Résolution de problèmes de complémentarité. : Application à un écoulement diphasique dans un milieu poreux

Ibtihel Ben Gharbia

UNIVERSITÉ PARIS DAUPHINE
CEREMADE
INRIA PARIS-ROCQUENCOURT
POMDAPI

N° attribué par la bibliothèque

THÈSE

présentée pour l'obtention du titre de
DOCTEUR EN SCIENCES
(Arrêté du 7 août 2006)

SPÉCIALITÉ : MATHÉMATIQUES APPLIQUÉES

Résolution de problèmes de complémentarité
Application à un écoulement diphasique dans un milieu poreux

Soutenue par : Ibtihel BEN GHARBIA

JURY

Directeurs de thèse : Jean Charles GILBERT
Directeur de Recherche à Inria Paris-Rocquencourt

Jérôme JAFFRÉ
Directeur de Recherche à Inria Paris-Rocquencourt

Rapporteurs : Mounir HADDOU
Professeur à l'Institut National des Sciences Appliquées de Rennes.

Olivier PIRONNEAU
Professeur à l'université Pierre et Marie Curie

Examinateur : Guy CHAVENT
Professeur émérite à l'université Paris Dauphine

Présentée et soutenue publiquement le 05/12/2012

Table des matières

À ma mère, mon père,
mes deux sœurs.

À toute ma famille.

Remerciements

Ce travail n'aurait jamais vu le jour sans le soutien, l'encouragement et le savoir de mes directeurs de thèse : Monsieur Jean Charles GILBERT, Directeur de Recherche à l'Inria Paris-Rocquencourt, et Monsieur Jérôme JAFFRÉ, Directeur de Recherche à l'Inria Paris-Rocquencourt et chef du projet Pomdapi, dont j'ai bénéficié pendant toutes mes années de recherche. Ils se sont montrés toujours disponibles et m'ont constamment aidée à y voir plus clair. Ils ont fait preuve à la fois d'une grande patience, de gentillesse, et d'un esprit responsable, critique et rigoureux. Je n'oublierai jamais leur sympathie et l'amitié que j'ai trouvé en eux. Qu'ils trouvent ici l'expression de ma profonde gratitude.

Mes sincères reconnaissances vont aux personnes qui m'ont fait l'honneur d'accepter d'être rapporteurs de ma thèse, et ainsi d'évaluer mon travail : Monsieur Mounir HADDOU, Professeur à l'Institut National des Sciences Appliquées (Rennes), et Monsieur Olivier PIRONNEAU, Professeur à l'Université Pierre et Marie Curie (Paris). Je remercie également Monsieur Guy CHAVENT, Professeur à l'Université Paris Dauphine, pour avoir accepté d'honorer ma soutenance de thèse de sa présence.

Ce travail a été soutenu par le GNR MoMas (CNRS, ANDRA, BRGM, CEA, EdF, IRSN) dont je voudrais saluer le dynamisme de ses membres, en particulier Monsieur Alain BOURGEAT et Madame Sylvie GRANET.

Mes plus vifs remerciements vont à Madame Jean E. ROBERTS, Directrice de Recherche à l'Inria Paris-Rocquencourt, pour son suivi précieux, ses qualités scientifiques et humaines qui ont largement contribué à l'aboutissement de ma thèse. J'exprime aussi ma reconnaissance à Michel KERN, à Caroline JAPHET, à Franǫis CLÉMENT et à Martin VOHRALIK pour leur soutien et leur gentillesse.

Je rends grâce à l'efficacité de l'assistante Nathalie BONTE avec laquelle j'ai eu l'occasion de travailler, et qui m'a rendu de nombreux et précieux services. Durant ces années, j'ai trouvé un environnement scientifique et humain d'une rare qualité au sein du projet POMDAPI. Merci à tous les membres de ce projet avec qui j'ai eu un très grand plaisir de travailler, plus particulièreme Phuong HOANG THI THAO, Mohamed RIAHI, Fatma CHEIKH et Ilyess AHMED. Merci aussi à mon ancienne collègue de bureau Alice CHICHE pour m'avoir épaulée et encouragée.

Mes remerciements s'adressent aussi à mon ancienne professeure et mon amie Hend BEN AMEUR pour ses encouragements continus, pour son soutien et pour les bons moments que nous avons passés ensemble.

Je remercie également Madame Yomna REBAI, mon ancienne professeure, de m'avoir encouragée et aidée à poursuivre mes études en France et de m'avoir toujours épaulée.

Merci à tous les gens que j'ai eu le plaisir de côtoyer durant ces années, à tous mes amis pour leur soutien moral et leur aide, plus particulièrement à Amel HAMZAOUI et Ines AYADI.

Je tiens aussi à remercier vivement ma chère amie Raouia TAKTAK, pour tout ce qu'elle a fait pour moi, pour le soutien qu'elle m'a toujours apporté, pour son écoute dans les moments les plus difficiles et pour les bons moments que l'on a passés ensemble.

Je suis également reconnaissante envers ma cousine Yosra et sa famille pour leur bienveillance et leur soutien durant mon séjour en France.

Mes plus profonds remerciements et reconnaissances vont à mes parents, à mes soeurs et à toute ma famille, à qui je dois tout. Tout au long de mon cursus, ils m'ont toujours soutenue, encouragée et aidée. Ils ont su me donner toutes les chances pour réussir.

Résumé

Les problèmes de complémentarité interviennent dans de nombreux domaines scientifiques : économie, mécanique des solides, mécanique des fluides. Ce n'est que récemment qu'ils ont commencé d'intéresser les chercheurs étudiant les écoulements et le transport en milieu poreux. Les problèmes de complémentarité sont un cas particulier des inéquations variationnelles. Dans cette thèse, on offre plusieurs contributions aux méthodes numériques pour résoudre les problèmes de complémentarité.

Dans la première partie de la thèse, on étudie les problèmes de complémentarité linéaires $0 \leqslant x \perp (Mx+q) \geqslant 0$ où l'inconnue x est dans $\mathbb{R}^n$ et où les données sont un vecteur q de $\mathbb{R}^n$ et une matrice M d'ordre n. L'existence et l'unicité de ce problème, quel que soit q, est obtenue si, et seulement si, la matrice M est une **P**-matrice. Une méthode très efficace pour résoudre les problèmes de complèmentarité est la méthode de Newton-min, une extension de la méthode de Newton aux problèmes non lisses.

Dans cette thèse, on montre d'abord, en construisant deux familles de contre-exemples, que la méthode de Newton-min ne converge pas pour la classe des **P**-matrices, sauf si $n = 1$ ou 2. Ensuite, on caractérise algorithmiquement la classe des **P**-matrices : c'est la classe des matrices qui sont telles que, quel que soit le vecteur q, l'algorithme de Newton-min ne fait pas de cycle de deux points. Enfin ces résultats de non-convergence nous ont conduit à construire une méthode de globalisation de l'algorithme de Newton-min dont nous avons démontré la convergence globale pour les **P**-matrices. Des résultats numériques montrent l'efficacité de cet algorithme et sa convergence polynomiale pour les cas considérés.

Dans la deuxième partie de cette thèse, nous nous sommes intéressés à un exemple de problème de complémentarité non linéraire concernant les écoulements en milieu poreux. Il s'agit d'un écoulement liquide-gaz à deux composants eau-hydrogène que l'on rencontre dans le cadre de l'étude du stockage des déchets radioactifs en milieu géologique. Nous présentons un modèle mathématique utilisant des conditions de complémentarité non linéaires décrivant ces écoulements. D'une part, nous proposons une méthode de résolution et un solveur pour ce problème. D'autre part, nous présentons les résultats numériques que nous avons obtenus suite à la simulation des cas-tests proposés par l'ANDRA (Agence Nationale pour la gestion des Déchets Radioactifs) et le GNR MoMaS. En particulier, ces résultats montrent l'efficacité de l'algorithme proposé et sa convergence quadratique pour ces cas-tests.

Mots-clés

Algorithme de Newton-min - Analyse non lisse - Caractérisation de la **P**-matricité - Convergence globale - Convergence locale - Convergence quadratique - Dissolution - Écoulement diphasique - Méthode de Newton non lisse - Milieu poreux - **M**-matrice - **NM**-matrice - **P**-matrice - Problème de complémentarité linéaire - Problème de complémentarité non linéaire - Stockage profond de déchets nucléaires.

Abstract

This manuscript deals with numerical methods for linear and nonlinear complementarity problems, and, more specifically, with solving gas phase appearance and disappearance modeled as a complementarity problem.

In the first part of this manuscript, we focused on the plain Newton-min method to solve the linear complementarity problem (LCP for short) $0 \leqslant x \perp (Mx+q) \geqslant 0$ that can be viewed as a nonsmooth Newton algorithm without globalization technique to solve the system of piecewise linear equations $\min(x, Mx+q) = 0$, which is equivalent to the LCP. When M is an **M**-matrix of order n, the algorithm was known to converge in at most n iterations. We show that this result no longer holds when M is a **P**-matrix of order $\geqslant 3$. On the one hand, we offer counter-examples showing that the algorithm may cycle in those cases. **P**-matrices are interesting since they are those ensuring the existence and uniqueness of the solution to the LCP for an arbitrary q. Incidentally, convergence occurs for a **P**-matrix of order 1 or 2. On the other hand, we provide a new algorithmic characterization of **P**-matricity : we show that a nondegenerate square real matrix M is a **P**-matrix if and only if, whatever is the real vector q, the Newton-min algorithm does not cycle between two points. In order to force the convergence of the Newton-min algorithm with **P**-matrices, we have derived a new method, which is robust, easy to describe, and simple to implement. It is globally convergent and the numerical results reported in this manuscript show that it outperforms a method of Harker and Pang.

In the second part of this manuscript, we consider the modeling of migration of hydrogen produced by the corrosion of the nuclear waste packages in an underground storage including the dissolution of hydrogen. It results in a set of nonlinear partial differential equations with nonlinear complementarity constraints. We show how to apply a robust and efficient solution strategy, the Newton-min method considered for LCP in the first part, to this geoscience problem and investigates its applicability and efficiency on this difficult problem. The practical interest of this solution technique is corroborated by numerical experiments from the Couplex Gas benchmark proposed by Andra and GNR MoMas. In particular, numerical results show that the Newton-min method is quadratically convergent for these problems.

Keywords

Dissolution - Global convergence - Linear complementarity problem - Local convergence - Nonsmooth function - **M**-matrix - Newton-min algorithm - **NM**-matrix - Nonlinear complementarity problem - Nonsmooth analysis - Nonsmooth Newton method - Nuclear waste underground storage - **P**-matricity characterization - **P**-matrix - Porous media - Quadratic convergence - Two-phase flow.

Chapitre 1

Introduction

1.1 The complementarity problem

A complementarity problem, abbreviated CP, in finite dimension is a system of finitely many inequalities in finitely many nonnegative variables along with a special equation that expresses the complementary relationship between the variables and corresponding inequalities. This complementarity condition is the key feature distinguishing the complementarity problem from a general inequality system. It lies at the heart of all constrained optimization problems in finite dimensions and provides a powerful framework for the modeling of equilibria of many kinds.

The systematic study of the complementarity problem began in the mid-1960s. In a span of five decades, the subject has developed into a very fruitful discipline in the field of mathematical optimization. The developments include a rich mathematical theory, a host of effective solution algorithms, a multitude of interesting connections to numerous disciplines, and a wide range of important applications in geoscience, engineering, and economics. The literature of the complementarity problems has benefited from contributions made by mathematicians (pure, applied, and computational), computer scientists, engineers of many kinds (civil, chemical, electrical, mechanical, and systems), and economists of diverse expertise (agricultural, computational, energy, financial, and spatial) [41, 16, 60]

In addition to optimization problems, many equilibrium problems from economics and important applied problems from diverse engineering fields can be profitably formu-

lated as CPs. We can quote for example : bimatrix games [86, 87], economic equilibrium problems [41, 40], free boundary problems [6], black oil model [22], phase appearance-disappearance problems [65, 13, 84] and dissolution-precipitation problems [19, 20].

The finite-dimensional variational inequality is a generalization of the complementarity problem [53, 77, 38, 76].

Mathematically, the complementarity problem consists in finding a vector $x \geqslant 0$ with n components such that $F(x) \geqslant 0$ and $x^\top F(x) = 0$. Here F is a function from $\mathbb{R}^n$ into itself, the inequalities have to be understood componentwise, and the sign $^\top$ denotes matrix transposition. The CP is often written in compact form as follows

$$0 \leqslant x \perp F(x) \geqslant 0. \tag{1.1.1}$$

Since the components of x and $F(x)$ must be nonnegative in (1.1.1), their perpendicularity with respect to the Euclidean scalar product required in the problem is equivalent to the nullity of the Hadamard product of the two vectors, that is

$$x \cdot F(x) = 0, \tag{1.1.2}$$

which mean that $x_i(F(x))_i$ must vanish for all index i.

There exist many different approaches to treat these complementarity problems, e.g., interior point methods [7, 8, 126] and active-set strategy [70]. Another approach is to rewrite the complementarity conditions as a system of smooth equations [110, 112, 111], as a set of nonsmooth equations [101, 73, 37] and also as constrained and unconstrained optimization problems [46, 68]. These formulations provide the basis for the development of the theory and algorithms for complementarity problems. For an overview of most of these methods, we refer to [38].

In this thesis, we are interested in the nonsmooth reformulation of the complementarity problem using the C-function minimum (see section 2.4.2). Due to the lack of differentiability, the assumptions for the use of classical Newton methods [35, 52] are not satisfied, but the so-called nonsmooth Newton method [26, 113] can be applied. Our interest in this solution approach was motivated by its efficiency.

A complementarily problem (1.1.1) with an affine defining function F is called a linear complementarity problem, abbreviated as LCP.

The linear complementarity problem

The linear complementarity problem refers to an inequality system with a rich mathematical theory. One particularly and well known context in which linear complementarity problems are found is the first-order optimality conditions of quadratic optimization problem [52].

In order to get a better insight into complementarity problems, we began our thesis's work by studying the linear problem.

We fix the notation by precising that the LCP consists in finding a vector $x \geqslant 0$ with n components such that $Mx+q \geqslant 0$ and $x^\top(Mx+q) = 0$. Here M is a real matrix of order n and q is a vector in $\mathbb{R}^n$.

Many algorithms have been proposed to solve problem LCP [103, 30]. They may be based on pivoting techniques [86, 29], which often suffer from the combinatorial aspect of the problem (i.e., the 2^n possibilities to realize $x \cdot (Mx+q) = 0$), on interior point methods, which originate from an algorithm introduced by Karmarkar in linear optimization [74, 1984] (see also [78, 1991] for one of the first accounts on the use of interior point methods to solve linear complementarity problems), on nonsmooth Newton approaches [38], such as the one considered here, and on the regularization of nonsmooth equations [92]. See [103, 30] for other iterative methods.

The linear complementarity problem is a special case of the nonlinear complementarity problem, abbreviated NCP. In the next section, we present a nonlinear complementarity problem encountered in liquid–gas flow in porous medium with two components hydrogen and water.

A nonlinear complementarity problem encountered in diphasic flow in porous media

A large community of scientists is concerned with the modeling of the mechanical and hydraulic behavior of a deep repository of radioactive waste, in order to understand its impact on the environment and human safety. This implies to be able to model and simulate complex phenomena such as the desaturation and resaturation of the geological medium and the migration of gas produced by the corrosion of nuclear waste packages contained in the storage, within complex heterogeneous and anisotropic domains including singular zones such as galleries and cell intersections. Moreover, materials with highly contrasted physical properties are involved and the time scale can be very long (from a thousand to

hundreds of thousands years).

Hence the simulation of these physical features happens to be a complex task, and their validation is a major concern to assess safety of the repository. Computational benchmarks, such as the Couplex Gaz benchmark [50], are useful for the definition of relevant physical models and numerical methods. Indeed, the Couplex Gaz benchmark has shown that the Darcy flow of two phases, the first one being an incompressible liquid phase and the second one the gaseous phase with two components (hydrogen-water), can lead to the appearance and disappearance of the gas phase, leading to the degeneracy of the equations satisfied by the saturation.

In order to overcome this difficulty, different choices for unknowns are available in the literature. An approach presented in [1], is to extend the notion of phase saturation, allowing for negative values, and values greater than one. In another method, developed in [127], primary variables are changed, depending on the value of the saturation : they are switched when, during a time step, the gas phase appears in the saturated zone. These methods remain sensitive to capillary pressure singularities.

In this thesis, we consider a system of nonlinear partial equations – conservation equations and Darcy laws – with nonlinear complementarity conditions describing the transfer of hydrogen between the two phases [65]. The advantage of this formulation is its validity whether the gas phase exists or not. Furthermore, variables do not have to be switched.

1.2 Thesis outline and contributions

Divided into two parts, the thesis contains nine chapters, followed by a bibliography and a subject index. The first part covers chapter two through chapter five, which give contributions to the analysis of numerical methods to solve linear complementarity problem. The second part consists of the remaining four chapters and studies an hydrogen-water flow in a porous medium where hydrogen can be present as a gas phase as well as being dissolved in liquid water. This is modeled as a nonlinear complementarity problem.

Chapter 2 serves two purposes : one, it introduces the important matrix classes in the study of LCP, (several of these classes are of interest because they characterize certain properties of LCPs) ; and two, it contains the background theory of nonsmooth analysis and algorithms.

In chapter 3, we focus on the plain Newton-min algorithm to solve the LCP, which

can be viewed as a nonsmooth Newton algorithm without globalization technique. When M is an **M**-matrix of order n, the algorithm is known to converge in at most n iterations. We show in this chapter that this result no longer holds when M is a **P**-matrix of order $\geqslant 3$, since then the algorithm may cycle. Incidentally, convergence occurs for a **P**-matrix of order 1 or 2. The content of chapter 3 is published in [10].

We provide, in chapter 4, a new algorithmic characterization of **P**-matricity. We show that a nondegenerate square real matrix M is a **P**-matrix if and only if, whatever is the real vector q, the Newton-min algorithm does not cycle between two points when it is used to solve the LCP. This characterization of **P**-matrices is proved in [11], a report under review for publication.

Since the plain Newton-min may cycle for **P**-matrices and since **P**-matrices are those ensuring the existence and uniqueness of the solution to the LCP for an arbitrary q, we have derived, in chapter 5, a new method for solving LCP for this matrix class. We proved that it is globally convergent for **P**-matrices and the numerical results reported in this chapter show that it outperforms a method proposed by Harker and Pang in 1991, which has some connections with our approach. This work is the object of a paper in preparation.

In the second part, we address one of the outstanding physical and mathematical problems in multiphase flow simulation : the appearance and the disappearance of the gas phase, leading to the degeneracy of the equations satisfied by the saturation.

We present, in chapter 6, a model for the migration, in a porous medium, of the hydrogen produced by the corrosion of nuclear waste packages in an underground storage including the dissolution of hydrogen.

Chapter 7 focuses on the nonlinear complementarity problem and its equivalent formulation using variational inequalities.

In the last two chapters (chapters 8 and 9), we present the discretization method and we show how to apply a robust and efficient solution strategy, the Newton-min method, to this geoscience problem. Finally, we show the efficiency and the accuracy of the formulation and of the Newton-min algorithm, on typical test cases used in order to simulate the storage of nuclear waste. In particular, numerical experiments show that the Newton-min method is quadratically convergent for these problems. The content of chapter 9 is accepted for publication in [13].

Première partie

Problème de complémentarité linéaire

Cette partie comporte les chapitres 2 à 5. Dans le chapitre 3 (qui reprend l'article publié [10]), nous montrons que l'algorithme de Newton-min peut cycler lorsque M est une **P**-matrice d'ordre $n \geqslant 3$. C'est très dommage car la **P**-matricité de M est la condition nécessaire et suffisante d'existence et d'unicité d'un problème de complémentarité linéaire. Nous avons toutefois démontré sa convergence pour une **P**-matrice d'ordre 1 ou 2. Dans le chapitre 4 (qui est l'article soumis [11]), nous présentons une nouvelle caractérisation algorithmique de la **P**-matricité. Plus précisément, nous montrons qu'une matrice M est une **P**-matrice si, et seulement si, quel que soit le vecteur q, l'algorithme de Newton-min ne fait pas de cycle de deux points lorsqu'il est utilisé pour résoudre le problème de complémentarité linéaire (2.1.3). Nous proposons dans le chapitre 5 un algorithme globalement convergent pour résoudre le problème de complémentarité linéaire (2.1.3) pour les **P**-matrices. Des tests numériques ont été réalisés montrant l'efficacité de cet algorithme et sa convergence polynomiale pour ces cas.

Chapitre 2

Cadre mathématique

2.1 Préliminaires

En mathématiques, et plus spécialement en recherche opérationnelle et en optimisation, un problème de complémentarité linéaire (PCL) est défini par la donnée d'une matrice $M \in \mathbb{R}^{n\times n}$ et d'un vecteur $q \in \mathbb{R}^n$. Il consiste à trouver un vecteur $x \in \mathbb{R}^n$ tel que ses composantes et celles de $z := Mx+q$ soient positives et tel que x et z soient orthogonaux pour le produit scalaire euclidien de $\mathbb{R}^n$:

$$x \geqslant 0, \qquad Mx+q \geqslant 0 \qquad \text{et} \qquad x^\top(Mx+q) = 0, \tag{2.1.1}$$

où $x^\top$ désigne le vecteur x transposé et la notation $u \geqslant 0$ signifie que toutes les composantes u_i du vecteur sont positives. Dès lors, l'orthogonalité requise de x et $Mx+q$ revient à demander que le produit de Hadamard de ces deux vecteurs soit nuls :

$$x \cdot (Mx+q) = 0. \tag{2.1.2}$$

On rappelle que le *produit de Hadamard* [63, 1991, page 73] de deux vecteurs $x, y \in \mathbb{R}^n$ est le vecteur $x \cdot y \in \mathbb{R}^n$ défini par

$$\forall i : \qquad (x \cdot y)_i = x_i y_i.$$

On écrit souvent le problème (2.1.1) de manière concise comme suit :

$$\mathrm{CL}(M,q) : \qquad 0 \leqslant x \perp (Mx+q) \geqslant 0. \tag{2.1.3}$$

Un PCL est dit linéaire parce que x intervient de manière linéaire dans les termes de gauche et de droite de (2.1.3) mais en réalité, c'est un problème non linéaire à cause de la relation (2.1.2) entre x et $Mx+q$. Ainsi, il n'y a pas de relation linéaire entre q et les solutions éventuelles du PCL. On retrouve donc le même abus de langage qu'en optimisation linéaire, problème auquel le PCL est apparenté (section 2.4.1).

Ces problèmes sont souvent NP ardus et donc difficiles à résoudre lorsque la dimension n du problème devient grande. La combinatoire du problème vient du fait qu'il faut déterminer quelles sont les composantes de la solution qui sont nulles et il y a alors 2^n possibilités de réaliser cela. On note

$$\mathrm{Sol}(M,q)$$

l'ensemble des solutions de $\mathrm{CL}(M,q)$. C'est une réunion de polyèdres convexes [66, 1983]. Un point x vérifiant $x \geqslant 0$ et $Mx+q \geqslant 0$ est dit *admissible* pour le problème $\mathrm{CL}(M,q)$ et l'ensemble

$$\mathrm{Adm}(M,q) := \{x \in \mathbb{R}^n : x \geqslant 0, \quad Mx+q \geqslant 0\}$$

est appelé *l'ensemble admissible* de ce problème. On dit que le problème est *réalisable* si $\mathrm{Adm}(M,q) \neq \varnothing$.

Les problèmes de complémentarité se sont d'abord manifestés dans les conditions d'optimalité des problèmes d'optimisation, les conditions de Karush, Kuhn et Tucker [52]. Elles permettent de modéliser des problèmes décrits par plusieurs systèmes d'équations qui sont en quelque sorte en compétition : celui qui est actif en un endroit et en un temps donnés, correspondant à un indice commun de x et de z, dépend de seuils qui sont ou non atteints : si le seuil $x_i = 0$ n'est pas atteint, c'est-à-dire que $x_i > 0$, l'équation $(Mx+q)_i = 0$ est active. Les propriétés des problèmes de complémentarité linéaire dépendent de celles de la matrice M. Ces propriétés sont très variées et ne dépendent pas d'un seul type de matrices. La situation est donc plus complexe et très différente de celle rencontrée dans la résolution d'un système d'équations linéaires, dont les propriétés dépendent pour beaucoup de l'inversibilité de la matrice définissant le système. Nous présentons, dans la section suivante, quelques grandes classes de matrices et les propriétés de $\mathrm{CL}(M,q)$ qui y sont associées.

2.2 Classes de matrices

Nous présentons ici les définitions et les propriétés de quelques classes de matrices, qui interviennent de manière essentielle dans ce manuscrit ou de manière anecdotique au cours des discussions. Le lecteur pressé pourra passer cette section technique en première lecture et y revenir en cas de besoin ; nous conseillons toutefois de lire la section 2.2.5 sur les **P**-matrices qui joueront un rôle important dans les chapitres 3, 4 et 5. Le tableau de la figure 2.2.1 pourra aussi donner une première idée du degré de raffinement de la théorie de la complémentarité. Dans la suite, on note $[\![1,N]\!] := \{1,\dots,N\}$ l'ensemble des N premiers entiers non nuls.

2.2.1 Classe des matrices non dégénérées

Définition 2.1 On dit qu'une matrice $M \in \mathbb{R}^{n\times n}$ est *non dégénérée* si elle vérifie l'une des conditions équivalentes de la proposition 2.2 [124, 2000, section 4]. On note **ND** l'*ensemble des matrices non-dégénérées*. □

Voici des propriétés caractérisant les matrices non-dégénérées.

Proposition 2.2 (matrice non-dégénérée) *Soit $M \in \mathbb{R}^{n\times n}$. Les propriétés suivantes sont équivalentes :*

(i) $\forall I \subset [\![1,n]\!]$, M_{II} *est inversible,*

(ii) *tout x vérifiant $x\cdot(Mx) = 0$ est nul.*

2.2.2 Classe des S-matrices

Définition 2.3 On dit qu'une matrice $M \in \mathbb{R}^{n\times n}$ est une **S**-*matrice* (d'après Stiemke) si elle vérifie l'une des conditions équivalentes de la proposition ci-dessous, c'est-à-dire si, quel que soit $q \in \mathbb{R}^n$, le problème de complémentarité linéaire (2.1.3) est réalisable. On note **S** l'*ensemble des* **S**-*matrices*. □

Voici des propriétés caractérisant la **S**-matricité.

Proposition 2.4 (S-matricité) *Soit $M \in \mathbb{R}^{n\times n}$. Les propriétés suivantes sont équivalentes :*

(i) $\forall q \in \mathbb{R}^n$, *le problème* $\mathrm{CL}(M,q)$ *est réalisable,*
(ii) *il existe un* $x \geqslant 0$ *tel que* $Mx > 0$,
(iii) *il existe un* $x > 0$ *tel que* $Mx > 0$.

2.2.3 Classe des Q-matrices

Pour $M \in \mathbb{R}^{n\times n}$, on note

$$Q_R(M) := \{q \in \mathbb{R}^n : \mathrm{CL}(M,q) \text{ est réalisable}\}, \tag{2.2.1}$$
$$Q_S(M) := \{q \in \mathbb{R}^n : \mathrm{CL}(M,q) \text{ a une solution}\}. \tag{2.2.2}$$

Ce sont deux cônes. En effet, d'une part, si $x \in \mathrm{Adm}(M,q)$ et $t > 0$, alors $tx \in \mathrm{Adm}(M,tq)$. D'autre part, si $\bar{x} \in \mathrm{Sol}(M,q)$ et $t > 0$, alors $t\bar{x} \in \mathrm{Sol}(M,tq)$. Évidemment, on a

$$Q_S(M) \subset Q_R(M), \tag{2.2.3}$$

sans que l'on ait nécessairement égalité (c'est le sujet de la proposition 2.6 ci-dessous). Avant d'énoncer la proposition, nous introduisons la classe des $\mathbf{Q}_0$-matrices.

Définition 2.5 On dit qu'une matrice $M \in \mathbb{R}^{n\times n}$ est une $\mathbf{Q}_0$-*matrice* si elle vérifie l'une des conditions équivalentes de la proposition 2.6 ci-dessous. On note $\mathbf{Q}_0$ l'*ensemble des* $\mathbf{Q}_0$-*matrices*. □

Voici des propriétés caractérisant la $\mathbf{Q}_0$-matricité.

Proposition 2.6 ($\mathbf{Q}_0$-matrice) *Soit $M \in \mathbb{R}^{n\times n}$. Les propriétés suivantes sont équivalentes :*

(i) *le problème* $\mathrm{CL}(M,q)$ *a une solution s'il est réalisable,*
(ii) $Q_S(M) = Q_R(M)$,
(iii) $Q_S(M)$ *est convexe.*

Terminons par la défintion des $\mathbf{Q}$-matrices.

Définition 2.7 On dit qu'une matrice $M \in \mathbb{R}^{n\times n}$ est une **Q**-*matrice* si $Q_S(M) = \mathbb{R}^n$ c'est-à-dire si $\mathrm{CL}(M,q)$ a une solution pour tout vecteur $q \in \mathbb{R}^n$. La caractérisation des **Q**-matrices se fait au moyen de celle des **S**-matrices et des $\mathbf{Q}_0$-matrices définies précédemment. On note **Q** l'*ensemble des* **Q**-*matrices*. □

Si l'on peut bien caractériser les **Q**-matrices, il est cependant plus difficile de pouvoir établir cette proprété numériquement en temps fini. Observons que

$$\boxed{\mathbf{Q} = \mathbf{Q}_0 \cap \mathbf{S}.} \tag{2.2.4}$$

2.2.4 Classe des $\mathbf{P}_0$-matrices

Définition 2.8 On dit qu'une matrice $M \in \mathbb{R}^{n\times n}$ est une $\mathbf{P}_0$-*matrice* si elle vérifie l'une des conditions équivalentes de la proposition 2.9 ci-dessous. On note $\mathbf{P}_0$ l'*ensemble des* $\mathbf{P}_0$-*matrices* [43, 1966]. □

Voici des propriétés caractérisant la $\mathbf{P}_0$-matricité.

Proposition 2.9 ($\mathbf{P}_0$-matricité) *Soit $M \in \mathbb{R}^{n\times n}$. Alors les propriétés suivantes sont équivalentes :*
- (i) $\forall I \subset [\![1,n]\!]$, $\det M_{II} \geqslant 0$,
- (ii) *pour tout $x \neq 0$, on peut trouver un indice i tel que $x_i \neq 0$ et $x_i(Mx)_i \geqslant 0$,*
- (iii) $\forall I \subset [\![1,n]\!]$, *les valeurs propres réelles de $M_{I,I}$ sont positives.*

Remarque 2.10 *L'équivalence $(i) \Leftrightarrow (ii)$ est due à Fiedler et Pták [43, 1966].*

2.2.5 Classe des P-matrices

Définition 2.11 On dit qu'une matrice $M \in \mathbb{R}^{n\times n}$ est une **P**-*matrice* si elle vérifie l'une des conditions équivalentes de la proposition 2.12 ci-dessus. On note **P** l'*ensemble des* **P**-*matrices*. □

La proposition 2.12 ci-dessous donne des conditions nécessaires et suffisantes sur M pour que le problème de complémentarité linéaire $\mathrm{CL}(M,q)$ ait une et une seule solution,

quel que soit $q \in \mathbb{R}^n$. L'implication $(i) \Rightarrow (ii)$ se montre comme suit. L'existence d'une solution s'obtient en considérant le problème d'optimisation quadratique *non*-convexe

$$\begin{cases} \inf x^\top (Mx+q) \\ Mx+q \geqslant 0 \\ x \geqslant 0. \end{cases} \tag{2.2.5}$$

Si $M \in \mathbf{P}$, ce problème est réalisable et borné ; il a donc une solution. On montre alors qu'en ses points stationnaires le critère est nul, si bien que ceux-ci sont solutions de $\mathrm{CL}(M,q)$. L'unicité implique alors que (2.2.5) a un unique point stationnaire.

Voici des propriétés caractérisant la **P**-matricité.

Proposition 2.12 (P-matricité) *Soit $M \in \mathbb{R}^{n\times n}$. Alors les propriétés suivantes sont équivalentes :*

(i) $\forall I \subset [\![1,n]\!]$, $\det M_{II} > 0$,
(ii) *tout x vérifiant $x \cdot (Mx) \leqslant 0$ est nul,*
(iii) $\forall I \subset [\![1,n]\!]$, *les valeurs propres réelles de M_{II} sont strictement positives,*
(iv) $\forall q \in \mathbb{R}^n$, *le problème $\mathrm{CL}(M,q)$ a une et une seule solution.*

La propriété (i) permet de vérifier qu'une matrice donnée est une **P**-matrice en examinant le signe de $2^n - 1$ déterminants, ce qui est beaucoup de travail ; on montre en fait que le problème de déterminer si une matrice donnée est une **P**-matrice est co-NP complet [31, 1994]. La propriété (ii) exprime qu'une **P**-matrice ne peut pas changer *tous* les signes d'un vecteur $x \neq 0$. L'équivalence $(i) \Leftrightarrow (iv)$ montre que les **P**-matrices sont pour le problème de complémentarité linéaire (2.1.3), ce que sont les matrices définies positives pour le problème d'optimisation quadratique sans contrainte.

Observons que si $M \in \mathbf{P}$ et $I \subset [\![1,N]\!]$, le point (i) de la proposition 2.12 affirme que $\det M_{II} > 0$ et donc que M_{II} est inversible.

Remarque 2.13 *Par les propriétés* (2.2.4) *et* (iv), *on voit que*

$$\boxed{\mathbf{P} \subset \mathbf{Q} \subset \mathbf{S}.} \tag{2.2.6}$$

En général, $\mathbf{P} \neq \mathbf{S}$. *Par exemple [30, 2009]*

$$M := \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

est une **S***-matrice car* $Me = (3,3) > 0$*, mais* $M \notin \mathbf{P}$ *car son déterminant est strictement négatif.*

Remarque 2.14 *Si* $M \notin \mathbf{P}$*, alors il existe un* $q \in \mathbb{R}^n$ *tel que l'une des deux situations exclusives suivantes ait lieu :*

- *soit* $\mathrm{CL}(M,q)$ *n'a pas de solution,*
- *soit* $\mathrm{CL}(M,q)$ *a plus d'une solution.*

Mais on ne peut pas s'assurer que, pour une matrice $M \notin \mathbf{P}$*, même symétrique et non-dégénérée, il existe un vecteur* q *tel que la première des deux situations ait lieu. En voici un contre-exemple.*

Contre-exemple 2.15 *Il existe une matrice symétrique* $M \notin \mathbf{P}$ *non-dégénérée, telle que* $\mathrm{CL}(M,q)$ *a une solution quel que soit* q*.* En voici une pour $n = 2$:

$$M = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

La matrice M est bien symétrique non-dégénérée et n'est pas dans **P**. Vérifions maintenant que $\mathrm{CL}(M,q)$ a bien une solution quel que soit $q \in \mathbb{R}^2$.

1. Si $q \geqslant 0$, alors $x = 0$ est solution.
2. Si $q_2 \geqslant 2q_1$ et $q_1 \leqslant 0$, alors $x = (-q_1, 0) \geqslant 0$ est solution. En effet, $Mx + q = (0, q_2 - 2q_1)$ est $\geqslant 0$ et complémentaire à x.
3. Si $q_1 \geqslant 2q_2$ et $q_2 \leqslant 0$, alors $x = (0, -q_2) \geqslant 0$ est solution. En effet, $Mx + q = (q_1 - 2q_2, 0)$ est $\geqslant 0$ et complémentaire à x.

On a bien ainsi couvert tous les $q \in \mathbb{R}^2$. □

2.2.6 Classe des Z-matrices

Définition 2.16 On dit qu'une matrice $M \in \mathbb{R}^{n\times n}$ est une **Z***-matrice* si $M_{ij} \leqslant 0$ pour $i \neq j$. On note **Z** l'*ensemble des* **Z***-matrices*. □

Le résultat suivant montre qu'un problème de complémentarité linéaire avec une matrice $M \in \mathbf{Z}$ est très simple, car on peut en trouver une solution en résolvant un problème d'optimisation linéaire. On rappelle que $u \in P$ est un *élément minimal* de P pour un ordre noté $\leqslant$, si $u \leqslant x$, pour tout $x \in P$; un tel élément, s'il existe, est nécessairement unique.

Proposition 2.17 *Soient $M \in \mathbf{Z}$ et $q \in \mathbb{R}^n$ tels que $\mathrm{CL}(M,q)$ soit réalisable. Alors $\mathrm{Adm}(M,q)$ a un élément minimal pour l'ordre $\leqslant$ de $\mathbb{R}^n$ et celui-ci est solution de $\mathrm{CL}(M,q)$. Il peut être calculé en résolvant le problème d'optimisation linéaire*

$$\begin{cases} \min\ p^\top x \\ Mx+q \geqslant 0 \\ x \geqslant 0, \end{cases} \tag{2.2.7}$$

dans lequel p est arbitraire dans $\mathbb{R}^n_{++}$.

Le résultat d'existence précédent a une réciproque qui caractérise les **Z**-matrices.

Proposition 2.18 (Z-matrice) *Les propriétés suivantes de $M \in \mathbb{R}^{n\times n}$ sont équivalentes :*

*(i) M est une **Z**-matrice,*

(ii) pour tout q rendant $\mathrm{CL}(M,q)$ réalisable, $\mathrm{Adm}(M,q)$ contient un minimum (pour l'ordre $\leqslant$ de $\mathbb{R}^n$) qui est solution de $\mathrm{CL}(M,q)$.

Remarque 2.19 *Le résultat précédent implique une observation faite par Chandrasekaran [21, 1970], à savoir que*

$$\mathbf{Z} \subset \mathbf{Q}_0.$$

2.2.7 Classe des M-matrices

Définition 2.20 On dit qu'une matrice $M \in \mathbb{R}^{n\times n}$ est une **M**-*matrice* si $M \in \mathbf{Z}$ et si M vérifie l'une des propriétés équivalentes de la proposition 2.21 ci-dessous. On note **M** l'*ensemble des* **M**-*matrices*. □

Proposition 2.21 (M-matrice) *Pour une matrice $M \in \mathbf{Z}$, les propriétés suivantes sont équivalentes :*

(i) $M \in \mathbf{P}$,
(ii) M *est inversible et* $M^{-1} \geqslant 0$,
(iii) *tous les valeurs propres de* M *ont une partie réelle* > 0,
(iv) $M \in \mathbf{S}$.

Beaucoup de schémas de discrétisation d'opérateurs différentiels du second ordre conduisent à des **M**-matrices [61, remarque 3.2] ; des exemples sont donnés dans [62, 1987] et [82, 1987]. Pour d'autres caractérisations des **M**-matrices, voir Fiedler et Pták [42, 1962].

2.2.8 Classe des NM-matrices

L'algorithme de Newton-min est l'algorithme de Newton non lisse (voir la section 2.5.3) sur l'équation

$$\min(x, Mx+q) = 0,$$

qui est équivalente à $\mathrm{CL}(M,q)$ (voir la section 2.4.2).

Définition 2.22 On dit qu'une matrice $M \in \mathbb{R}^{n\times n}$ est une **NM**-*matrice* si, quel que soit $q \in \mathbb{R}^n$, l'algorithme de Newton-min converge. On note **NM** l'*ensemble des* **NM**-*matrices*. □

Remarque 2.23 *Il sera montré au chapitre 4 que*

$$\mathbf{NM} \subset \mathbf{P}.$$

2.2.9 Classe des $\mathbf{R}_0$-matrices

Voici une classe des matrices qui intervient au chapitre 5 dans l'étude de la résolution des problèmes de complémentarité linéaire par la méthode de Newton non lisse. Soit $\Theta : \mathbb{R}^n \to \mathbb{R}$ la fonction de mérite quadratique du problème $\mathrm{CL}(M,q)$ définie par

$$\Theta(x) = \frac{1}{2} \| \min(x, Mx+q) \|^2. \tag{2.2.8}$$

On rappelle qu'une fonction définie sur $\mathbb{R}^n$ à valeurs réelles est dite *coercive* si elle a ses ensembles de sous-niveaux compacts, ce qui revient à dire qu'elle tend vers l'infini si $\|x\| \to \infty$.

Définition 2.24 On dit qu'une matrice $M \in \mathbb{R}^{n\times n}$ est une $\mathbf{R}_0$-*matrice* si elle vérifie les conditions équivalentes de la proposition 2.25 sont vérifiées. On note $\mathbf{R}_0$ l'*ensemble des* $\mathbf{R}_0$-*matrices*. □

La proposition suivante est démontrée dans [38, propositions 2.6.5 et 9.1.26].

Proposition 2.25 ($\mathbf{R}_0$-matricité) *Soit $M \in \mathbb{R}^{n\times n}$. Les propriétés suivantes sont équivalentes :*
- (i) *l'unique solution du problème* $\mathrm{CL}(M,0)$ *est la solution nulle,*
- (ii) *pour tout $q \in \mathbb{R}^n$, la fonction Θ est coercive,*
- (iii) *pour $q = 0$, la fonction Θ est coercive.*

Exemples

Voici quelques exemples de $\mathbf{R}_0$-matrices [38, proposition 9.1.26] :

- Les matrices non-dégénérées.
- Les matrices pour lesquelles, quel que soit $I \subset [\![1,n]\!]$, le seul $x_I \geqslant 0$ vérifiant $M_{I,I}x_I = 0$ est $x_I = 0$.
- Les matrices pour lesquelles, quel que soit $I \subset [\![1,n]\!]$, il existe un vecteur x_I tel que $M_{I,I}^\top x_I > 0$.
- Les matrices *strictement copositives*, c'est-à-dire celles M qui vérifient $x^\top Mx > 0$ pour tout $x \geqslant 0$ non nul.

2.2.10 Classes de matrices AS, SDP, DP, $\mathbf{P}_*(\kappa)$ et $\mathbf{P}_*$

Voici quelques classes (ou ensembles) de matrices qui ne sont pas directement rattachées à des propriétés particulières du problème $\mathrm{CL}(M,q)$, mais qui sont fréquemment rencontrées dans des approches algorithmiques.

Définition 2.26 On dit qu'une matrice $M \in \mathbb{R}^{n\times n}$ est *anti-symétrique* si $M^\top = -M$. On note **AS** l'*ensemble des matrices antisymétriques*. □

On voit facilement que $M \in \mathbf{AS}$ si et seulement si $x^\top Mx = 0$ pour tout x. Par ailleurs, comme les éléments diagonaux d'une **P**-matrice sont > 0 et que ceux d'une matrice anti-symétrique sont nuls, on a

$$\boxed{\mathbf{AS} \cap \mathbf{P} = \varnothing.} \tag{2.2.9}$$

Définition 2.27 On dit qu'une matrice $M \in \mathbb{R}^{n\times n}$ (non nécessairement symétrique) est *semi-définie positive* si $x^\top Mx \geqslant 0$ pour tout x. On note **SDP** l'*ensemble des matrices semi-définies positives*. □

Clairement, on a

$$\boxed{\mathbf{AS} \subsetneq \mathbf{SDP}.} \tag{2.2.10}$$

Définition 2.28 On dit qu'une matrice $M \in \mathbb{R}^{n\times n}$ est *définie positive* si $x^\top Mx > 0$ pour tout $x \neq 0$. On note **DP** l'*ensemble des matrices définies positives*. □

Remarque 2.29 *Par la proposition 2.9 (ii) des $\mathbf{P}_0$-matrices et la proposition 2.12 (ii) des* **P***-matrices, on voit que*

$$\boxed{\mathbf{DP} \subset \mathbf{P}} \qquad \textit{et} \qquad \boxed{\mathbf{SDP} \subset \mathbf{P}_0.}$$

Ces inclusions ne sont pas des égalités en général. Exemples [30, 2009, page 147] :

$$M := \begin{pmatrix} 1 & -3 \\ 0 & 1 \end{pmatrix} \in \mathbf{P} \setminus \mathbf{DP} \qquad \textit{et} \qquad M := \begin{pmatrix} 1 & -2 \\ 0 & 0 \end{pmatrix} \in \mathbf{P}_0 \setminus \mathbf{SDP},$$

car $e^\top Me = -1$. □

Voici une classe des matrices qui intervient dans l'étude de la résolution des problèmes de complémentarité linéaire par les méthodes de points intérieurs [78, 1991].

Définition 2.30 Soit $\kappa \geqslant 0$. On dit qu'une matrice $M \in \mathbb{R}^{n\times n}$ est une $\mathbf{P}_*(\kappa)$-*matrice* si pour tout $x \in \mathbb{R}^n$:

$$(1+4\kappa)\sum_{i\in I_+(x)} x_i(Mx)_i + \sum_{i\in I_-(x)} x_i(Mx)_i \geqslant 0, \tag{2.2.11}$$

où $I_+(x) = \{i : x_i(Mx)_i > 0\}$ et $I_-(x) = \{i : x_i(Mx)_i < 0\}$ (la dépendance en M de $I_+(x)$ et $I_-(x)$ n'est pas indiquée). On note $\mathbf{P}_*(\kappa)$ l'ensemble des $\mathbf{P}_*(\kappa)$-matrices. La plus petite valeur $\kappa \geqslant 0$ telle que $M \in \mathbf{P}_*(\kappa)$ est appelée le *handicap* de M.

On dit qu'une matrice $M \in \mathbb{R}^{n\times n}$ est une $\mathbf{P}_*$-*matrice* s'il existe un $\kappa \geqslant 0$ tel que $M \in \mathbf{P}_*(\kappa)$. On note

$$\mathbf{P}_* := \bigcup_{\kappa\geqslant 0} \mathbf{P}_*(\kappa)$$

l'ensemble des $\mathbf{P}_*$-matrices. □

Clairement, on a

$$0 \leqslant \kappa_1 \leqslant \kappa_2 \qquad \Longrightarrow \qquad \mathbf{P}_*(\kappa_1) \subset \mathbf{P}_*(\kappa_2) \subset \mathbf{P}_*.$$

La relation (2.2.11) s'écrit aussi

$$4\kappa \sum_{i\in I_+(x)} x_i(Mx)_i + x^\top Mx \geqslant 0,$$

si bien que l'on a

$$\boxed{\mathbf{P}_*(0) = \mathbf{SDP}.}$$

Remarque 2.31 *Selon [78, 1991], on a*

$$\boxed{\mathbf{P} \subsetneq \mathbf{P}_* \subsetneq \mathbf{P}_0.}$$

□

Pour résumer, la figure 2.2.1 illustre les propriétés d'inclusion de ces classes de matrices.

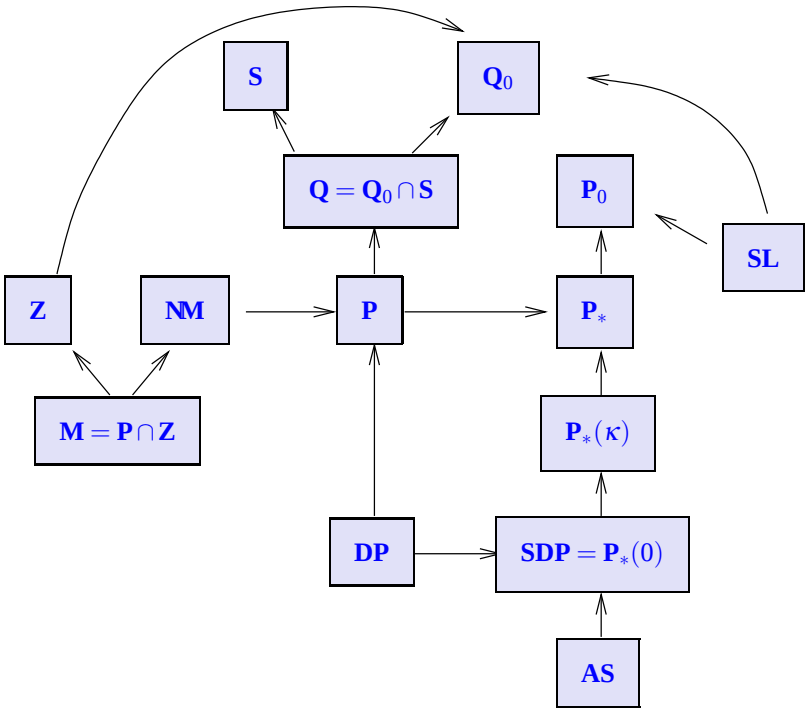


Figure 2.2.1 – Propriétés d'inclusion de quelques classes de matrices : la flèche $A \to B$ indique que $A \subset B$ (on peut cliquer sur la figure pour voir les propriétés de chaque classe de matrices)

Il n'y a pas d'algorithme idéal pour résoudre un problème de complémentarité linéaire, mais il y a un ensemble d'algorithmes qui sont, par leurs caractéristiques, plus ou moins adaptés à des classes particulières de problèmes. Nous citons, par exemple, les méthodes de pivotage qui ont une complexité exponentielle [102, 103, 86]. Il y a aussi les méthodes de points intérieurs [32, 67, 78, 88] et les méthodes non lisses auxquelles nous nous intéressons [24, 36, 96, 101, 26]

2.3 Complexité de CL(M,q)

Nous présentons ici quelques résultats de complexité concernant la difficulté de résoudre le problème CL(M,q), lorsque M est dans une classe donnée de matrices.

