zeta} &= \frac{-\beta + \Delta\bar{\eta}}{\lambda^*} \\ g &= \frac{\mathcal{M}_+}{2}\eta + \frac{\mathcal{M}_-}{2}\zeta + \bar{x} - \bar{y} = \bar{x} - \bar{y} + X\eta - Y\zeta \\ \bar{g} &= \frac{\mathcal{M}_+}{2}\eta + \frac{\mathcal{M}_-}{2}\bar{\zeta} + \bar{x} - \bar{y} = \bar{x} - \bar{y} + X\eta - Y\bar{\zeta} \\ \bar{g} &= Y(-\beta + \Delta\bar{\eta} - \bar{\zeta}) = Y(-\beta + \Delta\bar{\eta})\frac{\lambda^* - 1}{\lambda^*} = Y\bar{\zeta}(\lambda^* - 1) \end{aligned} \quad (\text{D.4})$$

where  $\eta \in [-1, +1]^{\mathcal{E}^0(x)}$  and  $\zeta \in [-1, +1]^{\mathcal{E}^-(x)}$  correspond to the parametrizations in  $[-1, +1]$ .  $\square$

We start with an observation which motivated the result of section D.1. The subset  $\mathcal{E}^0(x) := \{i \in \mathcal{E}^0(x) : H_i(x) = 0\}$  plays a slightly annoying role which justifies its introduction. Let  $x \in \mathbb{R}^n$ ,  $s \in \{\pm 1\}^{\mathcal{E}(x)}$  and  $J(s) \in \mathbb{R}^{n \times n}$  be the matrix defined by

$$J(s)_{\mathcal{F}(x) \cup \mathcal{G}(x),:} = \begin{bmatrix} F'_{\mathcal{F}(x)}(x) \\ G'_{\mathcal{G}(x)}(x) \end{bmatrix}, \quad J(s)_{\mathcal{E}(x),:} = \frac{s+e}{2} \cdot F'_{\mathcal{E}(x)}(x) + \frac{e-s}{2} \cdot G'_{\mathcal{E}(x)}(x). \quad (\text{D.5})$$

It is such that  $s_i = +1$  means  $J_{i,:} = F'_i$  and  $J_{i,:} = G'_i$  when  $s_i = -1$ . By definition of the B-differential of the componentwise minimum,  $J(s) \in \partial_B H(x) \iff \exists d, s \cdot V^\top d > 0$  where  $V^\top := G'_{\mathcal{E}(x)}(x) - F'_{\mathcal{E}(x)}(x)$ . Then, one can use the following equivalences (assuming here that  $\mathcal{E}^0(x) = \{i \in [1:n] : F_i(x) = 0 = G_i(x)\} = \emptyset$ , these indices have no impact on the final expression, so this assumption is not too important)<sup>2</sup>

$$\begin{aligned} J(s) \in \partial_B H(x) &\iff \exists d, s \cdot V^\top d > 0 \\ &\iff \exists d, s \cdot \text{Diag}(H_{\mathcal{E}(x)})^{-1} \text{Diag}(H_{\mathcal{E}(x)}) V^\top d > 0 \\ &\iff \exists d, s \cdot \text{Diag}(H_{\mathcal{E}(x)})^{-1} [\mathcal{M}_+ \mathcal{M}_-]^\top d > 0 \\ &\iff \exists d, s \cdot \text{Diag}(H_{\mathcal{E}(x)})^{-1} [2X - 2Y]^\top d > 0 \\ &\iff \exists d, s \cdot [X \ Y]^\top d > 0, \end{aligned} \quad (\text{D.6})$$

using the definitions recalled in (D.3) and rule D.2.1, then that

$$\begin{aligned} s_i H_i(x)^{-1} X_{:,i}^\top d > 0 &\iff s_i X_{:,i}^\top d > 0 \\ s_i H_i(x)^{-1} (-Y_{:,i})^\top d > 0 &\iff s_i Y_{:,i}^\top d > 0 \end{aligned}$$

because  $H_i(x) > 0$  for the indices of  $\mathcal{E}^0(x)$  and  $H_i(x) < 0$  for the indices of  $\mathcal{E}^-(x)$ . Furthermore, observe that, for  $s \in \{\pm 1\}^{\mathcal{E}(x)}$  (not necessarily  $s \in \mathcal{S}(V, 0) \iff J(s) \in \partial_B H(x)$ ), one can express  $J(s)^\top H(x)$  as

$$\begin{aligned} J(s)^\top H(x) &= F'_{\mathcal{F}(x)}(x)^\top F_{\mathcal{F}(x)}(x) + G'_{\mathcal{G}(x)}(x)^\top G_{\mathcal{G}(x)}(x) \\ &\quad + \frac{1}{2} [F'_{\mathcal{E}(x)}(x)^\top F_{\mathcal{E}(x)}(x) + G'_{\mathcal{E}(x)}(x)^\top G_{\mathcal{E}(x)}(x)] \\ &\quad + \frac{1}{2} [F'_{\mathcal{E}(x)}(x)^\top - G'_{\mathcal{E}(x)}(x)^\top] \text{Diag}(s) H_{\mathcal{E}(x)} \\ &= F'_{\mathcal{F}(x)}(x)^\top F_{\mathcal{F}(x)}(x) + G'_{\mathcal{G}(x)}(x)^\top G_{\mathcal{G}(x)}(x) \\ &\quad + \frac{1}{2} [F'_{\mathcal{E}(x)}(x)^\top F_{\mathcal{E}(x)}(x) + G'_{\mathcal{E}(x)}(x)^\top G_{\mathcal{E}(x)}(x)] + \frac{1}{2} [\mathcal{M}_+ \mathcal{M}_-] s \\ &= g_1 + [X - Y] s = g_1 + V \text{Diag}(H_{\mathcal{E}(x)}(x)) s \end{aligned} \quad (\text{D.7})$$

---

<sup>2</sup>For simplicity, we still write  $\partial_B H$  despite considering only the sign vectors found by the linearization, thus indicated by the matrix  $V$ .

---

(where indices of  $\mathcal{E}^0(x)$  would correspond to a column of  $X$  being zero, so the value of  $s_i$  is irrelevant)<sup>3</sup>. Recall that

$$\partial\theta(x) = \partial H(x)^\top H(x) = \text{conv}(\partial_B H(x))^\top H(x), \quad \partial_B H(x) = \text{ext}(\partial_C H(x))$$

where the first equality is Clarke's chain rule proposition, 2.3.19, the second is a property of Clarke's differential and the last is proposition 3.4.14. Therefore, we have that the elements of  $\partial_B H$  are defined by  $V$  (equivalently,  $[X \ Y]$  if  $\mathcal{E}^0(x) = \emptyset$ ), whereas the elements of  $\partial\theta(x)$  are governed by  $V \text{Diag}(H_{\mathcal{E}(x)}(x)) = [X \ -Y]$ .

Thus, the extremal elements of  $\partial\theta(x)$  (in  $\mathbb{R}^n$ ) correspond to the sign vectors that belong to  $\mathcal{S}(V, 0)$  (by restricting the  $s$  in (D.7) to  $\mathcal{S}(V, 0)$ ) **and** that correspond to vertices of the zonotope defined by  $[X \ -Y]$  (by the last line of (D.7)). In particular, this illustrates that  $\partial\theta(x)$  is a polytope (the convex hull of a finite number of points).<sup>4</sup> However, this does not imply that the vertices of  $\partial\theta(x)$  correspond to sign vectors in  $\mathcal{S}([X \ Y]) \cap \mathcal{S}([X \ -Y])$ : indeed, this set may even be empty.

**Counter-example D.2.2** (differences between  $[X \ Y]$  and  $[X \ -Y]$ ). Let

$$X = I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

By direct verification, one has

$$\begin{aligned} \mathcal{S}([X \ Y]) &= \left\{ (+, +, +, +), (+, -, +, +), (+, -, -, +), (+, +, +, -) \right\}, \\ &\quad \left\{ (-, -, -, -), (-, +, -, -), (-, +, +, -), (-, -, -, +) \right\}, \\ \mathcal{S}([X \ -Y]) &= \left\{ (-, +, +, +), (-, -, +, +), (-, -, +, -), (+, -, +, -) \right\}, \\ &\quad \left\{ (+, +, -, -), (-, +, -, -), (+, +, -, +), (-, +, -, +) \right\}, \\ \mathcal{S}([X \ Y]) \cap \mathcal{S}([X \ -Y]) &= \emptyset, \quad \mathcal{S}([X \ Y]) \cup \mathcal{S}([X \ -Y]) = \{\pm 1\}^4. \end{aligned}$$

In particular, we have that both zonotopes  $Z([X \ Y])$  and  $Z([X \ -Y])$  coincide, but  $[X \ -Y]$  evaluated with sign vectors of  $[X \ Y]$  (equivalently,  $[X \ Y]$  evaluated with sign vectors of  $[X \ -Y]$ ) do not have any vertex in common, see figure D.1.  $\square$

The meaning of this observation is that  $\partial\theta(x)$  has, in general, clearly less extremal points than  $|\mathcal{S}(V, 0)| = |\partial_B H(x)|$ . Naturally, if  $\mathcal{E}^{0+}(x) = \emptyset$  (eventually just  $\mathcal{E}^{0+}(x) \setminus \mathcal{E}^0(x)$ ),  $\mathcal{E}^-(x) = \emptyset$  or  $\mathcal{R}(X) \cap \mathcal{R}(Y) = \{0\}$ , the extremal points of  $\partial\theta(x)$  are in bijection with the elements of  $\mathcal{S}(V_{\mathcal{E}(x) \setminus \mathcal{E}^0(x)}, 0)$ <sup>5</sup>.

---

<sup>3</sup>In fact, for any  $s \in \mathcal{S}(V, 0)$ , there exists a  $s' \in \mathcal{S}(V_{:, \mathcal{E}(x) \setminus \mathcal{E}^0(x)}, 0)$  with  $s_{\mathcal{E}(x) \setminus \mathcal{E}^0(x)} = s'$  – by using the properties of the  $\mathcal{S}$ -tree with the indices of  $\mathcal{E}^0(x)$  last and removing the components of  $\mathcal{E}^0(x)$ .

<sup>4</sup>We believe the C-differential of  $\theta$  with the linearizations of  $F$  and  $G$  is also a zonotope but it is only a conjecture; one could be tempted to use  $\text{conv}(\{g_1 + [X \ -Y]s : s \in \mathcal{S}([X \ Y])\}) = g_1 + [X \ -Y]\text{conv}(\{s : s \in \mathcal{S}([X \ Y])\})$ , but the last term is a convex hull of (symmetric) points of the hypercube, which is unlikely to be easy to manipulate [262].

<sup>5</sup>Furthermore,  $s \in \mathcal{S}(V_{\mathcal{E}(x) \setminus \mathcal{E}^0(x)}, 0)$  implies that  $(s, s') \in \mathcal{S}(V, 0) \subseteq \{\pm 1\}^{\mathcal{E}(x)}$  by the induction properties of the  $\mathcal{S}$ -tree (chapter 3).

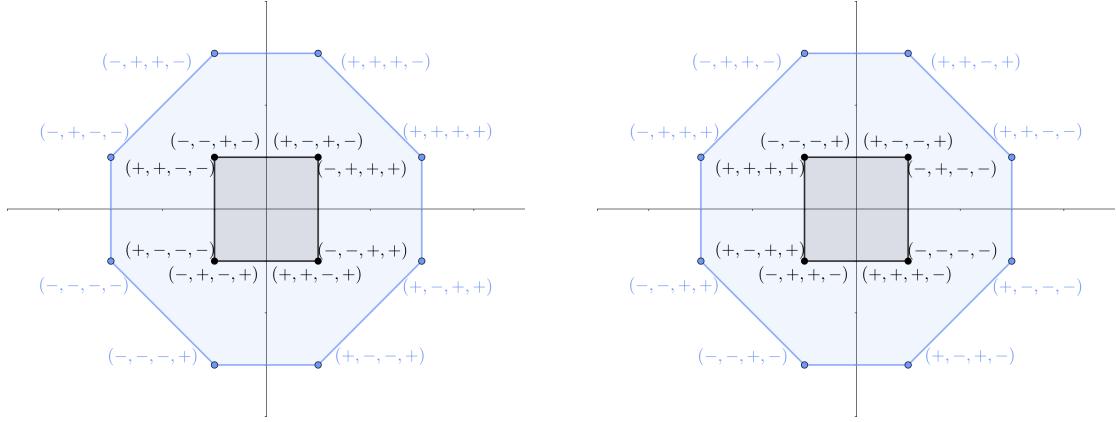


Figure D.1: Left:  $[X \ Y][-1, +1]^4$ , vertices in blue and other points in black (each dot is two sign vectors). Right:  $[X \ -Y][-1, +1]^4$ , vertices in blue and other points in black (each dot is two sign vectors). Schematically, the light blue corresponds to the zonotopes with  $[X \ Y]$  and  $[X \ -Y]$  and the black to  $\partial\theta(x)$ .

Said differently, some directions are positively multiplied ( $\mathcal{E}^{0+}(x)$ ) while some are negatively multiplied ( $\mathcal{E}^-(x)$ ), which disrupts the extremality. This is illustrated in counterexample D.2.3 and related to proposition D.1.1.

**Counter-example D.2.3** (multiplying by  $H$  disrupts extremality). Consider the following data, illustrated in figure D.2

$$F(x) = x = I_5x + 0, \quad G(x) = \begin{bmatrix} -1 & -16/5 & 0 & 0 & 0 \\ -2 & 21/5 & 0 & 0 & 0 \\ 0 & -5 & 1 & 0 & 0 \\ -2 & -3 & 0 & 1 & 0 \\ 3 & 1 & -1 & -1 & 0 \end{bmatrix} + \begin{bmatrix} 26/5 \\ -6/5 \\ 5 \\ 5 \\ -7 \end{bmatrix}.$$

Let  $x = [1; 1; -1; -1; 0]$ , clearly one has

$$F(x) = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 0 \end{bmatrix}, \quad G(x) = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ -1 \end{bmatrix} \implies \begin{cases} \mathcal{E}^{0+}(x) = \{1, 2\}, \\ \mathcal{E}^-(x) = \{3, 4\}, \\ \mathcal{F}(x) = \emptyset, \\ \mathcal{G}(x) = \{5\}. \end{cases}$$

---

According to (D.3) and rule 6.1.8, dropping some  $(x)$  for simplicity, one gets

$$\mathcal{M}_+ = (F'_{\mathcal{E}^{0+}(x)} - G'_{\mathcal{E}^{0+}(x)})^\top \text{Diag}(H_{\mathcal{E}^{0+}(x)}) = \begin{bmatrix} 2 & 2 \\ 16/5 & -16/5 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & -2 \\ -5 & -3 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 \\ \frac{16}{10} & -\frac{16}{10} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ \frac{5}{2} & \frac{3}{2} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\mathcal{M}_- = (F'_{\mathcal{E}^-(x)} - G'_{\mathcal{E}^-(x)})^\top \text{Diag}(H_{\mathcal{E}^-(x)}) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad Y = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

and one has

$$g_0(x) = 0 + G'_5(x)^\top G_5(x) + G'_{\{1,2,3,4\}}(x)^\top G_{\{1,2,3,4\}}(x) \\ = [-4 \ 8 \ 0 \ 0 \ 0]^\top,$$

as well as

$$\bar{x} - \bar{y} = g_0(x) + Xe - Ye = [-3 \ 4 \ 0 \ 0 \ 0]^\top.$$

The “zonotope approach”, using the variables  $X, Y$  and  $\bar{x} - \bar{y}$  is illustrated in figure D.3. Let us compute the differentials related to this situation. First, let us compute  $\partial_B H(x)$ . The corresponding matrix

$$V := (G'_{\mathcal{E}(x)}(x) - F'_{\mathcal{E}(x)}(x))^\top = \begin{bmatrix} -2 & -2 & 0 & -2 \\ -16/5 & 16/5 & -5 & -3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

is such that the sign vectors corresponding to Jacobian matrices are

$$\mathcal{S} = \left\{ \begin{array}{l} (+, +, +, +), \quad (+, +, -, +), \quad (-, +, -, +), \quad (-, +, -, -) \\ (-, -, -, -), \quad (-, -, +, -), \quad (+, -, +, -), \quad (+, -, +, +) \end{array} \right\}.$$

Indeed, the corresponding systems (without the zeros) read, up to a sign by symmetry,

$$\begin{bmatrix} 1 & 16/10 \\ 1 & -16/10 \\ 0 & 5/2 \\ 1 & 3/2 \end{bmatrix} d > 0, \quad \begin{bmatrix} 1 & 16/10 \\ 1 & -16/10 \\ 0 & -5/2 \\ 1 & 3/2 \end{bmatrix} d > 0, \quad \begin{bmatrix} -1 & -16/10 \\ 1 & -16/10 \\ 0 & -5/2 \\ 1 & 3/2 \end{bmatrix} d > 0, \quad \begin{bmatrix} 1 & 16/10 \\ -1 & 16/10 \\ 0 & 5/2 \\ 1 & 3/2 \end{bmatrix} d > 0,$$

where one can take  $d = \pm[2; 1]$ ,  $d = \pm[2; -1]$ ,  $d = \pm[1.55; -1]$ ,  $d = \pm[1; 1]$ . Then, the

Jacobian matrices  $J(s)$  given by (D.5) for  $s \in \mathcal{S}$ ,

$$\begin{aligned}
 J \begin{pmatrix} + \\ + \\ + \\ + \end{pmatrix} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 3 & 1 & -1 & -1 & 0 \end{bmatrix}, & J \begin{pmatrix} - \\ - \\ - \end{pmatrix} &= \begin{bmatrix} -1 & -16/5 & 0 & 0 & 0 \\ -2 & 21/5 & 0 & 0 & 0 \\ 0 & -5 & 1 & 0 & 0 \\ -2 & -3 & 0 & 1 & 0 \\ 3 & 1 & -1 & -1 & 0 \end{bmatrix}, \\
 J \begin{pmatrix} + \\ + \\ - \\ + \end{pmatrix} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 3 & 1 & -1 & -1 & 0 \end{bmatrix}, & J \begin{pmatrix} - \\ - \\ + \\ - \end{pmatrix} &= \begin{bmatrix} -1 & -16/5 & 0 & 0 & 0 \\ -2 & 21/5 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ -2 & -3 & 0 & 1 & 0 \\ 3 & 1 & -1 & -1 & 0 \end{bmatrix}, \\
 J \begin{pmatrix} - \\ + \\ - \\ + \end{pmatrix} &= \begin{bmatrix} -1 & -16/5 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 3 & 1 & -1 & -1 & 0 \end{bmatrix}, & J \begin{pmatrix} + \\ - \\ + \\ - \end{pmatrix} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -2 & 21/5 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ -2 & -3 & 0 & 1 & 0 \\ 3 & 1 & -1 & -1 & 0 \end{bmatrix}, \\
 J \begin{pmatrix} - \\ + \\ - \\ - \end{pmatrix} &= \begin{bmatrix} -1 & -16/5 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -5 & 1 & 0 & 0 \\ -2 & -3 & 0 & 1 & 0 \\ 3 & 1 & -1 & -1 & 0 \end{bmatrix}, & J \begin{pmatrix} + \\ - \\ + \\ + \end{pmatrix} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -2 & 21/5 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 3 & 1 & -1 & -1 & 0 \end{bmatrix}.
 \end{aligned}$$

Now, recall that  $\partial_B H(x) = \{J(s) : s \in \mathcal{S}\}$  and  $\partial\theta(x) = \partial H(x)^T H(x)$ , with:

$$\partial H(x)^T H(x) = \text{conv}(\partial_B H(x))^T H(x) = \text{conv}(\partial_B H(x)^T H(x))$$

since  $\cdot^T H(x)$  is an affine transformation. Now, the vectors  $(J^T H)$  involved are

$$\begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -2 \\ 5 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -4 \\ 9/5 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -2 \\ 24/5 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -4 \\ 8 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -4 \\ 3 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -2 \\ 31/5 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -4 \\ 16/5 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \quad (\text{D.8})$$

Now, let us detail the optimality process returning a  $\gamma_{\mathcal{E}^{0+}(x)}$  and a  $\gamma_{\mathcal{E}^{-}(x)}$  such that the value of  $g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^{-}(x)})$  is nonzero and verifies proposition 6.1.10. According to proposition C.0.7, the theorem is an equivalence and one has  $\lambda^* = 5$ , the corresponding point is  $\eta = [-1; -1]$ . Then, the projection on  $Z_y$  is given by  $\bar{y} + Y\zeta^*$  with  $\zeta^* = [1; -11/13]$ .

