Numerical resolution of algebraic systems with complementarity conditions. Application to the thermodynamics of compositional multiphase mixtures

Duc Thach Son Vu

Thèse de doctorat

NNT : 2020UPASG006

université PARIS-SACLAY

Numerical resolution of algebraic systems with complementarity conditions

Application to the thermodynamics of compositional multiphase mixtures

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 580 (STIC)
Sciences et technologies de l'information et de la communication
Spécialité de doctorat : Mathématiques et Informatique
Unité de recherche : IFPEN, Sciences et Technologies du Numérique, 92852, Rueil-Malmaison, France
Référent : CentraleSupélec

Thèse présentée et soutenue à Rueil-Malmaison, le 7 octobre 2020, par

VU Duc Thach Son

Composition du Jury

M. Abdel LISSER Professeur des universités, Université Paris-Saclay	Président
M. Samir ADLY Professeur des universités, Université de Limoges	Rapporteur
M. Didier AUSSEL Professeur des universités, Université de Perpignan	Rapporteur
Mme Hoai An LE THI Professeure des universités, Université de Lorraine	Examinatrice
M. Jean-Pierre DUSSAULT Professeur titulaire, Université de Sherbrooke	Examinateur
M. Quang Huy TRAN Ingénieur de recherche, HDR, IFP Energies nouvelles	Directeur de thèse
M. Mounir HADDOU Professeur des universités, INSA Rennes	Co-Directeur de thèse
Mme Ibtihel BEN GHARBIA Ingénieure de recherche, IFP Energies nouvelles	Co-encadrante

Acknowledgments

This thesis will never be completed without the support, encouragement, and knowledge from many people to whom I would like to express my deep gratitude.

This work took place in the Applied Mathematics Department of IFP Energies Nouvelles. This manuscript represents the culmination of three years very enriching, both from a scientific and human point of view. That's why I want in the first place to thank my three advisors who helped me to make this experience an exceptional adventure. Without their dedication to teaching, training, and growing me, I would not be here writing this acknowledgment.

First of all, from the bottom of my heart, my deepest gratitude goes to my thesis director Dr. Tran Quang Huy. He guided me through the three years enthusiastically with all his meticulous, dedication, and knowledge. I learned the way to write mathematics neatly and English usage from all the precise hand-writing papers he wrote for me; the way to stir and then still have an idea to escape a stuck point from all discussions with him or the way to organize my work well from all tips he gave me. He was a guide full of ideas for me during my Ph.D. I began my Ph.D. with many missing skills so everything even small I collected from him, I am grateful. His optimism and psychology always help me through the difficult periods of my Ph.D. He helped me to grow not only in my work but also in my life for the past 3 years.

I wish to express my gratitude most strongly to my thesis co-director Professor Mounir Haddou. I appreciate all his supports from being my master's teacher in Vietnam to becoming my Ph.D. advisor. Although not working directly much, but all the multi-day trips working with him in INSA Rennes always give me valuable experiences. His ideas and extensive knowledge have helped me a lot in building and completing my work. I also greatly appreciate his close friendship and valuable cooperation with French-Vietnam Master 2 program over the years. I am very proud to be his first Vietnamese Ph.D. student.

I send the most special thanks to my remaining advisor Dr. Ibtihel Ben Gharbia. I especially emphasize what I have learned in programming from her. It could just be a necessary space in a command line, a reminder for a comment, or a name for a file. I was enlightened a lot about programming more than when I was in college. I still remember many times when she spent hours late debugging or patiently teaching me a new programming language to me. Her strict times or friendly talks help me grow up a lot, I appreciate it. I am very proud of being her first Ph.D. student.

I am also thankful to the members of the jury for my Ph.D. thesis who examined and decided Ph.D. diploma for me, Prof. Abdel Lisser, Prof. Samir Adly, Prof. Didier Aussel, Prof. Le Thi Hoai An, and Prof. Jean-Pierre Dussault. This manuscript was also corrected by their remarks.

I send a thank to my colleagues, especially Ph.D. students, at the IFP Energies Nouvelles. In particular, thank Zakariae, Gouranga, Karine, Bastian, Nicolas, Henry, Arsene, Sylvie, Mani, Julien, Riad, Sabrina, Joelle, Guissel, Alexis, Karim, Jingang.

A special mention goes to my Vietnamese friends. I cannot imagine how dull the past three

years in France would have been without them. Thank Nguyen Van Thanh, Tran Thi Thoi, Do Minh Hieu, Phan Tan Binh, Nguyen Viet Anh, Nguyen Tien Dat, Phung Thanh Tam, Nguyen Thi Hoai Thuong, Ho Kieu Diem, Hoang Thi Kieu Loan, Le Tran Ngoc Tran, Tran Hoai Thuan, Cao Van Kien, Le Minh Duy, Ngo Tri Dat, Trinh Ngoc Tu, Nguyen Manh Quan, Nguyen Thi Thu Dieu, Cao Ngoc Yen Phuong.

Finally, I would like to express all my gratitude to my family, mom Ngo Thi Nga dad Vu Duc Ha younger sister Vu Thach Thao Phuong, and my girlfriend, also my fiancée Tran Thi Huong, for their unconditional encouragement and love. I always remember the midnight video calls in Vietnam because of the 5-6 hour gap with France or the few trips back to Vietnam. During the past three years, every day, they are the ones who listen and share with me all the joys as well as the sorrows or pressures I went through. Here is the last place I can lean on, also the motivation for me to move on and overcome challenges. Thank them too much for always being behind me in this journey, and next journeys...

Cuối cùng, tôi xin gửi lời cảm ơn sâu sắc nhất đến gia đình tôi, mẹ Ngô Thị Ngà ba Vũ Đức Hà em gái Vũ Thạch Thảo Phương, và người yêu của tôi, cũng là vợ sắp cưới Trần Thị Hương, đã động viên và yêu thương vô điều kiện. Tôi vẫn luôn nhớ những cuộc gọi video lúc nửa đêm ở Việt Nam vì khoảng cách 5-6 giờ với Pháp hay những chuyến trở về Việt Nam ít ỏi. Trong suốt ba năm qua, mỗi ngày, họ đều là những người lắng nghe và chia sẻ với tôi mọi niềm vui cũng như nỗi buồn hay áp lực mà tôi đã trải qua. Đây là nơi cuối cùng tôi có thể tựa vào, cũng là động lực để tôi bước tiếp và vượt qua thử thách. Cảm ơn họ rất nhiều vì đã luôn ở phía sau tôi trong hành trình này, và những hành trình tiếp theo...

Contents

Chapter 1

Introduction

Contents

Ce chapitre présente les motivations de la thèse ainsi que les principales contributions. Il fait aussi office de "résumé en français" requis par l'École Doctorale, d'où la différence dans la langue de rédaction avec les autres chapitres.

Nous décrivons d'abord en §1.1 le contexte général en partant de l'application métier à l'origine du problème, à savoir la simulation de réservoir. Nous y donnons un aperçu des modèles physiques utilisés et de leurs difficultés mathématiques au regard de la gestion de l'apparition et de la disparition des phases. Un accent particulier est mis sur la formulation unfiée*, où l'emploi des conditions de complémentarité permet de gagner en clarté et confort au prix de nouvelles difficultés d'ordre numérique, voire théorique pour certaines lois thermodynamiques comme les équations d'état cubiques.*

Une synthèse de l'état de l'art est ensuite fournie en §1.2 sur les méthodes de résolution numérique des systèmes algébriques contenant des conditions de complémentarité. Celles-ci sont divisées en deux catégories. La première comporte les méthodes non-lisses et semi-lisses dont fait partie Newton-min*, l'algorithme par défaut actuel dans les codes d'IFPEN. La seconde regroupe les méthodes de régularisation par les θ-fonctions de lissage ainsi que les méthodes de* points intérieurs*.*

Enfin, la dernière section §1.3 explique notre démarche et récapitule les résultats obtenus. Ce sera également l'occasion d'exposer le plan du mémoire.

1.1 Gestion des phases d'un mélange compositionnel

1.1.1 Simulation des écoulements polyphasiques multiconstituants

La simulation de réservoir est l'art d'utiliser les techniques numériques pour prédire le comportement des écoulements de fluides dans les milieux poreux, connaissant les conditions initiales et aux limites appropriées [8]. Née il y a plus d'un demi-siècle avec l'avènement des ordinateurs, elle est aujourd'hui devenue une technologie mature, avec un abondant catalogue de modèles adaptés aux différents besoins et une vaste panoplie de méthodes numériques performantes [26,28]. Jadis dédiées à la récupération des hydrocarbures dans le sous-sol, les mêmes équations sont depuis une décennie orientées vers des enjeux plus conformes à notre époque, comme la séquestration du dioxide de carbone dans les aquifères salines, le stockage de gaz dans les réservoirs géologiques ou l'enfouissement des déchets radioactifs...

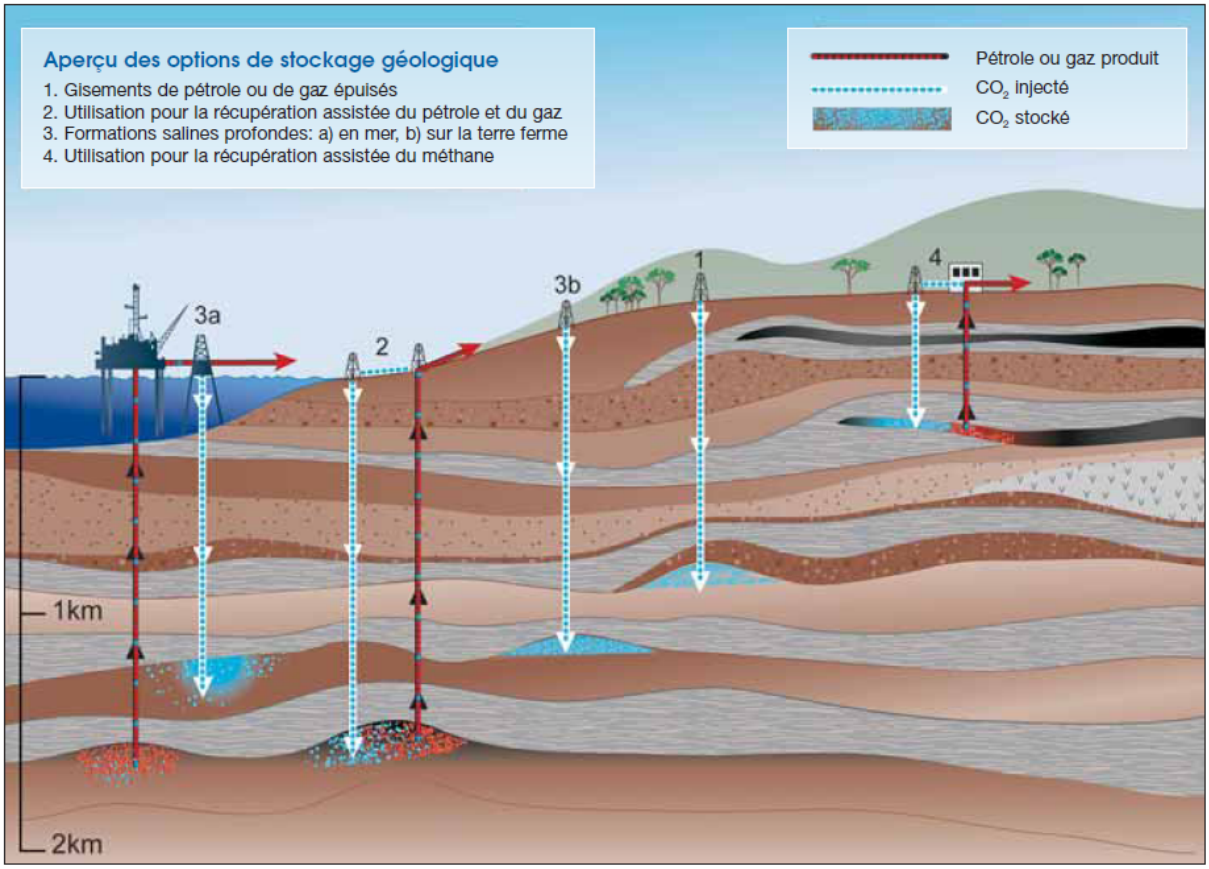


Figure 1.1: Diverses options de stockage sous-terrain. © GIEC 2005

Que ce soit pour le pétrole ou pour des finalités plus modernes, une caractéristique commune dans le cahier des charges que doit remplir un simulateur est sa capacité à traiter des cas "réalistes" faisant intervenir des dizaines ou des centaines d'*espèces* chimiques différentes. Même lorsque ces espèces ne réagissent pas entre elles, les lois régissant leur équilibre thermodynamique font que chaque espèce — ou chaque *constituant* — peut se retrouver sous une ou plusieurs phases différentes. Le concept de *phase* correspond grossièrement à l'intuition que nous avons

des états de la matière (gaz, liquide, solide), mais pas toujours (par exemple, l'huile est considérée comme phase distincte de l'eau). Les modèles qui prennent en compte cet aspect sont qualifiés de *polyphasiques compositionnels* ou polyphasiques *multiconstituants*. Dans la hiérarchie des modèles d'écoulement en milieu poreux, ce sont de loin les plus complexes[1].

La difficulté avec un mélange polyphasique compositionnel est que nous ne pouvons prévoir ni où et quand une nouvelle phase va apparaître, ni où et quand une ancienne phase va disparaître. Tout au mieux pouvons-nous poser les équations correspondant aux lois physiques considérées et "attendre que cela se passe". Or, la manière même de poser les équations fait débat. Pour expliquer ce point avec précision, nous allons introduire quelques notations en vue d'écrire... des équations. Soit

$$\mathcal{K} = \big\{\mathrm{I},\, \mathrm{II},\, \ldots,\, \mathrm{K}\big\}, \qquad \mathrm{K} \geqslant 2, \tag{1.1}$$

l'ensemble des constituants, et

$$\mathscr{P} = \big\{1,\, 2,\, \ldots,\, P\big\}, \qquad P \geqslant 2, \tag{1.2}$$

l'ensemble des phases virtuellement envisageables. S'il existe au moins une espèce $i \in \mathcal{K}$ dans la phase $\alpha \in \mathscr{P}$, celle-ci est dite *présente*. Le sous-ensemble $\Gamma(\boldsymbol{\chi}, t) \subset \mathscr{P}$ des phases présentes à une position $\boldsymbol{\chi} \in \mathbb{R}^3$ donnée et à un instant $t \in \mathbb{R}_+$ donné est appelé *contexte*. Ce dernier dépend ainsi de l'espace et du temps. Pour chaque phase présente $\alpha \in \Gamma$, on définit les fractions partielles x_α^j pour tout $j \in \mathcal{K}$, fonctions de $(\boldsymbol{\chi}, t)$. Celles-ci mesurent l'importance relative de chaque constituant au sein de la phase présente α.

Considérons le modèle d'écoulement polyphasique compositionnel en milieu poreux suivant, qui est très simpliste mais qui contient l'essence de la difficulté.

ÉTANT DONNÉS

$$\phi,\ \{\rho_\alpha^\circ\}_{\alpha\in\mathscr{P}},\ \{\Phi_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\mathscr{P}},\ \{\lambda_\alpha\}_{\alpha\in\mathscr{P}}\ ,$$

CHERCHER

$$\Gamma \subset \mathscr{P},\ \{\mathsf{S}_\alpha\}_{\alpha\in\Gamma} \geqslant 0,\ \{x_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\Gamma} \geqslant 0,\ \{\boldsymbol{u}_\alpha\}_{\alpha\in\Gamma},\ \mathsf{P}$$

fonctions de $(\boldsymbol{\chi}, t) \in \mathbf{D}_\chi \times \mathbb{R}_+$, où $\mathbf{D}_\chi \subset \mathbb{R}^3$ est un domaine borné, satisfaisant

- les lois de conservation massique

$$\phi \frac{\partial}{\partial t} \sum_{\beta\in\Gamma} \rho_\beta^\circ \mathsf{S}_\beta x_\beta^i + \mathrm{div}_{\boldsymbol{\chi}} \sum_{\beta\in\Gamma} \rho_\beta^\circ x_\beta^i \boldsymbol{u}_\beta = 0, \qquad \forall i \in \mathcal{K}\ ; \tag{1.3a}$$

- les relations bilans

$$\sum_{\beta\in\Gamma} \mathsf{S}_\beta - 1 = 0, \tag{1.3b}$$

$$\sum_{j\in\mathcal{K}} x_\alpha^j - 1 = 0, \qquad \forall \alpha \in \Gamma\ ; \tag{1.3c}$$

- les égalités de fugacité

$$x_\alpha^i \Phi_\alpha^i(\boldsymbol{x}_\alpha, \mathsf{P}) - x_\beta^i \Phi_\beta^i(\boldsymbol{x}_\beta, \mathsf{P}) = 0, \qquad \forall (i, \alpha, \beta) \in \mathcal{K} \times \Gamma \times \Gamma, \tag{1.3d}$$

où $\boldsymbol{x}_\alpha = (x_\alpha^{\mathrm{I}},\, \ldots,\, x_\alpha^{\mathrm{K}-1}) \in \mathbb{R}^{\mathrm{K}-1}$ est le vecteur des fractions partielles indépendantes ;

[1] bien plus que le modèle de *black-oil*, mieux connu du grand public mais qui n'en est qu'un cas très particulier.

- les lois de Darcy-Muskat
$$\boldsymbol{u}_\alpha = -\lambda_\alpha \nabla_{\boldsymbol{\chi}} \mathsf{P}, \qquad \forall \alpha \in \Gamma. \tag{1.3e}$$
- les conditions de Neumann homogènes sur le bord $\partial \mathbf{D}_\chi$.

La quantité ϕ représente le champ de *porosité*, supposé connu en fonction de l'espace $\boldsymbol{\chi}$. Les quantités ρ_α° représentent les *densités* (ou masse volumique) des phases, ici supposées constantes, ce qui correspond à un écoulement incompressible. Il n'y a ni gravité ni capillarité (une pression par phase) dans le modèle (1.3).

Les équations aux dérivées partielles (1.3a) expriment les lois fondamentales de conservation de chaque constituant $i \in \mathcal{K}$. Ces bilans matière sont suppléés par les identités (1.3b)–(1.3c) qui découlent de la définition des fractions partielles x_β^i et des *saturations* S_β, qui mesurent le taux de présence globale des phases et qui sont des inconnues. Les égalités de fugacité (1.3d) sont encore appelées *relations d'équilibre*, car traduisent l'équilibre thermodynamique pour chaque constituant $i \in \mathcal{K}$ à travers deux phases présentes. Les fonctions données Φ_α^i sont les *coefficients de fugacité*. Elles impliquent la pression P, qui est aussi un champ inconnu. Le gradient de ce champ apparaît dans les lois de Darcy-Muskat (1.3e) donnant empiriquement la vitesse de filtration $\boldsymbol{u}_\alpha$ de chaque phase. Le coefficient de proportionnalité λ_α entre $\nabla_{\boldsymbol{\chi}}\mathsf{P}$ et $\boldsymbol{u}_\alpha$ est une fonction donnée de la saturation S_α et de la composition partielle $\{x_\alpha^i\}_{i\in\mathcal{K}}$. Elle encapsule la perméabilité absolue, la perméabilité relative et la viscosité de la phase [26].

Si le contexte Γ est connu, on peut vérifier qu'il y a $|\Gamma|\mathrm{K} + 4|\Gamma| + 1$ équations scalaires pour $|\Gamma|\mathrm{K} + 4|\Gamma| + 1$ inconnues scalaires, où $|\Gamma|$ désigne le cardinal de Γ. Ceci montre que le système (1.3) est fermé. Le plus gênant est que le contexte Γ est aussi une inconnue, fonction de l'espace et du temps, alors qu'il n'y a pas vraiment d'équation qui permette de le déterminer sans ambiguïté. C'est là qu'il faut exploiter les conditions de positivité sur S_α et x_α^i. Lorsqu'on se donne une partie $\Gamma(\boldsymbol{\chi}, t)$ quelconque de $\mathscr{P}$ et qu'on résout le système, rien ne garantit que les saturations et les fractions partielles sont toutes positives. Le "bon" contexte — à supposer qu'il soit unique — est celui pour lequel $\mathsf{S}_\alpha \geqslant 0$ et $x_\alpha^i \geqslant 0$ pour tout $\alpha \in \Gamma$.

La plus mauvaise méthode pour trouver $\Gamma(\boldsymbol{\chi}, t)$ serait d'essayer de manière combinatoire tous les sous-ensembles de $\mathscr{P}$. Ici, cela est tout à fait exclu puisque ce sous-ensemble dépend des variables continues $(\boldsymbol{\chi}, t)$. Même après discrétisation en espace et en temps, le nombre de configurations à essayer serait astronomique ! Il vaut mieux partir d'une approximation initiale de $\Gamma(\boldsymbol{\chi}, t)$ qu'on corrige au fur et à mesure, en tenant compte des informations *a priori* dont on dispose sur l'écoulement. Souvent, une approximation initiale "raisonnable" peut être obtenue au moyen d'un *flash négatif* [1, 117] : on commence par supposer que toutes les phases sont présentes, i.e., $\Gamma(\boldsymbol{\chi}, t) = \mathscr{P}$; on résout le système et détecte pour chaque $(\boldsymbol{\chi}, t)$ les phases pour lesquelles la condition de positivité est respectée et ne garde que celles-ci dans le contexte actualisé.

Nous convenons d'appeler (1.3) la formulation en *variables naturelles* ou formulation de Coats, malgré un léger abus de vocabulaire. En fait, ce que les ingénieurs entendent par "formulation" n'est pas seulement un ensemble d'équations. L'usage de ce mot inclut aussi un choix de variables *primaires* et d'équations primaires, par opposition aux variables *secondaires* qui seront éliminées grâce aux équations secondaires. Le choix préconisé par Coats [30] est de prendre comme inconnues primaires P, $\{\mathsf{S}_\alpha\}_{\alpha\in\Gamma}$ et $\{x_\alpha^i\}_{\alpha\in\Gamma}$ et comme équations primaires (1.3a)–(1.3d). Les inconnues secondaires $\boldsymbol{u}_\alpha$ sont éliminées soit préalablement soit au niveau du système linéaire à l'intérieur de Newton par l'équation secondaire (1.3e). Cette étape n'est certes pas essentielle pour notre problème cible, qui est la gestion des changements de phase. Néanmoins, il demeure important dans les calculs pratiques parce qu'en diminuant la taille du système algébrique à

résoudre à chaque pas de temps, il permet de réduire notablement le temps de calcul. Il existe un grand nombre d'autres formulations possibles. Une revue assez complète a été effectuée dans [25, 116].