- Le problème CL(M,q) est NP-complet si $M \in \mathbf{P}_0$ [78, 1991].
- Le handicap d'une matrice $M \in \mathbf{P}_*$ peut être exponentiel en la taille en bit de M [32, 2010]. Ceci implique que les algorithmes (de points intérieurs en particulier), qui font intervenir κ polynomialement dans leur estimation de complexité ne sont pas polynomiaux.
- CL(M,q) peut être résolu en temps polynomial si M est définie positive (donc une **P**-matrice symétrique, propriété affirmée dans [100, 2002]) et même si M est semi-définie positive (par l'algorithme de l'ellipsoïde [89, 103]).
- Si M est une **M**-matrice, CL(M,q) peut être résolu en temps polynomial (par l'algorithme de Newton-min qui converge,dans ce cas, en au plus n itérations [71, 2004]).

2.4 Formulations équivalentes à CL(M,q)

2.4.1 Liens avec le problème d'optimisation quadratique

On considère le problème d'optimisation quadratique

$$\begin{cases} \inf\limits_{x\in\mathbb{R}^n} q^\top x + \dfrac{1}{2} x^\top M x \\ Ax \leqslant b, \end{cases} \tag{2.4.1}$$

où $q \in \mathbb{R}^n$, M est une matrice symétrique d'ordre n (non nécessairement semi-définie positive), A est une matrice de type $m \times n$ et $b \in \mathbb{R}^n$. Ses conditions d'optimalité du premier ordre s'écrivent pour un certain multiplicateur $\lambda \in \mathbb{R}^n$:

$$\begin{cases} q + Mx + A^\top \lambda = 0 \\ 0 \leqslant \lambda \perp (b - Ax) \geqslant 0. \end{cases} \tag{2.4.2}$$

Il s'agit d'un problème de complémentarité linéaire en (x,λ). Si M est inversible, on peut éliminer la première équation de (2.4.2) en la récrivant $x = -M^{-1}(A^\top \lambda + q)$, ce qui

conduit au problème de complémentarité linéaire en λ suivant :

$$0 \leqslant \lambda \perp (AM^{-1}A^{\top}\lambda + b + AM^{-1}q) \geqslant 0.$$

Ce problème a une et une seule solution λ si M est inversible et si $AM^{-1}A^{\top}$ est une **P**-matrice. En partant du (2.4.2) et en éliminant λ, et si les contraintes du problème (2.4.1) sont des contraintes de positivité de la variable x, alors on retrouve le problème $\mathrm{CL}(M,q)$ qui est équivalent aux conditions de Karush, Kuhn et Tucker du problème d'optimisation quadratique suivant

$$\begin{cases} \inf\limits_{x\in\mathbb{R}^n} q^{\top}x + \dfrac{1}{2}x^{\top}Mx \\ x \geqslant 0. \end{cases} \tag{2.4.3}$$

Un autre lien avec l'optimisation quadratique est le suivant : si la matrice M n'est pas une matrice symétrique, alors la relation entre $\mathrm{CL}(M,q)$ et le problème (2.4.3) ne tient plus. Dans ce cas, nous considérons le problème d'optimisation quadratique suivant

$$\begin{cases} \min\ x^{\top}(Mx+q) \\ x \geqslant 0 \\ Mx+q \geqslant 0. \end{cases} \tag{2.4.4}$$

Comme le coût de ce problème est borné inférieurement sur l'ensemble admissible (par zéro), ce problème (2.4.4) a toujours une solution [49]. On en déduit alors que

$x \in \mathrm{Sol}(M,q)$	$\Longleftrightarrow$	x est solution de (2.4.4) avec un coût optimal nul.

Cette observation montre que le problème de complémentarité linéaire est un cas particulier des problèmes quadratiques.

2.4.2 Formulations au moyen d'équations

On peut exprimer le problème de complémentarité $\mathrm{CL}(M,q)$ au moyen d'équations, en général non lisses, c'est-à-dire non différentiables. Les fonctions qui interviennent dans cette formulation et dont on cherche un zéro sont appelées des C-fonctions (C pour complémentarité). Cette formulation est utilisée par des algorithmes de résolution.

On dit que $f : \mathbb{R}^2 \to \mathbb{R}$ est une C-fonction si

$$f(a,b) = 0 \qquad \Longleftrightarrow \qquad ab = 0, \quad a \geqslant 0, \quad b \geqslant 0.$$

Une C-fonction permet donc de remplacer le problème $\mathrm{CL}(M,q)$ par le système d'équations

$$F(x) = 0,$$

où $F : \mathbb{R}^n \to \mathbb{R}^n$ a sa composante i définie par

$$F_i(x) = f(x_i, (Mx+q)_i).$$

On a intérêt à avoir des C-fonctions F non-différentiables, car elles conduisent à des algorithmes plus efficaces, pour au moins deux raisons : d'une part, les C-fonctions différentiables sont en général plus compliquées, plus non-linéaires ; d'autre part, les jacobiennes des C-fonctions différentiables sont en général singulières en des solutions dégénérées (et donc l'algorithme de Newton ne fonctionne pas) [72, 1995, théorème 3.1 et remarque 3.2].

On trouvera ci-après quelques exemples de C-fonctions et leurs propriétés.

2.4.2.1 La C-fonction de Fischer-Burmeister

La fonction *fonction de Fischer-Burmeister* [44, 45, 1992-1995] est définie par :

$$f_{FB}(a,b) = \sqrt{a^2+b^2} - (a+b).$$

On montre facilement qu'il sagit d'une C-fonction convexe et différentiable partout sauf en $(0,0)$. De plus, son carré est continûment différentiable. Elle est C^1 dans le voisinage d'un point non dégénéré. On a

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

Chen, Chen et Kanzow [23, 2000] considèrent ce qu'ils appellent la fonction de Fischer-Burmeister pénalisée :

$$f_{FBP}(a,b) = \lambda f_{FB}(a,b) + (1-\lambda)a^+b^+,$$

pour un $\lambda \in]0,1[$

L'intérêt principal de f_{FB} et f_{FBP} est que leurs carrés sont C^1, ce qui permet une globalisation classique des méthodes de Newton non lisses [113, 47]. Cependant la méthode introduit du mauvais conditionnement lorsque l'on se rapproche de la solution.

D'après [23, 101], les résultats numériques sont plutôt meilleurs avec f_{FBP} qu'avec f_{FB}.

2.4.2.2 La C-fonction de Mangasarian

La *fonction de Mangasarian* [90, 1976] est définie par

$$f_M(a,b) = \zeta(|a-b|) - \zeta(b) - \zeta(a),$$

où $\zeta : \mathbb{R} \to \mathbb{R}$ est strictement croissante et vérifie $\zeta(0) = 0$.

Elle peut être rendue différentiable par un choix approprié de ζ. Par exemple, elle est C^2 si $\zeta(t) = t^3$. On retrouve la fonction minimum (voir ci-dessous) en prenant $\zeta(t) \equiv t$, puisqu'alors $f_M(a,b) = -2\min(a,b)$.

2.4.2.3 La C-fonction minimum

La C-fonction la plus simple est la *fonction minimum* :

$$f_{\min}(a,b) = \min(a,b).$$

Elle semble avoir été proposée pour la première fois par Kostreva [80, 1976] et par Mangasarian [91, 1977]. On montre facilement qu'il s'agit bien d'une C-fonction. Dès lors

$$\min(x, Mx+q) = 0 \qquad \Longleftrightarrow \qquad x \in \mathrm{Sol}(M,q),$$

où la fonction min agit composante par composante : $\min(u,v)_i = \min(u_i, v_i)$ pour tout indice i.

La fonction min a été utilisé dans plusieurs travaux [106, 59, 107, 48, 123, 34]. Les tests numériques de [33, 71, 2004], montrent que l'utilisation de la C-fonction min donne de meilleurs résultats que celle de Fischer-Burmeister f_{FB}.

Comme la fonction min est souvent très efficace numériquement, nous l'avons choisie, dans ce travail, pour remplacer les conditions de complémentarité (2.1.1). Dans ce cas, le problème (2.1.3) devient équivalent à résoudre l'équation

$$H(x) = \min(x, Mx+q) = 0. \tag{2.4.5}$$

2.5 Analyse non lisse

Nous adoptons ici les notations habituelles et générales de l'analyse et de l'optimisation non lisse.

2.5.1 Ingrédients pour l'analyse non lisse

L'espace $\mathbb{R}^n$ est muni du produit scalaire euclidien $\langle x,y\rangle = x^\top y$. La norme associée est notée $||\cdot||$. On désigne par $B(x,r)$ la boule (ouverte) de centre $x \in \mathbb{R}^n$ et de rayon $r > 0$, définie par

$$B(x,r) := \{y \in \mathbb{R}^n : ||x-y|| < r\}.$$

Nous ne rappelons pas ici les notions fondamentales du calcul différentiel (différentiabilité au sens de Fréchet, au sens de Gâteaux, gradient, fonctions de classe C^n), mais nous fixons les quelques définitions supplémentaires indispensables à une généralisation de ces notions.

Définition 2.32 *Une fonction $f : \mathbb{R}^n \to \mathbb{R}$ est dite lipschitizienne sur $S \subset R^n$ s'il existe une constante $L > 0$ telle que, pour tout $y, z \in S$*

$$|f(y) - f(z)| \leqslant L||y-z||. \tag{2.5.1}$$

Une fonction f est dite localement lipschitizienne en $x \in R^n$ s'il existe un réel positif $\varepsilon > 0$ tel que f est lipschitizienne sur $B(x,\varepsilon)$.

Si f est lipschitizienne sur S (resp. localement lipschitizienne en x), toute valeur de L qui vérifie l'inégalité (2.5.1) de la définition 2.32 est appelée constante de Lipschitz de f sur S (resp. en x, ou au voisinage de x). Le caractère localement lipschitzien d'une fonction en x assure que son taux d'accroissement peut être contrôlé dans un voisinage de x. Une fonction localement lipschitzienne en un point n'est pas nécessairement diffrentiable en ce point (par exemple $x \to |x|$ en 0).

Terminons par la notion de dérivée directionnelle.

Définition 2.33 *Une fonction $f : \mathbb{R}^n \to \mathbb{R}$ admet une dérivée directionnelle (unilatérale) en $x \in \mathbb{R}^n$ dans la direction $d \in \mathbb{R}^n$ si la limite*

$$\lim_{\substack{t\to 0\\ t>0}} \frac{f(x+td)-f(x)}{t} \tag{2.5.2}$$

existe et est finie. On la note alors $f'(x,d)$. □

Si f est différentiable en x (au sens de Fréchet ou au sens de Gâteaux), alors elle admet des dérivés directionnelles dans toutes les directions d, et on a $f'(x,d) = f'(x)(d) = \nabla f(x)^\top d$ où $f'(x)$ est dérivée de f en x et $\nabla f(x)$ son gradient. Rappelons que la réciproque n'est pas vraie en général, sauf si les dérivées directionnelles sont continues.

Définition 2.34 *Une fonction* $f : \mathbb{R}^n \to \mathbb{R}$ *est dite sous-linéaire si elle est convexe et positivement homogène,* $f(tx) = t f(x)$ *pour tout* $x \in \mathbb{R}^n$ *et* $t > 0$. □

On montre facilement que, dans cette définition, l'hypothèse de convexité peut être remplacée par la sous-additivité de f :

$$f(x+y) \leqslant f(x) + f(y) \text{ pour tout } x,\ y \in \mathbb{R}^n.$$

Ainsi une fonction sous-linéaire est une fonction sous-additive et positivement homogène.

2.5.2 Sous-différentiel de Clarke

En 1975, Clarke définit des « gradients généralisés » pour une fonction localement lipschitzienne à valeurs réelles $f : \mathbb{R}^n \to \mathbb{R}$. Le but est de d'écrire le comportement à l'ordre un de fonctions non-différentiables. En 1976, Clarke étend la notion de sous-différentiel généralisé à des fonctions localement lipschitziennes à valeurs vectorielles. Rappelons-en la définition.

Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction localement lipschitzienne en $x \in R^n$, de constante de Lipschitz L, et ce dans toute la section 2.5.2. Cette hypothèse n'est pas suffisante pour parler de dérivées directionnelles de f, contrairement au cas convexe. Clarke [26], dont notre présentation reprend ici quelques résultats, définit la dérivée directionnelle généralisée de la façon suivante.

Définition 2.35 *La dérivée directionnelle généralisée en* x *dans la direction* d, *notée* $f^\circ(x,d)$ *est*

$$f^\circ(x;d) := \limsup_{\substack{x' \to x \\ t \downarrow 0}} \frac{f(x'+td) - f(x')}{t}.$$

□

Puisque f est localement lipschitzienne en $x \in R^n$, $f^\circ(x;d)$ est finie, et

$$|f^\circ(x;d)| \leqslant L\,||d||. \tag{2.5.3}$$

L'application $d \to f^\circ(x;d)$ hérite du caractère localement lipschitzien de f, de constante L, et celui-ci est même global. On montre de même que $d \to f^\circ(x;d)$ est sous-linéaire, de sorte qu'il est permis de définir un sous-différentiel pour les fonctions localement lipschitziennes.

Définition 2.36 *Le sous-différentiel de Clarke de f en x, noté $\partial f(x)$, est l'ensemble non vide de $\mathbb{R}^n$ défini par*

$$\partial f(x) = \{s \in \mathbb{R}^n \,:\, f^\circ(x;d) \geqslant \langle s,d\rangle,\ \forall d \in \mathbb{R}^n\}.$$

Les éléments de $\partial f(x)$ sont appelés les sous-gradients de Clarke (ou gradients généralisés) de f en x. □

Le sous-différentiel de Clarke est un ensemble compact et convexe. De plus, on a $\partial f(x) \subset B(0,L)$, d'après (2.5.3). Les dérivées directionnelles généralisées se déduisent de $\partial_C f(x)$ pour toute direction $d \in R^n$:

$$f^\circ(x;d) = \max_{s \in \partial f(x)} \langle s,d\rangle.$$

Le sous-différentiel de Clarke comme limites de gradients convergents

Si une fonction localement lipschitzienne n'est pas nécessairement différentiable, l'ensemble de ses points de non-différentiabilité est néanmoins de mesure nulle, comme l'affirme le théorème suivant, dû à Rademacher [26].

Théorème 2.37 *Soit $U \subset \mathbb{R}^n$ un ouvert, et $g : U \to \mathbb{R}^n$ une fonction localement lipschitzienne en tout point de U. Alors g est différentiable presque-partout sur $\mathbb{R}^n$ (pour la mesure de Lebesgue).*

Toute fonction $f : \mathbb{R}^n \to \mathbb{R}$ localement lipschitzienne sur $\mathbb{R}^n$ est donc différentiable partout, sauf sur Ω_{nd}, un ensemble de mesure nulle. La fonction f admet donc un gradient $\nabla f(x)$ en un point x arbitrairement proche de $x_{nd} \in \Omega_{nd}$. On a d'autre part la caractérisation du sous-différentiel de Clarke suivante [26, théorème 2.5.1]. Si une fonction

$f : \mathbb{R}^n \to \mathbb{R}$ est localement lipschitzienne sur $\mathbb{R}^n$ et différentiable sur $\mathbb{R}^n \backslash \Omega_{nd}$, alors

$$\partial f(x) \;=\; \text{co}\ \partial_{\text{B}} f(x), \tag{2.5.4}$$

où co désigne l'enveloppe convexe et $\partial_{\text{B}} f(x)$ est défini par

$$\partial_{\text{B}} f(x) = \{\lim f'(x_k) \,:\, x_k \to x,\ x_k \in \mathbb{R}^n \backslash \Omega_{nd},\ f'(x_k) \text{ converge}\}. \tag{2.5.5}$$

Cela permet d'étendre la définition du sous-différentiel de Clarke aux fonctions à valeurs vectorielles $f : \mathbb{R}^n \to \mathbb{R}^m$, par les mêmes formules (2.5.4) et (2.5.5) [26, section 2.6.1].

Finalement, le sous-différentiel produit pour les fonctions localement lipschitziennes à valeurs vectorielles $f : \mathbb{R}^n \to \mathbb{R}^m$, est définit par

$$\partial_{\times} f(x) := \partial f_1(x) \times \partial f_2(x) \times \cdots \times \partial f_m(x).$$

Remarque 2.38 *On observe que [38, proposition 7.1.14]*

$$\boxed{\partial f(x) \subseteq \partial_{\times} f(x).} \tag{2.5.6}$$

□

Comme $\text{co}(A \times B) = \text{co}(A) \times \text{co}(B)$ pour tous ensembles A et B, alors le sous-différentiel produit peut s'écrire comme suit

$$\partial_{\times} f(x) = \text{co}\left(\partial_{\text{B}} f_1(x) \times \partial_{\text{B}} f_2(x) \times \cdots \times \partial_{\text{B}} f_m(x)\right).$$

Règles de calcul

Nous donnons ici quelques règles de calcul utiles du sous-différentiel de Clarke. Les résultats ne sont souvent que des inclusions, mais il est possible d'obtenir des égalités en demandant une régularité plus forte que le caractère localement lipschitzien, dite régularité au sens de Clarke.

Définition 2.39 *Une fonction $f : \mathbb{R}^n \to \mathbb{R}$ localement lipschitzienne en x est dite régulière au sens de Clarke si elle admet des dérivées directionnelles $f'(x,d)$ dans toutes les directions $d \in \mathbb{R}^n$, et si de plus, on $f'(x;d) = f^\circ(x;d)$ pour tout $d \in \mathbb{R}^n$.*

En particulier, toute fonction convexe sur $\mathbb{R}^n$ et toute fonction continûment différentiable en x est régulière au sens de Clarke en x. D'autre part, si f est régulière au sens de Clarke et différentiable en x, alors $\partial f(x) = \{\nabla f(x)\}$.

Proposition 2.40 *Soit $(f_i)_{1\leqslant i\leqslant p}$ une famille finie de fonctions définie sur $\mathbb{R}^n$ à valeurs dans $\mathbb{R}$ et localement lipschitiziennes en $x \in \mathbb{R}^n$, et soit $(\alpha_i)_{1\leqslant i\leqslant p} \subset \mathbb{R}$. Alors la somme pondérée $\sum \alpha_i f_i$, est encore localement lipschitizienne en x, et on a*

$$\partial_C\Big(\sum_{i=1}^p \alpha_i f_i\Big) \subset \sum_{i=1}^p \alpha_i \partial_C f_i. \tag{2.5.7}$$

Remarque 2.41 *On a égalité dans l'équation (2.5.7) si une fonction f_i au plus n'est pas continûment différentiable en x. On a encore égalité si les f_i sont régulières au sens de Clarke, et si $\alpha_i \geqslant 0$ pour tout i.*

Proposition 2.42 *Soit $F : \mathbb{R}^n \to \mathbb{R}^m$ une fonction continûment différentiable en $x \in \mathbb{R}^n$, et $g : \mathbb{R}^m \to \mathbb{R}$ une fonction localement lipschitzienne en $F(x)$. Alors la composée $g \circ F$ est localement lipschitzienne en x, et on a*

$$\partial_C(g \circ F)(x) \subset \partial_C g(F(x)) \circ F'(x),$$

où $F'(x)$ désigne la différentielle de F en x. On a égalité si g est régulière au sens de Clarke en $F(x)$, et, dans ce cas, la composée $g \circ F$ est aussi régulière au sens de Clarke en x.

2.5.3 Schémas newtoniens pour problèmes non lisses

Soient Ω un *ouvert* de $\mathbb{R}^n$ et $F : \Omega \to \mathbb{R}^n$ une fonction non lisse dont on cherche un zéro, c'est-à-dire un point $x \in \mathbb{R}^n$ tel que

$$F(x) = 0.$$

On note $\mathscr{F}(\mathbb{R}^n, \mathbb{R}^n)$ l'ensemble des fonctions définies sur $\mathbb{R}^n$ à valeurs dans $\mathbb{R}^n$. Nous reprenons ici quelques éléments de la thèorie développée dans la section 7.2 de [38].

Définition 2.43 (schéma d'approximation newtonien) Un *schéma d'approximation newtonien* d'une fonction $F : \mathbb{R}^n \to \mathbb{R}^n$ en $x_0 \in \Omega$ est une application multivoque

$$J : \Omega_0 \twoheadrightarrow \mathscr{F}(\mathbb{R}^n, \mathbb{R}^n),$$

définie sur un voisinage $\Omega_0 \subset \Omega$ de x_0 telle que les conditions suivantes soient vérifiées :

(AN_1) pour tout $x \in \Omega_0$ et tout $J_x \in J(x)$, $J_x(0) = 0$,
(AN_2)

$$\limsup_{\substack{x \to x_0,\ x \neq x_0 \\ J_x \in J(x)}} \frac{F(x) + J_x(x_0 - x) - F(x_0)}{\|x - x_0\|} = 0. \tag{2.5.8}$$

□

Remarque 2.44 *L'application multivoque J joue dans l'algorithme de Newton le rôle de l'application dérivée F' du cas différentiable, mais $J(x)$ n'est pas formé d'applications* linéaires*, mais est un* sous-ensemble *de $\mathscr{F}(\mathbb{R}^n, \mathbb{R}^n)$. Les éléments $J_x \in J(x)$ sont donc des opérateurs* non *linéaires.*

Pour avoir une convergence quadratique, on aura besoin d'une hypothèse plus forte que (2.5.8)*; cette dernière n'assurant que la convergence superlinéaire.*

Définition 2.45 (schéma d'approximation newtonien fort) Un *schéma d'approximation newtonien fort* d'une fonction $F : \mathbb{R}^n \to \mathbb{R}^n$ en $x_0 \in \Omega$ est un schéma d'approximation newtonien J de F en x_0, dans lequel l'hypothèse (2.5.8) est renforcée en

(AN_2')

$$\limsup_{\substack{x\to x_0,\ x\neq x_0\\ J_x\in J(x)}} \frac{\|F(x)+J_x(x_0-x)-F(x_0)\|}{\|x-x_0\|^2} < \infty. \tag{2.5.9}$$

□

Pour avoir la convergence de l'algorithme de Newton schématique défini ci-dessous, on aura besoin d'une hypothèse de régularité du schéma d'approximation J au zéro x_0 de F considéré, qui est le pendant de la non-singularité de $F'(x_0)$ du cas différentiable.

Définition 2.46 (schéma d'approximation newtonien régulier) Un *schéma d'approximation newtonien régulier* d'une fonction $F:\mathbb{R}^n\to\mathbb{R}^n$ en $x_0\in\Omega$ est un schéma d'approximation newtonien J de F en x_0, formé d'*homéomorphismes uniformément lipschitziens* sur le voisinage Ω_0 de x_0, c'est-à-dire qu'il existe des constantes $L_J>0$ et $\varepsilon_J>0$ telles que pour tout $x\in\Omega_0$ et pour tout $J_x\in J(x)$, il existe des voisinages U_x et V_x, tels que $B(0,\varepsilon_J)\subset U_x$, $B(0,\varepsilon_J)\subset V_x$ et $J_x|_{U_x}:U_x\to V_x$ est un homéormorphisme lipschitzien dont l'inverse est lipschitzien de module L_J. □

Si F a un schéma d'approximation newtonien régulier J en un zéro x_* de F, on peut définir un algorithme de Newton qui générera une suite $\{x_k\}$ qui convergera vers x_*, pourvu que le premier itéré soit suffisament proche de x_*.

Algorithme 2.47 Newton schématique – une itération

On suppose qu'au début de l'itération k, on dispose d'un itéré $x_k\in\mathbb{R}^n$.

1. *Test d'arrêt* : si $F(x_k)=0$, arrêt de l'algorithme.
2. *Calcul du pas* : pour un opérateur J_k arbitraire dans $J(x_k)$, on calcule une solution d_k du système non linéaire suivant

$$F(x_k)+J_k(d_k)=0. \tag{2.5.10}$$

3. *Nouvel itéré* : $x_{k+1}:=x_k+d_k$.

L'algorithme suppose que le système *non* linéaire (2.5.10) est résoluble. On montrera en fait que, sous des hypothèses appropriées (en particulier x_k doit être assez proche d'un

zéro convenable de F), ce système a une solution unique. La consistance et la convergence *locale* de l'algorithme 2.47 sont assurées dans les conditions énoncées dans le théorème suivant [38, théorème 7.2.5].

Théorème 2.48 (convergence locale de Newton non lisse) *Soient Ω un ouvert de $\mathbb{R}^n$, $F : \Omega \to \mathbb{R}^n$ une fonction localement lipschitzienne sur Ω et $x_* \in \Omega$ tel que $F(x_*) = 0$. On suppose que F a un schéma d'approximation newtonien régulier J en x_*. Alors, il existe un voisinage V de x_* tel que si $x_0 \in V$, l'algorithme 2.47 de Newton schématique est bien défini et génère une suite $\{x_k\}$ convergeant superlinéairement vers x_*. Si le schéma d'approximation est fort, la convergence est quadratique.*

Dans la section suivante, nous nous intéressons à l'application de l'algorithme 2.47 à quelques classes particulières des fonctions non lisses : les fonctions B-différentiables, les functions semi-lisses et les fonctions différentiables par morceaux.

2.5.4 Fonctions non lisses et leur schéma newtonien

Nous présentons dans cette section, trois classes de fonctions non lisses et les schémas newtoniens associés. Nous ne suivons pas l'ordre chronologique de leur introduction, mais un ordre décroissant de difficulté de mise en œuvre. Les fonctions B-différentiables et l'algorithme de Bouligand-Newton associé requiert, en effet, de résoudre un système d'équations non linéaires à chaque itération. Dans les deux autres classes de fonctions (semi-lisses et différentiables par morceaux), l'itération de Newton ne requiert que la résolution d'un système linéaire. C'est dans la classe des fonctions C^1 par morceaux que se situe l'algorithme de Newton que nous utiliserons fréquemment par la suite. Nous ne couvrons pas toutes les méthodes de Newton non lisses développées ces dernières années, mais présentons ainsi une suite logique de méthodes conduisant à celle qui nous a intéressé dans cette thèse.

2.5.4.1 Fonctions B-différentiables

La B-différentiabilité a été introduite par Robinson [115]. Elle est un concept de différentiabilité plus faible que celui de Fréchet, dans lequel l'opérateur dérivée n'est pas requis d'être linéaire et borné, mais seulement positivement homogène et borné. La

lettre B fait référence à Bouligand. Cet affaiblissement important de la définition permet toutefois de préserver des propriétés importantes, telles que la B-différentiabilité en chaîne et la formule des accroissements finis. Contrairement à la Fréchet-différentiabilité, la B-différentiabilité n'est pas détruite par la prise du minimum ou du maximum d'un nombre fini de fonctions, ce qui est un atout dans certaines circonstances en particulier pour les problèmes de complémentarité. Présentons maintenant la définition de la B-différentiabilité.

Définition 2.49 *Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces normés. On dit qu'une fonction $f : \mathbb{E} \to \mathbb{F}$ est* B-différentiable *en $x \in \mathbb{E}$, s'il existe un opérateur $Bf(x) : \mathbb{E} \to \mathbb{F}$ positivement homogène (de degré un) et borné, tel que*

$$F(x+h) - F(x) - BF(x) \cdot h = o(h). \tag{2.5.11}$$

L'opérateur $BF(x)$ est appelé la B-dérivée de f en x.

On dit que $f : \mathbb{E} \to \mathbb{F}$ est B-différentiable sur un ouvert $\Omega \subset \mathbb{R}^n$ si f est B-différentiable en tout point de Ω. □

Pang [106, 1990] a étudié la méthode de Newton pour les fonctions B-différentiables et sa globalisation par recherche linéaire sur la fonction de mérite $x \mapsto \varphi(x) = \frac{1}{2}\|F(x)\|_2^2$. La direction d_k est obtenue comme solution du système *non* linéaire

$$BF(x_k)d_k = -F(x_k),$$

où $BF(x_k)$ est la B-dérivée de F en x_k. La direction d_k (si elle existe) est une direction descente de la fonction de mérite φ. Cependant, cette direction présente l'inconvénient d'être coûteuse numériquement du fait du système non linéaire qu'il faut résoudre pour la calculer [58, 59, 1990-1992]. Par exemple, dans le cas du PCL, le calcul d'une telle direction exige la résolution d'un problème de complémentarité linéaire [59, 1990].

Liens avec la différentiabilité directionnelle

Pour les fonctions localement lipschitziennes, la notion de B-différentiabilité est la même que celle de la différentiabilité directionnelle. On note $f'(x;h)$ la dérivée directionnelle en $x \in \mathbb{E}$ dans la direction $h \in \mathbb{E}$. Voici quelques liens avec la différentiabilié directionnelle [120, 1990].

Proposition 2.50 (i) *Si F est B-différentiable en x, alors F admet des dérivées di-*

rectionnelles en $x \in \mathbb{E}$ suivant toute direction $h \in \mathbb{E}$ et $F'(x;h) = Bf(x)h$.

(ii) *Si F est lipschitzienne dans un voisinage de x et si F admet des dérivées directionnelles en x suivant toute direction, alors f est B-différentiable en x.*

Nous présentons, dans la section suivante, la classe des fonctions semi-lisses. Pour ces fonctions, l'itération de Newton ne requiert que la résolution d'un système linéaire.

2.5.4.2 Fonctions semi-lisses

La notion de fonction semi-lisse a été introduite par Mifflin [99, 1977] et étendue aux fonctions multivoques par Qi et Sun [113, 1993].

Définition 2.51 Soient Ω un ouvert de $\mathbb{R}^n$ et $F : \Omega \to \mathbb{R}^n$ une fonction localement lipschitzienne sur Ω. On dit que F est *semi-lisse* en $x_0 \in \Omega$ si

(SL_1) F admet des dérivées directionnelles en x_0,
(SL_2) il existe un voisinage $\Omega_0 \subset \Omega$ de x_0 tel que

$$\forall x \in \Omega_0 : \quad F'(x; x-x_0) - F'(x_0; x-x_0) = o(\|x-x_0\|). \tag{2.5.12}$$

On dit que F est *semi-lisse* sur Ω si elle est semi-lisse en tout point de Ω. On dit que F est *fortement semi-lisse* en $x_0 \in \Omega$ si elle est semi-lisse en x_0 avec (SL_2) renforcée en

(SL_2') il existe un voisinage $\Omega_0 \subset \Omega$ de x_0 tel que

$$\forall x \in \Omega_0 : \quad F'(x; x-x_0) - F'(x_0; x-x_0) = O(\|x-x_0\|^2). \tag{2.5.13}$$

□

L'algorithme de Newton 2.47 appliqué à une fonction localement lipschitzienne quelconque, utilisant le sous-différentiel de Clarke (voir la section 2.5.2) comme schéma d'approximation newtonien (voir la section 2.5.3), ne converge pas nécessairement [38, 2003, exemple 7.4.1]. Les fonctions semi-lisses sont alors des fonctions localement lipschitziennes pour lesquelles l'algorithme de Newton converge avec un élément de sous-différentiel de Clarke [99, 113].

Exemple 2.52 *La C-fonction Fischer-Burmeister (voir la section 2.4.2.1) est une fonction semi-lisse [38].*

L'utilisation du sous-différentiel de Clarke $\partial F(x)$ comme schéma d'approximation newtonien a l'intérêt d'avoir ses éléments linéaires. Cependant, son utilisation dans l'algorithme 2.47 requiert que tous ses éléments soient inversibles ; par ailleurs, la non-singularité de $\partial F(x_*)$ en une solution x_* est requise pour avoir convergence de l'algorithme [38, théorème 7.5.3].

2.5.4.3 Fonctions différentiables par morceaux

On suppose dans cette section que F est $\mathscr{C}^1$ par morceaux. Les résultats obtenus sont utiles pour les problèmes de complémentarité représentés par la C-fonction min (section 2.4.2.3).

Définition 2.53 (fonction $\mathscr{C}^1_{\mathrm{M}}$) Soit Ω un ouvert de $\mathbb{R}^n$. On dit que $F : \Omega \to \mathbb{R}^n$ est $\mathscr{C}^1$ *par morceaux* si Ω est partitionné en un nombre fini de sous-domaines non vides $\mathscr{D}_1, \ldots, \mathscr{D}_p$ (deux à deux disjoints et dont l'union est Ω), qu'il existe des fonctions $F_i : \Omega_i \to \mathbb{R}^n$ définies dans un voisinage ouvert Ω_i de chaque sous-domaine $\mathscr{D}_i$ et de classe $\mathscr{C}^1$ sur Ω_i et que pour $x \in \mathscr{D}_i$, $F(x) = F_i(x)$. Si $F(x) = F_i(x)$ on dit la *fonction F_i est active* en x et on note

$$P(x) = \{i : F(x) = F_i(x)\},$$

l'ensemble des indices des fonctions actives en x. On note $\mathscr{C}^1_{\mathrm{M}}(\Omega, \mathbb{R}^n)$ l'ensemble des fonctions $\mathscr{C}^1$ par morceaux sur Ω à valeurs dans $\mathbb{R}^n$. □

Remarque 2.54 *Les fonctions différentiables par morceaux sont à la fois B-différentiables et semi-lisses.*

Pour ces fonctions, il est naturel de prendre comme schéma d'approximation newtonien la multifonction $J : \mathbb{R}^n \twoheadrightarrow \mathbb{R}^m$ définie en $x \in \mathbb{R}^n$ par [79]

$$J(x) = \{F_i'(x) : i \in P(x)\}.$$

Un premier intérêt vient du fait que tous les éléments de $J(x)$ sont linéaires. Un avantage de l'algorithme 2.47 tient au fait que le système (2.5.10) à résoudre à chaque itération devient *linéaire* contrairement aux algorithmes de B-Newton utilisant la B-différentiabilité [106, 59] (voir la section 2.5.4.1).

Pour que l'algorithme soit bien défini, il faudra que ce système ait une solution à chaque itération. En faisant l'hypothèse que les éléments de $J(x)$ sont inversibles dans le

voisinage d'un zéro de F et d'après le théorème 2.48, on obtient un algorithme localement convergent [38, théorème 7.2.15].

Dans cette thèse, nous nous sommes intéressés à la résolution des problèmes de complémentarité par l'algorithme de Newton non lisse appliqué à la fonction min, qu'on appelle Newton-min.

2.5.5 L'algorithme de Newton-min

Récriture sous forme d'équation non lisse

On considère l'algorithme de Newton-min pour résoudre (2.1.3) ou son équivalent sous forme d'équations

$$H(x) \equiv \min(x, Mx+q) = 0. \tag{2.5.14}$$

Quelques propriétés de la fonction min

Le lemme suivant montre que la fonction min est une fonction Lipschitzienne de module 1.

Lemme 2.55 *L'application* $\min : \mathbb{R}^{2n} \to \mathbb{R}^n : (x,y) \mapsto \min(x,y)$ *est non-expansive* (*i.e., lipschitzienne de module* 1) *en norme* ℓ_1 *: pour tout* (x,y), $(x',y') \in \mathbb{R}^{2n}$, *on a*

$$\|\min(x',y') - \min(x,y)\|_1 \leqslant \|(x',y') - (x,y)\|_1. \tag{2.5.15}$$

DÉMONSTRATION. En remarquant que $\min(x,y) = \frac{1}{2}(x+y-|x-y|)$, on a

$$\begin{aligned}
|\min(x',y') - \min(x,y)| &= \frac{1}{2}\big|x'-x+y'-y-|x'-y'|+|x-y|\big| \\
&\leqslant \frac{1}{2}\left(|x'-x|+|y'-y|+\big||x-y|-|x'-y'|\big|\right) \\
&\leqslant \frac{1}{2}(|x'-x|+|y'-y|+|x-y-x'+y'|) \\
&\leqslant |x'-x|+|y'-y| \\
&= \|(x',y') - (x,y)\|_1.
\end{aligned}$$

□

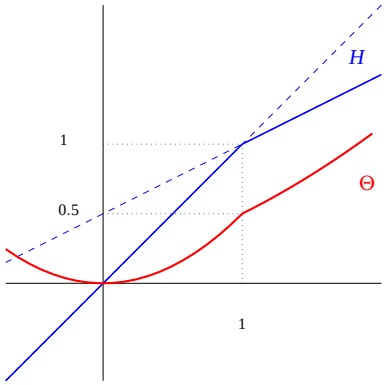
Régularité au sens de Clarke

Lemme 2.56 *La fonction de mérite ℓ_2 du problème* $\mathrm{CL}(M,q)$*, définie en* (2.2.8)*, n'est pas nécessairement régulière au sens de Clarke.*

DÉMONSTRATION. Prenons, par exemple,

$$n=1, \quad M=\frac{1}{2}, \quad \text{et} \quad q=\frac{1}{2}.$$

Alors, on a $\Theta(x)=\|H(x)\|^2/2$, avec $H(x)=\min\left(x,\frac{x+1}{2}\right)$ (voir la figure ci-dessous).



Dans ce cas, on peut facilement montrer que Θ n'est pas nécessairement régulière au sens de Clarke en $x=1$ parce que $\Theta^{\circ}(1;1)\neq\Theta'(1;1)$. En effet, il est clair que

$$\Theta'(1;1)=\lim_{t\downarrow 0}\frac{\Theta(1+t)-\Theta(1)}{t}=\lim_{t\downarrow 0}\frac{(2+t)^2/8-1/2}{t}=\frac{1}{2},$$

tandis que sa dérivée directionnelle généralisée s'écrit

$$\begin{aligned}
\Theta^\circ(1;1) &= \limsup_{\substack{x\to 1\\ t\downarrow 0}} \frac{\Theta(x+t)-\Theta(x)}{t} \\
&\geqslant \lim_{t\downarrow 0} \frac{\Theta((1-2t)+t)-\Theta(1-2t)}{t} \qquad [\text{pour } x=1-2t] \\
&= \lim_{t\downarrow 0} \frac{(1-t)^2-(1-2t)^2}{2t} \\
&= 1.
\end{aligned}$$

□

Jacobienne de Clarke de la fonction $H(x)=\min(x,Mx+q)$

Lemme 2.57 ($\partial_B H$) *Soit H définie en (2.4.5). La différentielle de Bouligand de H en x vérifie*

$$\partial_B H(x) \subset \{(I-D)+DM : \ D \text{ est diagonale avec } D_{ii}\in\{0,1\}\}. \tag{2.5.16}$$

En particulier, si $M\in\mathbf{P}$, tous les éléments de $\partial_B H(x)$ sont des **P***-matrices, donc des matrices inversibles.*

DÉMONSTRATION. Soit $\mathscr{D}$ l'ensemble des points de différentiabilité de H et $\{x^k\}$ une suite de $\mathscr{D}$ convergeant vers x, telle que $H'(x^k)\to J$. Il s'agit de montrer que J est dans l'ensemble à droite dans (2.5.16). En extrayant une sous-suite au besoin, on peut trouver un ensemble d'indices $I\subset[\![1,n]\!]$, indépendant de k, tel que $x_i^k\leqslant (Mx^k+q)_i$ pour $i\in I$ et $x_i^k>(Mx^k+q)_i$ pour $i\in I^c$. Comme $x^k\in\mathscr{D}$ par hypothèse, on obtient par la proposition 2.59 :

$$H'_I(x^k)=I_I \qquad \text{et} \qquad H'_{I^c}(x^k)=M_{I^c}.$$

À la limite, on trouve que $J_I=I_I$ et $J_{I^c}=M_{I^c}$, si bien que J est dans l'ensemble à droite dans (2.5.16) avec $D_{ii}=0$ si $i\in I$ et $D_{ii}=1$ si $i\in I^c$.

Par ailleurs, si $M \in \mathbf{P}$ et D est comme dans (2.5.16), alors $(I-D)+DM \in \mathbf{P}$ [3, lemme 2.1]. □

Dans l'article [25, 2009], on trouve un résultat plus précis sur le calcul du différentiel de Bouligand de $\min(F(x),G(x))$ (et aussi pour la fonction de Fischer-Burmeister 2.4.2.1).

Remarque 2.58 On n'a pas nécessairement égalité en (2.5.16). En effet, pour les données suivantes :

$$M = \begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}, \qquad q = 0 \qquad \text{et} \qquad \bar{x} = 0,$$

et pour $D = \mathrm{Diag}(1,0)$, la matrice

$$(I-D)+DM = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \notin \partial_{\mathrm{B}} H(\bar{x}) = \{I, M\}.$$

En effet, comme $x_1 = (Mx+q)_1 \Leftrightarrow x_1 = 0 \Leftrightarrow x_2 = (Mx+q)_2$, l'ensemble des points où H est différentiable s'écrit $\mathscr{D} = \{x \in \mathbb{R}^2 : x_1 \neq 0\}$. Par ailleurs, $x_1 < (Mx+q)_1 \Leftrightarrow x_1 > 0 \Leftrightarrow x_2 < (Mx+q)_2$ et $x_1 > (Mx+q)_1 \Leftrightarrow x_1 < 0 \Leftrightarrow x_2 > (Mx+q)_2$, si bien que le différentiel de Bouligand de H s'écrit $\partial_{\mathrm{B}}\Phi(\bar{x}) = \{I, M\}$.

Dans le cas présent, l'élément $(I-D)+DM$, avec $D = \mathrm{Diag}(1,0)$, n'est pas dans $\partial_{\mathrm{B}} H(\bar{x})$ parce que l'ensemble $\{x : x_1 > (Mx+q)_1, x_2 < (Mx+q)_2\}$ est vide. □

En appliquant le théorème 5 de Pang [106, 1990] au PCL et d'après le lemme 2.55, on obtient le résultat suivant, en particulier une caractérisation de la différentiabilité de la fonction $H(x) = \min(Mx+q, x)$ et de la fonction $\Theta(x) = \frac{1}{2}\|H(x)\|_2^2$.

Proposition 2.59 *(i) La fonction Θ admet des dérivées directionnelles, qui s'écrivent $\Theta'(x;d) = H(x)^\top H'(x;d)$, où*

$$H_i'(x;d) = \begin{cases} d_i & \text{if } i \in I_<(x) := \{i : x_i < (Mx+q)_i\} \\ \min(d_i, (Md)_i) & \text{if } i \in I_=(x) := \{i : x_i = (Mx+q)_i\} \\ (Md)_i & \text{if } i \in I_>(x) := \{i : x_i > (Mx+q)_i\}. \end{cases}$$

(*ii*) *H est F-différentiable en x si, et seulement si,* $M_{I_=(x)} = I_{I_=(x)}$ *et dans ce cas,*

$$H_i'(x) = \begin{cases} I_i & \text{if } i \in I_{\leqslant}(x) := \{i : x_i \leqslant (Mx+q)_i\} \\ M_i & \text{if } i \in I_{\geqslant}(x) := \{i : x_i \geqslant (Mx+q)_i\}. \end{cases}$$

Décrivons en détail l'algorithme Newton-min qui permet de résoudre le système d'équations non lisses (2.5.14).

Énoncé de l'algorithme

Algorithm 2.60 (Newton-min) Soit $x^1 \in \mathbb{R}^n$.

Pour $k = 2, 3, \ldots$, faire

1. Si x^{k-1} est une solution de $\mathrm{CL}(M,q)$, arrêt.
2. Choisir des ensembles d'indices complémentaires $A^k := A_0^k \cup A_1^k$ et $I^k := I_0^k \cup I_1^k$, où
$$\begin{aligned} A_0^k &:= \{i : x_i^{k-1} < (Mx^{k-1}+q)_i\}, \\ A_1^k \subset E^k &:= \{i : x_i^{k-1} = (Mx^{k-1}+q)_i\}, \\ I_0^k &:= \{i : x_i^{k-1} > (Mx^{k-1}+q)_i\}, \\ I_1^k &:= E^k \setminus A_1^k. \end{aligned}$$
3. Déterminer x^k comme solution de
$$\begin{cases} x_{A^k}^k = 0 \\ (Mx^k+q)_{I^k} = 0. \end{cases} \tag{2.5.17}$$

Cet algorithme est bien défini si M est *non-dégénérée*, c'est-à-dire si toutes les sous-matrices principales de M sont inversibles (voir la section 2.2.1). En effet, dans ce cas, la matrice

$$J_k = \begin{pmatrix} I_{A^kA^k} & 0 \\ M_{I^kA^k} & M_{I^kI^k} \end{pmatrix} \tag{2.5.18}$$

du système linéaire (2.5.17) est inversible (les lignes d'indices A_k de $J_k v = 0$ impliquent que $v_{A^k} = 0$, puis ses lignes d'indices I_k impliquent que $M_{I^k I^k} v_{I^k} = 0$ donc $v_{I^k} = 0$ par la non-singularité de $M_{I^k I^k}$).

Trois interprétations de l'algorithme

La question de savoir à quelle famille de méthodes on peut rattacher l'algorithme de Newton-min se pose et n'est pas sans importance. D'une part, cela permet de bénéficier des propriétés démontrées pour toutes les méthodes de la famille. D'autre part, cela permet d'apporter à la famille repérée des propriétés et des interrogations nouvelles qui se manifesteraient pour l'algorithme de Newton-min.

Première interprétation. Montrons d'abord que l'algorithme de Newton-min peut être rattaché aux méthodes de Newton non lisses pour résoudre l'équation non lisse (2.5.14).

Commençons par observer que nous pouvons reécrire l'algorithme comme suit

$$x^k = x^{k-1} - J_k^{-1} H(x^{k-1}), \tag{2.5.19}$$

où J_k est la matrice introduite dans (2.5.18). En effet, l'équation (2.5.19) peut aussi s'écrire comme suit

$$J_k(x^k - x^{k-1}) = -\min(x^{k-1}, Mx^{k-1} + q),$$

qui, avec les définitions de A_k et I_k, est équivalent au système suivant

$$\begin{aligned} x^k_{A^k} &= 0, \\ M_{I^k A^k}(x^k - x^{k-1})_{A^k} + M_{I^k I^k}(x^k - x^{k-1})_{I^k} &= -M_{I^k A^k} x^{k-1}_{A^k} - M_{I^k I^k} x^{k-1}_{I^k} - q_{I^k}. \end{aligned}$$

En simplifiant la deuxième équation, on retrouve la seconde relation du système (2.5.17).

Observons maintenant, que la « pseudo-jacobienne » J_k utilisée dans (2.5.19) est un élément du différentiel produit $\partial_\times H(x^k)$ de H en x^k. En effet, on a

$$\partial_\times H(x) = \{(I-D)I + DM : \ D \text{ est diagonale avec } D_{ii} \in [0,1]\}$$

et

$$J_k = (I - D_k)I + D_k M$$

avec $(D_k)_{ii} = 0$ si $i \in A_k$ et $(D_k)_{ii} = 1$ si $i \in I_k$. Dès lors, l'algorithme de Newton-min n'est pas un algorithme de Newton semi-lisse stricto sensu [113, 1993], puisque J_k n'est pas

nécessairement dans le différentiel de Clarke $\partial_C H(x^k)$. On peut toutefois le rattacher à la famille des algorithmes de Newton non lisses dans le sens où la « pseudo-jacobienne » J_k est dans un certain différentiel de H en x^k.

Deuxième interprétation. Si au lieu de (2.5.14), on prend le système dédoublé équivalent [48]

$$H_2(x,y) \equiv \begin{pmatrix} Mx+q-y \\ \min(x,y) \end{pmatrix} = 0, \tag{2.5.20}$$

l'algorithme de Newton-min est un algorithme de Newton semi-lisse stricto sensu pour ce système. En effet, le sous-différentiel de Clarke de H_2 s'écrit

$$\partial_C H_2(x,y) = \mathrm{co}\left\{ \begin{pmatrix} M & -I \\ I_A & 0 \\ 0 & I_{A^c} \end{pmatrix} : \{i : x_i < y_i\} \subset A \subset \{i : x_i \leqslant y_i\} \right\}.$$

L'intérêt du dédoublement des variables a donc été d'introduire une non-différentiabilité $(x,y) \mapsto \min(x,y)$ dont on peut aisément calculer le sous-différentiel de Clarke, alors que le calcul est malaisé pour H. En introduisant la variable auxiliaire $y^k := Mx^k + q$ et en prenant

$$J_k^2 := \begin{pmatrix} M & -I \\ I_{A_k} & 0 \\ 0 & I_{A_k^c} \end{pmatrix} \in \partial_C H_2(x^k, y^k),$$

avec A_k donné comme à l'étape 2 de l'algorithme de Newton-min, l'itération définissant $z^{k+1} := (x^{k+1}, y^{k+1})$ à partir de $z^k := (x^k, y^k)$ par

$$J_k^2(z^{k+1} - z^k) = -H_2(z^k)$$

peut-être considéré comme une itération de l'algorithme de Newton semi-lisse. Il est facile de vérifier que la partie en x de cette itération est aussi celle de l'algorithme de Newton-min.