Finally, observe that the optimal  $g$ , given by proposition 6.1.10, is equal to

$$\begin{aligned}
 g \left( \eta = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \zeta^* = \begin{bmatrix} 1 \\ -\frac{11}{13} \end{bmatrix} \right) &= \bar{x} - \bar{y} + X\eta - Y\zeta^* \\
 &= \begin{bmatrix} -3 \\ 4 \\ 0_3 \end{bmatrix} + \begin{bmatrix} -2 \\ 0 \\ 0_3 \end{bmatrix} - \begin{bmatrix} -\frac{11}{13} \\ \frac{16}{13} \\ 0_3 \end{bmatrix} = \begin{bmatrix} -4 - \frac{2}{13} \\ 2 + \frac{10}{13} \\ 0_3 \end{bmatrix}
 \end{aligned}$$

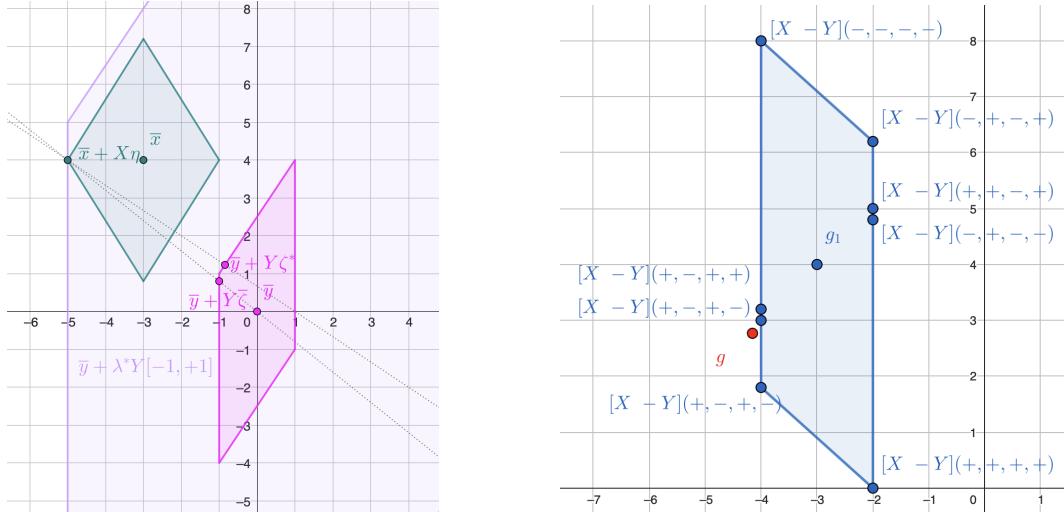


Figure D.2: Illustration of the zonotope aspect for a situation encountering multiple difficulties. On the left, the teal zonotope on top left is  $Z_x$  while  $Z_y$  is represented in the bottom right in magenta. The light purple represents the dilated version (by  $\lambda^*$ ) of  $Z_y$ . On the right, the blue zonotope is the representation of  $\partial\theta(x)$ , see (D.8), up to the three other components equal to zero (thus not shown). First, observe on the right picture that among the eight sign vectors corresponding to  $\partial_B H(x)$ , only four of them form the convex hull of the C-differential once multiplied by  $H$ . Moreover, the “neighbors” in the picture do not correspond to neighboring sign vectors. Finally, as described in a simpler example later, the method from appendix C returns a value of  $\mathcal{E}^{0+}(x)$  corresponding to the leftmost point in the teal area (the dilation of the bottom point on the boundary of  $Z_y$ , with  $\bar{\zeta}$ ) that corresponds to  $g$  (the projection is the top point with  $\zeta^*$ ) that is the red point in the right picture and is thus outside  $\partial\theta(x)$ ; this comes from the fact that the chosen signs are not in  $\mathcal{S}(V, 0)$ .

which is clearly not the convex combination of the vectors given in (D.8).  $\square$

This phenomenon is partly explained by proposition D.1.1.

### D.2.1 Detailed (simpler) counterexamples

Counterexample D.2.3 above shows the inadequacy of algorithm C.0.10 to find an element of  $\partial\theta(x)$  (though by construction, it helps finding a nonzero  $g$ ). The first counterexample below, in smaller dimension, shows that it is not an inherent flaw of algorithm C.0.10 itself. The following counterexamples discuss more technical points.

**Counter-example D.2.4** (wrong  $\gamma_{\mathcal{E}^{0+}(x)} \implies g \notin \partial\theta(x)$ ). Consider the following data

$$F(x) = x = I_3x + 0, \quad G(x) = \begin{bmatrix} 1 & -2 & 0 \\ 0 & -3 & 0 \\ 2 & -4 & 0 \end{bmatrix}, \quad x + \begin{bmatrix} -2 \\ -2 \\ -7 \end{bmatrix}.$$

Let  $x = [1; -1; 0]$ , clearly one has

$$F(x) = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \quad G(x) = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix} \implies \begin{cases} \mathcal{E}^{0+}(x) = \{1\}, & \mathcal{E}^-(x) = \{2\}, \\ \mathcal{F}(x) = \emptyset, & \mathcal{G}(x) = \{3\}. \end{cases}$$

According to (D.3) and rule 6.1.8, one gets

$$\begin{aligned} \mathcal{M}_+ &= (F'_{\mathcal{E}^{0+}(x)} - G'_{\mathcal{E}^{0+}(x)})^\top \text{Diag}(H_{\mathcal{E}^{0+}(x)}) = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}, & X &= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \\ \mathcal{M}_- &= (F'_{\mathcal{E}^-(x)} - G'_{\mathcal{E}^-(x)})^\top \text{Diag}(H_{\mathcal{E}^-(x)}) = \begin{bmatrix} 0 \\ -4 \\ 0 \end{bmatrix}, & Y &= \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}. \end{aligned}$$

Moreover, one has

$$\begin{aligned} g_0(x) &= 0 + G'_3(x)^\top G_3(x) + G'_{\{1,2\}}(x)^\top G_{\{1,2\}}(x) \\ &= [-1 \ 5 \ 0]^\top, \end{aligned}$$

as well as

$$\bar{x} - \bar{y} = g_0(x) + Xe - Ye = [-1 \ 4 \ 0]^\top.$$

The “zonotope approach”, using the variables  $X, Y$  and  $\bar{x} - \bar{y}$  is illustrated in figure D.3. Let us compute the differentials related to this situation. First, let us compute  $\partial_B H(x)$ . The corresponding matrix

$$V := (G'_{\mathcal{E}(x)}(x) - F'_{\mathcal{E}(x)}(x))^\top = \begin{bmatrix} 0 & 0 \\ -2 & -4 \\ 0 & 0 \end{bmatrix}$$

is such that the sign vectors corresponding to Jacobian matrices are

$$\mathcal{S} = \{(+, +), (-, -)\}.$$

Indeed, the corresponding systems read,

$$\begin{bmatrix} 0 & -2 & 0 \\ 0 & -4 & 0 \end{bmatrix} d > 0, \quad \begin{bmatrix} 0 & 2 & 0 \\ 0 & 4 & 0 \end{bmatrix} d > 0,$$

where one can take  $d = [0; -1; 0]$  and  $d = [0; 1; 0]$ . Then, the Jacobian matrices  $J(s)$  given by (D.5) for  $s \in \mathcal{S}$ ,

$$J \begin{pmatrix} - \\ - \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -4 & 0 \end{bmatrix}, \quad J \begin{pmatrix} + \\ + \end{pmatrix} = \begin{bmatrix} 1 & -2 & 0 \\ 0 & -3 & 0 \\ 2 & -4 & 0 \end{bmatrix}.$$

Now, recall that  $\partial_B H(x) = \{J(s) : s \in \mathcal{S}\}$  and  $\partial\theta(x) = \partial H(x)^\top H(x)$ , with:

$$\partial H(x)^\top H(x) = \text{conv}(\partial_B H(x))^\top H(x) = \text{conv}(\partial_B H(x)^\top H(x))$$

since  $\cdot^\top H(x)$  is an affine transformation. Now, the vectors involved in this convex hull are

$$\begin{bmatrix} -1 \\ 3 \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} -1 \\ 5 \\ 0 \end{bmatrix}. \quad (\text{D.9})$$

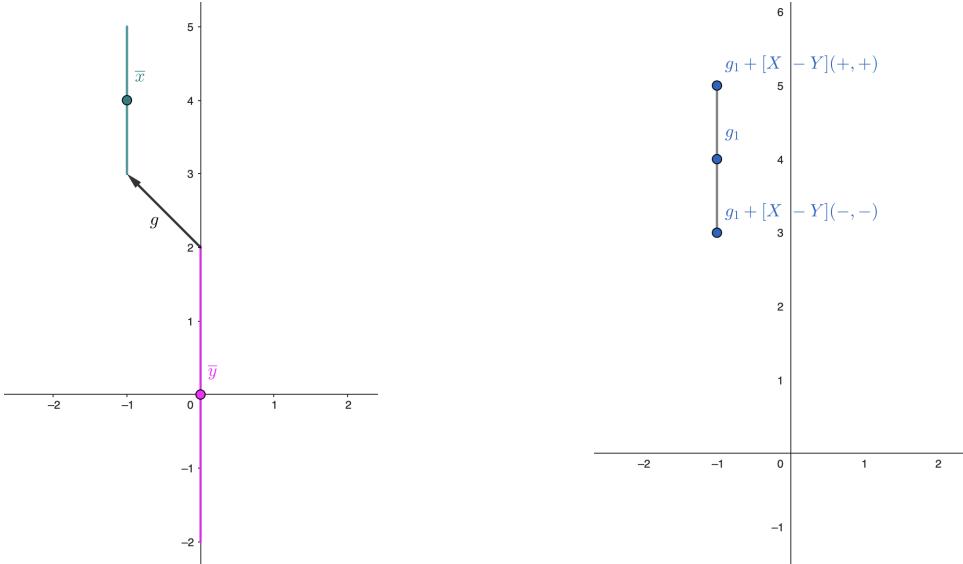


Figure D.3: Illustration of the counterexample. Left: corresponding zonotopes (teal for  $Z_x$ , magenta for  $Z_y$ ), the arrow represents  $g$  for  $\eta = -1$  which does not belong to  $\partial\theta(x)$ . Right: illustration of  $\partial\theta(x)$  and the elements in  $\partial_B H(x)^\top H(x)$ ;  $g_1$  is the center of the differential.

Finally, observe in figure D.3 that for any  $\gamma_{\mathcal{E}^0+(x)}$ , the projection is  $[0; 2; 0]$ , i.e.,  $\gamma_{\mathcal{E}^-(x)} = 1$ . Therefore, one has

$$g(\gamma_{\mathcal{E}^0+(x)}, \gamma_{\mathcal{E}^-(x)}) = [-1; 1 + 2\gamma_{\mathcal{E}^0+(x)}],$$

which belongs to  $\partial\theta(x)$  if and only if  $\gamma_{\mathcal{E}^0+(x)} = 1$ . This corresponds precisely to the point maximizing the distance, which is unique in this case.  $\square$

The following counterexamples detail difficulties encountered in proving that one can obtain an element of the differential. In particular, these difficulties are called “degeneracies”, since they are rather similar to degenerate linear optimization problems when the cost vector is orthogonal to a face of the feasible domain. The next example discusses a more elaborate situation, showing a limit of algorithm C.0.10: even without the projection (which ensures the resulting  $-g$  is a descent direction), the element may not be in the differential.

**Counter-example D.2.5** (degeneracies make  $\bar{g} \notin \partial\theta(x)$ ). Consider the following data

$$F(x) = x = I_5x + 0, \quad G(x) = \begin{bmatrix} 3 & 0 & -2 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix} + \begin{bmatrix} -4 \\ 2 \\ -2 \\ -2 \\ -15 \end{bmatrix}.$$

Let  $x = [1; -1; -1; -1; 0]$ , clearly one has

$$F(x) = \begin{bmatrix} 1 \\ -1 \\ -1 \\ -1 \\ 0 \end{bmatrix}, \quad G(x) = \begin{bmatrix} 1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix} \implies \begin{cases} \mathcal{E}^{0+}(x) = \{1\} \\ \mathcal{E}^-(x) = \{2, 3, 4\}, \\ \mathcal{G}(x) = \{5\}, \\ \mathcal{F}(x) = \emptyset. \end{cases}$$

According to (D.3) and rule 6.1.8, one gets

$$\begin{aligned} \mathcal{M}_+ &= (F'_{\mathcal{E}^{0+}(x)} - G'_{\mathcal{E}^{0+}(x)})^\top \text{Diag}(H_{\mathcal{E}^{0+}(x)}) = \begin{bmatrix} -2 \\ 0 \\ 2 \\ 0 \\ 0 \end{bmatrix}, & X &= \begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \\ \mathcal{M}_- &= (F'_{\mathcal{E}^-(x)} - G'_{\mathcal{E}^-(x)})^\top \text{Diag}(H_{\mathcal{E}^-(x)}) = \begin{bmatrix} -2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, & Y &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Moreover, one has

$$g_0(x) = 0 + G'_5(x)^\top G_5(x) + G'_{\{1,2,3,4\}}(x)^\top G_{\{1,2,3,4\}}(x), \\ = [1 \ 6 \ 3 \ 0 \ 0]^\top,$$

as well as

$$\bar{x} - \bar{y} = g_0(x) + Xe - Ye = [-1 \ 5 \ 3 \ 0 \ 0]^\top.$$

The “zonotope approach”, using the variables  $X$ ,  $Y$  and  $\bar{x} - \bar{y}$  is illustrated in figure D.4. Let us compute the differentials related to this situation. First, let us compute  $\partial_B H(x)$ . The corresponding matrix

$$V := (G'_{\mathcal{E}(x)}(x) - F'_{\mathcal{E}(x)}(x))^\top = \begin{bmatrix} 2 & -2 & 0 & 0 \\ 0 & 0 & -2 & 0 \\ -2 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

---

is such that the sign vectors corresponding to Jacobian matrices are

$$\left\{ \begin{array}{l} (+, +, +, +), (+, +, -, +), (+, -, +, +), (+, -, -, +), (+, -, +, -), (+, -, -, -) \\ (-, -, -, -), (-, -, +, -), (-, +, -, -), (-, +, +, -), (-, +, -, +), (-, +, +, +) \end{array} \right\}.$$

Indeed, the corresponding systems read, up to a sign by symmetry,

$$\begin{aligned} & \begin{bmatrix} -1 & 0 & 1 & 0_2^T \\ 2 & 0 & 0 & 0_2^T \\ 0 & 2 & 0 & 0_2^T \\ 0 & 0 & 2 & 0_2^T \end{bmatrix} d > 0, \quad \begin{bmatrix} -1 & 0 & 1 & 0_2^T \\ -2 & 0 & 0 & 0_2^T \\ 0 & 2 & 0 & 0_2^T \\ 0 & 0 & 2 & 0_2^T \end{bmatrix} d > 0, \quad \begin{bmatrix} -1 & 0 & 1 & 0_2^T \\ -2 & 0 & 0 & 0_2^T \\ 0 & 2 & 0 & 0_2^T \\ 0 & 0 & -2 & 0_2^T \end{bmatrix} d > 0, \\ & \begin{bmatrix} -1 & 0 & 1 & 0_2^T \\ 2 & 0 & 0 & 0_2^T \\ 0 & -2 & 0 & 0_2^T \\ 0 & 0 & 2 & 0_2^T \end{bmatrix} d > 0, \quad \begin{bmatrix} -1 & 0 & 1 & 0_2^T \\ -2 & 0 & 0 & 0_2^T \\ 0 & -2 & 0 & 0_2^T \\ 0 & 0 & 2 & 0_2^T \end{bmatrix} d > 0, \quad \begin{bmatrix} -1 & 0 & 1 & 0_2^T \\ -2 & 0 & 0 & 0_2^T \\ 0 & -2 & 0 & 0_2^T \\ 0 & 0 & -2 & 0_2^T \end{bmatrix} d > 0. \end{aligned}$$

where one can take (ignoring coordinates 4 and 5)  $d = \pm[1; 1; 2]$ ,  $d = \pm[-1; 1; 1]$ ,  $d = \pm[-2; 1; -1]$ ,  $d = \pm[1; -1; 2]$ ,  $d = \pm[-1; -1; 1]$ ,  $d = \pm[-2; -1; -1]$ . Then, the

Jacobian matrices  $J(s)$  given by (D.5) for  $s \in \mathcal{S}$ , are as follows.

$$\begin{aligned}
 J(+,+,+,-) &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \quad J(-,-,-,-) = \begin{bmatrix} 3 & 0 & -2 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \\
 J(+,+,-,+)&= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \quad J(-,-,+,-) = \begin{bmatrix} 3 & 0 & -2 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \\
 J(+,-,+,-)&= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \quad J(-,+,-,-) = \begin{bmatrix} 3 & 0 & -2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \\
 J(+,-,-,+)&= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \quad J(-,+,-,-) = \begin{bmatrix} 3 & 0 & -2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \\
 J(+,-,+,-)&= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \quad J(-,+,-,+)= \begin{bmatrix} 3 & 0 & -2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \\
 J(+,-,-,-)&= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}, \quad J(-,+,-,+)= \begin{bmatrix} 3 & 0 & -2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ 4 & -5 & -4 & -1 & 0 \end{bmatrix}.
 \end{aligned}$$

Now, recall that  $\partial_B H(x) = \{J(s) : s \in \mathcal{S}\}$  and  $\partial\theta(x) = \partial H(x)^T H(x)$ , with:

$$\partial H(x)^T H(x) = \text{conv}(\partial_B H(x))^T H(x) = \text{conv}(\partial_B H(x)^T H(x))$$

since  $\cdot^T H(x)$  is an affine transformation. Now, the vectors involved in this convex hull are

$$\begin{bmatrix} -3 \\ 4 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 6 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -3 \\ 6 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 4 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 4 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 6 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 4 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 5 \\ 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 4 \\ 5 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 6 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 6 \\ 5 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 4 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{D.10}$$

Now, let us detail algorithm C.0.10 and the corresponding problem: one has  $\lambda^* = 5$ , but there is degeneracy in the sense that the value of  $\eta$  is not relevant: any point in  $Z_x$  requires a dilation of  $\lambda^*$ .

Consider for instance  $\eta = -1$ , one gets  $z = [0; 5; 2]$  then  $\bar{\zeta} = [0; 1; 2/5]$ . The corresponding value of  $\bar{g}$  is  $\bar{g} = [0; 4; 8/5]$ , which does not belong to the convex combination of the above vectors. However, if another  $\eta$  is chosen (leading to the same value of  $\lambda^*$ ), such as  $\eta = +1$ , one gets  $\bar{\zeta} = [-2/5; 1; 4/5]$  and  $\bar{g} = [-8/5; 4; 16/5]$ , which does belong to the differential.

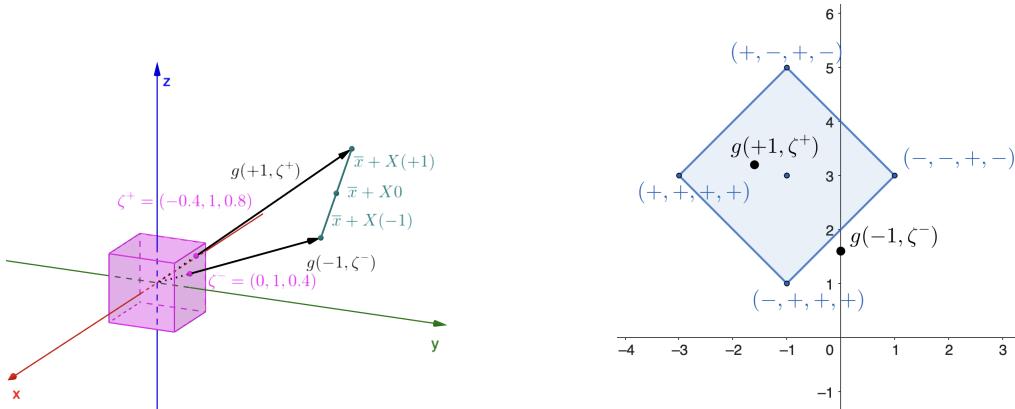


Figure D.4: Illustration of the counterexample. Left: corresponding zonotopes (teal for  $Z_x$ , magenta for  $Z_y$ ), the arrows represent  $g$  for  $\eta = -1$  which does not belong to  $\partial\theta(x)$  and  $\eta = +1$  which does. Right: illustration of  $\partial\theta(x)$  (in the plane  $x_2 = 4$ ). We observe, as seen in counterexample D.2.3, that not all sign vectors are extremal  $((+, -, +, +)$  and  $(-, +, +, -)$  both correspond to the middle blue dot) and that, depending on the value of  $\eta$ ,  $g$  may or may not belong to  $\partial\theta(x)$ . The remaining 6 sign vectors correspond to the face in the plane  $x_2 = 6$ .