La formulation en variables naturelles ou de Coats est celle implantée actuellement dans les logiciels d'IFPEN. Elle porte aussi le nom de formulation en *variable switching*. En effet, le jeu d'inconnues et d'équations n'est pas fixe et doit être constamment ajusté en fonction des changements locaux de contexte. Autrement dit, le "switching" se produit sans cesse pour chaque maille et à chaque pas de temps selon que les hypothèses émises sur le contexte sont violées ou non, à l'instar d'une méthode de type *active set* en optimisation sous contraintes. Il se produit même d'une itération de Newton à l'autre, en cas de négativité des saturations ou des fractions partielles, ce qui laisse de sérieux doutes au niveau théorique quant au système qu'on veut vraiment résoudre. Au niveau informatique, cette gestion dynamique est lourde à mettre en œuvre et consommatrice en temps de calcul. C'est là son inconvénient majeur.

1.1.2 Apports et revers des conditions de complémentarité

En 2011, une nouvelle formulation proposée par Lauser et al. [78] a retenu l'attention de la communauté des numériciens en écoulements polyphasiques compositionnels. À l'aide d'une notion de fractions partielles étendues et surtout des conditions de complémentarité, les auteurs parviennent à donner un traitement unifié aux phases présentes et absentes, d'où la dénomination de *formulation unifiée*. Voici ce que devient (1.3) dans la formulation unifiée.

ÉTANT DONNÉS

$$\phi,\ \{\rho_\alpha^\circ\}_{\alpha\in\mathscr{P}},\ \{\Phi_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\mathscr{P}},\ \{\lambda_\alpha\}_{\alpha\in\mathscr{P}}\ ,$$

CHERCHER

$$\{\mathsf{S}_\alpha\}_{\alpha\in\mathscr{P}},\ \{\xi_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\mathscr{P}},\ \{\boldsymbol{u}_\alpha\}_{\alpha\in\mathscr{P}},\ \mathsf{P}$$

fonctions de $(\boldsymbol{\chi}, t) \in \mathbb{R}^3 \times \mathbb{R}_+$ satisfaisant

- les lois de conservation massique

$$\phi\frac{\partial}{\partial t}\sum_{\beta\in\mathscr{P}}\rho_\beta^\circ\mathsf{S}_\beta\xi_\beta^i + \operatorname{div}_{\boldsymbol{\chi}}\sum_{\beta\in\mathscr{P}}\rho_\beta^\circ\xi_\beta^i\boldsymbol{u}_\beta = 0, \qquad \forall i\in\mathcal{K}\ ; \tag{1.4a}$$

- la conservation du volume

$$\sum_{\beta\in\mathscr{P}}\mathsf{S}_\beta - 1 = 0\ ; \tag{1.4b}$$

- les égalités de fugacité *étendue*

$$\xi_\alpha^i\Phi_\alpha^i(\boldsymbol{x}_\alpha,\mathsf{P}) - \xi_\beta^i\Phi_\beta^i(\boldsymbol{x}_\beta,\mathsf{P}) = 0, \qquad \forall(i,\alpha,\beta)\in\mathcal{K}\times\mathscr{P}\times\mathscr{P}, \tag{1.4c}$$

où les composantes de $\boldsymbol{x}_\alpha = (x_\alpha^{\mathrm{I}}, \ldots, x_\alpha^{\mathrm{K}-1}) \in \mathbb{R}^{\mathrm{K}-1}$ sont *définies* comme

$$x_\alpha^i = \frac{\xi_\alpha^i}{\sum_{j\in\mathcal{K}}\xi_\alpha^j}\ ; \tag{1.4d}$$

- les conditions de complémentarité

$$\min\left(\mathsf{S}_\beta,\, 1-\sum_{j\in\mathcal{K}}\xi_\beta^j\right)=0,\qquad \forall\beta\in\mathscr{P}\ ; \tag{1.4e}$$

- les lois de Darcy-Muskat

$$\boldsymbol{u}_\alpha=-\lambda_\alpha\nabla_{\boldsymbol{\chi}}\mathsf{P},\qquad \forall\alpha\in\mathscr{P}. \tag{1.4f}$$

Les fractions partielles x_α^i, auparavant définies seulement pour les phases présentes $\alpha\in\Gamma$, sont désormais remplacées par les fractions étendues ξ_α^i, définies pour toutes les phases $\alpha\in\mathscr{P}$. Le contexte Γ s'est totalement éclipsé du nouveau système. Si l'on veut le retrouver *a posteriori*, il suffit de chercher les phases α telles que $\mathsf{S}_\alpha>0$. Dans les relations d'équilibre étendues (1.4c), notons que le premier argument du coefficient de fugacité Φ_α^i doit être le vecteur des fractions étendues renormalisées par (1.4d), de sorte que les x_α^i ainsi calculés jouent encore le rôle de fractions partielles "classiques".

La véritable nouveauté réside dans les conditions de complémentarité (1.4e), qui expriment au fond que

$$\mathsf{S}_\beta\geqslant 0,\qquad 1-\sum_{j\in\mathcal{K}}\xi_\beta^j\geqslant 0,\qquad \mathsf{S}_\beta\left(1-\sum_{j\in\mathcal{K}}\xi_\beta^j\right)=0, \tag{1.5a}$$

ce qui peut encore s'écrire plus savamment comme

$$0\leqslant\mathsf{S}_\beta\perp 1-\sum_{j\in\mathcal{K}}\xi_\beta^j\geqslant 0. \tag{1.5b}$$

Autrement dit, au moins l'une des deux quantités est nulle tandis que l'autre doit garder le signe positif. Concrètement, si $\mathsf{S}_\beta>0$, à savoir si la phase β est présente, alors nécessairement $\sum_{j\in\mathcal{K}}\xi_\beta^j=1$. Il en résulte par (1.4c) que $\xi_\beta^i=x_\beta^i$, c'est-à-dire que les fractions étendues de la phase coïncident avec les fractions partielles classiques. Si $\mathsf{S}_\beta=0$, à savoir si la phase β est absente, on a *a priori* $\sum_{j\in\mathcal{K}}\xi_\beta^j\leqslant 1$. Dans le sous-cas $\sum_{j\in\mathcal{K}}\xi_\beta^j<1$, on parle d'absence *stricte* pour la phase β. Dans le sous-cas contraire, si $\sum_{j\in\mathcal{K}}\xi_\beta^j=1$, on a affaire à un point de *transition* qui marque la frontière entre la présence et l'absence de la phase β.

La formulation unifiée présente l'énorme avantage de travailler avec un jeu fixe d'équations et d'inconnues. Indiscutablement, ce confort est non-négligeable pour l'implémentation pratique. Sur le plan théorique, le cadre semble aussi plus satisfaisant, dans la mesure où les changements de phase sont automatiquement pris en charge par les conditions de complémentarité, ce qui évite entre autres d'avoir à recourir au flash négatif. Il en va ainsi dans de nombreux domaines, notamment en mécanique et en électronique [2], où les conditions de complémentarité s'imposent comme la façon la plus efficace pour exprimer un va-et-vient entre deux régimes de fonctionnement possibles pour un système. Un exemple récent à IFPEN où les conditions de complémentarité ont apporté une réelle avancée concerne la modélisation stratigraphique [102, 103]. Nous en étudierons un modèle très réduit en tant que banc d'essai pour nos méthodes numériques.

Plusieurs équipes se sont intéressées à la formulation unifiée pour les écoulements polyphasiques compositionnels. Outre l'Université de Stuttgart où l'idée a pris naissance, on peut citer Inria avec les travaux doctoraux de Ben Gharbia [11, 17] sur des lois de fugacité relativement simples, l'Université de Nice avec les travaux de Masson et ses co-auteurs [9, 86, 87] sur des lois de fugacité également simples mais en évoluant vers des modèles non-isothermes avec couplage. De

son côté, IFPEN s'est attaché à réaliser des comparaisons entre la formulation de Coats et celle de Lauser sur des cas d'écoulements réalistes, utilisant des coefficients de fugacité associés à des lois d'état cubiques [12, 13, 84, 101]. Ces comparaisons visent d'abord à valider les résultats obtenus par la formulation unifiée, puis à jauger de sa performance du point de vue de la robustesse (qui se manifeste notamment par la convergence de l'algorithme de résolution numérique). On observe qu'en cas de convergence pour la formulation unifiée, le temps de calcul est nettement meilleur, le facteur de gain se situant entre 3 et 10.

La thermodynamique serait-elle une nouvelle terre de conquête pour les conditions de complémentarité ? Nous n'en sommes pas encore là. Si les premiers succès sont prometteurs, ils s'accompagnent aussi d'un certain nombre de défauts mis en évidence lors des travaux précités. Le premier est imputable à la non-différentiabilité des conditions de complémentarité, ce qui empêche l'accès à la méthode de Newton classique. Bien entendu, on peut employer une variante de Newton avec une notion plus faible pour la matrice jacobienne. En l'occurrence, compte tenu de la fonction min pour exprimer la complémentarité (1.4e), c'est naturellement vers la méthode de *Newton-min* [3, 72] que se sont tournées toutes les équipes précédentes. Les détails de la méthode seront données en §1.2.1 et §4.2.3. Pour le moment, faisons le constat que sur certains cas difficiles, par exemple quand le pas de temps est trop grand, Newton-min souffre d'un phénomène de *cyclage* : les itérés oscillent de manière périodique entre quelques états, souvent deux ou trois. Cette pathologie s'explique directement à partir de la discontinuité des dérivées sur des exemples "jouets", comme en §4.4. En somme, à moins de disposer d'une meilleure méthode de résolution du système en formulation unifiée, on n'a fait que reporter la difficulté du problème de départ sur les épaules du solveur non-linéaire.

En marge de cette obstruction générique, commune à tous les systèmes non-différentiables, le déploiement de la formulation unifiée (1.4) se heurte également à un obstacle plus subtil, spécifique à certaines lois de fugacité pourtant couramment utilisées en thermodynamique. À vrai dire, nous n'en étions pas conscients au début et ne l'avons découvert que suite aux nombreux "plantages" du code. Mais il est utile de l'évoquer ici afin de compléter le tableau des difficultés.

Soit

$$\Omega = \left\{ \boldsymbol{x} = (x^{\mathrm{I}}, \ldots, x^{\mathrm{K}-1}) \in \mathbb{R}^{\mathrm{K}-1} \mid x^{\mathrm{I}} > 0, \ldots, x^{\mathrm{K}-1} > 0, 1 - x^{\mathrm{I}} - \ldots - x^{\mathrm{K}-1} > 0 \right\} \tag{1.6}$$

le domaine du vecteur des fractions partielles indépendantes et considérons les mélanges diphasiques, où les phases de $\mathscr{P} = \{G, L\}$ sont le gaz et le liquide. Dans la famille des équations d'état cubiques, les coefficients de fugacité $\Phi_G^i(\boldsymbol{x})$ et $\Phi_L^i(\boldsymbol{x})$ — pour alléger, on omet la dépendance par rapport à la pression P — sont définies par l'intermédiaire d'une équation du troisième degré. Prenons l'exemple de la loi de Van der Waals, où cette équation s'écrit

$$Z^3(\boldsymbol{x}) - [B(\boldsymbol{x}) + 1]Z^2(\boldsymbol{x}) + A(\boldsymbol{x})Z(\boldsymbol{x}) - A(\boldsymbol{x})B(\boldsymbol{x}) = 0, \tag{1.7}$$

où les fonctions $A(\cdot)$, $B(\cdot)$ sont données. Lorsque l'équation admet trois racines réelles, on les nomme

$$Z_L(\boldsymbol{x}) \leqslant Z_I(\boldsymbol{x}) \leqslant Z_G(\boldsymbol{x}).$$

Cette définition de $Z_G(\cdot)$ et $Z_L(\cdot)$ permet de calculer les coefficients de fugacité par

$$\begin{aligned} \ln \Phi_\alpha^i(\boldsymbol{x}) = {} & \frac{B(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} B(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B(\boldsymbol{x})} \left[Z_\alpha(\boldsymbol{x}) - 1 \right] - \ln \left[Z_\alpha(\boldsymbol{x}) - B(\boldsymbol{x}) \right] \\ & + \left[\frac{B(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} B(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B(\boldsymbol{x})} - \frac{2A(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} A(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{A(\boldsymbol{x})} \right] \frac{A(\boldsymbol{x})}{Z_\alpha(\boldsymbol{x})}, \end{aligned} \tag{1.8}$$

pour $i \in \mathcal{K}$ et $\alpha \in \{G, L\}$, où les composantes de $\boldsymbol{\delta}^i = (\delta_{i,\mathrm{I}}, \ldots, \delta_{i,\mathrm{K}-1})$ sont des symboles de Kronecker. Malheureusement, la région des $\boldsymbol{x} \in \Omega$ où la cubique (1.7) possède trois racines réelles ne couvre pas tout Ω. Elle n'en est qu'une modeste partie. Dans le reste de Ω, soit on ne peut définir que $Z_L(\boldsymbol{x})$ mais pas $Z_G(\boldsymbol{x})$, soit vice-versa. Par conséquent, lorsqu'on décrète une égalité de type

$$\xi_G^i \Phi_G^i(\boldsymbol{x}_G) - \xi_L^i \Phi_L^i(\boldsymbol{x}_L) = 0, \tag{1.9}$$

deux scénarios peuvent *grosso modo* se produire. Si les deux phases sont présentes, chaque vecteur $\boldsymbol{x}_\alpha$ se trouve dans le domaine de définition de Z_α et des Φ_α^i. Les deux fugacités étendues au premier membre sont bien définies et l'on peut espérer l'existence d'une solution. Si l'une des phase est absente, disons G, alors seul $\boldsymbol{x}_L$ se trouve dans le domaine de définition de Z_L et la valeur de $\xi_L^i \Phi_L^i(\boldsymbol{x}_L)$ peut ne pas se trouver dans l'ensemble image de $\xi_G^i \Phi_G^i(\boldsymbol{x}_G)$. Dans ce cas, il n'y a pas de solution au système. Pour tenter de satisfaire l'égalité, il faudra faire sortir $\boldsymbol{x}_G$ du domaine de Z_G, ce qui ne pourra se faire sans un prolongement de la fonction Φ_G^i.

L'explication que nous venons de faire s'appuie sur les coefficients de fugacité dans le but d'être la plus courte possible. En §3.3.1, un éclairage supplémentaire sera fourni en termes de fonctions de Gibbs et de leurs gradients, qui sont des grandeurs plus fondamentales et qui permettront d'approfondir notre compréhension de cette difficulté.

Il peut être soutenu que le même défaut des lois cubiques devrait causer le même préjudice à la formulation en variables naturelles. Il n'en est rien. Dans la formulation de Coats, si le contexte est correctement deviné, nous n'avons pas besoin de calculer quoi que ce soit en rapport avec la phase évanescente. En l'absence d'une phase, l'équation (1.9) n'existe pas dans le système et le problème ci-dessus n'est pas pertinent. Si le contexte est mal deviné, nous avons la possibilité de nous rattraper en changeant le contexte. La formulation en variables naturelles n'a pas à aller chercher l'information là où celle-ci n'existe pas. La formulation unifiée s'inflige cette mission impossible, de par sa vocation — ou sa prétention — à traiter toutes les phases sur un pied d'égalité.

Nous avons dit plus haut qu'une "formulation" vient avec un choix de variables primaires et d'équations primaires. Dans la formulation de Lauser, les variables primaires sont P, $\{\mathsf{S}_\alpha\}_{\alpha \in \mathscr{P}}$, $\{\varphi^i\}_{i \in \mathcal{K}}$, où φ^i est la valeur commune de la fugacité étendue de l'espèce i à travers les phases. Les fractions étendues ξ_α^i sont alors prééliminées par l'inversion du système local $\mathrm{K} \times \mathrm{K}$

$$\xi_\alpha^i \Phi_\alpha^i(\boldsymbol{x}_\alpha) = \varphi^i, \qquad i \in \mathcal{K}, \tag{1.10}$$

dans chaque phase α. Pour les mêmes raisons qu'avant, à cause de la construction par équation d'état cubique des Φ_α^i, $\alpha \in \{G, L\}$, le système (1.10) n'a pas toujours de solution pour tout $\boldsymbol{\varphi} = (\varphi^{\mathrm{I}}, \ldots, \varphi^{\mathrm{K}})$. Les essais numériques de [84, 101] corroborent cette remarque.

1.1.3 Objectifs de la thèse

En dépit de ces deux difficultés majeures, nous avons la conviction que la formulation unifiée présente un fort potentiel pour améliorer la performance des simulateurs d'écoulement polyphasique compositionnel. En soi, formuler de manière unifiée le problème au niveau continu est déjà un progrès considérable. Il serait dommage de s'arrêter en si bon chemin. Pour "transformer l'essai" et aller au bout de l'intérêt de la formulation unifiée, nous devons relever deux défis :

1. Mettre au point une méthode de résolution numérique des systèmes d'équations contenant des conditions de complémentarité, en remplacement de Newton-min. La nouvelle méthode doit avoir une meilleure garantie de convergence et être aussi robuste que possible par rapport aux paramètres du problème, ainsi qu'au point initial.

2. Mettre en place des remèdes éventuellement *ad hoc* pour contourner la difficulté inhérente aux équations d'état cubiques, indépendamment de toute méthode numérique de résolution. Le cas échéant, préciser les conditions mathématiques favorables à l'existence et l'unicité d'une solution dans la formulation unifiée.

Sur le deuxième objectif, il n'y a à notre connaissance aucun travail antérieur, la difficulté ayant été identifiée "en cours de route". Sur le premier objectif, en revanche, il y a une volumineuse littérature.

1.2 Méthodes existantes pour les conditions de complémentarité

Après discrétisation de (1.4) par un schéma Euler implicite en temps et un schéma de type volumes finis en espace sur un domaine borné muni de conditions aux limites appropriées, nous devons résoudre à chaque pas de temps un système de la forme

$$\Lambda(X) = 0, \tag{1.11a}$$
$$\min(G(X),\, H(X)) = 0, \tag{1.11b}$$

dans laquelle $X \in \mathbb{R}^n$ est l'inconnue et $\Lambda : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^{\ell-m}$, $G : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^m$ et $H : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^m$ sont des fonctions continûment différentiables sur le domaine ouvert $\mathcal{D}$. Rappelons que la fonction min dans (1.11b), qui agit composante par composante, n'est qu'une astuce algébrique commode pour exprimer la complémentarité

$$0 \leqslant G(X) \perp H(X) \geqslant 0.$$

Pour être encore plus concis, posons

$$F(X) = \begin{bmatrix} \Lambda(X) \\ \min(G(X),\, H(X)) \end{bmatrix} \in \mathbb{R}^\ell, \tag{1.12a}$$

de sorte que le système à résoudre devient

$$F(X) = 0, \tag{1.12b}$$

où F n'est pas différentiable partout. Nous distinguons deux catégories de méthodes pour la résolution de (1.12), que nous passons rapidement en revue ci-après en faisant référence au chapitre §4 pour de plus amples détails.

1.2.1 Méthodes de Newton non-lisses

Pour une fonction F continûment différentiable, la méthode de Newton

$$X^{k+1} = X^k - [\nabla F(X^k)]^{-1} F(X^k) \tag{1.13}$$

correspond à la recherche d'un zéro du modèle d'approximation locale

$$\bar{X} \mapsto F(\bar{X}) + \nabla F(X^k)(\bar{X} - X^k) \tag{1.14}$$

au voisinage de X^k. Il existe une théorie de Newton *non-lisse* [47, §7.2] qui généralise le modèle local (1.14) en un schéma d'approximation de Newton

$$\bar{X} \mapsto F(X^k) + T(X^k, \bar{X} - X^k), \tag{1.15}$$

où chaque $T(X, \cdot)$ provient d'un ensemble $\mathcal{T}(X)$ soumis à des conditions techniques [Définition 4.5] qui garantissent le caractère bien défini et la convergence locale [Théorème 4.2] à taux quadratique [Théorème 4.3] de l'algorithme généralisé [Algorithme 4.1]. Le lecteur trouvera les énoncés précis de cette théorie en §4.2.1. En réalité, cette théorie de Newton non-lisse est avant tout un cadre opérationnel abstrait qui ne donne pas lieu à un algorithme concret. On ne demande même pas que $T(X^k, \cdot)$ soit linéaire !

Pour avoir un objet plus "palpable", il faut se restreindre aux fonctions F lipschitziennes pour lesquelles on peut définir la sous-différentielle de Bouligand $\partial_B F$ et celle de Clarke ∂F [Définition 4.7], qui est l'enveloppe convexe de la première. Cela ouvre la voie à l'approximation locale linéaire

$$\bar{X} \mapsto F(\bar{X}) + M^k(\bar{X} - X^k) \tag{1.16}$$

où $M^k \in \partial F(X^k)$. Cependant, on ne peut vérifier les hypothèses techniques du cadre non-lisse [Définition 4.5] que pour une sous-classes de fonctions lipschitziennes, définies alors [Définition 4.8] comme les fonctions semi-lisses [92, 105]. Dans ce cas, on parle de méthode de Newton *semi-lisse* [Algorithme 4.2], avec le caractère bien défini [Théorèmes 4.4] et les bons résultats de convergence [Théorème 4.5]. Là encore, les énoncés précis se trouvent en §4.2.2.

Un cas particulier important d'algorithme semi-lisse est la méthode de Newton-min [Algorithm 4.3]. Dans le cas du système (1.12), il est en effet possible de montrer [Proposition 4.2] que les matrices de $\partial_B F(X)$ sont de la forme

$$M = \begin{bmatrix} \nabla\Lambda(X) \\ \nabla\!\!\!\nabla \end{bmatrix}, \qquad \nabla\!\!\!\nabla \in \mathbb{R}^{m\times\ell}, \tag{1.17a}$$

dans laquelle la α-ième ligne de $\nabla\!\!\!\nabla$ pour $\alpha \in \{1, \ldots, m\}$ est

$$\nabla\!\!\!\nabla_\alpha = \begin{cases} \nabla G_\alpha(X) & \text{if } G_\alpha(X) < H_\alpha(X), \\ \nabla G_\alpha(X) \text{ or } \nabla H_\alpha(X) & \text{if } G_\alpha(X) = H_\alpha(X), \\ \nabla H_\alpha(X) & \text{if } G_\alpha(X) > H_\alpha(X). \end{cases} \tag{1.17b}$$

Les défauts de la méthode Newton-min ont été soulignés en §1.1.2. Ils ont été également formalisés dans [11, 14]. Un autre inconvénient avec Newton-min est qu'il est difficile de le "globaliser" par une recherche linéaire afin d'atteindre un comportement globalement convergent.