Troisième interprétation. Comme la fonction min est une fonction différentiable par morceaux, on peut aussi voir l'algorithme de Newton-min comme un algorithme de Newton sur une fonction $\mathscr{C}^1$ par morceaux. En effet, on peut montrer facilement que J_k défini en (2.5.18) est la matrice jacobienne des fonctions actives (voir la définition 2.53) de la fonction H au point x_k.

Propriétés connues de l'algorithme

Comme le nombre de partitions (A, I) est fini (mais exponentiel), alors ou bien l'algorithme converge en un nombre fini d'itérations, ou bien il diverge. Par ailleurs, les prorpiétés suivantes ont été démontrées.

- Si M est une **M**-matrice, l'algorithme converge en au plus n itérations, quel que soit l'itéré initial [71, 2004, théorème 3.2].
- Si M est une petite perturbation d'une **M**-matrice, l'algorithme converge en un nombre fini d'itérations, quel que soit l'itéré initial [61, 1993, théorème 3.4].
- Si M est une **P**-matrice, l'algorithme converge en une d'itération, si l'itéré initial est suffisamment proche de la solution [61, 1993, théorème 3.1].

Chapitre 3

Nonconvergence of the plain Newton-min algorithm for linear complementarity problems with a P-matrix

Dans ce chapitre, nous nous intéressons à l'algorithme Newton-min, utilisé pour résoudre le problème de complémentarité linéaire (PCL) $0 \leqslant x \perp (Mx+q) \geqslant 0$ qui peut être interprété comme un algorithme de Newton non lisse sans globalisation cherchant à résoudre le système d'équations linéaires par morceaux $\min(x, Mx+q) = 0$, qui est équivalent au PCL. Lorsque M est une **M**-matrice d'ordre n, on sait que l'algorithme converge en au plus n itérations. Nous montrons dans ce chapitre que ce résultat ne tient plus lorsque M est une **P**-matrice d'ordre $n \geqslant 3$; l'algorithme peut en effet cycler dans ce cas. On a toutefois la convergence de l'algorithme pour une **P**-matrice d'ordre 1 ou 2. Les résultats de ce chapitre sont repris de l'article [10], publié dans *Mathematical Programming*, que nous reproduisons ci-dessous.

3.1 Introduction

The linear complementarity problem (LCP) consists in finding a vector $x \geqslant 0$ with n components such that $Mx+q \geqslant 0$ and $x^\top(Mx+q) = 0$. Here M is a real matrix of order n, q is a vector in $\mathbb{R}^n$, the inequalities have to be understood componentwise, and the sign $^\top$ denotes matrix transposition. The LCP is often written in compact form as follows

$$\mathrm{LC}(M,q): \qquad 0 \leqslant x \perp (Mx+q) \geqslant 0.$$

This problem is known to have a unique solution for any $q \in \mathbb{R}^n$ if and only if M is a **P**-matrix [118, 30], i.e., a matrix with positive principal minors : $\det M_{II} > 0$ for all nonempty $I \subset \{1,\ldots,n\}$. We denote by **P** the set of **P**-matrices. Other classes of matrices M intervening in the discussion below are the class **Z** of **Z**-matrices (which have nonpositive off-diagonal elements : $M_{ij} \leqslant 0$ for all $i \neq j$) and the class $\mathbf{M} := \mathbf{P} \cap \mathbf{Z}$ of **M**-matrices (they are called **K**-matrices in [30]).

Since the components of x and $Mx+q$ must be nonnegative in $\mathrm{LC}(M,q)$, their perpendicularity with respect to the Euclidean scalar product required in the problem is equivalent to the nullity of the Hadamard product of the two vectors, that is

$$x \cdot (Mx+q) = 0. \tag{3.1.1}$$

Recall that the Hadamard product $u \cdot v$ of two vectors u and v is the vector having its ith component equal to $u_i v_i$. A point x such that (3.1.1) holds is here called a *node* or is said *to satisfy complementarity*. Since, for a node x, either x_i or $(Mx+q)_i$ vanishes, for all indices i, there are at most 2^n nodes for a *nondegenerate* matrix M, which is a matrix having nonsingular principal submatrices. On the other hand, a point x such that x and $Mx+q$ are nonnegative (resp. positive) is said to be *feasible* (resp. *strictly feasible*). A solution to $\mathrm{LC}(M,q)$ is therefore a feasible node.

Many algorithms have been proposed to solve problem $\mathrm{LC}(M,q)$ [103, 30]. They may be based on pivoting techniques [86, 29], which often suffer from the combinatorial aspect of the problem (i.e., the 2^n possibilities to realize (3.1.1)), on interior point methods, which originate from an algorithm introduced by Karmarkar in linear optimization [74, 1984] (see also [78, 1991] for one of the first accounts on the use of interior point methods to solve linear complementarity problems), and on nonsmooth Newton approaches [38], such as the one considered here. See [103, 30] for other iterative methods.

The algorithm we consider in this paper maintains the complementarity condition (3.1.1) at each iteration, while feasibility is obtained at convergence (when this one occurs). As a result, all the iterates are nodes, except possibly the first one, and the algorithm terminates as soon as it has found a feasible iterate. More specifically, suppose that the current iterate x is a node. The algorithm first defines index sets A^+ and I^+ associated with the next iterate x^+ ; in its simplest form, it takes

$$A^+ := \{i : x_i \leqslant (Mx+q)_i\} \qquad \text{and} \qquad I^+ := \{i : x_i > (Mx+q)_i\}. \tag{3.1.2}$$

Then it computes x^+ by solving the linear system formed of the equations

$$x^+_{A^+} = 0 \qquad \text{and} \qquad (Mx^+ + q)_{I^+} = 0. \tag{3.1.3}$$

To have a well defined algorithm, an assumption on M is necessary so that this system has a solution for any choice of complementary sets A^+ and I^+, and vector $q : M$ must be nondegenerate.

The motivation sustaining the algorithm is that it can be viewed as a semismooth Newton method to solve the system of piecewise linear equations

$$\min(x, Mx+q) = 0, \tag{3.1.4}$$

in which the minimum operator 'min' acts componentwise : $[\min(x,y)]_i = \min(x_i, y_i)$. On the one hand, since, for a and $b \in \mathbb{R}$, $\min(a,b) = 0$ if and only if $0 \leqslant a \perp b \geqslant 0$, the system (3.1.4) is indeed equivalent to problem LC(M,q) (see [80, 1976] and [91, 1977, lemma 2.1] for instance). On the other hand, it is indeed clear that the function in (3.1.4) is differentiable at a point x without *doubly active index* (i.e., without index i such that $x_i = (Mx+q)_i$) and that its Jacobian matrix is the one used in the linear system (3.1.3) ; when there are doubly active indices, the Jacobian used in (3.1.3) is determined by the choice (3.1.2). This description makes it natural to call *Newton-min* the algorithm that updates x by the formulas (3.1.2)–(3.1.3).

Here are some remarks about algorithm (3.1.2)–(3.1.3). First, note that the algorithm has a principle quite different from the one used by an interior point approach, which generates strictly feasible points, while complementarity is obtained at the limit. Note also that if, in the local analysis of the method, it is important to allow the first iterate x^1 not to be a node, in this paper, x^1 will always be assumed to be a node. Finally, observe that a consequence of the fact that the algorithm only generates nodes is that it is equivalent to

say that it converges or that it converges in a finite number of iterations or that it does not cycle (algorithm (3.1.2)–(3.1.3) is a Markov process).

The algorithm sketched above and that we further explore in this paper can be traced back at least to the algorithm (6.2)–(6.4) of Aganagić in [3, 1984], who considers a linear complementarity problem that reads instead $0 \leqslant (Xx) \perp (Yx+q) \geqslant 0$, in which X and Y must be jointly **M**-regular. Here, it is not important to be precise about the meaning of this property of joint **M**-regularity, but to know that if $X = I$, as in $\mathrm{LC}(M,q)$, then the property is equivalent to the **M**-matricity of $Y \equiv M$. When $X = I$ and $Y \equiv M$, the algorithm (6.2)–(6.4) of Aganagić [3] is exactly the algorithm (3.1.2)–(3.1.3) above, which is proven in [3, theorem 6.2] to be *monotonically* (in the sense that $x^k \leqslant x^{k+1}$ for $k \geqslant 2$) and *globally* (i.e., x^1 may be arbitrary) convergent to the unique solution of $\mathrm{LC}(M,q)$, provided $M \in \mathbf{M}$ (in [3], the first iterate x^1 is supposed to be zero, but this assumption is not necessary). Under the conditions that $x^1 = 0$ and $M \in \mathbf{M}$, this algorithm is actually identical to the one proposed and analyzed earlier by Chandrasekaran [21, 1970] [3, remark 2]. Although not presented in that manner, the algorithm proposed by Bergounioux, Ito, and Kunisch [15, 1999] to solve a strongly convex quadratic optimal control problem, under the name of *primal-dual active set strategy* (PDAS), can be viewed as an extension of algorithm (3.1.2)–(3.1.3) to an *infinite* dimensional problem (see [14, 2000] for a comparison with interior point methods); the authors prove the convergence of the algorithm on a discretized version of the problem under some conditions. In [61, 2003], Hintermüller, Ito, and Kunisch consider a quadratic optimization problem with simple bounds, which can actually be written as a linear complementarity problem like $\mathrm{LC}(M,q)$; they establish the equivalence between the PDAS strategy and the semismooth Newton method, i.e., algorithm (3.1.2)–(3.1.3); the algorithm is also shown to be *locally superlinearly* convergent when $M \in \mathbf{P}$, and to be *monotonically* (in the sense that $\sum_i x_i^k \leqslant \sum_i x_i^{k+1}$) and *globally* convergent when M is in the set that we denote here by

$$\mathbf{M}_\varepsilon := \{M : M \text{ is a matrix near an } \mathbf{M}\text{-matrix of the same order}\}. \tag{3.1.5}$$

In $\mathbf{M}_\varepsilon$, the level of proximity to an **M**-matrix is left unprecise (and depends on the considered matrix M), but in the proof of [61, theorem 3.4], this proximity is sufficiently tight to imply $\mathbf{M}_\varepsilon \subset \mathbf{P}$. Another interesting property of algorithm (3.1.2)–(3.1.3), proved by Kanzow [71, 2004], is its convergence in at most n iterations when $M \in \mathbf{M}$ and the first iterate is a node. When applied to algorithm (3.1.2)–(3.1.3), the result of Fischer and Kanzow [48, 1996] shows that a solution to $\mathrm{LC}(M,q)$, if any, is reached in one step, provided

x^1 is sufficiently close to that solution and M is a nondegenerate matrix. To conclude this review of results, we would like to cite the quadratic local convergence of Newton's method for piecewise C^1 functions proved by Kojima and Shindo [79, 1986], which is related to the formulation (3.1.4) of problem LC(M,q).

This paper presents examples of nonconvergence of the Newton-min algorithm when M is a **P**-matrix. These counter-examples hold for the *plain* (or undamped) Newton-min algorithm, i.e., without the use of globalization techniques such as linesearch or trust regions. One may believe that, as a Newton-like method to solve a nonlinear (and nonsmooth) system of equations, this is not a good strategy. We share this opinion, in general. However the algorithm deals with a piecewise linear function, and it has been shown to be convergent without globalization techniques when M is a **P**-matrix sufficiently near an **M**-matrix (see the discussion in the previous paragraph). Therefore, searching for the weakest assumptions for which these convergence properties hold seems to us a valid topic. The examples in this paper show that it is not enough to require the **P**-matricity for M.

The paper is structured as follows. In the next section, we are more specific on the definition of the algorithm, by making it a slightly more flexible than in the description (3.1.2)–(3.1.3) above. Some elementary properties of the algorithm, useful in the sequel, are also given. Section 3.3 describes and analyzes the examples of nonconvergence of the plain Newton-min algorithm with a **P**-matrix, when $n \geqslant 3$. These counter-examples work for both definitions of the algorithm, those of sections 3.1 and 3.2. In them, the algorithm can be forced to cycle and visit p nodes, with a p that can be chosen arbitrarily in $\{3,\ldots,n\}$. We consider successively the cases when n is odd and even, which require a different analysis. These counter-examples readily imply that the convergence radius of the plain Newton-min algorithm for solving LC(M,q) with $M \in \mathbf{P}$, which is known to be > 0 [61, 2003], can be arbitrarily small (corollary 3.10). In section 3.4, the plain Newton-min algorithm is claimed to converge for a **P**-matrix when $n = 1$ or $n = 2$, a crumb of consolation.The paper concludes with the perspective section 3.5.

3.2 The algorithm

The plain Newton-min algorithm described in this section generates points that satisfy the complementarity conditions in (3.1.1), while the nonnegativity conditions in LC(M,q) are satisfied when the solution is reached. The starting point x^1 may or may not satisfy this

complementarity condition.

Algorithm 3.1 (plain Newton-min) Let $x^1 \in \mathbb{R}^n$.

For $k = 2, 3, \ldots$, do the following.

1. If x^{k-1} is a solution to $\mathrm{LC}(M,q)$, stop.
2. Choose complementary index sets $A^k := A_0^k \cup A_1^k$ and $I^k := I_0^k \cup I_1^k$, where

$$\begin{aligned} A_0^k &:= \{i : x_i^{k-1} < (Mx^{k-1}+q)_i\}, \\ A_1^k \subset E^k &:= \{i : x_i^{k-1} = (Mx^{k-1}+q)_i\}, \\ I_0^k &:= \{i : x_i^{k-1} > (Mx^{k-1}+q)_i\}, \\ I_1^k &:= E^k \setminus A_1^k. \end{aligned}$$

3. Determine x^k as a solution to

$$\left\{ \begin{array}{l} x_{A^k}^k = 0 \\ (Mx^k+q)_{I^k} = 0. \end{array} \right. \tag{3.2.1}$$

The algorithm is well defined if M is nondegenerate. This assumption will be generally reinforced by supposing that $M \in \mathbf{P}$. We recall from [42, 30] that

$$M \in \mathbf{P} \qquad \Longleftrightarrow \qquad \text{any } x \text{ verifying } x \cdot (Mx) \leqslant 0 \text{ vanishes.} \tag{3.2.2}$$

Note that algorithm 3.1 is more flexible than the one presented in section 3.1, in that the doubly active indices (those in E^k) can be chosen to belong either to A^k or I^k. If this choice is not entirely determined by the current point x^k, the generated sequence $\{x^k\}$ may no longer be a Markov process. We will not be more precise here, however, since this choice has actually no impact on the counter-examples given in this paper, for which the algorithm never generates iterates with doubly active indices.

In this paper, we always assume that x^1 is a node ; then there are two complementary subsets A^1 and I^1 of $\{1, \ldots, n\}$, such that $x_{A^1}^1 = 0$ and $(Mx^1+q)_{I^1} = 0$. In other contexts, in particular for studying the local convergence of the algorithm [61], it is better not to make this assumption.

By the selection of the index sets in step 2, one certainly has for $k \geqslant 2$:

$$x_{A^k}^{k-1} \leqslant (Mx^{k-1}+q)_{A^k} \qquad \text{and} \qquad x_{I^k}^{k-1} \geqslant (Mx^{k-1}+q)_{I^k}. \tag{3.2.3}$$

Therefore,

$$x_{A^k}^{k-1} \leqslant 0 \qquad \text{and} \qquad (Mx^{k-1}+q)_{I^k} \leqslant 0 \tag{3.2.4}$$

must hold [61, 71]. Indeed, for $i \in A^k$, $x_i^{k-1} \leqslant (Mx^{k-1}+q)_i$ by (3.2.3) and, since either $x_i^{k-1} = 0$ or $(Mx^{k-1}+q)_i = 0$ by complementarity, one necessarily has $x_i^{k-1} \leqslant 0$, which proves the first inequality in (3.2.4). A similar reasoning yields the second inequality in (3.2.4).

We denote by e^i the ith vector of the canonical basis of $\mathbb{R}^n$: its jth component is equal to 1 if $j = i$ and to 0 otherwise.

3.3 Nonconvergence for $n \geqslant 3$

In this section, we show that the plain Newton-min algorithm described in section 3.2 may not converge if M is a **P**-matrix and $n \geqslant 3$. We start in section 3.3.1 with the case when n is odd and provide an example of a **P**-matrix M, a vector q, and a starting point, for which the algorithm makes cycles having n nodes. In section 3.3.2, we consider the case when n is even and construct another class of **P**-matrices, which can be viewed as perturbations of the matrices in the first example, for which a cycle having n nodes is also possible. We conclude in section 3.3.3 by constructing examples for which the plain Newton-min algorithm makes cycles visiting p nodes, with p arbitrary in $\{3,\dots,n\}$.

3.3.1 Cycles with an odd number of nodes

In this section, we show the nonconvergence of the plain Newton-min algorithm for problems having the following features.

Example 3.2 Let $n \geqslant 2$. The matrix $M \in \mathbb{R}^{n\times n}$ and the vector $q \in \mathbb{R}^n$ are given by

$$M = \begin{pmatrix} 1 & & & \alpha \\ \alpha & \ddots & & \\ & \ddots & 1 & \\ & & \alpha & 1 \end{pmatrix} \qquad \text{and} \qquad q = 1,$$

where 1 denotes the vector of all ones and the elements of M that are not represented are zeros. More precisely, $M_{ij} = 1$ if $i = j$, $M_{ij} = \alpha$ if $i = (j \bmod n) + 1$, and $M_{ij} = 0$ otherwise. Since $q \geqslant 0$, a solution to $\mathrm{LC}(M,q)$ for these data is $x = 0$. □

The matrix M and the vector q in example 3.2 have already been used by Morris [100] (in that paper, $\alpha = 2$, M is the transpose of the one here, and $q = -1$), although we arrived at them in a different manner, as explained in section 3.4. For completeness and precision, we start by studying the **P**-matricity of the matrix M in example 3.2 (the result for n odd and $\alpha = 2$ was already claimed in [100] without a detailed proof).

Lemma 3.3 *Consider the matrix M in example 3.2. If n is even, $M \in \mathbf{P}$ if and only if $|\alpha| < 1$. If n is odd, $M \in \mathbf{P}$ if and only if $\alpha > -1$.*

PROOF. Observe first that if $\alpha \leqslant -1$, the nonzero vector $x = 1$ is such that $x \cdot (Mx) = (1+\alpha)1 \leqslant 0$; hence $M \notin \mathbf{P}$.

Observe next that $M \in \mathbf{P}$ when $-1 < \alpha < 0$. Indeed, let $x \in \mathbb{R}^n$ be such that $x \cdot (Mx) \leqslant 0$ or equivalently

$$x_1(x_1 + \alpha x_n) \leqslant 0, \quad x_2(x_2 + \alpha x_1) \leqslant 0, \quad \ldots, \quad x_n(x_n + \alpha x_{n-1}) \leqslant 0. \tag{3.3.1}$$

If $x_n > 0$, using (3.3.1) from right to left shows that all the components of x are positive and verify

$$0 < x_n \leqslant |\alpha| x_{n-1} \leqslant |\alpha|^2 x_{n-2} \leqslant \cdots |\alpha|^{n-1} x_1 \leqslant |\alpha|^n x_n.$$

Since $|\alpha| < 1$, this is incompatible with $x_n > 0$. Having $x_n < 0$ is not possible either (just multiply x by -1 and use the same argument). Hence $x_n = 0$ and using (3.3.1) from left to right shows that $x = 0$. The **P**-matricity now follows from (3.2.2).

When $\alpha = 0$, M is the identity matrix and is therefore a **P**-matrix.

Suppose now that n is even. If $\alpha \geqslant 1$, the nonzero vector x, defined by $x_i = (-1)^{i+1}$, is such that $x \cdot (Mx) = (1-\alpha)1 \leqslant 0$; hence $M \notin \mathbf{P}$. Let us now show, by contradiction, that $M \in \mathbf{P}$ if $0 < \alpha < 1$, assuming that there is a nonzero x satisfying $x \cdot (Mx) \leqslant 0$, hence that (3.3.1) holds. Then, as above, all the components of x are nonzero and one can assume that $x_n > 0$. Starting with the rightmost inequality of (3.3.1), one obtains by induction for $i = 1, \ldots, n-1$:

$$x_{n-i} \leqslant -\frac{1}{\alpha^i} x_n < 0 \quad (\text{for } i \text{ odd}) \qquad \text{and} \qquad x_{n-i} \geqslant \frac{1}{\alpha^i} x_n > 0 \quad (\text{for } i \text{ even}).$$

Since n is even, $x_1 \leqslant -(1/\alpha^{n-1})x_n < 0$ and, using the first inequality of (3.3.1), $x_n \geqslant -(1/\alpha)x_1 \geqslant (1/\alpha^n)x_n$, which is in contradiction with $0 < \alpha < 1$ and $x_n > 0$.

Suppose finally that n is odd and $\alpha > 0$. Again, for proving that $M \in \mathbf{P}$, we argue by contradiction, assuming that there is a nonzero x such that $x \cdot (Mx) \leqslant 0$, which is equivalent to (3.3.1). As above, one can assume that $x_n > 0$. Starting with the rightmost inequality in (3.3.1), one can specify by induction the sign $\sigma(x_i)$ of the x_i's :

$$\sigma(x_{n-1}) = -1, \quad \sigma(x_{n-2}) = 1, \quad \ldots, \quad \sigma(x_1) = (-1)^{n-1} = 1,$$

since n is odd. Finally, the first inequality in (3.3.1) gives $\sigma(x_n) = -\sigma(x_1) = -1$, in contradiction with $x_n > 0$. Hence $x = 0$ and $M \in \mathbf{P}$ by (3.2.2). □

Recall that e^k denotes the kth basis vector of $\mathbb{R}^n$.

Lemma 3.4 *Suppose that $n \geqslant 2$ and consider problem $\mathrm{LC}(M,q)$ in which M and q are given in example 3.2 with $\alpha > 1$. When applied to that problem $\mathrm{LC}(M,q)$ and started from $x^1 = -e^1$, algorithm 3.1 cycles by visiting in order the n nodes $x^k = -e^k$, $k = 1, \ldots, n$.*

PROOF. The proof proceeds by induction. By assumption, $x^1 = -e^1$.

Suppose now that $x^{k-1} = -e^{k-1}$ for some $k \in \{2, \ldots, n\}$ and let us show that $x^k = -e^k$. There hold

$$x^{k-1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad \text{and} \qquad Mx^{k-1} + q = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ 1-\alpha \\ 1 \\ \vdots \\ 1 \end{pmatrix},$$

where the component -1 (resp. 0) is at position $k-1$ in x^{k-1} (resp. in $Mx^{k-1}+q$). The update rules of algorithm 3.1 and $\alpha > 1$ then imply that $I^k = \{k\}$ and $x^k = -e^k$.

We still have to show that $x^{n+1} = -e^1$. From $x^n = -e^n$, one deduces that

$$x^n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -1 \end{pmatrix} \qquad \text{and} \qquad Mx^n + q = \begin{pmatrix} 1-\alpha \\ 1 \\ \vdots \\ 1 \\ 0 \end{pmatrix}.$$

Again, the update rules of algorithm 3.1 and $\alpha > 1$ show that $I^{n+1} = \{1\}$ and $x^{n+1} = -e^1$. Therefore algorithm 3.1 cycles. □

By combining lemmas 3.3 and 3.4, one can see that algorithm 3.1 may cycle when M is a **P**-matrix of odd order $n \geqslant 3$: this is the case when the algorithm is started at $x^1 = -e^1$ and when M and q are given by example 3.2, with n odd and $\alpha > 1$. When n is even, the condition $\alpha > 1$ used in lemma 3.4 prevents the matrix M from being a **P**-matrix. It is not difficult to show, however, that the algorithm can also cycle when n is even, $n \geqslant 4$, and $M \in \mathbf{P}$, by using the construction given in the proof of proposition 3.9 : the cycle visits an odd number of nodes (hence $< n$).

3.3.2 Cycles with an even number of nodes

The next family of examples is obtained by perturbing the matrices of example 3.2 with a parameter β, in order to construct cycles having n nodes for an even oder **P**-matrix.

Example 3.5 Let $n \geqslant 3$. The matrix $M \in \mathbb{R}^{n\times n}$ and the vector $q \in \mathbb{R}^n$ are given by

$$M = \begin{pmatrix} 1 & & & \beta & \alpha \\ \alpha & \ddots & & & \beta \\ \beta & \ddots & 1 & & \\ & \ddots & \alpha & 1 & \\ & & \beta & \alpha & 1 \end{pmatrix} \qquad \text{and} \qquad q = 1,$$

where the elements of M that are not represented are zeros or, more precisely, for $i,\ j \in$

$\{1,\dots,n\}$:

$$M_{ij} = \begin{cases} 1 & \text{if } i = j \\ \alpha & \text{if } i = (j \bmod n) + 1 \\ \beta & \text{if } i = ((j+1) \bmod n) + 1 \\ 0 & \text{otherwise.} \end{cases}$$

Since $q \geqslant 0$, a solution to $\mathrm{LC}(M,q)$ for these data is $x = 0$. □

Conditions for having an n-node cycle with algorithm 3.1 on problem $\mathrm{LC}(M,q)$ with M and q given by example 3.5 are very simple to express.

Lemma 3.6 *Suppose that $n \geqslant 3$ and consider problem $\mathrm{LC}(M,q)$ in which M and q are given in example 3.5 with $\alpha > 1$ and $\beta < 1$. When applied to that problem $\mathrm{LC}(M,q)$ and started from $x^1 = -e^1$, algorithm 3.1 cycles by visiting in order the n nodes defined by $x^k = -e^k$, $k = 1,\dots,n$.*

PROOF. The proof is quite similar to the one of lemma 3.4 and proceeds by induction. By assumption, $x^1 = -e^1$.

Suppose now that $x^{k-1} = -e^{k-1}$ for some $k \in \{2,\dots,n-1\}$ and let us show that $x^k = -e^k$. there hold

$$x^{k-1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad \text{and} \qquad Mx^{k-1} + q = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ 1-\alpha \\ 1-\beta \\ 1 \\ \vdots \\ 1 \end{pmatrix},$$

where the component -1 (resp. 0) is at position $k-1$ in x^{k-1} (resp. in $Mx^{k-1}+q$). The update rules of algorithm 3.1, $\alpha > 1$, and $\beta < 1$ then imply that $I^k = \{k\}$ and $x^k = -e^k$.

Therefore, by induction

$$x^{n-1} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -1 \\ 0 \end{pmatrix} \qquad \text{and} \qquad Mx^{n-1} + q = \begin{pmatrix} 1-\beta \\ 1 \\ \vdots \\ 1 \\ 0 \\ 1-\alpha \end{pmatrix}.$$

The update rules of algorithm 3.1, $\alpha > 1$, and $\beta < 1$ then imply that $I^n = \{n\}$, so that

$$x^n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -1 \end{pmatrix} \qquad \text{and} \qquad Mx^n + q = \begin{pmatrix} 1-\alpha \\ 1-\beta \\ 1 \\ \vdots \\ 1 \\ 0 \end{pmatrix}.$$

Again, the update rules of algorithm 3.1, $\alpha > 1$, and $\beta < 1$ show that $I^{n+1} = \{1\}$. Therefore, $x^{n+1} = -e^1$ and algorithm 3.1 cycles. □

We have now to examine whether the conditions on α and β given by lemma 3.6 are compatible with the **P**-matricity of the matrix M in example 3.5. Actually, the conditions on α and β ensuring the **P**-matricity of that matrix M are nonlinear and much more complex to write than the simple conditions on α given in lemma 3.3; in particular, their number depends on the dimension n. The shaded regions in figure 3.3.1 show the intersections with the box $[-2,2] \times [-2,2]$ of the sets $\mathscr{P}$ formed of the (α,β) pairs for which the matrix M is a **P**-matrix, when its order is $n = 3$, 4, 5, and 6. Of course, by lemma 3.3, $\mathscr{P}$ contains the set $\{(\alpha,\beta) : -1 < \alpha < 1,\ \beta = 0\}$ when n is even and the set $\{(\alpha,\beta) : -1 < \alpha,\ \beta = 0\}$ when n is odd. It is not clear at this point, however, whether, for any $n \geqslant 3$, these regions will contain points with $\alpha > 1$ and $\beta < 1$, which are the conditions highlighted by lemma 3.6. Lemma 3.8 below shows that this is actually the case since $\mathscr{P}$ always contains the interiors of the nonconvex polyhedron and the ellipse represented in figure 3.3.1, which are independent of n.

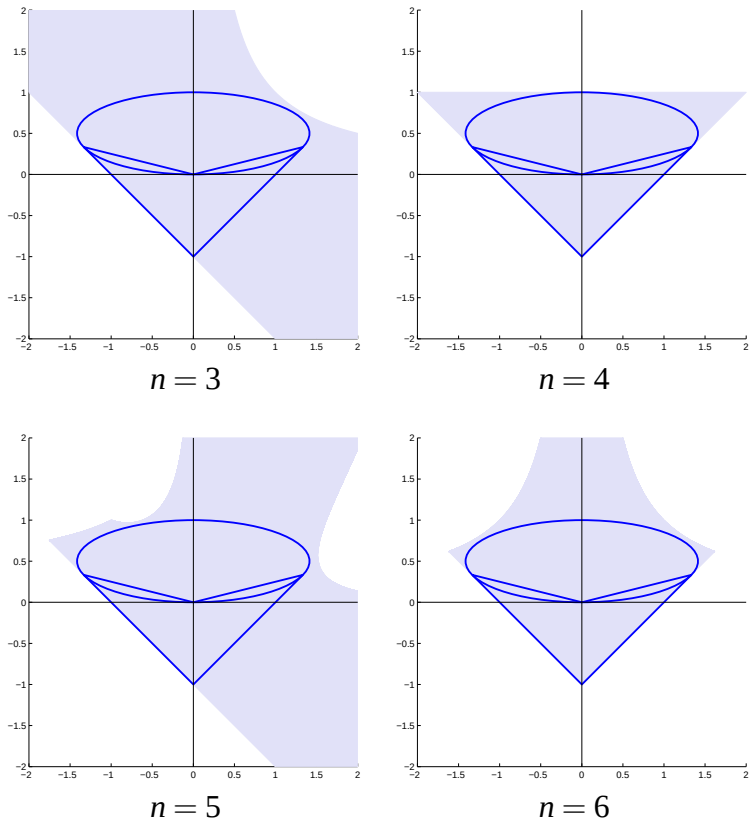


Figure 3.3.1 – The shaded regions are formed of the (α,β) pairs in $[-2,2]\times[-2,2]$ for which the matrix M of example 3.5 is a **P**-matrix, when its order is $n = 3$, 4, 5, and 6; these regions are nonconvex but star-shaped with respect to $(0,0)$, which is the point corresponding to the identity matrix. According to lemma 3.8, the interiors of the represented nonconvex polyhedron and ellipse, which are independent of n, are always contained in these regions, for any $n \geqslant 3$.

Lemma 3.7 *Consider the matrix M in example 3.5. $M \in \mathbf{P}$ only if $1+\alpha+\beta > 0$. If $n = 4k$ for some $k \in \mathbb{N}^*$, $M \in \mathbf{P}$ only if $|\alpha| - 1 < \beta < 1$. If $n = 4k+2$ for some $k \in \mathbb{N}^*$, $M \in \mathbf{P}$ only if $|\alpha| - 1 < \beta$.*

PROOF. Observe first that if $1+\alpha+\beta \leqslant 0$, the nonzero vector $x = 1$ is such that $x \cdot (Mx) = (1+\alpha+\beta)1 \leqslant 0$; hence $M \notin \mathbf{P}$.

Suppose that n is even. If $1-\alpha+\beta \leqslant 0$, the nonzero vector x, defined by $x_i = (-1)^i$, is such that $x \cdot (Mx) = (1-\alpha+\beta)1 \leqslant 0$; hence $M \notin \mathbf{P}$.

Suppose that n is a multiple of 4. If $\beta \geqslant 1$, the nonzero vector x, whose ith component is zero when i is even and takes the value $(-1)^{(i-1)/2}$ when i is odd, is such that $x \cdot (Mx) = (1-\beta)|x| \leqslant 0$; hence $M \notin \mathbf{P}$. □

To prepare the proof of lemma 3.8, we write the circulant matrix M in example 3.5 as follows

$$M = I + \beta J^{n-2} + \alpha J^{n-1},$$

where I denotes the $n \times n$ identity matrix and J is the elementary circulant $n \times n$ matrix

$$J = \begin{pmatrix} 0 & 1 & & 0 \\ & 0 & \ddots & \\ & & \ddots & 1 \\ 1 & & & 0 \end{pmatrix}.$$

More precisely, $J_{ij} = 1$ if $j = (i \bmod n) + 1$ and $J_{ij} = 0$ otherwise. It is well known [54, formula (4.7.10)] that J is diagonalizable on $\mathbb{C}$, meaning that there is a diagonal matrix $D \in \mathbb{C}^{n \times n}$ and a nonsingular matrix $P \in \mathbb{C}^{n \times n}$ such that $J = PDP^{-1}$; in addition its eigenvalues are the nth roots of unity : $D := \mathrm{Diag}(1, w^1, \ldots, w^{n-1})$, where $w = \mathrm{e}^{2\pi i/n}$ and i is the imaginary unit.

Lemma 3.8 *The set of (α, β) pairs ensuring that $M \in \mathbf{P}$ in example 3.5 contains the set $\{(\alpha, \beta) : |\alpha| - 1 < \beta < |\alpha|/4$ or $\alpha^2 + 8(\beta - \frac{1}{2})^2 < 2\}$.*

PROOF. We only have to show that when α and β satisfy

$$|\alpha|-1<\beta<\frac{|\alpha|}{4} \qquad \text{or} \qquad \alpha^2+8\left(\beta-\frac{1}{2}\right)^2<2, \tag{3.3.2}$$

the matrix $M+M^\top$ is positive definite, since then M is clearly a **P**-matrix [75, p. 175].

For any integer p, $(J^p)^\top = J^{n-p}$. Therefore $M+M^\top = 2I+\alpha J+\beta J^2+\beta J^{n-2}+\alpha J^{n-1}$ and, using $J = PDP^{-1}$, we obtain

$$M+M^\top = P\left(2I+\alpha D+\beta D^2+\beta D^{n-2}+\alpha D^{n-1}\right)P^{-1},$$

This identity shows that the eigenvalues of the symmetric matrix $M+M^\top$ are the (necessarily real) numbers

$$\lambda_k := 2+\alpha e^{2k\pi i/n}+\beta e^{4k\pi i/n}+\beta e^{2k(n-2)\pi i/n}+\alpha e^{2k(n-1)\pi i/n},$$

for $k=0,\ldots,n-1$. Using $e^{2kp\pi i/n}+e^{2k(n-p)\pi i/n} = 2\cos(2kp\pi/n)$ (for p integer) and $\cos 2\theta = 2\cos^2\theta-1$, we obtain

$$\lambda_k = 2+2\alpha\cos(2k\pi/n)+4\beta\cos^2(4k\pi/n)-2\beta.$$

We see that the desired positivity of the eigenvalues λ_k depends on the positivity of the following polynomial on $[-1,1]$:

$$t\mapsto\varphi(t) = 2\beta t^2+\alpha t+(1-\beta).$$

We denote by $t_\pm := [-\alpha\pm(\alpha^2-8\beta(1-\beta))^{1/2}]/(4\beta)$ the roots of φ when $\beta\neq 0$ and consider in sequence the three possible cases, identifying in each case conditions that ensure the positivity of φ on $[-1,1]$.

- Case $\beta = 0$. Then φ is positive on $[-1,1]$ if $|\alpha|<1$.
- Case $\beta>0$. There are two subcases.
  - If φ has no real root, i.e., if $\alpha^2<8\beta(1-\beta)$ or equivalently $\alpha^2+8(\beta-\frac{1}{2})^2<2$, then φ is clearly positive on $\mathbb{R}$.
  - If φ has two (possibly equal) real roots $t_\pm$, i.e., if $\alpha^2\geqslant 8\beta(1-\beta)$, then these verify $t_-\leqslant t_+$ and φ is positive on $[-1,1]$,

  - either if $1 < t_-$, which is ensured if $-\alpha - 1 < \beta < -\alpha/4$,
  - or if $t_+ < -1$, which is ensured if $\alpha - 1 < \beta < \alpha/4$.
- Case $\beta < 0$. Then, φ has two real roots $t_\pm$, which verify $t_+ < t_-$, and φ is positive on $[-1,1]$ if both $t_+ < -1$ and $1 < t_-$ holds, which is ensured if $|\alpha| - 1 < \beta < 0$.

By gathering the above conditions, we obtain (3.3.2). □

3.3.3 Nonconvergence

The nonconvergence result is summarized in proposition 3.9, in which it is shown that cycles made of p nodes are possible when $n \geqslant 3$ and $p \in \{3,\dots,n\}$. It is then shown with proposition 3.10 that when $n \geqslant 3$, although the plain Newton-min algorithm is known to converge locally, i.e., when the starting point is near the solution to $\mathrm{LC}(M,q)$, the radius of convergence can be made as small as desired by modifying q ; this makes the convergence of the plain Newton-min algorithm unlikely.

Proposition 3.9 (nonconvergence for $n \geqslant 3$) *When $n \geqslant 3$, algorithm 3.1 may fail to converge when trying to solve $\mathrm{LC}(M,q)$ with a* **P***-matrix M. A cycle made of p nodes is possible, for an arbitrary $p \in \{3,\dots,n\}$.*

PROOF. Since the plain Newton-min algorithm visits only a finite number of nodes, it fails to converge if and only if it cycles.

When $n \geqslant 3$ is odd, a cycle made of n nodes is possible on problem $\mathrm{LC}(M,q)$ with M and q given by example 3.2, and $\alpha > 1$: lemma 3.3 shows that $M \in \mathbf{P}$ and lemma 3.4 shows that a cycle is possible.

When $n \geqslant 4$ is even, a cycle made of n nodes is possible on problem $\mathrm{LC}(M,q)$ with M and q given by example 3.5, and α and β satisfying $\alpha > 1$ and $\alpha - 1 < \beta < \alpha/4$ (hence $\beta < 1/3$) : lemma 3.6 shows that a cycle is possible and lemma 3.8 shows that $M \in \mathbf{P}$.

When $n \geqslant 3$ and $p \in \{3,\dots,n\}$, consider a $p \times p$ matrix $\tilde{M} \in \mathbf{P}$, a vector $\tilde{q} \in \mathbb{R}^p$, and a starting point $\tilde{x}^1 \in \mathbb{R}^p$, such that algorithm 3.1 applied to problem $\mathrm{LC}(\tilde{M},\tilde{q})$ and starting at $\tilde{x}^1$ generates iterates $\tilde{x}^k$ forming a cycle made of p nodes (this is possible by what has just been proven). With obvious notation, define

$$M = \begin{pmatrix} \tilde{M} & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & I_{n-p} \end{pmatrix}, \qquad q = \begin{pmatrix} \tilde{q} \\ 0_{n-p} \end{pmatrix}, \qquad \text{and} \qquad x^1 = \begin{pmatrix} \tilde{x}^1 \\ 0_{n-p} \end{pmatrix}.$$

The **P**-matricity of M is clear, by observing that $x \cdot (Mx) \leqslant 0$ implies $x = 0$. Denote by x^k the kth iterate generated by algorithm 3.1 on $\mathrm{LC}(M,q)$ starting from x^1. Observe first that when an index $i > p$, there holds $x_i^k = 0$, whenever $i \in I^k$ or $i \in A^k$; therefore the generated iterates $x^k \in \mathbb{R}^p \times \{0_{n-p}\}$. Hence, if the same rule as the one used by algorithm 3.1 on $\mathbb{R}^p$ is used to decide whether an index $i \in E^k$ will be considered as being in I^k or A^k, the iterates x^k will be $(\tilde{x}^k, 0_{n-p})$. Obviously, as the $\tilde{x}^k$'s, these iterates also form a cycle made of p nodes. □

To make the above nonconvergence result more concrete, we provide two examples of **P**-matrices M_n, of order $n = 3$ and $n = 4$ respectively, which make algorithm 3.1 fail with an n-cycle, when it starts at $x^1 = (-1, 0, \ldots, 0)$ to solve problem $\mathrm{LC}(M_n, 1)$:

$$M_3 := \begin{pmatrix} 1 & 0 & 2 \\ 2 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix} \qquad \text{and} \qquad M_4 := \begin{pmatrix} 1 & 0 & 1/2 & 4/3 \\ 4/3 & 1 & 0 & 1/2 \\ 1/2 & 4/3 & 1 & 0 \\ 0 & 1/2 & 4/3 & 1 \end{pmatrix}. \tag{3.3.3}$$

We have used the lemmas 3.3 and 3.4 for constructing M_3 and the lemmas 3.6 and 3.8 for designing M_4.

The *convergence radius* of an iterative algorithm for finding a solution $\bar{x}$ to a given problem is the largest $\rho > 0$ such that the algorithm converges to $\bar{x}$ when the initial iterate is in the ball of center $\bar{x}$ and radius ρ. The plain Newton-min algorithm is known to be locally convergent [61, 2003]. By scaling the variables in the previous examples, however, one can make the convergence radius of the plain Newton-min algorithm as small as desired. Even though this observation is straightforward, we highlight it in the following corollary to stress the fact that without modification the plain Newton-min algorithm may have little chance to converge.

Corollary 3.10 (small convergence radius for $n \geqslant 3$) *When $n \geqslant 3$, the convergence radius of algorithm 3.1 to solve $\mathrm{LC}(M,q)$ with a **P**-matrix M may be arbitrarily small.*

PROOF. Take an example of matrix $M \in \mathbf{P}$ and vector $q \geqslant 0$ such that algorithm 3.1 cycles when it tries to solve $\mathrm{LC}(M,q)$ from some nonzero x^1 (this is possible by using one of the problems considered in the proof of proposition 3.9). The convergence radius of algorithm 3.1 for that problem is therefore less than $\|x^1\|$ (the solution is 0).

Now, algorithm 3.1 starting at $\tilde{x}^1 = \varepsilon x^1$, for some $\varepsilon > 0$, for solving $\mathrm{LC}(M, \varepsilon q)$ generates the iterates $\tilde{x}^k = \varepsilon x^k$ and therefore cycles. The convergence radius for the plain Newton-min algorithm on this new problem is less than $\varepsilon \|x^1\|$, which can be made arbitrarily small by letting $\varepsilon \downarrow 0$. □

3.4 Convergence for $n = 1$ or 2

If the plain Newton-min algorithm does not necessary converge for a **P**-matrix M of order $n \geqslant 3$, it does converge when $n = 1$ or 2.

Proposition 3.11 (convergence for $n = 1$ or $n = 2$) *Suppose that M is a* **P***-matrix and that $n = 1$ or $n = 2$. Then the plain Newton-min algorithm converges.*

The proof of this proposition can be found in the full report [9]. It was obtained by highlighting first necessary conditions for having a cycle made of 3 nodes. It is then shown that, when M is a **P**-matrix, the algorithm cannot do cycles made of 2 distinct nodes and that the conditions for making a cycle made of 3 distinct nodes cannot be satisfied when $n = 2$ (when $n = 1$ such a 3-cycle does not exist).

In this section, we prove the convergence of the plain Newton-min algorithm, algorithm 3.1, when M is a **P**-matrix of order 1 or 2. The proof for $n = 1$ is straightforward. The one presented for $n = 2$ is indirect but highlights the origin of the counter-example 3.2 and allows us to present some properties of the plain Newton-min algorithm.

The convergence of the plain Newton-min algorithm when $n = 1$ is a direct consequence of the following reassuring and elementary property, already proven in [15, theorem 2.1] in the case where the complementarity problem expresses the optimality conditions of an infinite dimensional quadratic optimization problem.

Lemma 3.12 (stagnation at a solution) *Suppose that M is nondegenerate. Then, a node is a solution to $\mathrm{LC}(M, q)$ if and only if the plain Newton-min algorithm, without its stopping test in step 1, starting at that node, takes the same node as the next iterate.*

PROOF. Let x be a node of problem $\mathrm{LC}(M,q)$. We denote by A^+ and I^+ the index sets determined in step 2 of algorithm 3.1 and by x^+ the next iterate.

If x is a solution, then $x \geqslant 0$ and $Mx+q \geqslant 0$. There hold $x^+_{A^+} = 0$ by (3.2.1) and $x_{A^+} = 0$ by (3.2.4) and the nonnegativity of x, so that $x^+_{A^+} = x_{A^+}$. Similarly, there hold $(Mx^+ + q)_{I^+} = 0$ by (3.2.1) and $(Mx+q)_{I^+} = 0$ by (3.2.4) and the nonnegativity of $Mx+q$, so that $0 = (M(x^+-x))_{I^+} = M_{I^+I^+}(x^+-x)_{I^+}$ [since $(x^+-x)_{A^+} = 0$] and therefore $(x^+-x)_{I^+} = 0$ by the nonsingularity of $M_{I^+I^+}$. We have shown that $x^+ = x$.

Conversely, assume that $x^+ = x$. By (3.2.3), there hold $x_{A^+} \leqslant (Mx+q)_{A^+}$ and $x_{I^+} \geqslant (Mx+q)_{I^+}$. Since $x_{A^+} = x^+_{A^+} = 0$ [by (3.2.1)] and $(Mx+q)_{I^+} = (Mx^+ + q)_{I^+} = 0$ [by (3.2.1)], we get $x \geqslant 0$ and $Mx+q \geqslant 0$; hence x is a solution. □

We consider now the case when $n = 1$ and M is nondegenerate. For such a trivial problem, a direct proof of convergence is obviously possible. The proof given below uses instead lemma 3.12, which is nevertheless useful elsewhere.

Proposition 3.13 (convergence for $n = 1$) *Suppose that $n = 1$, that M is nondegenerate $(M \neq 0)$, and that $\mathrm{LC}(M,q)$ has a solution. Then the plain Newton-min algorithm converges.*

PROOF. Without restriction, it can be assumed that the first iterate x^1 is a node. If x^1 is a solution, the algorithm stops at that point (by step 1 of algorithm 3.1). If x^1 is not the solution, there is another node that is solution (by assumption). Since there are no more than 2 nodes (since $n = 1$), the algorithm takes the solution as the next iterate (by lemma 3.12) and stops there. □

Since a **P**-matrix is nondegenerate, the elementary preceding result applies when $M \in \mathbf{P}$. When $M \neq 0 \in \mathbb{R}$ but $\mathrm{LC}(M,q)$ has no solution, the algorithm cycles by visiting alternatively the 2 nodes of the problem, i.e., $x = -q/M < 0$ and $x = 0$.

We start the study of the convergence of the plain Newton-min algorithm when $n = 2$ by showing that the algorithm cannot do cycles made of 2 nodes (lemma 3.14). We have seen with examples 3.2 and 3.5 that the algorithm can do a 3-cycle when $n \geqslant 3$, but this implies some conditions that are highlighted by lemma 3.15. We finally show that these conditions cannot be satisfied when $n = 2$ and, as a result, that the algorithm must converge (proposition 3.16).