Indeed, since both these values have a second component equal to 4, they must be a convex combination of the elements having a second component also equals to 4 (since the others vectors have a second component equal to 6). Now, it is clear to see that

$$\begin{bmatrix} 0 \\ 8/5 \end{bmatrix} \notin \text{conv} \left( \begin{bmatrix} -3 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ 5 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right) = \begin{bmatrix} -1 \\ 3 \end{bmatrix} + B(0, 2)_{\|\cdot\|_1},$$

$$\begin{bmatrix} -8/5 \\ 16/5 \end{bmatrix} \in \text{conv} \left( \begin{bmatrix} -3 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ 5 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right) = \begin{bmatrix} -1 \\ 3 \end{bmatrix} + B(0, 2)_{\|\cdot\|_1},$$

which confirms the observation.  $\square$

The following counterexample aims at describing a technical difficulty encountered in propositions D.2.7 and D.2.8. It is obtained as a lifting of counterexample D.2.4.

**Counter-example D.2.6** (degeneracies). Consider the following data:

$$F(x) = x = I_5x + 0, \quad G(x) = \begin{bmatrix} 1 & -2 & 0 & 0 & 0 \\ 0 & 1 & -2 & 0 & 0 \\ 0 & -4 & 1 & 0 & 0 \\ 0 & 0 & -4 & 1 & 0 \\ 2 & -2 & 0 & 1 & 0 \end{bmatrix} + \begin{bmatrix} 2 \\ -2 \\ 4 \\ -4 \\ -2 \end{bmatrix}.$$

Let  $x = [1; 1; -1; -1; 0]$ , clearly one has

$$F(x) = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 0 \end{bmatrix}, \quad G(x) = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ -1 \end{bmatrix} \implies \begin{cases} \mathcal{E}^{0+}(x) = \{1, 2\}, \\ \mathcal{E}^-(x) = \{3, 4\}, \\ \mathcal{F}(x) = \emptyset, \\ \mathcal{G}(x) = \{5\}. \end{cases}$$

According to (D.3) and rule 6.1.8, one gets

$$\begin{aligned} \mathcal{M}_+ &= (F'_{\mathcal{E}^{0+}(x)} - G'_{\mathcal{E}^{0+}(x)})^\top \text{Diag}(H_{\mathcal{E}^{0+}(x)}) = \begin{bmatrix} 0 & 0 \\ 2 & 0 \\ 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, & X &= \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \\ \mathcal{M}_- &= (F'_{\mathcal{E}^-(x)} - G'_{\mathcal{E}^-(x)})^\top \text{Diag}(H_{\mathcal{E}^-(x)}) = \begin{bmatrix} 0 & 0 \\ -4 & 0 \\ 0 & -4 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, & Y &= \begin{bmatrix} 0 & 0 \\ 2 & 0 \\ 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

Moreover, one has

$$\begin{aligned} g_0(x) &= 0 + G'_5(x)^\top G_5(x) + G'_{\{1,2,3,4\}}(x)^\top G_{\{1,2,3,4\}}(x) \\ &= [-1 \ 5 \ 1 \ 0 \ 0]^\top, \end{aligned}$$

as well as

$$\bar{x} - \bar{y} = g_0(x) + Xe - Ye = [-1 \ 4 \ 0 \ 0 \ 0]^\top.$$

The “zonotope approach”, using the variables  $X$ ,  $Y$  and  $\bar{x} - \bar{y}$  is illustrated in figure D.5. Let us compute the differentials related to this situation. First, consider  $\partial_B H(x)$ . The corresponding matrix

$$V := (G'_{\mathcal{E}(x)}(x) - F'_{\mathcal{E}(x)}(x))^\top = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -2 & 0 & -4 & 0 \\ 0 & -2 & 0 & -2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

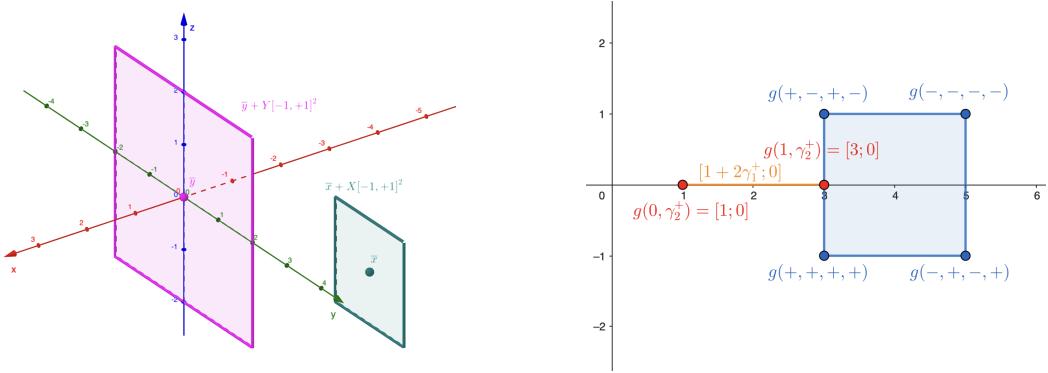


Figure D.5: Illustration of the degeneracies, mostly obtained in counterexamples by artificially adding a dimension (the third dimension is omitted in the right picture, it is in the plane  $x_1 = -1$ ). On the left, the zonotope aspect described by the previous equations. On the right, the illustration of the gradients  $g$  obtained in this situation: the blue square represents the differential of  $\theta$ ; in particular, the vertices correspond to sign vectors that are not “adjacent”. Here, it is due to the fact the vectors of  $X$  and  $Y$  are colinear, but the absence of adjacency can also occur without this particular case. The orange segment represents the possible  $g$  obtained by the choices of  $\mathcal{E}^{0+}(x)$ , the red points its vertices. In particular, the vertices of the differential in blue correspond to the zonotope expressed in (D.7).

is such that the sign vectors corresponding to Jacobian matrices are

$$\mathcal{S} = \{(+, +, +, +), (-, -, -, -), (+, -, +, -), (-, +, -, +)\}.$$

Indeed, the corresponding systems read, up to a sign by symmetry,

$$\pm \begin{bmatrix} 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 \\ 0 & -4 & 0 & 0 & 0 \\ 0 & 0 & -4 & 0 & 0 \end{bmatrix} d > 0, \quad \pm \begin{bmatrix} 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & -4 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 \end{bmatrix} d > 0,$$

where one can take  $d = \pm[0; -1; -1; 0; 0]$  and  $d = \pm[0; -1; 1; 0; 0]$ . Then, the Jacobian matrices  $J(s)$  given by (D.5) for  $s \in \mathcal{S}$ ,

$$J \begin{pmatrix} + \\ + \\ + \\ + \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 2 & -2 & 0 & -1 & 0 \end{bmatrix}, \quad J \begin{pmatrix} - \\ - \\ - \\ - \end{pmatrix} = \begin{bmatrix} 1 & -2 & 0 & 0 & 0 \\ 0 & 1 & -2 & 0 & 0 \\ 0 & -4 & 1 & 0 & 0 \\ 0 & 0 & -4 & 1 & 0 \\ 2 & -2 & 0 & -1 & 0 \end{bmatrix},$$

$$J \begin{pmatrix} + \\ - \\ + \\ - \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -4 & 1 & 0 \\ 2 & -2 & 0 & -1 & 0 \end{bmatrix}, \quad J \begin{pmatrix} - \\ + \\ - \\ + \end{pmatrix} = \begin{bmatrix} 1 & -2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 2 & -2 & 0 & -1 & 0 \end{bmatrix}.$$

Now, recall that  $\partial_B H(x) = \{J(s) : s \in \mathcal{S}\}$  and  $\partial\theta(x) = \partial H(x)^\top H(x)$ , with:

$$\partial H(x)^\top H(x) = \text{conv}(\partial_B H(x))^\top H(x) = \text{conv}(\partial_B H(x)^\top H(x))$$

since  $\cdot^\top H(x)$  is an affine transformation. Now, the vectors involved in this convex hull are

$$\begin{bmatrix} -1 \\ 3 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 5 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 3 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 5 \\ -1 \\ 0 \\ 0 \end{bmatrix}. \quad (\text{D.11})$$

Now, let us detail where the degeneracies intervene. Let  $\gamma_{\mathcal{E}^0+(x)} = (\gamma_1^+, \gamma_2^+)$ , after applying the projection solving proposition 6.1.10, one has

$$g(\gamma_{\mathcal{E}^0+(x)}, \gamma_{\mathcal{E}^-(x)}) = [-1; 2 + 2\gamma_1^+ - 1; 0; 0; 0]^\top = [-1; 1 + 2\gamma_1^+; 0; 0; 0]^\top.$$

This is illustrated on figure D.5. Observe that there is no unicity of the  $\gamma_{\mathcal{E}^0+(x)}$  maximizing the distance between both zonotopes: the choice of  $\gamma_2^+$  is irrelevant: whatever the value of  $\gamma_2^+$  is, the same  $g$  is obtained after the projection.

Consider an extremal point, i.e.,  $\eta_2 \in \{\pm 1\}$  with  $\eta_1 = 1$ . Then, the projection corresponds to  $\zeta = (1, \pm 1/2)$ . Therefore, the vertices of the face containing  $\zeta$  are  $\hat{\zeta}^+ := (+1, +1)$  and  $\hat{\zeta}^- := (+1, -1)$  for both  $\eta_2 = 1$  or  $\eta_2 = -1$ .

In the proof, one would ideally have that the corresponding sign vectors belong in  $\mathcal{S}$ :

$$\eta_2 = +1 \implies (+, +, +, +), (+, +, +, -); \quad \eta_2 = -1 \implies (+, -, +, +), (+, -, +, -)$$

and for either values of  $\gamma_2^+$ , one of the two signs is not in  $\mathcal{S}$ . Since this phenomenon cannot be avoided (if it occurs), the proof proceeds as follows. For the vertex  $\hat{\zeta}^+$ , a direction  $\hat{d}^+ \in \mathcal{R}(Y)$  is such that  $\hat{d}_1^+ = 0$ ,  $\hat{d}_2^+ > 0$  and  $\hat{d}_3^+ > 0$ . For the vertex  $\hat{\zeta}^-$ , a direction  $\hat{d}^- \in \mathcal{R}(Y)$  is such that  $\hat{d}_1^- = 0$ ,  $\hat{d}_2^- > 0$  and  $\hat{d}_3^- < 0$ . Moreover, one has  $\bar{c} := [0; 1; 0]$ . Therefore, for  $\hat{\zeta}^+$  and  $\hat{d}^+$ , one has  $\hat{c}_1 = 0$ ,  $\hat{c}_2 > 1 > 0$ , and  $\hat{c}_3 > 0$  (the two remaining dimensions are irrelevant), meaning  $\tilde{\eta}_2 = +1$ . Similarly, for the other vertex  $\hat{\zeta}^-$ ,  $\hat{d}_1^- = 0$ ,  $\hat{d}_2^- > 0$  and  $\hat{d}_3^- < 0$ , leading to  $\tilde{\eta}_2 = -1$ .

Observe that the choice of  $\eta_2$  does not change the value of  $g$ : only the theoretical proof wants to find one that is a convenient convex combination. Observe that the value of  $\eta_2 = 0$  is the half-sum of two elements in the B-differential:

$$[-1; 3; 0; 0; 0] = \frac{1}{2}([-1; 3; -1; 0; 0] + [-1; 3; 1; 0; 0]).$$

However, if the teal zonotope was tilted to the top / bottom, one would need a suitable choice instead of 0. For instance, if  $\bar{x} = [-1; 4; 1/2]$ , the value of  $g$  after the projection is

---

unchanged (for a fixed  $\eta_1$ ). However, one then has  $\zeta_2 = (1 + 2\eta_2)/4$ , and the “system” to solve to find a convex combination of vertices reads

$$\begin{pmatrix} +1 \\ \eta_2 \\ +1 \\ (1 + 2\eta_2)/4 \end{pmatrix} = t \begin{pmatrix} +1 \\ +1 \\ +1 \\ +1 \end{pmatrix} + (1 - t) \begin{pmatrix} +1 \\ -1 \\ +1 \\ -1 \end{pmatrix} \iff \begin{cases} \eta_2 = 2t - 1 \\ 2\eta_2 + 1 = 4(2t - 1) \end{cases}$$

which is solved by  $(\eta_2, t) = (1/2, 3/4)$ .  $\square$

## D.2.2 Degeneracies and (theoretical) corrections

First, we consider a simpler property dealing with the  $\bar{\cdot}$  variables from chapter C.

**Proposition D.2.7** (property of  $\bar{\zeta}$  and  $\bar{g}$ ). *With the same notations, there exists a modification  $\tilde{\eta}$  of  $\bar{\eta}$  such that  $\bar{g} = \bar{x} + X\tilde{\eta} - \bar{y} - Y\bar{\zeta} \in \partial\theta(x)$ .*

*Proof.* Recall that one has  $\bar{x} + X\eta = \bar{y} + \lambda^*Y\bar{\zeta}$  and let  $\bar{g} := \bar{x} + X\bar{\eta} - \bar{y} - Y\bar{\zeta}$ . It is clear that  $\bar{y} + Y\bar{\zeta}$  belongs to the boundary of  $Z_y$  therefore to the interior of a unique face  $\bar{F}$ , see proposition B.1.4. By propositions B.2.2 and B.2.6, let  $I^* := \{i \in \mathcal{E}^-(x) : \bar{\zeta}_i \in \{-1, +1\}\}$  and  $I^{\bar{F}} := \{i \in \mathcal{E}^-(x) : \bar{\zeta}_i \in (-1, +1)\}$ . Using proposition B.1.7, let  $\bar{c}$  be an associated “normal” to  $\bar{F}$ . By proposition B.2.3, one has

$$i \in I^* \implies \bar{\zeta}_i Y_{:,i}^\top \bar{c} > 0, \quad i \in I^{\bar{F}} \implies Y_{:,i}^\top \bar{c} = 0.$$

Clearly,  $\bar{y} + Y\bar{\zeta}$  is a convex combination of the vertices of  $\bar{F}$ , which have the expression  $\bar{y} + Y_{:,I^*} \bar{\zeta}_{I^*} + Y_{:,I^{\bar{F}}} \hat{\zeta}_{I^{\bar{F}}}$ , i.e.,

$$\bar{F} = \bar{y} + Y_{:,I^*} \bar{\zeta}_{I^*} + \text{conv}(\{Y_{:,I^{\bar{F}}} \hat{\zeta}_{I^{\bar{F}}}, \hat{\zeta}_{I^{\bar{F}}} \in \mathcal{S}(Y_{:,I^{\bar{F}}})\})$$

( $\bar{y} + Y\bar{\zeta}$  is a convex combination of the  $\bar{y} + Y\hat{\zeta}$ ). Therefore, if one could show that all the Jacobian matrices corresponding to the signs  $[\bar{\eta}; \hat{\zeta}]$  are in  $\partial_B H(x)$ , then one would have

$$\begin{bmatrix} \bar{\eta} \\ \bar{\zeta} \end{bmatrix} = \sum t_i \begin{bmatrix} \bar{\eta} \\ \hat{\zeta}^i \end{bmatrix}, \bar{J} := J([\bar{\eta}; \bar{\zeta}]) = \sum t_i J([\bar{\eta}; \hat{\zeta}^i]), \bar{J}^\top H(x) = \left[ \sum t_i J([\bar{\eta}; \hat{\zeta}^i]) \right]^\top H(x).$$

However, as described in counterexample D.2.5, this may not hold due to “degeneracies”. Let us detail a way to proceed. First, we observe the indices of  $\mathcal{E}^{0+}(x)$ . Let us first show that  $\bar{\eta}_i X_{:,i}^\top \bar{c} \geq 0$ . Suppose that there exists a  $i \in \mathcal{E}^{0+}(x)$  such that  $\bar{\eta}_i X_{:,i}^\top \bar{c} < 0$ . With similar arguments as in the previous appendices, the point  $\bar{x} + X\bar{\eta} - 2\bar{\eta}_i X_{:,i} \in Z_x$  would be “beyond” the hyperplane orthogonal to  $\bar{c}$  and further from  $Z_y$ , which contradicts the optimality of  $\lambda^*$ . Therefore,  $\bar{c}$  is such that

$$\forall i \in \mathcal{E}^{0+}(x), \bar{\eta}_i X_{:,i}^\top \bar{c} \geq 0, \quad \forall i \in I^*, \bar{\zeta}_i y_i^\top \bar{c} > 0, \quad \forall i \in I^{\bar{F}}, y_i^\top \bar{c} = 0.$$

Now, let us focus on the indices such that  $X_{:,i}^\top \bar{c} = 0$ . We show it implies  $X_{:,i} \in L_{\bar{F}} = \text{aff}(F) - \text{aff}(F)$ . Indeed, let  $\bar{H} := \bar{c}^\perp + z^{\bar{F}}$  for  $z^{\bar{F}} := \bar{y} + Y_{:,I^*} \bar{\zeta}_{I^*}$  the center of  $\bar{F}$ ,  $z^{\bar{F}} \in \bar{F}$ , one has  $\bar{F} = Z_y \cap \bar{H}$ . Let

$$\begin{aligned} Z_y^* &:= \bar{y} + \lambda^* Y[-1, +1]^{\mathcal{E}^-(x)} = \bar{y} + \lambda^*(Z_y - \bar{y}), \\ \bar{F}^* &:= \bar{y} + \lambda^*(\bar{F} - \bar{y}) \quad \text{and} \\ \bar{H}^* &:= \bar{c}^\perp + \bar{y} + \lambda^*(z^{\bar{F}} - \bar{y}) \end{aligned}$$

be the relevant elements linked to the dilation by  $\lambda^*$  of the zonotope  $Z_y$ . One has  $\bar{F}^* = Z_y^* \cap \bar{H}^*$ , see figure D.6.

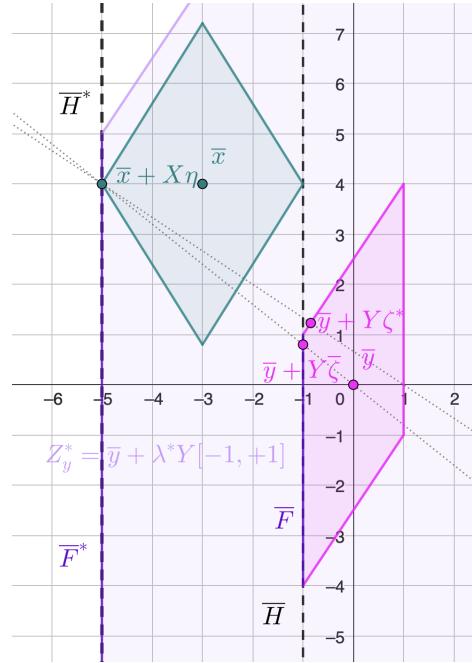


Figure D.6: Illustration of the dilated quantities, with the data from counterexample D.2.3.

Therefore, suppose that for some  $i \in \mathcal{E}^{0+}(x)$ ,  $X_{:,i}^\top \bar{c} = 0$  (i.e.,  $X_{:,i} \in \bar{c}^\perp = \bar{H} - \bar{H}$ ) and  $X_{:,i} \notin L_{\bar{F}}$ . Then, for  $\delta \in [0, 2]$ , the point  $z^* := \bar{x} + X\eta - \delta\eta_i X_{:,i}$  in  $Z_x$  by definition, belongs to  $\bar{H}^*$  since

$$\begin{aligned} z^* &= \bar{x} + X\eta - \delta\eta_i X_{:,i} = \bar{y} + \lambda^* Y \bar{\zeta} - \delta\eta_i X_{:,i} \\ &= \bar{y} + \lambda^* Y_{:,I^*} \bar{\zeta}_{I^*} + \lambda^* Y_{:,I^F} \bar{\zeta}_{I^F} - \delta\eta_i X_{:,i} \\ &= \bar{y} + \lambda^*(z^{\bar{F}} - \bar{y}) + \lambda^* Y_{:,I^F} \bar{\zeta}_{I^F} - \delta\eta_i X_{:,i} \\ &\in \bar{y} + \lambda^*(z^{\bar{F}} - \bar{y}) + c^\perp = \bar{H}^*, \end{aligned}$$

where the first equality comes from the primal-dual problem, the second splits the indices of  $\mathcal{E}^-(x)$ , the third rearranges the terms to make  $z^{\bar{F}}$  appear, and the last line come from the properties of  $\bar{c}$  and the definition of  $\bar{H}^*$ . However,

$$\begin{aligned} z^* &= \bar{x} + X\eta - \delta\eta_i X_{:,i} = \bar{y} + \lambda^* Y \bar{\zeta} - \delta\eta_i X_{:,i} \\ &\in \bar{F}^* - \delta\eta_i X_{:,i} \end{aligned}$$

---

thus  $z^* \notin \overline{F}^*$  since we supposed  $X_{:,i} \notin L_{\overline{F}}$ . Therefore, using  $\overline{F}^* = Z_y^* \cap \overline{H}^*$ ,

$$z^* \notin \overline{F}^*, z^* \in \overline{H}^* \implies z^* \notin Z_y^*.$$

This means there exists a point of  $Z_x$  that is not in  $Z_y^*$ , which is a contradiction with the optimality of  $\lambda^*$ . Now, define  $\mathcal{E}^{0+}(x)_0 := \{i \in \mathcal{E}^{0+}(x) : X_{:,i}^\top \bar{c} = 0\}$ ,  $i \in \mathcal{E}^{0+}(x)_0 \Rightarrow X_{:,i} \in L_{\overline{F}}$ . This means the indices of  $\mathcal{E}^{0+}(x)_0$  are not relevant in the sense the value of the corresponding  $\eta_i$  could be different without changing the optimality of  $\lambda^*$ .