1.2.2 Méthodes de régularisation

À l'opposé des méthodes non-lisses ou semi-lisses, les méthodes de régularisation tentent d'abord de lisser la fonction F, ce qui introduit un paramètre de régularisation qu'il faudra faire tendre vers 0. Une régularisation de F est la donnée d'une famille de fonctions

$$\left\{ \widetilde{F}(\cdot;\nu) : \ \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^\ell, \ \ \nu > 0 \right\} \tag{1.18}$$

telle que : (i) $\widetilde{F}(\cdot;\nu)$ soit continûment différentiable en X pour tout $\nu > 0$; (ii) $\widetilde{F}(\cdot;\nu)$ soit continue par rapport à ν, selon un certain sens fonctionnel ; (iii) $\lim_{\nu\downarrow 0} \widetilde{F}(\cdot;\nu) = F(\cdot)$, toujours selon un certain sens fonctionnel. À partir d'une valeur courante pour le couple (X^k, ν^k), la stratégie consiste à :

1. Résoudre $\widetilde{F}(X^{k+1};\nu^k) = 0$ en l'inconnue X^{k+1} par la méthode de Newton classique, utilisant X^k comme point initial. Très souvent, pour gagner en temps de calcul, on ne fait qu'une seule itération de Newton.

2. Diminuer le paramètre de régularisation de ν^k à ν^{k+1} à l'aide d'une règle heuristique. Recommencer jusqu'à ce que $F(X^{k+1}) = 0$.

Parmi les nombreuses régularisations possibles d'une condition de complémentarité

$$0 \leqslant v \perp w \geqslant 0, \tag{1.19}$$

où v et w sont des scalaires, celles utilisant les θ-fonctions sont particulièrement élégantes. Elles consistent à traduire d'abord (1.19) sous l'une des formes équivalentes [Lemmes 4.2 et 4.3]

$$v \geqslant 0, \qquad w \geqslant 0, \qquad \mathfrak{S}(v) + \mathfrak{S}(w) \leqslant 1 \tag{1.20a}$$

ou

$$v \geqslant 0, \qquad w \geqslant 0, \qquad \mathfrak{S}(v) + \mathfrak{S}(w) = \mathfrak{S}(v+w), \tag{1.20b}$$

dans lesquelles

$$\mathfrak{S}(t) = \begin{cases} 0 & \text{if } \ t = 0, \\ 1 & \text{if } \ t > 0. \end{cases} \tag{1.21}$$

est la fonction saut[2]. Ensuite, on approche (1.20a)–(1.20b) par

$$v \geqslant 0, \qquad w \geqslant 0, \qquad \theta_\nu(v) + \theta_\nu(w) = 1 \tag{1.22a}$$

ou

$$v \geqslant 0, \qquad w \geqslant 0, \qquad \theta_\nu(v) + \theta_\nu(w) = \theta_\nu(v+w), \tag{1.22b}$$

en utilisant

$$\theta_\nu(t) := \theta\left(\frac{t}{\nu}\right), \qquad \nu > 0, \tag{1.23}$$

comme régularisation de $\mathfrak{S}$, obtenue par contraction d'une fonction "père" $\theta : \mathbb{R}_+ \to [0,1)$ continue, croissante, concave et vérifiant [Définition 4.11]

$$\theta(0) = 0, \qquad \lim_{t\to+\infty} \theta(t) = 1. \tag{1.24}$$

Initiée par Haddou et ses co-auteurs [7, 55], l'approximation de la complémentarité par les θ-fonctions ont trouvé un usage polyvalent dans de nombreux problèmes appliqués [19, 56, 57, 93]. En pratique, pour appliquer cette régularisation au problème (1.12), il est recommandé d'introduire les variables d'écart $V = G(X)$ et $W = H(X)$ avant de considérer le système régularisé

$$\Lambda(X) = 0, \tag{1.25a}$$
$$G(X) - V = 0, \tag{1.25b}$$
$$H(X) - W = 0, \tag{1.25c}$$
$$\nu\left[\theta_\nu(V) + \theta_\nu(W) - 1\right] = 0. \tag{1.25d}$$

[2] *step function* en anglais.

Dans la dernière équation, la fonction θ agit composante par composante, tandis que la prémultiplication par ν sert à prévenir l'explosion les dérivées lorsque $\nu \downarrow 0$.

Les méthodes de points intérieurs [54,118], réputées pour leur grande efficacité en programmation linéaire grâce notamment à leur complexité polynomiale, peuvent s'interpréter comme des méthodes de régularisation. Nous nous intéressons plus particulièrement aux méthodes *primales-duales* [119], dans lesquelles les variables primales (inconnues de départ) et duales (multiplicateurs de Lagrange) jouissent du même statut. Lorsqu'on décortique une méthode de points intérieurs de type primal-dual, on s'aperçoit qu'il s'agit au fond d'une méthode de résolution du système algébrique des conditions d'optimalité de Karush-Kuhn-Tucker (KKT). Le fait que ce système provient d'un problème de minimisation sous contraintes d'inégalité compte finalement peu dans la méthode. Cela laisse donc entrevoir la perspective de transposer ces méthodes au cas d'un système général contenant des conditions de complémentarité.

Le problème de départ (1.12) est remplacé par la suite des problèmes régularisés

$$\Lambda(X) = 0, \tag{1.26a}$$
$$G(X) - V = 0, \tag{1.26b}$$
$$H(X) - W = 0, \tag{1.26c}$$
$$V \odot W - \nu \mathbf{1} = 0, \tag{1.26d}$$

où $\odot$ désigne le produit composante par composante et $\mathbf{1} \in \mathbb{R}^m$ est le vecteur dont toutes les composantes sont égales à 1. De manière plus concise, ce problème s'écrit

$$\mathbf{F}(\mathbf{X}\,; \nu) = 0, \tag{1.27a}$$

avec

$$\mathbf{X} = \begin{bmatrix} X \\ V \\ W \end{bmatrix} \in \mathbb{R}^{\ell+2m}, \qquad \mathbf{F}(\mathbf{X}\,; \nu) = \begin{bmatrix} \Lambda(X) \\ G(X) - V \\ H(X) - W \\ V \odot W - \nu \mathbf{1} \end{bmatrix} \in \mathbb{R}^{\ell+2m}. \tag{1.27b}$$

La méthode génère alors une suite $\mathbf{X}^k = (X^k, V^k, W^k)$ ainsi qu'une suite auxiliaire $\nu^k > 0$ telles que

$$(X^k,\, V^k,\, W^k) \to (\bar{X},\, G(\bar{X}),\, H(\bar{X})), \qquad \nu^k \to 0,$$

où $\bar{X}$ est un zéro de F. De surcroît, la première suite doit satisfaire la condition de stricte positivité

$$V^k > 0, \qquad W^k > 0,$$

pour tout $k \geqslant 0$.

De ce principe général, plusieurs méthodes peuvent être déduites. La plus simple est celle dite *à un pas* [Algorithme 4.5], dont l'esprit est fidèle à celui des méthodes de régularisation : on fait une itération de Newton à ν^k fixé pour trouver X^{k+1}, puis on met à jour ν^{k+1} "à la louche" selon l'une des règles empiriques (4.77) ou une autre. Une méthode plus sophistiquée, qui comporte deux étapes [Algorithme 4.6], est inspirée de l'algorithme de Mehrotra [88], référence incontournable en optimisation. Dans cet algorithme, le paramètre ν^k est toujours égal à la mesure de centralité $\langle V^k, W^k \rangle / m$ de l'itéré courant, où $\langle \cdot, \cdot \rangle$ désigne le produit scalaire. À la première étape, surnommée *prédicteur*, on fait fi de ν^k et cherche à atteindre immédiatement la cible ultime, qui correspond à $\nu = 0$, en faisant un pas de Newton (4.78) puis en tronquant la direction obtenue pour respecter la positivité. Quelle que soit l'issue de cette tentative audacieuse, un facteur de

recentrage σ^k est évalué par l'heuristique (4.82) afin de viser l'objectif mieux adapté $\nu = \sigma^k \nu^k$ dans la seconde étape, appelée *correcteur*. Ce facteur d'adaptation σ^k est un ingrédient essentiel de l'algorithme. La dernière étape incorpore également une correction du second ordre dans les équations dans le but de gagner en précision et se termine par une autre troncature, toujours en vue de rester dans le domaine strictement intérieur.

1.3 Démarche, contributions et plan du mémoire

Dans cette thèse, nous avons pris le parti de nous focaliser sur le sous-problème de l'*équilibre des phases*, extrait d'un modèle d'écoulement complet comme (1.4). L'avantage d'étudier d'abord ce sous-problème est qu'il est plus petit et qu'il ne dépend pas de $(\boldsymbol{\chi}, t)$. En escamotant ainsi l'écoulement, on peut mieux se concentrer sur la thermodynamique pure. Une fois les difficultés appréhendées et résolues, nous reviendrons bien entendu au modèle d'écoulement complet dans les simulations numériques.

1.3.1 Étude du problème de l'équilibre des phases

Les deux premières sections du chapitre §2 introduisent ce sous-problème de l'équilibre des phases, de manière délibérément indépendante du modèle d'écoulement complet retenu puisqu'il peut y en avoir plusieurs. Montrons ici le lien entre le sous-problème de l'équilibre de phases et le modèle (1.4). Soit

$$\rho = \sum_{\alpha \in \mathscr{P}} \rho_\alpha^\circ \mathsf{S}_\alpha \tag{1.28}$$

la densité totale. Définissons les fractions de phase

$$Y_\beta = \frac{\rho_\beta^\circ \mathsf{S}_\beta}{\rho}, \qquad \beta \in \mathscr{P}, \tag{1.29}$$

ainsi que les compositions

$$c^i = \frac{1}{\rho} \sum_{\beta \in \mathscr{P}} \rho_\beta^\circ \mathsf{S}_\beta \xi_\beta^i, \qquad i \in \mathcal{K}. \tag{1.30}$$

Il est alors facile de voir que ces deux types de fractions sont reliées par la relation bilan

$$c^i = \sum_{\beta \in \mathscr{P}} Y_\beta \xi_\beta^i, \qquad \forall i \in \mathcal{K}. \tag{1.31}$$

D'autre part, à cause de (1.29) et comme $\rho_\beta^\circ / \rho > 0$, la condition de complémentarité (1.4e) équivaut encore à

$$\min\left(Y_\beta,\, 1 - \sum_{j \in \mathcal{K}} \xi_\beta^j\right) = 0, \qquad \forall \beta \in \mathscr{P}. \tag{1.32}$$

Les relations (1.31), (1.32) auxquelles se joignent les relations d'équilibre étendues (1.4c) forment un système qui n'est autre que la formulation unifiée (2.37)–(2.39) du problème de l'équilibre des phases. La scission avec le modèle complet d'écoulement est réalisée en considérant que les compositions $\{c^i\}_{i \in \mathcal{K}}$ ainsi que la pression P sont données.

La section §2.3 regroupe plusieurs résultats originaux concernant la formulation unifiée du problème de l'équilibre des phases. Ces résultats s'expriment le plus naturellement lorsqu'on utilise les fonctions d'énergie de Gibbs, dont le rôle central est ainsi mis en exergue.

- En §2.3.1, nous montrons qu'elle permet de retrouver rigoureusement le critère du plan tangent [Théorème 2.1], certes connu des physiciens mais dont la démonstration dans les ouvrages de thermodynamique suit un cheminement tout à fait différent.

- En §2.3.2, nous mettons en avant un lien fort et jusqu'à présent méconnu entre la formulation unifiée et la minimisation d'une énergie de Gibbs modifiée du mélange, exprimée directement en fonction des fractions étendues [Théorèmes 2.3 et 2.4]. Il n'y a pas équivalence parfaite, mais nous prouvons que la formulation unifiée correspond à un choix pour les fractions des phases absentes parmi une infinité possible de minimiseurs. Ce choix est de surcroît naturel, puisqu'il est obtenu par limite continue de solutions dans lesquelles les phases sont présentes.

- En §2.3.3, nous émettons des hypothèses raisonnables [Hypothèses 2.2] afin d'assurer l'existence et l'unicité des fractions étendues dans deux configurations particulières mais importantes. Elles requièrent notamment la stricte convexité des fonctions de Gibbs et seront indispensables pour la suite des développements théoriques.

À partir de la section §2.4, nous nous restreignons à un mélange diphasique. En §2.4.1, nous définissons deux notions de dégénérescence pour les solutions, à savoir les points de transition et les points azéotropiques, qui seront exclues plus tard des théorèmes. En §2.4.2, nous examinons le cas particulier d'un mélange binaire (à deux composantes), pour lequel nous démontrons l'existence et l'unicité d'une solution pour la formulation unifiée.

1.3.2 Analyse de convexité des lois simples et prolongement des lois cubiques

Le chapitre §3 pousse plus loin l'étude du problème de l'équilibre des phases en prenant en compte l'expression explicite de quelques lois physiques spécifiques habituellement utilisées par la fonction d'énergie de Gibbs. La première section §3.1 s'intéresse à la question de savoir si les Hypothèses 2.2 sont satisfaites pour certaines lois simples. La réponse est positive inconditionnellement pour la loi de Henry [Proposition 3.1], conditionnellement pour les lois de Margules [Proposition 3.2] et Van Laar [Proposition 3.3]. Pour ces dernières, nous déterminons la région dans l'espace des paramètres pour laquelle la fonction de Gibbs associée est strictement convexe.

Les lois d'état cubiques, très prisées par les ingénieurs réservoir pour leur précision, font l'objet de la section §3.2. Comme cela est rappelé en §3.2.1, leur construction passe par une équation du troisième degré dépendant de deux paramètres. Nous examinons plus en profondeur la loi de Van der Waals en §3.2.2 et celle de Peng-Robinson en §3.2.3. Pour chaque loi,

- nous donnons l'expression des coefficients de fugacité pour une loi de mélange générale [Théorèmes 3.1 et 3.4] en supposant que la racine de l'équation cubique correspondant à la phase considérée existe ;

- nous élucidons le comportement de l'équation cubique en fonction de la criticité des paramètres [Théorèmes 3.2 et 3.5], à partir de quoi nous énonçons les règles permettant d'attribuer une phase à une racine [Définitions 3.2 et 3.3] en régime sous-critique ;

- nous identifions dans le plan des paramètres la frontière entre la zone à une racine réelle et celle à trois racines réelles [Théorèmes 3.3 et 3.6], ce qui sera extrêmement utile pour la suite.

Le troisième point est tout à fait nouveau. Le matériel des deux premiers points existe plus ou moins dans les livres de thermodynamique, mais nous en avons cherché des démonstrations plus rigoureuses. Ceci nous a conduit notamment à déterminer la valeur exacte des paramètres critiques de Peng-Robinson, dont la littérature ne donne en général que des approximations décimales.

Vu la complexité des lois cubiques, la question de la stricte convexité des fonctions de Gibbs associées ne sera guère abordée. À la place, nous examinons dans la section §3.3 une question plus urgente et plus vitale concernant la limitation des domaines de définition des fonctions de Gibbs. En effet, comme expliqué rapidement en §1.1.2 et repris pas à pas en §3.3.1, cette particularité des lois d'état cubiques est un handicap sérieux pour la formulation unifiée, car elle est susceptible de mettre en défaut l'existence d'une solution quand l'une des phases est absente.

Nous proposons d'y remédier en prolongeant les fonctions de Gibbs à tout le domaine des fractions par deux méthodes. La première, dite *directe* et détaillée en §3.3.2, est trop intimement liée au cas binaire et se généralise difficilement au cas d'un nombre quelconque d'espèces. La seconde, dite *indirecte* et développée en §3.3.3, manipule les racines au lieu des fractions et s'avère mieux adaptée au cas multicompositionnel. L'idée de base est que quand la cubique n'a qu'une seule racine réelle associée à une certaine phase, on peut utiliser la partie réelle (commune) des deux autres racines complexes (conjuguées) comme "racine" associée à l'autre phase. En envoyant cette valeur dans les formules de la fonction de Gibbs, on obtient un prolongement continu. Cette stratégie, justifiée par des propriétés favorables [Lemmes 3.4 et 3.5], donne d'excellents résultats numériques.

1.3.3 Élaboration de la méthode des points intérieurs non-paramétrique

Les méthodes de régularisation évoquées en §1.2.2 et détaillées en §4.3 sont séduisantes sur le papier et donnent d'ailleurs des résultats acceptables la plupart du temps. Elles ont toutes néanmoins un défaut en commun : il n'y pas de recette miracle pour piloter la suite des paramètres de régularisation ν^k vers 0. Une règle heuristique qui fonctionne bien sur un problème peut échouer piteusement sur un autre. L'utilisateur doit essayer plusieurs suites ν^k avant de savoir laquelle convient le mieux à son problème.

Ce constat nous incite à concevoir en §5.1 une nouvelle méthode, appelée *nonparametric interior-point method* (NPIPM), dans laquelle la mise à jour de ν est "automatique" et couplée avec celle des inconnues $\mathbf{X} = (X, V, W)$. Pour cela, nous devons accomplir une nouvelle "unification", cette fois entre $\mathbf{X}$ et ν. Concrètement, on pose

$$\mathbb{X} = \begin{bmatrix} \mathbf{X} \\ \nu \end{bmatrix}, \qquad \mathbb{F}(\mathbb{X}) = \begin{bmatrix} \mathbf{F}(\mathbf{X}\,;\nu) \\ \frac{1}{2}\|V^-\|^2 + \frac{1}{2}\|W^-\|^2 + \eta\nu + \nu^2 \end{bmatrix}, \tag{1.33a}$$

où $\eta > 0$ est un petit paramètre fixé une fois pour toutes et

$$\|V^-\|^2 = \sum_{\alpha=1}^{m} \min^2(V_\alpha, 0), \qquad \|W^-\|^2 = \sum_{\alpha=1}^{m} \min^2(W_\alpha, 0), \tag{1.33b}$$

et on cherche à résoudre

$$\mathbb{F}(\mathbb{X}) = 0. \tag{1.34}$$

La construction de $\mathbb{F}$ est faite de sorte que tout zéro $\bar{\mathbb{X}} = (\bar{\mathbf{X}}, \bar{\nu})$ de $\mathbb{F}$ tel que $\bar{\nu} > -\eta/2$ vérifie

$$\bar{\nu} = 0, \qquad \mathbf{F}(\bar{\mathbf{X}}\,;0) = 0, \qquad V^- = W^- = 0.$$

Comme expliqué en détail en §5.1.1, la raison d'être du terme linéaire $\eta\nu$ dans la dernière équation est d'éviter une racine double en $\bar{\nu} = 0$ et d'assurer ainsi une convergence quadratique.

Puisque $\mathbb{F}$ est différentiable, on peut appliquer la méthode de Newton classique

$$\mathbb{X}^{k+1} = \mathbb{X}^k - [\nabla \mathbb{F}(\mathbb{X}^k)]^{-1} \mathbb{F}(\mathbb{X}^k), \tag{1.35}$$

combinée avec une recherche linéaire de type Armijo pour tenter d'assurer une convergence globale [Algorithm 5.1]. La théorie de convergence globale à laquelle nous faisons appel, due à Bonnans [21], est rappelée en §5.1.2. Elle repose de manière essentielle sur l'hypothèse de régularité du zéro, par laquelle on entend que la matrice jacobienne $\nabla \mathbb{F}(\bar{\mathbb{X}})$ est non-singulière.

En application de cette théorie, nous nous attachons en §5.2 à vérifier la régularité des zéros du problème de l'équilibre des phases pour un mélange diphasique compositionnel en formulation unifiée. Notre résultat principal [Théorème 5.3], acquis au prix de laborieuses transformations de déterminants, est que sous l'hypothèse de stricte convexité des fonctions de Gibbs, toute solution du problème est régulière à l'exception des points transitionnels et des points azéotropiques. En marge de la preuve générale en §5.2.1, nous indiquons également une démonstration plus courte pour le cas des lois de Henry en §5.2.2.

1.3.4 Comparaison numérique de plusieurs méthodes sur plusieurs modèles

Le chapitre §6 relate enfin les expériences numériques que nous avons menées sur plusieurs modèles physiques avec conditions de complémentarité en utilisant plusieurs méthodes numériques. Les quatre premiers modèles, traités en §6.1, sont considérés comme "simples" du fait du faible nombre d'équations et d'inconnues. Le premier d'entre eux, en §6.1.1, ne relève pas de la thermodynamique mais de la géologie, et plus exactement de la stratigraphie dont nous avons eu un aperçu en §4.4. Les deux suivants, en §6.1.2–§6.1.3, correspondent au modèle (2.77) pour l'équilibre d'un mélange diphasique respectivement binaire (à deux constituants) et ternaire (à trois constituants). Le dernier de la série des modèles "simples", en §6.1.4, est une variante du modèle binaire avec une évolution temporelle imposée à la composition c et où la valeur du pas de temps Δt influe sur la raideur du système à résoudre. C'est un avant-goût du modèle "complet" (6.42).

En ce qui concerne les méthodes numériques, nous procédons de la manière suivante. En partant du premier modèle, nous essayons toutes les méthodes envisagées. Si une méthode ne donne pas de résultat satisfaisant, elle est éliminée de la liste. Nous passons alors au modèle suivant et essayons les méthodes qui restent. Ainsi de suite...

La deuxième section §6.2 porte sur un modèle d'écoulement "complet" dont nous présentons seulement les équations algébriques et aux dérivées partielles au niveau continu, la discrétisation spatiale par un schéma de volumes finis étant longue et pouvant être consultée par ailleurs. Ce modèle est loin d'être le plus réaliste : il manque plusieurs effets physiques importants comme la gravité et la capillarité (différence de pression entre les deux phases). Néanmoins, il est suffisamment complexe pour créer des difficultés à Newton-min et NPIPM. Contrairement aux cas simples où NPIPM surpasse sans conteste Newton-min, ici la situation est plus délicate. Il y a certes quelques scénarios pour lesquels NPIPM converge sans que Newton-min ne le fasse. Mais en général, l'amélioration apportée par NPIPM est faible et parfois NPIPM peut faire légèrement moins bien en nombre d'itérations. Nous avançons quelques explications à ces observations en considérant la spécificité des problèmes d'évolution au regard de l'initialisation des algorithmes.