Lemma 3.14 (no 2-cycle) *If M is a* **P***-matrix, then the plain Newton-min algorithm does not make cycles formed of* 2 *distinct nodes.*

PROOF. We argue by contradiction, assuming that the algorithm visits in order the following nodes $x^1 \to x^2 \to x^1$, with $x^1 \neq x^2$. Since the algorithm goes from x^1 to x^2 and from x^2 to x^1, the definition of the sets A^k and I^k in step 2 of the algorithm implies that

$$x^1_{A^2} \leqslant (Mx^1+q)_{A^2} \qquad \text{and} \qquad x^1_{I^2} \geqslant (Mx^1+q)_{I^2}, \tag{3.4.1}$$
$$x^2_{A^1} \leqslant (Mx^2+q)_{A^1} \qquad \text{and} \qquad x^2_{I^1} \geqslant (Mx^2+q)_{I^1}. \tag{3.4.2}$$

After possible rearrangement of the component order, we get

$$x^2-x^1 = \begin{pmatrix} 0_{A^1\cap A^2} \\ x^2_{A^1\cap I^2} \\ 0_{I^1\cap A^2} \\ x^2_{I^1\cap I^2} \end{pmatrix} - \begin{pmatrix} 0_{A^1\cap A^2} \\ 0_{A^1\cap I^2} \\ x^1_{I^1\cap A^2} \\ x^1_{I^1\cap I^2} \end{pmatrix} = \begin{pmatrix} 0_{A^1\cap A^2} \\ x^2_{A^1\cap I^2} \\ -x^1_{I^1\cap A^2} \\ (x^2-x^1)_{I^1\cap I^2} \end{pmatrix}.$$

Observe that the components of x^2-x^1 with indices in $A^1\cap I^2$ are nonpositive since $x^2_{A^1\cap I^2} \leqslant (Mx^2+q)_{A^1\cap I^2}$ [by $(3.4.2)_1$] $=0$ [by $(3.2.1)_2$] and that the components with indices in $I^1\cap A^2$ are nonnegative since $-x^1_{I^1\cap A^2} \geqslant -(Mx^1+q)_{I^1\cap A^2}$ [by $(3.4.1)_1$] $=0$ [by $(3.2.1)_2$]. On the other hand, by $(3.2.1)_2$, there holds

$$M(x^2-x^1) = \begin{pmatrix} (Mx^2)_{A^1\cap A^2} \\ -q_{A^1\cap I^2} \\ (Mx^2)_{I^1\cap A^2} \\ -q_{I^1\cap I^2} \end{pmatrix} - \begin{pmatrix} (Mx^1)_{A^1\cap A^2} \\ (Mx^1)_{A^1\cap I^2} \\ -q_{I^1\cap A^2} \\ -q_{I^1\cap I^2} \end{pmatrix} = \begin{pmatrix} (M(x^2-x^1))_{A^1\cap A^2} \\ -(Mx^1+q)_{A^1\cap I^2} \\ (Mx^2+q)_{I^1\cap A^2} \\ 0 \end{pmatrix}.$$

In this vector, the components with indices in $A^1\cap I^2$ are nonnegative since $-(Mx^1+q)_{A^1\cap I^2} \geqslant -x^1_{A^1\cap I^2}$ [by $(3.4.1)_2$] $=0$ [by $(3.2.1)_1$] and the components with indices in $I^1\cap A^2$ are nonpositive since $(Mx^2+q)_{I^1\cap A^2} \leqslant x^2_{I^1\cap A^2}$ [by $(3.4.2)_2$] $=0$ [by $(3.2.1)_1$]. Therefore

$$(x^2-x^1)\cdot M(x^2-x^1) \leqslant 0.$$

Since $M\in\mathbf{P}$, there holds $x^1=x^2$ by (3.2.2), contradicting the initial assumption. □

Lemma 3.15 (necessary conditions for a 3-cycle) *Suppose that M is a* **P***-matrix and that the plain Newton-min algorithm cycles by visiting in order the three distinct nodes $x^1 \to x^2 \to x^3$. Then the following three sets of indices must be nonempty*

$$\begin{cases} (A^1 \cap I^2 \cap I^3) \cup (I^1 \cap A^2 \cap A^3) \\ (I^1 \cap A^2 \cap I^3) \cup (A^1 \cap I^2 \cap A^3) \\ (I^1 \cap I^2 \cap A^3) \cup (A^1 \cap A^2 \cap I^3). \end{cases} \tag{3.4.3}$$

PROOF. We use the same technique as in the proof of lemma 3.14. Since the algorithm goes from x^1 to x^2 and from x^2 to x^3, one has from step 2 of the algorithm that

$$x^1_{A^2} \leqslant (Mx^1+q)_{A^2} \qquad \text{and} \qquad x^1_{I^2} \geqslant (Mx^1+q)_{I^2} \tag{3.4.4}$$

$$x^2_{A^3} \leqslant (Mx^2+q)_{A^3} \qquad \text{and} \qquad x^2_{I^3} \geqslant (Mx^2+q)_{I^3}. \tag{3.4.5}$$

Using $(3.2.1)_1$, there holds

$$x^2 - x^1 = \begin{pmatrix} 0_{A^1\cap A^2\cap A^3} \\ 0_{A^1\cap A^2\cap I^3} \\ x^2_{A^1\cap I^2\cap A^3} \\ x^2_{A^1\cap I^2\cap I^3} \\ 0_{I^1\cap A^2\cap A^3} \\ 0_{I^1\cap A^2\cap I^3} \\ x^2_{I^1\cap I^2\cap A^3} \\ x^2_{I^1\cap I^2\cap I^3} \end{pmatrix} - \begin{pmatrix} 0_{A^1\cap A^2\cap A^3} \\ 0_{A^1\cap A^2\cap I^3} \\ 0_{A^1\cap I^2\cap A^3} \\ 0_{A^1\cap I^2\cap I^3} \\ x^1_{I^1\cap A^2\cap A^3} \\ x^1_{I^1\cap A^2\cap I^3} \\ x^1_{I^1\cap I^2\cap A^3} \\ x^1_{I^1\cap I^2\cap I^3} \end{pmatrix} = \begin{pmatrix} 0_{A^1\cap A^2\cap A^3} \\ 0_{A^1\cap A^2\cap I^3} \\ x^2_{A^1\cap I^2\cap A^3} \\ x^2_{A^1\cap I^2\cap I^3} \\ -x^1_{I^1\cap A^2\cap A^3} \\ -x^1_{I^1\cap A^2\cap I^3} \\ (x^2-x^1)_{I^1\cap I^2\cap A^3} \\ (x^2-x^1)_{I^1\cap I^2\cap I^3} \end{pmatrix} \cdot \begin{matrix} [0] \\ [0] \\ [-] \\ [+] \\ [+] \\ [+] \end{matrix}$$

The extra column on the right gives the sign of each component, when appropriate ; this one is deduced by arguments similar to those in the proof of lemma 3.14, that is

$$x^2_{A^1\cap I^2\cap A^3} \leqslant (Mx^2+q)_{A^1\cap I^2\cap A^3} = 0 \qquad [\text{by } (3.4.5)_1 \text{ and } (3.2.1)_2]$$

$$x^2_{A^1\cap I^2\cap I^3} \geqslant (Mx^2+q)_{A^1\cap I^2\cap I^3} = 0 \qquad [\text{by } (3.4.5)_2 \text{ and } (3.2.1)_2]$$

$$x^1_{I^1\cap A^2} \leqslant (Mx^1+q)_{I^1\cap A^2} = 0 \qquad [\text{by } (3.4.4)_1 \text{ and } (3.2.1)_2].$$

On the other hand, using $(3.2.1)_2$, there holds

The sign of the components given in the extra column on the right is justified as follows :

$$\begin{aligned}
&(Mx^1+q)_{A^1\cap I^2} \leqslant x^1_{A^1\cap I^2} = 0 && \text{[by } (3.4.4)_2 \text{ and } (3.2.1)_1\text{]}\\
&(Mx^2+q)_{I^1\cap A^2\cap A^3} \geqslant x^2_{I^1\cap A^2\cap A^3} = 0 && \text{[by } (3.4.5)_1 \text{ and } (3.2.1)_1\text{]}\\
&(Mx^2+q)_{I^1\cap A^2\cap I^3} \leqslant x^2_{I^1\cap A^2\cap I^3} = 0 && \text{[by } (3.4.5)_2 \text{ and } (3.2.1)_1\text{]}.
\end{aligned}$$

Taking the Hadamard product of the two vectors now gives

$$(x^2-x^1)\cdot M(x^2-x^1) = \begin{pmatrix}
0_{A^1\cap A^2\cap A^3} & [0]\\
0_{A^1\cap A^2\cap I^3} & [0]\\
-x^2_{A^1\cap I^2\cap A^3}\cdot(Mx^1+q)_{A^1\cap I^2\cap A^3} & [-]\\
-x^2_{A^1\cap I^2\cap I^3}\cdot(Mx^1+q)_{A^1\cap I^2\cap I^3} & [+]\\
-x^1_{I^1\cap A^2\cap A^3}\cdot(Mx^2+q)_{I^1\cap A^2\cap A^3} & [+]\\
-x^1_{I^1\cap A^2\cap I^3}\cdot(Mx^2+q)_{I^1\cap A^2\cap I^3} & [-]\\
0_{I^1\cap I^2\cap A^3} & [0]\\
0_{I^1\cap I^2\cap I^3} & [0]
\end{pmatrix},$$

where the extra column on the right gives the sign of each components. Therefore, if the index set $(A^1\cap I^2\cap I^3)\cup(I^1\cap A^2\cap A^3)$ were empty, we would have $(x^2-x^1)\cdot M(x^2-x^1)\leqslant 0$, which, with the **P**-matricity of M and (3.2.2), would imply that $x_2 = x_1$, in contradiction with the initial assumption. We have proven that the first index set in (3.4.3) is nonempty.

To show that the second index set in (3.4.3) is nonempty, we use the fact that the algorithm goes from x^2 to x^3 and from x^3 to x^1. Therefore, by cycling the indices in the result just obtained, we see that $(A^2\cap I^3\cap I^1)\cup(I^2\cap A^3\cap A^1)\neq\varnothing$; this corresponds to the second index set in (3.4.3). By cycling the indices again, we obtain $(A^3\cap I^1\cap I^2)\cup(I^3\cap A^1\cap A^2)\neq\varnothing$; this corresponds to the third index set in (3.4.3). □

Example 3.2 was actually obtained for $n = 3$, by forcing the 3 sets in (3.4.3) to be nonempty by setting

$$I^1\cap A^2\cap A^3 = \{1\},\quad A^1\cap I^2\cap A^3 = \{2\},\quad \text{and}\quad A^1\cap A^2\cap I^3 = \{3\},$$

which yields $I^k = \{k\}$.

Proposition 3.16 (convergence for $n = 2$) *Suppose that M is a* **P***-matrix and that $n = 2$. Then the plain Newton-min algorithm converges.*

PROOF. We know that the algorithm converges if it does not make a p-cycle, a cycle made of $p \geqslant 2$ distinct nodes. It cannot make a 2-cycle by lemma 3.14. By lemma 3.15, to make a 3-cycle, the three sets in (3.4.3) must be nonempty; but these sets are disjoint; since $n = 2$, one of them must be empty, so that the algorithm does not make a 3-cycle. Therefore, the algorithm will visit a 4th node if it has not found the solution on the first 3 visited nodes. This last node is then the solution, since the solution exist ($M \in \mathbf{P}$) and there are at most $2^n = 4$ distinct nodes ($M \in \mathbf{P}$). □

3.5 Perspectives

The work presented in this paper can be pursued along at least two directions. One possibility is to better mark out the set of matrices for which the plain Newton-min method converges. According to [61, 2003, theorem 3.4] and the counter-examples of the present paper, this set contains the set $\mathbf{M}_\varepsilon$ of matrices sufficiently near an **M**-matrix but not all of the larger set **P** of **P**-matrices, see the discussion around (3.1.5). Such a study may result in the identification of an analytically well defined set of matrices or it may be a long process with an endless refinement. In the first case, it would be nice to see whether membership in that new matrix class can be determined in polynomial time, knowing that recognizing a **P**-matrix is a co-NP-complete problem [31, 124].

Another possibility is to modify the algorithm to force its convergence for **P**-matrices or an even larger class of matrices. Being able to deal with **P**-matrices is important for at least three reasons. First, linear complementarity problems with a **P**-matrix are encountered in practice [119, 2004]. Next, this is exactly the class of matrices that ensure the existence and uniqueness of the solution to the LCP [118, 30], which forces us to pay attention to these matrices. Finally, the possibility to find a polynomial algorithm to solve the LCP with a **P**-matrix still seems to be an open question. Some authors argue that such an algorithm might exist; see Morris [100, 2002], who refers to a contribution by Megiddo [97, 1988], himself citing an unpublished related note of Solow, Stone and Tovey [122, 1987]. The possibility that a modified version of the plain Newton-min algorithm might have

the desired polynomiality property cannot be excluded. A natural remedy would be to add a globalization technique (linesearch or trust regions, see [28, 17] for example) to the plain Newton-min algorithm in order to force its convergence ; see [59, 64, 1990, 2009], for contributions along that direction. This globalization is not straightforward since the Newton-min direction may not be a descent direction of the standard ℓ_2 merit function associated with (3.1.4) [12].

Acknowledgments

We would like to thank P. Knabner and S. Kräutle for having drawn our attention on the Newton-min algorithm and its efficiency on practical problems [19, 81, 2009, 2010], and M. Hintermüller, Ch. Kanzow, N. Metla [98, 2008], and A. Griewank for various exchanges of mails and conversations on the topics. Also, we wish to thank the two referees for their helpful comments. This work was partially supported by the MoMaS group (PACEN/CNRS, ANDRA, BRGM, CEA, EDF, IRSN).

Chapitre 4

An algorithmic characterization of P-matricity

Nous montrons dans ce chapitre qu'une matrice M est une **P**-matrice si, et seulement si, quel que soit le vecteur q, l'algorithme de Newton-min ne fait pas de cycle de deux points lorsqu'il est utilisé pour résoudre le problème de complémentarité linéaire $0 \leqslant x \perp (Mx+q) \geqslant 0$. Les résultats de ce chapitre sont repris de l'article [11], en révision pour publication dans *SIAM Journal on Matrix Analysis and Applications*.

4.1 Introduction

Being given a positive integer n, a matrix $M \in \mathbb{R}^{n\times n}$, and a vector $q \in \mathbb{R}^n$, the *linear complementarity problem* consists in determining a vector $x \in \mathbb{R}^n$ such that

$$x \geqslant 0, \qquad Mx+q \geqslant 0, \qquad \text{and} \qquad x^\top(Mx+q) = 0.$$

Inequalities have to be understood componentwise ; for example $x \geqslant 0$ means $x_i \geqslant 0$ for all indices $i \in [\![1,n]\!] := \{1,\dots,n\}$. The Euclidean scalar product of two vectors u and v is denoted by $u^\top v = \sum_i u_i v_i$. This problem is sometimes written in compact form as follows

$$\mathrm{LC}(M,q) : \qquad 0 \leqslant x \perp (Mx+q) \geqslant 0,$$

where the sign $\perp$ is used to denote the perpendicularity with respect to the Euclidean scalar product.

Let M_{IJ} denote the submatrix of a matrix M formed of its rows with indices in I and its columns with indices in J. An $n \times n$ real matrix M is a **P**-*matrix* if its principal minors are positive : for all $I \subset [\![1,n]\!]$, $\det M_{II} > 0$ (by convention $\det M_{\varnothing\varnothing} = 1$). The class of **P**-matrices is denoted by **P** (the order n of the matrices is implicit and assumed fixed in that notation). These matrices have an eminent role in linear complementarity problems since it can be shown that $M \in \mathbf{P}$ if and only if $\mathrm{LC}(M,q)$ has one and only one solution, whatever is q [118, 30, 1958]. Another characterization of **P**-matricity, which will be useful below, is the following [42, 30, 1962] :

$$M \in \mathbf{P} \qquad \Longleftrightarrow \qquad \text{any } x \text{ verifying } x \cdot (Mx) \leqslant 0 \text{ vanishes,} \tag{4.1.1}$$

where we have denoted by $u \cdot v$ the *Hadamard product* of the vectors u and v, which is a vector whose ith component is $u_i v_i$.

There are many other equivalent conditions for a matrix to be in **P**, than the three given above [3, 63, 117, 30, 116]. This paper gives still another characterization of **P**-matricity, which is in terms of a property of the algorithm for solving $\mathrm{LC}(M,q)$ that is described in section 4.2, the Newton-min algorithm. We show that $M \in \mathbf{P}$ if and only if the Newton-min algorithm does not cycle between two distinct points, whatever is q, when it is used to solve $\mathrm{LC}(M,q)$.

4.2 The Newton-min algorithm

Let I be a subset of $[\![1,n]\!]$. We denote by $I^c := [\![1,n]\!] \setminus I$ the complementary set of I in $[\![1,n]\!]$ and by $|I|$ the cardinality of I. For a vector $v \in \mathbb{R}^n$, v_I is the vector in $\mathbb{R}^{|I|}$ whose components are the components v_i's of v with index $i \in I$.

The *Newton-min algorithm* is a short name for the nonsmooth Newton method [99, 113, 125, 1977-1993] for solving the nonsmooth piecewise linear equation

$$\min(x, Mx+q) = 0,$$

which is equivalent to $\mathrm{LC}(M,q)$ [80, 91, 1976-1977]. More specifically, at the current iterate $x \in \mathbb{R}^n$, the algorithm determines a set of indices

$$I := \{i \in [\![1,n]\!] : x_i > (Mx+q)_i\} \tag{4.2.1}$$

and computes the next iterate x^+ as the unique solution to the system

$$x^+_{I^c} = 0 \qquad \text{and} \qquad (Mx^+ + q)_I = 0. \tag{4.2.2}$$

The uniqueness of the solution of the system (4.2.2) is certainly ensured if M is *nondegenerate*, meaning that the principal minors of M do not vanish (we denote by $\mathbf{ND}$ the set of nondegenerate matrices), so that this wellposedness condition of the algorithm will always be assumed. In that case,

$$x^+_I = -M_{II}^{-1} q_I.$$

A more general form of the algorithm puts also in I some of the indices i such that $x_i = (Mx+q)_i$ (see section 2 in [11] for instance), but we do not consider this version of the algorithm below; numerically, the difference is not essential and the present version is more natural (the linear systems to solve have a smaller size $|I|$). The Newton-min algorithm can be traced back at least to Aganagić in [3, 1984]; see paragraph 7 of the introduction of [11] for more details on its origin and its developments.

By definition (see (4.2.2)), except possibly for the initial iterate, the Newton-min algorithm only visits points x satisfying $x \cdot (Mx+q) = 0$ or equivalently

$$x_{I^c} = 0 \qquad \text{and} \qquad (Mx+q)_I = 0, \tag{4.2.3}$$

for some (possibly empty) index set $I \subset [\![1,n]\!]$. The point x satisfying (4.2.3) is denoted by

$$x^{(I)}$$

and is called a *node*. Clearly $x^{(\varnothing)} = 0$. Since there are 2^n different index sets I, there are *at most* 2^n nodes (two different index sets may yield the same node; for example, there is a single node, zero, if and only if $q = 0$).

Since the Newton-min algorithm is a *Markov process* in x (i.e., the next iterate only depends on the current one) and only visits nodes and since the number of nodes is finite, either the algorithm converges or it cycles by visiting a finite number of distinct nodes repetitively. The identification of the conditions of convergence of the Newton-min algorithm may, therefore, go through the analysis of the conditions that prevent cycles from occurring. The case of the 2-cycles is considered in the next section (for $k \geqslant 2$, a *k-cycle* is a cycle made of k distinct nodes). Let us formalize this a bit more.

For $k \geqslant 2$, we denote by $\mathbf{NM}_k$ the set of nondegenerate matrices $M \in \mathbb{R}^{n\times n}$ such that the Newton-min algorithm does not make k-cycles when it is used to solve $\mathrm{LC}(M,q)$, whatever is q. Therefore, for the reasons given at the beginning of the previous paragraph,

$$\mathbf{NM} := \bigcap_{k\geqslant 2} \mathbf{NM}_k \tag{4.2.4}$$

is the class of nondegenerate matrices $M \in \mathbb{R}^{n\times n}$ such that the Newton-min algorithm converges, whatever is q and the initial point. In this paper, we prove that

$$\mathbf{NM}_2 = \mathbf{P}.$$

Since $\mathbf{NM}$ is included in $\mathbf{NM}_2$, this identity implies in particular that

$$\mathbf{NM} \subset \mathbf{P}, \tag{4.2.5}$$

i.e., the set of matrices ensuring the convergence of the Newton-min algorithm, whatever is q and the initial point, is contained in $\mathbf{P}$. It has been shown in [9, 11, 2009] that $\mathbf{P} \subset \mathbf{NM}$ if $n = 1$ or $n = 2$ (hence $\mathbf{P} = \mathbf{NM}$ in that case) and that $\mathbf{P} \nsubseteq \mathbf{NM}$ if $n \geqslant 3$ (hence the inclusion (4.2.5) is strict in that case).

We recall that an $\mathbf{M}$-*matrix* is a $\mathbf{P}$-matrix with nonpositive off-diagonal elements ($M_{ij} \leqslant 0$ when $i \neq j$). According to [3, 1984, theorem 6.2],

$$\mathbf{M} \subset \mathbf{NM}.$$

It is also known that $M \in \mathbf{NM}$ if M is sufficiently close to an $\mathbf{M}$-matrix [61, 2003].

4.3 A dual characterization of the absence of 2-cycle

The Newton-min algorithm makes a 2-*cycle* if it generates successively the iterates

$$x^{(I)} \to x^{(J)} \to x^{(I)} \to x^{(J)} \to \cdots,$$

for distinct nodes $x^{(I)} \neq x^{(J)}$ associated with index sets I and $J \subset [\![1,n]\!]$. The dual characterization of the absence of 2-cycle of the Newton-min algorithm given in this section is a first step in the derivation of our main result (theorem 4.4).

We start by recalling Motzkin's theorem of the alternative (see [56, 2010, theorem 3.15 or 7.17] for instance), on which the dual characterization of proposition 4.2 rests.

Lemma 4.1 (Motzkin) *Let $A \in \mathbb{R}^{m_A \times n}$ and $B \in \mathbb{R}^{m_B \times n}$ be two matrices with the same number of columns. Then, there is a vector $x \in \mathbb{R}^n$ satisfying*

$$Ax < 0 \qquad \text{and} \qquad Bx \leqslant 0 \tag{4.3.1}$$

if and only if there is no vector $\alpha \in \mathbb{R}^{m_A}_+ \setminus \{0\}$ and $\beta \in \mathbb{R}^{m_B}_+$ such that

$$A^\top \alpha + B^\top \beta = 0. \tag{4.3.2}$$

Strict inequalities on vectors must also be understood componentwise; hence $Ax < 0$ in (4.3.1) means $(Ax)_i < 0$ for all $i \in [\![1, m_A]\!]$. The (components of the) vectors α and β in the lemma will be called *Motzkin multipliers* below. There is one such scalar multiplier for each inequality in (4.3.1).

The next proposition gives a dual condition on the matrix M such that the Newton-min algorithm does not cycle between two given nodes when it is used to solve $\mathrm{LC}(M,q)$, whatever is the vector q. The dual aspect of condition (4.3.3) comes from lemma 4.1 and therefore involves, like in (4.3.2), the transpose of (modified) submatrices of M. The *symmetric difference* of two index sets I and $J \subset [\![1,n]\!]$ is denoted by

$$I \triangle J := (I \cap J^c) \cup (I^c \cap J) = (I \cup J) \setminus (I \cap J).$$

Proposition 4.2 (no cycle $x^{(I)} \to x^{(J)} \to x^{(I)}$) *Suppose that $M \in \mathbf{ND}$ and let be given two different subsets I and $J \subset [\![1,n]\!]$. Then the following conditions are equivalent :*

(*i*) *there is an* $\alpha \in \mathbb{R}_+^{|I \triangle J|} \setminus \{0\}$ *such that*

$$\begin{pmatrix} M_{(I\cap J^c)(I\cap J^c)} & -M_{(I\cap J^c)(I^c\cap J)} \\ -M_{(I^c\cap J)(I\cap J^c)} & M_{(I^c\cap J)(I^c\cap J)} \end{pmatrix}^\top \alpha$$
$$\geqslant \begin{pmatrix} -M_{(I\cap J)(I\cap J^c)} & M_{(I\cap J)(I^c\cap J)} \end{pmatrix}^\top M^{-\mathsf{T}}_{(I\cap J)(I\cap J)} \begin{pmatrix} -M_{(I\cap J^c)(I\cap J)} \\ M_{(I^c\cap J)(I\cap J)} \end{pmatrix}^\top \alpha, \quad (4.3.3)$$

(*ii*) *whatever is q, the Newton-min algorithm does not make the cycle* $x^{(I)} \to x^{(J)} \to x^{(I)}$ *when it is used to solve* $\mathrm{LC}(M,q)$.

PROOF. Let us first specify the conditions on q such that the Newton-min algorithm makes the cycle $x^{(I)} \to x^{(J)} \to x^{(I)}$ when it is used to solve $\mathrm{LC}(M,q)$. Let $x^1 = x^{(I)}$, so that by (4.2.3),

$$\begin{cases} x^1_I = -M_{II}^{-1} q_I \\ x^1_{I^c} = 0 \end{cases} \quad \text{and} \quad \begin{cases} (Mx^1+q)_I = 0 \\ (Mx^1+q)_{I^c} = q_{I^c} - M_{I^cI} M_{II}^{-1} q_I. \end{cases}$$

We have used the nonsingularity of M_{II}, which comes from the nondegeneracy of M. Now, by definition of the Newton-min algorithm (4.2.1)-(4.2.2), the next iterate $x^2 = x^{(J)}$ if and only if

$$\underbrace{-(M_{II}^{-1}q_I)_{J^c} \leqslant 0}_{\beta_{I\cap J^c}}, \qquad \underbrace{(M_{II}^{-1}q_I)_J < 0}_{\alpha'_{I\cap J}},$$
$$\underbrace{q_{I^c\cap J} < M_{(I^c\cap J)I} M_{II}^{-1} q_I}_{\alpha_{I^c\cap J}}, \quad \text{and} \quad M_{(I\cup J)^cI} M_{II}^{-1} q_I \leqslant q_{(I\cup J)^c}. \quad (4.3.4)$$

The vectors under the braces are Motzkin multipliers, which will be used below. In that case

$$\begin{cases} x^2_J = -M_{JJ}^{-1} q_J \\ x^2_{J^c} = 0 \end{cases} \quad \text{and} \quad \begin{cases} (Mx^2+q)_J = 0 \\ (Mx^2+q)_{J^c} = q_{J^c} - M_{J^cJ} M_{JJ}^{-1} q_J. \end{cases}$$

We have used the nonsingularity of M_{JJ}, which comes from the nondegeneracy of M.

Now, $x^3 = x^{(I)}$, if and only if

$$\begin{gathered}\underbrace{-(M_{JJ}^{-1}q_J)_{I^c} \leqslant 0}_{\beta_{I^c\cap J}}, \qquad \underbrace{(M_{JJ}^{-1}q_J)_I < 0}_{\alpha''_{I\cap J}},\\ \underbrace{q_{I\cap J^c} < M_{(I\cap J^c)J}M_{JJ}^{-1}q_J}_{\alpha_{I\cap J^c}}, \qquad \text{and} \qquad M_{(I\cup J)^c J}M_{JJ}^{-1}q_J \leqslant q_{(I\cup J)^c}.\end{gathered} \tag{4.3.5}$$

The vectors under the braces are Motzkin multipliers, which will be used below. We have shown that the Newton-min algorithm makes the cycle $x^{(I)} \to x^{(J)} \to x^{(I)}$ if and only if q satisfies the linear inequalities in (4.3.4)-(4.3.5).

Observe now that the components of q with indices in $(I\cup J)^c$ intervene only in the last inequalities in (4.3.4)-(4.3.5) and that these inequalities can be satisfied by taking these components of q sufficiently large. Therefore, below, we do not have to consider the satisfiability of these last inequalities. This is the reason why we have not assigned Motzkin multipliers to these inequalities.

By Motzkin's theorem of the alternative (lemma 4.1), there is a q satisfying the linear inequalities in (4.3.4)-(4.3.5) if and only if one cannot find

$$(\alpha, \alpha', \alpha'', \beta) \in \mathbb{R}_+^{|I\triangle J|} \times \mathbb{R}_+^{|I\cap J|} \times \mathbb{R}_+^{|I\cap J|} \times \mathbb{R}_+^{|I\triangle J|}$$

(these are the vectors under the braces in (4.3.4)-(4.3.5)) such that $(\alpha, \alpha', \alpha'') \neq 0$ and

- $(M_{II}^{-\mathsf{T}})_{I\cap J^c}\begin{pmatrix}-\beta_{I\cap J^c}\\ \alpha'_{I\cap J}\end{pmatrix} - (M_{II}^{-\mathsf{T}})_{I\cap J^c}M^{\top}_{(I^c\cap J)I}\alpha_{I^c\cap J} + \alpha_{I\cap J^c} = 0,$
- $(M_{II}^{-\mathsf{T}})_{I\cap J}\begin{pmatrix}-\beta_{I\cap J^c}\\ \alpha'_{I\cap J}\end{pmatrix} - (M_{II}^{-\mathsf{T}})_{I\cap J}M^{\top}_{(I^c\cap J)I}\alpha_{I^c\cap J} + (M_{JJ}^{-\mathsf{T}})_{I\cap J}\begin{pmatrix}\alpha''_{I\cap J}\\ -\beta_{I^c\cap J}\end{pmatrix}$
  $- (M_{JJ}^{-\mathsf{T}})_{I\cap J}M^{\top}_{(I\cap J^c)J}\alpha_{I\cap J^c} = 0,$
- $\alpha_{I^c\cap J} + (M_{JJ}^{-\mathsf{T}})_{I^c\cap J}\begin{pmatrix}\alpha''_{I\cap J}\\ -\beta_{I^c\cap J}\end{pmatrix} - (M_{JJ}^{-\mathsf{T}})_{I^c\cap J}M^{\top}_{(I\cap J^c)J}\alpha_{I\cap J^c} = 0,$

where A_R denotes the submatrix of a matrix A formed of its rows with indices in R. With the notation

$$u := \begin{pmatrix}-\beta_{I\cap J^c}\\ \alpha'_{I\cap J}\end{pmatrix} - M^{\top}_{(I^c\cap J)I}\alpha_{I^c\cap J} \quad \text{and} \quad v := \begin{pmatrix}\alpha''_{I\cap J}\\ -\beta_{I^c\cap J}\end{pmatrix} - M^{\top}_{(I\cap J^c)J}\alpha_{I\cap J^c} \tag{4.3.6}$$

the system above reads equivalently

$$\begin{pmatrix} (M_{II}^{-\mathsf{T}})_{I\cap J^c}u \\ (M_{II}^{-\mathsf{T}})_{I\cap J}u \\ 0 \end{pmatrix} + \begin{pmatrix} \alpha_{I\cap J^c} \\ 0 \\ \alpha_{I^c\cap J} \end{pmatrix} + \begin{pmatrix} 0 \\ (M_{JJ}^{-\mathsf{T}})_{I\cap J}v \\ (M_{JJ}^{-\mathsf{T}})_{I^c\cap J}v \end{pmatrix} = 0. \tag{4.3.7}$$

If we multiply the first two equations to the left by $M_{II}^{\top}$, if we multiply the last two equations to the left by $M_{JJ}^{\top}$, and if we use the second equation, we obtain the equivalent system

$$u + M_{(I\cap J^c)I}^{\top}\alpha_{I\cap J^c} - M_{(I\cap J)I}^{\top}(M_{II}^{-\mathsf{T}})_{I\cap J}u = 0, \tag{4.3.8a}$$
$$v + M_{(I^c\cap J)J}^{\top}\alpha_{I^c\cap J} + M_{(I\cap J)J}^{\top}(M_{II}^{-\mathsf{T}})_{I\cap J}u = 0. \tag{4.3.8b}$$

Now the rows with indices in $I\cap J$ of these last two equations read

$$\alpha'_{I\cap J} - M_{(I^c\cap J)(I\cap J)}^{\top}\alpha_{I^c\cap J} + M_{(I\cap J^c)(I\cap J)}^{\top}\alpha_{I\cap J^c} - M_{(I\cap J)(I\cap J)}^{\top}(M_{II}^{-\mathsf{T}})_{I\cap J}u = 0,$$
$$\alpha''_{I\cap J} - M_{(I\cap J^c)(I\cap J)}^{\top}\alpha_{I\cap J^c} + M_{(I^c\cap J)(I\cap J)}^{\top}\alpha_{I^c\cap J} + M_{(I\cap J)(I\cap J)}^{\top}(M_{II}^{-\mathsf{T}})_{I\cap J}u = 0.$$

By adding these equations, we obtain $\alpha'_{I\cap J} + \alpha''_{I\cap J} = 0$, which implies that $\alpha'_{I\cap J} = \alpha''_{I\cap J} = 0$ since the components of these two vectors are nonnegative. One deduces then from any of these last equations that

$$(M_{II}^{-\mathsf{T}})_{I\cap J}u = M_{(I\cap J)(I\cap J)}^{-\mathsf{T}}\left(M_{(I\cap J^c)(I\cap J)}^{\top}\alpha_{I\cap J^c} - M_{(I^c\cap J)(I\cap J)}^{\top}\alpha_{I^c\cap J}\right). \tag{4.3.9}$$

Now the $I\cap J^c$ component of equation (4.3.8a) and the $I^c\cap J$ component of equation (4.3.8b) read

$$\begin{aligned} &- M_{(I^c\cap J)(I\cap J^c)}^{\top}\alpha_{I^c\cap J} + M_{(I\cap J^c)(I\cap J^c)}^{\top}\alpha_{I\cap J^c} \\ &\qquad - M_{(I\cap J)(I\cap J^c)}^{\top}M_{(I\cap J)(I\cap J)}^{-\mathsf{T}}\left(M_{(I\cap J^c)(I\cap J)}^{\top}\alpha_{I\cap J^c} - M_{(I^c\cap J)(I\cap J)}^{\top}\alpha_{I^c\cap J}\right) \\ &\qquad = \beta_{I\cap J^c}, \end{aligned} \tag{4.3.10a}$$

$$\begin{aligned} &- M_{(I\cap J^c)(I^c\cap J)}^{\top}\alpha_{I\cap J^c} + M_{(I^c\cap J)(I^c\cap J)}^{\top}\alpha_{I^c\cap J} \\ &\qquad + M_{(I\cap J)(I^c\cap J)}^{\top}M_{(I\cap J)(I\cap J)}^{-\mathsf{T}}\left(M_{(I\cap J^c)(I\cap J)}^{\top}\alpha_{I\cap J^c} - M_{(I^c\cap J)(I\cap J)}^{\top}\alpha_{I^c\cap J}\right) \\ &\qquad = \beta_{I^c\cap J}. \end{aligned} \tag{4.3.10b}$$

We have deduced the system (4.3.10) and $\alpha'_{I\cap J}=\alpha''_{I\cap J}=0$ from (4.3.8). Reciprocally, to show that there has been no loss of information in that operation, let us show that one can recover the system (4.3.8) from (4.3.10) and $\alpha'_{I\cap J}=\alpha''_{I\cap J}=0$, provided we define u and v by (4.3.6). Indeed, let us denote by w the right hand side of (4.3.9) and let us first show that the identity (4.3.9) is verified :

$$\begin{aligned}
(M_{II}^{-\mathsf{T}})_{I\cap J}u &= (M_{II}^{-\mathsf{T}})_{I\cap J}\begin{pmatrix} -\beta_{I\cap J^c}-M^{\top}_{(I^c\cap J)(I\cap J^c)}\alpha_{I^c\cap J} \\ -M^{\top}_{(I^c\cap J)(I\cap J)}\alpha_{I^c\cap J}\end{pmatrix} \quad [(4.3.6),\ \alpha'_{I\cap J}=0]\\
&= (M_{II}^{-\mathsf{T}})_{I\cap J}\begin{pmatrix} -M^{\top}_{(I\cap J^c)(I\cap J^c)}\alpha_{I\cap J^c}+M^{\top}_{(I\cap J)(I\cap J^c)}w \\ -M^{\top}_{(I\cap J^c)(I\cap J)}\alpha_{I\cap J^c}+M^{\top}_{(I\cap J)(I\cap J)}w\end{pmatrix} \quad \begin{bmatrix}\text{(4.3.10a) and}\\ \text{definition of } w\end{bmatrix}\\
&= (M_{II}^{-\mathsf{T}})_{I\cap J}M_{II}^{\top}\begin{pmatrix}-\alpha_{I\cap J^c}\\ w\end{pmatrix}\\
&= w.
\end{aligned}$$

Now the $I\cap J^c$ component of (4.3.8a) is a consequence of (4.3.10a) ; the $I\cap J$ components of (4.3.8a) and (4.3.8b) are consequences of $\alpha'_{I\cap J}=\alpha''_{I\cap J}=0$ and (4.3.9) ; and the $I^c\cap J$ component of (4.3.8b) is a consequence of (4.3.10b).

Therefore, getting rid of the vector $\beta\in\mathbb{R}_+^{|I\triangle J|}$ in (4.3.10), we see that there is a q such that the Newton-min algorithm makes the cycle $x^{(I)}\to x^{(J)}\to x^{(I)}$ when it is used to solve $\mathrm{LC}(M,q)$ if and only if one cannot find an $\alpha\in\mathbb{R}_+^{|I\triangle J|}\setminus\{0\}$ such that

$$\begin{aligned}
&\begin{pmatrix} M_{(I\cap J^c)(I\cap J^c)} & -M_{(I\cap J^c)(I^c\cap J)} \\ -M_{(I^c\cap J)(I\cap J^c)} & M_{(I^c\cap J)(I^c\cap J)}\end{pmatrix}^{\top}\alpha\\
&\qquad\geqslant \begin{pmatrix}-M_{(I\cap J)(I\cap J^c)} & M_{(I\cap J)(I^c\cap J)}\end{pmatrix}^{\top}M^{-\mathsf{T}}_{(I\cap J)(I\cap J)}\begin{pmatrix}-M_{(I\cap J^c)(I\cap J)}\\ M_{(I^c\cap J)(I\cap J)}\end{pmatrix}^{\top}\alpha.
\end{aligned}$$

The equivalence $(i)\Leftrightarrow(ii)$ follows. □

Observe that, as expected, condition (4.3.3) is symmetric in I and J, in the sense that the permutation $I\leftrightarrow J$ does not modify the condition.

Proposition 4.2 is apparently difficult to use if the goal is to characterize the class $\mathbf{NM}_2$ of matrices. It requires indeed to consider all the possible pairs of distinct subsets I and $J\subset[\![1,n]\!]$. In addition, for each of these pairs, each choice of $\alpha\in\mathbb{R}_+^{|I\triangle J|}\setminus\{0\}$ in

(4.3.3) may yield different conditions on M that make the Newton-min algorithm avoid the 2-cycle $x^{(I)} \to x^{(J)} \to x^{(I)}$. This looks like a long-term and hazardous task. We show in theorem 4.4 below, however, that it is equivalent to avoid all the 2-cycles or to avoid the small subset of 2-cycles $x^{(I)} \to x^{(J)} \to x^{(I)}$, in which $I = J \cup \{i\}$ with $i \notin J$. In other words, this small subset of 2-cycles contains all the information on M that is needed to prevent the Newton-min algorithm from making 2-cycles. A precious interest of the choice $I = J \cup \{i\}$ is that α is then a positive *scalar*, which can be eliminated from the inequality (4.3.3). The task of characterizing the matrices M in $\mathbf{NM}_2$ is then much easier. It is shown in the next section that these matrices are the **P**-matrices.

4.4 A characterization of P-matricity

Let us start by an elementary lemma.

Lemma 4.3 *Suppose that I and J are two index sets included in $[\![1,n]\!]$ such that $J \setminus I \neq \varnothing$ and that $x^{(I)} = x^{(J)}$ for some q. Then the Newton-min algorithm* (4.2.1)-(4.2.2) *does not make the null displacement $x^{(I)} \to x^{(J)}$ when it is used to solve* $\mathrm{LC}(M,q)$.

PROOF. We argue by contradiction, assuming that the Newton-min algorithm goes from $x^{(I)}$ to $x^{(J)}$. Then, $(Mx^{(I)}+q)_{I^c \cap J} < x^{(I)}_{I^c \cap J}$ [by the algorithm rule (4.2.1)] and $x^{(I)}_{I^c \cap J} = 0$ [by the definition of $x^{(I)}$, see $(4.2.3)_1$], so that

$$(Mx^{(I)}+q)_{I^c \cap J} < 0.$$

On the other hand,

$$(Mx^{(J)}+q)_{I^c \cap J} = 0,$$

by the definition of $x^{(J)}$, see $(4.2.3)_2$. Therefore $(Mx^{(I)}+q)_{I^c \cap J} \neq (Mx^{(J)}+q)_{I^c \cap J}$, contradicting the fact that $x^{(I)} = x^{(J)}$ for the given q. □

One comment on this lemma. It is not difficult to see that if the node $x^{(I)}$ is a solution to $\mathrm{LC}(M,q)$, then the next index set, here denoted J, computed by the Newton-min algorithm by (4.2.1) is smaller than I, hence $J \setminus I = \varnothing$ and lemma 5.3 does not apply to that case. In particular, if $q = 0$, then $J = \varnothing$ whatever is I. Incidentally, it is also known that if $x^{(I)}$ is a

solution to $\mathrm{LC}(M,q)$ and if M is nondegenerate, then $x^{(J)} = x^{(I)}$, where $x^{(J)}$ is the iterate following $x^{(I)}$ (see lemma 4.1 in [9] and the references thereof).

We denote by $\mathrm{cof}(M)$ the *cofactor matrix* of a matrix $M \in \mathbb{R}^{n\times n}$, whose element $[\mathrm{cof}(M)]_{ij}$ is the *cofactor* $\mathrm{cof}(M_{ij})$ of the element M_{ij} of M, that is

$$\mathrm{cof}(M_{ij}) := (-1)^{i+j} \det M_{([\![1,n]\!]\setminus\{i\})([\![1,n]\!]\setminus\{j\})}. \tag{4.4.1}$$

We use the notation $\mathrm{cof}_{II}(M_{ij})$ for the cofactor of the element M_{ij} in M_{II}. Recall [83, 1987, chapter VI] that for any index i and j :

$$\det M = \textstyle\sum_{i'} M_{i'j}\,\mathrm{cof}(M_{i'j}) = \sum_{j'} M_{ij'}\,\mathrm{cof}(M_{ij'}) \tag{4.4.2}$$

and that

$$M^{-1} = (\det M)^{-1}\,\mathrm{cof}(M^\top). \tag{4.4.3}$$

Our main result is given in theorem 4.4 below. The implication $(i) \Rightarrow (ii)$ of the theorem was already proven in [9, 2009, lemma 4.3], but that part of the paper was not selected by the refereeing process for appearing in the published version of the paper [11, 2012].

Theorem 4.4 (a characterization of P-matricity) *Suppose that $M \in$ ND. Then the following conditions are equivalent :*

- (i) *$M \in \mathbf{P}$,*
- (ii) *for any q, the Newton-min algorithm does not cycle between two different nodes when it is used to solve $\mathrm{LC}(M,q)$,*
- (iii) *for any q, for any subset $J \subset [\![1,n]\!]$, and for any index $i \in [\![1,n]\!] \setminus J$, the Newton-min algorithm does not cycle between the nodes $x^{(J)}$ and $x^{(J\cup\{i\})}$ when it is used to solve $\mathrm{LC}(M,q)$,*
- (iv) *for any subset $J \subset [\![1,n]\!]$ and any index $i \in [\![1,n]\!] \setminus J$, there holds*

$$M_{ii} \geqslant M_{\{i\}J} M_{JJ}^{-1} M_{J\{i\}}. \tag{4.4.4}$$

PROOF. $[(i) \Rightarrow (ii)]$ We prove the contrapositive, assuming that the algorithm visits in order the following nodes $x^{(I)} \to x^{(J)} \to x^{(I)}$, for some I and $J \subset [\![1,n]\!]$ and some $q \in \mathbb{R}^n$ such that $x^{(I)} \neq x^{(J)}$. We simplify the notation by setting $x^1 := x^{(I)}$ and $x^2 := x^{(J)}$. Since the

Newton-min algorithm goes from x^1 to x^2 and from x^2 to x^1, the very definition (4.2.1)-(4.2.2) of the algorithm implies that

$$x^1_{J^c} \leqslant (Mx^1+q)_{J^c} \qquad \text{and} \qquad x^1_J > (Mx^1+q)_J, \tag{4.4.5}$$
$$x^2_{I^c} \leqslant (Mx^2+q)_{I^c} \qquad \text{and} \qquad x^2_I > (Mx^2+q)_I. \tag{4.4.6}$$

After a possible rearrangement of the component order, we get

$$x^2-x^1 = \begin{pmatrix} 0_{I\cap J^c} \\ x^2_{I\cap J} \\ x^2_{I^c\cap J} \\ 0_{I^c\cap J^c} \end{pmatrix} - \begin{pmatrix} x^1_{I\cap J^c} \\ x^1_{I\cap J} \\ 0_{I^c\cap J} \\ 0_{I^c\cap J^c} \end{pmatrix} = \begin{pmatrix} -x^1_{I\cap J^c} \\ (x^2-x^1)_{I\cap J} \\ x^2_{I^c\cap J} \\ 0_{I^c\cap J^c} \end{pmatrix} \cdot \begin{matrix} [+] \\ [?] \\ [-] \\ [0] \end{matrix}$$

The extra column on the right gives the sign of each component, when appropriate : the components of x^2-x^1 with indices in $I\cap J^c$ are nonnegative since $-x^1_{I\cap J^c} \geqslant -(Mx^1+q)_{I\cap J^c}$ [by $(4.4.5)_1$] $=0$ [by $(4.2.2)_2$] and that the components with indices in $I^c\cap J$ are nonpositive since $x^2_{I^c\cap J} \leqslant (Mx^2+q)_{I^c\cap J}$ [by $(4.4.6)_1$] $=0$ [by $(4.2.2)_2$]. On the other hand, by $(4.2.2)_2$, there holds

$$M(x^2-x^1) = \begin{pmatrix} (Mx^2)_{I\cap J^c} \\ -q_{I\cap J} \\ -q_{I^c\cap J} \\ (Mx^2)_{I^c\cap J^c} \end{pmatrix} - \begin{pmatrix} -q_{I\cap J^c} \\ -q_{I\cap J} \\ (Mx^1)_{I^c\cap J} \\ (Mx^1)_{I^c\cap J^c} \end{pmatrix} = \begin{pmatrix} (Mx^2+q)_{I\cap J^c} \\ 0 \\ -(Mx^1+q)_{I^c\cap J} \\ (M(x^2-x^1))_{I^c\cap J^c} \end{pmatrix} \cdot \begin{matrix} [-] \\ [0] \\ [+] \\ [?] \end{matrix}$$

The extra column on the right gives the sign of each component, when appropriate : the components with indices in $I\cap J^c$ are nonpositive since $(Mx^2+q)_{I\cap J^c} \leqslant x^2_{I\cap J^c}$ [by $(4.4.6)_2$] $=0$ [by $(4.2.2)_1$] and the components with indices in $I^c\cap J$ are nonnegative since $-(Mx^1+q)_{I^c\cap J} \geqslant -x^1_{I^c\cap J}$ [by $(4.4.5)_2$] $=0$ [by $(4.2.2)_1$]. Therefore

$$(x^2-x^1)\cdot M(x^2-x^1) \leqslant 0.$$

Since $x^1 \neq x^2$, M cannot be a **P**-matrix (see (4.1.1)).

$[(ii)\Rightarrow(iii)]$ Let $q\in\mathbb{R}^n$, $J\subset[\![1,n]\!]$, and $i\in[\![1,n]\!]\setminus J$. Set $I:=J\cup\{i\}$. If $x^{(I)}\neq x^{(J)}$, then (iii) is a clear consequence of (ii). If $x^{(I)}=x^{(J)}$, then (iii) is a consequence of lemma 5.3, according to which the Newton min does not go from $x^{(J)}$ to $x^{(I)}$ because $I\setminus J\neq\varnothing$.

$[(iii) \Rightarrow (iv)]$ Let J and i be like in (iv), and set $I = J \cup \{i\}$. By (iii), whatever is q, the Newton-min algorithm does not cycle between the nodes $x^{(I)}$ and $x^{(J)}$ when it is used to solve $\mathrm{LC}(M,q)$. Then, the implication $(ii) \Rightarrow (i)$ of proposition 4.2 shows that there is an $\alpha > 0$ such that (4.2) holds. Since $I \cap J^c = \{i\}$, $I^c \cap J = \varnothing$, $I \cap J = J$, $I \triangle J = \{i\}$, and α is a positive scalar that can be eliminated from (4.2), this inequality simplifies in (4.4.4) (use also the fact that $M_{\{i\}J} M_{JJ}^{-1} M_{J\{i\}}$ is a scalar, hence equal to its transpose).

$[(iv) \Rightarrow (i)]$ We prove by induction that $\det M_{II} > 0$ for any $I \subset [\![1,n]\!]$, which is equivalent to $M \in \mathbf{P}$. By applying (iv) with $J = \varnothing$, we obtain $M_{ii} > 0$ for a nondegenerate matrix, so that $\det M_{II} > 0$ when $|I| = 1$. Now, assume that J and i are chosen like in (iv), that $I = J \cup \{i\}$, that $\det M_{JJ} > 0$ (induction assumption), and let us show that $\det M_{II} > 0$, which will conclude the proof of $(iv) \Rightarrow (i)$.