Finally, for any  $\hat{\zeta}$  corresponding to the vertex  $\bar{y} + Y\hat{\zeta}$  of  $\overline{F}$ , let  $\hat{d} \in \mathcal{R}(Y)$  be a direction such that  $\hat{\zeta}_i y_i^\top \hat{d} > 0$ , i.e., a direction verifying the vertex. Let  $\hat{c} := \bar{c} + \varepsilon \hat{d}$  for some small positive  $\varepsilon$ . Since  $\hat{d}$  is taken in an open set, up to small modifications (then modifications of  $\varepsilon$ ) one can assume that  $X_{:,i}^\top \hat{c} \neq 0$  for  $i \in \mathcal{E}^{0+}(x)_0$ . Let  $\hat{\eta}_i = \text{sgn}(X_{:,i}^\top \hat{c})$  for  $i \in \mathcal{E}^{0+}(x)_0$ , one has, using the properties of  $\hat{d}$  and  $\bar{c}$

$$\begin{cases} i \in I^* \implies \hat{c}^\top \hat{\zeta}_i y_i = \bar{c}^\top \hat{\zeta}_i y_i + \varepsilon \hat{d}^\top \hat{\zeta}_i y_i \stackrel{>0}{>} 0, \\ i \in I^{\overline{F}} \implies \hat{c}^\top \hat{\zeta}_i y_i = \bar{c}^\top \hat{\zeta}_i y_i + \varepsilon \hat{d}^\top \hat{\zeta}_i y_i \stackrel{=0}{>} 0, \end{cases}$$

and

$$\begin{cases} i \in \mathcal{E}^{0+}(x)_0 \implies \hat{c}^\top \hat{\eta}_i X_{:,i} = \bar{c}^\top \hat{\eta}_i X_{:,i} + \varepsilon \hat{d}^\top \hat{\eta}_i X_{:,i} \stackrel{>0}{>} 0, \\ i \in \mathcal{E}^{0+}(x)_+ \implies \hat{c}^\top \hat{\eta}_i X_{:,i} = \bar{c}^\top \hat{\eta}_i X_{:,i} + \varepsilon \hat{d}^\top \hat{\eta}_i X_{:,i} \stackrel{>0}{>} 0, \end{cases}$$

meaning that we have justified that each of the Jacobian matrices corresponding to the sign vectors  $[\hat{\eta}; \hat{\zeta}]$  are in  $\partial_B H(x)$ . Now, write the convex combination

$$\bar{\zeta} = \sum t_j \hat{\zeta}^j, \quad t_j \geq 0, \quad \sum t_j = 1$$

and let  $\tilde{\eta} := \sum t_i \hat{\eta}_i$ . By construction, the sign vectors

$$(\hat{\eta}; \hat{\zeta}) = (\eta_{\mathcal{E}^{0+}(x)_+}; \hat{\eta}_{\mathcal{E}^{0+}(x)_0}; \bar{\zeta}_{I^*}; \hat{\zeta}_{I^{\overline{F}}})$$

correspond to Jacobian matrices of  $\partial_B H(x)$ , so using  $(\tilde{\eta}; \bar{\zeta}) = \sum t_j (\hat{\eta}^j; \hat{\zeta}^j)$  one has that  $\bar{g} := \bar{x} - \bar{y} + X\tilde{\eta} - Y\bar{\zeta} \in \partial\theta(x)$ .  $\square$

Observe that neither the “initial” values  $(\bar{\eta}, \bar{\zeta})$  nor the modified ones  $(\tilde{\eta}, \bar{\zeta})$  necessarily verify proposition 6.1.10, in the sense one needs to compute  $\theta'$  explicitly to ensure  $g$  is a descent direction. The following proposition details how a specific value of  $\eta$ , after the projection to obtain  $\zeta$ , can return an element of  $\partial\theta(x)$  (the degeneracies intervene in the proof but no modification of the  $\eta$  is required).

The difficulty is that since  $\zeta$  is obtained by the projection of a point depending on  $\eta$ , both quantities must be changed simultaneously.

**Proposition D.2.8** (one element of the differential). *With the same notations, consider the following problems*

$$\max_{\eta \in [-1,+1]^{\mathcal{E}^0+(x)}} \min_{\zeta \in [-1,+1]^{\mathcal{E}^-(x)}} \|g(\eta, \zeta)\|^2 / 2 = \max_{\eta \in [-1,+1]^{\mathcal{E}^0+(x)}} \text{dist}(\bar{x} + X\eta, \bar{y} + Y[-1,+1]^{\mathcal{E}^-(x)})^2 / 2$$

and let  $(\eta^{**}, \zeta^{**})$  be a solution. Then  $g(\eta^{**}, \zeta^{**}) \in \partial\theta(x)$ .

Observe that the inner minimization, which is a distance / projection, corresponds to proposition 6.1.10.

*Proof.* First, since the involved function is continuous and the sets are convex compact, the problems have solutions. The solution  $\eta^{**}$  to the outer problem is obtained at an extremal point, i.e.,  $\eta^{**} \in \{-1,+1\}^{\mathcal{E}^0+(x)}$ : indeed, since the maximized function is convex, the solution is a vertex of the polytope. Now, if  $Z_x \subseteq Z_y$ , i.e., the point  $x$  is (strongly)  $\theta$ -stationary, one gets  $g(\eta^{**}, \zeta^{**}) = 0$  which is coherent since  $0 \in \partial\theta(x)$ . Otherwise,  $g \neq 0$  and the reasoning is similar to the one in the proof of proposition D.2.7.

Now, the projection of  $\bar{x} + X\eta^{**}$  onto  $Z_y$  belongs to (the boundary of)  $Z_y$ . Using lemma B.1.4, this projection belongs to the relative interior of a face  $F_y^*$ . Using proposition B.2.2, this face  $F_y^*$  is a zonotope generated by the indices in  $\mathcal{E}^-(x)_{I^F}$  and centered by the ones in  $\mathcal{E}^-(x)_{I^*}$  (noted  $I^F$  and  $I^*$  for simplicity, see figure D.6):

$$F_y^* = \bar{y} + Y_{:, \mathcal{E}^-(x)_{I^*}} \zeta_{\mathcal{E}^-(x)_{I^*}}^{**} + Y_{:, \mathcal{E}^-(x)_{I^F}} [-1, +1]^{I^F} = \bar{y} + \tilde{y} + Y_{:, \mathcal{E}^-(x)_{I^F}} [-1, +1]^{I^F}.$$

Moreover, let  $L_y = \text{aff}(F_y^*) - \text{aff}(F_y^*)$  be the linear subspace spanned by  $F_y^*$ . Using proposition B.2.6, let the projection of  $\bar{x} + X\eta^{**}$  be  $\bar{y} + Y\zeta^{**}$ , with  $\zeta_{I^*}^{**} \in \{\pm 1\}^{I^*}$  and  $\zeta_{I^F}^{**} \in (-1, +1)^{I^F}$ . Let  $c^* = g := \bar{x} + X\eta^{**} - \bar{y} - Y\zeta^{**}$ , using that  $\bar{y} + Y\zeta^{**}$  is the projection of  $\bar{x} + X\eta^{**}$  on  $Z_y$ ,  $c^*$  verifies proposition B.2.8, i.e.,  $\zeta_i^{**} y_i^\top c^* \geq 0$ . One also has that  $Z_y \subseteq \{z \in \mathbb{R}^n : z^\top c^* \leq (\bar{y} + Y\zeta^{**})^\top c^*\}$  since  $c^*$  is an outward-pointing normal. Let us show the same property holds for the indices of  $\mathcal{E}^{0+}(x)$ .

The relevance of the  $^{**}$  variables comes from the fact that  $Z_x \subseteq \{z \in \mathbb{R}^n : z^\top c^* \leq (\bar{x} + X\eta^{**})^\top c^*\}$ . Indeed, if there was a point of  $Z_x$  in the other half-space, its distance to  $Z_y$  would be greater using the inclusion  $Z_y \subseteq \{z \in \mathbb{R}^n : z^\top c^* \leq (\bar{y} + Y\zeta^{**})^\top c^*\}$ .

Suppose that there exists some  $i \in \mathcal{E}^{0+}(x)$  such that  $\eta_i^{**} X_{:,i}^\top c^* < 0$ . Then, the point  $\bar{x} + X\eta^{**} - 2\eta_i^{**} X_{:,i} \in Z_x$  but would be such that

$$(\bar{x} + X\eta^{**} - 2\eta_i^{**} X_{:,i})^\top c^* = (\bar{x} + X\eta^{**})^\top c^* - 2\eta_i^{**} X_{:,i}^\top c^* > (\bar{x} + X\eta^{**})^\top c^*,$$

which is a contradiction. Therefore,  $\eta_i^{**} X_{:,i}^\top c^* \geq 0$ .

Finally, the question of the degeneracies, i.e., the indices  $i \in \mathcal{E}^{0+}(x)$  such that  $X_{:,i}^\top g = 0$ , cannot be treated as easily as in proposition D.2.7. Indeed, if one modifies only the parameter  $\eta$ , then the couple  $\eta, \zeta$  does not correspond to  $g = c^*$  anymore. Therefore, one needs to justify that both can be modified simultaneously to ensure  $g$  is not modified, since it must remain the same to benefit from the projection property, i.e., proposition 6.1.10.

Now, let  $\mathcal{E}^{0+}(x) := \mathcal{E}^{0+}(x)_+ \cup \mathcal{E}^{0+}(x)_0 = \{i \in \mathcal{E}^{0+}(x) : \eta_i^{**} X_{:,i}^\top g > 0\} \cup \{i \in \mathcal{E}^{0+}(x) : X_{:,i}^\top g = 0\}$ . The degenerate indices,  $\mathcal{E}^{0+}(x)_0$ , correspond to a face  $F_x^*$  of  $Z_x$  and to values of  $\eta_{\mathcal{E}^{0+}(x)_0}$  that may be changed without modification of the optimal value (one must also modify  $\zeta$ ). As in proposition D.2.7, we have that  $\text{aff}(F_x^*) - \text{aff}(F_x^*) \subseteq \text{aff}(F_y^*) - \text{aff}(F_y^*)$ , i.e., the span of  $F_x^*$  is contained in the one of  $F_y^*$ , and  $F_x^* - g \subseteq F_y^*$ .<sup>6</sup> Then, let  $\tilde{x} := X_{:, \mathcal{E}^{0+}(x)_+} \eta_{\mathcal{E}^{0+}(x)_+}^{**}$ ,  $\tilde{y} := Y_{:, I^*} \zeta_{I^*}^{**}$ , one has:

$$F_x^* := \bar{x} + \tilde{x} + X_{:, \mathcal{E}^{0+}(x)_0}[-1, +1]^{\mathcal{E}^{0+}(x)_0},$$

$$g := \bar{x} + \tilde{x} + X_{:, \mathcal{E}^{0+}(x)_0} \eta_{\mathcal{E}^{0+}(x)_0}^{**} - \bar{y} - \tilde{y} - Y_{:, I^*} \zeta_{I^*}^{**} = \tilde{w} + X_{:, \mathcal{E}^{0+}(x)_0} \eta_{\mathcal{E}^{0+}(x)_0}^{**} - Y_{:, I^*} \zeta_{I^*}^{**},$$

where  $\tilde{w} = \bar{x} - \bar{y} + \tilde{x} - \tilde{y}$ , and let  $\tilde{g} := X_{:, \mathcal{E}^{0+}(x)_0} \eta_{\mathcal{E}^{0+}(x)_0}^{**} - Y_{:, I^*} \zeta_{I^*}^{**}$  so that  $g = \tilde{w} + \tilde{g}$ . Any parameter  $\eta = (\eta_{\mathcal{E}^{0+}(x)_+}^{**}, \eta_{\mathcal{E}^{0+}(x)_0})$  for some arbitrary  $\eta_{\mathcal{E}^{0+}(x)_0}$  gives, after the projection, the same value  $g$ , though the corresponding  $\zeta_{I^*}$  depends on  $\eta_{\mathcal{E}^{0+}(x)_0}$ . One wants to obtain a certain tuple

$$(\tilde{\eta}^{**}, \tilde{\zeta}^{**}) = ((\eta_{\mathcal{E}^{0+}(x)_+}^{**}, \tilde{\eta}_{\mathcal{E}^{0+}(x)_0}^{**}), (\zeta_{I^*}^{**}, \tilde{\zeta}_{I^*}^{**}))$$

(so in particular  $g(\eta^{**}, \zeta^{**}) = g(\tilde{\eta}^{**}, \tilde{\zeta}^{**})$ ) where the values  $\eta_{\mathcal{E}^{0+}(x)_+}^{**}$  and  $\zeta_{I^*}^{**}$  are unchanged,  $\tilde{\eta}_{\mathcal{E}^{0+}(x)_0}^{**}$  and  $\tilde{\zeta}_{I^*}^{**}$  correspond to a certain convex combination of elements of  $\partial_B H(x)$ .

As in proposition D.2.7, let the vertices of  $F_y^*$  be denoted by  $\bar{y} + Y \hat{\zeta}^j$  for  $j \in J$  with  $J$  some index set. In particular,  $\hat{\zeta}_{I^*}^j = \zeta_{I^*}^{**}$  for all  $j \in J$ . For some  $j \in J$ , since  $\bar{y} + Y \hat{\zeta}^j$  is a vertex of  $Z_y$ , let  $\hat{c}^j := c^F + \varepsilon d_F^j$  where  $c^F$  is a normal vector to  $F_y^*$  given by proposition B.1.7 and  $d_F^j \in L_y$  is a direction verifying the vertex  $\hat{\zeta}_{I^*}^j$  in the subset  $I^*$ .<sup>7</sup> The vector  $\hat{c}^j$  is a verifying direction of  $\hat{\zeta}^j$ :

$$\begin{cases} i \in I^* \implies \zeta_i^{**} y_i^\top \hat{c}^j = \underbrace{\zeta_i^{**} y_i^\top c^F}_{>0} + \underbrace{\varepsilon \zeta_i^{**} y_i^\top d_F^j}_{=\varepsilon\dots} > 0, \\ i \in I^* \implies \hat{\zeta}_i^j y_i^\top \hat{c}^j = \underbrace{\hat{\zeta}_i^j y_i^\top c^F}_{=0} + \underbrace{\varepsilon \hat{\zeta}_i^j y_i^\top d_F^j}_{>0} > 0, \end{cases}$$

where the top left inequality comes from the properties of the normal, the bottom left equality from the fact  $c^F$  is a normal to the face and the bottom right inequality from the fact  $d_F^j$  is chosen as a verifying direction. Therefore, one gets that for all  $i \in \mathcal{E}^-(x)$ ,  $\hat{\zeta}_i^j y_i^\top \hat{c}^j > 0$ . Now, in the subspace  $L_y$ , up to small modifications of  $d_F^j$ , one can assume that  $X_{:,i}^\top \hat{c}^j \neq 0$  for  $i \in \mathcal{E}^{0+}(x)_0$  (still maintaining  $d_F^j \in L_y$  since these  $X_{:,i}$  also belong to  $L_y$ ). Then, for  $i \in \mathcal{E}^{0+}(x)_0$ , let  $\hat{\eta}_i^j := \text{sgn}(X_{:,i}^\top \hat{c}^j)$ , one has, using the properties of  $\hat{d}^j$  and  $c^F$ ,

$$\begin{cases} i \in \mathcal{E}^{0+}(x)_0 \implies \hat{\eta}_i X_{:,i}^\top \hat{c}^j = \underbrace{\hat{\eta}_i X_{:,i}^\top c^F}_{=0} + \underbrace{\varepsilon \hat{\eta}_i X_{:,i}^\top \hat{d}^j}_{>0} > 0, \\ i \in \mathcal{E}^{0+}(x)_+ \implies \hat{\eta}_i X_{:,i}^\top \hat{c}^j = \underbrace{\hat{\eta}_i X_{:,i}^\top c^F}_{>0} + \underbrace{\varepsilon \hat{\eta}_i X_{:,i}^\top \hat{d}^j}_{=\varepsilon\dots} > 0, \end{cases}$$

<sup>6</sup>Otherwise, by using the directions of  $X$  not spanned by  $Y$ , the distance increases which contradicts the optimality.

<sup>7</sup>Meaning that  $(\hat{\zeta}_{I^*}^j)_i y_i^\top d > 0$  for all  $i \in I^*$ .

meaning the sign vector  $(\eta_{\mathcal{E}^{0+}(x)_+}^{**}, \hat{\eta}_{\mathcal{E}^{0+}(x)_0}^j, \zeta_{I^*}^{**}, \hat{\zeta}_{IF}^j)$  corresponds to a  $s \in \mathcal{S}([X Y])$  or equivalently a matrix in  $\partial_B H(x)$  by (D.6)<sup>8</sup>. For the symmetric (in  $F^*$ ) vertex  $\bar{y} + Y_{:, I^*} \hat{\zeta}_{I^*} - Y_{:, IF} \hat{\zeta}_{IF}$ , one can take the direction  $c^F - \varepsilon d_F^j$ , so for this symmetric vertex one obtains  $-\hat{\eta}_{\mathcal{E}^{0+}(x)_0}^j$ .

This construction results in an association between (all) the vertices of  $F_y^*$  and some of  $F_x^*$ . Recall that  $F_x^* - g \subseteq F_y^*$ , which reads

$$\begin{aligned} \bar{x} + \tilde{x} + X_{:, \mathcal{E}^{0+}(x)_0}[-1, +1]^{\mathcal{E}^{0+}(x)_0} - \tilde{w} - \tilde{g} &\subseteq \bar{y} + \tilde{y} + Y_{IF}[-1, +1]^{IF} \\ \iff -\tilde{g} + X_{:, \mathcal{E}^{0+}(x)_0}[-1, +1]^{\mathcal{E}^{0+}(x)_0} &\subseteq +Y_{:, IF}[-1, +1]^{IF} \end{aligned} \quad (\text{D.12})$$

and represents a zonotope inclusion in smaller dimension and with less indices. Let us show now a useful property of the directions  $z^j := (X_{:, \mathcal{E}^{0+}(x)_0} \hat{\eta}_{\mathcal{E}^{0+}(x)_0}^j - Y_{:, IF} \hat{\zeta}_{IF}^j)$  for  $j \in J$ : one has  $\tilde{g} = X_{:, \mathcal{E}^{0+}(x)_0} \eta_{\mathcal{E}^{0+}(x)_0}^{**} - Y_{:, IF} \zeta_{IF}^{**} \in \text{conv}\{z^j : j \in J\}$ . This can be used as follows, with  $t_j \geq 0$ ,  $j \in J$  and  $\sum_J t_j = 1$ :

$$\begin{aligned} \tilde{g} &= \sum_{j \in J} t_j z^j = \sum_{j \in J} t_j (X_{:, \mathcal{E}^{0+}(x)_0} \hat{\eta}_{\mathcal{E}^{0+}(x)_0}^j - Y_{:, IF} \hat{\zeta}_{IF}^j) \\ \tilde{w} + \tilde{g} &= \sum_{j \in J} t_j (\tilde{w} + X_{:, \mathcal{E}^{0+}(x)_0} \hat{\eta}_{\mathcal{E}^{0+}(x)_0}^j - Y_{:, IF} \hat{\zeta}_{IF}^j) \\ g &= \sum_{j \in J} t_j (g_1 + X[\eta_{\mathcal{E}^{0+}(x)_+}^{**}; \hat{\eta}_{\mathcal{E}^{0+}(x)_0}^j] - Y[\zeta_{I^*}^{**}; \hat{\zeta}_{IF}^j]) \\ &= \sum_{j \in J} t_j (g_1 + [X - Y][\eta_{\mathcal{E}^{0+}(x)_+}^{**}; \hat{\eta}_{\mathcal{E}^{0+}(x)_0}^j; \zeta_{I^*}^{**}; \hat{\zeta}_{IF}^j]), \end{aligned}$$

which indicates that  $g$  is a convex combination of elements of the form  $g_1 + [X - Y]s$  with  $s \in \mathcal{S}$ , as described by (D.7), meaning  $g \in \partial\theta(x)$ . Finally, let us justify that

$$\tilde{g} = X_{:, \mathcal{E}^{0+}(x)_0} \eta_{\mathcal{E}^{0+}(x)_0}^{**} - Y_{:, IF} \zeta_{IF}^{**} \in \text{conv}\{z^j : j \in J\}.$$

Suppose the inclusion does not hold. Then, since  $\tilde{g}$  and the convex hull are convex compact sets, there exists a strict separator vector  $\bar{d}$  such that

$$\forall z \in \text{conv}\{z^j : j \in J\}, \bar{d}^\top \tilde{g} > \bar{d}^\top z.$$

By convexity, take  $z = z^j$  for every  $j \in J$ , the previous condition equivalently reads

$$\begin{aligned} \forall j \in J, \bar{d}^\top \tilde{g} &> \bar{d}^\top z^j \\ \iff \forall j \in J, \bar{d}^\top \tilde{g} &> \bar{d}^\top (X_{:, \mathcal{E}^{0+}(x)_0} \hat{\eta}_{\mathcal{E}^{0+}(x)_0}^j - Y_{:, IF} \hat{\zeta}_{IF}^j). \end{aligned}$$

In particular, since  $\{z^j, j \in J\}$  is symmetric by construction:

$$\begin{aligned} \forall j \in J, \bar{d}^\top \tilde{g} &> \bar{d}^\top (Y_{:, IF} \hat{\zeta}_{IF}^j - X_{:, \mathcal{E}^{0+}(x)_0} \hat{\eta}_{\mathcal{E}^{0+}(x)_0}^j) \\ \iff \forall j \in J, \bar{d}^\top (\tilde{g} + X_{:, \mathcal{E}^{0+}(x)_0} \hat{\eta}_{\mathcal{E}^{0+}(x)_0}^j) &> \bar{d}^\top (Y_{:, IF} \hat{\zeta}_{IF}^j). \end{aligned}$$

---

<sup>8</sup>If  $\mathcal{E}^0(x) \neq \emptyset$ , the corresponding  $X_{:, i}$  are zero but these do not intervene in the optimization problems – these indices can be removed by using the same argument as under (D.7).