Part I

Thermodynamic setting

Chapter 2

Phase equilibrium for multicomponent mixtures

Contents

Nous exposons le problème de l'équilibre des phases pour un mélange polyphasique compositionnel, dont la résolution numérique constitue la motivation de cette thèse. Par rapport aux présentations usuelles en thermodynamique, la nôtre se focalise sur les vraies inconnues que sont les fractions de phase et d'espèce, omettant souvent d'indiquer les grandeurs fixées que sont la pression et la température.

Après rappel de quelques notions préliminaires en §2.1, nous introduisons en §2.2 deux formulations pour ce problème. La première, dite formulation naturelle*, fait appel à une gestion dynamique des variables. La seconde, appelée* formulation unifiée*, permet de travailler avec un jeu fixe d'inconnues et d'équations au moyen des conditions de complémentarité. Nous établissons en §2.3 quelques propriétés originales de la formulation unifiée, en particulier sa relation avec la minimisation de l'énergie de Gibbs.*

En nous restreignant ensuite au cadre diphasique en §2.4, nous donnons la forme définitive au modèle à résoudre numériquement dans cette thèse. Nous examinons le cas particulier des mélanges à deux constituants, pour lesquels nous mettons en avant quelques propriétés supplémentaires, notamment la construction géométrique par Gibbs de la solution exacte, que nous redémontrons rigoureusement à partir de la formulation unifiée.

2.1 Preliminary notions

We start by reviewing some prerequisites on the thermodynamics of multiphase multicomponent mixtures. This also gives us the opportunity to introduce the mathematical notations that will be used throughout this manuscript.

2.1.1 Material balance

2.1.1.1 Species, phases and context

A multicomponent mixture is a physical system consisting of several chemically distinct components or *species*. Such a system arises in many real-life applications such as transport of hydrocarbons or subsurface energy storage, where the components may be, for instance, hydrogen (H_2), water (H_2O), carbon dioxide (CO_2), methane (CH_4)... To think of the mixture in a more abstract way, let us designate by

$$\mathcal{K} = \big\{\mathrm{I}, \mathrm{II}, \ldots, \mathrm{K}\big\}, \qquad \mathrm{K} \geqslant 2, \tag{2.1}$$

the set of its species, labeled by Roman numerals. The total number of components $\mathrm{K} = |\mathcal{K}|$ usually ranges from tens to hundreds, so that sometimes partial aggregation or lumping is necessary to reduce complexity.

Each component $i \in \mathcal{K}$ may be present under one or many *phases*, hence the denomination of multiphase multicomponent mixtures. Intuitively, a phase is more or less a state of matter, e.g., gas (G), liquid (L), oil (O), solid (S)... However, this notion is more subtle, especially at high pressure [39]. Again, to lay down an abstract framework, let us consider

$$\mathscr{P} = \big\{1, 2, \ldots, P\big\}, \qquad P \geqslant 2, \tag{2.2}$$

the set of all virtually possible phases, labeled by Arabic numerals. The choice of $\mathscr{P}$ within a model is the (difficult) task of physicists: P should be large enough to take into account the appearance of new phases in models with time evolution, but not too large for computations to remain feasible. Most commonly, the maximum number of possible phases $P = |\mathscr{P}|$ is about 3 in IFPEN's simulations.

Let $n_\alpha^i \geqslant 0$ be the number of moles[1] of component $i \in \mathcal{K}$ existing under phase $\alpha \in \mathscr{P}$. Then,

$$n_\alpha = \sum_{i \in \mathcal{K}} n_\alpha^i \tag{2.3}$$

is the number moles of matter within phase α. If $n_\alpha = 0$, the phase α is said to be *absent*. Indeed, it does not exist. If $n_\alpha > 0$, the phase α is said to be *present*. The subset of present phases, namely,

$$\Gamma = \{\alpha \in \mathscr{P} \mid \ n_\alpha > 0\} \subset \mathscr{P} \tag{2.4}$$

is referred to as the *context*. Since the statement of the phase equilibrium problem in this chapter is static and local, the context seems to share the same features. Nevertheless, in flow models where the n_α^i's vary in time and space, the context also depends on time and space.

[1] A mole of substance is defined as exactly $6.02214076 \cdot 10^{23}$ particles (atoms, molecules, ions, electrons), the latter number being the Avogadro constant.

2.1.1.2 Phasic, partial and global fractions

By summing (2.3) over the phases, we obtain

$$n = \sum_{\alpha\in\mathscr{P}} n_\alpha = \sum_{\alpha\in\mathscr{P}} \sum_{i\in\mathcal{K}} n_\alpha^i \tag{2.5}$$

as the total number of moles of matter in the mixture. Naturally, it is assumed that $n > 0$; otherwise, the system is empty. This allows us to define the *phasic* fraction

$$Y_\alpha = \frac{n_\alpha}{n} = \frac{n_\alpha}{\sum_{\beta\in\mathscr{P}} n_\beta} \in [0,1] \tag{2.6}$$

of phase $\alpha \in \mathscr{P}$. Thus, the phase can be characterized as absent or present depending on whether $Y_\alpha = 0$ or $Y_\alpha > 0$. Of course,

$$\sum_{\alpha\in\Gamma} Y_\alpha = 1. \tag{2.7}$$

If a phase α is present, that is, $n_\alpha > 0$ or equivalently $Y_\alpha > 0$, then it is possible to define

$$x_\alpha^i = \frac{n_\alpha^i}{n_\alpha} = \frac{n_\alpha^i}{\sum_{j\in\mathcal{K}} n_\alpha^j} \in [0,1] \tag{2.8}$$

as the *partial* fraction of component $i \in \mathcal{K}$ within phase $\alpha \in \Gamma$. From definition (2.8), it follows that

$$\sum_{i\in\mathcal{K}} x_\alpha^i = 1 \tag{2.9}$$

for all $\alpha \in \Gamma$. Note that this notion does not make sense for an absent phase $\alpha \notin \Gamma$, at least from a quick inspection of (2.8), which gives rise to the indeterminate form 0/0. Surprisingly, the unified formulation of §2.2.2 will enable us to assign a well-defined value to x_α^i even for a vanishing phase, subject to some technical conditions. This will be done in §2.3.3.

By reversing the order of summation in (2.5), we have

$$n = \sum_{i\in\mathcal{K}} \sum_{\alpha\in\mathscr{P}} n_\alpha^i = \sum_{i\in\mathcal{K}} n^i, \tag{2.10}$$

where the newly defined quantity

$$n^i = \sum_{\alpha\in\mathscr{P}} n_\alpha^i \tag{2.11}$$

represents the total number of moles of component i across all phases. Then,

$$c^i = \frac{n^i}{n} = \frac{n^i}{\sum_{j\in\mathcal{K}} n^j} \in [0,1] \tag{2.12}$$

is called the *global* fraction of component i inside the mixture. Needless to say,

$$\sum_{i\in\mathcal{K}} c^i = 1. \tag{2.13}$$

By dividing (2.11) by n, restricting summation in the right-hand sides to present phases and artificially inserting n_α in each summand, we end up with

$$c^i = \sum_{\alpha\in\Gamma} Y_\alpha x_\alpha^i. \tag{2.14}$$

Given the context Γ, the phasic fractions $\{Y_\alpha\}_{\alpha\in\Gamma}$ and the partial fractions $\{x_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\mathscr{P}}$, it is straightforward to calculate the global composition $\{c^i\}_{i\in\mathcal{K}}$ by (2.14). The phase equilibrium problem takes exactly the opposite direction: given the global composition $\{c^i\}_{i\in\mathcal{K}}$ satisfying (2.13), is it possible to find the context Γ, the phasic fractions $\{Y_\alpha\}_{\alpha\in\Gamma}$ and the partial fractions $\{x_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\mathscr{P}}$ satisfying (2.7), (2.9) and (2.14) beside positivity? Obviously, we do not have enough equations yet. The missing ones are addressed below.

2.1.2 Chemical equilibrium

2.1.2.1 Gibbs energy, chemical potential and Gibbs-Duhem conditions

The behavior of each phase $\alpha \in \mathscr{P}$ is governed by a single fundamental function

$$\mathsf{G}_\alpha \ : \ \mathbb{R}_+^{\mathrm{K}} \to \mathbb{R}$$

known as the *Gibbs free energy* of the phase. The Gibbs energy is the Legendre-conjugate of the internal energy with respect to volume and entropy [115], which makes it a function of the number of moles, the pressure and the temperature. Therefore, it is well suited to the study of systems at fixed pressure and temperature[2]. We require G_α to be as smooth as necessary.

With respect to the number of moles, this function must be *extensive*. This actually means that it must be homogeneous of degree 1, i.e.,

$$\mathsf{G}_\alpha(\lambda n_\alpha^{\mathrm{I}},\, \lambda n_\alpha^{\mathrm{II}},\, \ldots,\, \lambda n_\alpha^{\mathrm{K}}) = \lambda \mathsf{G}_\alpha(n_\alpha^{\mathrm{I}},\, n_\alpha^{\mathrm{II}},\, \ldots,\, n_\alpha^{\mathrm{K}}), \qquad \text{for all } \ \lambda > 0. \tag{2.15}$$

Then, Euler's homogeneous function theorem —derived by differentiating (2.15) with respect to λ and by putting $\lambda = 1$ in the result— asserts that

$$\mathsf{G}_\alpha(n_\alpha^{\mathrm{I}},\, n_\alpha^{\mathrm{II}},\, \ldots,\, n_\alpha^{\mathrm{K}}) = \sum_{j\in\mathcal{K}} n_\alpha^j \, \frac{\partial \mathsf{G}_\alpha}{\partial n_\alpha^j}(n_\alpha^{\mathrm{I}},\, n_\alpha^{\mathrm{II}},\, \ldots,\, n_\alpha^{\mathrm{K}}). \tag{2.16}$$

Furthermore, the functions

$$\mu_\alpha^j = \frac{\partial \mathsf{G}_\alpha}{\partial n_\alpha^j} \tag{2.17}$$

can be shown to be homogeneous of degree 0, i.e.,

$$\mu_\alpha^j(\lambda n_\alpha^{\mathrm{I}},\, \lambda n_\alpha^{\mathrm{II}},\, \ldots,\, \lambda n_\alpha^{\mathrm{K}}) = \mu_\alpha^j(n_\alpha^{\mathrm{I}},\, n_\alpha^{\mathrm{II}},\, \ldots,\, n_\alpha^{\mathrm{K}}), \qquad \text{for all } \ \lambda > 0. \tag{2.18}$$

Each function μ_α^j is the *chemical potential* of component $j \in \mathcal{K}$ within phase $\alpha \in \mathscr{P}$. Note that G_α and the μ_α^j's are defined for all phases, present or absent, since here the n_α^i's are dummy arguments.

Differentiating the Euler relation

$$\mathsf{G}_\alpha = \sum_{i\in\mathcal{K}} n_\alpha^i \, \mu_\alpha^i \tag{2.19}$$

with respect to n_α^j yields

$$\mu_\alpha^j = \sum_{i\in\mathcal{K}} \delta_{j,i} \mu_\alpha^i + \sum_{i\in\mathcal{K}} n_\alpha^i \, \frac{\partial \mu_\alpha^i}{\partial n_\alpha^j},$$

[2] This is also why we shall not explicitly write down the dependency of the Gibbs energy with respect to the pressure P and the temperature T.

from which it is deduced that

$$\sum_{i\in\mathcal{K}} n_\alpha^i \,\frac{\partial \mu_\alpha^i}{\partial n_\alpha^j} = 0, \qquad \text{for all } \; j \in \mathcal{K}. \tag{2.20}$$

Identity (2.20), called the *Gibbs-Duhem condition*, can be regarded as a compatibility requirement to be prescribed on K given 0-homogeneous functions μ_α^i so that they can correctly play the role of chemical potentials for a *bona fide* Gibbs energy function.

We now wish to express (2.19)–(2.20) in terms of the partial fractions x_α^i defined in (2.8). Again, since we are interested in functional relationships, we can put aside our concerns about an absent phase and carry out calculations for all phases $\alpha \in \mathscr{P}$. Plugging

$$\lambda = \frac{1}{n_\alpha}$$

into (2.15) and (2.18) results in

$$\mathsf{G}_\alpha(n_\alpha^{\mathrm{I}}, n_\alpha^{\mathrm{II}}, \ldots, n_\alpha^{\mathrm{K}}) = n_\alpha \mathsf{G}_\alpha(x_\alpha^{\mathrm{I}}, x_\alpha^{\mathrm{II}}, \ldots, x_\alpha^{\mathrm{K}}), \tag{2.21a}$$

$$\mu_\alpha^i(n_\alpha^{\mathrm{I}}, n_\alpha^{\mathrm{II}}, \ldots, n_\alpha^{\mathrm{K}}) = \quad \mu_\alpha^i(x_\alpha^{\mathrm{I}}, x_\alpha^{\mathrm{II}}, \ldots, x_\alpha^{\mathrm{K}}). \tag{2.21b}$$

Because of (2.9), the quantities $x_\alpha^{\mathrm{I}}, x_\alpha^{\mathrm{II}}, \ldots, x_\alpha^{\mathrm{K}}$ are not independent. We select the first $\mathrm{K}-1$ partial fractions

$$\boldsymbol{x}_\alpha = (x_\alpha^{\mathrm{I}}, \ldots, x_\alpha^{\mathrm{K}-1}) \; \in \overline{\Omega} \subset \mathbb{R}^{\mathrm{K}-1}$$

as independent variables. Whenever a x_α^{K} turns up in any formula, it should be interpreted as

$$x_\alpha^{\mathrm{K}} = 1 - x_\alpha^{\mathrm{I}} - \ldots - x_\alpha^{\mathrm{K}-1}.$$

The domain of $\boldsymbol{x}_\alpha$ is the closure of

$$\Omega = \big\{\boldsymbol{x} = (x^{\mathrm{I}}, \ldots, x^{\mathrm{K}-1}) \in \mathbb{R}^{\mathrm{K}-1} \mid x^{\mathrm{I}} > 0, \ldots, x^{\mathrm{K}-1} > 0, \, 1 - x^{\mathrm{I}} - \ldots - x^{\mathrm{K}-1} > 0\big\}, \tag{2.22a}$$

namely,

$$\overline{\Omega} = \big\{\boldsymbol{x} = (x^{\mathrm{I}}, \ldots, x^{\mathrm{K}-1}) \in \mathbb{R}^{\mathrm{K}-1} \mid x^{\mathrm{I}} \geqslant 0, \ldots, x^{\mathrm{K}-1} \geqslant 0, \, 1 - x^{\mathrm{I}} - \ldots - x^{\mathrm{K}-1} \geqslant 0\big\}. \tag{2.22b}$$

Although this choice somehow breaks the symmetry, it is commonly resorted to in practice. Introduce for each phase α the *intensive* or *molar* Gibbs energy and chemical potentials

$$g_\alpha \; : \; \overline{\Omega} \to \mathbb{R}, \qquad \mu_\alpha^i \; : \; \Omega \to \mathbb{R},$$

defined as

$$g_\alpha(\boldsymbol{x}_\alpha) = \mathsf{G}_\alpha(x_\alpha^{\mathrm{I}}, x_\alpha^{\mathrm{II}}, \ldots, x_\alpha^{\mathrm{K}}), \tag{2.23a}$$

$$\mu_\alpha^i(\boldsymbol{x}_\alpha) = \mu_\alpha^i(x_\alpha^{\mathrm{I}}, x_\alpha^{\mathrm{II}}, \ldots, x_\alpha^{\mathrm{K}}), \tag{2.23b}$$

In (2.23b), we have slightly abused notation by reusing the same symbol μ_α^i in the left-hand side. We require g_α and μ_α^i to be as smooth as necessary over Ω. Moreover, g_α is assumed to be extendable by continuity to the closure $\overline{\Omega}$, but not the μ_α^i's which usually blow up on $\partial\Omega$.

The following statement summarizes some identities between g_α and μ_α^i that would be most helpful in the sequel.

Lemma 2.1 (Connection between molar Gibbs energy and chemical potentials)**.** *For all* $\boldsymbol{x}_\alpha \in \Omega$*:*

1. *The molar Gibbs energy is related to the chemical potentials by*

$$g_\alpha(\boldsymbol{x}_\alpha) = \sum_{i=\mathrm{I}}^{\mathrm{K}} x_\alpha^i \, \mu_\alpha^i(\boldsymbol{x}_\alpha). \tag{2.24a}$$

2. *Each chemical potential can be deduced from the molar Gibbs energy by*

$$\mu_\alpha^j(\boldsymbol{x}_\alpha) = g_\alpha(\boldsymbol{x}_\alpha) + \nabla_{\boldsymbol{x}_\alpha} g_\alpha(\boldsymbol{x}_\alpha) \cdot (\boldsymbol{\delta}^j - \boldsymbol{x}_\alpha), \qquad \text{for all } \ j \in \mathcal{K}, \tag{2.24b}$$

   where the Kronecker vector $\boldsymbol{\delta}^i$ *is defined as* $\boldsymbol{\delta}^j = (\delta_{j,1}, \delta_{j,2}, \ldots, \delta_{j,\mathrm{K}-1}) \in \mathbb{R}^{\mathrm{K}-1}$.

3. *The gradient of the molar Gibbs energy is given from the chemical potentials by*

$$\frac{\partial g_\alpha}{\partial x_\alpha^j}(\boldsymbol{x}_\alpha) = \mu_\alpha^j(\boldsymbol{x}_\alpha) - \mu_\alpha^{\mathrm{K}}(\boldsymbol{x}_\alpha), \qquad \text{for all } \ j \in \mathcal{K}\backslash\{\mathrm{K}\}. \tag{2.24c}$$

4. *The gradients of the chemical potentials satisfy the* Gibbs-Duhem condition

$$\sum_{i=\mathrm{I}}^{\mathrm{K}} x_\alpha^i \, \nabla_{\boldsymbol{x}_\alpha} \mu_\alpha^i(\boldsymbol{x}_\alpha) = 0. \tag{2.24d}$$

Chứng minh. To prove (2.24a), we just have to divide (2.19) by n_α and to make use of (2.23). From definition (2.17), we have

$$\mu_\alpha^j(\boldsymbol{x}_\alpha) = \frac{\partial}{\partial n_\alpha^j}(n_\alpha g_\alpha(\boldsymbol{x}_\alpha)) = g_\alpha(\boldsymbol{x}_\alpha) + n_\alpha \sum_{i=\mathrm{I}}^{\mathrm{K}-1} \frac{\partial g_\alpha}{\partial x_\alpha^i}(\boldsymbol{x}_\alpha) \frac{\partial x_\alpha^i}{\partial n_\alpha^j}$$

for $j \in \mathcal{K}$. But

$$\frac{\partial x_\alpha^i}{\partial n_\alpha^j} = \frac{\partial}{\partial n_\alpha^j}\left(\frac{n_\alpha^i}{n_\alpha}\right) = \frac{\delta_{i,j} n_\alpha - n_\alpha^i}{(n_\alpha)^2} = \frac{\delta_{j,i} - x_\alpha^i}{n_\alpha}.$$

Plugging this into the previous equation yields

$$\mu_\alpha^j(\boldsymbol{x}_\alpha) = g_\alpha(\boldsymbol{x}_\alpha) + \sum_{i=\mathrm{I}}^{\mathrm{K}-1} (\delta_{j,i} - x_\alpha^i) \frac{\partial g_\alpha}{\partial x_\alpha^i}(\boldsymbol{x}_\alpha),$$

of which (2.24b) is just a condensed vector form. Let us now subtract the last potential

$$\mu_\alpha^{\mathrm{K}}(\boldsymbol{x}_\alpha) = g_\alpha(\boldsymbol{x}_\alpha) + \nabla g_\alpha(\boldsymbol{x}_\alpha) \cdot (\boldsymbol{\delta}^{\mathrm{K}} - \boldsymbol{x}_\alpha)$$

from each μ_α^j, $j \in \mathcal{K}\backslash\{\mathrm{K}\}$, given by (2.24b). This cancels out $g_\alpha(\boldsymbol{x}_\alpha)$ and $\boldsymbol{x}_\alpha$. Since $\boldsymbol{\delta}^{\mathrm{K}} = (0, 0, \ldots, 0)$, we are left with (2.24c). To derive the Gibbs-Duhem condition (2.24d), we start from (2.24a) and differentiate both sides with respect to x_α^j, $j \in \mathcal{K}\backslash\{\mathrm{K}\}$. This leads to

$$\frac{\partial g_\alpha}{\partial x_\alpha^j} = \sum_{i=\mathrm{I}}^{\mathrm{K}-1} \left(\delta_{j,i} \mu_\alpha^i + x_\alpha^i \frac{\partial \mu_\alpha^i}{\partial x_\alpha^j}\right) - \mu_\alpha^{\mathrm{K}} + x_\alpha^{\mathrm{K}} \frac{\partial \mu_\alpha^{\mathrm{K}}}{\partial x_\alpha^j},$$

the minus sign in the right-hand side being due to $\partial x_\alpha^{\mathrm{K}} / \partial x_\alpha^j = -1$. This can rearranged as

$$\frac{\partial g_\alpha}{\partial x_\alpha^j} = \mu_\alpha^j - \mu_\alpha^{\mathrm{K}} + \sum_{i=\mathrm{I}}^{\mathrm{K}} x_\alpha^i \frac{\partial \mu_\alpha^i}{\partial x_\alpha^j}.$$

By virtue of (2.24c), the sum in the right-hand side above must vanish. In other words,

$$\sum_{i=\mathrm{I}}^{\mathrm{K}} x_\alpha^i \, \frac{\partial \mu_\alpha^i}{\partial x_\alpha^j}(\boldsymbol{x}_\alpha) = 0, \qquad \text{for all } \; j \in \{\mathrm{I}, \ldots, \mathrm{K}-1\}, \tag{2.25}$$

which is the component-wise version of (2.24d). □

2.1.2.2 Equilibrium conditions, fugacity and fugacity coefficient

In a multicomponent mixture without any chemical reaction (also called *non-reactive*), the presence of two phases $(\alpha, \beta) \in \Gamma \times \Gamma$ implies that some equilibrium conditions must be achieved. According to thermodynamics, these conditions are the equalities across the two phases of pressure, temperature, and the chemical potentials corresponding to each component $i \in \mathcal{K}$. In other words,

$$\mu_\alpha^i(\boldsymbol{x}_\alpha) = \mu_\beta^i(\boldsymbol{x}_\beta), \qquad \text{for all } \; (i, \alpha, \beta) \in \mathcal{K} \times \Gamma \times \Gamma. \tag{2.26}$$

These are the missing equations for the phase equilibrium problem. Since pressure and temperature are identical across the phases α and β, we can keep omitting them as arguments of the μ^i's in (2.26).