Let us denote the indices in J by j_k, $k \in [\![1, |J|]\!]$. Using the cofactor matrix of M_{JJ} in (4.4.4) and the induction assumption $\det M_{JJ} > 0$, one gets

$$
\begin{aligned}
0 \;\leqslant\; & M_{ii} \det M_{JJ} - M_{\{i\}J} \operatorname{cof}(M_{JJ}^{\top}) M_{J\{i\}} && [(4.4.4),\ (4.4.3),\ \det M_{JJ} > 0] \\
=\; & M_{ii} \det M_{JJ} - \textstyle\sum_{k=1}^{|J|} \sum_{l=1}^{|J|} M_{ij_k} \operatorname{cof}_{JJ}([M_{JJ}]_{j_l j_k}) M_{j_l i} \\
=\; & M_{ii} \det M_{JJ} - \textstyle\sum_{k=1}^{|J|} \sum_{l=1}^{|J|} M_{ij_k} (-1)^{l+k} \det M_{(J\setminus\{j_l\})(J\setminus\{j_k\})} M_{j_l i} && [(4.4.1)] \\
=\; & M_{ii} \det M_{JJ} + \textstyle\sum_{k=1}^{|J|} (-1)^{k+|J|+1} M_{ij_k} \sum_{l=1}^{|J|} M_{j_l i} (-1)^{l+|J|} \det M_{(J\setminus\{j_l\})(J\setminus\{j_k\})} \\
=\; & M_{ii} \det M_{JJ} + \textstyle\sum_{k=1}^{|J|} (-1)^{k+|J|+1} M_{ij_k} \det\left(M_{J(J\setminus\{j_k\})}, M_{J\{i\}}\right) && [(4.4.2)] \\
=\; & \det M_{II} && [(4.4.2)].
\end{aligned}
$$

Therefore $\det M_{II} > 0$ by the nondegeneracy of M. □

Even though the following consequence of theorem 4.4 is clear and was already summarized by formula (4.2.5) in the introduction, we quote it in a corollary to make easier a possible future citation.

It is clear that $\mathbf{NM}$ is included in the set $\mathbf{ND} \cap \mathbf{Q}$of nondegenerate matrices ensuring that $\mathrm{LC}(M,q)$ has a solution, whatever is q; indeed, if $M \in \mathbf{ND} \setminus \mathbf{Q}$, one can find a vector q such that $\mathrm{LC}(M,q)$ has no solution, in which case, the Newton-min algorithm has no other choice than cycling (we recall from lemma 4.1 in [9] that, when $M \in \mathbf{ND}$, the sequence generated by the Newton-min algorithm can be stationnary only at a solution). The inclusion (4.4.7), however, was not clear to us, before theorem 4.4 was estabilised.

Corollary 4.5 (NM is in P) *The set of matrices in $M \in \mathbf{ND}$ ensuring the convergence of the Newton-min algorithm when it is used to solve* $\mathrm{LC}(M,q)$*, whatever is the vector q and the initial point, is included in* $\mathbf{P}$*. More compactly*

$$\mathbf{NM} \subset \mathbf{P}. \tag{4.4.7}$$

PROOF. This is because $\mathbf{NM}$ is the set of matrices in $M \in \mathbf{ND}$ ensuring the convergence of the Newton-min algorithm when it is used to solve $\mathrm{LC}(M,q)$, whatever is the vector q (see the discussion before formula (4.2.4)), because $\mathbf{NM} \subset \mathbf{NM}_2$ by the definition (4.2.4) of $\mathbf{NM}$, and because $\mathbf{NM}_2 = \mathbf{P}$ by theorem 4.4. □

The implication $(ii) \Rightarrow (i)$ of theorem 4.4, according to which only the $\mathbf{P}$-matrices in $\mathbf{ND}$ prevent the Newton-min algorithm from cycling between two nodes, is ultimately based on Motzkin's theorem of the alternative, which supports the implication $(iii) \Rightarrow (iv)$ of theorem 4.4, while the implication $(ii) \Rightarrow (iii)$ is straightforward and the implication $(iv) \Rightarrow (i)$ has an algebraic nature. Motzkin's theorem is not constructive whereas, for $M \in \mathbf{ND} \setminus \mathbf{P}$, it would certainly be interesting (and reassuring) to be able to construct a q and to select two nodes such that the Newton-min algorithm cycles between these nodes when it is used to solve $\mathrm{LC}(M,q)$. This is the goal of the next proposition, which takes inspiration from the contrapositive of the implications $(iii) \Rightarrow (iv) \Rightarrow (i)$ of theorem 4.4 : if $M \in \mathbf{ND} \setminus \mathbf{P}$, there are a vector $q \in \mathbb{R}^n$, a subset $J \subset [\![1,n]\!]$, and an index $i \in [\![1,n]\!] \setminus J$ such that the Newton-min algorithm cycles between the nodes $x^{(J)}$ and $x^{(J\cup\{i\})}$. According to the contrapositive of the implication $(iv) \Rightarrow (i)$, the index sets J and $I := J \cup \{i\}$ are such that $\det M_{JJ} > 0$ and $\det M_{II} < 0$, which provides a means to select the index sets. One still has to find the vector q and this is precisely what the next proposition brings by exhibiting a whole family of vectors q such that this cycle occurs.

We adopt the notation $t^+ := \max(0,t)$ for $t \in \mathbb{R}$. The operator 'max' is supposed to act componentwise on vectors.

Proposition 4.6 (2-cycle for $M \notin \mathbf{P}$) *Suppose that $M \in \mathbf{ND} \setminus \mathbf{P}$. Then*

1) *there are two index sets I and $J \subset [\![1,n]\!]$ and an index $i \in [\![1,n]\!]$ such that $I = J \cup \{i\}$, $\det M_{II} < 0$, and $\det M_{JJ} > 0$,*
2) *for any two index sets I and $J \subset [\![1,n]\!]$ and an index $i \in [\![1,n]\!]$ having the properties given in point 1, the Newton-min algorithm cycles between $x^{(I)}$ and $x^{(J)}$ when the components of q are determined in order as follows*

$$q_J = -M_{JJ} e^J, \tag{4.4.8}$$

$$q_i = -M_{iJ} e^J - \varepsilon, \quad \text{with } 0 < \varepsilon < \frac{|\det M_{II}|}{\max_{j \in J} [\mathrm{cof}_{II}(M_{ij})]^+}, \tag{4.4.9}$$

$$q_{I^c} \geqslant \max\left(M_{I^c J} M_{JJ}^{-1} q_J, M_{I^c I} M_{II}^{-1} q_I\right), \tag{4.4.10}$$

where e^J is the vector of all ones in $\mathbb{R}^{|J|}$.

PROOF. 1) Since $M \notin \mathbf{P}$, some principal minor of M is negative. Then one can choose an index set I with the smallest cardinal number $|I|$ such that $\det M_{II} < 0$. Since $|I| \geqslant 1$, one can choose an index $i \in I$ and set $J := I \setminus \{i\}$ (J may be empty). The properties in point 1 are verified for the selected sets I and J, and the index i (recall that $\det M_{\varnothing\varnothing} = 1$ by convention).

2) Let the index sets I and J and the index i be such that $I = J \cup \{i\}$, $\det M_{II} < 0$, and $\det M_{JJ} > 0$. Suppose that q satisfies (4.4.8)-(4.4.10). Let $x^1 := x^{(J)}$, so that

$$\begin{cases} x^1_J = -M_{JJ}^{-1} q_J \\ x^1_{J^c} = 0 \end{cases} \quad \text{and} \quad \begin{cases} (Mx^1 + q)_J = 0 \\ (Mx^1 + q)_{J^c} = q_{J^c} - M_{J^c J} M_{JJ}^{-1} q_J. \end{cases}$$

For a q satisfying the assumption, there hold

$$\begin{aligned} M_{JJ}^{-1} q_J = -e^J < 0 \qquad & [(4.4.8)], \\ q_i - M_{iJ} M_{JJ}^{-1} q_J = -\varepsilon < 0 \qquad & [(4.4.9)], \\ q_{I^c} - M_{I^c J} M_{JJ}^{-1} q_J \geqslant 0 \qquad & [(4.4.10)]. \end{aligned}$$

These inequalities imply that the next iterate visited by the Newton-min algorithm (4.2.1)-

(4.2.2) is $x^2 := x^{(I)}$, so that

$$\begin{cases} x^2_I = -M_{II}^{-1} q_I \\ x^2_{I^c} = 0 \end{cases} \qquad \text{and} \qquad \begin{cases} (Mx^2+q)_I = 0 \\ (Mx^2+q)_{I^c} = q_{I^c} - M_{I^cI} M_{II}^{-1} q_I. \end{cases}$$

We now want to show that appropriate inequalities on q are verified that ensure that the iterate following x^2 is x^1. Observe that by (4.4.8) and (4.4.9)

$$M_{II}^{-1} q_I = -M_{II}^{-1} M_{IJ} e^J - \varepsilon\, M_{II}^{-1} e^{I,i} = -(e^I - e^{I,i}) - \varepsilon\, (\det M_{II})^{-1} \operatorname{cof}_{II}(M_{iI})^\top,$$

where $e^{I,i} \in \mathbb{R}^{|I|}$ is a vector whose components are all zero, except the one at the position of i in I whose value is 1. Therefore

$$\forall\, j \in J: \qquad (M_{II}^{-1} q_I)_j = -1 - \varepsilon\, (\det M_{II})^{-1} \operatorname{cof}_{II}(M_{ij}) < 0, \tag{4.4.11}$$

since $-\varepsilon\, (\det M_{II})^{-1} \operatorname{cof}_{II}(M_{ij}) \leqslant \varepsilon\, |\det M_{II}|^{-1} \max\limits_{j\in J} [\operatorname{cof}_{II}(M_{ij})]^+ < 1$, by the choice of ε in (4.4.9). On the other hand,

$$(M_{II}^{-1} q_I)_i = -\varepsilon\, (\det M_{II})^{-1} \operatorname{cof}_{II}(M_{ii}) = -\varepsilon\, (\det M_{II})^{-1} (\det M_{JJ}) > 0, \tag{4.4.12}$$

since $(\det M_{II}) < 0$ and $(\det M_{JJ}) > 0$ by assumption. Finally, by (4.4.10), there holds

$$q_{I^c} - M_{I^cI} M_{II}^{-1} q_I \geqslant 0. \tag{4.4.13}$$

The inequalities (4.4.11), (4.4.12), and (4.4.13) imply that the iterate following x^2 is indeed x^1. Hence the Newton-min algorithm cycles between the nodes $x^{(J)}$ and $x^{(I)}$. □

Thanks to proposition 4.6, the equivalence $(i) \Leftrightarrow (ii)$ of proposition 4.4 can now be proven without proposition 4.2 and Motzkin's theorem of the alternative, so that one can wonder whether section 4.3 is still useful. We have maintained it in the paper for two reasons. First, proposition 4.2 has brought a clear indication on the way of choosing the index sets I and J in proposition 4.6, whose origin could be obscure otherwise. Next, the dual characterization of the absence of 2-cycle that it provides may be useful in the extention of this work.

4.5 Discussion and perspectives

A natural extention of the present work would consist in looking at the possibility to give a *simple* algebraic description of the classes $\mathbf{NM}_k$, for $k \geqslant 3$, and $\mathbf{NM}$. An intrigating feature of the approach followed in this paper to characterize $\mathbf{NM}_2$ is its intrinsic conjonctive-disjonctive nature. It is conjonctive in the sense that it provides conditions to satisfy for each possible 2-cycle that the matrix must prevent. Its disjonctive aspect comes from the use of Motzkin's theorem of the alternative (lemma 4.1), which provides a possibility to avoid a given cycle for each acceptable choice of multipliers α in proposition 4.2, and there may be many. For example, when $M \in \mathbf{P}$, the same approach shows that the Newton-min algorithm does not make the 3-cycle $x^{(\{i\})} \to x^{(\{j\})} \to x^{(\{k\})} \to x^{(\{i\})}$, for three different indices i, j, and $k \in [\![1,n]\!]$, when it is used to solve $\mathrm{LC}(M,q)$, whatever is $q \in \mathbb{R}^n$, if and only if *one* of the following eight conditions holds

$$
\begin{array}{l}
M_{ik}M_{jj} \leqslant M^+_{ij}M_{jk}, \quad M_{ji}M_{kk} \leqslant M^+_{jk}M_{ki}, \quad M_{kj}M_{ii} \leqslant M^+_{ki}M_{ij}, \\
M^+_{ji}M_{ik} \leqslant M_{jk}M_{ii}, \quad M^+_{kj}M_{ji} \leqslant M_{ki}M_{jj}, \quad M^+_{ik}M_{kj} \leqslant M_{ij}M_{kk}, \\
M^+_{ik}M^+_{kj}M^+_{ji} \leqslant M_{ii}M_{jj}M_{kk}, \quad \text{or} \quad M^+_{ij}M^+_{jk}M^+_{ki} > M_{ii}M_{jj}M_{kk}.
\end{array}
$$

Note that conditions 1, 2, 3, and 7 are satisfied by an $\mathbf{M}$-matrix. We don't know whether this disjonctive form of the conditions disappears for the whole classes $\mathbf{NM}_k$ or $\mathbf{NM}$, i.e., when all the possible cycles must be avoided, as it does for $\mathbf{NM}_2 = \mathbf{P}$.

Another natural question is whether the membership to $\mathbf{NM}$ can be determined in polynomial time ; we recall that recognizing a $\mathbf{P}$-matrix is a co-NP-complete problem [31, 124, 1994-2000]. It would also be important to know whether the Newton-min algorithm solves the linear complementarity problem $\mathrm{LC}(M,q)$ in polynomial time when $M \in \mathbf{NM}$; recall that it does when $M \in \mathbf{M}$ [71, 2004].

Chapitre 5

A globally convergent modified Newton-min algorithm for solving linear complementarity problems with a P-matrix

Dans ce chapitre, nous proposons une méthode de globalisation de l'algorithme de Newton-min 3.1 afin de forcer sa convergence pour les problèmes de complémentarité avec **P**-matrice. Nous démontrons également sa convergence globale pour cette classe de matrices et nous présentons des résultats numériques montrant son efficacité et sa convergence polynomiale pour les cas considérés. Ce travail fait l'objet d'un article en préparation.

5.1 Introduction

Given a real matrix M of order n and a vector q in $\mathbb{R}^n$, the linear complementarity problem (LCP) is to find a nonnegative vector x with n components such that $Mx+q \geqslant 0$ and $x^\top(Mx+q) = 0$. Here and below, the inequalities have to be understood componentwise, and the sign $^\top$ denotes matrix transposition. The LCP is often written in compact form as follows

$$\mathrm{LC}(M,q): \qquad 0 \leqslant x \perp Mx+q \geqslant 0. \tag{5.1.1}$$

In equation form, it can be written

$$H(x) \equiv \min(x, Mx+q) = 0. \tag{5.1.2}$$

Many algorithms have been proposed to solve problem LC(M,q) [74, 103, 69, 30]. Here, we are interested in the plain Newton-min algorithm (see chapter 3), which can be viewed as a semismooth Newton algorithm without globalization technique to solve the system of piecewise linear equations (5.1.2). Recall from section 2.5.5 that this algorithm first defines complementary index sets A^+ and I^+, which in their simplest form, read

$$A^+ := \{i : x_i \leqslant (Mx+q)_i\} \qquad \text{and} \qquad I^+ := \{i : x_i > (Mx+q)_i\}. \tag{5.1.3}$$

Then, it computes a direction d by solving the linear system formed of the equations

$$d_{A^+} = -x^+_{A^+} \qquad \text{and} \qquad d_{I^+} = -x_{I^+} - M^{-1}_{I^+I^+} q_{I^+}. \tag{5.1.4}$$

Next, the solution x^+ is given by the following relation

$$x^+ = x + d. \tag{5.1.5}$$

Since the Newton-min algorithm may cycle for **P**-matrices (see chapter 3) and in order to force its convergence, a natural remedy would be to add a globalization technique. Contrary to the differentiable case, the Newton-min direction d computed above is not necessary a descent direction of the natural merit function $x \mapsto \|\min(x, Mx+q)\|^2/2$ associated with problem (5.1.2) at an arbitrary vector x (see section 5.2.1). We had thought [85] to impose the non-negativity of the unknown x at each iteration which was observed to ensure the descent of the computed directions. We observed afterwords that this idea was actually already used in the paper by Pang and Gabriel [108], in which they have proposed an iterative method for solving the nonlinear complementarity problem (see Part II). This method involves solving a sequence of non negatively constrained convex quadratic optimization problems of the least-square type [108].

In this chapter, we present a different approach that is globally convergent for solving the LCP with a **P**-matrix. This method consists in minimizing a piecewise affine merit function subject to non-negativity constraints. Furthermore, in the proposed approach the descent direction is computed as a solution to a linear optimization problem ; it is therefore computationally tractable and slightly less computationally expensive than the one

proposed by Pang and Gabriel [108](the latter requires to solve a quadratic problem at each iteration).

The organization of this chapter is as follows. In section 5.2, we present various natural directions and discuss their respective interests. The new algorithm is described in section 5.3. We prove its global convergence. Finally, in order to demonstrate the practical efficiency of our method, we report, in section 5.4, the numerical results of some computational experiments and we also compare our approach with a method of Harker and Pang [106, 59].

5.2 A globalization of the Newton-min algorithm

We are interested here in the globalization of the Newton-min algorithm by line-search. To do this, we have studied various directions, which are natural in some sense. This results in the selection of one of them, which has better theoretical and numerical properties.

We begin our discussion with the first natural direction which is the semismooth direction defined by (5.1.4). Next comes the B-Newton direction, which is known to have good properties, although it is computationally expensive and it is not clear whether it yields global convergence. Finally (section 5.2.3), we present the direction that we prefer for both the global convergence it ensures and its numerical efficiency.

5.2.1 The semismooth direction

A natural merit function for problem LCP (5.1.1) when it is solved by Newton-min iteration on the function H, is the function

$$\Theta(x) = \frac{1}{2}\|H(x)\|^2, \tag{5.2.1}$$

where $\|\cdot\|$ denotes the Euclidean norm. It is natural since when H is smooth, the Newton direction for finding a zero of H is a descent direction of Θ, whatever is x.

We recall that M is said to be an $\mathbf{R}_0$-matrix (see section 2.2.9) if the homogeneous linear complementarity problem $0 \leqslant x \perp Mx \geqslant 0$ has no nonzero solution and that a function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is said to be *coercive* if the following equivalent properties are verified :

- $\forall v \in \mathbb{R}$, the sublevel set $\{x \in \mathbb{R}^n : \varphi(x) \leqslant v\}$ is *bounded*.
- $\lim_{||x||\to+\infty} \varphi(x) = +\infty$.

We recall the next proposition from [38, propositions 2.6.5 and 9.1.26].

Proposition 5.1 *The merit function $x \mapsto \Theta(x) = \|\min(x, Mx+q)\|$ is coercive if and only if $M \in \mathbf{R}_0$.*

Some differentiability propreties

Due to the presence of the "min" operator, the functions H and Θ are not differentiable in the traditional sense of Fréchet. Various differentiability properties of H and Θ have been derived in [106] ; it will be useful for us to summarize them here. By translating Pang's theorem [106, 1990, theorem 5] to the LCP, one obtains the following result.

Proposition 5.2 *Let $\Theta : \mathbb{R}^n \to \mathbb{R}$ be defined at $x \in \mathbb{R}^n$ by (5.2.1).*

(i) The function Θ has directional derivatives, which read for any $d \in \mathbb{R}^n$: $\Theta'(x;d) = H(x)^\top H'(x;d)$, where

$$H_i'(x;d) = \begin{cases} d_i & \text{if } i \in A_0(x) := \{i : x_i < (Mx+q)_i\} \\ \min(d_i, (Md)_i) & \text{if } i \in E(x) := \{i : x_i = (Mx+q)_i\} \\ (Md)_i & \text{if } i \in I_0(x) := \{i : x_i > (Mx+q)_i\}. \end{cases} \tag{5.2.2}$$

(ii) H is F-differentiable at x if and only if $M_{E(x)} = I_{E(x)}$.

On the convexity of Θ

We show in this section that a necessary and sufficient condition for Θ to be convex whatever is q is that $M = I$. Hence, the piecewise linearity of H does not result in general in a "nice" merit function Θ.

Lemma 5.3 *If A and B are two matrices such that $Ax = Bx$ for all $x \in \mathbb{R}^n_{++}$, then $A = B$.*

PROOF. Let $x \in \mathbb{R}^n$. One can write $x = (x^+ + e) - (x^- + e)$, where $x^+ = \max(x, 0)$, $x^- = \max(-x, 0)$, and e is the vector of all ones. Since $x^+ + e > 0$ and $x^- + e > 0$, the assumption

on A and B yields $Ax = A(x^+ + e) - A(x^- + e) = B(x^+ + e) - B(x^- + e) = Bx$. Hence $A = B$. □

Proposition 5.4 *The merit function* $x \mapsto \Theta(x) = \frac{1}{2}\|\min(x, Mx+q)\|^2$ *is convex for any* $q \in \mathbb{R}^n$ *if and only if* $M = I$.

PROOF. [⇒] Suppose that $M \neq I$. By lemma 5.3, one can find $x \in \mathbb{R}^n$ such that

$$x > 0 \qquad \text{and} \qquad (I - M)x \neq 0.$$

Define $q = (I - M)x$. We now show that $\Theta'(x;x) + \Theta'(x;-x) < 0$, which will imply that Θ is not convex. Note that $H(x) := \min(x, Mx+q) = x$. Using proposition 5.2, we have

$$\Theta'(x;x) + \Theta'(x;-x) = x^\top \big(\min(x, Mx) - \max(x, Mx)\big) < 0,$$

since $\min(x, Mx) \leqslant \max(x, Mx)$, $\min(x, Mx) \neq \max(x, Mx)$, and $x > 0$.

[⇐] If M is the identity matrix, $\Theta(x) = \frac{1}{2}\|\min(x, x+q)\|^2 = \frac{1}{2}\|x - q^-\|^2$. Hence Θ is a convex quadratic function. □

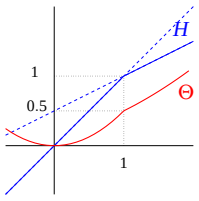
Clarke subdifferentiability of Θ

Recall that Θ is said to be *quasidifferentiable* [109, 1971] or *Clarke regular* [26, 1983, definition 2.3.4] at x if

$$\forall d \in \mathbb{R}^n : \quad \Theta'(x;d) \text{ exists and } \Theta'(x;d) = \Theta^\circ(x;d).$$

Lemma 5.5 *The function* Θ *is not necessarily Clarke regular.*

PROOF. Here is indeed an example in which Θ is not Clarke regular :



$$n = 1, \quad M = \frac{1}{2}, \quad \text{and} \quad q = \frac{1}{2}.$$

In this case, $\Theta(x) = [\min(x,(x+1)/2)]^2/2$. Clearly

$$\Theta'(1;1) = \lim_{t\downarrow 0} \frac{\Theta(1+t)-\Theta(1)}{t} = \lim_{t\downarrow 0} \frac{(2+t)^2/8-1/2}{t} = \frac{1}{2},$$

while the Clarke directional derivative reads

$$\begin{aligned}
\Theta^\circ(1;1) &= \limsup_{\substack{x\to 1\\ t\downarrow 0}} \frac{\Theta(x+t)-\Theta(x)}{t} \\
&\geqslant \lim_{t\downarrow 0} \frac{\Theta((1-2t)+t)-\Theta(1-2t)}{t} \qquad [\text{taking } x = 1-2t] \\
&= \lim_{t\downarrow 0} \frac{(1-t)^2-(1-2t)^2}{2t} \\
&= 1.
\end{aligned}$$

Since $\Theta^\circ(1;1) \neq \Theta'(1;1)$, Θ is not Clarke regular at $x = 1$. □

On the semismooth Newton direction

We prove in the next proposition that the semismooth direction is a descent direction if x is a node. We recall that a point x said to be a node if (2.1.2) holds, in other words, if it satisfies complementarity.

Proposition 5.6 *Let $\Theta : \mathbb{R}^n \to \mathbb{R}$ be defined at $x \in \mathbb{R}^n$ by (5.2.1). If x is a node and if d is the direction defined by (5.1.4), then $\Theta'(x;d) = -2\Theta(x)$, which is < 0 if $H(x) \neq 0$.*

PROOF. Using proposition 5.2, we have

$$\Theta'(x;d) = H(x)^\top H'(x;d) = x_{A_0}^\top d_{A_0} + x_E^\top \min(d_E,(Md)_E) + (Mx+q)_{I_0}^\top (Md)_{I_0}.$$

Obviously, $x_{A_0}^\top d_{A_0} = -\|x_{A_0}\|_2^2$. In addition, $x^+ = x+d$ verifies $(Mx^+ + q)_{I_0} = 0$. Hence

$$(Md)_{I_0} = -(Mx+q)_{I_0}. \tag{5.2.3}$$

We conclude that $(Mx+q)_{I_0}^\top (Md)_{I_0} = -\|(Mx+q)_{I_0}\|_2^2$. We achieve the proof by showing that the middle term vanishes : since for a node, either x_i or $(Mx+q)_i$ must vanish, one must have $x_E = (Mx+q)_E = 0$. □

We prove in the next proposition that the semismooth direction is a descent direction if x is nonnegative.

Proposition 5.7 *Let $\Theta : \mathbb{R}^n \to \mathbb{R}$ be defined at $x \in \mathbb{R}^n$ by (5.2.1). If $x \geqslant 0$ and if d is the direction defined by (5.1.4), then $\Theta'(x;d) \leqslant -2\Theta(x)$.*

PROOF. Like before, $\Theta'(x;d)$ is given by

$$\begin{aligned}\Theta'(x;d) &= x_{A_0}^\top d_{A_0} + x_E^\top \min(d_E, (Md)_E) + (Mx+q)_{I_0}^\top (Md)_{I_0} \\ &= x_{A_0}^\top d_{A_0} - \|x_E\|^2 + x_E^\top \min((x+d)_E, (Mx+q+Md)_E) \\ &\quad + (Mx+q)_{I_0}^\top (Md)_{I_0},\end{aligned}$$

where we have taken into account the fact that $x_E = (Mx+q)_E$.

To analyze the term $x_{A_0}^\top d_{A_0}$, we observe that $d_{A_0} = -x_{A_0}$ so $x_{A_0}^\top d_{A_0} = -\|x_{A_0}\|_2^2$. The term $(Mx+q)_{I_0}^\top (Md)_{I_0}$ is analyzed in a similar manner, by observing $(Md)_{I_0} = -(Mx+q)_{I_0}$. Hence, $(Mx+q)_{I_0}^\top (Md)_{I_0} = -\|(Mx+q)_{I_0}\|^2$. With these computations, the directional derivative becomes

$$\Theta'(x;d) \leqslant -2\,\Theta(x) + x_E^\top \min((x+d)_E, (Mx+q+Md)_E).$$

We now show that the second term in the right hand side above is nonpositive, which will conclude the proof. For $i \in E$, we have by using $\min((x+d)_i, (Mx+q+Md)_i) \leqslant (x+d)_i = 0$ and $x_i \geqslant 0$:

$$x_i \min((x+d)_i, (Mx+q+Md)_i) \leqslant 0.$$

□

In a linesearch or trust-region method, the iterates are not necessary nodes and may have negative components, so that propositions 5.6 and 5.7 are not sufficient. In other words, to be able to implement a linesearch method, using the Newton-min directions, it

is necessary to show that these directions are descent directions of Θ at an arbitrary point, as this is the case for the Newton direction when H is smooth. Unfortunately, this is not the case. Indeed, the next example shows that the often suggested semismooth Newton direction (see Kanzow [71]) computed by

$$\begin{cases} (x+d)_i = 0 & \text{if } x_i < (Mx+q)_i \\ (Mx+q+Md)_i = 0 & \text{if } x_i \geqslant (Mx+q)_i \end{cases} \tag{5.2.4}$$

is not necessary a descent direction of Θ if x is not a node. of the unknowns in the B-differentiable direction (5.2.5) of section 5.2.2.

Example 5.8 Consider problem (5.1.1) with

$$n = 3, \qquad M = \begin{pmatrix} 1 & 0 & 2 \\ 2 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix}, \qquad q = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \text{and} \qquad x = \begin{pmatrix} -1/2 \\ -1/2 \\ 0 \end{pmatrix}.$$

Note that the circulant matrix M is a **P**-matrix (see section 2.2.5).

Since $Mx+q = (1/2, -1/2, 0)$, there hold $A_0 = \{1\}$, $I_0 = \varnothing$, and $E = \{2,3\}$, so that the solution to (5.2.4) is $d = (1/2, -1/2, 1)$. Then $\Theta(x) = 1/4$ and, for $t \geqslant 0$, $\Theta(x+td) = \|(-1+t, -1-t, 0)\|^2/8 = (1+t^2)/4$, so that Θ increases along d. □

Clearly, the reason why a semismooth Newton direction may not be a descent direction of Θ at x comes from the indices in $E(x) := \{i_i = (Mx+q)_i\}$. In a linesearch method, it seems obvious to avoid this flaw, simply by requiring the linesearch to avoid the rare points x with nonempty index set $E(x)$. According to the numerical experiments of Harker and Pang [60, 1990], it seems that such a simple strategy does not yield a polynomial algorithm. For these reasons, we looked at other types of directions. Before presenting our proposition in section 5.2.3, we recall, in the next section, the definition of the B-Newton direction, which is known to be a descent direction of Θ; it is not useful in practice, however, because of the complexity of its computation.

5.2.2 The B-Newton direction

The B-Newton direction is based on the notion of B-differentiability, which it introduced in section 2.5.4.1 It is known from [106, 59] that the B-differentiable Newton direction (2.5.11) is a descent direction of Θ. This *B-Newton direction* d is defined at x as a

solution to $H(x)+BH(x)d=0$, which can be written as the following linear complementarity problem

$$\begin{cases} (x+d)_{A_0}=0 \\ (Mx+q+Md)_{I_0}=0 \\ 0\leqslant (x+d)_E \perp (Mx+q+Md)_E \geqslant 0, \end{cases} \tag{5.2.5}$$

where A_0, I_0 and E are defined in (5.2.2). For the indices in A_0 and I_0, the equation clearly writes like in (5.2.5). Observe indeed that while for indices in E,

$$\begin{aligned} 0 &= \min((x+d)_E,(Mx+q+Md)_E) \\ &= H_E(x)+\min(d_E,(Md)_E) \qquad [x_E=(Mx+q)_E] \\ &= H_E(x)+BH_E(x)d \qquad [BH_E(x)d=H'_E(x;d)]. \end{aligned}$$

We recall the next result from [59, 1990, p. 271].

Proposition 5.9 *The system* (5.2.5) *has a unique solution if the pair* (I_0,E) *satisfies the following properties.*
(i) *The submatrix* $M_{I_0I_0}$ *is nonsingular,*
(ii) *The Schur complement* $(M_{EE}-M_{EI_0}M_{I_0I_0}^{-1}M_{I_0E})$ *is a* **P***-matrix.*

PROOF. Let $x^+=x+d$. If $M_{I_0I_0}$ is nonsingular, the first two equations in (5.2.5) yield

$$x^+_{A_0}=0 \qquad \text{and} \qquad x^+_{I_0}=-M_{I_0I_0}^{-1}M_{I_0E}x^+_E-M_{I_0I_0}^{-1}q_{I_0}.$$

Therefore the full system amounts to solving the following linear complementarity problem in x^+_E :

$$0\leqslant x^+_E \perp \left(M_{EE}-M_{EI_0}M_{I_0I_0}^{-1}M_{I_0E}\right)x^+_E+\left(q_E-M_{EI_0}M_{I_0I_0}^{-1}q_{I_0}\right)\geqslant 0,$$

which has a unique solution if the Schur complement is a **P**-matrix. □

If $E=[\![1,n]\!]$, problem (5.2.5) is as hard to solve as the original problem, which does not make such a direction d particularly attractive. For instance, example 5.10 below gives an easy problem and a point x, for which problem (5.2.5) is exactly the same as the original problem (5.1.1).

Example 5.10 Consider the problem

$$M = \tfrac{1}{2} I, \quad q = \tfrac{1}{2} \cdot 1, \quad \text{and} \quad x = 1.$$

where I the identity matrix of order n and 1 denotes the vector $\in \mathbb{R}^n$ of all ones. Since $Mx+q = 1$, there hold $A_0 = I_0 = \varnothing$, and $E = [\![1, n]\!]$. If we introduce $z := 1+d$, problem (5.2.5) clearly reads

$$0 \leqslant z \perp \tfrac{1}{2}(z+1) \geqslant 0,$$

which is exactly the original problem. □

Since the B-Newton direction is too expensive to compute (it requires to solve an LCP), Harker and Pang [60, 1990] have proposed an algorithm that only generates *non-degenerate iterates* (i.e., without doubly active index), so that a single linear system needs to be solved at each iteration. The iterates are found by carefully choosing the stepsize along the Newton direction. Their algorithm is, however, probably not polynomial, since it exhibits poor iteration counters on some LCP examples. In the next section 5.2.3, we propose another descent direction, that is defined by a linear optimization problem and is therefore computationally tractable.

5.2.3 The ℓ_1-norm direction

The strategy consists in solving the linear complementarity problem (5.1.1), written as follows

$$\begin{cases} \inf\ \Theta_1(x) \\ x \geqslant 0 \\ Mx+q \geqslant 0. \end{cases} \tag{5.2.6}$$

where Θ_1 is the ℓ_1-norm of $H(x)$. It is clear that (5.2.6) is equivalent to $\mathrm{LC}(M,q)$ if this problem has a solution. The constraint $x \geqslant 0$ is added for ensuring descent of the computed directions (see proposition 5.7), while the constraint $Mx+q$ is added for a symmetry reason and because it gives to Θ_1 an interesting property. Observe indeed that if $x \geqslant 0$ and $Mx+q \geqslant 0$ are maintained, the ℓ_1-norm of $H(x)$ is linear in $H(x)$ and it is defined as follows

$$\Theta_1(x) := \|H(x)\|_1 = e^\top H(x) = \textstyle\sum_{i=1}^n \min(x_i, (Mx+q)_i),$$

where e is the vector of all ones.

We note by

$$X := \{x \in \mathbb{R}^n : x \geqslant 0,\ Mx+q \geqslant 0\},$$

the feasible set of the problem.

5.2.3.1 Study of the formulation

The formulation (5.2.6) has the attractive features given in proposition 5.11.

Proposition 5.11 *Assume that $X \neq \varnothing$. Then the merit function Θ_1 has the following properties.*
(i) It is piecewise linear and concave on X.
(ii) It is coercive on X if and only if $M \in \mathbf{R}_0$.

PROOF. [*(i)*] For a fixed $i \in [\![1,n]\!]$, $x \in \mathbb{R}^n \mapsto (x_i, (Mx+q)_i) \in \mathbb{R}^2$ is affine and $(u,v) \in \mathbb{R}^2 \mapsto \min(u,v) \in \mathbb{R}$ is concave. Hence, on X, Θ_1 is concave as a sum of concave functions.

[*(ii)*] If $M \notin \mathbf{R}_0$, there exists a nonzero direction d such that $0 \leqslant d \perp (Md) \geqslant 0$ (see section 2.2.9). Then X is unbounded since, when $x \in X \neq \varnothing$, X contains the half-line of points $x^t := x + td$ for all $t \geqslant 0$. Now, for $t \geqslant 0$ sufficiently large and $i \in [\![1,n]\!]$, there holds

$$\min(x_i^t, (Mx^t+q)_i) = \begin{cases} x_i & \text{if } d_i = 0 \text{ and } (Md)_i > 0 \\ (Mx+q)_i & \text{if } d_i > 0 \text{ and } (Md)_i = 0 \\ \min(x_i, (Mx+q)_i) & \text{if } d_i = 0 \text{ and } (Md)_i = 0. \end{cases}$$

Therefore $\Theta_1(x^t)$ is bounded when $t \to +\infty$, implying that Θ_1 is not coercive on X.

Suppose now that $M \in \mathbf{R}_0$. Then Θ is coercive (proposition 5.1). Now the fact that $\Theta_1(x) = \|H(x)\|_1$ on X and the equivalence of norms on $\mathbb{R}^n$ imply that Θ_1 is coercive on X. □

The concavity property of Θ_1 is interesting, since it implies that there will be no need to use line-search to obtain monotonic decrease of Θ_1 at each iteration; the unit step size is always suitable provided it maintains the next iterate in the feasible set X. This also implies that the global convergence of the algorithm will not be prevented by step-sizes converging to zero, only the quality of the descent direction will intervene. Now, minimizing a concave function on a polyhedron highlights the possible combinatorial aspect of the formulation, which is less attractive. It is therefore important to identify a

class of matrices for which this formulation does not introduce parasitic stationary points (this is the subject of proposition 5.13 below).

The coercivity property of Θ_1 on X is also interesting, since any algorithm forcing the decrease of Θ_1 on X at each iteration will generate bounded sequences, making the convergence analysis easier. Recall that $\mathbf{R}_0$ is a very large class of matrices, including the **P**-matrices.

The next lemma gives another property of **P**-matrices that will be used in the proof of proposition 5.13.

Lemma 5.12 *If M is a* **P***-matrix, then for any partition (I, J, K) of $[\![1,n]\!]$, the system*

$$\begin{cases} x_I = 0 \\ (Mx+q)_J = 0 \\ 0 \leqslant x_K \perp (Mx+q)_K \geqslant 0. \end{cases} \tag{5.2.7}$$

has (at least) a solution.

PROOF. Since M_{JJ} is nonsingular as a principal submatrix of M, the first two equations in (5.2.7) yield

$$x_J = -M_{JJ}^{-1} M_{JK} x_K - M_{JJ}^{-1} q_K.$$

Therefore the full system amounts to solving the following linear complementarity problem in x_K :

$$0 \leqslant x_J \perp M_{JJ}^s x_J + \left(q_K - M_{KJ} M_{JJ}^{-1} q_J\right) \geqslant 0, \tag{5.2.8}$$

where $M_{JJ}^s := M_{KK} - M_{KJ} M_{JJ}^{-1} M_{JK}$.

Since $M_{(J\cup K)(J\cup K)}$ is a **P**-matrix as a principal submatrix of M and M_{JJ}^s is a **P**-matrix as the Schur complement of M_{JJ} in the **P**-matrix $M_{J\cup K)(J\cup K)}$ (proposition 5.9), the linear complementarity problem (5.2.8) has a solution. □

Below, we say that $d \in \mathbb{R}^n$ *is a feasible direction at* $x \in X$ if $x + td \in X$ for some $t > 0$ (and therefore for all sufficiently small $t > 0$, by the convexity of X). We denoted by $\mathrm{Sol}(M,q)$ the set of solutions of the linear complementarity problem $\mathrm{LC}(M,q)$.

Proposition 5.13 (no parasitic stationary point) *If M is a* **P***-matrix and if a point $x \in X$ is such that $\Theta_1'(x;d) \geqslant 0$ for all feasible direction d at x, then $x \in \mathrm{Sol}(M,q)$.*

PROOF. Let A_0, I_0, and E three index sets defined by (5.2.2). By proposition 5.12, the following linear complementarity problem in $d \in \mathbb{R}^n$

$$\left\{ \begin{array}{l} (x+d)_{A_0} = 0 \\ (Mx+q+Md)_{I_0} = 0 \\ 0 \leqslant (x+d)_E \perp (Mx+q+Md)_E \geqslant 0 \end{array} \right. \tag{5.2.9}$$

has a solution, say $\hat{d}$. Observe that, with the definition of A_0, I_0 and E and the fact that $x \in X$:

$$\begin{array}{cc} (x+\hat{d})_{A_0 \cup E} \geqslant 0, & x_{I_0} > (Mx+q)_{I_0} \geqslant 0, \\ (Mx+q+M\hat{d})_{I_0 \cup E} \geqslant 0, & (Mx+q)_{A_0} > x_{A_0} \geqslant 0. \end{array}$$

Then $x+t\hat{d} \geqslant 0$ and $Mx+q+tM\hat{d} \geqslant 0$ for some $t > 0$, which shows that the direction $\hat{d}$ is feasible at x.

On the other hand, by the third condition in (5.2.9) and $x_E = (Mx+q)_E$, one has

$$0 = \min((x+\hat{d})_E, (Mx+q+M\hat{d})_E) = x_E + \min(\hat{d}_E, (M\hat{d})_E). \tag{5.2.10}$$

Therefore

$$\begin{array}{rcll} \Theta_1'(x;\hat{d}) & = & e^\top H'(x;\hat{d}) & \\ & = & e_{A_0}^\top \hat{d}_{A_0} + e_E^\top \min(\hat{d}_E, (M\hat{d})_E) + e_{I_0}^\top (M\hat{d})_{I_0} & \text{[proposition 5.2]} \\ & = & -e_{A_0}^\top x_{A_0} - e_E^\top x_E - e_{I_0}^\top (Mx+q)_{I_0} & \text{[(5.2.9) and (5.2.10)]} \\ & = & -\Theta_1(x). & (5.2.11) \end{array}$$

We can now conclude. Since, by assumption, $\Theta_1'(x;d) \geqslant 0$ for all feasible d and since $\hat{d}$ has been shown to be feasible, it follows that $\Theta_1'(x;\hat{d}) \geqslant 0$. Then, since $\Theta_1(x) \geqslant 0$ by the feasibility of x, (5.2.11) implies that $\Theta_1(x) = 0$, meaning that $x \in \mathrm{Sol}(M,q)$. □

5.2.3.2 *Search direction problem*

Let us now design an algorithm for minimizing Θ_1 on X. Let $J \equiv J(x)$ by a "Jacobian" at x of the nonsmooth map $x \mapsto H(x) := \min(x, Mx+q)$, whose ith line is

$$J_i = \begin{cases} I_i & \text{if } i \in A = A_0 \cup A_1 \\ M_i & \text{if } i \in I = I_0 \cup I_1, \end{cases} \tag{5.2.12}$$

where the index sets A_0, I_0, and E are defined by (5.2.2), and $A_1 \cup I_1$ is an arbitrary partition of E. We have denoted by I_i (resp. M_i) the ith row of the identity matrix (resp. of M). Below, we shall say that J is associated with the index set A, which satisfies $A_0 \subset A \subset A_0 \cup E$. A direction d is then defined as a solution to the linear optimization problem

$$\begin{cases} \inf\ e^\top J d \\ x + d \geqslant 0 \\ (Mx+q) + Md \geqslant 0. \end{cases} \tag{5.2.13}$$

It can be viewed as an SQP-like direction for problem (5.2.6), in which the "curvature" information is neglected (see part III in [17, 2006] for an introduction to the SQP algorithm). Note that the objective of (5.2.6) is not differentiable, so that its "curvature" is not well defined, and that the constraints are affine. The optimality conditions of (5.2.13) can be written

$$\begin{cases} (a) & J^\top e - s - M^\top z = 0 \\ (b) & 0 \leqslant (x+d) \perp s \geqslant 0 \\ (c) & 0 \leqslant (Mx+q+Md) \perp z \geqslant 0, \end{cases} \tag{5.2.14}$$

where $s \in \mathbb{R}^n$ [resp. $z \in \mathbb{R}^n$] is the vector of multipliers associated with the constraint $x+d \geqslant 0$ [resp. $(Mx+q)+Md \geqslant 0$].

Proposition 5.14 (well posedness and directional derivative) *Let $x \in X$. Then, problem (5.2.13) has a solution and any solution d to problem (5.2.13) is a non-strict descent direction of Θ_1 at x that satisfies*

$$-\Theta_1(x) \leqslant \Theta_1'(x;d) \leqslant e^\top J d \leqslant 0.$$

PROOF. To show that the linear optimization problem (5.2.13) has a solution, it is sufficient to show that it is feasible and that its lagrangian dual is feasible. Since $x \in X$, the problem is feasible (by taking $d = 0$). As for the dual, it can be written

$$\begin{cases} \sup\ -x^\top s - (Mx+q)^\top z \\ s \geqslant 0 \\ z \geqslant 0 \\ s + M^\top z = J^\top e. \end{cases} \tag{5.2.15}$$

This problem is clearly feasible : take $s_i = 1$ if $i \in A$, $s_i = 0$ if $i \in I$ and $z = e - s$. Therefore problem (5.2.13) has a solution.

On the other hand, by proposition 5.2, one has

$$\Theta_1'(x;d) = e^\top H'(x;d) = e_{A_0}^\top d_{A_0} + e_E^\top \min(d_E, (Md)_E) + e_{I_0}^\top (Md)_{I_0}. \tag{5.2.16}$$

Now, by the inequality constraints of problem (5.2.13), $d \geqslant -x$ and $Md \geqslant -(Mx+q)$, so that, $\min(d_E, (Md)_E) \geqslant -x_E = -(Mx+d)_E$ and because $e \geqslant 0$, we obtain

$$\Theta_1'(x;d) \geqslant -e_A^\top x_A - e_I^\top (Mx+q)_I = -\Theta_1(x).$$

To show that $\Theta'(x;d)$ is nonpositive, we start with the expression in the right hand side of (5.2.16) :

$$\begin{array}{rcll} \Theta_1'(x;d) & \leqslant & e_A^\top d_A + e_I^\top (Md)_I & [e_E \geqslant 0] \\ & = & e^\top Jd & [\text{definition of } J] \\ & \leqslant & 0 & [0 \text{ is feasible for (5.2.13)}]. \end{array}$$

□

Remark 5.15 *The proof of proposition 5.14 has shown that the bound $\Theta_1'(x;d) \geqslant -\Theta_1(x)$ is essentially due to the constraints of* (5.2.13).

Let us note that, if $x \in X$ is a solution to $\mathrm{LC}(M,q)$, then the triple (d,s,z), with $d = 0$, $s = (e_A, 0_I)$, and $z = (0_A, e_I)$, is a solution to (5.2.14). This is clear for the complementarity

conditions (b) and (c) since x_A and $(Mx+q)_I = 0$ by the definition of A and I, while for (a), one has indeed

$$J^\top e - s - M^\top z = \begin{pmatrix} e_A + M_{IA}^\top e_I \\ M_{II}^\top e_I \end{pmatrix} - \begin{pmatrix} e_A \\ 0_I \end{pmatrix} - \begin{pmatrix} M_{IA}^\top e_I \\ M_{II}^\top e_I \end{pmatrix} = 0.$$

Proposition 5.16 below shows a kind of reciprocal to this observation : if $x \in X$ is not a solution to $\mathrm{LC}(M,q)$, then one can find a J such that $d = 0$ is not a solution to (5.2.13).

Proposition 5.16 (descent direction) *If M is a* **P***-matrix and if $x \in X$ is not a solution to $\mathrm{LC}(M,q)$, then for some matrix $J \in \mathbb{R}^{n\times n}$ satisfying (5.2.12), any solution d to (5.2.13) satisfies $\Theta_1'(x;d) < 0$.*

PROOF. Let (A_0, E, I_0) be the partition of $[\![1,n]\!]$ associated with $x \in X$ by (5.2.2). Since $M \in \mathbf{P}$ and since $x \in X$ is not a solution to $\mathrm{LC}(M,q)$, there is a feasible direction $\hat{d}$ at x such that $\Theta_1'(x;\hat{d}) < 0$ (proposition 5.13). Therefore, one can choose a $t > 0$ such that $x^t := x + t\hat{d}$ satisfies

$$x^t \in X, \qquad x^t_{A_0} < (Mx^t+q)_{A_0}, \qquad \text{and} \qquad x^t_{I_0} > (Mx^t+q)_{I_0}. \tag{5.2.17}$$

Let us now determine a Jacobian J satisfying (5.2.12) at x :

$$J_i = \begin{cases} I_i & \text{if } i \in A := \{i : x_i^t \leqslant (Mx^t+q)_i\} \\ M_i & \text{if } i \in I := \{i : x_i^t > (Mx^t+q)_i\}. \end{cases}$$

By this definition of A and I, and by (5.2.17), there hold

$$A_0 \subset A \subset A_0 \cup E \qquad \text{and} \qquad I_0 \subset I \subset I_0 \cup E.$$

Therefore

$$\begin{aligned} \Theta_1(x) + te^\top J\hat{d} &= e_A^\top x_A + e_I^\top (Mx+q)_I + te^\top J\hat{d} \\ &= e_A^\top x_A^t + e_I^\top (Mx^t+q)_I \qquad [\text{definition of } J] \\ &= \Theta_1(x^t) \qquad [\text{definitions of } A \text{ and } I]. \end{aligned}$$

This yields $e^\top J\hat{d} = \Theta_1'(x;\hat{d}) < 0$. Now, $t\hat{d}$ is feasible for (5.2.13) (since $x^t \in X$ by (5.2.17)), so that any solution d to the linear problem (5.2.13) verifies $e^\top Jd \leqslant e^\top J(t\hat{d}) < 0$ and the result follows from proposition 5.14. □

5.3 A convergent algorithm

In this section, we give a detailed description of the algorithm we propose for solving the linear complementarity problem (5.1.1).