---

Finally, since the convex sets are compact, one can make a small modification of  $\bar{d}$  (remaining in  $L_y$ ) to ensure that  $\bar{d}^\top z^j \neq 0$  for every  $j \in J$ . Then, define the following sign vector:

$$\bar{s} := \text{sgn}(Y_{:,I^F}^\top \bar{d}) \in \{\pm 1\}^{I^F},$$

which is the (partial) sign vector corresponding to the vertex verified by  $\bar{d}$ . Then, consider the strict inequality for the index  $\bar{j}$  corresponding to  $\bar{s}$  (which exists since  $J$  covers the vertices of  $F_y^*$  and  $\bar{s}$  is among those vertices). Since  $\bar{z} := Y_{:,I^F} \bar{s}$  maximizes  $\bar{d}^\top z$  for  $z \in Y_{:,I^F} [-1, +1]^{I^F}$ , this means  $\tilde{g} + X_{:,E^{0+}(x)_0} \hat{\eta}_{E^{0+}(x)_0}^{\bar{j}}$  does not belong to  $Y_{:,I^F} [-1, +1]^{I^F}$ . By symmetry,  $-\tilde{g} - X_{:,E^{0+}(x)_0} \hat{\eta}_{E^{0+}(x)_0}^{\bar{j}}$  also does not belong to  $Y_{:,I^F} [-1, +1]^{I^F}$ . This is a contradiction with the inclusion of the zonotopes in (D.12) with variable  $-\eta_{E^{0+}(x)_0}^{\bar{j}}$ , up to the translation from  $\pm g$ .  $\square$

**Remark D.2.9** (local and global minima). A (strict) local maxima  $\eta$  of the distance function also returns an element  $g \in \partial\theta(x)$ .

Indeed,  $g$  is a nonstrict normal as described in proposition B.2.8. By local maximality of the distance, one can show (as for global optimality) that  $\eta_i X_{:,i}^\top g \geq 0$ . From there, treating the degenerate indices can be done as in the main proof.  $\square$

**Example D.2.10** (Correct values for the differential). Consider example D.2.3. The values of  $\bar{x} + X\eta \in \mathbb{R}^2$ , or equivalently those of  $\eta \in [-1, +1]^2$  such that, after the projection,  $g(\gamma_{E^{0+}(x)}, \gamma_{E^{-}(x)}(\gamma_{E^{0+}(x)})) \in \partial\theta(x)$  are indicated in figure D.7.

On this relatively simple example, we see that obtaining an element of the C-differential may not be that easy, since the solutions form a complicated set.

However, recall that  $g(\gamma_{E^{0+}(x)}, \gamma_{E^{-}(x)}(\gamma_{E^{0+}(x)}))$  is obtained with the projection to ensure  $-g(\gamma_{E^{0+}(x)}, \gamma_{E^{-}(x)}(\gamma_{E^{0+}(x)}))$  is a descent direction. There may be many other combinations of  $\gamma_{E^{0+}(x)}$  ( $\eta$ ) and  $\gamma_{E^{-}(x)}$  ( $\zeta$ ) such that  $g(\gamma_{E^{0+}(x)}, \gamma_{E^{-}(x)})$  is a descent direction and/or an element of  $\partial\theta(x)$ .

In the following picture, we arbitrarily fix  $\zeta = 0$ : clearly,  $\zeta$  is not obtained by the projection, so a priori  $-g(\gamma_{E^{0+}(x)}, \gamma_{E^{-}(x)})$  may not be a descent direction. Actually, we see that whatever  $\eta$  is,  $-g(\gamma_{E^{0+}(x)}, \gamma_{E^{-}(x)})$  is a descent direction but the values of  $\eta$  such that  $g(\gamma_{E^{0+}(x)}, \gamma_{E^{-}(x)}) \in \partial\theta(x)$  change.

Now, we arbitrarily fix  $\zeta = e$ : clearly,  $\zeta$  is not obtained by the projection, so a priori  $-g(\gamma_{E^{0+}(x)}, \gamma_{E^{-}(x)})$  may not be a descent direction. Actually, we see that for some choice of  $\eta$  is,  $-g(\gamma_{E^{0+}(x)}, \gamma_{E^{-}(x)})$  is not a descent direction; the values of  $\eta$  such that  $g(\gamma_{E^{0+}(x)}, \gamma_{E^{-}(x)}) \in \partial\theta(x)$  also change.

In the pictures, we see that most often,  $\theta'(x, -g) \leq 0$ . However, in another situation where the zonotopes would be much closer ( $Z_x$  partly contained in  $Z_y$ ), it would be more likely that more combinations of weights would lead to  $-g$  being an ascent direction.  $\square$

## BIBLIOGRAPHY

---

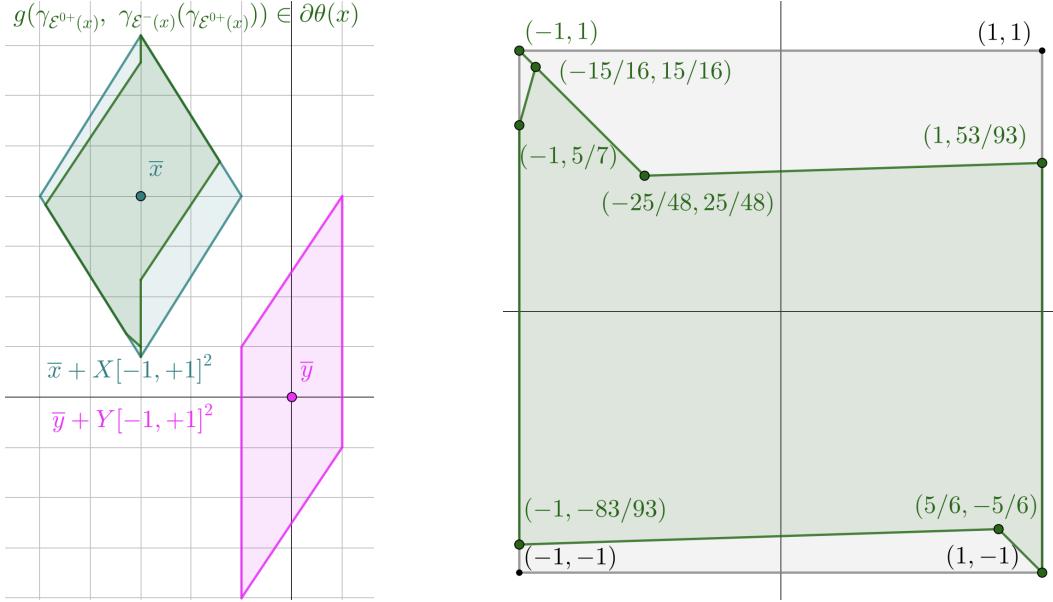


Figure D.7: Left:  $Z_y$  in magenta,  $Z_x$  in teal, and the subset of  $Z_x$  in green corresponding to the  $\gamma_{\mathcal{E}^0(x)}$ 's ( $\eta$ 's) such that the  $\gamma_{\mathcal{E}^-(x)}$ 's ( $\zeta$ 's) obtained after projection yield a  $g \in \partial\theta(x)$ . Right: corresponding values in  $[-1,+1]^2$  (i.e., with  $\eta$ ). Recall that the topmost point of the left figure corresponds to  $\eta = (1, -1)$  and the leftmost point of the left figure corresponds to  $\eta = (-1, -1)$ , which explains the change of orientation (to recover a shape similar to the left picture, turn by  $+3\pi/4$  counterclockwise then apply axial symmetry vertically).

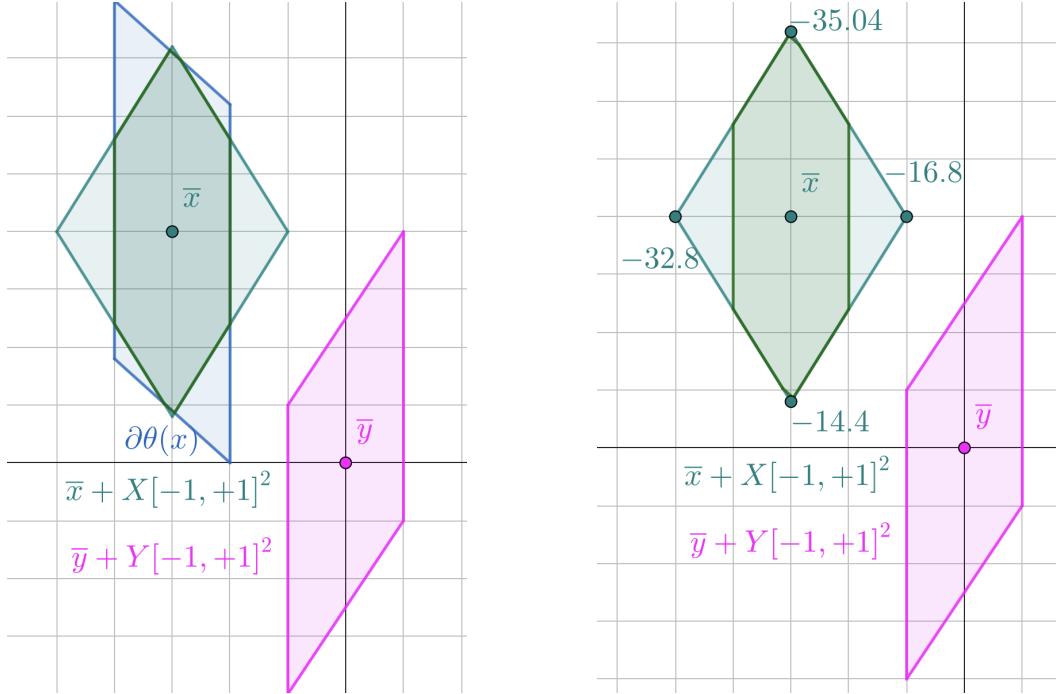


Figure D.8: In magenta,  $Z_y$  and in teal,  $Z_x$ . Left:  $\partial\theta(x)$  in blue, the intersection with  $Z_x$  in darker green corresponds to the  $\eta$  with  $g(\eta, \zeta = 0) \in \partial\theta(x)$ . Right: values of  $g'(x, -g(\gamma_{\mathcal{E}^0(x)}, \gamma_{\mathcal{E}^-(x)}))$  for the extremal  $\eta$ 's: every  $g$  with  $\zeta = 0$  is a descent direction.

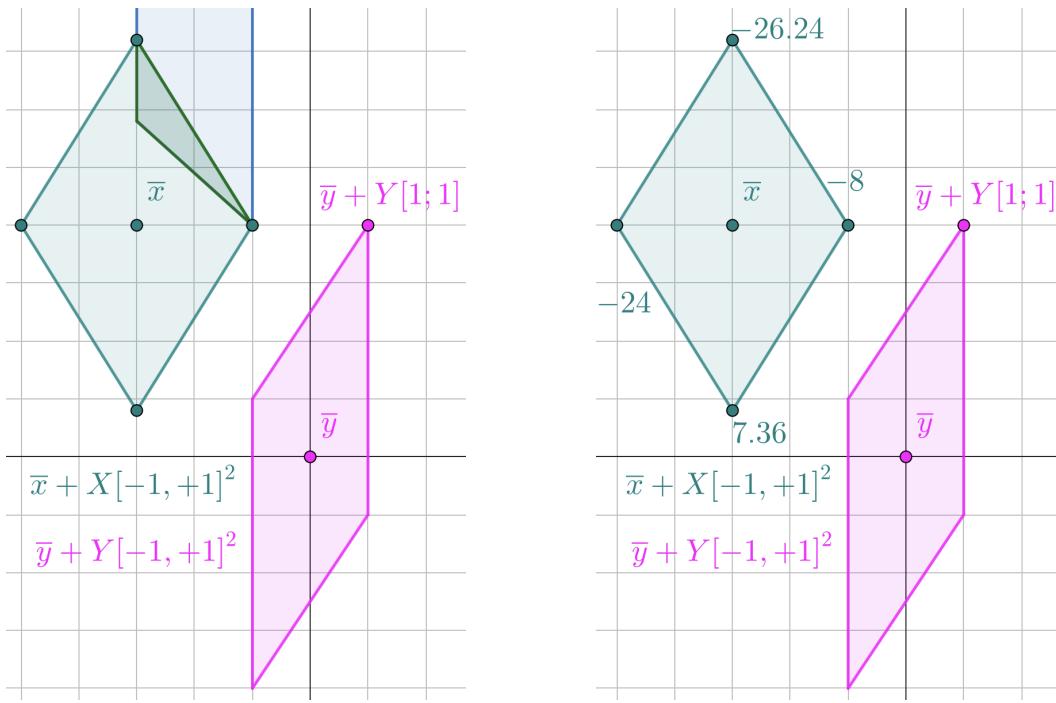


Figure D.9: In magenta,  $Z_y$  and in teal,  $Z_x$ . Left:  $\partial\theta(x)$  in blue, the intersection with  $Z_x$  in darker green corresponds to the  $\eta$  with  $g(\eta, \zeta = 0) \in \partial\theta(x)$ . Right: values of  $\theta'(x, -g(\gamma_{\mathcal{E}^{0+}(x)}, \gamma_{\mathcal{E}^-(x)}))$  for the extremal  $\eta$ 's: for some specific  $\eta$ ,  $-g$  is an ascent direction.

# Bibliography

- [1] L. Abdallah, M. Haddou, and T. Migot. “Solving Absolute Value Equation Using Complementarity and Smoothing Functions”. In: *Journal of Computational and Applied Mathematics* 327 (Jan. 2018), pp. 196–207. ISSN: 03770427. doi: 10.1016/j.cam.2017.06.019 (cit. on p. 43).
- [2] Vincent Acary and 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 (cit. on pp. 2, 3).
- [3] Muhamed Aganagić. “Newton’s Method for Linear Complementarity Problems”. In: *Mathematical Programming* 28.3 (Oct. 1984), pp. 349–362. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF02612339 (cit. on pp. 28, 57, 225).
- [4] Marcelo Aguiar and Swapneel Mahajan. *Topics in Hyperplane Arrangements*. Vol. 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 (cit. on pp. 5, 47, 73, 75, 76, 78, 143, 148).
- [5] Jan Harold Alcantara and 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 (June 2022), p. 114092. ISSN: 03770427. doi: 10.1016/j.cam.2022.114092 (cit. on pp. 28, 36).
- [6] Jan Harold Alcantara, Chen-Han Lee, Chieu Thanh Nguyen, Yu-Lin Chang, and Jein-Shan Chen. “On Construction of New NCP Functions”. In: *Operations Research Letters* 48.2 (Mar. 2020), pp. 115–121. ISSN: 01676377. doi: 10.1016/j.orl.2020.01.002 (cit. on pp. 4, 19).
- [7] Gerald L. Alexanderson and John E. Wetzel. “Arrangements of Planes in Space”. In: *Discrete Mathematics* (1981), pp. 219–240 (cit. on pp. 47, 78).
- [8] Xavier Allamigeon, Stéphane Gaubert, and Frédéric Meunier. “Tropical Complementarity Problems and Nash Equilibria”. In: *SIAM Journal on Discrete Mathematics* 37.3 (Sept. 2023), pp. 1645–1665. ISSN: 0895-4801, 1095-7146. doi: 10.1137/21M1446861 (cit. on p. 2).
- [9] Mihai Anitescu and Florian Alexandru Potra. “Formulating Dynamic Multi-Rigid-Body Contact Problems with Friction as Solvable Linear Complementarity Problems”. In: *Nonlinear Dynamics* 14 (1997), pp. 231–247 (cit. on p. 2).

- 
- [10] Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. “Lower Bounds for Non-Convex Stochastic Optimization”. In: *Mathematical Programming* 199.1-2 (May 2023), pp. 165–214. ISSN: 0025-5610, 1436-4646. doi: 10.1007/s10107-022-01822-7 (cit. on p. 45).
  - [11] Larry Armijo. “Minimization of Functions Having Lipschitz Continuous First Partial Derivatives”. In: *Pacific Journal of Mathematics* 16.1 (Jan. 1966), pp. 1–3. ISSN: 0030-8730, 0030-8730. doi: 10.2140/pjm.1966.16.1 (cit. on p. 25).
  - [12] Christos A. Athanasiadis. “Characteristic Polynomials of Subspace Arrangements and Finite Fields”. In: *Advances in Mathematics* 122.2 (Sept. 1996), pp. 193–233. ISSN: 00018708. doi: 10.1006/aima.1996.0059 (cit. on pp. 5, 48, 53, 129, 133, 143, 146).
  - [13] David Avis and Komei Fukuda. “A Pivoting Algorithm for Convex Hulls and Vertex Enumeration of Arrangements and Polyhedra”. In: *Discrete Computational Geometry* 8 (1992), pp. 295–31. ISSN: 0179-5376, 1432-0444. doi: 10.1007/BF02293050 (cit. on pp. 51, 143, 152, 218).
  - [14] David Avis and Komei Fukuda. “Reverse Search for Enumeration”. In: *Discrete Applied Mathematics* 65 (Mar. 1996), pp. 21–46. doi: 10.1016/0166-218X(95)00026-N (cit. on pp. 6, 51, 61, 73, 76, 85).
  - [15] Pierre Baldi. “Deep Learning in Biomedical Data Science”. In: *Annual Review of Biomedical Data Science* 1.1 (July 2018), pp. 181–205. ISSN: 2574-3414, 2574-3414. doi: 10.1146/annurev-biodatasci-080917-013343 (cit. on p. 69).
  - [16] Pierre Baldi and Roman Vershynin. “Polynomial Threshold Functions, Hyperplane Arrangements, and Random Tensors”. In: *SIAM Journal on Mathematics of Data Science* 1.4 (Jan. 2019), pp. 699–729. ISSN: 2577-0187. doi: 10.1137/19M1257792 (cit. on pp. 53, 61, 69).
  - [17] Antoine Bambade, Fabian Schramm, Sarah El-Kazdadi, Stéphane Caron, Adrien Taylor, and Justin Carpentier. “PROXQP: An Efficient and Versatile Quadratic Programming Solver for Real-Time Robotics Applications and Beyond”. In: (2023), p. 17 (cit. on p. 228).
  - [18] Laurence Beaude, Konstantin Brenner, Simon Lopez, Roland Masson, and 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 (June 2019), pp. 443–470. ISSN: 1420-0597, 1573-1499. doi: 10.1007/s10596-018-9794-9 (cit. on p. 2).
  - [19] Amir Beck and 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), pp. 56–79. ISSN: 1052-6234, 1095-7189. doi: 10.1137/18M1217760 (cit. on pp. 44, 219, 231).
  - [20] Ibtihel Ben Gharbia. “Résolution de Problèmes de Complémentarité.: Application à Un Écoulement Diphasique Dans Un Milieu Poreux”. PhD thesis. Université Paris Dauphine Paris IX, 2012 (cit. on pp. 2, 13, 28, 29, 33, 208).