For a solid phase, μ_α^i is a constant. For fluid phases such as gas, liquid and oil, the chemical potential takes the form

$$\mu_\alpha^i(\boldsymbol{x}_\alpha) = \ln(x_\alpha^i \Phi_\alpha^i(\boldsymbol{x}_\alpha)), \tag{2.27}$$

in which Φ_α^i is called the *fugacity coefficient* of component i in phase α. Note, however, that it depends on the partial concentrations of the other components as well. As for the quantity

$$f_\alpha^i(\boldsymbol{x}_\alpha) = x_\alpha^i \Phi_\alpha^i(\boldsymbol{x}_\alpha), \tag{2.28}$$

it is known as the *fugacity* of component i in phase α. The equality of chemical potentials (2.26) is then equivalent to that of fugacities

$$x_\alpha^i \Phi_\alpha^i(\boldsymbol{x}_\alpha) = x_\beta^i \Phi_\beta^i(\boldsymbol{x}_\beta), \qquad \text{for all } \; (i, \alpha, \beta) \in \mathcal{K} \times \Gamma \times \Gamma. \tag{2.29}$$

In practice, the fugacity coefficients Φ_α^i are given empirically or inferred from an equation of state. This will be elaborated on in chapter §3.

Remark 2.1. In physics textbooks, chemical potentials and fugacities are defined as

$$\widehat{\mu}_\alpha^i(\boldsymbol{x}_\alpha, \mathsf{P}, \mathsf{T}) = \widehat{\mu}_\bullet^i(\mathsf{P}, \mathsf{T}) + \mathsf{R}\mathsf{T} \ln(x_\alpha^i \Phi_\alpha^i(\boldsymbol{x}_\alpha, \mathsf{P}, \mathsf{T})), \tag{2.30a}$$

$$\widehat{f}_\alpha^i(\boldsymbol{x}_\alpha, \mathsf{P}, \mathsf{T}) = x_\alpha^i \Phi_\alpha^i(\boldsymbol{x}_\alpha, \mathsf{P}, \mathsf{T})\mathsf{P}, \tag{2.30b}$$

where P is the pressure, T the temperature, R the universal gas constant and $\mu_\bullet^i(\mathsf{P}, \mathsf{T})$ a reference ideal value. Since P and T are equal across the phases, they drop out from the equality of chemical potentials and we have the equivalence

$$\widehat{\mu}_\alpha^i(\boldsymbol{x}_\alpha, \mathsf{P}, \mathsf{T}) = \widehat{\mu}_\beta^i(\boldsymbol{x}_\beta, \mathsf{P}, \mathsf{T}) \Leftrightarrow \; \mu_\alpha^i(\boldsymbol{x}_\alpha) = \mu_\beta^i(\boldsymbol{x}_\beta).$$

The form (2.27) has the advantage of highlighting the influence of partial fractions at fixed (P, T). Opting for (2.27)–(2.28) instead of keeping (2.30) amounts to working with the molar Gibbs energy function g_α instead of

$$\widehat{g}_\alpha(\boldsymbol{x}_\alpha, \mathsf{P}, \mathsf{T}) = \sum_{i \in \mathcal{K}} \widehat{\mu}_\bullet^i(\mathsf{P}, \mathsf{T}) x_\alpha^i + \mathsf{R}\mathsf{T} g_\alpha(\boldsymbol{x}_\alpha).$$

The two functions differ from each other by an additive affine function and a multiplicative constant. □

Substituting the form (2.27) into (2.24a), we obtain

$$g_\alpha(\boldsymbol{x}_\alpha) = \sum_{i=\mathrm{I}}^{\mathrm{K}} x_\alpha^i \ln x_\alpha^i + \sum_{i=\mathrm{I}}^{\mathrm{K}} x_\alpha^i \ln \Phi_\alpha^i(\boldsymbol{x}_\alpha) \tag{2.31}$$

The first sum in the right-hand side, $\sum_{j=\mathrm{I}}^{\mathrm{K}} x_\alpha^j \ln x_\alpha^j$, is called the *ideal* part. The second sum, denoted by

$$\Psi_\alpha(\boldsymbol{x}_\alpha) = \sum_{i=\mathrm{I}}^{\mathrm{K}} x_\alpha^i \ln \Phi_\alpha^i(\boldsymbol{x}_\alpha), \tag{2.32}$$

is called the *excess* part or the *excess Gibbs energy*. In this perspective, a fluid phase α is assimilated to a "perturbation" of the ideal gas. Whenever we want to modify the Gibbs function, we should act only on the excess part. We shall adopt this point of view in chapter §3.

Owing to the regularity assumptions made on g_α and μ_α^i, the functions

$$\Psi_\alpha \ : \ \overline{\Omega} \to \mathbb{R}, \qquad \ln \Phi_\alpha^i \ : \ \Omega \to \mathbb{R},$$

are also as smooth as necessary, with Ψ_α extendable by continuity to $\overline{\Omega}$ but not the $\ln \Phi_\alpha^i$'s. The very useful relations between Ψ_α and $\ln \Phi_\alpha^i$ are similar to those between g_α and μ_α^i.

Lemma 2.2 (Connection between molar excess Gibbs energy and logarithm of fugacity coefficients)**.** *For all $\boldsymbol{x}_\alpha \in \Omega$:*

1. *Each fugacity coefficient can be deduced from the excess Gibbs energy by*

$$\ln \Phi_\alpha^j(\boldsymbol{x}_\alpha) = \Psi_\alpha(\boldsymbol{x}_\alpha) + \nabla_{\boldsymbol{x}_\alpha} \Psi_\alpha(\boldsymbol{x}_\alpha) \cdot (\boldsymbol{\delta}^j - \boldsymbol{x}_\alpha), \qquad \textit{for all } \ j \in \mathcal{K}, \tag{2.33a}$$

   where the Kronecker vector $\boldsymbol{\delta}^i$ is defined as $\boldsymbol{\delta}^j = (\delta_{j,1}, \delta_{j,2}, \ldots, \delta_{j,\mathrm{K}-1}) \in \mathbb{R}^{\mathrm{K}-1}$.

2. *The gradient of the excess Gibbs energy is given from the fugacity coefficents by*

$$\frac{\partial \Psi_\alpha}{\partial x_\alpha^j}(\boldsymbol{x}_\alpha) = \ln \Phi_\alpha^j(\boldsymbol{x}_\alpha) - \ln \Phi_\alpha^{\mathrm{K}}(\boldsymbol{x}_\alpha), \qquad \textit{for all } \ j \in \mathcal{K}\backslash\{\mathrm{K}\}. \tag{2.33b}$$

3. *The gradients of the fugacity coefficents satisfy the* Gibbs-Duhem condition

$$\sum_{i=\mathrm{I}}^{\mathrm{K}} x_\alpha^i \, \nabla_{\boldsymbol{x}_\alpha} \{\ln \Phi_\alpha^i\}(\boldsymbol{x}_\alpha) = 0. \tag{2.33c}$$

Chứng minh. The proof is straightforward. For each identity from Lemma 2.1, we just have to separate the ideal part from the excess part. The ideal part vanishes trivially. □

A given family of positive real-valued functions $\{\Phi_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\mathscr{P}}$ is said to be *admissible* if, for each $\alpha \in \mathscr{P}$, there exists a Gibbs energy function g_α such that they are the fugacity coefficients. This implies, in particular, that the functions Φ_α^i satisfy the Gibbs-Duhem condition (2.33c).

2.2 Two mathematical formulations

Equipped with the preliminary notions and notations of §2.1, we are now in a position to rigorously state the phase equilibrium problem in two different ways: the "traditional" one and the "modern" one.

2.2.1 Variable-switching formulation

Let us write down a first formulation before commenting on it.

GIVEN

$$\mathcal{K},\ \mathscr{P},\ \{\Phi_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\mathscr{P}} \text{ admissible},$$
$$\{c^i\}_{i\in\mathcal{K}} \in [0,1] \ \text{ subject to } \ \textstyle\sum_{i\in\mathcal{K}} c^i = 1,$$

FIND

$$\Gamma \subset \mathscr{P},\ \{Y_\alpha\}_{\alpha\in\Gamma} \in (0,1],\ \{x_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\Gamma} \in [0,1]$$

so as to satisfy

- the material balances

$$\sum_{\beta\in\Gamma} Y_\beta - 1 = 0, \tag{2.34a}$$

$$\sum_{j\in\mathcal{K}} x_\alpha^j - 1 = 0, \qquad \forall \alpha \in \Gamma, \tag{2.34b}$$

$$\sum_{\beta\in\Gamma} Y_\beta x_\beta^i - c^i = 0, \qquad \forall i \in \mathcal{K}; \tag{2.34c}$$

- the fugacity equalities

$$x_\alpha^i \Phi_\alpha^i(\boldsymbol{x}_\alpha) - x_\beta^i \Phi_\beta^i(\boldsymbol{x}_\beta) = 0, \qquad \forall (i,\alpha,\beta) \in \mathcal{K}\times\Gamma\times\Gamma. \tag{2.35}$$

This first formulation has the advantage of being "natural," insofar as it uses the variables that have been introduced so far. It also bears the name of *natural variable* formulation. The price to be paid for naturality is that the context Γ is itself an unknown. To circumvent this major difficulty, we have to start by making an "educated guess" for Γ. At every fixed Γ, we attempt to solve the algebraic equations (2.34)–(2.35): this is what physicists call a *flash* —or a (P,T)-flash to be more accurate in our case. After exiting the flash, we check the positivity of Y_α and the non-negativity of x_α^i, for $\alpha \in \Gamma$. Should one of these fractions have the wrong sign, we must update Γ in some "smart" way and go for another flash! The number of unknowns and equations for a flash (2.34)–(2.35), as well as their significance, strongly depend on the assumption currently made about the context Γ. Understandably, this approach is also qualified as the *variable-switching* formulation.

Remark 2.2. Another reason for calling it this way is that in most multiphase multicomponent flow models of interest, there are many (coupled) equilibrium problems to be solved: one per cell and per time-step. Since even the correct context changes in space and in time, the size and the structure of the global system to be solved at each time-step keeps evolving. The choice of relevant unknowns and equations then turns out to be delicate. To this end, Coats [30] advocated a set of "natural" variables for some multiphase flow models in porous media. But the heart of Coats' strategy, when boiled down to a single phase equilibrium problem, is exactly what we described above. □

At first sight, there seems to be a lot redundancy in (2.34)–(2.35). A natural question to ask is how many independent equations we do have for a given Γ, and whether or not this number is equal to that of the unknowns in the same context.

Proposition 2.1. *For a fixed context* $\Gamma \in \mathscr{P}$*, system* (2.34)–(2.35) *contains* $(\mathrm{K}+1)\gamma$ *unknowns and* $(\mathrm{K}+1)\gamma$ a priori *independent equations, where* $\mathrm{K} = |\mathcal{K}|$ *and* $\gamma = |\Gamma|$.

Chứng minh. There are γ unknowns $\{Y_\alpha\}_{\alpha\in\Gamma}$ and $\mathrm{K}\gamma$ unknowns $\{x_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\Gamma}$. Hence, the number of unknowns is $\gamma + \mathrm{K}\gamma = (\mathrm{K}+1)\gamma$.

It can be observed that by summing (2.34c) over $i \in \mathcal{K}$, permuting the order of the double sum and invoking (2.34b), we obtain (2.34a) thanks to the assumption $\sum_{i\in\mathcal{K}} c^i = 1$. Thus, equation (2.34a) can be obtained from the remaining ones and should be left out of the system. To eliminate redundancy in the fugacity equalities, we fix a phase $\beta \in \Gamma$ and require (2.35) to hold for all $\alpha \in \Gamma\backslash\{\beta\}$. The resulting system

$$\sum_{j\in\mathcal{K}} x_\alpha^j - 1 = 0, \qquad \forall \alpha \in \Gamma, \tag{2.36a}$$

$$\sum_{\beta\in\Gamma} Y_\beta x_\beta^i - c^i = 0, \qquad \forall i \in \mathcal{K}; \tag{2.36b}$$

$$x_\alpha^i \Phi_\alpha^i(\boldsymbol{x}_\alpha) - x_\beta^i \Phi_\beta^i(\boldsymbol{x}_\beta) = 0, \qquad \forall (i,\alpha) \in \mathcal{K}\times\Gamma\backslash\{\beta\}, \tag{2.36c}$$

plainly contains

$$\gamma + \mathrm{K} + \mathrm{K}(\gamma - 1) = (\mathrm{K}+1)\gamma$$

equations. The independance of the fugacity equalities (2.36c) is a hypothesis to be made on the physical properties of the species. □

There is a vast literature on numerical methods [89–91, 117] for the flash problem (2.36) at fixed Γ. In addition to the classical and generic Newton-Raphson method [6, 115], many special purpose algorithms have been dedicated to the flash problem. These are iterative methods based on various kinds of substitution [61], the most famous of them being the Rachford-Rice substitution [106]. Regarding the update of the context Γ, it is recommended to start with the highest number of possible phases, i.e., $\Gamma = \mathscr{P}$. In case of failure, one of the phases whose phasic fraction has the wrong sign is taken out. The procedure continues until a flash is successful or until there remains a single phase. There exist many variants [23, 75] to this general philosophy.

2.2.2 Unified formulation

To avoid the annoyance of dynamically handling the context, Lauser *et al.* [78] put forward an alternate formulation for the phase equilibrium problem. Let us write it down before commenting on its advantages.

GIVEN

$$\mathcal{K},\ \mathscr{P},\ \{\Phi_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\mathscr{P}} \text{ admissible},$$
$$\{c^i\}_{i\in\mathcal{K}} \in [0,1] \ \text{ subject to } \ \textstyle\sum_{i\in\mathcal{K}} c^i = 1,$$

FIND

$$\{Y_\alpha\}_{\alpha\in\mathscr{P}} \in (0,1],\ \ \{\xi_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\mathscr{P}} \in [0,1]$$

so as to satisfy

- the material balances

$$\sum_{\beta\in\mathscr{P}} Y_\beta - 1 = 0, \tag{2.37a}$$

$$\sum_{\beta\in\mathscr{P}} Y_\beta \xi_\beta^i - c^i = 0, \qquad \forall i \in \mathcal{K}; \tag{2.37b}$$

- the *extended* fugacity equalities

$$\xi_\alpha^i \Phi_\alpha^i(\boldsymbol{x}_\alpha) - \xi_\beta^i \Phi_\beta^i(\boldsymbol{x}_\beta) = 0, \qquad \forall (i,\alpha,\beta) \in \mathcal{K}\times\mathscr{P}\times\mathscr{P}, \tag{2.38a}$$

  where the components of $\boldsymbol{x}_\alpha = (x_\alpha^{\mathrm{I}}, \ldots, x_\alpha^{\mathrm{K}-1}) \in \mathbb{R}^{\mathrm{K}-1}$ are *defined* as

$$x_\alpha^i = \frac{\xi_\alpha^i}{\sum_{j\in\mathcal{K}} \xi_\alpha^j}; \tag{2.38b}$$

- the complementarity conditions

$$\min\left(Y_\beta,\, 1 - \sum_{j\in\mathcal{K}} \xi_\beta^j\right) = 0, \qquad \forall \beta \in \mathscr{P}. \tag{2.39}$$

In this second formulation, the partial fractions x_α^i have been replaced by a new notion, that of *extended* fractions ξ_α^i. The latter are defined over $(i,\alpha) \in \mathcal{K}\times\mathscr{P}$ instead of being restricted to $(i,\alpha) \in \mathcal{K}\times\Gamma$. Although the connection between extended fractions and partial fractions is given by the renormalization (2.38b), the x_α^i's here are merely auxiliary variables that can be eliminated by inserting (2.38b) into (2.38a). The complementarity conditions (2.39) means that, for each $\beta \in \mathscr{P}$,

$$Y_\beta \geqslant 0, \qquad 1 - \sum_{j\in\mathcal{K}} \xi_\beta^j \geqslant 0, \qquad Y_\beta\left(1 - \sum_{j\in\mathcal{K}} \xi_\beta^j\right) = 0. \tag{2.40}$$

As a consequence, for each phase $\beta \in \mathscr{P}$, there are three possible regimes:

▷ $Y_\beta > 0$.
Phase β is present. This implies $\sum_{j\in\mathcal{K}} \xi_\beta^j = 1$ and by virtue of (2.38b), $\xi_\beta^i = x_\beta^i$ for all $i \in \mathcal{K}$. In other words, the extended fractions corresponding to a present phase coincide with the usual partial fractions.

▷ $1 - \sum_{j\in\mathcal{K}} \xi_\beta^j > 0$.
This entails $Y_\beta = 0$, i.e., phase β is absent. Since $\sum_{j\in\mathcal{K}} \xi_\beta^j < 1$, we have $\xi_\beta^i \neq x_\beta^i$. The extended fractions corresponding to an absent phase do not coincide in general with the usual partial fractions (barring from the exception below).

▷ $Y_\beta = 0$ and $1 - \sum_{j\in\mathcal{K}} \xi_\beta^j = 0$.
This happens at the frontier between those solutions for which phase β is present and those solutions for which phase β is absent. At such a *transition* point, phase β starts appearing or disappearing.

It is legitimate to be concerned about the origin of the sign condition $1 - \sum_{j\in\mathcal{K}} \xi_\beta^j \geqslant 0$. After all, it seems to bring a new piece of information that was clearly not included in the variable-switching formulation (2.34)–(2.35). As will be proven in §2.3.1, this condition ensures a stability property known as the *tangent plane criterion* by physicists. It can also be related to the minimization of the Gibbs energy of the mixture, as will be done in §2.3.2.

The ability of the formulation (2.37)–(2.39) to deal with all possible configurations (arising from the presence or the absence of each phase) in the same manner accounts for the name of *unified formulation*. The context Γ no longer appears in the statement of the problem. It can be determined *a posteriori* by collecting those phases α for which $Y_\alpha > 0$. The unified formulation has turned an intricate combinatorial problem into a fixed set of equations and unknowns, with which it is definitely more convenient to work with. Let us clarify the number of unknowns and independent equations of (2.37)–(2.39).

Proposition 2.2. *System* (2.37)–(2.39) *contains* $(\mathrm{K}+1)P$ *unknowns and* $(\mathrm{K}+1)P$ a priori *independent equations, where* $\mathrm{K} = |\mathcal{K}|$ *and* $P = |\mathscr{P}|$.

Chứng minh. There are P unknowns $\{Y_\alpha\}_{\alpha\in\mathscr{P}}$ and $\mathrm{K}P$ unknowns $\{\xi_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\mathscr{P}}$. Hence, the number of unknowns is $P + \mathrm{K}P = (\mathrm{K}+1)P$.

It can be observed that by summing (2.37b) over $i \in \mathcal{K}$, permuting the order of the double sum, we obtain

$$\sum_{\beta\in\mathscr{P}} Y_\beta \sum_{i\in\mathcal{K}} \xi_\beta^i - \sum_{i\in\mathcal{K}} c^i = 0. \tag{2.41}$$

By virtue of the third part of (2.40), which results from the complementarity condtions (2.39), we have

$$Y_\beta \sum_{i\in\mathcal{K}} \xi_\beta^i = Y_\beta.$$

Then, with the help of $\sum_{i\in\mathcal{K}} c^i = 1$, equation(2.41) becomes

$$\sum_{\beta\in\mathscr{P}} Y_\beta - 1 = 0,$$

which is none other than (2.37a). The latter equation is therefore redundant and should be left out of the system. To eliminate redundancy in the extended fugacity equalities, we fix a phase $\beta \in \mathscr{P}$ and require (2.38a) to hold for all $\alpha \in \mathscr{P}\backslash\{\beta\}$. The resulting system

$$\sum_{\beta\in\mathscr{P}} Y_\beta \xi_\beta^i - c^i = 0, \qquad \forall i \in \mathcal{K}; \tag{2.42a}$$

$$\xi_\alpha^i \Phi_\alpha^i(\boldsymbol{x}_\alpha) - \xi_\beta^i \Phi_\beta^i(\boldsymbol{x}_\beta) = 0, \qquad \forall (i,\alpha) \in \mathcal{K}\times\mathscr{P}\backslash\{\beta\}, \tag{2.42b}$$

$$\min\left(Y_\beta,\, 1 - \sum_{j\in\mathcal{K}} \xi_\beta^j\right) = 0, \qquad \forall \beta \in \mathscr{P}, \tag{2.42c}$$

in which the x_α^i's are seen as functions of the ξ_α^i's by means of (2.38b), contains

$$\mathrm{K} + \mathrm{K}(P-1) + P = (\mathrm{K}+1)P$$

equations. The independance of the extended fugacity equalities (2.42b) is a hypothesis to be made on the physical properties of the species. □

REMARK 2.3. To solve (2.42) in practice, Lauser *et al.* [77, 78] advocated using the common values $\{\varphi^i\}_{i\in\mathcal{K}}$ of extended fugacity across phases as main unknowns. This gives rise to a two-level algorithm. In the inner level, we solve P nonlinear systems of size $\mathrm{K}\times\mathrm{K}$

$$\xi_\alpha^i \Phi_\alpha^i(\boldsymbol{x}_\alpha) = \varphi^i, \qquad \forall i \in \mathcal{K}, \tag{2.43}$$

one for each $\alpha\in\mathscr{P}$. These local inversions express the extended fractions as implicit functions $\xi_\alpha^i(\boldsymbol{\varphi})$ of the extended fugacity vector $\boldsymbol{\varphi} = (\varphi^{\mathrm{I}}, \dots \varphi^{\mathrm{K}}) \in \mathbb{R}_+^{\mathrm{K}}$. In the outer level, we solve one nonlinear system of size $(\mathrm{K}+P)\times(\mathrm{K}+P)$ consisting of the remaining equations

$$\textstyle\sum_{\beta\in\mathscr{P}} Y_\beta \xi_\beta^i(\boldsymbol{\varphi}) - c^i = 0, \qquad \forall i \in \mathcal{K}, \tag{2.44a}$$

$$\min\big(Y_\beta,\, 1 - \textstyle\sum_{j\in\mathcal{K}} \xi_\beta^j(\boldsymbol{\varphi})\big) = 0, \qquad \forall \beta \in \mathscr{P}. \tag{2.44b}$$

This approach, the interest of which is to involve only "small" systems, was followed by subsequent works at IFPEN [12, 13, 84, 101]. The difficulty, however, lies in the computation of the gradients of the ξ_α^i's with respect to $\boldsymbol{\varphi}$, which are necessary for solving (2.44) *via* the Newton method. Analytically or numerically, these gradient evaluations are expensive. In view of this previous experience, we have preferred to tackle (2.42) in a more direct way. □

2.3 Properties of the unified formulation

The unified formulation enjoys many remarkable properties that seem to be unknown so far, at least to our knowledge. In particular, it achieves a deep connection with some classical results in thermodynamics. In this section, we are going to carefully derive these properties.