Algorithm 5.17 The algorithm generates pairs $(x, \mathscr{C})$ formed of an iterate $x \in X$ and a collection of discarded index sets $\mathscr{C} = \{A_1, \ldots, A_m\}$, where each $A = A_i$ is a subset of $[\![1, n]\!]$, with which we associate the closed convex polyhedron

$$P_A := \{x \in X : x_A \leqslant (Mx+q)_A,\ x_{A^c} \geqslant (Mx+q)_{A^c}\}.$$

At the beginning, the first iterate x is given in X and $\mathscr{C}$ is set to $\varnothing$. One iteration, from $(x, \mathscr{C})$ to $(x^+, \mathscr{C}^+)$, works as follows.

1. If all the sets A such that $x \in P_A$ are in $\mathscr{C}$, exit from the solver, otherwise, pick an index set $A \notin \mathscr{C}$ such that $x \in P_A$.
2. Set $J_A := I_A$, $J_{A^c} := M_{A^c}$, and solve (5.2.13) for d.
3. If $e^\top J d < 0$, set $x^+ := x + d$ and $\mathscr{C}^+ := \varnothing$ (the iteration is completed) otherwise, set $\mathscr{C} := \mathscr{C} \cup \{A\}$ and pursue the iteration at step 1.

Let us now give some comments on Algorithm 5.17 in order to get a better insight into the method.

1. In step 3 : when $d = 0$, instead of setting $\mathscr{C} = \varnothing$, one could set $\mathscr{C} = \{A\}$ if both x and x^+ are in P_A, since then this one has already been explored by the finishing iteration.
2. The algorithm copes with the combinatorics at the current point in an inelegant and straightforward manner, since it explores all the possible ways of dealing with the indices in E, looking for a descent direction d, hence satisfying $\Theta'(x; d) < 0$. Therefore, this algorithm is not polynomial.
3. The next iterate $x^+ \in X$, since d is defined as a solution to the linear optimization problem (5.2.13), then $x^+ = x + d \geqslant 0$ and $(Mx+q) + Md = Mx^+ q \geqslant 0$.

The main interest of this algorithm is therefore to show that global convergence is possible with the min C-function (Theorem 5.18) and a **P**-matrix M.

5.3.1 Convergence

We are now ready to establish the global convergence of the algorithm described above.

Theorem 5.18 (convergence of algorithm 5.17) *If M is a* **P***-matrix, then algorithm 5.17 is well defined and generates a sequence that converges to the unique solution to* $\mathrm{LC}(M,q)$ *in a finite number of iterations.*

PROOF. Since $M \in \mathbf{P} \subset \mathbf{S}$, the linear optimization problem (5.2.13) that must be solved at each iteration has always a solution d (proposition 5.14), so that algorithm 5.17 is well defined.

Let us now show that when the algorithm exits in step 1, x is the solution to the problem. In the exit situation, for all the pseudo-Jacobians J associated with any set A such that $x \in P_A$, the computed solution d to (5.2.13) is such that $e^\top Jd = 0$ (otherwise by the first option in step 3, $\mathscr{C}$ would have been set to $\varnothing$ and the algorithm would have generated an iterate different from x). By proposition 5.16, this means that x is the solution to $\mathrm{LC}(M,q)$.

Observe now that the algorithm cannot stagnate at a non-solution point x, since after a finite number of iterations it has explored all the possible index sets A such that $x \in P_A$ and then either terminates in step 1 with a solution or goes to a different point in step 3.

Let us now show that, the iterate x^+ following x is such that

$$\Theta_1(x^+) \leqslant \min\{\Theta_1(x) : x \in P_A\} \qquad \text{and} \qquad \Theta_1(x^+) < \Theta_1(x). \tag{5.3.1}$$

Observe first that the very definition of the step $x^+ - x$ by (5.2.13) shows that x^+ minimizes on X_2 the function $\varphi_A : z \mapsto \Theta_1(x) + e^\top J(z-x)$, in which J is determined in (5.2.12) by the choice of an index set A such that $x \in P_A$. Observe also that $\Theta_1 = \varphi_A$ on P_A, since for $z \in P_A$:

$$\begin{aligned}
\varphi_A(z) &= \Theta_1(x) + e^\top J(z-x) \\
&= e_A^\top x_A + e_I^\top (Mx+q)_I + e_A^\top (z-x)_A + e_I^\top M_I(z-x) \qquad [x \in P_A] \\
&= e_A^\top z_A + e_I^\top (Mz+q)_I \\
&= e^\top \min(z, Mz+q) \qquad [z \in P_A] \\
&= e^\top H(z) \\
&= \Theta_1(z).
\end{aligned} \tag{5.3.2}$$

Now, because x^+ may not be in P_A, the first inequality in (5.3.1) is not straightforward ; it results from

$$\begin{aligned}
\Theta_1(x^+) &\leqslant \Theta_1(x) + \Theta_1'(x; x^+ - x) && [\text{concavity of } \Theta_1 \text{ (proposition 5.11)}] \\
&\leqslant \Theta_1(x) + e^\top J(x^+ - x) && [\text{proposition 5.14}] \qquad (5.3.3) \\
&= \varphi_A(x^+) && [\text{definition of } \varphi_A] \\
&= \min\{\varphi_A(x) : x \in X_2\} && [x^+ \text{ minimizes } \varphi_A \text{ on } X_2] \\
&\leqslant \min\{\varphi_A(x) : x \in P_A\} && [P_A \subset X_2] \\
&= \min\{\Theta_1(x) : x \in P_A\} && [\Theta_1 = \varphi_A \text{ on } P_A \text{ by (5.3.2)}],
\end{aligned}$$

while the second inequality in (5.3.1) results from (5.3.3) and the fact that $e^\top J(x^+ - x) < 0$ (step 3 of the algorithm).

One can now conclude the convergence proof. Observe first that any iterate, say x', following x^+, if any, is not in P_A (since $\Theta_1(x') < \Theta_1(x^+) \leqslant \min\{\Theta_1(x) : x \in P_A\}$). Then, because there is a finite number of polyhedra like P_A, the algorithm will end up with an iterate x in a polyhedron P_A containing the solution. In that case, since x^+ minimizes Θ_1 on P_A, x^+ is the solution to the problem. Hence, the solution is found in a fnite number of iterations. □

5.4 Numerical examples

In this section, we present some numerical results for algorithm 5.17, which have been tried to several examples. We have implemented the algorithm in MATLAB and the termination criterion used is $||H(x^k)|| \leqslant 10^{-7}$.

The first two examples consists of LCP's with matrices M for which the plain Newton-min algorithm 3.1 cycles (see section 3.3). The first example (example 3.2) is defined in section 3.3.1, we have taken $\alpha = 2$. The second example (example 3.5) is defined in the section 3.3.2. For the parameters, we have taken the values $\alpha = 4/3$ and $\beta = 1/2$. In these cases, algorithm 5.17 converges in just one iteration for several values of the dimension n as indicated by Table 5.4.1.

Dimension \ Example	Example 3.2	Example 3.5
$n = 50$	1	1
$n = 100$	1	1
$n = 200$	1	1
$n = 500$	1	1

Table 5.4.1 – Number of iterations for example 3.2 and example 3.5

As can be seen from Table 5.4.1, our algorithm has no difficulties to solve these problems.

Example 5.19 *Murty's example ([103, 1988])*

$$M_1 = \begin{pmatrix} 1 & 2 & 2 & \cdots & 2 \\ 0 & 1 & 2 & \cdots & 2 \\ 0 & 0 & 1 & \cdots & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \qquad \textit{and} \qquad q = -1.$$

Obviously, the matrix M_1 is a **P***-matrix ; and even a* **PSD***-matrix (since $M + M^\top = 2ee^\top$, so that $x^\top Mx = \frac{1}{2}x^\top(M+M^\top)x = (e^\top x)^2 \geqslant 0$). The chosen starting point is $x^0 = 0$ and the solution to this problem is $x^* = (0,\ldots,0,1)$.*

Example 5.20 *Fathi's example ([39, 1979])*

$$M_2 = \begin{pmatrix} 1 & 2 & 2 & \cdots & 2 \\ 2 & 5 & 6 & \cdots & 6 \\ 2 & 6 & 9 & \cdots & 10 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 2 & 6 & 10 & \cdots & 4(n-1)+1 \end{pmatrix} \qquad \textit{and} \qquad q = -1.$$

Since $M_2 = M_1 M_1^\top$ and $M_1 \in \mathbf{P}$, M_2 is symmetric positive definite. Hence, M_2 is a **P***-matrix. The solution to this problem is $x^* = (1,0,\cdots,0)$. The chosen starting point is $x^0 = 0$.*

Examples 5.19 and 5.20 are standard test problems for which Lemke's pivot algorithm [86] is known to run in exponential time; see [103, 1988, chapter 6]. Furthermore, the number of iterations needed by the algorithm of Harker and Pang [60] (see section 5.2.2) seems to increase exponentially with the dimension n of these problems. On the other hand, the number of iterations needed by our algorithm to solve problem 5.19 for several values of the dimension n is indicated in Table 5.4.2. Table 5.4.3 reports the results of running the algorithm 5.17 on Fathi example 5.20. As one can see from the tables, our algorithm takes very few iterations to converge to the solutions. On the other hand, our algorithm clearly outperforms Harker and Pang's method

Dimension \ Algorithm	Harker and Pang	Algorithm 5.17
$n = 8$	9	1
$n = 16$	20	1
$n = 32$	72	1
$n = 64$	208	1
$n = 128$	> 300	1

Table 5.4.2 – Number of iterations for Murty's example 5.19

Dimension \ Algorithm	Harker and Pang	Algorithm 5.17
$n = 8$	8	2
$n = 16$	16	2
$n = 32$	32	2
$n = 64$	65	2
$n = 128$	63	2

Table 5.4.3 – Number of iterations for Fathi's example 5.20

Example 5.21 *In this example, M_1, M_2, and q are the same as in Murty's example 5.19 and Fathi's example 5.20, and the starting point x^0 is randomly generated with entries*

in $[-1000, 1000]$. *For each values of n, ten examples have been generated in this way and the average of the iterations needed by our algorithm are summarized in Table 5.4.4.*

Dimension \ Example	Murty's example	Fathi's example
$n = 50$	1	5
$n = 100$	1	7
$n = 200$	1	11
$n = 500$	1	13

Table 5.4.4 – Number of iterations for randomly generated starting point.

One observe that convergence occurs rapidly, despite the distance of the first iterate from the solution. Of course other experiment should be undertaken to assert the quality of the algorithm. We plan to do so in a next future.

5.5 Conclusion and perspectives

In this chapter, we have derived a new method for solving the linear complementarity problem $\mathrm{LC}(M,q)$ with a **P**-matrix. We have demonstrated that this method is robust easy to describe, and simple to implement. It is globally convergent and the numerical results reported above show that its performance is better than the ones of Harker and Pang's method. So, the algorithm 5.17 is a useful tool for solving $\mathrm{LC}(M,q)$ with a **P**-matrix. On the other hand, since it explores all the possible ways of dealing with the indices in E, this algorithm is not polynomial.

The work presented in this chapter can be pursued along at least two directions. One possibility is to modify the algorithm in order to find a polynomial algorithm to solve the LCP with a P-matrix, for example, by selecting the descent piece in a clever manner in order to avoid exploring all of them. Another possibility is to modify the globalization technique and to use trust regions instead of the line-search technique.

Deuxième partie

Écoulement liquide-gaz en milieu poreux

La seconde partie de ce mémoire, qui regroupe les chapitres 6 à 9, traite des écoulements diphasiques en milieu poreux et plus particulièrement du cas du mélange eau-hydrogène dans le cadre de l'étude du stockage des déchets radioactifs en milieu géologique. On y élabore un modèle mathématique utilisant des conditions de complémentarité non linéaires décrivant ces écoulements. Une méthode de résolution efficace et un solveur robuste pour les problèmes de complémentarité non linéaires sont construits. Ce travail donne lieu à la simulation des cas-tests proposés par le GNR MoMaS [1] et l'ANDRA [2].

Nous présentons, dans le chapitre 6, une formulation permettant d'éviter la dégénérescence mathématique lors de la disparition de la phase gazeuse. Dans le chapitre 7, nous proposons ainsi une méthode de résolution et un solveur pour ce problème. Nous présentons également différents cas-tests permettant de montrer le bon traitement de l'apparition et de la disparition de la phase gazeuse et l'éfficacité de notre solveur (voir chapitres 8 et 9). Nous comparons les résultats obtenus lors du benchmark « Écoulement diphasique » dans le cadre du GNR MoMaS.

1. Le Groupement MoMaS : Modélisations Mathématiques et Simulations Numériques liées aux problèmes de gestion des déchets nucléaires [114]
2. Agence Nationale pour la gestion des Déchets Radioactifs [5]

Chapitre 6

Écoulement diphasique avec échange entre les phases comme un problème de complémentarité non linéaire

Nous présentons dans ce chapitre une formulation mathématique consistante qui permet de décrire un mélange eau-hydrogène dans un milieu poreux en présence de deux phases liquide et gaz, mais aussi lorsque la phase gaz est absente. L'approche proposée conduit à un problème d'équations aux dérivées partielles non linéraires avec des conditions de complémentarité non linéaires.

6.1 Problématique

Comme toute activité industrielle, l'industrie nucléaire produit des déchets. L'objectif fondamental de la gestion à long terme des déchets est de protéger, sur une très grande durée, l'homme et l'environnement des risques associés à ces déchets. Des précautions considérables sont prises pour les gérer. Ceux-ci sont, bien entendu, soumis à une réglementation très précise et à des contrôles fréquents et approfondis.

Le stockage consiste à confiner ces déchets dans une formation géologique profonde pour s'opposer à la dissémination des radionucléides qu'elle contient. Ce confinement devra s'effectuer sur des grandes échelles de temps (jusqu'à plusieurs centaines de milliers d'années), de manière passive, c'est-à-dire sans nécessiter de maintenance ou de

surveillance à très long terme.

Le travail de cette partie s'inscrit dans le contexte du projet de stockage en formation géologique profonde de déchets radioactifs à haute activité et vie longue. Les études inhérentes au stockage de déchets en formation géologique profonde menées par l'ANDRA s'appuient sur les caractéristiques du site et s'attachent à évaluer les conditions dans lesquelles on pourrait construire, exploiter et surveiller un site de stockage.

Un site de stockage géologique présente une grande diversité de phénomènes physiques. Un aspect problématique est la production, par le stockage lui-même, d'hydrogène issu de la dégradation chimique de certains matériaux du stockage. Ce phénomène génère une certaine quantité d'hydrogène et peut s'étaler sur une longue période de temps. Cet hydrogène peut être dissous dans l'eau résidente ou être sous forme gazeuse. L'objectif de la modélisation est donc de comprendre ce qui adviendra de cet hydrogène, son éventuelle accumulation et sa migration.

Dans un premier temps la quantité d'hydrogène produite sera faible, l'hydrogène sera dissous, il n'y aura qu'une phase liquide et l'écoulement sera monophasique. Mais dans un deuxième temps, l'hydrogène apparaîtra aussi sous forme gazeuse et l'écoulement sera diphasique liquide-gaz. Dans un troisième temps, lorsque la production d'hydrogène s'arrêtera et que l'hydrogène aura migré, la phase gazeuse disparaîtra et l'écoulement redeviendra monophasique. Ces changements sont locaux.

Pour rendre compte de ces apparitions et disparitions de phases nous introduirons une formulation élégante et efficace utilisant des contraintes de complémentarité. (voir chapitre 7). Nous proposons ainsi une méthode de résolution et un solveur pour cette formulation. Nous présenterons également différents cas tests permettant de montrer le bon traitement de l'apparition et de la disparition de la phase gazeuse.

6.2 Physique du problème

On décrit dans cette section les différents phénomènes dans le transport dans un milieu poreux saturé par « un fluide » composé de deux phases comprenant deux composants. Dans notre cas, la migration de l'hydrogène au sein d'un site de stockage géologique, les deux phases considérées sont la phase liquide et la phase gazeuse ; les deux composants considérés sont l'eau (constituant essentiel de la phase liquide) et l'hydrogène (constituant essentiel de la phase gazeuse). Pour chaque composant, deux types de transports sont possibles au sein d'une phase : *la convection* due à l'écoulement de la phase et *la diffusion*

due aux différences de concentration du composant dans la phase. De plus l'interaction des deux phases fait intervenir deux phénomènes supplémentaires : d'une part un phénomène capillaire, au niveau des pores, qui impose une différence entre les pressions de chaque phase, et d'autre part un phénomène thermodynamique qui relie les compositions de chaque phase [95].

Les écoulements souterrains s'intéressent à la circulation de l'eau à travers les pores d'un sol ou une roche déformable. Il y a une forte interaction entre l'eau et le comportement des sites de stockage souterrain des déchets nucléaires.

Trois hypothèses essentielles sont faites dans la modélisation physique qui suit : le milieu poreux est supposé indéformable ; la température du fluide est supposée constante ; le fluide est supposé être à l'équilibre thermodynamique à chaque instant.

6.2.1 Le milieu poreux

Un milieu poreux est un milieu composé d'une matrice solide et d'un espace vide pouvant être occupé par un ou plusieurs fluides. On observe ce milieu poreux à une échelle telle qu'on peut le considérer comme un milieu continu. Un point de l'espace représente un volume élémentaire représentatif (VER) de mileu poreux et on définit la porosité en ce point comme le rapport dans ce volume élémentaire entre le volume de pore où les fluides peuvent circuler et le volume total de ce volume élémentaire (volume de pore + volume de roche). On note cette quantité ϕ et l'on suppose qu'elle ne dépend que de l'espace (milieu poreux indéformable).

Lorsque l'espace poreux est occupé par une seule phase fluide (liquide ou gaz), l'écoulement est dit monophasique et lorsqu'il est occupé par plusieurs phases fluides, l'écoulement est dit multiphasique.

6.2.2 Deux phases et deux composants dans un écoulement en milieu poreux

L' écoulement que nous considérons est supposé composé de deux phases, phase liquide et phase gaz, avec deux composants, l'eau et l'hydrogène. Les quantités relatives aux phases seront munies de l'indice $i = \ell$ pour la phase liquide ou $i = g$ pour la phase gazeuse, et les quantités relatives aux composants seront munies de l'indice $j = w$ pour le composant eau ou $j = h$ pour le composant hydrogène. Pour chaque phase $i = \ell, g$, on

note p_i sa pression et ρ_i sa densité massique. On définit enfin la saturation de la phase is_i comme le rapport entre le volume occupé par la phase i et le volume total de pore pour un volume élémentaire représentatif. Puisque le volume poreux est entièrement occupé par les fluides , on a la relation suivante

$$s_\ell + s_g = 1, \quad 0 \leqslant s_i \leqslant 1, \quad i = \ell, g.$$

Chaque phase est-elle même constituée de deux composants : eau et hydrogène. La composition de chaque phase i peut être caractérisée par les concentrations massiques des composants eau et hydrogène au sein de cette phase. On notera ρ_j^i la concentration massique du composant $j = w, h$, dans la phase $i = \ell, g$. La densité massique de chaque phase correspond à la somme des concentrations massiques de tous les composants qu'elle contient. On a donc

$$\rho_i = \rho_w^i + \rho_h^i, \quad i = \ell, g.$$

6.2.3 Conservation de la masse pour chacun des composants

Dans chaque phase, la migration du composant j est due pour une part au transport par l'écoulement de la phase et pour une autre part à la diffusion moléculaire au sein de la phase. On a donc les équations aux dérivées partielles suivantes, qui traduisent la conservation de la masse pour l'eau et l'hydrogène,

$$\begin{aligned} &\frac{\partial}{\partial t}(\phi \rho_w^\ell s_\ell + \phi \rho_w^g s_g) + \operatorname{div}(\rho_w^\ell \mathbf{q}_\ell + \rho_w^g \mathbf{q}_g + J_w^\ell + J_w^g) = Q_w, \\ &\frac{\partial}{\partial t}(\phi s_\ell \rho_h^\ell + \phi s_g \rho_h^g) + \operatorname{div}(\rho_h^\ell \mathbf{q}_\ell + \rho_h^g \mathbf{q}_g + J_h^\ell + J_h^g) = Q_h, \end{aligned} \tag{6.2.1}$$

où Q_j, $j = w, h$, est le débit de composant j sortant, $\mathbf{q}_\ell$ et $\mathbf{q}_g$ sont les vitesses de Darcy des phases liquide et gazeuse et J_j^i est le flux diffusif effectif du composant j, $j = w, h$, dans la phase i, $i = \ell, g$. Les expressions de $\mathbf{q}_i$ et J_j^i seront donnńes dans les paragraphes suivants.

6.2.4 La loi de Darcy généralisée

L'écoulement monophasique dans un milieu poreux saturé est décrit par la loi de Darcy. Pour un fluide de densité massique ρ, de viscosité dynamique μ et de pression

p la loi de Darcy exprime **q**, la vitesse d'écoulement du fluide saturant, par

$$\mathbf{q} = -\frac{1}{\mu}K(x)(\nabla p - \rho g \nabla z),$$

où g est l'accélération de la pesanteur et $K(x)$ le tenseur de perméabilité intrinsèque absolue du milieu poreux au point x et z la cote au point d'espace x. La loi de Darcy-Muskat généralise cette relation aux écoulements multiphasiques en introduisant une notion de fonction de perméabilité relative qui s'écrit

$$\mathbf{q}_i = -K(x)\frac{k_{ri}(s_i)}{\mu_i}(\nabla p_i - \rho_i g \nabla z), \quad i = \ell, g, \tag{6.2.2}$$

où $k_{ri}(s_i)$ est la fonction de perméabilité relative de la phase i, c'est une fonction à valeur dans $[0;1]$, croissante et qui satisfait les deux égalités suivantes :

$$k_{ri}(s_i = 0) = 0 \quad \text{et} \quad k_{ri}(s_i = 1) = 1.$$

En d'autres termes, la loi de Darcy-Muskat est similaire à la loi de Darcy à la différence que la perméabilité absolue K est remplacé par une perméabilité "effective" $K(x)k_{ri}$. La figure 6.2.1 représente l'allure des courbes de $k_{r\ell}$ et k_{rg}.

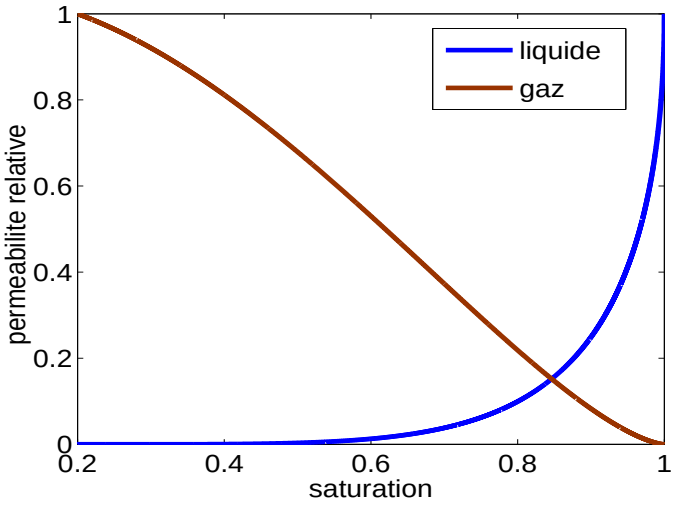


Figure 6.2.1 – Modèle de van Genuchten pour la perméabilité relative.

6.2.5 Diffusion moléculaire

Notons M^j la masse molaire du composant j, on définit c_i^i la concentration molaire du composant j dans la phase i par

$$c_j^i = \frac{\rho_j^i}{M^j}, \quad j = w,h, \quad i = \ell,g.$$

La concentration molaire de la phase est alors la somme des concentrations molaires des composants présents dans la phase ; dans le cas eau/hydrogène, on a donc

$$c_i = c_w^i + c_h^i = \frac{\rho_w^i}{M^w} + \frac{\rho_h^i}{M^h}, \quad i = \ell,g.$$

Et la fraction molaire du composant j dans la phase i est le rapport de la concentration molaire du composant sur la concentration molaire de la phase :

$$\chi_h^i = \frac{c_h^i}{c_i}, \quad \chi_w^i = \frac{c_w^i}{c_i}; \quad \chi_w^i + \chi_h^i = 1, \quad i = \ell,g. \tag{6.2.3}$$

Finalement, on définit le flux diffusif massique du composant j dans la phase i (voir équations (6.2.1)), par

$$J_j^i = -\phi M^j s_i c_i D_j^i \nabla \chi_j^i, \quad j = w,h, \quad i = \ell,g.$$

où D_j^i est le coefficient de diffusion moléculaire du composant j dans la phase i. Remarquons enfin que la diffusion moléculaire dans une phase n'apporte aucune contribution au déplacement globale de cette phase, i.e.

$$J_h^i + J_w^i = 0. \tag{6.2.4}$$

Dans notre cas, la relation (6.2.3) implique que $\nabla \chi_w^i = -\nabla \chi_h^i$.

6.2.6 Pression capillaire

De manière générale, la pression capillaire est définie comme la différence des pressions de deux fluides immiscibles de part et d'autre de l'interface les séparant. Cette différence est due à la courbure de l'interface, la pression du fluide du côté concave de

l'interface étant supérieure à celle du côté convexe. Dans un tube (ou dans un pore), le sens de la courbure de l'interface entre les deux fluides provient de la capacité plus ou moins grande de chaque fluide à "mouiller" la paroi solide. Ainsi dans le cas liquide/gaz, le liquide est (à de rares exceptions près) le fluide le plus mouillant et donc la pression du gaz est supérieure à celle du liquide. Les expérimentations font néanmoins apparaître que la pression capillaire dépend essentiellement de la saturation et ceci d'une manière monotone. Plusieurs facteurs affectent les propriétés capillaires des milieux poreux, parmi lesquels on peut peut noter : la dimension et la distribution des pores ; les fluides et les solides impliqués. Dans toute la suite, on exprimera la pression capillaire comme une fonction univoque de la saturation. Pour une configuration liquide-gaz, on définit donc

$$p_c(s_\ell) = p_g - p_\ell \geqslant 0.$$

Toujours dans le cas liquide/gaz, on observe qu'il existe une saturation résiduelle en liquide, $s_{l,res}$, qui correspond à la saturation où le liquide ne forme plus une phase continue dans les pores et ne peut alors plus s'écouler. La perte de continuité hydraulique implique que la pression capillaire peut croître indéfiniment sans pour autant réduire la saturation du liquide s_ℓ sous sa valeur résiduelle $s_{l,res}$. Autrement dit, on a

$$\lim_{\substack{s \to s_{l,res} \\ s > s_{l,res}}} p_c(s_\ell) = +\infty.$$

La pression capillaire est une fonction décroissante. Dans le modèle de van Genuchten [51] la pression capillaire s'annule pour la saturation en eau maximale avec une pente verticale. La figure 6.2.2 représente l'allure d'une courbe de pression capillaire.

6.3 Modélisation

6.3.1 Hypothèses physiques du modèle

Nous avons présenté, dans la section précédente, l'ensemble des phénomènes physiques impliqués dans l'écoulement diphasique eau/hydrogène en milieux poreux. Cette modélisation physique suppose que le milieu poreux est indéformable (porosité constante). Des hypothèses simplificatrices ont aussi été introduites, nous avons supposé

- que l'écoulement du mélange fluide est isotherme, i.e. la température est constante en tous points et en tous temps,

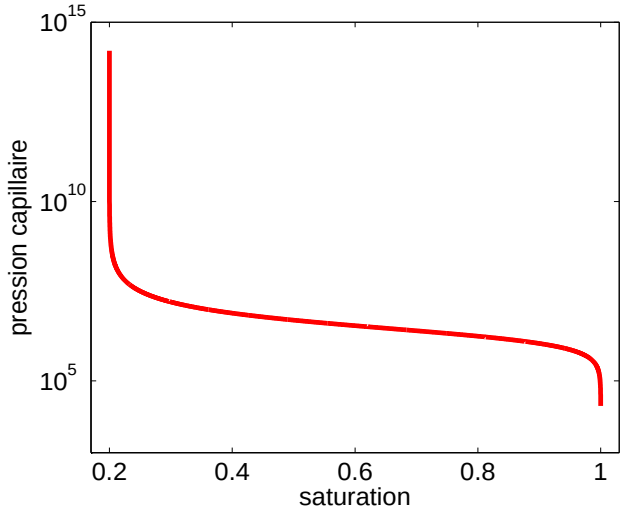


Figure 6.2.2 – Modèle de van Genuchten pour la pression capillaire.

- que l'eau est incompressible, i.e. ρ_w^ℓ est constante en tous points et en tous temps,
- que le gaz est légèrement compressible i.e. $\rho_g = C_g p_g$ où C_g est une constante de compressibilité,
- que l'eau n'est présente que dans la phase liquide tandis que l'hydrogène peut être présent dans les deux phases i.e.

$$\rho_w^g = 0, \quad \rho_g = \rho_h^g, \quad \chi_h^g = \frac{c_h^g}{c_g} = 1, \quad \chi_w^g = 0, \quad J_h^g = J_w^g = 0.$$

  Et on a aussi d'après la relation (6.2.4), $J_h^\ell = -J_w^\ell$.
- que pour la phase liquide, l'eau est le solvant et l'hydrogène le soluté et que la solution liquide est une solution diluée idéale. Ceci implique $c_h^l \ll c_w^l$, et on aura donc

$$\chi_h^\ell = \frac{c_h^\ell}{c^\ell} = \frac{c_h^\ell}{c_h^\ell + c_w^\ell} \approx \frac{c_h^\ell}{c_w^\ell} = \frac{M^w}{M^h \rho_w^\ell} \rho_h^\ell.$$

6.3.2 Équations de l'écoulement

En intégrant les hypothèse simplificatrices présentées dans la section 6.3.1 aux équations de conservation de la masse de chaque composant (6.2.1), on obtient les équations simplifiées suivantes

$$\begin{aligned} &\frac{\partial}{\partial t}(\phi \rho_w^\ell s_\ell) + \operatorname{div}(\rho_w^\ell \mathbf{q}_\ell - J_h^\ell) = Q_w, \\ &\frac{\partial}{\partial t}(\phi s_\ell \rho_h^\ell + \phi s_g \rho_h^g) + \operatorname{div}(\rho_h^\ell \mathbf{q}_\ell + \rho_h^g \mathbf{q}_g + J_h^\ell) = Q_h. \end{aligned} \tag{6.3.1}$$

Étant donné que le problème se pose dans un régime où la phase liquide sera toujours présente (seule la phase gazeuse peut être amenée à disparaître), il est commode de prendre comme inconnue principale, outre la saturation de la phase liquide s_ℓ, la pression dans la phase liquide p_ℓ.

6.3.3 Formulation utilisant la loi de Henry

Concernant les échanges entre phases, la quantité d'hydrogène dissout dans la phase liquide en présence de la phase gazeuse est donnée par la loi de Henry

$$H p_g = \rho_h^\ell, \quad \text{avec} \quad H = \tilde{H}(T) M^h.$$

où $\tilde{H}(T)$ est la constante de Henry et

$$\rho_h^\ell = \frac{M^h \rho_w^\ell}{M^w} \chi_h^\ell = C_\ell \chi_h^\ell \quad \text{avec} \quad C_\ell = \frac{M^h \rho_w^\ell}{M^w}.$$

Pour traiter en même temps le cas où la phase gazeuse n'existe pas, on peut formuler le problème sous la forme d'un problème de complémentarité de la façon suivante. En effet, soit la phase gazeuse existe, $1 - s_\ell > 0$ et la loi d'Henry s'applique, soit la phase gazeuse n'existe pas, $s_\ell = 1, p_c(s_\ell) = 0$ et $H p_\ell - \rho_h^\ell > 0$ ce qui signifie que pour une pression dans la phase liquide donnée p_ℓ la densité massique d'hydrogène est trop petite pour que l'hydrogène soit partiellement gazeux, ou encore que, pour une densité massique d'hydrogène donnée ρ_ℓ^h la pression p_ℓ est trop grande pour que l'hydrogène soit partiellement gazeux. On obtient ainsi les contraintes de complémentarité

$$(1 - s_\ell)\Big(H(p_\ell + p_c(s_\ell)) - C_\ell \chi_h^\ell\Big) = 0, \quad 1 - s_\ell \geqslant 0, \quad H(p_\ell + p_c(s_\ell)) - C_\ell \chi_h^\ell \geqslant 0. \tag{6.3.2}$$

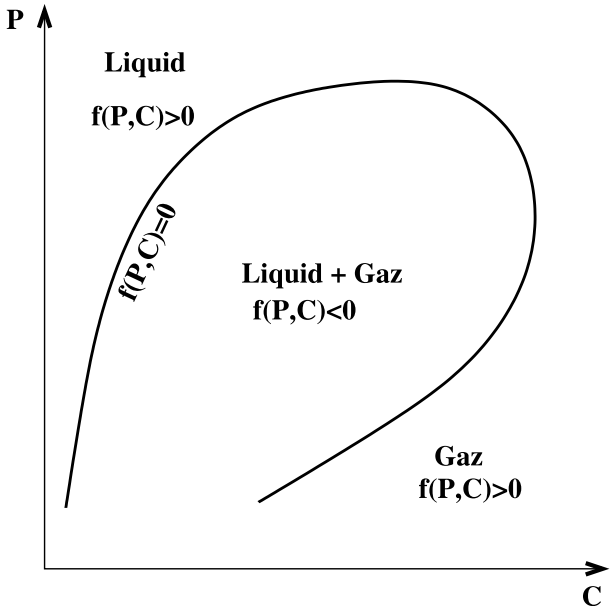


Figure 6.3.1 – Diagramme de phase usuel.

Ainsi le problème de la dissolution de l'hydrogène dans l'eau peut être formulé sous la forme d'un problème de complémentarité avec comme inconnues principales s_ℓ, p_ℓ, χ_h^ℓ avec s_ℓ, χ_h^ℓ des quantités comprises entre zéro et un, le problème étant singulier en $s_\ell = 1$ à cause de la tangente verticale de la pression capillaire en ce point.

Pour justifier cette formulation en terme de conditions de complémentarité, on peut utiliser un diagramme de phase.

6.3.4 Diagrammes de phase

Le diagramme de phase [65] est une figure telle que celle représentée sur la figure 6.3.1 où une courbe $f(C;P) = 0$ sépare les zones où le système est monophasique $f(P;C) > 0$ - liquide ou gaz - de celles où le système est diphasique $f(P;C) < 0$ - liquide + gaz. Il permet de répartir deux composants, par exemple l'hydrogène et l'eau, entre les deux phases liquide et gaz. Les variables P et C sont définies dans le tableau 6.3.1. On notera que ces variables et la saturation sont continues au travers de la courbe $f(P,C) = 0$. Dans

	liquid	liquid + gas	gas
	$s_\ell = 1$	$0 < s_\ell < 1$	$s_\ell = 0$
P	p_ℓ	$p_g = p_\ell + p_c(s_\ell)$	p_g
C	ρ_h^ℓ	$\rho_h^\ell + \rho_g$	ρ_g

Table 6.3.1 – Définitions des variables P et C dans les zones monophasique et diphasique.

la suite, on va se concentrer sur l'interface entre la zone liquide et la zone diphasique. La loi de Henry n'est définie que dans la zone diphasique. Dans cette zone, on a donc

$$HP - C = Hp_g - (\rho_h^\ell + \rho_g) = -\rho_g < 0.$$

et lorsqu'on tend vers la limite de séparation entre les zones liquide et diphasique, on a $HP - C = -\rho_g \to 0$. Au dessus de cette droite se trouve la zone liquide $HP - C = Hp_\ell - \rho_h^\ell > 0$. Ainsi, l'utilisation de la loi de Henry revient à remplacer la partie de la courbe de séparation $f(C;P) = 0$ qui sépare la zone liquide de la zone diphasique par la droite $HP - C = 0$. Ceci correspond au diagramme de phase de la figure 6.3.2.

6.3.5 Problème de complémentrarité non linéaire

Puisque la phase liquide sera toujours présente, on choisit comme inconnues principales s_ℓ, p_ℓ ainsi que χ_h^ℓ.

Notree problème s'écrit donc sous la forme du système suivant, formé d'équations aux dérivées partielles non linéaires et de conditions de complémentarité non linéaires

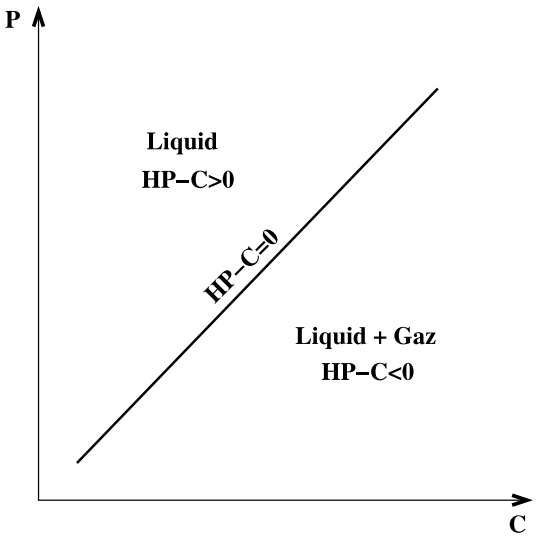


Figure 6.3.2 – Diagramme de phase usuel.

suivantes

$$
\begin{aligned}
&\frac{\partial}{\partial t}(\rho_w^\ell \phi\, s_\ell) + \mathrm{div}(\rho_w^\ell \mathbf{q}_\ell - J_h^\ell) = Q_w,\\
&\frac{\partial}{\partial t}(\phi s_\ell C_\ell \chi_h^\ell + \phi(1-s_\ell) C_g(p_\ell + p_c(s_\ell))) + \mathrm{div}(C_\ell \chi_h^\ell \mathbf{q}_\ell + C_g(p_\ell + p_c(s_\ell))\mathbf{q}_g + J_h^\ell) = Q_h,\\
&\mathbf{q}_\ell = -K(x) k_\ell(s_\ell)(\nabla p_\ell - (\rho_w^\ell + C_\ell \chi_h^\ell) g \nabla z),\\
&\mathbf{q}_g = -K(x) k_g(1-s_\ell)(\nabla(p_\ell + p_c(s_\ell)) - C_g(p_\ell + p_c(s_\ell)) g \nabla z),\\
&j_h^\ell = -M^h \phi\, s_\ell \,(\frac{\rho_w^\ell}{M_w} + \frac{C_\ell}{M_h}\chi_h^\ell) D_h^\ell \nabla \chi_h^\ell,\\
&(1-s_\ell)\left(M^h H(p_\ell + p_c(s_\ell)) - C_\ell \chi_h^\ell\right) = 0, \quad 1 - s_\ell \geqslant 0, \quad H(p_\ell + p_c(s_\ell)) - C_\ell \chi_h^\ell \geqslant 0.
\end{aligned}
\tag{6.3.3}
$$

D'une part, lors de l'apparition et la disparition de la phase gazeuse, cette formulation permet de modéliser le passage d'un écoulement monophasique à un écoulement diphasique et vice versa. D'autre part, grâce au choix pertinent des inconnues principales

(s_ℓ, p_ℓ, χ_h^ℓ), ces différents types d'écoulement possèdent les mêmes variables.

Chapitre 7

Problème de complémentarité non linéaire

7.1 Présentation du problème

De nombreux problèmes non linéaires contiennent des conditions de complémentarité qui sont de la forme :

$$F(x)^\top G(x) = 0, \quad F(x) \geqslant 0, \quad G(x) \geqslant 0.$$

où F et G : $\mathbb{R}^n \to \mathbb{R}^p$ $(p \leqslant n)$ sont des fonctions non linéaires différentiables des inconnues x. On peut citer les problèmes de contact, de changement de phase, de dissolution-précipitation et d'écoulement gaz-liquide avec apparition ou disparition d'une phase (voir chapitre 6).

Dans cette dernière discipline, ces problèmes sont souvent résolus numériquement de façon empirique en annulant certaines de leurs composantes de manière complémentaire, en testant au cours des itérations de Newton si le bon choix de composantes nulles à été fait et en espérant avoir la positivité à la convergence. Ces méthodes se heurtent à la combinatoire représentée par les 2^p manières d'annuler les composantes de $F(x)$ et de $G(x)$ de manière complémentaire. Elles ne reposent pas sur une analyse précise et sont d'ailleurs souvent mises en défaut.

Les méthodes de résolution des problèmes de complémentarité ont connu un essor important ces dernières années suites aux travaux des numériciens [20, 94, 57, 84]. L'état

de l'art est maintenant décrit dans des articles de synthèse et monographies, y compris lorsque l'optimisation entre en jeu [70, 105, 38, 104]. On y trouve essentiellement trois techniques : les techniques d'activation [70], les approches par points intérieurs [7, 8, 126] les méthodes de l'analyse non lisse [104, 105, 38]. La première technique est souvent considérées comme moins efficaces que les deux autres, si l'on ne dispose pas d'une estimation des functions actives (c'est-à-dire nulles) en la solution. Lorsque les systèmes non linéaires à résoudre sont issus de la discrétisation en espace et en temps d'une équation d'évolution, comme dans notre cas, les bornes actives (localisées en espace) au temps précédent sont souvent une bonne estimation de celles qui le sont au temps courant. Notre travail considère la troisième famille des techniques numériques.

Dans le contexte décrit ci-dessus, il y a plusieurs difficultés non classiques. La première difficulté provient de certains modèles de pression capillaire ou de perméabilité relative, tels que ceux de Van Genuchten (voir les figures 6.2.1 et 6.2.2), dans lesquels les non-linéarités sont des fonctions de la saturation qui présentent des tangentes verticales aux extrémités des intervalles de définition. Les problèmes contraints représentent une deuxième difficulté puisque la saturation ou la fraction molaire sont des quantités comprises entre 0 et 1. La résolution des problèmes de complémentarité est une troisième difficulté. Un des objectifs de cette partie est de proposer un algorithme alliant efficacité et robustesse, du type Newton non lisse, permettant de résoudre le problème de changement de phase en considération (voir le chapitre 6). Cet algorithme est implémenté, en Matlab, sous la forme d'un solveur, qu'on appelle Newton-min.

7.2 Formulations canoniques des problèmes de complémentarité non linéaires

7.2.1 De la complémentarité aux inéquations variationnelles

Dans cette section, nous nous intéressons aux formulations équivalentes d'un problème de complémentarité non linéaire de manière à les situer comme des cas particuliers des problèmes d'inéquations variationnelles.

La formulation qui nous intéresse est la suivante

$$\begin{cases} H(x) = 0 \\ 0 \leqslant F(x) \perp G(x) \geqslant 0, \end{cases} \tag{7.2.1}$$

où $F : \mathbb{R}^n \to \mathbb{R}^p$, $G : \mathbb{R}^n \to \mathbb{R}^p$, $H : \mathbb{R}^n \to \mathbb{R}^{n-p}$ et la deuxième relation exprime que les vecteurs $F(x)$ et $G(x)$ doivent avoir leurs composantes positives et doivent être orthogonaux (pour le produit scalaire euclidien). Les dimensions sont liées au fait que les contraintes de complémentarité comptent pour p relations ($F_i(x)G_i(x) = 0$ pour $i = 1, \ldots, p$, tandis que la positivité de $F(x)$ et $G(x)$ n'intervient pas dans le décompte des équations) et que l'on a besoin de n équations pour déterminer les n inconnues $x \in \mathbb{R}^n$. On pourrait augmenter la souplesse du modèle en y ajoutant des inégalités, mais nous n'en avons pas eu besoin.

Le problème défini par la formulation (7.2.1) est un cas particulier du problème général de complémentarité non linéaire

$$0 \leqslant F^*(x) \perp G^*(x) \geqslant 0. \tag{7.2.2}$$

En effet, la formulation (7.2.1) peut se réécrire sous la forme

$$0 \leqslant \begin{pmatrix} F(x) \\ H(x) \end{pmatrix} \perp \begin{pmatrix} G(x) \\ 1_{n-p} \end{pmatrix} \geqslant 0,$$

où 1_{n-p} est le vecteur de $\mathbb{R}^{n-p}$ dont toutes les composantes valent 1. D'un point de vue numérique, il est plus efficace de considérer directement (7.2.1) en se passant de l'artifice ci-dessus, puisqu'alors la combinatoire porte sur un vecteur plus petit (de dimension p plutôt que n).

On peut encore réécrire le problème de complémentarité (7.2.2), sous la forme

$$0 \leqslant x \perp F^{**}(x) \geqslant 0. \tag{7.2.3}$$

Il est clair que (7.2.3) est un cas particulier de (7.2.2), obtenu en prenant l'identité pour G. Inversement, (7.2.2) peut s'exprimer comme un problème (7.2.3) par

$$0 \leqslant \begin{pmatrix} u \\ v \\ y \\ a \\ b \end{pmatrix} \perp \begin{pmatrix} 0 \\ 0 \\ G(u-v) \\ y - F^*(u-v) \\ -y + F^*(u-v) \end{pmatrix} \geqslant 0,$$

où on a décomposé x en $u - v$ avec u et v positifs, a et b sont des variables auxiliaires sans signification et les deux dernières relations sont utilisées pour imposer $y = F^*(x)$. Ici aussi l'écriture de (7.2.2), et a fortiori de (7.2.1), sous la forme (7.2.3) se paie par une

augmentation du nombre des conditions de complémentarité et donc de la combinatoire du problème, ce qui devrait détériorer l'efficacité numérique.

Les problèmes équivalents (7.2.1), (7.2.2) et (7.2.3) sont des cas particuliers du *problème d'inégalités variationnelles* [38]. Un tel problème s'écrit

$$\begin{cases} x \in K \\ F^{**}(x)^\top (y-x) \geqslant 0, \quad \forall y \in K, \end{cases} \tag{7.2.4}$$

où $F^{**} : \mathbb{R}^n \to \mathbb{R}^n$ et K est une partie de $\mathbb{R}^n$. En fait, le problème d'inéquations variationnelles (7.2.5) est équivalent à un problème de complémentarité si K est un cône.

Observons d'abord que, lorsque K est un cône, le problème (7.2.4) devient

$$K \ni x \perp F^{**}(x) \in K^+, \tag{7.2.5}$$

qui signifie que $x \in K$, $F^{**}(x) \in K^+ := \{y \in \mathbb{R}^n : y^\top x \geqslant 0,\ \forall x \in K\}$ (le *cône dual positif* de K) et x est orthogonal à $F^{**}(x)$. Pour obtenir (7.2.5) à partir de (7.2.4), il suffit en effet de prendre $y = \dfrac{x}{2} \in K$ et $y = 2x \in K$ pour voir que $x^\top F^{**}(x) = 0$; on en déduit que $y^\top F^{**}(x) \geqslant 0$ pour tout $y \in K$, ce qui montre que $F^{**}(x) \in K^+$. La réciproque (7.2.5) $\Rightarrow$ (7.2.4) est évidente.

Maintenant, en considérant le problème (7.2.5), on voit que (7.2.3) est un cas particulier de (7.2.5) obtenu lorsque K est l'*orthant positif* $\mathbb{R}^n_+ := \{x \in \mathbb{R}^n : x \geqslant 0\}$.

7.2.2 Transformation en problème d'inéquation variationnelle

Nous avons vu que le problème (7.2.1) peut s'écrire comme un problème d'inéquations variationnelles avec un ensemble conique K au moyen des transformations (7.2.1) $\to$ (7.2.2) $\to$ (7.2.3) $\to$ (7.2.5). Procéder de la sorte conduisait à un problème d'inéquations variationnelles de grande taille. Nous présentons ci-dessous une transformation de ce type mais plus compacte. Concernant cette formulation, on prend $K := \mathbb{R}^n \times \mathbb{R}^p_+$

$$\begin{cases} (x,u) \in K \\ \begin{pmatrix} H(x) \\ G(x)-u \\ F(x) \end{pmatrix}^\top \begin{pmatrix} y-x \\ v-u \end{pmatrix} \geqslant 0, \qquad \forall\, (y,v) \in K, \end{cases} \tag{7.2.6}$$

Nous montrons que les formulations (7.2.1) et (7.2.6) sont équivalentes.

[(7.2.1) $\Rightarrow$ (7.2.6)] : soit x solution de (7.2.1). Posons $u := G(x) \in \mathbb{R}^p_+$, on a alors pour tout $v \in \mathbb{R}^p_+$, $F(x)^\top (v-u) = F(x)^\top v \geqslant 0$. On obtient donc (7.2.6).