- [21] Ibtihel Ben Gharbia, Joëlle Ferzly, Martin Vohralík, and 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 (Mar. 2023), p. 114765. ISSN: 03770427. doi: 10.1016/j.cam.2022.114765 (cit. on pp. 2, 16).
- [22] Ibtihel Ben Gharbia and 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), pp. 349–364. ISSN: 0025-5610, 1436-4646. doi: 10.1007/s10107-010-0439-6 (cit. on p. 57).
- [23] Ibtihel Ben Gharbia and Jean Charles Gilbert. "An Algorithmic Characterization of \$P\$-Matricity". In: *SIAM Journal on Matrix Analysis and Applications* 34.3 (Jan. 2013), pp. 904–916. ISSN: 0895-4798, 1095-7162. doi: 10.1137/120883025 (cit. on pp. 13, 29, 57).
- [24] Ibtihel Ben Gharbia and 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), pp. 800–813. ISSN: 0895-4798, 1095-7162. doi: 10.1137/18M1168522 (cit. on pp. 13, 29, 57).
- [25] Ibtihel Ben Gharbia and Jérôme Jaffré. "Gas Phase Appearance and Disappearance as a Problem with Complementarity Constraints". In: *Mathematics and Computers in Simulation* 99 (May 2014), pp. 28–36. ISSN: 03784754. doi: 10.1016/j.matcom.2013.04.021 (cit. on p. 57).
- [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 (cit. on p. 90).
- [27] Hanspeter Bieri and Walter Nef. "A Recursive Sweep-Plane Algorithm, Determining All Cells of a Finite Division of  $R^d$ ". In: *Computing* 28 (1982), pp. 189–198 (cit. on pp. 6, 51, 85, 143).
- [28] Anders Björner, Michel Las Vergnas, Bernd Sturmfels, Neil White, and Günter Ziegler. *Oriented Matroids*. Cambridge, UK: Cambridge University Press, 2000. ISBN: 0-521-77750-X (cit. on p. 48).
- [29] J Frédéric Bonnans, Jean Charles Gilbert, Claude Lemaréchal, and Claudia A Sagastizábal. *Numerical Optimization: Theoretical and Practical Aspects*. second. Universitext. Berlin: Springer-Verlag, 2006. ISBN: 3-540-35445-X (cit. on pp. 87, 90, 171, 181, 272).
- [30] J. Frédéric Bonnans. "Local Analysis of Newton-type Methods for Variational Inequalities and Nonlinear Programming". In: *Applied Mathematics & Optimization* 29.2 (Mar. 1994), pp. 161–186. ISSN: 0095-4616. doi: 10.1007/BF01204181 (cit. on p. 30).
- [31] J. Frédéric Bonnans, Jean Charles Gilbert, Claude Lemaréchal, and Claudia A Sagastizábal. *Optimisation Numérique – Aspects théoriques et pratiques*. Mathématiques et Applications 27. Springer Verlag, Berlin, 1997 (cit. on pp. 87, 90).
- [32] Jonathan M. Borwein and 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 (cit. on pp. 57, 69).

- 
- [33] Marie-Charlotte Brandenburg, Jesús De Loera, and Chiara Meroni. *The Best Ways to Slice a Polytope*. July 2024. doi: 10.1090/mcom/4006 (cit. on p. 143).
- [34] David Bremner, Komei Fukuda, and Ambros Marzetta. “Primal–Dual Methods for Vertex and Facet Enumeration”. In: *Discrete & Computational Geometry* 20.3 (Oct. 1998), pp. 333–357. ISSN: 0179-5376. doi: 10.1007/PL00009389 (cit. on pp. 51, 218).
- [35] Taylor Brysiewicz, Holger Eble, and Lukas Kühne. “Computing Characteristic Polynomials of Hyperplane Arrangements with Symmetries”. In: *Discrete & Computational Geometry* 70.4 (Dec. 2023), pp. 1356–1377. ISSN: 0179-5376, 1432-0444. doi: 10.1007/s00454-023-00557-2 (cit. on pp. 6, 48, 50, 53, 96, 129, 133, 137, 139, 143, 199, 245, 252).
- [36] Hannes Buchholzer, Christian Kanzow, Peter Knabner, and 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), pp. 193–221. ISSN: 0926-6003, 1573-2894. doi: 10.1007/s10589-010-9379-6 (cit. on p. 2).
- [37] R. Creighton Buck. “Partition of Space”. In: *American Mathematical Monthly* 50 (1943), pp. 541–544. ISSN: 0002-9890, 1930-0972. doi: 10.2307/2303424 (cit. on p. 47).
- [38] Quan M. Bui and 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 (Apr. 2020), p. 109163. ISSN: 00219991. doi: 10.1016/j.jcp.2019.109163 (cit. on p. 2).
- [39] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. “Lower Bounds for Finding Stationary Points I”. In: *Mathematical Programming* 184.1-2 (Nov. 2020), pp. 71–120. ISSN: 0025-5610, 1436-4646. doi: 10.1007/s10107-019-01406-y (cit. on pp. 45, 46).
- [40] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. “Lower Bounds for Finding Stationary Points II: First-Order Methods”. In: *Mathematical Programming* 185.1-2 (Feb. 2021), pp. 315–355. ISSN: 0025-5610. doi: 10.48550/arXiv.1711.00841. arXiv: 1711.00841 [math] (cit. on p. 45).
- [41] Frédéric Cazals and Sébastien Loriot. “Computing the Arrangement of Circles on a Sphere, with Applications in Structural Biology”. In: *Computational Geometry* 42.6-7 (Aug. 2009), pp. 551–565. ISSN: 09257721. doi: 10.1016/j.comgeo.2008.10.004 (cit. on p. 143).
- [42] Michal Černý, Miroslav Rada, Jaromír Antoch, and Milan Hladík. “A Class of Optimization Problems Motivated by Rank Estimators in Robust Regression”. In: *Optimization* 71 (2022), pp. 2241–2271. doi: 10.48550/arXiv.1910.05826. arXiv: 1910.05826 [math] (cit. on pp. 52, 73).
- [43] Bintong Chen, Xiaojun Chen, and Christian Kanzow. “A Penalized Fischer-Burmeister NCP-function: Theoretical Investigation and Numerical Results”. In: *Mathematical Programming* 88.1 (June 2000), pp. 211–216. ISSN: 0025-5610. doi: 10.1007/PL00011375 (cit. on p. 37).
- [44] Bintong Chen and 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), pp. 1168–1190 (cit. on pp. 3, 16).

- [45] Xiaojun Chen. "Superlinear Convergence of Smoothing Quasi-Newton Methods for Nonsmooth Equations". In: *Journal of Computational and Applied Mathematics* 80.1 (Apr. 1997), pp. 105–126. ISSN: 03770427. doi: 10.1016/S0377-0427(97)80133-1 (cit. on pp. 4, 42).
- [46] Xiaojun Chen, Zuhair Nashed, and Liqun Qi. "Smoothing Methods and Semismooth Methods for Nondifferentiable Operator Equations". In: *SIAM Journal on Numerical Analysis* 38.4 (Jan. 2000), pp. 1200–1216. ISSN: 0036-1429, 1095-7170. doi: 10.1137/S0036142999356719 (cit. on pp. 4, 42).
- [47] Xiaojun Chen, Liqun Qi, and 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), pp. 519–540. ISSN: 0025-5718, 1088-6842. doi: 10.1090/S0025-5718-98-00932-6 (cit. on pp. 3, 4, 41).
- [48] Xiaojun Chen and Shuhuang Xiang. "Sparse Solutions of Linear Complementarity Problems". In: *Mathematical Programming* 159.1-2 (Sept. 2016), pp. 539–556. ISSN: 0025-5610, 1436-4646. doi: 10.1007/s10107-015-0950-x (cit. on p. 1).
- [49] Sung Jin Chung. "NP-Completeness of the Linear Complementarity Problem". In: *Journal of Optimization Theory and Applications* 60.3 (Mar. 1989), pp. 393–399. ISSN: 0022-3239, 1573-2878. doi: 10.1007/BF00940344 (cit. on pp. 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 (cit. on pp. 87, 90, 182).
- [51] Frank H. Clarke. *Optimization and Nonsmooth Analysis*. second. Vol. 1. Classics in Applied Mathematics. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics (SIAM), 1990. ISBN: 0-89871-256-4 (cit. on pp. 4, 19–23, 57–59, 119, 220, 222).
- [52] Kenneth L Clarkson and Peter W Shor. "Applications of Random Sampling in Computational Geometry, II". In: *Discrete & Computational Geometry* 4 (1989), pp. 387–421. ISSN: 0179-5376, 1432-0444. doi: 10.1007/BF02187740 (cit. on pp. 48, 78).
- [53] Kenneth L. Clarkson. "New Applications of Random Sampling in Computational Geometry". In: *Discrete & Computational Geometry* 2.2 (June 1987), pp. 195–222. ISSN: 0179-5376, 1432-0444. doi: 10.1007/BF02187879 (cit. on p. 48).
- [54] Richard W. Cottle and George B. Dantzig. "A Generalization of the Linear Complementarity Problem". In: *Journal of Combinatorial Theory* 8.1 (Jan. 1970), pp. 79–90. ISSN: 00219800. doi: 10.1016/S0021-9800(70)80010-2 (cit. on pp. 2, 57).
- [55] Richard Warren Cottle. "Linear Complementarity since 1978". In: *Variational Analysis and Applications*. Nonconvex Optimization and Its Applications 79.1 (2005), pp. 239–257 (cit. on p. 2).
- [56] Richard Warren Cottle. "Nonlinear Programs with Positively Bounded Jacobians". ? Berkeley, USA: University of California, 1964 (cit. on p. 1).
- [57] Richard Warren Cottle. "Nonlinear Programs with Positively Bounded Jacobians". In: *SIAM Journal on Applied Mathematics* 14.1 (Jan. 1966), pp. 1–12. doi: 10.1137/0114012 (cit. on p. 1).

- 
- [58] Richard Warren Cottle, Jong-Shi Pang, and 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 (cit. on pp. 1, 2, 12, 16, 57, 222, 223).
- [59] Thomas M. Cover and 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 (cit. on p. 89).
- [60] Gregory E. Coxson. “The P-matrix Problem Is Co-NP-complete”. In: *Mathematical Programming* 64.1-3 (Mar. 1994), pp. 173–178. ISSN: 0025-5610, 1436-4646. doi: 10 . 1007/BF01582570 (cit. on pp. 3, 13).
- [61] Henry H. Crapo and Gian-Carlo Rota. *On the Foundations of Combinatorial Theory: Combinatorial Geometries*. The M.I.T. Press, Cambridge, Mass.-London, 1970 (cit. on pp. 5, 47, 143).
- [62] Jad Dabaghi, Vincent Martin, and Martin Vohralík. “Adaptive Inexact Semismooth Newton Methods for the Contact Problem Between Two Membranes”. In: *Journal of Scientific Computing* 84.2 (Aug. 2020), p. 28. ISSN: 0885-7474, 1573-7691. doi: 10 . 1007/s10915-020-01264-3 (cit. on p. 2).
- [63] Aris Daniilidis, Mounir Haddou, Tri Minh Le, and Olivier Ley. “Solving Nonlinear Absolute Value Equations”. In: (2024) (cit. on pp. 16, 43).
- [64] Jesús De Loera, Jörg Rambau, and Francisco Santos. *Triangulations - Structures for Algorithms and Applications*. Algorithms and Computation in Mathematics 25. Berlin: Springer-Verlag, 2010. ISBN: 978-3-642-12970-4 (cit. on p. 47).
- [65] Tecla De Luca, Francisco Facchinei, and Christian Kanzow. “A Semismooth Equation Approach to the Solution of Nonlinear Complementarity Problems”. In: *Mathematical Programming* 75.3 (Dec. 1996), pp. 407–439. ISSN: 0025-5610, 1436-4646. doi: 10 . 1007 / BF02592192 (cit. on pp. 24, 25, 35).
- [66] Tecla De Luca, Francisco Facchinei, and Christian Kanzow. “A Theoretical and Numerical Comparison of Some Semismooth Algorithms for Complementarity Problems”. In: *Computational Optimization and Applications* 16 (Jan. 2000), pp. 173–205. doi: 10 . 1023 / A : 1008705425484 (cit. on pp. 28, 35, 60).
- [67] Wolfram Decker, Christian Eder, Claus Fieker, Max Horn, and Michael Joswig. *The Computer Algebra System OSCAR: Algorithms and Examples*. 1st ed. Vol. 32. Algorithms and {C}omputation in {M}athematics. Springer, 2024. ISBN: 1431-1550 (issn) (cit. on p. 6).
- [68] The Sage Developers. *Sagemath, the Sage Mathematics*. 2024 (cit. on pp. 6, 50, 143).
- [69] Elizabeth D. Dolan and Jorge J. Moré. “Benchmarking Optimization Software with Performance Profiles”. In: *Mathematical Programming* 91.2 (Jan. 2002), pp. 201–213. ISSN: 0025-5610, 1436-4646. doi: 10 . 1007/s101070100263 (cit. on p. 196).
- [70] György Dósa, István Szalkai, and 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), pp. 383–392 (cit. on pp. 50, 97, 154).

- [71] Jean-Pierre Dussault, Mathieu Frappier, and 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 (Dec. 2019), pp. 359–380. ISSN: 21924406. doi: 10.1007/s13675-019-00116-6 (cit. on p. 57).
- [72] Jean-Pierre Dussault, Mathieu Frappier, and Jean Charles Gilbert. “Polyhedral Newton-min Algorithms for Complementarity Problems”. In: *Mathematical Programming* (2025) (cit. on pp. 7, 9, 31, 33, 57, 203–206, 225, 232).
- [73] Jean-Pierre Dussault and 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 (May 2023), 1221–1238\*. doi: 10.1007/s10107-022-01860-1 (cit. on pp. 28, 57).
- [74] Jean-Pierre Dussault, Jean Charles Gilbert, and Baptiste Plaquevent-Jourdain. *Computing the B-differential of the Componentwise Minimum of Two Vector Functions - Partial Description by Linearization*. Tech. rep. Inria Paris, Université de Sherbrooke, 2025 (cit. on pp. 74, 75, 83, 84, 104).
- [75] Jean-Pierre Dussault, Jean Charles Gilbert, and Baptiste Plaquevent-Jourdain. *ISF and BD-IFFMIN*. 2023 (cit. on pp. 6, 84, 96, 104, 196).
- [76] Jean-Pierre Dussault, Jean Charles Gilbert, and Baptiste Plaquevent-Jourdain. *ISF and BD-IFFMIN - 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 (cit. on pp. 6, 84, 196).
- [77] Jean-Pierre Dussault, Jean Charles Gilbert, and Baptiste Plaquevent-Jourdain. “On the B-differential of the Componentwise Minimum of Two Affine Vector Functions”. In: *Mathematical Programming Computation* (2025) (cit. on pp. 6, 55, 107, 143–145, 148, 152–154, 157, 164, 165, 169, 173, 174, 176, 184, 193, 195, 196, 235).
- [78] Jean-Pierre Dussault, Jean Charles Gilbert, and 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 (cit. on pp. 59, 60, 69, 75, 76, 78, 80, 83, 92, 106, 171).
- [79] Jean-Pierre Dussault, Jean Charles Gilbert, and Baptiste Plaquevent-Jourdain. “Primal and Dual Approaches for the Chamber Enumeration of Real Hyperplane Arrangements”. In: *(submitted)* (2025) (cit. on pp. 7, 85, 95, 97, 104, 141).
- [80] Jean-Pierre Dussault, Jean Charles Gilbert, and 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 (cit. on pp. 144, 148, 151, 152, 155, 166, 188–190, 193).
- [81] Herbert Edelsbrunner. *Algorithms in Combinatorial Geometry*. Vol. 10. EATCS Monographs on Theoretical Computer Science. Berlin: Springer-Verlag, 1987. ISBN: 3-540-13722-X (cit. on pp. 5, 47, 73, 78, 143, 218).

- 
- [82] Herbert Edelsbrunner and Leonidas J. Guibas. “Topologically Sweeping an Arrangement”. In: *Journal of Computer and system Sciences* 38 (1989), pp. 165–194. ISSN: 2543-991X, 2080-5519. doi: 10.14708/wm.v48i2.316 (cit. on p. 51).
- [83] Herbert Edelsbrunner, Joseph O’ROURKE, and Raimund Seidel. “CONSTRUCTING ARRANGEMENTS OF LINES AND HYPERPLANES WITH APPLICATIONS”. In: *SIAM Journal on Computation* 15.2 (1986), pp. 341–363 (cit. on pp. 6, 51, 73, 85, 143).
- [84] Kenny Erleben and 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 (cit. on p. 2).
- [85] Francisco Facchinei and Christian Kanzow. “A Nonsmooth Inexact Newton Method for the Solution of Large-Scale Nonlinear Complementarity Problems”. In: *Mathematical Programming* 76.3 (Mar. 1997), pp. 493–512. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF02614395 (cit. on p. 35).
- [86] Francisco Facchinei and Jong-Shi Pang, eds. *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 (cit. on pp. 1–4, 12, 16, 17, 25, 29–32, 34, 38, 41, 57, 58, 228).
- [87] Francisco Facchinei and João Soares. “A New Merit Function For Nonlinear Complementarity Problems And A Related Algorithm”. In: *SIAM Journal on Optimization* 7.1 (Feb. 1997), pp. 225–247. ISSN: 1052-6234, 1095-7189. doi: 10.1137/S1052623494279110 (cit. on pp. 30, 34, 57, 120, 126, 223).
- [88] Yahya Fathi. “Computational Complexity of LCPs Associated with Positive Definite Symmetric Matrices”. In: *Mathematical Programming* 17.1 (Dec. 1979), pp. 335–344. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF01588254 (cit. on p. 3).
- [89] J.-A. Ferrez, Komei Fukuda, and 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), pp. 35–50. ISSN: 03772217. doi: 10.1016/j.ejor.2003.04.011 (cit. on p. 52).
- [90] Michael C. Ferris, Christian Kanzow, and Todd S. Munson. “Feasible Descent Algorithms for Mixed Complementarity Problems”. In: *Mathematical Programming* 86.3 (Dec. 1999), pp. 475–497. ISSN: 0025-5610. doi: 10.1007/s101070050101 (cit. on p. 3).
- [91] Michael C. Ferris and Jong-Shi Pang. “Engineering and Economic Applications of Complementarity Problems”. In: *SIAM Review* 39.4 (Jan. 1997), pp. 669–713. ISSN: 0036-1445, 1095-7200. doi: 10.1137/S0036144595285963 (cit. on p. 2).
- [92] Andreas Fischer. “A Special Newton-type Optimization Method”. In: *Optimization* 24 (Jan. 1992), pp. 269–284. doi: 10.1080/02331939208843795 (cit. on pp. 4, 17, 34).
- [93] Andreas Fischer. “Solution of Monotone Complementarity Problems with Locally Lipschitzian Functions”. In: *Mathematical Programming* 76.3 (Mar. 1997), pp. 513–532. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF02614396 (cit. on p. 35).