2.3.1 Behavior of tangent planes

Valuable insights can be gained by transforming the extended fugacity equalities (2.38a) into another form, the geometric significance of which is clearer. Before doing so, let us set the scene by introducing some concepts and notations. Recall that

$$\overline{\Omega} = \big\{\boldsymbol{x} = (x^{\mathrm{I}}, \dots, x^{\mathrm{K}-1}) \in \mathbb{R}^{\mathrm{K}-1} \mid x^{\mathrm{I}} \geqslant 0,\ \dots,\ x^{\mathrm{K}-1} \geqslant 0,\ 1 - x^{\mathrm{I}} - \dots - x^{\mathrm{K}-1} \geqslant 0\big\}$$

defined in (2.22b), is the domain of the (renormalized) partial fractions. In $\overline{\Omega}\times\mathbb{R}$, the generic element is denoted by $(\boldsymbol{x}, y)$. To each molar Gibbs energy function $g_\alpha : \overline{\Omega}\to\mathbb{R}$, we associate its graph

$$\mathcal{G}_\alpha = \big\{(\boldsymbol{x},\, y) \in \overline{\Omega}\times\mathbb{R} \mid y = g_\alpha(\boldsymbol{x})\big\}. \tag{2.45}$$

Note that we have not specified the phase subscript for the variable $\boldsymbol{x}$, since we intend to visualize several graphs on the same domain. For an interior point $\boldsymbol{x}_\alpha\in\Omega$, we designate by $T_{\boldsymbol{x}_\alpha}\mathcal{G}_\alpha$ the tangent hyperplane to $\mathcal{G}_\alpha$ at $\boldsymbol{x}_\alpha$. This tangent hyperplane, which exists thanks to the regularity assumptions on g_α, is the graph of the affine function $T_{\boldsymbol{x}_\alpha} g_\alpha : \mathbb{R}^{\mathrm{K}-1}\to\mathbb{R}$ defined as

$$T_{\boldsymbol{x}_\alpha} g_\alpha(\boldsymbol{x}) = g_\alpha(\boldsymbol{x}_\alpha) + \nabla_{\boldsymbol{x}} g_\alpha(\boldsymbol{x}_\alpha)\cdot(\boldsymbol{x} - \boldsymbol{x}_\alpha). \tag{2.46}$$

In general, $T_{\boldsymbol{x}_\alpha} g_\alpha$ and $T_{\boldsymbol{x}_\alpha}\mathcal{G}_\alpha$ cannot be defined in this way for $\boldsymbol{x}_\alpha\in\partial\Omega$, as $\nabla_{\boldsymbol{x}} g_\alpha(\boldsymbol{x}_\alpha)$ blows up.

Although the existence of a solution to the unified formulation (2.37)–(2.39) is not yet guaranteed, let us assume that $(\{\bar{Y}_\alpha\}_{\alpha\in\mathscr{P}},\ \{\bar{\xi}_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\mathscr{P}})$ is a solution satisfying $\bar{\boldsymbol{x}}_\alpha\in\Omega$ for all $\alpha\in\mathscr{P}$ and let us try to learn as much as we can about it.

Theorem 2.1. *For any pair $(\alpha, \beta) \in \mathscr{P} \times \mathscr{P}$ of phases, present or absent:*

1. *The* K *potentials in phase β are equal to their counterparts in phase α shifted by a same constant. More specifically, for all $j \in \mathcal{K}$,*

$$\mu_\beta^j(\bar{\boldsymbol{x}}_\beta) = \mu_\alpha^j(\bar{\boldsymbol{x}}_\alpha) + [\ln \bar{\sigma}_\alpha - \ln \bar{\sigma}_\beta], \tag{2.47a}$$

   where

$$\bar{\sigma}_\alpha = \sum_{i \in \mathcal{K}} \bar{\xi}_\alpha^i. \tag{2.47b}$$

2. *The two tangents hyperplanes $T_{\bar{\boldsymbol{x}}_\alpha}\mathcal{G}_\alpha$ and $T_{\bar{\boldsymbol{x}}_\beta}\mathcal{G}_\beta$ are parallel. More accurately, there holds the equality of gradients*

$$\nabla_{\boldsymbol{x}} g_\alpha(\bar{\boldsymbol{x}}_\alpha) = \nabla_{\boldsymbol{x}} g_\beta(\bar{\boldsymbol{x}}_\beta). \tag{2.47c}$$

Chứng minh. For each phase $\alpha \in \mathscr{P}$, let us define σ_α as in (2.47b), so that for all $j \in \mathcal{K}$, we have

$$\bar{\xi}_\alpha^j = \bar{\sigma}_\alpha \bar{x}_\alpha^j$$

in view of the normalization (2.38b). The extended fugacity equalities (2.38a) then become

$$\bar{\sigma}_\alpha \bar{x}_\alpha^j \Phi_\alpha^j(\bar{\boldsymbol{x}}_\alpha) = \bar{\sigma}_\beta \bar{x}_\beta^j \Phi_\beta^j(\bar{\boldsymbol{x}}_\beta). \tag{2.48}$$

Taking the natural logarithm of both sides and recalling definition (2.27) of the fugacity coefficient, we obtain

$$\ln \bar{\sigma}_\alpha + \mu_\alpha^j(\bar{\boldsymbol{x}}_\alpha) = \ln \bar{\sigma}_\beta + \mu_\beta^j(\bar{\boldsymbol{x}}_\beta). \tag{2.49}$$

From this, we deduce (2.47a). Subtracting the last equality

$$\ln \bar{\sigma}_\alpha + \mu_\alpha^{\mathrm{K}}(\bar{\boldsymbol{x}}_\alpha) = \ln \bar{\sigma}_\beta + \mu_\beta^{\mathrm{K}}(\bar{\boldsymbol{x}}_\beta).$$

from (2.49) and recalling (2.24c) [Lemma 2.1], we have

$$\frac{\partial g_\alpha}{\partial x^j}(\bar{\boldsymbol{x}}_\alpha) = \frac{\partial g_\beta}{\partial x^j}(\bar{\boldsymbol{x}}_\beta)$$

for all $j \in \{\mathrm{I}, \mathrm{II}, \ldots, \mathrm{K}-1\}$. This completes the proof for (2.47c). □

The first part of Theorem 2.1 indicates that, in general, there is no equality of chemical potentials, computed using the renormalized partial fractions. Equality holds in fact for *extended* chemical potentials, defined as $\ln(\xi_\alpha^i \Phi_\alpha^i(\boldsymbol{x}_\alpha))$. The second part of Theorem 2.1 is more interesting. Let us investigate this aspect further by making an additional assumption on one of the phases.

Theorem 2.2 (Tangent plane criterion)**.** *Assume that a phase $\alpha \in \mathscr{P}$ is present, i.e., $\bar{Y}_\alpha > 0$. Then, for any other phase $\beta \in \mathscr{P}$, absent or present,*

$$T_{\bar{\boldsymbol{x}}_\beta} g_\beta(\boldsymbol{x}) \geqslant T_{\bar{\boldsymbol{x}}_\alpha} g_\alpha(\boldsymbol{x}), \qquad \textit{for all } \boldsymbol{x} \in \mathbb{R}^{\mathrm{K}-1}, \tag{2.50}$$

where $T_{\bar{\boldsymbol{x}}_\alpha} g_\alpha$ and $T_{\bar{\boldsymbol{x}}_\beta} g_\beta$ are the linearized expansions defined in (2.46)*. In other words, the tangent hyperplane $T_{\bar{\boldsymbol{x}}_\beta}\mathcal{G}_\beta$ lies above or coincide with the tangent hyperplane $T_{\bar{\boldsymbol{x}}_\alpha}\mathcal{G}_\alpha$.*

Chứng minh. From equality (2.47a), we have

$$\mu_\beta^{\mathrm{K}}(\bar{\boldsymbol{x}}_\beta) = \mu_\alpha^{\mathrm{K}}(\bar{\boldsymbol{x}}_\alpha) + C_{\alpha\beta}, \qquad C_{\alpha\beta} = \ln \bar{\sigma}_\alpha - \ln \bar{\sigma}_\beta.$$

Since $\bar{Y}_\alpha > 0$, the complementarity condition (2.39) entails $\bar{\sigma}_\alpha = \sum_{j\in\mathcal{K}} \bar{\xi}_\alpha^j = 1$, hence $\ln \bar{\sigma}_\alpha = 0$. For any other $\beta \in \mathscr{P}$, we have $\bar{\sigma}_\beta = \sum_{j\in\mathcal{K}} \bar{\xi}_\beta^j \leqslant 1$, also by virtue of (2.39). Therefore, $\ln \bar{\sigma}_\beta \leqslant 0$ and $C_{\alpha\beta} \geqslant 0$. Thus,

$$\mu_\beta^{\mathrm{K}}(\bar{\boldsymbol{x}}_\beta) \geqslant \mu_\alpha^{\mathrm{K}}(\bar{\boldsymbol{x}}_\alpha).$$

Using (2.24b) from Lemma 2.1, we can rewrite the previous inequality as

$$g_\beta(\bar{\boldsymbol{x}}_\beta) - \nabla_{\boldsymbol{x}} g_\beta(\bar{\boldsymbol{x}}_\beta) \cdot \bar{\boldsymbol{x}}_\beta \geqslant g_\alpha(\bar{\boldsymbol{x}}_\beta) - \nabla_{\boldsymbol{x}} g_\alpha(\bar{\boldsymbol{x}}_\alpha) \cdot \bar{\boldsymbol{x}}_\alpha. \tag{2.51}$$

On the other hand, taking the dot product of the equality of gradients (2.47c) with any $\boldsymbol{x} \in \Omega$, we have

$$\nabla_{\boldsymbol{x}} g_\beta(\bar{\boldsymbol{x}}_\beta) \cdot \boldsymbol{x} = \nabla_{\boldsymbol{x}} g_\alpha(\bar{\boldsymbol{x}}_\alpha) \cdot \boldsymbol{x}. \tag{2.52}$$

Adding together (2.51) and (2.52), we end up with

$$g_\beta(\bar{\boldsymbol{x}}_\beta) + \nabla_{\boldsymbol{x}} g_\beta(\bar{\boldsymbol{x}}_\beta) \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}_\beta) \geqslant g_\alpha(\bar{\boldsymbol{x}}_\beta) + \nabla_{\boldsymbol{x}} g_\alpha(\bar{\boldsymbol{x}}_\alpha) \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}_\alpha)$$

which is the desired result (2.50). □

This result is notoriously known in thermodynamics as the *tangent plane criterion* [89]. It is usually derived by physicists from a local analysis of phase stability (see §2.3.2). Theorem 2.2 testifies to the fact that this stability property is already encoded in the unified formulation *via* the sign of $1 - \sum_{j\in\mathcal{K}} \bar{\xi}_\beta^j$. If phase β is "strictly" absent, namely, if $1 - \sum_{j\in\mathcal{K}} \bar{\xi}_\beta^j > 0$ and $\bar{Y}_\beta = 0$, then the tangent hyperplane $T_{\bar{\boldsymbol{x}}_\beta}\mathcal{G}_\beta$ will lie strictly above $T_{\bar{\boldsymbol{x}}_\alpha}\mathcal{G}_\alpha$.

Let us now push one step further by looking at the case of several present phases. Let $\bar{\Gamma}$ be the set of all $\alpha \in \mathscr{P}$ such that $\bar{Y}_\alpha > 0$. Its cardinal is denoted by $\bar{\gamma} = |\bar{\Gamma}|$.

Corollary 2.1 (Common tangent hyperplane)**.** *At a solution of the unified formulation satisfying $\bar{\boldsymbol{x}}_\alpha \in \Omega$ for all $\alpha \in \mathscr{P}$, the $\bar{\gamma}$ tangent hyperplanes $\{T_{\bar{\boldsymbol{x}}_\alpha}\mathcal{G}_\alpha\}_{\alpha\in\bar{\Gamma}}$ are all the same. Moreover,*

$$\boldsymbol{c} = (c^{\mathrm{I}}, \ldots, c^{\mathrm{K}-1}) \in \mathrm{int}(\mathrm{Conv}(\{\bar{\boldsymbol{x}}_\alpha\}_{\alpha\in\bar{\Gamma}})), \tag{2.53}$$

i.e., the global composition point belongs to the open convex hull spanned by the $\bar{\gamma}$ points $\{\bar{\boldsymbol{x}}_\alpha\}_{\alpha\in\bar{\Gamma}}$. Finally, a necessary condition for this solution to be unique is that

$$\bar{\gamma} \leqslant \mathrm{K}. \tag{2.54}$$

Chứng minh. Let $(\alpha, \beta) \in \bar{\Gamma} \times \bar{\Gamma}$. Applying Theorem 2.2 twice and switching their roles, we have $T_{\bar{\boldsymbol{x}}_\beta} g_\beta(\boldsymbol{x}) \geqslant T_{\bar{\boldsymbol{x}}_\alpha} g_\alpha(\boldsymbol{x})$ and $T_{\bar{\boldsymbol{x}}_\alpha} g_\alpha(\boldsymbol{x}) \geqslant T_{\bar{\boldsymbol{x}}_\beta} g_\beta(\boldsymbol{x})$, whence $T_{\bar{\boldsymbol{x}}_\alpha} g_\alpha(\boldsymbol{x}) = T_{\bar{\boldsymbol{x}}_\beta} g_\beta(\boldsymbol{x})$ for all $\boldsymbol{x} \in \Omega$. Thus, $T_{\bar{\boldsymbol{x}}_\alpha}\mathcal{G}_\alpha = T_{\bar{\boldsymbol{x}}_\beta}\mathcal{G}_\beta$. The material balance (2.37b) reads

$$c^i = \sum_{\beta\in\mathscr{P}} \bar{Y}_\beta \bar{\xi}_\beta^i = \sum_{\alpha\in\bar{\Gamma}} \bar{Y}_\alpha \bar{x}_\alpha^i,$$

where the last equality comes from retaining only those summands in the context, where the two notions of extended and partial fractions coincide. Extracting the first $\mathrm{K} - 1$ components from the above equation yields

$$\boldsymbol{c} = \sum_{\alpha\in\bar{\Gamma}} \bar{Y}_\alpha \bar{\boldsymbol{x}}_\alpha. \tag{2.55}$$

Since $\bar{Y}_\alpha > 0$ and $\sum_{\alpha\in\Gamma} \bar{Y}_\alpha = 1$, the point $\boldsymbol{c}$ belongs to the interior of $\mathrm{Conv}(\{\bar{\boldsymbol{x}}_\alpha\}_{\alpha\in\bar{\Gamma}})$, the dimension of which is at most $\bar{\gamma} - 1$. The weights $\{\bar{Y}_\alpha\}_{\alpha\in\bar{\Gamma}}$ of this convex combination are solutions of a linear system of K equations in $\bar{\gamma}$ unknowns. If $\bar{\gamma} > \mathrm{K}$, the matrix of the linear system has a nonzero kernel. Moving along a direction in this kernel with a small enough step, it is possible to find another set of weights satisfying the system while remaining positive. □

From this *common tangent plane* property, a purely geometric procedure can be devised in order to build a solution of the phase equilibrium formulated by (2.37)–(2.39). The construction involves the lower convex envelope of the function $\boldsymbol{x} \mapsto \min_{\alpha\in\mathscr{P}} g_\alpha(\boldsymbol{x})$. More details will be given in §2.4.2 for two-phase binary mixtures. Regarding condition (2.54), it is automatically satisfied when $P \leqslant \mathrm{K}$, which turns out to be true in practice: there are about two or three phases at most for tens to hundreds of components.

2.3.2 Connection with Gibbs energy minimization

The previous section §2.3.1 has revealed the benefits of imposing $1 - \sum_{i\in\mathcal{K}} \xi_\alpha^i \geqslant 0$ from the beginning, by means of the unified formulation (2.37)–(2.39). This enabled us to recover all the well-known properties of the solutions. We would like, however, to better understand where this sign information comes from.

2.3.2.1 On the origin of the sign condition

In the literature, the condition $1 - \sum_{i\in\mathcal{K}} \xi_\alpha^i \geqslant 0$ is customarily derived from a phase stability analysis. The most commonly cited reference is Michelsen [89], in relation to the tangent plane criterion. A more mathematical presentation was recently given by Ben Gharbia-Flauraud [12]. The idea is the following: starting from single-phase α, we wonder if the mixture would be "tempted" to split into two phases. The difference in the Gibbs energies between the new configuration and the old one is minimized with respect to all virtually possible compositions of a would-be new phase β. Phase α is said to be *stable* if the smallest value of this difference is positive. This gives rise to a condition on the composition of the fictitious phase β at which the minimum is reached. This condition is finally expressed in terms of the extended fractions, defined to be a rescaled version of the mole numbers in phase β.

This classical analysis suffers from a few limitations. First, it is restricted to two phases. Second, it is local: the Gibbs energy difference under study must be linearized via a first-order Taylor expansion, before minimizing. Third, the notion of extended fractions appears only at the end, in a very *ad hoc* way. It would be far more satisfying if we could derive a more direct connection between the unified formulation (2.37)–(2.39) and a multiphase multicomponent Gibbs enery minimization problem expressed in terms of the extended fractions ξ_α^i, without any linearization.

We claim that such a quest is attainable. In this section, we are going to show that every solution of the unified formulation is necessarily a critical point of some constrained minimization problem $(\mathcal{P})$ stated below. The quantities $1 - \sum_{i\in\mathcal{K}} \xi_\alpha^i$ will then appear to be the Lagrange multipliers associated with the constraints $Y_\alpha \geqslant 0$. Conversely, while not every critical point of the minimization problem $(\mathcal{P})$ is a solution of the unified formulation, some "natural" choice of critical points satisfies the unified formulation. This result, which does not seem to be known in the community, sheds a new light into the complementarity conditions (2.39).

2.3.2.2 Towards a novel interpretation

In order to state the minimization problem, we need to introduce a new Gibbs function. For each phase $\alpha \in \mathscr{P}$, let $\mathfrak{g} : \mathbb{R}^{\mathrm{K}}_{+} \to \mathbb{R}$ be the *extended* molar Gibbs energy defined as

$$\mathfrak{g}_\alpha(\xi^{\mathrm{I}}_\alpha, \ldots, \xi^{\mathrm{K}}_\alpha) = \sum_{i\in\mathcal{K}} \xi^i_\alpha \ln(\xi^i_\alpha \Phi^i_\alpha(\boldsymbol{x}_\alpha)). \tag{2.56}$$

For normalized fractions, $\mathfrak{g}_\alpha(x^{\mathrm{I}}_\alpha, \ldots, x^{\mathrm{K}}_\alpha) = g_\alpha(\boldsymbol{x}_\alpha)$. Thus, g_α extends the intensive Gibbs function g_α to the domain of extended fractions. It should not be confused with the extensive Gibbs function

$$\mathsf{G}_\alpha(\xi^{\mathrm{I}}_\alpha, \ldots, \xi^{\mathrm{K}}_\alpha) = \sum_{i\in\mathcal{K}} \xi^i_\alpha \ln(x^i_\alpha \Phi^i_\alpha(\boldsymbol{x}_\alpha)), \tag{2.57}$$

introduced in (2.15)–(2.16), using n^i_α in place of ξ^i_α. Unlike g_α and G_α, the extended function $\mathfrak{g}_\alpha$ is neither intensive nor extensive. But it has many handy properties summarized in the following Lemma. For convenience, we shall from now on be using the notations

$$\boldsymbol{\xi}_\alpha = (\xi^{\mathrm{I}}_\alpha, \ldots, \xi^{\mathrm{K}}_\alpha), \qquad \sigma_\alpha = \sum_{i\in\mathcal{K}} \xi^i_\alpha. \tag{2.58}$$

Lemma 2.3. *For all $\boldsymbol{\xi}_\alpha \in \mathbb{R}^{\mathrm{K}}_{+}$ and $j \in \mathcal{K}$,*

$$\frac{\partial \mathfrak{g}_\alpha}{\partial \xi^j_\alpha}(\boldsymbol{\xi}_\alpha) = \ln(\xi^j_\alpha \Phi^j_\alpha(\boldsymbol{x}_\alpha)) + 1, \tag{2.59a}$$

$$\mathfrak{g}_\alpha(\boldsymbol{\xi}_\alpha) = \sum_{i\in\mathcal{K}} \xi^i_\alpha \frac{\partial \mathfrak{g}_\alpha}{\partial \xi^i_\alpha}(\boldsymbol{\xi}_\alpha) - \sigma_\alpha, \tag{2.59b}$$

$$1 = \sum_{i\in\mathcal{K}} \xi^i_\alpha \frac{\partial \ln(\xi^i_\alpha \Phi^i_\alpha)}{\partial \xi^j_\alpha}(\boldsymbol{\xi}_\alpha). \tag{2.59c}$$

Chứng minh. Inserting $\xi^i_\alpha = \sigma_\alpha x^i_\alpha$ into (2.56), we find

$$\mathfrak{g}_\alpha(\boldsymbol{\xi}_\alpha) = \sum_{i\in\mathcal{K}} \xi^i_\alpha [\ln(x^i_\alpha \Phi^i_\alpha(\boldsymbol{x}_\alpha)) + \ln \sigma_\alpha] = \mathsf{G}_\alpha(\boldsymbol{\xi}_\alpha) + \sigma_\alpha \ln \sigma_\alpha.$$

Differentiating this equality with respect to ξ^j_α, we have

$$\frac{\partial \mathfrak{g}_\alpha}{\partial \xi^j_\alpha}(\boldsymbol{\xi}_\alpha) = \frac{\partial \mathsf{G}_\alpha}{\partial \xi^j_\alpha}(\boldsymbol{\xi}_\alpha) + \ln \sigma_\alpha + 1 = \ln(x^j_\alpha \Phi^j_\alpha(\boldsymbol{x}_\alpha)) + \ln \sigma_\alpha + 1,$$

which proves (2.59a). Multiplying (2.59a) by ξ^j_α and summing over $j \in \mathcal{K}$, we arrive at

$$\sum_{i\in\mathcal{K}} \xi^i_\alpha \frac{\partial \mathfrak{g}_\alpha}{\partial \xi^i_\alpha}(\boldsymbol{\xi}_\alpha) = \sum_{i\in\mathcal{K}} \xi^i_\alpha \ln(\xi^i_\alpha \Phi^j_\alpha(\boldsymbol{x}_\alpha)) + \sum_{i\in\mathcal{K}} \xi^i_\alpha = \mathfrak{g}_\alpha(\boldsymbol{\xi}_\alpha) + \sigma_\alpha,$$

which proves (2.59b). The last relation (2.59c) follows from

$$\sum_{i\in\mathcal{K}} \xi^i_\alpha \frac{\partial \ln(\xi^i_\alpha \Phi^i_\alpha)}{\partial \xi^j_\alpha}(\boldsymbol{\xi}_\alpha) = \sum_{i\in\mathcal{K}} \xi^i_\alpha \frac{\partial \ln(x^i_\alpha \Phi^i_\alpha)}{\partial \xi^j_\alpha}(\boldsymbol{\xi}_\alpha) + \sum_{i\in\mathcal{K}} \xi^i_\alpha \frac{\partial \ln \sigma_\alpha}{\partial \xi^j_\alpha},$$

in the right-hand side of which the first summand vanishes thanks to the Gibbs-Duhem condition (2.20) and the second summand boils down to 1. □

Equipped with this new Gibbs function, we can now consider the following minimization problem ($\mathcal{P}$).