[(7.2.6) $\Rightarrow$ (7.2.1)] : soit (x,u) solution de (7.2.6). En prenant $v = u$ et y quelconque, on obtient $H(x) = 0$ et $G(x) = u \geqslant 0$. En prenant $y = x$, on a $F(x)^\top (v-u) \geqslant 0$, pour tout $v \geqslant 0$. Si $v = 2u \geqslant 2$ et $v = \dfrac{u}{2} \geqslant 0$, on obtient $F(x)^\top u = 0$, donc $F(x)^\top G(x) = 0$. Enfin, on a alors aussi $F(x)^\top v \geqslant 0$ pour tout $v \geqslant 0$, donc $F(x) \geqslant 0$.

7.3 Résolution des systèmes d'équations non lisses

De nombreuses études se sont efforcées de mettre au point des algorithmes pour résoudre les problèmes de complémentarité [38]. Elles ont parfois des approches algorithmiques communes, telles que les méthodes d'activation des contraintes, les méthodes de l'analyse non lisse et les algorithmes de points intérieurs. Dans cette thèse, nous nous sommes intéressés à la résolution des problèmes de complémentarité non linéaires par l'algorithme de Newton non lisse appliqué à la fonction min (voir les sections 2.5 et 2.4.2.3), que nous appelons Newton-min non linéaire.

Nous présentons maintenant un algorithme qui permet de résoudre le système d'équations non linéaires $F(x) = 0$ dans lequel $F : \mathbb{R}^n \to \mathbb{R}^n$ est une fonction non lisse et différentiable par morceaux (voir la section 2.5.4.3). On rappelle que pour ces fonctions, le schéma d'approximation newtonien 2.43 est $J : \mathbb{R}^n \twoheadrightarrow \mathbb{R}^m$ définie en $x \in \mathbb{R}^n$ par

$$J(x) = \{F_i'(x) : i \in P(x)\},$$

avec $P(x) = \{i : F(x) = F_i(x)\}$ l'ensemble des indices des fonctions $F_i(x)$ actives en x.

Algorithme 7.1 Newton lisse par morceaux – une itération

On suppose qu'au début de l'itération k, on dispose d'un itéré $x_k \in \mathbb{R}^n$.

1. *Test d'arrêt* : si $F(x_k) = 0$, arrêt de l'algorithme.
2. *Calcul du pas* : pour un indice $i \in P(x_k)$, on calcule une solution d_k du système linéaire suivant

$$F(x_k) + F_i'(x_k) d_k = 0. \tag{7.3.1}$$

3. *Nouvel itéré* : $x_{k+1} := x_k + d_k$.

La consistance et la convergence *locale* de l'algorithme 7.1 sont assurées dans les conditions énoncées dans le théorème suivant [38, théorème 7.2.15].

Théorème 7.2 (convergence locale de Newton lisse par morceaux) *Soient Ω un ouvert de $\mathbb{R}^n$, $F : \Omega \to \mathbb{R}^n$ une fonction différentiable et $x_* \in \Omega$ un point vérifiant $F(x_*) = 0$. On suppose que $F_i'(x_*)$ est inversible pour tout $i \in J(x_*)$. Alors, il existe un voisinage V de x_* tel que si $x_0 \in V$, l'algorithme 7.1 de Newton non lisse est bien défini et génère une suite $\{x_k\}$ convergeant superlinéairement vers x_*. Si toutes les fonctions actives de F au voisinage de x_* sont localement lipschitziennes en x_*, la convergence est alors quadratique.*

Chapitre 8

Simulation numérique

Nous avons construit, dans le chapitre 6, un modèle pour les écoulements liquide-gaz en milieu poreux capable de prendre en compte l'apparition et la disparition de la phase gazeuse. Le modèle obtenu consiste en un système d'équations aux dérivées partielles non linéaires auxquelles sont adjointes des conditions de complémentarité non linéaires. La première partie de ce chapitre présente une technique de discrétisation permettant la simulation numérique de ce modèle. Dans ce contexte, on propose une mise en œuvre numérique du modèle utilisant une discrétisation implicite en temps et une méthode de volumes finis pour la discrétisation en espace. Nous proposons, dans une seconde partie, une méthode de résolution robuste et efficace de type Newton non lisse (voir la section 7.3). Nous présentons, dans une troisième partie, différents cas-tests qui illustrent les capacités du modèle et du code développé à prendre en compte l'apparition/disparition de la phase gazeuse en milieu poreux homogène et en milieu poreux hétérogène. En particulier, les expériences numériques montrent que la méthode de Newton-min se révéle efficace et converge quadratiquement pour ces problèmes.

8.1 Discrétisation

Rappelons le système décrit dans le chapitre 7 et qui est donné par les équations suivantes

$$\begin{aligned}
&\frac{\partial}{\partial t}(\phi s_\ell) + \operatorname{div}(\mathbf{q}_\ell - \frac{1}{\rho_w^\ell} J_h^\ell) = \frac{Q_w}{\rho_w^\ell},\\
&\frac{\partial}{\partial t}(\phi s_\ell C_\ell \chi_h^\ell) + \phi(1-s_\ell) C_g(p_\ell + p_c(s_\ell))) + \operatorname{div}(C_\ell \chi_h^\ell \mathbf{q}_\ell + C_g(p_\ell + p_c(s_\ell))\mathbf{q}_g + J_h^\ell) = Q_h,\\
&(1-s_\ell)\big(M^h H(p_\ell + p_c(s_\ell)) - C_\ell \chi_h^\ell\big) = 0, \quad 1 - s_\ell \geqslant 0, \quad H(p_\ell + p_c(s_\ell)) - C_\ell \chi_h^\ell \geqslant 0.
\end{aligned} \tag{8.1.1}$$

8.1.1 Discrétisation en temps

Pour la discrétisation en temps, on note Δt le pas de temps. nous avons décomposé l'intervalle de simulation en N_t pas de temps de longueur Δt, et on indice les temps $t = 0, \ldots, n\Delta t$ par $n = 0, 1, \ldots, N_t$. Nous avons utilisé un schéma d'Euler implicite en temps.

8.1.2 Discrétisation en espace

En espace, nous avons discrétisé l'intervalle $(0, L)$ en N_x intervalles de longueur h et nous avons utilisé une méthode de volumes finis centrés sur les mailles avec décentrage amont des perméabilités relatives. Les notations des indices sont définies sur la figure (8.1.1).

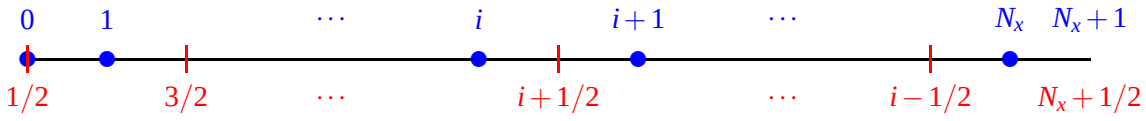


Figure 8.1.1 – Discrétisation en espace

Pour le composant eau, l'équation de conservation discrétisée est

$$\phi\ \rho_w^\ell \frac{s_i^{n+1} - s_i^n}{\Delta t} + \frac{\rho_w^\ell\,(q_{\ell,i+1/2}^{n+1} - q_{\ell,i-1/2}^{n+1}) - (J_{i+1/2}^{n+1} - J_{i-1/2}^{n+1})}{h} = 0, \quad i = 1,\ldots,Nx,$$

avec

$$q_{\ell,1/2}^{n+1} = -Kk_{\ell,1/2}^{*n+1}\,\big(\frac{p_{\ell 1}^{n+1} - p_{\ell 0}^{n+1}}{h/2} - (\rho_w^\ell + C_\ell \chi_0^{n+1})g\big),$$

$$q_{\ell,i+1/2}^{n+1} = -K\,k_{\ell,i+1/2}^{*n+1}\,\big(\frac{p_{\ell,i+1}^{n+1} - p_{\ell,i}^{n+1}}{h} - (\rho_w^\ell + C_\ell \chi_i^{n+1})g\big), \quad i = 2,\ldots,Nx-1,$$

$$q_{\ell,2N_x+1/2}^{n+1} = -K\,k_{\ell,N_x+1/2}^{*n+1}\,\big(\frac{p_{\ell,N_x+1}^{n+1} - p_{\ell,N_x}^{n+1}}{h/2} - (\rho_w^\ell + C_\ell \chi_{i+1}^{n+1})\,g\big).$$

Pour le composant hydrogène, l'équation de conservation discrétisée est

$$\phi\,[\frac{C_\ell\,(s_i^{n+1}\,\chi_i^{n+1} - s_i^n\,\chi_i^n}{\Delta t} + \frac{\rho_{gi}^{n+1}\,(1-s_i^{n+1}) - \rho_{gi}^n\,(1-s_i^n)}{\Delta t}]$$
$$+ C_\ell \frac{q_{\ell,i+1/2}^{n+1}\,\chi_{i+1}^{n+1} - q_{\ell,i-1/2}^{n+1}\,\chi_i^{n+1}}{h}$$
$$+ \frac{\rho_{g,i+1/2}^{n+1}\,q_{g,i+1/2}^{n+1} - \rho_{g,i-1/2}^{n+1}\,q_{g,i-1/2}^{n+1}}{h} + \frac{J_{i+1/2}^{n+1} - J_{i-1/2}^{n+1}}{h} = \frac{Q_i^{n+1}}{h},\ i = 1,\ldots,Nx,$$

avec

$$
\begin{aligned}
&\rho_{gi}^{n+1} = C_g(p_i^{n+1} + p_c(s_i^{n+1})),\\
&\rho_{g,i+1/2}^{n+1} = (\rho_{gi}^{n+1} + \rho_{g,i+1}^{n+1})/2,\\
&q_{g,1/2}^{n+1} = -Kk_{g,1/2}^{*n+1}\,\big(\frac{p_1^{n+1} + p_c(s_1^{n+1}) - p_0^{n+1} - p_c(s_0^{n+1})}{h/2} - \rho_{g,1/2}^{n+1} g\big),\\
&q_{g,i+1/2}^{n+1} = -Kk_{g,i+1/2}^{*n+1}\,\big(\frac{p_{i+1}^{n+1} + p_c(s_{i+1}^{n+1}) - p_i^{n+1} - p_c(s_i^{n+1})}{h} - \rho_{g,i+1/2}^{n+1} g\big),\ i = 2,\ldots,Nx-1,\\
&q_{g,N_x+1/2}^{n+1} = -Kk_{g,N_x+1/2}^{*n+1}\,\big(\frac{p_{N_x+1}^{n+1} + p_c(s_{N_x+1}^{n+1}) - p_{N_x}^{n+1} - p_c(s_{N_x}^{n+1})}{h} - \rho_{g,N_x+1/2}^{n+1} g\big),\\
&J_{1/2}^{n+1} = -M^h\,\phi\,D_h^l s_0^{n+1}(\frac{\rho_w^\ell}{M_w} + \frac{C_\ell}{M_h}\chi_0^{n+1}(\frac{\chi_1^{n+1} - \chi_0^{n+1}}{h/2}),\\
&J_{i+1/2}^{n+1} = -M^h\phi D_h^l s_i^{n+1}(\frac{\rho_w^\ell}{M_w} + \frac{C_\ell}{M_h}\chi_i^{n+1}(\frac{\chi_{i+1}^{n+1} - \chi_i^{n+1}}{h}),\quad i = 2,\ldots,Nx-1,\\
&J_{Nx+1/2}^{n+1} = -M^h\phi D_h^l s_{Nx}^{n+1}(\frac{\rho_w^\ell}{M_w} + \frac{C_\ell}{M_h}\chi_{Nx}^{n+1}(\frac{\chi_{Nx+1}^{n+1} - \chi_{Nx}^{n+1}}{h/2}).
\end{aligned}
$$

Dans la suite, pour simplifier les notations, on pose $N = N_x$ et on note

- $x \in \mathbb{R}^{3N}$, le vecteur des inconnues s_ℓ, p_ℓ, χ_h^ℓ,
- $H : \mathbb{R}^{3N} \to \mathbb{R}^{2N}$, les équations de conservations discrétisées,
- $F : \mathbb{R}^{3N} \to \mathbb{R}^N$, la fonction discrétisée $1 - s_\ell$,
- $G : \mathbb{R}^{3N} \to \mathbb{R}^N$, la fonction discrétisée $H(p_\ell + p_c(s_\ell)) - C_\ell\chi_h^\ell$.

Alors à chaque pas de temps Δt, le problème peut s'écrire sous la forme compacte suivante

$$
\begin{aligned}
&H(x) = 0,\\
&F(x)^\top G(x) = 0,\quad F(x) \geqslant 0,\quad G(x) \geqslant 0.
\end{aligned}
\tag{8.1.2}
$$

Il s'agit d'un système d'équations aux dérivées partielles non linéaires avec des conditions de complèmentarité non linéaires.

8.2 Méthode de résolution

Comme dans le cas linéaire et pour des raisons d'éfficacité numérique, nous avons remplacé les conditions de complémentarité non linéaires par la C-fonction minimum (voir 2.4.2). La deuxième relation du système (8.1.2) est alors équivalente à

$$\varphi(x) = \min((F(x), G(x)) = 0,$$

et le problème (8.1.2) peut s'écrire comme suit

$$\begin{cases} H(x) = 0, \\ \varphi(x) = 0. \end{cases} \tag{8.2.1}$$

C'est un système non différentiable à cause de la fonction min, présente dans la seconde équation, et qui est une fonction linéaire par morceaux. Nous proposons de le résoudre par une méthode de Newton non lisse présentée dans la section suivante.

8.2.1 L'algorithme de Newton-min non linéaire

Nous présontons ici en détail l'algorithme de Newton-min pour résoudre le système non linéaire (8.2.1). On note par $\psi(x)$ la fonction définie comme suit

$$\psi(x) := \begin{pmatrix} H(x) \\ \varphi(x) \end{pmatrix}$$

et par $\varepsilon > 0$ la tolérance utilisée pour l'acceptation d'une solution approchée.

Soit $x^1 \in \mathbb{R}^{3N}$ donné. Pour $k = 2, 3, \ldots$, faire

1) Si $||\psi||_2 \leqslant \varepsilon$, arrêt. La solution du pas temps courant est donné par $x = x^k$.

2) Choisir deux ensembles d'indices complémentaires A^k et I^k par

$$\begin{aligned} A^k &:= \{i : G_i(x_n^k) \leqslant F_i(x^k)\} \subset \{1, \ldots, N\}, \\ I^k &:= \{i : G_i(x^k) > F_i(x^k)\} \subset \{1, \ldots, N\}. \end{aligned} \tag{8.2.2}$$

3) Selectionner un élément J_{x^k} tel que

$$(J_{x^k})_i = \begin{cases} F_i'(x^k) & \text{si } i \in A^k, \\ G_i'(x^k) & \text{si } i \in I^k. \end{cases}$$

4) Soit x^{k+1} solution de

$$\begin{aligned} &H(x^k) + H^{'}(x^k)(x^{k+1} - x^k) = 0, \\ &\varphi(x^k) + J_{x^k}(x^{k+1} - x^k) = 0. \end{aligned}$$

On remarque que, comme dans le cas de la méthode de Newton pour trouver un zéro d'une fonction lisse, un unique système linéaire doit être résolu à chaque itération.

8.2.2 Calcul de la matrice jacobienne

8.2.2.1 Différentiation automatique

Les outils de différentiation automatique peuvent ou bien utiliser la transformation de source, ou bien la surcharge d'opérateur. Concernant la transformation de code source, l'utilisateur fournit des fichiers sources à un logiciel de différentiation, qui fabrique de nouveaux fichiers sources permettant le calcul de la dérivée de la fonction calculée par le programme de départ. Un outil de différentiation automatique par surcharge d'opérateurs permet de différentier un programme en imposant au compilateur d'interpréter les opérations de calcul de manière à réaliser l'opération habituelle et celle du calcul de la dérivée directionnelle. L'utilisateur de cet outil doit modifier légèrement ses sources pour s'y adapter et la dérivation a lieu pendant l'exécution. Étant donné que notre programme est écrit en Matlab, nous avons utilisé le logiciel ADMAT [1] [27]. C'est un outil de différentiation automatique par surcharge d'opérateurs. Les outils par surcharge d'opérateurs sont plus faciles à implémenter.

1. ADMAT : an Automatic Differentiation toolbox for MATLAB

8.2.2.2 Principe

La bibliothèque de différentiation automatique ADMAT fournit de nouvelles fonctionnalités que l'utilisateur doit intégrer à bon escient dans son code. Il faut déterminer la partie du code à différentier, en définir précisément les entrées et les sorties, et demander explicitement le calcul de dérivation suivant le mode choisi. Le programme recalcule alors les valeurs ainsi que les dérivées.

L'outil est basé sur les dérivations des opérations binaires $(+, -, \times, \ldots)$, des fonctions standards telles que $(\exp, \sin, \log, \ldots)$ et les règles de composition. La dérivation est possible en mode direct, dans l'objectif par exemple de calculer une matrice jacobienne colonne par colonne, et en mode inverse, pour un calcul de matrice jacobienne ligne par ligne, ce qui est intéressant lorsqu'il y a moins de lignes que de colonnes.

Pour fonctionner en mode direct, une bibliothèque de différentiation automatique par surcharge d'opérateurs définit une classe (ou plus simplement un type de données) **deriv** qui contient au moins :

- Les attributs ou variables de classe **val** et **deriv**, représentant respectivement une valeur et une dérivée directionnelle, c'est-à-dire pour chaque variable *x* déclarée dans le code de départ, deux variables sont déclarées dans le code d'arrivée, une valeur **x.value** et sa dérivée directionnelle **x.deriv**.
- Une surcharge des opérateurs binaires $(+,\ -,\ \times,\ \ldots)$ et une surcharge des fonctions standards mathématiques $(\exp,\ \sin,\ \log, \ldots)$, c'est-à-dire que ces opérateurs et ces fonctions pourront être utilisées avec la classe **deriv** en argument et que la bibliothèque de différentiation définit ce que doit être le résultat de l'application de ces opérateurs ou de ces fonctions.

Exemple 8.1 *Nous présentons un exemple très simple du calcul de la dérivée d'une fonction définie par un produit de deux arguments en utilisant ADMAT. Soit* $F = u * v$*, on obtient comme résultat retourné*

```
F.value = u.value*v.value.
F.deriv = u.deriv*v.value + u.value*v.deriv.
```

8.3 Expériences Numériques

Nous présentons dans cette section les résultats de simulation que l'on obtient en utilisant la procédure de résolution qu'on vient de présenter (voir section 8.2.1) sur différents cas test. Tous ces cas test sont ainsi définis sur une géométrie bidimensionnelle trés simple et les effets de la gravité sont occultés (écoulement horizontal : l'altitude z est constante). Dans tous les cas test, on considérera un écoulement eau/hydrogène dont les caractéristiques spécifiques aux fluides et aux composants sont données dans le tableau 8.3.1.

Paramètre	Valeur
T	$303 \quad K$
D_ℓ^h	$3\,10^{-9} \quad m^2/s$
μ_ℓ	$1\,10^{-3} \quad Pa.s$
μ_g	$99.10^{-6} \quad Pa.s$
$H(T=303K)$	$7.65 \quad 10^{-6} mol/Pa/m^3$
M_w	$10^{-2} \quad kg/mol$
M_h	$2\,10^{-3} \quad kg/mol$
ρ_ℓ^w	$10^3 \quad kg/m^3$

Table 8.3.1 – Valeurs des caractéristiques des fluides et des composants.

Les fonctions de perméabilité relative et la loi de pression capillaire seront à chaque fois données par le modèle de van Genuchten-Mualem et s'exprimeront donc ainsi :

$$\begin{aligned} p_c &= P_r \left(S_{le}^{-1/m} - 1 \right)^{1/n}, \\ k_{rl} &= \sqrt{S_{le}} \left(1 - \left(1 - S_{le}^{1/m} \right)^m \right)^2, \\ k_{rg} &= \sqrt{1 - S_{le}} \left(1 - S_{le}^{1/m} \right)^{2m}, \\ S_{le} &= \frac{S_l - S_{lr}}{1 - S_{lr} - S_{gr}} \quad \text{avec} \quad m = 1 - \frac{1}{n}. \end{aligned}$$

8.3.1 Cas-test 1 : apparition et disparition de la phase gazeuse

Ce premier cas test vise à illustrer la capacité du modèle et du solveur à prendre en compte l'apparition et la disparition de la phase gazeuse depuis un état totalement saturé en liquide. On considère pour cela un domaine rectangulaire de dimension L_x (problème assimilable à du 1D) (voir figure 8.3.1).

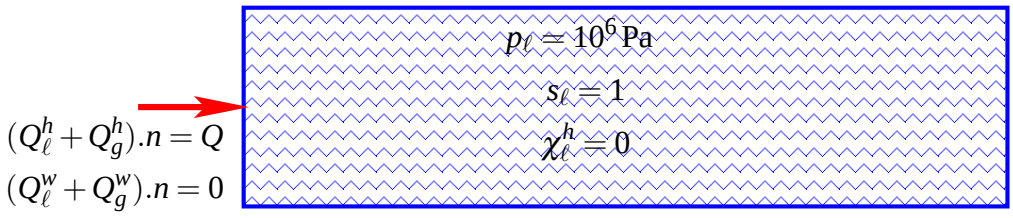


Figure 8.3.1 – Définition de la géométrie du cas test 1.

On suppose que le milieu poreux est homogène et initialement saturé en eau pure auquel on impose une injection d'hydrogène en entrée sur la durée $[0; T_{inj}]$ et un état d'eau pure de pression fixée en sortie. Les caractéristiques du milieu poreux et la géométrie sont données dans le tableau 8.3.2.

Conditions initiales et conditions aux limites

Les conditions initiales sont $s_\ell(t=0) = 1$, $\chi_h^\ell(t=0) = 0$ et $p_\ell(t=0) = 10^6$ Pa. Concernant les conditions aux bords gauche, un flux d'hydrogène donné par $\rho_h^\ell \mathbf{q}_\ell + \rho_h^g \mathbf{q}_g + j_h^\ell = 5.57\,10^{-6}$ kg/m^2 /year. À partir de cette condition, on peut déduire la saturation. Nous imposons également un flux d'eau nul $\rho_w^\ell \mathbf{q}_\ell - j_h^\ell = 0$. À droite, la pression de liquide est fixée, $p_\ell = 10^6$ Pa, et la saturation de liquide est égale à $s_\ell = 1$.

Résultats numériques

La simulation de ce cas-test a été effectuée sur une durée allant de $t = 0$ ans à $t = T_{fin} = 10^6$ ans ; le pas d'espace a été pris constant égal à 1 m. Une description fine des résultats de ce cas test, pendant la période d'injection et pendant la période post-injection, ainsi que la convergence quadratique de Newton-min dans ce cas, sont présentées dans l'article [13] qui est reproduit dans le chapitre 9.

Porous medium parameters		Fluid characteristics parameters	
Paramètre	Valeur	Paramètre	Valeur
K	$5\,10^{-20}$ m^2	T	303 K
ϕ	0.15 (-)	D_ℓ^h	$3\,10^{-9}$ m^2/s
P_r	$2\,10^6$ Pa	μ_ℓ	$1\,10^{-9}$ Pa/s
n	1.49 (-)	μ_g	$9\,10^{-9}$ Pa/s
S_{lr}	0.4 (-)	$H(T = 303\text{K})$	$7.65\,10^{-6}$ mol/Pa/m^3
S_{gr}	0 (-)	M_w	10^{-2} kg/mol
		M_h	$2\,10^{-3}$ kg/mol
		ρ_ℓ^w	10^3 kg/m^3

Table 8.3.2 – Valeurs des caractéristiques des fluides et des composants : cas eau/hydrogène.

Comparaisons des résultats

Ce cas test est proposée dans le cadre du benchmark « Écoulement diphasique » dans le cadre du CNR MoMas [114]. Cette étude a été réalisé par 5 autres équipes utilisant des codes différents :

1- Aster à EDF par S. Granet (EDF).

2- Code IRSN en développement par F. Smaï (IRSN).

3- M++ (Meshes, Multigrid) par T. Müller et E. Marchand de l'université d'Erlangen-Nuremberg (Erlangen).

4- DUNE avec le module de discrétisation dune-pdelab par R. Neumann de l'université d'Heidelberg (Heidelberg).

5- Code en MATLAB par I. Mozolevski (UFSC).

Les schémas numériques utilisés sont de type volumes finis et éléments finis, pour plus d'informations voir les différentes présentations dans [114, 50]. Cependant, nous ne possédons pas d'information sur les pas de temps utilisés par les différentes équipes pour réaliser cette étude. La figure 8.3.2 représente l'évaluation temporelle de la pression de liquid, de la pression de gaz et de la saturation de gas à l'entrée obtenues par les six équipes.

Ces résultats ont été présentés lors du worshop GNR MoMas [55].

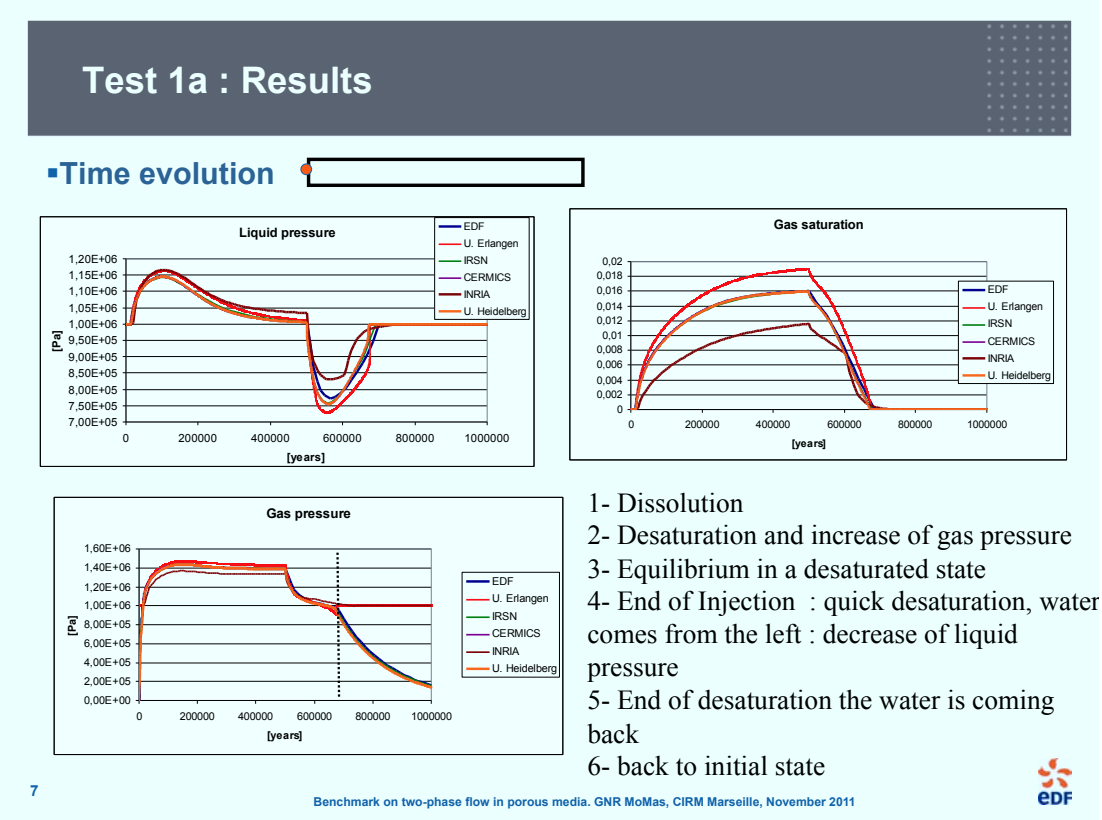


Figure 8.3.2 – Benchmark - comparaison de résultats : Pression du liquide et de gaz et saturation de gaz en fonction de temps.

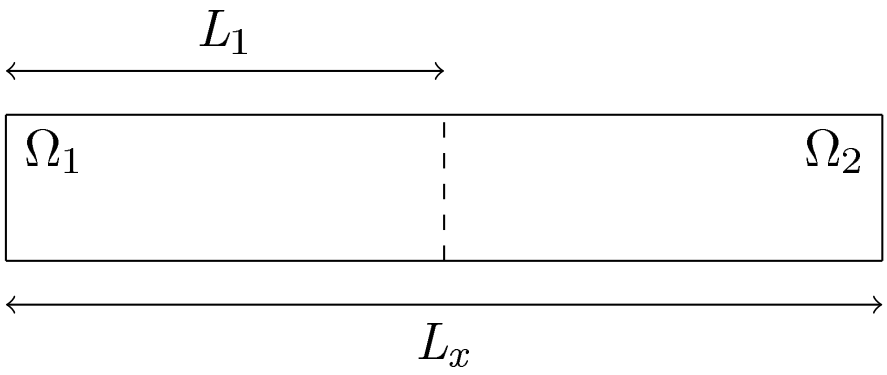


Figure 8.3.3 – Définition de la géométrie du cas test 2.

Notre formulation a permis de traiter correctement l'apparition et la disparition de phase pour un écoulement en milieu poreux et de retrouver des résultats similaires à ceux disponibles dans la littérature [121] et à ceux obtenus par les autres équipes. Les différences observeées pour les résultats sur les saturations en gaz à l'entrée s'explique par les pas de temps choisis par les différents équipes.

8.3.2 Cas-test 2 : milieu hétérogène

Ce second cas test est une variation du premier. On y reprend le concept d'une injection d'hydrogène, constante dans le temps ici, dans un milieu poreux initialement saturé en eau pure, à la différence que le milieu poreux n'est plus homogène mais désormais constitué de deux matériaux homogènes adjacents Ω_1 et Ω_2(voir figure 8.3.3) avec $L_x = 200m$ et $L_1 = 100\,m$. Les données de ce cas test sont indiquées dans le tableau 8.3.3.

Conditions initiales et conditions aux limites

Les conditions initiales sont $s_\ell\,(t=0)=1$, $\chi_h^\ell\,(t=0)=0$ and $p_\ell\,(t=0)=10^6$ Pa. Les conditions aux bords gauche, un flux d'hydrogène donné par $\rho_h^\ell \mathbf{q}_\ell + \rho_h^g \mathbf{q}_g + j_h^\ell = 5.57\,10^{-6}$ kg/m^2/year. À partir de cette condition, on peut déduire la saturation. On impose également un flux d'eau nul $\rho_w^\ell \mathbf{q}_\ell - j_h^\ell = 0$. À droite, la pression de liquide est fixée, $p_\ell = 10^6$ Pa, et la saturation de liquide est égale à $s_\ell = 1$. Les conditions à l'interface sont

Milieu poreux		Milieu poreux	
Paramètre	Valeur sur Ω_1	Paramètre	Valeur sur Ω_2
K	10^{-18} m^2	K	$5\,10^{-20}$ m^2
ϕ	0.3 (-)	ϕ	0.15 (-)
P_r	$2\,10^6$ Pa	P_r	$15\,10^6$ Pa
n	1.54 (-)	n	1.49 (-)
S_{lr}	0.01 (-)	S_{lr}	0.4 (-)
S_{gr}	0 (-)	S_{gr}	0 (-)

Table 8.3.3 – Valeurs des caractéristiques des fluides et des composants : cas eau/hydrogène.

- la continuité de la pression de liquid : $p_\ell^{(1)} = p_\ell^{(2)}$ avec $p_\ell^{(1)}$ et $p_\ell^{(1)}$ les pressions de liquide de part et d'autre de l'interface,
- la continuité du flux d'eau : $\mathbf{q}_w^{(1)} = \mathbf{q}_w^{(2)}$ avec $\mathbf{q}_w^{(1)}$ et $\mathbf{q}_w^{(2)}$ les flux d'eau de part et d'autre de l'interface,
- la continuité du flux d'hydrogène : $\mathbf{q}_h^{(1)} = \mathbf{q}_h^{(2)}$ avec $\mathbf{q}_h^{(1)}$ et $\mathbf{q}_h^{(2)}$ les flux d'hydrogène de part et d'autre de l'interface,
- la continuité de la pression de liquide : $p_c^{(1)}(s_\ell^{(1)}) = p_c^{(2)}(s_\ell^{(2)})$ avec $p_c^{(1)}$ et $p_c^{(2)}$ la loi de pression capillaires de chaque roche et $s_\ell^{(1)}$ et $s_\ell^{(2)}$ les saturations de liquide de part et d'autres de l'interface.

Résultats numériques

La simulation de ce cas test a été effectuée sur une durée allant de $t = 0$ ans à $t = T_{fin} = 10^6$ ans ; le pas d'espace a été pris constant égal à 1 m le pas de temps $\Delta t =$ 2000 ans. Nous avons représenté, en différents temps, les profils de la pression du liquide p_ℓ, de la concentration d'hydrogène dissout, et de la saturation en gaz s_g. L'évolution de ces différentes grandeurs est qualitativement analogue à celle observée durant la période d'injection du cas test 1(voir 9.4). À la différence remarquable des profils de saturation en gaz qui présentent une discontinuité à l'interface entre les deux milieux poreux.

Dans ce cas, on distingue quatre phases au cours de la simulation (voir Figures 8.3.4 8.3.5 et 8.3.6) :

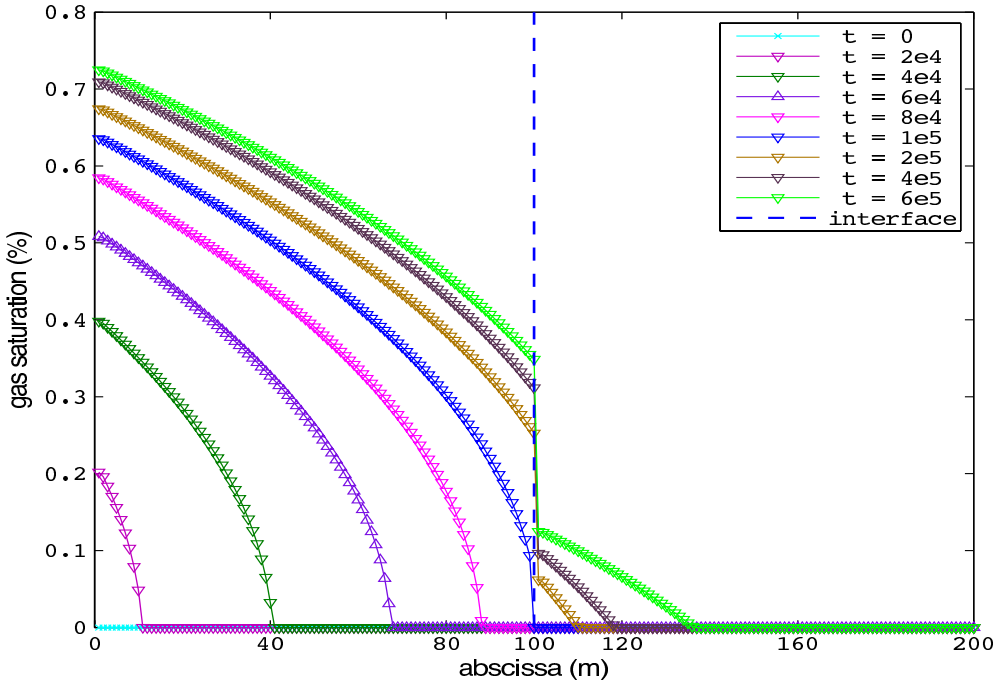


Figure 8.3.4 – Profile de la saturation de gaz.

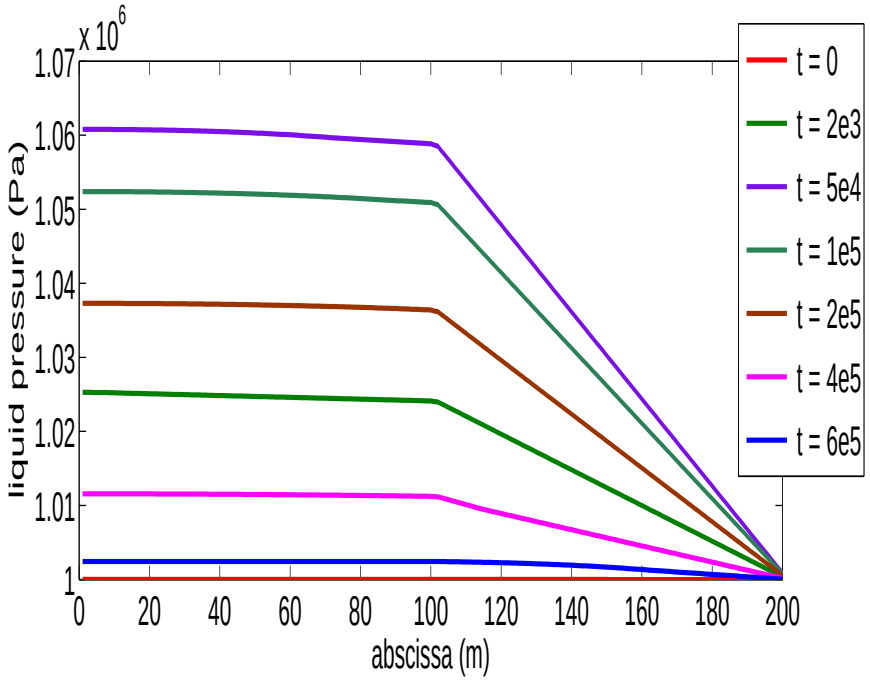


Figure 8.3.5 – Profile de la pression de liquide.

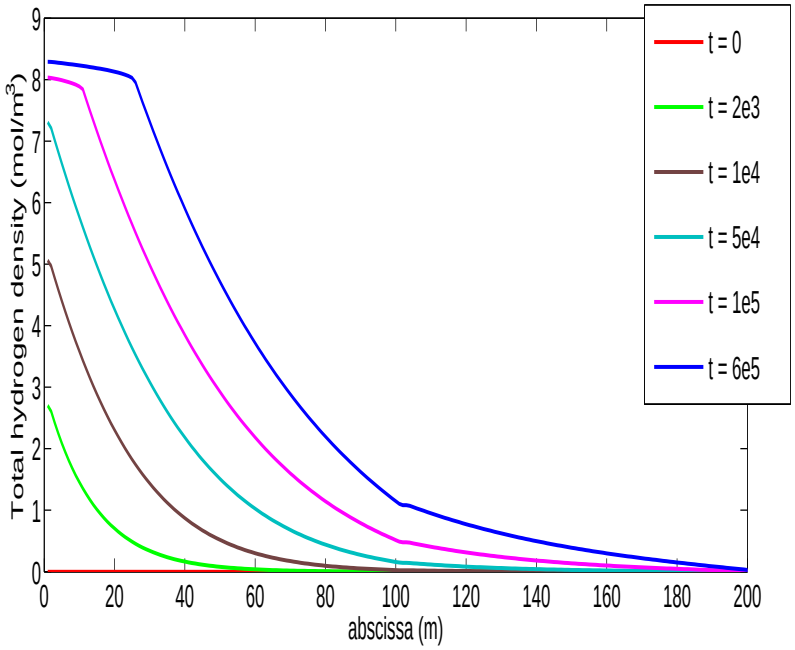


Figure 8.3.6 – Profile de la pression de liquide.

- Pour $0 < t < 4.10^4$ ans : la saturation en gaz et la pression du liquide restent constantes sur tout le domaine tandis que la concentration d'hydrogène augmente.
- Pour $410^4 < t < 6.10^4$ ans : la pression du liquide et la concentration d'hydrogne augmentent sur l'ensemble du domaine. La phase gazeuse apparaît et la saturation en gaz augmente aussi sur une partie croissante du domaine.
- Pour $6.10^4 < t < 2.10^5$ ans, l'évolution se poursuit dans le même sens. Le front de saturation dépasse l'interface entre les deux milieux poreux. La saturation en gaz est non nulle de part et d'autre de l'interface et on observe une discontinuité de la saturation au niveau de cette interface. Ceci s'explique par la différence des lois de pression capillaire de chaque roche.
- Pour $2.10^5 < t < 10^6$ ans : la concentration d'hydrogène et la saturation en gaz continuent de croître tandis que la pression liquide diminue. De la même manière que dans le cas test 1, le système poursuit son évolution vers un état stationnaire.

Convergence quadratique de Newton-min

La figure 8.3.7 montre la convergence quadratique de l'algorithme de Newton-min dans ce cas aussi. Comme mentionné dans le thèoreme 2.37, on peut avoir une convergence quadratique locale, au moins pour les pas de temps qui sont suffisamment petits. En effet, on peut vérifier dans cette figure, qu'à chaque pas de temps, le résidu est réduit de 10^{-5} à 10^{-10} en une itération, ce qui est le signe d'une convergence quadratique.

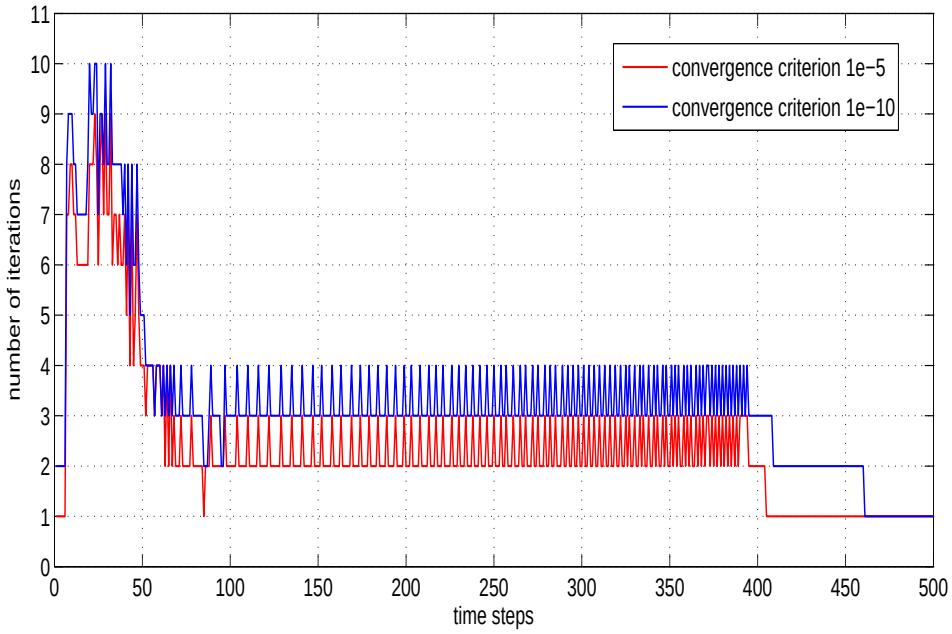


Figure 8.3.7 – Convergence quadratique de Newton-min.

8.4 Conclusion

Nous avons montré dans ce chapitre que le modèle proposé pour les écoulements liquide-gaz avec échange entre les phases en milieu poreux ainsi que la méthode de résolution sont capables de prendre en compte l'apparition/disparition de la phase gazeuse. Nous montrons comment appliquer une stratégie robuste (pas d'échec constanté) et efficace (rapidité de la convergence), la méthode de Newton-min, pour résoudre ces problèmes de géoscience. En particulier, les expériences numériques montrent que la méthode de Newton-min converge quadratiquement pour ces problèmes.

Chapitre 9

Gas phase appearance and disappearance as a problem with complementarity constraints

Dans ce chapitre, nous présentons un article [13], en révision pour publication dans *Mathematics and Computers in Simulation*. Cet article contient la description de la méthode et des résultats numériques obtenus pour le cas test 8.3.1 inspiré du benchmark couplex-Gaz.

9.1 Introduction

The couplex-Gas benchmark [50] was proposed by Andra (French National Inventory of Radioactive Materials and Waste) [5] and the research group MoMaS (Mathematical Modeling and Numerical Simulation for Nuclear Waste Management Problems) [114] in order to improve the simulation of the migration of hydrogen produced by the corrosion of nuclear waste packages in an underground storage. This is a system of two-phase (liquid-gas) flow with two components (hydrogen-water). The benchmark generated some interest and engineers encountered difficulties in handling the appearance and disappearance of the phases. The resulting formulation [65] is a set of partial differential equations with nonlinear complementarity constraints. Even though they appear in several problems of flow and transport in porous media like the black oil model presented in [22]

or transport problems with dissolution-precipitation [81, 20, 93], complementarity problems are not usually identified as such in hydrogeology and, to circumvent the solution of complementarity conditions, problems are often solved by reformulating the problem as in [18, 2, 4]. However the solution of complementarity problems is an active field in optimization [10, 38, 60] and we draw from the know-how of this scientific community. A similar path is followed bin papers like [94, 57, 84]. The application of a nonsmooth Newton method [61, 71], sometimes called the Newton-min algorithm, to solve nonlinear complementarity problem is described. We will demonstrate through a test case, the ability of our model and our solver to efficiently cope with appearance or/and disappearance of one phase.

In the section 9.2, we introduce the formulation of the problem and in the section 9.3 we describe the numerical method. In the section 9.4, we present and discuss a numerical experiment.

9.2 Problem formulation

This section gives a precise formulation of the mathematical model for the application that was outlined in the introduction. We consider a problem where the gas phase can disappear while the liquid phase is always present.

9.2.1 Fluid phases

Let ℓ and g be the respective indices for the liquid phase and the gas phase. Darcy's law reads

$$\mathbf{q}_i = -K(x)k_i(s_i)(\nabla p_i - \rho_i g \nabla z), \quad i = \ell, g, \tag{9.2.1}$$

where K is the absolute permeability. For each phase $i = \ell, g$, s_i is the saturation and $k_i = \dfrac{k_{ri}(s_i)}{\mu_i}$ is the mobility with k_{ri} the relative permeability and μ_i the viscosity (assumed to be constant). The mobility k_i is an increasing function of s_i such that $k_i(0) = 0$, $i = \ell, g$. Assuming that the phases occupy the whole pore space, the phase saturations satisfy

$$0 \leqslant s_i \leqslant 1, \quad s_\ell + s_g = 1.$$

The phase pressures are related through the capillary pressure law

$$p_c(s_\ell) = p_g - p_\ell \geqslant 0,$$

assuming that the gas phase is the non-wetting phase. The capillary pressure is a decreasing function of the saturation s_ℓ.

In the following, we will choose s_ℓ and p_ℓ as the main variables since we assume that the liquid phase cannot disappear for the problem under consideration.

9.2.2 Fluid components

We consider two components, water and hydrogen, identified by the indices $j = w, h$. The mass density of the phase is

$$\rho_i = \rho_w^i + \rho_h^i, \quad i = \ell, g.$$

From M^w and M^h, the water and hydrogen molar masses, we define the molar concentration of phase i :

$$c_i = c_w^i + c_h^i, \quad c_j^i = \frac{s_i \rho_j^i}{M^j}, \quad j = w, h, \quad i = \ell, g. \tag{9.2.2}$$

The molar fractions are

$$\chi_h^i = \frac{c_h^i}{c_i}, \quad \chi_w^i = \frac{c_w^i}{c_i}, \quad i = \ell, g. \tag{9.2.3}$$

Obviously,

$$\chi_w^i + \chi_h^i = 1, \quad i = \ell, g. \tag{9.2.4}$$

We assume that the liquid phase may contain both components, while the gas phase contains only hydrogen, that is the water does not vaporize. In this situation we have

$$\rho_w^g = 0, \quad \rho_g = \rho_h^g, \quad \chi_h^g = \frac{c_h^g}{c_g} = 1, \quad \chi_w^g = 0.$$

For the liquid phase, we assume that the water is the solvent and the hydrogen is the solute and that the quantity of hydrogen dissolved in the liquid is small, that is $c_h^l \ll c_w^l$. So we have

$$\chi_h^\ell \approx \frac{c_h^\ell}{c_w^\ell} = \frac{M^w}{M^h \rho_w^\ell} \rho_h^\ell.$$

A third main unknown will be χ_h^ℓ, in addition to s_ℓ and p_ℓ.

9.2.3 Conservation of mass

We introduce the molecular diffusion flux for the diffusion of hydrogen in the liquid phase

$$j_h^\ell = -\phi M^h s_\ell c_\ell D_h^\ell \nabla \chi_h^\ell \tag{9.2.5}$$

where D_h^ℓ is a molecular diffusion coefficient.

Conservation of mass applied to each component, water and hydrogen, gives

$$\begin{aligned} &\frac{\partial}{\partial t}(\phi \rho_w^\ell s_\ell) + \mathrm{div}(\rho_w^\ell \mathbf{q}_\ell - j_h^\ell) = Q_w, \\ &\frac{\partial}{\partial t}(\phi s_\ell \rho_h^\ell + \phi s_g \rho_h^g) + \mathrm{div}(\rho_h^\ell \mathbf{q}_\ell + \rho_h^g \mathbf{q}_g + j_h^\ell) = Q_h. \end{aligned} \tag{9.2.6}$$

We assume also that the gas is slightly compressible, that is $\rho_g = C_g p_g$ with C_g the compressibility constant, and that the liquid phase is incompressible, that is ρ_w^ℓ is constant.