- [94] Andreas Fischer and Houyuan Jiang. "Merit Functions for Complementarity and Related Problems: A Survey". In: *Computational Optimization and Applications* 17 (Dec. 2000), pp. 159–182. doi: 10.1023/A:1026598214921 (cit. on pp. 4, 18).
- [95] Andreas Fischer and Christian Kanzow. "On Finite Termination of an Iterative Method for Linear Complementarity Problems". In: *Mathematical Programming* 74.3 (Sept. 1996), pp. 279–292. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF02592200 (cit. on pp. 28–30).
- [96] Robert M. Freund and James B. Orlin. "On the Complexity of Four Polyhedral Set Containment Problems". In: *Mathematical Programming* 33.2 (Nov. 1985), pp. 139–145. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF01582241 (cit. on p. 218).
- [97] Komei Fukuda. "From the Zonotope Construction to the Minkowski Addition of Convex Polytopes". In: *Journal of Symbolic Computation* 38.4 (Oct. 2004), pp. 1261–1272. ISSN: 07477171. doi: 10.1016/j.jsc.2003.08.007 (cit. on p. 51).
- [98] Masao Fukushima. "Equivalent Differentiable Optimization Problems and Descent Methods for Asymmetric Variational Inequality Problems". In: *Mathematical Programming* 53.1-3 (Jan. 1992), pp. 99–110. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF01585696 (cit. on pp. 3, 4, 38, 39).
- [99] Aurél Galántai. "Properties and Construction of NCP Functions". In: *Computational Optimization and Applications* 52.3 (July 2012), pp. 805–824. ISSN: 0926-6003, 1573-2894. doi: 10.1007/s10589-011-9428-9 (cit. on pp. 4, 18).
- [100] Yu Gao, Haiming Song, Xiaoshen Wang, and 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 (cit. on p. 2).
- [101] Ewgenij Gawrilow and Michael Joswig. "Polymake: A Framework for Analyzing Convex Polytopes." In: *Polytopes—Combinatorics and Computation (Oberwolfach, 1997)*. DMV Sem. 29. Basel: Birkhäuser, 2000, pp. 43–73. ISBN: 3-7643-6351-7 (cit. on p. 50).
- [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] (cit. on p. 46).
- [103] Carl Geiger and Christian Kanzow. "On the Resolution of Monotone Complementarity Problems". In: *Computational Optimization and Applications* 5.2 (Mar. 1996), pp. 155–173. ISSN: 0926-6003, 1573-2894. doi: 10.1007/BF00249054 (cit. on p. 35).
- [104] Helmut Gfrerer and Jiří V. Outrata. "On a Semismooth\* Newton Method for Solving Generalized Equations". In: *SIAM Journal on Optimization* 31.1 (Jan. 2021), pp. 489–517. ISSN: 1052-6234, 1095-7189. doi: 10.1137/19M1257408 (cit. on p. 41).
- [105] Jean Charles Gilbert. *Fragments d'Optimisation Différentiable - Théories et Algorithmes*. /. Vol. 1. /. /: /, 2021 (cit. on pp. 14, 26, 90, 107, 181, 182, 268, 272).
- [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 (cit. on p. 87).

- 
- [107] Allen A. Goldstein. “Optimization of Lipschitz Continuous Functions”. In: *Mathematical Programming* 13.1 (Dec. 1977), pp. 14–22. ISSN: 0025-5610, 1436-4646. doi: 10 . 1007 / BF01584320 (cit. on p. 45).
- [108] Paul Gordan. “Über die Auflösung linearer Gleichungen mit reellen Coefficienten.” In: *Mathematische Annalen* (1873), pp. 23–28. doi: 10 . 1007/BF01442864 (cit. on pp. 50, 66, 143, 145).
- [109] Nicholas Ian Mark Gould and Jennifer Scott. “A Note on Performance Profiles for Benchmarking Software”. In: *ACM Transactions on Mathematical Software* 43.2 (June 2017), pp. 1–5. ISSN: 0098-3500, 1557-7295. doi: 10 . 1145 / 2950048 (cit. on p. 196).
- [110] M. Seetharama Gowda and Roman Sznajder. “The Generalized Order Linear Complementarity Problem”. In: *SIAM Journal on Matrix Analysis and Applications* 15.3 (July 1994), pp. 779–795. ISSN: 0895-4798. doi: 10 . 1137 / S0895479892237859 (cit. on p. 57).
- [111] Daniel R. Grayson and Michael E. Stillman. *Macaulay2, a Software System for Research in Algebraic Geometry*. 2024 (cit. on pp. 6, 50, 143).
- [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 (cit. on p. 69).
- [113] Luigi Grippo, Francesco Lampariello, and Stefano Lucidi. “A Nonmonotone Line Search Technique for Newton’s Method”. In: *SIAM Journal on Numerical Analysis* 23.4 (Aug. 1986), pp. 707–716. ISSN: 0036-1429, 1095-7170. doi: 10 . 1137 / 0723046 (cit. on p. 25).
- [114] Branko Grünbaum. *Convex Polytopes*. Interscience Publishers John Wiley & Sons, Inc. Vol. 16. Pure and Applied Mathematics. New York: AMS, Providence, RI, 1967 (cit. on pp. 52, 73, 78).
- [115] Osman Güler. *Foundations of Optimization*. Vol. 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 (cit. on p. 145).
- [116] Mounir Haddou. “A New Class of Smoothing Methods for Mathematical Programs with Equilibrium Constraints”. In: *Pacific Journal of Optimization* 5 (2009), pp. 87–95 (cit. on p. 42).
- [117] Mounir Haddou and Patrick Maheux. “Smoothing Methods for Nonlinear Complementarity Problems”. In: *Journal of Optimization Theory and Applications* 160.3 (Mar. 2014), pp. 711–729. ISSN: 0022-3239, 1573-2878. doi: 10 . 1007 / s10957 - 013 - 0398 - 1 (cit. on pp. 4, 42).
- [118] Dan Halperin and Micha Sharir. “Arrangements”. In: *Handbook of Discrete and Computational Geometry*. Jacob E. Goodman and Joseph O’Rourke and Csaba D. Tóth. CRC Press - Taylor & Francis Group, 2018 (cit. on pp. 5, 47, 73, 143).
- [119] Shih-Ping Han, Jong-Shi Pang, and Narayan Rangaraj. “Globally Convergent Newton Methods for Nonsmooth Equations”. In: *Mathematics of Operations Research* 17.3 (Aug. 1992), pp. 586–607. ISSN: 0364-765X, 1526-5471. doi: 10 . 1287 / moor . 17 . 3 . 586 (cit. on pp. 21, 31, 32).

- [120] Patrick T. Harker and Jong-Shi Pang. “Finite-Dimensional Variational Inequality and Nonlinear Complementarity Problems: A Survey of Theory, Algorithms and Applications”. In: *Mathematical Programming* 48.1-3 (Mar. 1990), pp. 161–220. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF01582255 (cit. on p. 2).
- [121] Patrick T. Harker and Baichun Xiao. “Newton’s Method for the Nonlinear Complementarity Problem: A B-differentiable Equation Approach”. In: *Mathematical Programming* 48.1-3 (Mar. 1990), pp. 339–357. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF01582262 (cit. on p. 15).
- [122] Xiahui He and 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). Ed. by Vyacheslav Kalashnikov, pp. 1–8. ISSN: 1024-123X, 1563-5147. doi: 10.1155/2019/2912301 (cit. on pp. 2, 3, 40).
- [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 (cit. on p. 19).
- [124] Michael Hintermüller, Kazufumi Ito, and Karl Kunisch. “The Primal-Dual Active Set Strategy as a Semismooth Newton Method”. In: *SIAM Journal on Optimization* 13.3 (Jan. 2002), pp. 865–888. ISSN: 1052-6234, 1095-7189. doi: 10.1137/S1052623401383558 (cit. on pp. 3, 40).
- [125] Michael Hintermüller and 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), pp. 868–902. ISSN: 1052-6234, 1095-7189. doi: 10.1137/080720681 (cit. on p. 16).
- [126] Jean-Baptiste Hiriart-Urruty and 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 (cit. on pp. 57, 225).
- [127] Tim Hoheisel, Christian Kanzow, Boris S. Mordukhovich, and Hung M. Phan. “Generalized Newton’s Method Based on Graphical Derivatives”. In: *Nonlinear Analysis: Theory, Methods & Applications* 75.3 (Feb. 2012), pp. 1324–1340. ISSN: 0362546X. doi: 10.1016/j.na.2011.06.039 (cit. on pp. 4, 41).
- [128] Stefan Hüeber, Georg Stadler, and 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), pp. 572–596. ISSN: 1064-8275, 1095-7197. doi: 10.1137/060671061 (cit. on pp. 2, 3).
- [129] Stefan Hüeber and 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 (July 2005), pp. 3147–3166. ISSN: 00457825. doi: 10.1016/j.cma.2004.08.006 (cit. on p. 2).
- [130] Anwar A Irmatov. “Arrangements of Hyperplanes and the Number of Threshold Functions”. In: *Acta Applicandae Mathematicae* 68 (2001), pp. 211–226. doi: 10.1023/A:1012087813557 (cit. on p. 53).

- 
- [131] Kazufumi Ito and Karl Kunisch. “On a Semi-Smooth Newton Method and Its Globalization”. In: *Mathematical Programming* 118.2 (May 2009), pp. 347–370. ISSN: 0025-5610, 1436-4646. doi: 10 . 1007 / s10107 - 007 - 0196 - 3 (cit. on pp. 3, 40).
  - [132] Alexey F. Izmailov and 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 (cit. on pp. 30, 34, 57, 60).
  - [133] Michael I. Jordan, Tianyi Lin, and Manolis Zampetakis. *On the Complexity of Deterministic Nonsmooth and Nonconvex Optimization*. Nov. 2022. doi: 10 . 48550 / arXiv . 2209 . 12463. arXiv: 2209 . 12463 [math] (cit. on p. 45).
  - [134] Norman H. Josephy. *Newton’s Method for Generalized Equations*. Tech. rep. ADA077096. wisconsin: University of Madison, 1979, p. 37 (cit. on pp. 15, 29).
  - [135] C. Kanzow. “Nonlinear Complementarity as Unconstrained Optimization”. In: *Journal of Optimization Theory and Applications* 88.1 (Jan. 1996), pp. 139–155. ISSN: 0022-3239, 1573-2878. doi: 10 . 1007 / BF02192026 (cit. on p. 39).
  - [136] Christian Kanzow and Masao Fukushima. “Solving Box Constrained Variational Inequalities by Using the Natural Residual with D-gap Function Globalization”. In: *Operations Research Letters* 23.1-2 (Aug. 1998), pp. 45–51. ISSN: 01676377. doi: 10 . 1016 / S0167 - 6377 (98) 00023 - 6 (cit. on pp. 40, 60).
  - [137] Christian Kanzow and Helmut Kleinmichel. “A New Class of Semismooth Newton-Type Methods for Nonlinear Complementarity Problems”. In: *Computational Optimization and Applications* 11 (1998), pp. 227–251 (cit. on p. 36).
  - [138] Christian Kanzow, Nobuo Yamashita, and Masao Fukushima. “New NCP-Functions and Their Properties”. In: *Journal of Optimization Theory and Applications* 94.1 (July 1997), pp. 115–135. ISSN: 0022-3239. doi: 10 . 1023 / A : 1022659603268 (cit. on pp. 4, 18, 28).
  - [139] Stepan Karamardian. “Generalized Complementarity Problem”. In: *Journal of Optimization Theory and Applications* 8.3 (Sept. 1971), pp. 161–168. ISSN: 0022-3239, 1573-2878. doi: 10 . 1007 / BF00932464 (cit. on p. 2).
  - [140] Lars Kastner and Marta Panizzut. “Hyperplane Arrangements in Polymake”. In: *Mathematical Software – ICMS 2020*. Ed. by Anna Maria Bigatti, Jacques Carette, James H. Davenport, Michael Joswig, and Timo De Wolff. Vol. 12097. Cham: Springer International Publishing, 2020, pp. 232–240. ISBN: 978-3-030-52199-8 978-3-030-52200-1. doi: 10 . 1007 / 978 - 3 - 030 - 52200 - 1 \_ 23 (cit. on pp. 6, 50, 143).
  - [141] Leonid G Khachiyan, Endre Boros, Khaled M. Elbassioni, Vladimir A. Gurvich, and Kazuhisa Makino. “On the Complexity of Some Enumeration Problems for Matroids”. In: *SIAM Journal on Discrete Mathematics* 19.4 (Jan. 2005), pp. 966–984. ISSN: 0895-4801, 1095-7146. doi: 10 . 1137 / S0895480103428338 (cit. on pp. 50, 89, 141, 177).
  - [142] Robert John Kingan and 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, pp. 287–295. ISBN: 0-8218-3761-3 (cit. on p. 50).

- [143] Kolja Knauer, Luis Pedro Montejano, and Jorge Luis Ramírez Alfonsín. “How Many Circuits Determine an Oriented Matroid?” In: *Combinatorica* 38.4 (Aug. 2018), pp. 861–885. ISSN: 0209-9683, 1439-6912. doi: 10.1007/s00493-016-3556-x (cit. on p. 50).
- [144] Masakazu Kojima, Nimrod Megiddo, Toshihito Noma, and 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 (cit. on pp. 3, 14, 15).
- [145] Masakazu Kojima, Shinji Mizuno, and Akiko Yoshise. “A Polynomial-Time Algorithm for a Class of Linear Complementarity Problems”. In: *Mathematical Programming* 44.1-3 (May 1989), pp. 1–26. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF01587074 (cit. on pp. 3, 15).
- [146] Masakazu Kojima and 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 (Dec. 1986), pp. 352–375. doi: 10.15807/jorsj.29.352 (cit. on pp. 26, 57).
- [147] Michael Martin Kostreva. “Direct Algorithms for Complementarity Problems”. PhD thesis. Ann Arbor, MI: Rensselaer Polytechnic Institute, 1976 (cit. on p. 17).
- [148] Lukas Kühne. “The Universality of the Resonance Arrangement and Its Betti Numbers”. In: *Combinatorica* 43.2 (Apr. 2023), pp. 277–298. ISSN: 0209-9683, 1439-6912. doi: 10.1007/s00493-023-00006-x (cit. on pp. 53, 96).
- [149] Adrian Kulmburg and Matthias Althoff. “On the Co-NP-completeness of the Zonotope Containment Problem”. In: *European Journal of Control* 62 (Nov. 2021), pp. 84–91. ISSN: 09473580. doi: 10.1016/j.ejcon.2021.06.028 (cit. on pp. 219, 271, 272, 275).
- [150] Bernd Kummer. “NEWTON’s METHOD FOR NON-DIFFERENTIABLE FUNCTIONS”. In: *Advances in Mathematical Optimization*. Ed. by J. Guddat Et Al. De Gruyter, Dec. 1988, pp. 114–125. ISBN: 978-3-11-247992-6. doi: 10.1515/9783112479926-011 (cit. on p. 24).
- [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 (cit. on pp. 78, 141, 143).
- [152] Carlton E. Lemke. “Bimatrix Equilibrium Points and Mathematical Programming”. In: *Management Science* 11.7 (May 1965), pp. 681–689. ISSN: 0025-1909, 1526-5501. doi: 10.1287/mnsc.11.7.681 (cit. on 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 (July 1944), pp. 164–168. ISSN: 0033-569X, 1552-4485. doi: 10.1090/qam/10666 (cit. on p. 26).
- [154] Li-Zhi Liao, Houduo Qi, and Liqun Qi. “Solving Nonlinear Complementarity Problems with Neural Networks: A Reformulation Method Approach”. In: *Journal of Computational and Applied Mathematics* 131.1-2 (June 2001), pp. 343–359. ISSN: 03770427. doi: 10.1016/S0377-0427(00)00262-4 (cit. on p. 36).

- 
- [155] Zhi-Quan Luo, Olvi Leon Mangasarian, Jun Ren, and Mikhail V. Solodov. “New Error Bounds for the Linear Complementarity Problem”. In: *Mathematics of Operations Research* 19.4 (Nov. 1994), pp. 880–892. ISSN: 0364-765X, 1526-5471. doi: 10.1287/moor.19.4.880 (cit. on p. 38).
- [156] Zhi-Quan Luo and 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, pp. 204–225. ISBN: 0-89871-391-9 (cit. on p. 18).
- [157] Changfeng Ma, Jia Tang, and Xiaohong Chen. “A Globally Convergent Levenberg–Marquardt Method for Solving Nonlinear Complementarity Problem”. In: *Applied Mathematics and Computation* 192 (2007), pp. 370–381 (cit. on pp. 4, 44).
- [158] Mend-Amar Majig and Masao Fukushima. “Restricted-Step Josephy–Newton Method for General Variational Inequalities with Polyhedral Constraints”. In: *Pacific Journal of Optimization* 6 (May 2010), p. 15 (cit. on p. 39).
- [159] Olvi Leon Mangasarian. “Equivalence of the Complementarity Problem to a System of Nonlinear Equations”. In: *SIAM Journal on Applied Mathematics* 31.1 (July 1976), pp. 89–92. ISSN: 0036-1399, 1095-712X. doi: 10.1137/0131009 (cit. on pp. 4, 17, 18).
- [160] Olvi Leon Mangasarian and Robert R. Meyer. “Absolute Value Equations”. In: *Linear Algebra and its Applications* 419.2-3 (Dec. 2006), pp. 359–367. ISSN: 00243795. doi: 10.1016/j.laa.2006.05.004 (cit. on p. 16).
- [161] Olvi Leon Mangasarian and Michael V Solodov. “A Linearly Convergent Derivative-Free Descent Method for Strongly Monotone Complementarity Problems”. In: *Computational Optimization and Applications* 14 (1999), pp. 5–16. ISSN: 0926-6003, 1573-2894. doi: 10.1023/A:1008752626695 (cit. on pp. 38, 39).
- [162] Olvi Leon Mangasarian and Mikhail V. Solodov. “Nonlinear Complementarity as Unconstrained and Constrained Minimization”. In: *Mathematical Programming* 62.1-3 (Feb. 1993), pp. 277–297. ISSN: 0025-5610. doi: 10.1007/BF01585171 (cit. on p. 38).
- [163] Estelle Marchand, Torsten Müller, and 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 (June 2012), pp. 691–708. ISSN: 1420-0597, 1573-1499. doi: 10.1007/s10596-012-9279-1 (cit. on pp. 2, 57).
- [164] Estelle Marchand, Torsten Müller, and 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 (Apr. 2013), pp. 431–442. ISSN: 1420-0597, 1573-1499. doi: 10.1007/s10596-013-9341-7 (cit. on pp. 2, 57).
- [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 (June 1963), pp. 431–441. ISSN: 0368-4245, 2168-3484. doi: 10.1137/0111030 (cit. on p. 26).

- [166] Arnaud Mary and 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 (cit. on pp. 50, 89).
- [167] Peter McMullen. "On Zonotopes". In: *Transactions of the American Mathematical Society* 159 (Sept. 1971), pp. 91–109. doi: 10.2307/1996000 (cit. on pp. 52, 218, 265).
- [168] Nimrod Megiddo. "A Monotone Complementarity Problem with Feasible Solutions but No Complementary Solutions". In: *Mathematical Programming* 12.1 (Dec. 1977), pp. 131–132. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF01593775 (cit. on p. 14).
- [169] Nimrod Megiddo. *A Note on the Complexity of P-matrix LCP and Computing an Equilibrium*. Tech. rep. RJ 6439 (62557). San Jose, CA, USA: Almaden Research Center, 1988, pp. 1–6 (cit. on p. 14).
- [170] Arturo Merino and Torsten Mütze. "Traversing Combinatorial 0/1-Polytopes via Optimization". In: *SIAM Journal on Computing* 53.5 (Oct. 2024), pp. 1257–1292. ISSN: 0097-5397, 1095-7111. doi: 10.1137/23M1612019 (cit. on p. 51).
- [171] Robert Mifflin. "Semismooth and Semiconvex Functions in Constrained Optimization". In: *SIAM Journal on Control and Optimization* 15.6 (Nov. 1977), pp. 959–972. ISSN: 0363-0129, 1095-7138. doi: 10.1137/0315061 (cit. on p. 23).
- [172] Edward Minieka. "Finding the Circuits of a Matroid". In: *JOURNAL OF RESEARCH of the National Bureau of Standards* 80B.3 (1976) (cit. on p. 50).
- [173] George J Minty. "Montone (Nonlinear) Operators in Hilbert Spaces". In: *Duke Mathematical Journal* 29 (1962), pp. 341–346 (cit. on p. 15).
- [174] Shinji Mizuno, Akiko Yoshise, and Takeshi Kikuchi. "PRACTICAL POLYNOMIAL TIME ALGORITHMS FOR LINEAR COMPLEMENTARITY PROBLEMS". In: *Journal of the Operations Research Society of Japan* 32.1 (1989), pp. 75–92. ISSN: 0453-4514, 2188-8299. doi: 10.15807/jorsj.32.75 (cit. on 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 (cit. on p. 24).
- [176] Boris S. Mordukhovich. *Variational Analysis and Generalized Differentiation. I*. Vol. 1. Grundlehren Der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Berlin: Springer-Verlag, 2006. ISBN: 978-3-540-25437-9 3-540-25437-4 (cit. on p. 24).
- [177] Boris S. Mordukhovich. *Variational Analysis and Generalized Differentiation. II*. Vol. 2. Grundlehren Der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Berlin: Springer-Verlag, 2006. ISBN: 978-3-540-25438-6 3-540-25438-2 (cit. on p. 24).
- [178] Theodore S. Motzkin. *Beiträge zur Theorie der linearen Ungleichungen*. Tech. rep. Jerusalem, Israel: University Basel, 1936 (cit. on pp. 50, 144, 145).