GIVEN

$$\mathcal{K},\ \mathscr{P},\ \{\Phi_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\mathscr{P}} \text{ admissible},$$
$$\{c^i\}_{i\in\mathcal{K}} \in [0,1] \text{ subject to } \textstyle\sum_{i\in\mathcal{K}} c^i = 1,$$

FIND

$$\min_{\substack{\{Y_\alpha\}_{\alpha\in\mathscr{P}}\\ \{\boldsymbol{\xi}_\alpha\}_{\alpha\in\mathscr{P}}}} \sum_{\alpha\in\mathscr{P}} Y_\alpha \mathfrak{g}_\alpha(\boldsymbol{\xi}_\alpha) \tag{2.60a}$$

subject to

$$\sum_{\alpha\in\mathscr{P}} Y_\alpha - 1 = 0, \tag{2.60b}$$
$$\sum_{\alpha\in\mathscr{P}} Y_\alpha \xi_\alpha^i - c^i = 0, \qquad \forall i \in \mathcal{K}, \tag{2.60c}$$
$$-Y_\alpha \leqslant 0, \qquad \forall \alpha \in \mathscr{P}. \tag{2.60d}$$

The objective function in (2.60a) represents a notion of extended Gibbs energy for the mixture. The equality constraints (2.60b)–(2.60c) are exactly the material balances (2.37) of the unified formulation. This time, there is no redundancy since we have not imposed the complementarity conditions (2.39).

Let u, $\{v^i\}_{i\in\mathcal{K}}$ and $\{w_\alpha\}_{\alpha\in\mathscr{P}}$ be the Lagrange multipliers associated respectively with the constraints (2.60b), (2.60c) and (2.60d). The Lagrangian of the minimization problem (2.60) reads

$$\begin{aligned}\mathscr{L}(\{Y_\alpha\}, \{\boldsymbol{\xi}_\alpha\}, u, \{v^i\}, \{w_\alpha\}) = &\sum_{\alpha\in\mathscr{P}} Y_\alpha \mathfrak{g}_\alpha(\boldsymbol{\xi}_\alpha)\\ &+ u\Big(\sum_{\alpha\in\mathscr{P}} Y_\alpha - 1\Big) + \sum_{i\in\mathcal{K}} v^i \Big(\sum_{\alpha\in\mathscr{P}} Y_\alpha \xi_\alpha^i - c^i\Big) - \sum_{\alpha\in\mathscr{P}} w_\alpha Y_\alpha.\end{aligned}$$

The saddle-points of $\mathscr{L}$ are given by the Karush-Kuhn-Tucker (KKT) conditions [22, 94]

$$\mathfrak{g}_\beta(\boldsymbol{\xi}_\beta) + u + \sum_{i\in\mathcal{K}} v^i \xi_\beta^i - w_\beta = 0, \qquad \forall \beta \in \mathscr{P} \tag{2.61a}$$
$$Y_\beta \left[\frac{\partial \mathfrak{g}_\beta}{\partial \xi_\beta^j}(\boldsymbol{\xi}_\beta) + v^j\right] = 0, \qquad \forall (j,\beta) \in \mathcal{K}\times\mathscr{P}, \tag{2.61b}$$
$$\sum_{\alpha\in\mathscr{P}} Y_\alpha - 1 = 0, \tag{2.61c}$$
$$\sum_{\alpha\in\mathscr{P}} Y_\alpha \xi_\alpha^i - c^i = 0, \qquad \forall i \in \mathcal{K}, \tag{2.61d}$$
$$\min(Y_\beta, w_\beta) = 0, \qquad \forall \beta \in \mathscr{P}. \tag{2.61e}$$

The last equation (2.61e) expresses the complementarity between each inequality constraint (2.60d) and its Lagrange multiplier at optimality. It can be rephrased as

$$Y_\beta \geqslant 0, \qquad w_\beta \geqslant 0, \qquad Y_\beta\, w_\beta = 0.$$

A set of values $\{(Y_\alpha, \boldsymbol{\xi}_\alpha)\}_{\alpha\in\mathscr{P}}$ is said to be a *critical point* for the minimization problem (2.60) if there exists a set of values $(u, \{v^i\}_{i\in\mathcal{K}}, \{w_\alpha\}_{\alpha\in\mathscr{P}})$ such that the KKT optimality system (2.61) is satisfied.

2.3.2.3 From one formulation to the other

We first show that it is easy to go from the unified formulation to the minimization problem.

Theorem 2.3. *Every solution* $\{(\bar{Y}_\alpha, \bar{\boldsymbol{\xi}}_\alpha)\}_{\alpha\in\mathscr{P}}$ *of the unified formulation* (2.37)–(2.39) *is a critical point of the minimization problem* (2.60)*, with*

$$\bar{u} = 1, \qquad \bar{v}^j = -[\ln(\bar{\varphi}^j) + 1], \qquad \bar{w}_\beta = 1 - \bar{\sigma}_\beta, \tag{2.62}$$

where $\bar{\varphi}^j$ *is the common value of the extended fugacity* $\bar{\xi}^j_\alpha \Phi^j_\alpha(\bar{\boldsymbol{x}}_\alpha)$ *across all phases* $\alpha \in \mathscr{P}$*.*

Chứng minh. Let $\{(\bar{Y}_\alpha, \bar{\boldsymbol{\xi}}_\alpha)\}_{\alpha\in\mathscr{P}}$ be a solution of (2.37)–(2.39). The material balances (2.61c)–(2.61d) are naturally met, owing to (2.37). The equality of extended fugacities (2.38a) makes it possible to define $\bar{v}^j = -[\ln(\varphi^j) + 1]$ in the way described in the Theorem. This choice of $\bar{v}^j$ trivially fulfills (2.61b) because of (2.59a). The choice of $\bar{w}_\beta$ implies (2.61e) because of (2.39). It remains to check (2.61a). To this end, we use Lemma 2.3 to write

$$\mathfrak{g}_\beta(\bar{\boldsymbol{\xi}}_\beta) + \bar{u} + \sum_{i\in\mathcal{K}} \bar{v}^i \bar{\xi}^i_\beta - \bar{w}_\beta = \sum_{i\in\mathcal{K}} \bar{\xi}^i_\beta \frac{\partial \mathfrak{g}_\beta}{\partial \bar{\xi}^i_\beta}(\bar{\boldsymbol{\xi}}_\beta) - \bar{\sigma}_\beta + 1 - \sum_{i\in\mathcal{K}} \bar{\xi}^i_\beta \frac{\partial \mathfrak{g}_\beta}{\partial \bar{\xi}^i_\beta}(\bar{\boldsymbol{\xi}}_\beta) - (1 - \bar{\sigma}_\beta) = 0.$$

This completes the proof. □

In the reverse direction, things do not go as smoothly. The main difficulty lies in the indetermination of the extended fractions for an absent phase.

Theorem 2.4. *Let* $\{(\widetilde{Y}_\alpha, \widetilde{\xi}_\alpha)\}_{\alpha\in\mathscr{P}}$ *be a critical point of the minimization problem* (2.60)*.*

1. *If two phases* $(\alpha, \beta) \in \mathscr{P} \times \mathscr{P}$ *are both present, i.e.,* $\widetilde{Y}_\alpha > 0$ *and* $\widetilde{Y}_\beta > 0$*, then*

$$\widetilde{\sigma}_\alpha = \widetilde{\sigma}_\beta = 1, \qquad \widetilde{\xi}^i_\alpha \Phi^i_\alpha(\widetilde{\boldsymbol{x}}_\alpha) = \widetilde{\xi}^i_\beta \Phi^i_\beta(\widetilde{\boldsymbol{x}}_\beta) \quad \textit{for all } i \in \mathcal{K}. \tag{2.63}$$

   This implies that the complementarity condition (2.39) *holds for both phases and that the extended fugacity equalities* (2.38a) *hold between the two phases considered.*

2. *If phase* α *is present and phase* β *is absent, i.e.,* $\widetilde{Y}_\alpha > 0$ *and* $\widetilde{Y}_\beta = 0$*, then*

$$\widetilde{\sigma}_\alpha = 1, \qquad \sum_{i\in\mathcal{K}} \widetilde{\xi}^i_\beta \big[\, \ln(\widetilde{\xi}^i_\beta \Phi^i_\beta(\widetilde{\boldsymbol{x}}_\beta)) - \ln(\widetilde{\xi}^i_\alpha \Phi^i_\alpha(\widetilde{\boldsymbol{x}}_\alpha)) \,\big] + 1 - \widetilde{\sigma}_\beta \geqslant 0. \tag{2.64}$$

   In general, the complementarity condition (2.39) *does not hold for phase* β *and the extended fugacity equalities* (2.38a) *do not hold between* α *and* β*. But the complementarity condition* (2.39) *is automatically met for phase* β *as soon as the extended fugacity equalities* (2.38a) *hold between* α *and* β*.*

Chứng minh. Let $\{(\widetilde{Y}_\alpha, \widetilde{\xi}_\alpha)\}_{\alpha\in\mathscr{P}}$, $(\widetilde{u}, \{\widetilde{v}^i\}_{i\in\mathcal{K}}, \{\widetilde{w}_\alpha\}_{\alpha\in\mathscr{P}})$ be a solution of the KKT system (2.61). First, assume that $\widetilde{Y}_\alpha > 0$ and $\widetilde{Y}_\beta > 0$. It is then possible to simplify by $\widetilde{Y}$ in (2.61b) to obtain

$$\frac{\partial \mathfrak{g}_\alpha}{\partial \xi^j_\alpha}(\widetilde{\boldsymbol{\xi}}_\alpha) + \widetilde{v}^j = 0, \qquad \frac{\partial \mathfrak{g}_\beta}{\partial \xi^j_\beta}(\widetilde{\boldsymbol{\xi}}_\beta) + \widetilde{v}^j = 0.$$

From this, it is deduced that

$$\frac{\partial \mathfrak{g}_\alpha}{\partial \xi^j_\alpha}(\widetilde{\boldsymbol{\xi}}_\alpha) = \frac{\partial \mathfrak{g}_\beta}{\partial \xi^j_\beta}(\widetilde{\boldsymbol{\xi}}_\beta) = -\widetilde{v}^j.$$

According to (2.59a) [Lemma 2.3], this is equivalent to the equality of extended fugacities (2.38a), rewritten in the second part of (2.63). On the other hand, $\widetilde{Y}_\alpha > 0$ implies $\widetilde{w}_\alpha = 0$ by (2.61e). Equation (2.61a) then becomes

$$\mathfrak{g}_\alpha(\widetilde{\boldsymbol{\xi}}_\alpha) + \widetilde{u} - \sum_{i\in\mathcal{K}} \widetilde{\xi}^i_\alpha \frac{\partial \mathfrak{g}_\alpha}{\partial \widetilde{\xi}^i_\alpha}(\widetilde{\boldsymbol{\xi}}_\alpha) = 0.$$

Combining this with (2.59b) [Lemma 2.3], we infer that $\widetilde{\sigma}_\alpha = \widetilde{u}$. Repeating the same reasoning for β, we also get $\widetilde{\sigma}_\beta = \widetilde{u}$. Hence, $\widetilde{\sigma}_\alpha = \widetilde{\sigma}_\alpha$. This means that $\widetilde{\sigma}$ takes on the same value $\widetilde{u}$ in all present phases. Let $\widetilde{\Gamma}$ be set of $\pi \in \mathscr{P}$ such that $\widetilde{Y}_\pi > 0$. Note that $\widetilde{\Gamma} \neq \varnothing$ because of (2.61c). Summing (2.61d) over $i \in \mathcal{K}$ and permuting the order of summation yields

$$0 = \sum_{i\in\mathcal{K}} \sum_{\pi\in\mathscr{P}} \widetilde{Y}_\pi \widetilde{\xi}^i_\pi - \sum_{i\in\mathcal{K}} c^i = \sum_{\pi\in\mathscr{P}} \widetilde{Y}_\pi \widetilde{\sigma}_\pi - 1 = \widetilde{u} \sum_{\pi\in\widetilde{\Gamma}} \widetilde{Y}_\pi - 1 = \widetilde{u} - 1.$$

Therefore, $\widetilde{u} = 1$, which proves the first part of (2.63).

Assume now that $\widetilde{Y}_\alpha > 0$ and $\widetilde{Y}_\beta = 0$. It is no longer possible to divide (2.61b) by $\widetilde{Y}_\beta$ to retrieve information on the extended fugacities. Likewise, we now simply have $w_\beta \geqslant 0$ from (2.61e). Equation (2.61a) for phase β leads to

$$\mathfrak{g}_\beta(\widetilde{\boldsymbol{\xi}}_\beta) + \widetilde{u} + \sum_{i\in\mathcal{K}} \widetilde{v}^i \widetilde{\xi}^i_\beta = \widetilde{w}_\beta \geqslant 0.$$

Because phase α is present, $\widetilde{\sigma}_\alpha = \widetilde{u} = 1$ and $\widetilde{v}^i = -[\partial \mathfrak{g}_\alpha / \partial \xi^i_\alpha](\widetilde{\boldsymbol{\xi}}_\alpha)$. Invoking (2.59b) [Lemme 2.3] for phase β, we can transform the above equality into

$$\sum_{i\in\mathcal{K}} \widetilde{\xi}^i_\beta \left[\frac{\partial \mathfrak{g}_\beta}{\partial \xi^i_\beta}(\widetilde{\boldsymbol{\xi}}_\beta) - \frac{\partial \mathfrak{g}_\alpha}{\partial \xi^i_\alpha}(\widetilde{\boldsymbol{\xi}}_\alpha) \right] - \widetilde{\sigma}_\beta + 1 \geqslant 0.$$

This is none other than the second part of (2.64). □

To fully grasp the meaning of Theorem 2.4, it is capital to observe that when a critical point of (2.60) has a vanishing phase $\beta \in \mathscr{P}$ for which $\widetilde{Y}_\beta = 0$, the corresponding extended fractions $\widetilde{\boldsymbol{\xi}}_\beta$ cannot be uniquely determined. Indeed, $\widetilde{\boldsymbol{\xi}}_\beta$ plainly does not contribute to neither the objective function (2.60a) nor the constraint (2.60c) at fixed $\widetilde{Y}_\beta = 0$. To put it another way, changing $\widetilde{\boldsymbol{\xi}}_\beta$ to any other vector $\mathbb{R}^{\mathrm{K}}_+$ will provide another acceptable critical point. Thus, as soon as there is a critical point of (2.60) for which $\widetilde{Y}_\beta = 0$, there are in fact an infinity of such critical points. Among this infinity of critical points, only those for which

$$\widetilde{\xi}^i_\beta \Phi^i_\beta(\widetilde{\boldsymbol{x}}_\beta) = \widetilde{\xi}^i_\alpha \Phi^i_\alpha(\widetilde{\boldsymbol{x}}_\alpha) \quad \text{for all} \quad i \in \mathcal{K}, \tag{2.65}$$

where α is present phase ($\widetilde{Y}_\alpha > 0$), will be also solutions of the unified formulation (2.37)–(2.39). Combining this with Theorem 2.3, we can interpret the unified formulation as a set of equations that is slightly "stronger" than that of the KKT system for the critical points. It is stronger in the sense that it helps selecting some special critical points —and hopefully just one— among the infinity of possible critical points that appear when one of the phases disappears.

2.3.2.4 A continuity principle

We now give an argument to assert that the critical points thus selected by the unified formulation are "natural" ones. By this, we mean that the additional conditions (2.65) to be prescribed on the extended fractions of an absent phase β can be interpreted as the limit of a continuous process during which β was present before vanishing. To build up this process, let us reformulate the minimization problem $(\mathcal{P})$ or (2.60) as the *bilevel* or *hierarchical* problem

$$\min_{Y_\beta} \min_{\substack{\{Y_\alpha\}_{\alpha\in\mathscr{P}\backslash\{\beta\}} \\ \{\boldsymbol{\xi}_\alpha\}_{\alpha\in\mathscr{P}}}} \sum_{\alpha\in\mathscr{P}\backslash\{\beta\}} Y_\alpha \mathfrak{g}_\alpha(\boldsymbol{\xi}_\alpha) + Y_\beta \mathfrak{g}_\beta(\boldsymbol{\xi}_\beta) \tag{2.66a}$$

subject to

$$\sum_{\alpha\in\mathscr{P}\backslash\{\beta\}} Y_\alpha + Y_\beta - 1 = 0, \tag{2.66b}$$

$$\sum_{\alpha\in\mathscr{P}\backslash\{\beta\}} Y_\alpha \xi_\alpha^i + Y_\beta \xi_\beta^i - c^i = 0, \qquad \forall i \in \mathcal{K}, \tag{2.66c}$$

$$-Y_\alpha \leqslant 0, \qquad \forall \alpha \in \mathscr{P}\backslash\{\beta\}. \tag{2.66d}$$

The constraints (2.66b)–(2.66d) are imposed on the inner minimization problem $(\mathcal{P}_{Y_\beta})$

$$\min_{\substack{\{Y_\alpha\}_{\alpha\in\mathscr{P}\backslash\{\beta\}} \\ \{\boldsymbol{\xi}_\alpha\}_{\alpha\in\mathscr{P}}}} \sum_{\alpha\in\mathscr{P}\backslash\{\beta\}} Y_\alpha \mathfrak{g}_\alpha(\boldsymbol{\xi}_\alpha) + Y_\beta \mathfrak{g}_\beta(\boldsymbol{\xi}_\beta) \tag{2.67}$$

for a fixed $Y_\beta \geqslant 0$. To begin with, consider $(\mathcal{P}_{Y_\beta})$ for a fixed and small enough $Y_\beta > 0$. The KKT optimality conditions for (2.66b)–(2.67) are

$$\mathfrak{g}_\alpha(\boldsymbol{\xi}_\alpha) + u + \sum_{i\in\mathcal{K}} v^i \xi_\alpha^i - w_\alpha = 0, \qquad \forall \alpha \in \mathscr{P}\backslash\{\beta\} \tag{2.68a}$$

$$Y_\alpha \left[\frac{\partial \mathfrak{g}_\alpha}{\partial \xi_\alpha^i}(\boldsymbol{\xi}_\alpha) + v^i \right] = 0, \qquad \forall (i,\alpha) \in \mathcal{K} \times \mathscr{P}, \tag{2.68b}$$

$$\sum_{\alpha\in\mathscr{P}\backslash\{\beta\}} Y_\alpha + Y_\beta - 1 = 0, \tag{2.68c}$$

$$\sum_{\alpha\in\mathscr{P}\backslash\{\beta\}} Y_\alpha \xi_\alpha^i + Y_\beta \xi_\beta^i - c^i = 0, \qquad \forall i \in \mathcal{K}, \tag{2.68d}$$

$$\min(Y_\alpha,\, w_\alpha) = 0, \qquad \forall \alpha \in \mathscr{P}\backslash\{\beta\}. \tag{2.68e}$$

Note that (2.68a) and (2.68e) do not make sense for β since Y_β is not a variable for the inner problem, but that (2.68b) do make sense for (i,β) since ξ_β^i is a variable with respect to which minimization is carried out. Assume that for each small enough $Y_\beta > 0$ there is a unique critical point. We designate it by

$$\{\widetilde{Y}_\alpha(Y_\beta)\}_{\alpha\in\mathscr{P}\backslash\{\beta\}},\ \{\widetilde{\boldsymbol{\xi}}_\alpha(Y_\beta)\}_{\alpha\in\mathscr{P}}$$

to lay emphasis on its dependency with respect to Y_β. Setting $\alpha = \beta$ in (2.68b), we are allowed to divide by $Y_\beta > 0$ in order to obtain

$$-\widetilde{v}(Y_\beta) = \frac{\partial \mathfrak{g}_\beta}{\partial \xi_\beta^i}(\widetilde{\boldsymbol{\xi}}_\alpha(Y_\beta)) = \ln(\widetilde{\xi}_i^\beta(Y_\beta)\Phi(\widetilde{\boldsymbol{x}}_\beta(Y_\beta))) \qquad \text{for all } \ i \in \mathcal{K}.$$

Setting α in (2.68b) to another present phase (which necessary exists since $Y_\beta < 1$) and simplifying by $Y_\alpha > 0$, we have

$$-\widetilde{v}(Y_\beta) = \frac{\partial \mathfrak{g}_\alpha}{\partial \xi_\alpha^i}(\widetilde{\boldsymbol{\xi}}_\alpha(Y_\beta)) = \ln(\widetilde{\xi}_i^\alpha(Y_\beta)\Phi(\widetilde{\boldsymbol{x}}_\alpha(Y_\beta))) \qquad \text{for all } \ i \in \mathcal{K}.$$

From the last two equalities, it follows that

$$\widetilde{\xi}_i^\beta(Y_\beta)\Phi(\widetilde{\boldsymbol{x}}_\beta(Y_\beta)) = \widetilde{\xi}_i^\alpha(Y_\beta)\Phi(\widetilde{\boldsymbol{x}}_\alpha(Y_\beta)) \qquad \text{for all } \ i \in \mathcal{K}.$$

Now, we let $Y_\beta \downarrow 0$. If all of the quantities involved in the above equality have finite limits, we clearly end up with (2.65). The values assigned to the extended fractions in an absent phase in the unified formulation are thus based on a continuity principle for the critical point.