9.2.4 Nonlinear complementarity constraints

Next, we apply Henry's law which says that, at a constant temperature, the amount of a given gas that dissolves in a given type and volume of liquid is directly proportional to the partial pressure of that gas in equilibrium with that liquid.

In the presence of the gas phase, Henry's law reads $H p_g = \rho_h^\ell$, where $H = \mathrm{H}(T) M^h$ with H(T) is the Henry law constant, depending only on the temperature.

There are two possible cases : the gas phase exists : $1 - s_\ell > 0$, Henry's law applies and $H(p_\ell + p_c(s_\ell)) - \rho_h^\ell = 0$, or the gas phase does not exist, $s_\ell = 1$ and $H(p_\ell + p_c(1)) - \rho_h^\ell \geqslant 0$ which says that for a given pressure p_ℓ the concentration ρ_h^ℓ is too small for the hydrogen component to be partly gaseous, or conversely for a given concentration ρ_h^ℓ the pressure p_ℓ is too large for the hydrogen component to be partly gaseous.

These cases can be written as complementary constraints

$$(1 - s_\ell)\big(H(p_\ell + p_c(s_\ell)) - \rho_h^\ell\big) = 0, \;\; 1 - s_\ell \geqslant 0, \;\; H(p_\ell + p_c(s_\ell)) - \rho_h^\ell \geqslant 0. \tag{9.2.7}$$

Finally we end up with a system of nonlinear partial differential equations (conservation equations (9.2.6) and Darcy laws (9.2.1)) with the nonlinear complementarity constraints (9.2.7) describing the transfer of hydrogen between the two phases, the unknowns being s_ℓ, p_ℓ, and χ_h^ℓ. This formulation has the advantage of being valid whether the gas phase exists or not [65].

9.3 Discretization and solution method

We use a first order Euler implicit scheme for time discretization and cell-centered finite volumes for space discretization. We denote by N, the number of degrees of freedom for s_ℓ, p_ℓ and χ_h^ℓ which is equal to the number of cells. We introduce

- $x \in \mathbb{R}^{3N}$, the vector of unknowns for s_ℓ, p_ℓ, χ_h^ℓ,
- $\mathscr{H} : \mathbb{R}^{3N} \to \mathbb{R}^{2N}$, the discretized conservation equations,
- $\mathscr{F} : \mathbb{R}^{3N} \to \mathbb{R}^{N}$, the discretized function $1 - s_\ell$,
- $\mathscr{G} : \mathbb{R}^{3N} \to \mathbb{R}^{N}$, the discretized function $H(p_\ell + p_c(s_\ell)) - \dfrac{M^h \rho_w^\ell}{M^w} \chi_h^\ell$.

Then at each time step the problem can be written in compact form

$$\begin{aligned} &\mathscr{H}(x) = 0, \\ &\mathscr{F}(x)^\top \mathscr{G}(x) = 0, \quad \mathscr{F}(x) \geqslant 0, \quad \mathscr{G}(x) \geqslant 0, \end{aligned} \tag{9.3.1}$$

where the inequalities have to be understood component-wise.

9.3.1 A non-smooth system using the Minimum function

It is well known that complementarity conditions, consisting of equations and inequalities, can be expressed equivalently by an equation via a complementarity function [38](C-function). Let

$$\begin{array}{rccc} \varphi & : & \mathbb{R}^N \times \mathbb{R}^N & \to & \mathbb{R}^N \\ & & (a,b) & \mapsto & \min(a,b) \end{array}$$

be the minimum function, in which the min operator acts component-wise. This is a C-function, in the sense that it satisfies

$$\varphi(a,b) = 0 \qquad \Longleftrightarrow \qquad a \geqslant 0, \quad b \geqslant 0, \quad a^\top b = 0. \tag{9.3.2}$$

Other typical scalar C-functions [38] are

- the Fisher-Burmeister function : $\varphi(a,b) = \sqrt{a^2+b^2} - a - b$,
- $\varphi(a,b) = -ab + \overset{2}{\min}(0,a) + \overset{2}{\min}(0,b)$.

Using this minimum function, we can write the complementarity problem (9.3.1) as

$$\begin{aligned} &\mathscr{H}(x) = 0, \\ &\varphi(\mathscr{F}(x), \mathscr{G}(x)) = 0. \end{aligned} \tag{9.3.3}$$

Hence, the resulting system of mass conservation (differential) equations and equilibrium conditions is fully free of inequalities (pure set of equations). The only drawback of the introduction of a complementarity problem is that the problem is no longer C^1, since $\varphi \notin C^1(\mathbb{R}^{2N}, \mathbb{R}^N)$, while the typical assumption for having the local quadratic convergence of Newton's algorithm requires to have a "C^1 function with a Lipchitz-continuous derivative". However, it is well known, especially in the community of optimization, that the assumptions can be weakened in several ways, for example by only assuming strong semi-smoothness. In the next section we give the definition of semi-smoothness from [26, 38].

9.3.2 Semi-smoothness

Let $\psi : \mathbb{R}^N \to \mathbb{R}^N$ be a locally lipschitz-continuous function. Then, by Rademacher's theorem[38], there is a dense subset $D \subset \mathbb{R}^N$ on which f is differentiable. The B-subdifferential of ψ at a point $x \in \mathbb{R}^N$ is the set

$$\partial_B \psi(x) := \{J \in \mathbb{R}^{N\times N} \quad | \quad J = \lim_{k\to\infty} \psi'(x_k),\ (x_k) \subset D,\ x_k \to x\},$$

where ψ' is the derivative of ψ. The *generalized Jacobian* of ψ at x [26] is the set

$$\partial \psi(x) = \operatorname{co} \partial_B \psi(x),$$

where $\operatorname{co} S$ denotes the convex hull of a set S. Now, the function ψ is said to be semi-smooth at x if ψ is directionally differentiable at x and

$$Jd - \psi'(x;d) = o(||d||),$$

for any $d \to 0$ and for any $J \in \partial \psi(x+d)$, where $\psi'(x;d)$ denotes the directional derivative of ψ at x in the direction of d. Analogously, ψ is called *strongly semi-smooth* at x, if

$$Jd - \psi'(x;d) = o(||d||^2).$$

ψ is called (strongly) semi-smooth if ψ is (strongly) semi-smooth at any point $x \in \mathbb{R}^N$.

It is well known that the minimum function and the Fisher-Burmeister function are strongly semi-smooth. One can then solve system (9.3.3) using the nonsmooth Newton's method, called the Newton-min method [10, 11] when the min function is used. The Newton-min method can also be regarded as an active set strategy [61].

9.3.3 The Newton-min algorithm

We now give an exact statement of the Newton-min algorithm for solving the nonlinear system of equation (9.3.3).

Below $\partial\varphi(x)$ denotes the generalized Jacobian of φ at a point x. Let Res be the residual of $\psi(x)$ where $\psi(x) := \begin{pmatrix} \mathscr{H}(x) \\ \varphi(x) \end{pmatrix}$ and ε be a stopping criterion for Res.

Let $x^1 \in \mathbb{R}^N$. For $k = 2, 3, \ldots$, do the following.

1) If Res $\leqslant \varepsilon$, stop.
2) Define the complementary index sets A^k and I^k by

$$A^k := \{i : \mathscr{G}_i(x^k) < \mathscr{F}_i(x^k)\}, \quad I^k := \{i : \mathscr{G}_i(x^k) \geqslant \mathscr{F}_i(x^k)\}.$$

3) Select an element such that its ith line is equal to $\mathscr{F}_i'(x^k)$ [resp. $\mathscr{G}_i'(x^k)$] if $\mathscr{F}_i(x^k) \leqslant \mathscr{G}_i(x^k)$ [resp. $\mathscr{F}_i(x^k) > \mathscr{G}_i(x^k)$].
4) Let x^{k+1} be a solution to

$$\begin{aligned} &\mathscr{H}(x^k) + \mathscr{H}'(x^k)(x^{k+1} - x^k) = 0, \\ &\varphi(x^k) + \mathscr{J}_x^k(x^{k+1} - x^k) = 0, \qquad \mathscr{J}_x^k \in \partial\varphi(x^k). \end{aligned}$$

Note that, as in a smooth Newton method, only one linear system has to be solved at each Newton iteration.

Furthermore the Newton-min method satisfies also a quadratic convergence property. Indeed, a theorem[38] says that if x^* is a solution to the system $\psi(x) = 0$, such that J is

nonsingular for all J determined by the choice of A and I at the point x^*, then for any initial value sufficiently close to x^*, the Newton-min method generates a sequence that converges superlinearly to x^*. If F' and G' are locally Lipschitz at x^*, the the convergence rate is quadratic.

We have not yet proved the hypothesis of non-singularity of J for our system but we observed the quadratic convergence in our numerical experiments.

9.4 Numerical experiment

9.4.1 A problem inspired from the Couplex Gas benchmark

We consider a one-dimensional core with length L = 200 m, initially saturated with liquid ($s_\ell = 1$) and containing no hydrogen ($\chi_h^\ell = 0$). Hydrogen is injected at a given rate on the left. After a while the hydrogen injection is stopped. The problem is then to simulate the migration of hydrogen and to illustrate the gas appearance and disappearance phenomena.

We calculate spatial evolutions of the liquid pressure, the total hydrogen molar density and the the gas saturation along the line. Computations are performed from the initial time up to the stationary state.

The core is supposed to be homogenous porous medium. The capillary pressure function p_c and the relative permeability functions, k_{rl} and k_{rg}, are given by the Van Genuchten-Mualem model [51] :

$$p_c = P_r \left(S_{le}^{-1/m} - 1 \right)^{1/n},$$
$$k_{rl} = \sqrt{S_{le}} \left(1 - \left(1 - S_{le}^{1/m} \right)^m \right)^2, \quad k_{rg} = \sqrt{1 - S_{le}} \left(1 - S_{le}^{1/m} \right)^{2m},$$

with $S_{le} = \dfrac{S_l - S_{lr}}{1 - S_{lr} - S_{gr}}$ and $m = 1 - \dfrac{1}{n}$, and where parameters P_r, n, S_{lr} and S_{gr} depend on the porous medium. The parameters describing the porous medium and the fluid characteristics are given in Table 9.4.1. Fluid temperature is fixed to $T = 303$ K.

Initial conditions are $S_\ell(t=0) = 1$, $\chi_h^\ell(t=0) = 0$ and $p_\ell(t=0) = 10^6$ Pa. For boundary conditions on the left, the hydrogen flow rate is given, $\rho_h^\ell \mathbf{q}_\ell + \rho_h^g \mathbf{q}_g + j_h^\ell = 5.57\,10^{-6}$ kg/m^2/year. From this condition, one can deduce the saturation. Still on the left, we impose

Porous medium parameters		Fluid characteristics parameters	
Parameter	Value	Parameter	Value
K	$5\,10^{-20}$ m^2	T	303 K
ϕ	0.15 (-)	D_ℓ^h	$3\,10^{-9}$ m^2/s
P_r	$2\,10^6$ Pa	μ_ℓ	$1\,10^{-9}$ Pa.s
n	1.49 (-)	μ_g	$9\,10^{-9}$ Pa.s
S_{lr}	0.4 (-)	$H(T = 303\text{K})$	$7.65\,10^{-6}$ mol/Pa/m^3
S_{gr}	0 (-)	M_w	10^{-2} kg/mol
		M_h	$2\,10^{-3}$ kg/mol
		ρ_w^ℓ	10^3 kg/m^3

Table 9.4.1 – Values of porous medium fluid characteristics.

a zero water flow rate $\rho_w^\ell \mathbf{q}_\ell - j_h^\ell = 0$. On the right, the liquid pressure is given, $p_\ell = 10^6$ Pa, and the liquid saturation is set to $s_\ell = 1$.

9.4.2 Results and comments

For the numerical simulation below we divided the space interval into 200 intervals of equal length and we used a constant time step of 5000 years. During the simulation, we can identify four important periods, three periods during injection and one period after injection.

During injection (figures 9.4.1, 9.4.2 and 9.4.3) : $\mathbf{0 < t < 5.10^5}$ years

- Period 1 ($\mathbf{0 < t < 2\,10^4}$ years) : only the hydrogen density increases (Figure 9.4.1, green curves), while the liquid pressure and the gas saturation stay constant (Figures 9.4.2 and 9.4.3, green curve) ; the whole domain is saturated with water $(s_g = 0)$.
- Period 2 ($\mathbf{2\,10^4 \leqslant t \leqslant 1.5\,10^5}$ years) : at $t = 2\,10^4$, the gas phase appears ($s_g > 0$). During this period, the liquid pressure increases (Figures 9.4.3, blue curves) and pressure gradients are non zero which corresponds to a displacement of both phases. The total hydrogen density and the gas saturation increase (Figures 9.4.1 and 9.4.2, blue curves) and the unsatured area grows.

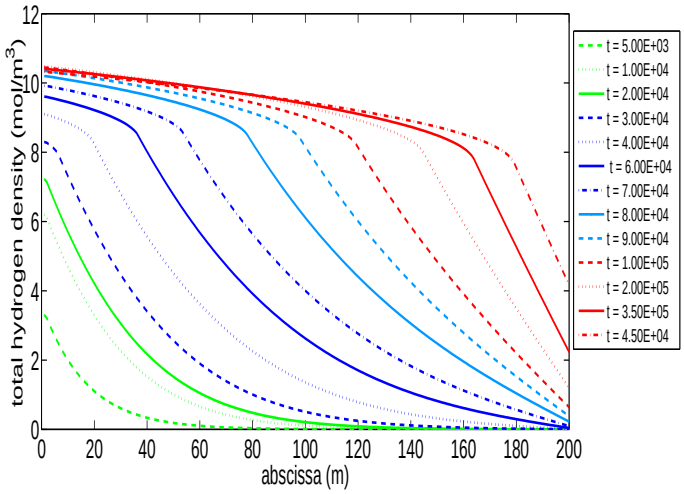


Figure 9.4.1 – Spatial evolution of hydrogen density at several times t (in years) during hydrogen injection.

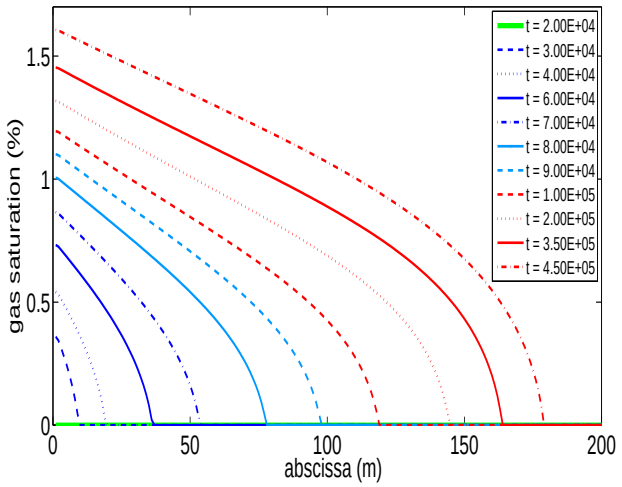


Figure 9.4.2 – Spatial evolution of gas saturation at several times t (in years) during hydrogen injection.

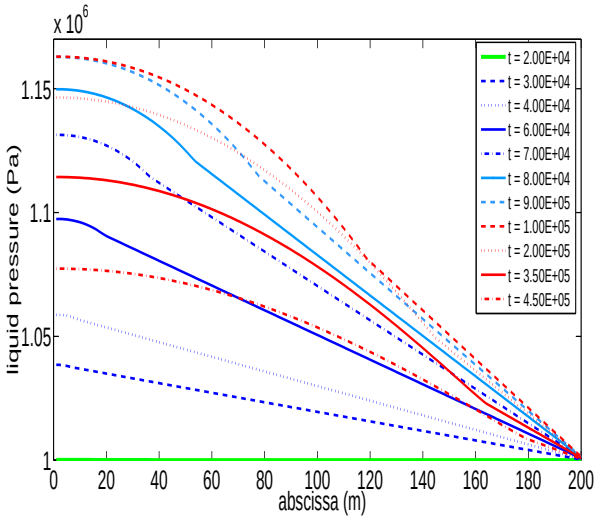


Figure 9.4.3 – Spatial evolution of liquid pressure at several times t (in years) during hydrogen injection.

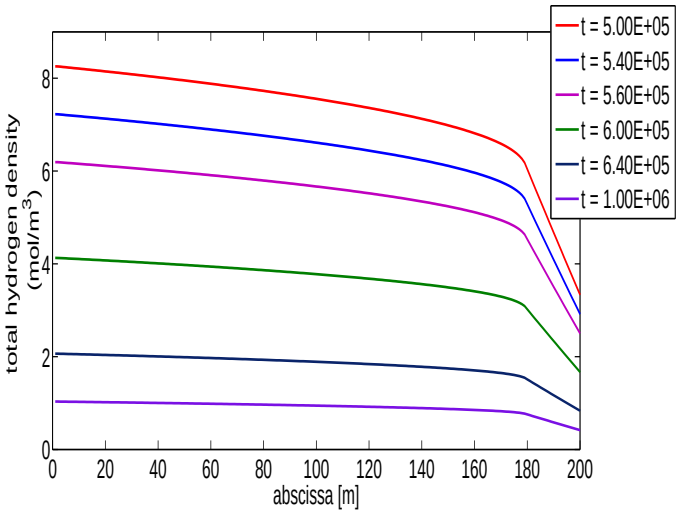


Figure 9.4.4 – Spatial evolution of hydrogen density at several times t (in years) after hydrogen injection is stopped.

- Period 3 ($\mathbf{1.5\ 10^5 < t < 5\ 10^5}$ years) : while the total hydrogen density and the gas saturation continue to increase (Figures 9.4.1 and 9.4.2, red curves) ; the liquid pressure and the pressure gradient decrease since there is no water injection (Figure 9.4.3, red curves).

After injection (Figures 9.4.4, 9.4.5 and 9.4.6) :

- Period 4 ($\mathbf{t > 5\ 10^5}$ years) : cell by cell, starting from the right, the gas saturation decreases and after a while, the gas phase disappears (Figure 9.4.5). At the end of the simulation the system reaches a stationary state (Figure 9.4.4) and the liquid pressure gradient goes to zero (Figure 9.4.6).

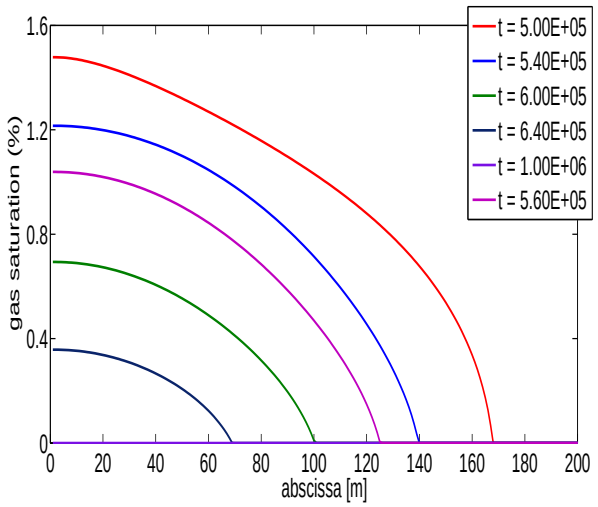


Figure 9.4.5 – Spatial evolution of gas saturation at several times t (in years) after hydrogen injection is stopped.

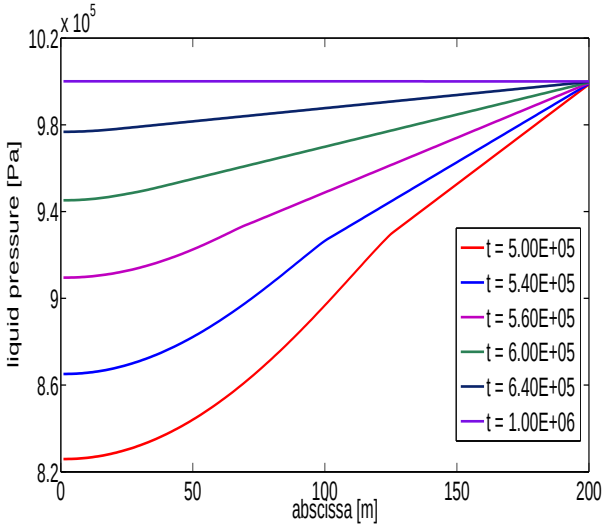


Figure 9.4.6 – Spatial evolution of liquid pressure at several times t (in years) after hydrogen injection is stopped.

9.4.3 Quadratic convergence

The figure 9.4.7 shows the number of Newton-min iterations per time step for two convergence criterions, $\varepsilon_1 = $ 1.e-5 (red curve) and $\varepsilon_2 = $ 1.e-10 (blue curve). The points are connected with a straight line. As mentioned at the end of section 9.3.3, one can expect local quadratic convergence, at least for time steps which are sufficiently small. In Figure 9.4.7, we can observe this quadratic convergence. Indeed one can verify in this figure that, at each time step, the residue goes from 1.e-5 to 1.e-10 in one iteration.

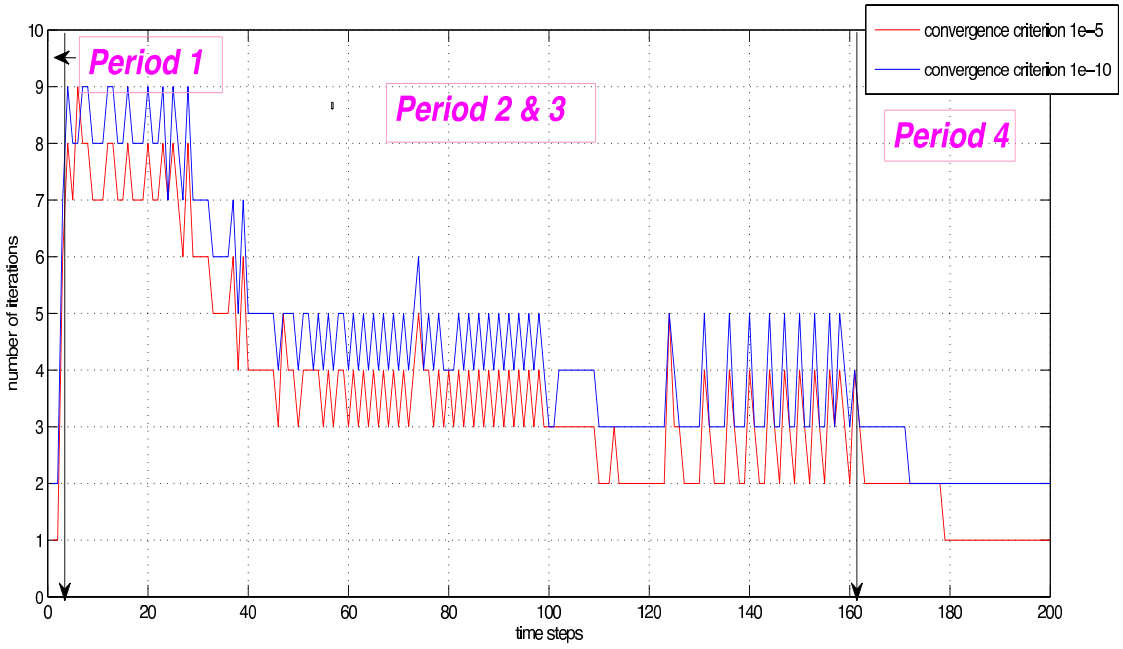


Figure 9.4.7 – Quadratic convergence of Newton-min : number of Newton-min iterations per time step for two convergence criterions, 1.e-5 (red curve) and 1.e-10 (blue curve).

9.5 Conclusion

We have studied a solution procedure for a model describing a system of two-phase (liquid-gas) flow in porous media with two components (hydrogen-water) where hydrogen can dissolve in the liquid phase. The problem is formulated as a nonlinear complementarity problem and is solved with the Newton-min method. We considered an example of a Couplex-Gas benchmark and we showed the ability of our solver to describe the appearance and disappearance of the gas phase during the migration of hydrogen. We also discussed the quadratic convergence of the Newton-min method. A theoretical justification for this quadratic convergence and other benchmark examples are under investigation.

Bibliographie

[1] A. ABADPOUR AND M. PANFILOV, *Method of negative saturations for multiphase compositional flow with oversaturated zones*, OF EUROMECH 499 P. éditon, (2010), pp. 101–119. 18

[2] ——, *symptotic decomposed model of two-phase compositional flow in porous media : Analytical front tracking method for riemann problem*, Transport in Porous Media, 82 (2010), pp. 547–565. 170

[3] M. AGANAGIĆ, *Newton's method for linear complementarity problems*, Mathematical Programming, 28 (1984), pp. 349–362. 56, 64, 86, 87, 88

[4] B. AMAZIANE, S. ANTONTSEV, L. PANKRATOV, AND A. PIATNITSKI, *Homogenization of immiscible compressible two-phase flow in porous media : Application to gas migration in a nuclear waste repository*, Multiscale Modeling and Simulation, 8 (2010), pp. 2023–2047. 170

[5] ANDRA, *French National Inventory of Radioactive Materials and Waste*. http://www.andra.fr/international/. 127, 169

[6] C. BAIOCCHI AND A. CAPELO, *Variational and Quasivariational Inequalities : Applications to Free Boundary Problems*, John Wiley, 1984. 16

[7] S. BELLAVIA, M. MACCONI, AND B. MORINI, *An affine scaling trust-region approach to bound-constrained nonlinear systems*, Applied Numerical Mathematics, 44 (2003), pp. 257–431. 16, 144

[8] S. BELLAVIA AND B. MORINI, *An interior global method for nonlinear systems with simple bounds*, Optimization Methods and Software, 20 (2005), pp. 1–22. 16, 144

[9] I. BEN GHARBIA AND J. GILBERT, *Nonconvergence of the plain Newton-min algorithm for linear complementarity problems with a P-matrix – The full report*,

Rapport de Recherche 7160, INRIA, BP 105, 78153 Le Chesnay, France, 2009. [preprint]. 78, 88, 95, 97

[10] ——, *Nonconvergence of the plain Newton-min algorithm for linear complementarity problems with a P-matrix*, Mathematical Programming, 134 (2012), pp. 349–364. [doi]. 19, 23, 61, 170, 175

[11] ——, *An algorithmic characterization of P-matricity*, SIAM Journal on Matrix Analysis and Applications (revised), (2012). [preprint]. 19, 23, 85, 87, 88, 95, 175

[12] ——, 2012. In preparation. 84

[13] I. BEN GHARBIA AND J. JAFFRÉ, *Gas phase appearance and disappearance as a problem with complementarity constraints*, Mathematics and Computers in Simulation (revised), (2012). [preprint]. 16, 19, 157, 169

[14] M. BERGOUNIOUX, M. HADDOU, M. HINTERMÜLLER, AND K. KUNISCH, *A comparison of a Moreau-Yosida-based active set strategy and interior point methods for constrained optimal control problems*, SIAM Journal on Optimization, 11 (2000), pp. 495–521. 64

[15] M. BERGOUNIOUX, K. ITO, AND K. KUNISCH, *Primal-dual strategy for constrained optimal control problems*, SIAM Journal on Control and Optimization, 37 (1999), pp. 1176–1194. 64, 78

[16] S. BILLUPS AND K. MURTY, *Complementarity problems*, Journal of Computational and Applied Mathematics, 124 (2000), pp. 303–318. 15

[17] J. BONNANS, J. GILBERT, C. LEMARÉCHAL, AND C. SAGASTIZÁBAL, *Numerical Optimization – Theoretical and Practical Aspects* (second edition), Universitext, Springer Verlag, Berlin, 2006. 84, 116

[18] A. BOURGEAT, M. JURAK, AND F. SMAÏ, *Two phase partially miscible flow and transport modeling in porous media ; application to gas migration in a nuclear waste repository*, Computational Geoscience, 13 (2009), pp. 29–42. 170

[19] H. BUCHHOLZER, C. KANZOW, P. KNABNER, AND S. KRÄUTLE, *Solution of reactive transport problems including mineral precipitation-dissolution reactions by a semismooth Newton method*, Tech. Rep. 288, Institute of Mathematics, University of Würzburg, Würzburg, 2009. 16, 84

[20] ——, *Solution of reactive transport problems including mineral precipitation-dissolution reactions by a semismooth Newton method*, Computational Optimization and Applications, 50 (2011), pp. 193–221. 16, 143, 170

[21] R. CHANDRASEKARAN, *A special case of the complementary pivot problem*, Opsearch, 7 (1970), pp. 263–268. 32, 64

[22] G. CHAVENT AND J. JAFFRÉ, *Mathematical Models ans Finite Elements for Reservoir Simulation*, North Holland, 1986. 16, 169

[23] B. CHEN, X. CHEN, AND C. KANZOW, *A penalized Fischer-Burmeister NCP-function*, Mathematical Programming, 88 (2000), pp. 211–216. 40, 41

[24] C. CHEN AND O. MANGASARIAN, *Smoothing methods for convex inequalities and linear complementarity problems*, Mathematical Programming, 71 (1995), pp. 51–69. [doi]. 37

[25] X. CHEN AND C. ZHANG, *Smoothing projected gradient method and its application to stochastic linear complementarity problem*, SIAM Journal on Optimization, 20 (2009), pp. 627–649. 56

[26] F. CLARKE, *Optimization and Nonsmooth Analysis*, Classics in Applied Mathematics, 5, SIAM, Philadelphia, PA, USA, 1990. second edition. 16, 37, 43, 44, 45, 107, 174

[27] T. COLEMAN AND A. VERMA, *ADMAT : An Automatic Differentiation Toolbox for MATLAB*, (1998). 154

[28] A. CONN, N. GOULD, AND P. TOINT, *Trust-Region Methods*, MPS-SIAM Series on Optimization 1, SIAM and MPS, Philadelphia, 2000. 84

[29] R. COTTLE AND G. DANTZIG, *Complementarity pivot theory of mathematical programming*, Linear Algebra and its Applications, 1 (1968), pp. 103–125. 17, 62

[30] R. COTTLE, J.-S. PANG, AND R. STONE, *The linear complementarity problem*, no. 60 in Classics in Applied Mathematics, SIAM, Philadelphia, PA, USA, 2009. 17, 30, 35, 62, 66, 83, 86, 104

[31] G. COXSON, *The P-matrix problem is co-NP-complete*, Mathematical Programming, 64 (1994), pp. 173–178. 30, 83, 101

[32] E. DE KLERK AND M. NAGY, *On the complexity of computing the handicap of a sufficient matrix*, rapport de recherche, Tilburg University, The Netherlands, 2010. 37, 38

[33] J. DE LOS REYES AND K. KUNISCH, *A comparison of algorithms for control constrained optimal control of the burgers equation*, Calcolo, 41 (2004), pp. 203–225. 41

[34] T. DE LUCA, F. FACCHINEI, AND C. KANZOW, *A theoretical and numerical comparison of some semismooth algorithms for complementarity problems*, Computational Optimization and Applications, 16 (2000), pp. 173–205. 41

[35] P. DEUFLHARD, *Newton Methods for Nonlinear Problems – Affine Invariance and Adaptative Algorithms*, no. 35 in Computational Mathematics, Springer, Berlin, 2004. 16

[36] F. FACCHINEI, A. FISCHER, AND C. KANZOW, *Inexact Newton methods for semismooth equations with applications to variational inequality problems*, in Nonlinear Optimization and Applications, G. Di Pillo and F. Giannessi, eds., Plenum Press, New York, 1996, pp. 125–139. 37

[37] F. FACCHINEI AND C. KANZOW, *A nonsmooth inexact Newton method for the solution of large-scale nonlinear complementarity problems*, Mathematical Programming, 76 (1997), pp. 493–512. 16

[38] F. FACCHINEI AND J.-S. PANG, *Finite-Dimensional Variational Inequalities and Complementarity Problems* (two volumes), Springer Series in Operations Research, Springer, 2003. 16, 17, 34, 45, 47, 49, 51, 52, 53, 62, 106, 144, 146, 147, 148, 170, 173, 174, 175

[39] Y. FATHY, *Computatiional complexity of LCPs associated with positive definite symmetric matrices*, Mathematical Programming, 17 (1979), pp. 335–344. 122

[40] M. FERRIS AND K. SINAPIROMSARAN, *Formulating and solving nonlinear programs as mixed complementarity problems*, Lecture Notes in Economics and Mathematical Systems, 481 (2000), p. 132Ð148. 16

[41] M. C. FERRIS AND J. S. PANG, *Engineering and economic applications of complementarity problems*, SIAM Review, 39 (1997), pp. 669–713. 15, 16

[42] M. FIEDLER AND V. PTÁK, *On matrices with nonpositive off-diagonal elements and principal minors*, Czechoslovak Mathematics Journal, 12 (1962), pp. 382–400. 33, 66, 86

[43] ——, *Some generalizations of positive definiteness and monotonicity*, Numerische Mathematik, 9 (1966), pp. 163–172. 29

[44] A. FISCHER, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284. 40

[45] ——, *A Newton-type method for positive semidefinite linear complementarity problems*, Journal of Optimization Theory and Applications, 86 (1995), pp. 585–608. 40

[46] ——, *New constrained optimization reformulation of complementarity problems*, Journal of Optimization Theory and Applications, 97 (1998), pp. 105–117. 16

[47] A. FISCHER AND J. HOUYUAN, *Merit functions for complementarity and related problems : a survey*, Computational Optimization and Applications, 17 (2000), pp. 159–182. 41

[48] A. FISCHER AND C. KANZOW, *On finite termination of an iterative method for linear complementarity problems*, Mathematical Programming, 74 (1996), pp. 279–292. 41, 59, 64

[49] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Research Logistics Quarterly, 3 (1956), pp. 95–110. 39

[50] FRENCH NATIONAL INVENTORY OF RADIOACTIVE MATERIALS, WASTE, AND RESEARCH GROUP MOMAS, *The Couplex-Gas Benchmark*. http://www.gdrmomas.org/ex_qualifications.html/. 18, 158, 169

[51] M. V. GENUCHTEN, *A closed form equation for predicting the hydraulic conductivity of unsaturated soils*, Soil Science Society of America Journal, 44 (1980), pp. 892–898. 135, 176

[52] J. GILBERT, *Éléments d'Optimisation Différentiable – Théorie et Algorithmes*, Syllabus de cours à l'ENSTA, Paris, 2012. [internet]. 16, 17, 26

[53] R. GLOWINSKI, J. LIONS, AND R. TREMOLIÈRES, *Numerical Analysis of Variational Inequalities*, North-Holland, Amsterdam, New York, 1981. 16

[54] G. GOLUB AND C. V. LOAN, *Matrix Computations* (third edition), The Johns Hopkins University Press, Baltimore, Maryland, 1996. 74

[55] S. GRANET, *Présentation : Benchmark multiphasique*, 2011. http://www.gdrmomas.org/Activites/2011/. 159

[56] O. GÜLER, *Foundations of Optimization*, no. 258 in Graduate Texts in Mathematics, Springer, 2010. [doi]. 89

[57] C. HAGER AND B. WOHLMUTH, *Semismooth Newton methods for variational problems with inequality constraints*, GAMM-Mitt, 33 (2010), pp. 8–24. 143, 170

[58] S.-H. HAN, J.-S. PANG, AND N. RANGARAJ, *Globally convergent Newton methods for nonsmooth equations*, Mathematics of Operations Research, 17 (1992), pp. 586–607. 50

[59] P. HARKER AND J.-S. PANG, *A damped-Newton method for the linear complementarity problem*, in Computational Solution of Nonlinear Systems of Equations, E. Allgower and K. Georg, eds., no. 26 in Lecture in Applied Mathematics, AMS, Providence, RI, 1990. 41, 50, 52, 84, 105, 110, 111

[60] ——, *Finite-dimensional variational inequality and nonlinear complementarity problems : A survey of theory, algorithms and applications*, Mathematical Programming, 48 (1990), pp. 161–220. 15, 110, 112, 123, 170

[61] M. HINTERMÜLLER, K. ITO, AND K. KUNISCH, *The primal-dual active set strategy as a semismooth Newton method*, SIAM Journal on Optimization, 13 (2003), pp. 865–888. 33, 60, 64, 65, 66, 67, 77, 83, 88, 170, 175

[62] R. HOPPE, *Multigrid algorithms for variational inequalities*, SIAM Journal on Numerical Analysis, 24 (1987), pp. 1046–1065. 33

[63] R. HORN AND C. JONHSON, *Topics in Matrix Analysis*, Cambridge University Press, New York, NY, USA, 1991. 25, 86

[64] K. ITO AND K. KUNISCH, *On a semi-smooth Newton method and its globalization*, Mathematical Programming, 118 (2009), pp. 347–370. 84

[65] J. JAFFRÉ AND A. SBOUI, *Henry' law and gas phase disappearance*, Transport in Porous Media, 82 (2010), pp. 521–526. 16, 18, 138, 169, 172

[66] M. JANSSEN, *On the structure of the solution set of a linear complementarity problem*, Cahiers Centre Études Rech. Opér., 25 (1983), pp. 41–48. 26

[67] J. JI AND F. POTRA, *An infeasible-interior-point method for the P_*-matrix LCP*, report on computational mathematics, Department of Mathematics, The University of Iowa, Iowa City, IA 52242, USA, 1994. 37

[68] C. KANZOW, *Nonlinear complementarity as unconstrained optimization*, Journal of Optimization Theory and Applications, 88 (1996), pp. 139–155. 16

[69] ——, *Some noninterior continuation methods for linear complementarity problems*, SIAM Journal on Matrix Analysis and Applications, 17 (1996), pp. 851–868. 104

[70] C. KANZOW, *An active set-type Newton method for constrained nonlinear systems*, in Complementarity : applications, algorithms and extensions, M. Ferris, O. Mangasarian, and J. Pang, eds., Kluwer Acad. Publ., Dordrecht, 2001, pp. 179–200. 16, 144

[71] ——, *Inexact semismooth Newton methods for large-scale complementarity problems*, Optimization Methods and Software, 19 (2004), pp. 309–325. 38, 41, 60, 64, 67, 101, 110, 170

[72] C. KANZOW AND H. KLEINMICHEL, *A class of Newton-type methods for equality and inequality constrained optimization*, Optimization Methods and Software, 5 (1995), pp. 173–198. 40

[73] ——, *A new class of semismooth Newton-type methods for nonlinear complementarity problems*, Computational Optimization and Applications, 11 (1998), pp. 117–251. 16

[74] N. KARMARKAR, *A new polynomial-time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395. 17, 62, 104

[75] R. KELLOGG, *On complex eigenvalues of M and P matrices*, Numerische Mathematik, 19 (1972), pp. 170–175. 75

[76] N. KIKUCHI AND J. T. ODEN, *Contact Problems in Elasticity : A Study of Variational Inequalities and Finite Element Methods*, Studies in Applied and Numerical Mathematics, Siam Philadelphia, 1988. 16

[77] D. KINDERLEHRER AND G. STAMPACCHIA, *An introduction to variational inequalities and their applications*, SIAM, Philadelphia, 2000. 16

[78] M. KOJIMA, N. MEGIDDO, T. NOMA, AND A. YOSHISE, *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*, no. 538 in Lecture Notes in Computer Science, Springer-Verlag, Berlin, 1991. 17, 35, 36, 37, 38, 62

[79] M. KOJIMA AND S. SHINDO, *Extension of Newton and quasi-Newton methods to systems of PC^1 equations*, Journal of Operations Research Society of Japan, 29 (1986), pp. 352–375. 52, 65

[80] M. KOSTRENA, *Direct algorithms for complementarity problems*, PhD thesis, Rensselaer Polytechnic Institute, Troy, New York, 1976. 41, 63, 86

[81] S. KRÄUTLE, *The semismooth Newton method for multicomponent reactive transport with minerals*, tech. rep., Department of Mathematics, University of Erlangen-Nuremberg, Erlangen, Germany, 2010. 84, 170

[82] Y. KUZNETSOV, P. NEITTAANMÄKI, AND P. TARVAINEN, *Overlapping block methods for obstacle problems with convection-diffusion operators*, In M.C. Ferris, J.-S. Pang, éditeurs, Complementarity and Variational Problems, SIAM, Philadelphia, USA, 1987. 33

[83] S. LANG, *Linear algebra*, Undergraduate Texts in Mathematics, Springer, 1987. (third edition). 95

[84] A. LAUSER, C. HAGER, R. HELMIG, AND B. WOHLMUTH, *A new approach for phase transitions in miscible multi-phase flow in porous media*, Advances in Water Resources, 34 (2011), pp. 957–966. 16, 143, 170

[85] C. LEMARÉCHAL, I. BEN GHARBIA, AND J. GILBERT, *Suggestions in discussions*, 2011. 104

[86] C. LEMKE, *Bimatrix equilibrium points and mathematical programming*, Management Science, 11 (1965), pp. 681–689. 16, 17, 37, 62, 123

[87] C. LEMKE AND J. HOWSON, *Equilibrium points of bimatrix games*, Management Science, 12 (1964), pp. 413–423. 16

[88] S. LEYFFER, G. LÓPEZ-CALVA, AND J. NOCEDAL, *Interior methods for mathematical programs with complementarity constraints*, SIAM Journal on Optimization, 17 (2006), pp. 52–77. 37

[89] N. M. M. KOJIMA AND Y. YE, *An interior point potential reduction algorithm for the linear complementarity problem*, Mathematical Programming, 54 (1994), pp. 267–279. 38

[90] O. MANGASARIAN, *Equivalence of the complementarity to a system of nonlinear equations*, SIAM Journal on Applied Mathematics, 31 (1976), pp. 96–119. 41

[91] ——, *Solution of symmetric linear complementarity problems by iterative methods*, Journal of Optimization Theory and Applications, 22 (1977), pp. 465–485. 41, 63, 86

[92] O. L. MANGASARIAN, *The ill-posed linear complementarity problem*, 1995. 17

[93] E. MARCHAND, T. MÜLLER, AND P. KNABNER, *Fully coupled generalized hybrid-mixed finite element approximation of two-phase two-component flow in porous media. Part I : Mathematical model*, Submitted, (2012). 170

[94] ——, *Fully coupled generalized hybrid-mixed finite element approximation of two-phase two-component flow in porous media. Part II : Numerical scheme and numerical results*, Computational Geosciences, 16 (2012), pp. 691–708. 143, 170

[95] G. D. MARSILY, *Hydrogéologie quantitative*, Collection sciences de la terre, Masson, 1981. 131

[96] J. MARTÍNEZ AND L. QI, *Inexact Newton methods for solving nonsmooth equations*, Journal of Computational and Applied Mathematics, 60 (1995), pp. 127–145. 37

[97] N. MEGIDDO, *A note on the complexity of P-matrix LCP and computing an equilibrium*, Tech. Rep. RJ 6439 (62557), 1988. IBM Research, Almaden Research Center, 650 Harry Road, San Jose, CA, USA. 83

[98] N. METLA, *The Sequential Quadratic Programming Method for Elliptic Optimal Control Problems with Mixed Control-State Constraints*, PhD thesis, Johann Radon Institute for Computational and Applied Mathematics, Johannes Kepler Universität, Linz, Austria, 2008. 84

[99] R. MIFFLIN, *Semismooth and semiconvex functions in constrained optimization*, SIAM Journal on Control and Optimization, 15 (1977), pp. 957–972. 51, 86

[100] W. MORRIS, *Randomized pivot algorithms for P-matrix linear complementarity problems*, Mathematical Programming, 92A (2002), pp. 285–296. 38, 68, 83

[101] T. MUNSON, F. FACCHINEI, M. FERRIS, A. FISCHER, AND C. KANZOW, *The semismooth algorithm for large scale complementarity problems*, INFORMS Journal on Computing, 13 (2001), pp. 294–311. 16, 37, 41

[102] K. MURTY, *Computational complexity of complementarity pivot methods*, Mathematical Programming Study, 7 (1978), pp. 61–73. 37

[103] ——, *Linear Complementarity, Linear and Nonlinear Programming*, Heldermann Verlag, Berlin, 1988. 17, 37, 38, 62, 104, 122, 123

[104] J. OUTRATA, *Mathematical programs with equilibrium constraints : theory and numerical methods*, in Nonsmooth Mechanics of Solids, J. Haslinger and G. Stavroulakis, eds., no. 485 in CISM Courses and Lectures, Springer, 2006, pp. 221–274. 144

[105] J. OUTRATA, M. KOČVARA, AND J. ZOWE, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*, Kluwer Academic Publishers, Dordrecht, 1998. 144

[106] J.-S. PANG, *Newton's method for B-differentiable equations*, Mathematics of Operations Research, 15 (1990), pp. 311–341. 41, 50, 52, 56, 105, 106, 110

[107] ——, *A B-differentiable equation-based, globally and locally quadratically convergent algorithm for nonlinear programs, complementarity and variational inequality problems*, Mathematical Programming, 51 (1991), pp. 101–131. 41

[108] J. S. PANG AND S. A. GABRIEL, *NE/SQP : A robust algorithm for the nonlinear complementarity problem*, Mathematical Programming, 60 (1993), pp. 295–337. 104, 105

[109] B. PSHENICHNYI, *Necessary Conditions for en Extremum*, Marcel Dekker, New York, 1971. 107

[110] P.TSENG, *Analysis of a non-interior continuation method based on chen-mangasarian smoothing functions for complementarity problems*, in Reformuation : Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, Kluwer Academic Publishers, 1998, pp. 381–404. 16

[111] H. QI AND L. LIAO, *A smoothing Newton method for general nonlinear complementarity problems*, Computational Optimization and Application, 17 (2000), pp. 231–254. 16

[112] L. QI, D. SUN, AND G. ZHOU, *A new look at smoothing Newton methods for nonlinear complementarity problems and box constrained variational inequalities*, Mathematical Programming, 87 (2000), pp. 1–35. 16

[113] L. QI AND J. SUN, *A nonsmooth version of Newton's method*, Mathematical Programming, 58 (1993), pp. 353–367. 16, 41, 51, 58, 86

[114] RESEARCH GROUP MOMAS, *Mathematical Modeling and Numerical Simulation for Nuclear Waste Management Problems*. http://www.gdrmomas.org/. 127, 158, 169

[115] S. ROBINSON, *Local structure of feasible sets in nonlinear programming, part III : stability and sensitivity. mathematica*, Mathematical Programming Study, 30 (1987), pp. 45–66. 49

[116] J. ROHN, *On Rump's characterization of P-matrices*, Optimization Letters, 6 (2012), pp. 1017–1020. [doi]. 86

[117] S. M. RUMP, *On P-matrices*, Linear Algebra and its Applications, 363 (2003), pp. 237–250. [doi]. 86

[118] H. SAMELSON, R. THRALL, AND O. WESLER, *A partition theorem for the Euclidean n-space*, Proceedings of the American Mathematical Society, 9 (1958), pp. 805–807. 62, 83, 86

[119] U. SCHÄFER, *A linear complementarity problem with a P-matrix*, SIAM Review, 46 (2004), pp. 189–201. 83

[120] A. SHAPIRO, *On concepts of directional differentiability*, Journal of Optimization Theory and Applications, 66 (1990), pp. 477–487. 50

[121] F. SMAÏ, *Apparition-disparition de phase dans un écoulement diphasique eau/hydrogène en milieu poreux : Injection de gaz dans un milieu saturé en eau pure*, 2009. http://sources.univ-lyon1.fr/cas_test/. 161

[122] D. SOLOW, R. STONE, AND C. TOVEY, *Solving LCP on P-matrices is probably not NP-hard*. Unpublished note, 1987. 83

[123] D. SUN, J. HAN, AND Y. ZHAO, *On the finite termination of the damped-Newton algorithm for the linear complementarity problem*, Acta Mathematica Numerica applicatae, 21 (1998), pp. 148–154. 41

[124] P. TSENG, *Co-NP-completeness of some matrix classification problems*, Mathematical Programming, 88 (2000), pp. 183–192. 27, 83, 101

[125] M. ULBRICH, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, no. 11 in MPS-SIAM Series on Optimization, SIAM Publications, Philadelphia, PA, USA, 2011. [doi]. 86

[126] S. WRIGHT, *Primal-Dual Interior-Point Methods*, SIAM Publication, Philadelphia, 1997. 16, 144

[127] Y. S. WUA AND P. A. FORSYTH, *On the selection of primary variables in numerical formulation for modeling multiphase flow in porous media*, J. Contam. Hydrol, 48 (2001), pp. 277–304. 18

Index

Vu :
Le Président
M.

Vu :
Les Suffrageants
M.

Vu et permis d’imprimer :
Le Vice-Président du Conseil Scientifique Chargé de la Recherche de l’Université Paris Dauphine