- 
- [179] Todd S. Munson, Francisco Facchinei, Michael C. Ferris, Andreas Fischer, and Christian Kanzow. “The Semismooth Algorithm for Large Scale Complementarity Problems”. In: *INFORMS Journal on Computing* 13.4 (Nov. 2001), pp. 294–311. ISSN: 1091-9856, 1526-5528. doi: 10.1287/ijoc.13.4.294.9734 (cit. on pp. 3, 34, 40).
- [180] Katta G. Murty. “Computational Complexity of Complementary Pivot Methods”. In: *Complementarity and Fixed Point Problems* 7 (1978), pp. 61–73. doi: 10.1007/BFb0120782 (cit. on 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 (cit. on pp. 1, 12, 14, 57).
- [182] Katta G. Murty and Santosh N. Kabadi. “Some NP-complete Problems in Quadratic and Nonlinear Programming”. In: *Mathematical Programming* 39 (1987), pp. 117–129. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF02592948 (cit. on p. 45).
- [183] John Lawrence Nazareth and Liqun Qi. “Globalization of Newton’s Method for Solving Nonlinear Equations”. In: *Numerical Linear Algebra with Applications* 3.3 (May 1996), pp. 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 (cit. on p. 26).
- [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 (cit. on p. 45).
- [185] Foundation Inc. OEIS. *The Online Encyclopedia of Integer Sequences*. 2025 (cit. on p. 167).
- [186] Peter Orlik and Louis Solomon. “Combinatorics and Topology of Complements of Hyperplanes”. In: *Inventiones Mathematicae* 56.2 (Feb. 1980), pp. 167–189. ISSN: 0020-9910, 1432-1297. doi: 10.1007/BF01392549 (cit. on pp. 47, 143).
- [187] Peter Orlik and Hiroaki Terao. *Arrangement of Hyperplanes*. Vol. 300. Grundlehren Der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Madison: Springer-Verlag Berlin Heidelberg GmbH, 1992. ISBN: 3-540-55259-6 (cit. on pp. 5, 47, 143, 148, 159).
- [188] James M. Ortega and Werner C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. 2nd ed. Classics in Applied Mathematics 30. Philadelphia, PA, USA: SIAM, 2000. ISBN: 0-89871-461-3 (cit. on p. 24).
- [189] OSCAR. *OSCAR – Open Source Computer Algebra Research System, Version 1.0.0*. The OSCAR Team. 2024 (cit. on pp. 6, 143).
- [190] El Hassene Osmani, Mounir Haddou, Lina Abdallah, and 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, May 2021, pp. 1–8. ISBN: 978-1-6654-4103-2. doi: 10.1109/ICOA51614.2021.9442649 (cit. on pp. 4, 42, 43).
- [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 (cit. on pp. 6, 48, 66, 79, 141, 153).

- [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 (July 1991), pp. 101–131. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF01586928 (cit. on pp. 28, 30–32, 57, 222).
- [193] Jong-Shi Pang. “Complementarity Problems”. In: *Handbook of Global Optimization*. Vol. 2. Nonconvex Optimization and Its Applications. Dordrecht: Kluwer, 1995, pp. 271–338 (cit. on pp. 2, 57).
- [194] Jong-Shi Pang. “Inexact Newton Methods for the Nonlinear Complementarity Problem”. In: *Mathematical Programming* 36.1 (Oct. 1986), pp. 54–71. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF02591989 (cit. on pp. 4, 28).
- [195] Jong-Shi Pang. “Newton’s Method for B-Differentiable Equations”. In: *Mathematics of Operations Research* 15.2 (May 1990), pp. 311–341. ISSN: 0364-765X, 1526-5471. doi: 10.1287/moor.15.2.311 (cit. on pp. 4, 28, 30, 57, 123).
- [196] Jong-Shi Pang and Steven A. Gabriel. “NE/SQP: A Robust Algorithm for the Nonlinear Complementarity Problem”. In: *Mathematical Programming* 60.1-3 (June 1993), pp. 295–337. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF01580617 (cit. on pp. 31–33, 223).
- [197] Jong-Shi Pang, Shih-Ping Han, and Narayan Rangaraj. “Minimization of Locally Lipschitzian Functions”. In: *SIAM Journal on Optimization* 1.1 (Feb. 1991), pp. 57–82. ISSN: 1052-6234, 1095-7189. doi: 10.1137/0801006 (cit. on pp. 4, 16, 21, 31, 32, 44).
- [198] Jong-Shi Pang and Liqun Qi. “Nonsmooth Equations: Motivation and Algorithms”. In: *SIAM Journal on Optimization* 3.3 (Aug. 1993), pp. 443–465. ISSN: 1052-6234, 1095-7189. doi: 10.1137/0803021 (cit. on pp. 27, 60).
- [199] Ji-Ming Peng. “Equivalence of Variational Inequality Problems to Unconstrained Minimization”. In: *Mathematical Programming* 78.3 (Sept. 1997), pp. 347–355. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF02614360 (cit. on p. 40).
- [200] Sandra Pieraccini, Maria Grazia Gasparo, and Aldo Pasquali. “Global Newton-type Methods and Semismooth Reformulations for NCP”. In: *Applied Numerical Mathematics* 44.3 (Feb. 2003), pp. 367–384. ISSN: 01689274. doi: 10.1016/S0168-9274(02)00169-1 (cit. on p. 36).
- [201] Knot Pipatsrisawat and Adnan Darwiche. “On the Power of Clause-Learning SAT Solvers as Resolution Engines”. In: *Artificial Intelligence* 175.2 (Feb. 2011), pp. 512–525. ISSN: 00043702. doi: 10.1016/j.artint.2010.10.002 (cit. on p. 240).
- [202] Elijah Polak and Liqun Qi. “Globally and Superlinearly Convergent Algorithm for Minimizing a Normal Merit Function”. In: *SIAM Journal on Control and Optimization* 36.3 (May 1998), pp. 1005–1019. ISSN: 0363-0129, 1095-7138. doi: 10.1137/S0363012996310245 (cit. on p. 40).
- [203] Alexander Postnikov and Richard P. Stanley. “Deformations of Coxeter Hyperplane Arrangements”. In: *Journal of Combinatorial Theory, Series A* 91.1-2 (Mar. 2000), pp. 544–597. ISSN: 00973165. doi: 10.1006/jcta.2000.3106 (cit. on pp. 48, 53, 96, 193, 257).
- [204] Liqun Qi. “Convergence Analysis of Some Algorithms for Solving Nonsmooth Equations”. In: *Mathematics of Operations Research* 18.1 (1993), pp. 227–244 (cit. on pp. 4, 21, 22, 24, 27, 28, 30, 31, 57, 59, 60, 84).

- 
- [205] Liqun Qi. "Trust Region Algorithms for Solving Nonsmooth Equations". In: *SIAM Journal of Optimization* 5.1 (1995), pp. 219–230 (cit. on pp. 4, 43).
- [206] Liqun Qi and Jie Sun. "A Nonsmooth Version of Newton's Method". In: *Mathematical Programming* 58 (1993), pp. 353–367 (cit. on pp. 23, 27, 30, 57, 118).
- [207] Liqun Qi and Jie Sun. "A Trust Region Algorithm for Minimization of Locally Lipschitzian Functions". In: *Mathematical Programming* 66.1-3 (Aug. 1994), pp. 25–43. ISSN: 0025-5610, 1436-4646. doi: 10.1007/BF01581136 (cit. on pp. 28, 31, 33).
- [208] Miroslav Rada and 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), pp. 455–473. ISSN: 0895-4801, 1095-7146. doi: 10.1137/15M1027930 (cit. on pp. 6, 51, 58, 59, 85–88, 95, 96, 98, 103, 104, 143, 144, 147, 169–171, 193, 195, 196, 199, 250).
- [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 (cit. on pp. 19, 57, 110).
- [210] Manuel Radons and Josué Tonelli-Cueto. "Generalized Perron Roots and Solvability of the Absolute Value Equation". In: *SIAM Journal on Matrix Analysis and Applications* 44.4 (Dec. 2023), pp. 1645–1666. ISSN: 0895-4798, 1095-7162. doi: 10.1137/22M1517184 (cit. on 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 (May 1994), pp. 352–389. ISSN: 0364-765X, 1526-5471. doi: 10.1287/moor.19.2.352 (cit. on p. 15).
- [212] Jörg Rambau. "Symmetric Lexicographic Subset Reverse Search for the Enumeration of Circuits, Cocircuits, and Triangulations up to Symmetry". In: (2023), pp. 1–41 (cit. on pp. 50, 51, 89, 137–139, 143, 177, 199, 245, 252, 253).
- [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), pp. 1038–1058. ISSN: 0179-5376, 1432-0444. doi: 10.1007/s00454-023-00541-w (cit. on p. 50).
- [214] Jörg Rambau. "TOPCOM: TRIANGULATIONS OF POINT CONFIGURATIONS AND ORIENTED MATROIDS". In: *Mathematical Software*. Beijing, China: WORLD SCIENTIFIC, July 2002, pp. 330–340. ISBN: 978-981-238-048-7 978-981-277-717-1. doi: 10.1142/9789812777171\_0035 (cit. on pp. 6, 50, 143, 177, 253).
- [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), pp. 405–422. ISSN: 00246115. doi: 10.1112/plms/s1-19.1.405 (cit. on pp. 5, 47, 73, 78, 143).
- [216] Stephen M. Robinson. "Generalized Equations and Their Solutions, Part II. Applications to Nonlinear Programming". In: *Mathematical Programming Study* (1982), pp. 200–221. ISSN: 0303-3929. doi: 10.1287/88f6b0d7-91b7-4653-a41a-102ab53cf8e3 (cit. on pp. 3, 4, 15).

## BIBLIOGRAPHY

---

- [217] Stephen M. Robinson. “Generalized Equations and Their Solutions. Part I. Basic Theory.” In: *Mathematical Programming Study* 10 (1979), pp. 128–141. ISSN: 0303-3929. doi: 10.1007/bfb0120850 (cit. on pp. 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), pp. 45–66 (cit. on pp. 23, 56).
- [219] Stephen M. Robinson. “Normal Maps Induced by Linear Transformations”. In: *Mathematics of Operations Research* 17.3 (Aug. 1992), pp. 691–714. ISSN: 0364-765X, 1526-5471. doi: 10.1287/moor.17.3.691 (cit. on p. 15).
- [220] Stephen M. Robinson. “Strongly Regular Generalized Equations”. In: *Mathematics of Operations Research* 5.1 (Feb. 1980), pp. 43–62. ISSN: 0364-765X, 1526-5471. doi: 10.1287/moor.5.1.43 (cit. on pp. 3, 15, 30, 223).
- [221] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton 28. Princeton, NJ: Princeton University Press, 1970 (cit. on pp. 14, 57, 71, 72, 152).
- [222] Siegfried M. Rump. “On P-matrices”. In: *Linear Algebra and its Applications* 363 (Apr. 2003), pp. 237–250. doi: 10.1016/S0024-3795(01)00590-0 (cit. on p. 13).
- [223] Sadra Sadraddini and Russ Tedrake. *Linear Encodings for Polytope Containment Problems*. Mar. 2019. arXiv: 1903.05214 [math] (cit. on pp. 218, 219, 271–273, 276).
- [224] Romesh Saigal. *Linear Programming - A Modern Integrated Analysis*. 48. June 1995 (cit. on pp. 181, 182).
- [225] Hans Samelson, Robert M. Thrall, and 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 (cit. on pp. 12, 13).
- [226] Holger Scheel and Stefan Scholtes. “Mathematical Programs with Complementarity Constraints: Stationarity, Optimality, and Sensitivity”. In: *Mathematics of Operations Research* 25.1 (Feb. 2000), pp. 1–22. ISSN: 0364-765X, 1526-5471. doi: 10.1287/moor.25.1.1.15213 (cit. on p. 16).
- [227] Ludwig Schläfli. *Gesammelte mathematische Abhandlungen*. Springer, Basel: Birkhäuser, 1950 (cit. on pp. 5, 47, 79, 143, 165).
- [228] Jürgen Schmidhuber. “Deep Learning in Neural Networks: An Overview”. In: *Neural Networks* 61 (Jan. 2015), pp. 85–117. ISSN: 08936080. doi: 10.1016/j.neunet.2014.09.003 (cit. on pp. 53, 69).
- [229] Paul D. Seymour. “A Note on Hyperplane Generation”. In: *Journal of Combinatorial Theory, Series B* 61 (1994), pp. 88–91. doi: 10.1006/jctb.1994.1033 (cit. on p. 50).
- [230] Alexander Shapiro. “On Concepts of Directional Differentiability”. In: *Journal of Optimization Theory and Applications* 66.3 (Sept. 1990), pp. 477–487. ISSN: 0022-3239, 1573-2878. doi: 10.1007/BF00940933 (cit. on pp. 23, 28, 229).
- [231] Nora Helena Sleumer. “Hyperplane Arrangements: Construction, Visualization and Applications”. PhD thesis. Zurich, Switzerland: Swiss Federal Institute of Technology, 2000 (cit. on pp. 51, 73).

- 
- [232] Nora Helena Sleumer. “Output-Sensitive Cell Enumeration in Hyperplane Arrangements”. In: *Algorithm Theory – SWAT’98*. Ed. by Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen, Stefan Arnborg, and Lars Ivansson. Vol. 1432. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 300–309. ISBN: 978-3-540-64682-2 978-3-540-69106-8. doi: 10.1007/BFb0054377 (cit. on pp. 6, 51, 73, 76, 85, 143, 152).
- [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 (cit. on pp. 27, 57).
- [234] Mikhail V. Solodov and Benar Fux Svaiter. “A New Projection Method for Variational Inequality Problems”. In: *SIAM Journal on Control and Optimization* 37.3 (Jan. 1999), pp. 765–776. ISSN: 0363-0129. doi: 10.1137/S0363012997317475 (cit. on p. 28).
- [235] Mikhail V. Solodov and 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), pp. 605–625. ISSN: 1052-6234. doi: 10.1137/S1052623498337546 (cit. on p. 28).
- [236] Richard P. Stanley. “An Introduction to Hyperplane Arrangements”. In: *Geometric Combinatorics*. 1st ed. Vol. 13. IAS/Park City Math. Ser. Providence, Rhode Island: Amer. Math. Soc., 2007, pp. 389–496. ISBN: 978-0-8218-3736-8 0-8218-3736-2 (cit. on pp. 5, 47, 73, 78, 143).
- [237] Richard P. Stanley. *Enumerative Combinatorics*. second. Vol. 1. Cambridge Studies in Advanced Mathematics. Cambridge, UK: Cambridge University Press, 2012 (cit. on pp. 5, 47, 148, 167).
- [238] Richard P. Stanley. *Enumerative Combinatorics*. second. Vol. 2. Cambridge Studies in Advanced Mathematics. Cambridge, UK: Cambridge University Press, 2024. ISBN: 978-1-009-26249-1 978-1-009-26248-4 (cit. on pp. 5, 47, 143).
- [239] Jakob Steiner. “Einige Gesetze über die Theilung der Ebene und des Raumes.” In: *J. Reine Angew. Math* (1826), pp. 349–364 (cit. on pp. 5, 47, 73, 78, 143).
- [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), pp. 525–538. ISSN: 0022-3239, 1573-2878. doi: 10.1007/BF00939395 (cit. on p. 14).
- [241] Pudukkottai K. Subramanian. “Gauss-Newton Methods for the Complementarity Problem”. In: *Journal of Optimization Theory and Applications* 77.3 (June 1993), pp. 467–482. ISSN: 0022-3239, 1573-2878. doi: 10.1007/BF00940445 (cit. on p. 18).
- [242] Defeng Sun, Masao Fukushima, and 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), pp. 452–473 (cit. on p. 40).
- [243] Defeng Sun and Liqun Qi. “On NCP-functions”. In: *Computational Optimization and Applications* 13 (1999), pp. 201–220. doi: 10.1023/A:1008669226453 (cit. on pp. 13, 15, 28, 37, 42).
- [244] Panjie Tian, Zhensheng Yu, and Yue Yuan. “A Smoothing Levenberg-Marquardt Algorithm for Linear Weighted Complementarity Problem”. In: *AIMS Mathematics* 8.4 (2023), pp. 9862–9876. ISSN: 2473-6988. doi: 10.3934/math.2023498 (cit. on p. 43).

- [245] Paul Tseng. “Co-NP-completeness of Some Matrix Classification Problems”. In: *Mathematical Programming* 88.1 (June 2000), pp. 183–192. ISSN: 0025-5610, 1436-4646. doi: 10.1007/s101070000159 (cit. on p. 13).
- [246] Paul Tseng, Nobuo Yamashita, and Masao Fukushima. “EQUIVALENCE OF COMPLEMENTARITY PROBLEMS TO DIFFERENTIABLE MINIMIZATION: A UNIFIED APPROACH”. In: *SIAM Journal of Optimization* 6.2 (May 1996), pp. 446–460 (cit. on p. 39).
- [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 (cit. on p. 3).
- [248] Duc Thach Son Vu. “Numerical Resolution of Algebraic Systems with Complementarity Conditions. Application to the Thermodynamics of Compositional Multiphase Mixtures”. PhD thesis. Université Paris-Saclay, 2020 (cit. on p. 42).
- [249] Duc Thach Son Vu, Ibtihel Ben Ghabria, Mounir Haddou, and 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 (Dec. 2021), pp. 1243–1274. ISSN: 03784754. doi: 10.1016/j.matcom.2021.07.015 (cit. on pp. 2, 4, 42).
- [250] Dominic J. A. Welsh. *Complexity: Knots, Colourings and Counting*. Vol. 186. University of Oxford: Cambridge University Press, Aug. 1993. ISBN: 978-0-521-45740-8 (cit. on p. 51).
- [251] Walter Wenzel, Nihat Ay, and Frank Pasemann. “Hyperplane Arrangements Separating Arbitrary Vertex Classes in N-Cubes”. In: *Advances in Applied Mathematics* 25.3 (Oct. 2000), pp. 284–306. ISSN: 01968858. doi: 10.1006/aama.2000.0701 (cit. on pp. 53, 69, 96).
- [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 (July 1982), pp. 401–405. ISSN: 0022-3239, 1573-2878. doi: 10.1007/BF00935279 (cit. on p. 17).
- [253] Robert Owen Winder. “Partitions of N-Space by Hyperplanes”. In: *SIAM Journal on Applied Mathematics* 14.4 (July 1966), pp. 811–818. ISSN: 0036-1399, 1095-712X. doi: 10.1137/0114068 (cit. on pp. 48, 53, 61, 78–80, 143, 164, 165).
- [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 (cit. on p. 15).
- [255] Shuhuang Xiang and Xiaojun Chen. “Computation of Generalized Differentials in Nonlinear Complementarity Problems”. In: *Computational Optimization and Applications* 50.2 (Oct. 2011), pp. 403–423. ISSN: 0926-6003, 1573-2894. doi: 10.1007/s10589-010-9349-z (cit. on pp. 31, 59, 61, 75, 83, 84, 120, 145).
- [256] Nobuo Yamashita and Masao Fukushima. “On Stationary Points of the Implicit Lagrangian for Nonlinear Complementarity Problems”. In: *Journal of Optimization Theory and Applications* 84.3 (Mar. 1995), pp. 653–663. ISSN: 0022-3239, 1573-2878. doi: 10.1007/BF02191990 (cit. on p. 38).
- [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 (?) (Cit. on pp. 5, 48, 69, 78, 129, 143, 148, 164, 167).

- 
- [258] Chao Zhang, Xiaojun Chen, and Naihua Xiu. “Global Error Bounds for the Extended Vertical LCP”. In: *Computational Optimization and Applications* 42.3 (Apr. 2009), pp. 335–352. ISSN: 0926-6003, 1573-2894. doi: 10.1007/s10589-007-9134-9 (cit. on p. 57).
  - [259] Ju-liang Zhang and Jian Chen. “A Smoothing Levenberg–Marquardt Type Method for LCP”. In: *Journal of Computational Mathematics* (2004), pp. 735–752 (cit. on p. 43).
  - [260] Ju-liang Zhang and Xiangsun Zhang. “A Smoothing Levenberg–Marquardt Method for NCP”. In: *Applied Mathematics and Computation* 178.2 (July 2006), pp. 212–228. ISSN: 00963003. doi: 10.1016/j.amc.2005.11.036 (cit. on pp. 4, 43).
  - [261] Shuzi Zhou and Zhanyong Zou. “A New Iterative Method for Discrete HJB Equations”. In: *Numerische Mathematik* 111.1 (Nov. 2008), pp. 159–167. ISSN: 0029-599X, 0945-3245. doi: 10.1007/s00211-008-0166-6 (cit. on p. 2).
  - [262] Günter M. Ziegler. *Lectures on 0/1-Polytopes*. Sept. 1999. arXiv: math/9909177 (cit. on p. 284).
  - [263] Günter M. Ziegler. *Lectures on Polytopes*. 7th. Vol. 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 (cit. on pp. 49, 52, 153, 218, 265).
  - [264] Günter M. Ziegler, Laura Anderson, and Kolja Knauer. “Oriented Matroids Today”. In: *The Electronic Journal of Combinatorics* 1000 (2024), p. 73. ISSN: 1077-8926. doi: 10.37236/25 (cit. on p. 48).