2.3.3 Well-definedness of extended fractions

Let us go back to the equality of gradients (2.47c) and ask ourselves the following question. Assume that phase β is well determined, namely, the extended fractions $\{\xi_\beta^i\}_{i\in\mathcal{K}}$ are known. Under which hypotheses on the Gibbs energy g_α would it be possible to invert the relation

$$\nabla_{\boldsymbol{x}} g_\alpha(\boldsymbol{x}_\alpha) = \nabla_{\boldsymbol{x}} g_\beta(\boldsymbol{x}_\beta)$$

in order to get the partial fractions $\boldsymbol{x}_\alpha$, from which the extended fractions $\{\xi_\alpha^i\}_{i\in\mathcal{K}}$ could also be calculated? Mathematically, this makes sense insofar as we have a $(\mathrm{K}-1)\times(\mathrm{K}-1)$ nonlinear system. Before elaborating on the requirements to be imposed on g_α, let us point out two instances where this issue crucially arises.

2.3.3.1 Two essential issues

The first situation occurs when the solution is single-phase, say, in phase β. Put another way, $\bar{Y}_\beta = 1$ and $\bar{Y}_\alpha = 0$ for all $\alpha \in \mathscr{P}\backslash\{\beta\}$. By (2.37b), rewritten as (2.55), we have $\bar{\boldsymbol{x}}_\beta = \boldsymbol{c}$. Assume $\boldsymbol{c} \in \Omega$. After Theorem 2.1, the extended fractions in a vanishing phase $\alpha \in \mathscr{P}\backslash\{\beta\}$ satisfy

$$\nabla_{\boldsymbol{x}} g_\alpha(\bar{\boldsymbol{x}}_\alpha) = \nabla_{\boldsymbol{x}} g_\beta(\boldsymbol{c}), \tag{2.69a}$$

$$\ln \bar{\sigma}_\alpha + \mu_\alpha^{\mathrm{K}}(\bar{\boldsymbol{x}}_\alpha) = \mu_\beta^{\mathrm{K}}(\boldsymbol{c}), \tag{2.69b}$$

where we recall that

$$\bar{\sigma}_\alpha = \sum_{j\in\mathcal{K}} \bar{\xi}_\alpha^j, \qquad \bar{x}_\alpha^i = \frac{\bar{\xi}_\alpha^i}{\bar{\sigma}_\alpha}.$$

If (2.69a) could be uniquely inverted, i.e., if we had the legitimacy to write

$$\bar{\boldsymbol{x}}_\alpha = [\nabla_{\boldsymbol{x}} g_\alpha]^{-1}(\nabla_{\boldsymbol{x}} g_\beta(\boldsymbol{c})),$$

then we could easily deduce from (2.69b) that

$$\bar{\sigma}_\alpha = \exp[\mu_\beta^{\mathrm{K}}(\boldsymbol{c}) - \mu_\alpha^{\mathrm{K}}(\bar{\boldsymbol{x}}_\alpha)], \qquad \bar{\xi}_\alpha^i = \bar{\sigma}_\alpha \bar{x}_\alpha^i,$$

and phase α would be entirely determined. We refer to this first situation as the *vanishing phases* problem.

The second situation takes place in Lauser's suggestion for using the extended fugacities, as mentioned in Remark 2.3. By means of similar operations (taking the log of both sides, introducing the sum of extended fractions, using the connection between the potentials and the molar Gibbs energy), the inner system (2.43) can be transformed into

$$\nabla_{\boldsymbol{x}} g_\alpha(\boldsymbol{x}_\alpha) = \{\ln \varphi^i - \ln \varphi^{\mathrm{K}}\}_{1 \leqslant j \leqslant \mathrm{K}-1}, \tag{2.70a}$$

$$\ln \sigma_\alpha + \mu_\alpha^{\mathrm{K}}(\boldsymbol{x}_\alpha) = \ln \varphi^{\mathrm{K}}, \tag{2.70b}$$

which displays exactly the same structure as (2.69). Our ability to solve (2.70) for all reasonable inputs $\boldsymbol{\varphi} \in \mathbb{R}_+^{\mathrm{K}}$ relies on the existence of an unambiguous reciprocal function $[\nabla_{\boldsymbol{x}} g_\alpha]^{-1}$. We refer to this second situation as the *local fugacity inversion* problem.

2.3.3.2 Two sets of assumptions

The claimed superiority of the unified formulation over the variable-switching formulation rests upon its capability to assign well-determined values to the extended fractions in the absent phases. Failing to do so would ultimately defeat its purpose. The short calculation above demonstrates that this capability cannot be taken for granted. Additional assumptions need to be made in order to ensure that the unified formulation works properly. Below is the most natural one.

Hypotheses 2.1. The gradient map $\nabla_{\boldsymbol{x}} g_\alpha : \Omega \to \mathbb{R}^{\mathrm{K}-1}$ is a homemorphism. In other words, it is a continuous bijection as well as its inverse $[\nabla_{\boldsymbol{x}} g_\alpha]^{-1} : \mathbb{R}^{\mathrm{K}-1} \to \Omega$.

The following noteworthy example hints that Hypotheses 2.1 is neither unrealistic nor unreachable.

Proposition 2.3. *The molar Gibbs energy function of an ideal gas*

$$g_\alpha(\boldsymbol{x}_\alpha) = \sum_{i=\mathrm{I}}^{\mathrm{K}} x_\alpha^i \ln x_\alpha^i, \tag{2.71}$$

where $x_\alpha^{\mathrm{K}} = 1 - x_\alpha^{\mathrm{I}} - \ldots - x_\alpha^{\mathrm{K}-1}$*, satisfies Hypotheses 2.1.*

Chứng minh. To alleviate notations, let us omit the phase subscript α of $\boldsymbol{x}$. The gradient $\nabla_{\boldsymbol{x}} g_\alpha : \Omega \to \mathbb{R}^{\mathrm{K}-1}$ is given by

$$\nabla_{\boldsymbol{x}} g_\alpha(\boldsymbol{x}) = (\ln x^{\mathrm{I}} - \ln x^{\mathrm{K}}, \ldots, \ln x^{\mathrm{K}-1} - \ln x^{\mathrm{K}}). \tag{2.72}$$

This map is continuous over Ω. For any given $\boldsymbol{u} = (u^{\mathrm{I}}, \ldots, u^{\mathrm{K}-1}) \in \mathbb{R}^{\mathrm{K}-1}$, the nonlinear system $\nabla_{\boldsymbol{x}} g_\alpha(\boldsymbol{x}) = \boldsymbol{u}$ can be turned into the $\mathrm{K} \times \mathrm{K}$ linear system

$$\begin{aligned} x^{\mathrm{I}} - \exp(u^{\mathrm{I}}) x^{\mathrm{K}} &= 0, \\ &\vdots \\ x^{\mathrm{K}-1} - \exp(u^{\mathrm{K}-1}) x^{\mathrm{K}} &= 0, \\ x^{\mathrm{I}} + \ldots + x^{\mathrm{K}-1} + x^{\mathrm{K}} &= 1. \end{aligned}$$

The first $\mathrm{K} - 1$ components of the solution are

$$x^{\mathrm{I}} = \frac{\exp(u^{\mathrm{I}})}{1 + \sum_{i=\mathrm{I}}^{\mathrm{K}-1} \exp(u^i)}, \quad \ldots, \quad x^{\mathrm{K}-1} = \frac{\exp(u^{\mathrm{K}-1})}{1 + \sum_{i=\mathrm{I}}^{\mathrm{K}-1} \exp(u^i)}.$$

This defines a unique continuous inverse map $[\nabla_{\boldsymbol{x}} g_\alpha]^{-1} : \mathbb{R}^{\mathrm{K}-1} \to \Omega$. □

Unfortunately, Hypotheses 2.1 may not be easy to check for fluids other than an ideal gas. Therefore, it could be more convenient to consider some stronger but more convenient hypotheses.

Hypotheses 2.2. The gradient map $\nabla_{\boldsymbol{x}} g_\alpha : \Omega \to \mathbb{R}^{\mathrm{K}-1}$ is surjective. Moreover, the molar Gibbs energy $g_\alpha : \Omega \to \mathbb{R}$ is *strictly* convex, that is, it satisfies one of the two conditions below, which are equivalent for a twice differentiable function:

(a) For all $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega \times \Omega$ with $\boldsymbol{x} \neq \boldsymbol{y}$,

$$\big\langle \nabla_{\boldsymbol{x}} g_\alpha(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} g_\alpha(\boldsymbol{y}),\, \boldsymbol{x} - \boldsymbol{y} \big\rangle > 0. \tag{2.73}$$

(b) For all $\boldsymbol{x} \in \Omega$, the Hessian matrix $\nabla^2_{\boldsymbol{xx}} g_\alpha(\boldsymbol{x})$ is definite positive.

We refer the reader to [24, 109] for the notion of strict convexity and for the equivalence between the two conditions (a) and (b) for twice differentiable functions. Surjectivity provides existence of a solution $\boldsymbol{x} \in \Omega$ to $\nabla_{\boldsymbol{x}} g_\alpha(\boldsymbol{x}) = \boldsymbol{u} \in \mathbb{R}^{\mathrm{K}-1}$. Strict convexity enforces uniqueness of such a solution. Again, the case of an ideal gas suggests that this is not an unreasonable assumption.

Proposition 2.4. *The molar Gibbs energy function of an ideal gas, defined by* (2.71)*, is strictly convex.*

Chứng minh. Again, we drop the phase subscript α for clarity. From the expression (2.72) of the gradient, the Hessian matrix can be found to be

$$\nabla^2_{\boldsymbol{xx}} g(\boldsymbol{x}) = \frac{1}{x^{\mathrm{K}}} \mathbf{E} + \mathrm{Diag}\left(\frac{1}{x^{\mathrm{I}}}, \ldots, \frac{1}{x^{\mathrm{K}-1}} \right),$$

where $\mathbf{E}$ is the matrix whose all entries are equal to 1. It follows that, for a generic $\boldsymbol{v} \in \mathbb{R}^{\mathrm{K}-1}$,

$$\big\langle \nabla^2_{\boldsymbol{xx}} g(\boldsymbol{x})\boldsymbol{v},\, \boldsymbol{v} \big\rangle = \frac{1}{x^{\mathrm{K}}} |v^{\mathrm{I}} + \ldots + v^{\mathrm{K}-1}|^2 + \sum_{i=1}^{\mathrm{K}-1} \frac{|v^i|^2}{x^i}.$$

When $\boldsymbol{x} \in \Omega$, it is obvious that $\big\langle \nabla^2_{\boldsymbol{xx}} g(\boldsymbol{x})\boldsymbol{v},\, \boldsymbol{v} \big\rangle > 0$ for all $\boldsymbol{v} \neq \mathbf{0}$. □

To conclude this section, Hypotheses 2.2 set the framework in which we can guarantee that the extended fractions introduced in the unified formation are well-defined. Strict convexity of the molar Gibbs energy will also be of great help in proving non-singularity of the solution of the unified formulation in chapter §5.

2.4 Two-phase mixtures

Throughout the rest of this manuscript, we shall study only the two-phase case, for which

$$\mathscr{P} = \{G,\, L\}, \qquad P = 2. \tag{2.74}$$

The two-phase case is sufficiently representative of the numerical difficulties we wish to address, while simple enough to make implementations faster. The new labels G (gas) and L (liquid) are aimed at being more meaningful and fixing ideas. They have no consequence on the ensuing mathematical developments.

2.4.1 The multicomponent case

Let us write down the corresponding unified formulation in the simplest way possible. System (2.42) is now reduced to

$$Y_G\xi_G^i + Y_L\xi_L^i - c^i = 0, \qquad \forall i \in \mathcal{K}, \tag{2.75a}$$
$$\xi_G^i\Phi_G^i(\boldsymbol{x}_G) - \xi_L^i\Phi_L^i(\boldsymbol{x}_L) = 0, \qquad \forall i \in \mathcal{K}, \tag{2.75b}$$
$$\min\left(Y_G,\, 1 - \textstyle\sum_{j\in\mathcal{K}} \xi_G^j\right) = 0, \tag{2.75c}$$
$$\min\left(Y_L,\, 1 - \textstyle\sum_{j\in\mathcal{K}} \xi_L^j\right) = 0. \tag{2.75d}$$

2.4.1.1 A further reduction

This $(2\mathrm{K}+2)\times(2\mathrm{K}+2)$ nonlinear system can be further simplified as follows. As we already know that the phasic fractions will automatically satisfy $Y_G + Y_L = 1$, let us choose one of them as unknown and express the other as its complement to 1, that is,

$$Y_G = Y, \qquad Y_L = 1 - Y. \tag{2.76}$$

This enables us to work with the $(2\mathrm{K}+1)\times(2\mathrm{K}+1)$ nonlinear system

$$\boxed{\begin{aligned}
Y\xi_G^i + (1-Y)\xi_L^i - c^i &= 0, \qquad \forall i \in \mathcal{K}\backslash\{\mathrm{K}\}, && (2.77a)\\
\xi_G^i\Phi_G^i(\boldsymbol{x}_G) - \xi_L^i\Phi_L^i(\boldsymbol{x}_L) &= 0, \qquad \forall i \in \mathcal{K}, && (2.77b)\\
\min\left(Y,\, 1 - \textstyle\sum_{j\in\mathcal{K}} \xi_G^j\right) &= 0, && (2.77c)\\
\min\left(1-Y,\, 1 - \textstyle\sum_{j\in\mathcal{K}} \xi_L^j\right) &= 0. && (2.77d)
\end{aligned}}$$

in the unknowns $(Y, \boldsymbol{\xi}_G, \boldsymbol{\xi}_L) \in \mathbb{R}\times\mathbb{R}^{\mathrm{K}}\times\mathbb{R}^{\mathrm{K}}$. It is important to point out that in the material balances (2.77a), there are now only $\mathrm{K}-1$ equations. As system (2.77) has one less unknown than (2.75), it should also have one less equation. We have decided to leave aside the material balance of the last component K. Let us rationalize this decision.

From the complementarity condition (2.77c), it follows that

$$Y = Y\sum_{j\in\mathcal{K}} \xi_G^j = Y\sum_{j\in\mathcal{K}\backslash\{\mathrm{K}\}} \xi_G^j + Y\xi_G^{\mathrm{K}}.$$

Hence,

$$Y\sum_{j\in\mathcal{K}\backslash\{\mathrm{K}\}} \xi_G^j = Y - Y\xi_G^{\mathrm{K}}. \tag{2.78a}$$

Likewise, starting from the complementarity condition (2.77d), we have

$$(1-Y)\sum_{j\in\mathcal{K}\backslash\{\mathrm{K}\}} \xi_G^j = (1-Y) - (1-Y)\xi_L^{\mathrm{K}}. \tag{2.78b}$$

Summing the material balances (2.77a) over $i \in \mathcal{K}\backslash\{\mathrm{K}\}$, invoking (2.78) and recalling that $\sum_{j\in\mathcal{K}\backslash\{\mathrm{K}\}} c^j = 1 - c^{\mathrm{K}}$ yield

$$[Y - Y\xi_G^{\mathrm{K}}] + [(1-Y) - (1-Y)\xi_L^{\mathrm{K}}] - (1 - c^{\mathrm{K}}) = 0.$$

After simplification and a change of sign, we obtain the material balance of component K. Thus, the "forgotten" equation can be in fact recovered from those prescribed in (2.77).

2.4.1.2 Two kinds of singularity

There are two kinds of singular solutions to which we should pay attention. These are noteworthy not only from the physical standpoint, but also from the mathematical perspective. Indeed, the determinant of some Jacobian matrix vanishes at the singular solutions, which will be unfavorable for numerical methods, as will be seen later. The first family of singular solutions was already briefly mentioned in §2.2.2.

Definition 2.1. A solution $(\bar{Y}, \bar{\boldsymbol{\xi}}_G, \bar{\boldsymbol{\xi}}_L) \in \mathbb{R} \times \mathbb{R}^{\mathrm{K}} \times \mathbb{R}^{\mathrm{K}}$ of (2.77) is said to be a *transition* point when both arguments of one of the complementarity conditions vanish simultaneously, that is,

$$\bar{Y} = 0, \qquad 1 - \sum_{i \in \mathcal{K}} \bar{\xi}^i_G = 0, \tag{2.79a}$$

or

$$\bar{Y} = 1, \qquad 1 - \sum_{i \in \mathcal{K}} \bar{\xi}^i_L = 0. \tag{2.79b}$$

In the two-phase framework, such a point marks the change in the nature of the solution, from a two-phase regime to a single-phase regime or vice-versa. To avoid ambiguity due to transition points, we say that the gas phase G is *strictly* absent if $Y = 0$ and $1 - \sum_{i \in \mathcal{K}} \bar{\xi}^i_G > 0$. Likewise, we say that the liquid phase L is *strictly* absent if $Y = 1$ and $1 - \sum_{i \in \mathcal{K}} \bar{\xi}^i_L > 0$.

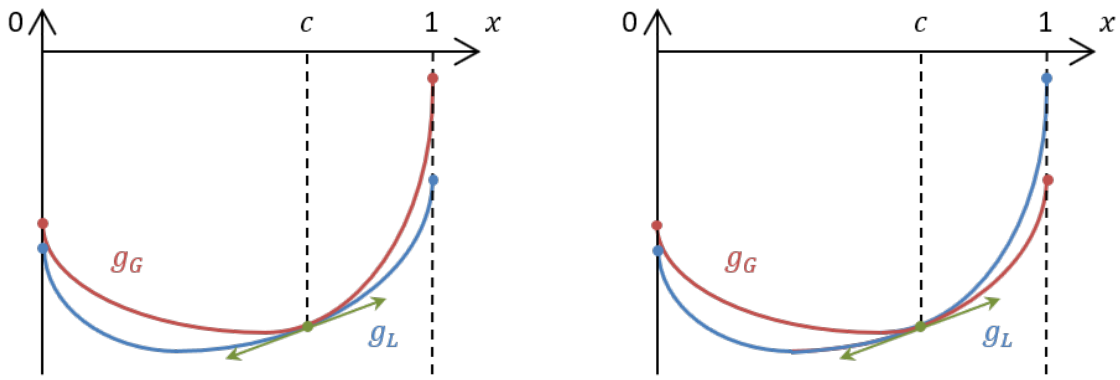


Figure 2.1: Azeotropic compositions for a two-phase two-component mixture.

Definition 2.2. A global composition $\boldsymbol{c} \in \Omega$, where $\Omega \subset \mathbb{R}^{\mathrm{K}-1}$ is the open domain of fractions defined in (2.22a), is said to be *azeotropic* if the Gibbs hypersurfaces $\mathcal{G}_G$ and $\mathcal{G}_L$, defined in (2.45), are tangent to each other at $\boldsymbol{c}$. In other words if $T_{\boldsymbol{c}}\mathcal{G}_G = T_{\boldsymbol{c}}\mathcal{G}_L$, or equivalently,

$$g_G(\boldsymbol{c}) = g_L(\boldsymbol{c}), \qquad \nabla_{\boldsymbol{x}} g_G(\boldsymbol{c}) = \nabla_{\boldsymbol{x}} g_L(\boldsymbol{c}). \tag{2.80}$$

Note that $\boldsymbol{c}$ alone is not responsible for azeotropy. It also takes the two Gibbs functions to behave in a peculiar way to satisfy (2.80). If azeotropy occurs at some $\boldsymbol{c} \in \Omega$, then it is easily seen that

$$(\bar{Y}, \bar{\xi}^i_G, \bar{\xi}^i_L) = (Y, c^i, c^i) \tag{2.81}$$

is a solution of (2.77) for all $Y \in [0, 1]$. This infinity of solutions is undertermined with respect to the phasic fraction Y. Physically speaking, since the two phases have identical proportions of

species, they can no longer be distinguished from each other. Therefore, it is no longer possible to tell how much of a phase is globally present in the mixture[3]. The second kind of singularity consists of azeotropic solutions (2.80)–(2.81). An illustration of azeotropic configurations is given in Figure 2.1 for a two-component mixture, with $\mathrm{K} = 2$ and $\Omega = (0, 1)$.

2.4.2 The binary case

The special case of two-phase two-component mixtures, for which in addition to (2.74),

$$\mathcal{K} = \big\{\mathrm{I}, \mathrm{II}\big\}, \qquad \mathrm{K} = 2, \tag{2.82}$$

is called *binary*. Thanks to its simplicity, analytical calculations can performed and geometric constructions to worked out, which helps gaining intuition into the phase equilibrium problem.

2.4.2.1 Notations and assumptions

As $\mathrm{K} - 1 = 1$, the domain of partial fractions is $\overline{\Omega} = [0, 1]$. For conciseness, we shall be using the symbols c instead of c^{I}, x_G instead of x_G^{I} and x_L instead of x_L^{I}. The vectors $\boldsymbol{c} = (c^{\mathrm{I}})$, $\boldsymbol{x}_G = (x_G^{\mathrm{I}})$ and $\boldsymbol{x}_L = (x_L^{\mathrm{I}})$ will also be written as c, x_G and x_L. We recall that

$$x_G = \frac{\xi_G^{\mathrm{I}}}{\xi_G^{\mathrm{I}} + \xi_G^{\mathrm{II}}}, \qquad x_L = \frac{\xi_L^{\mathrm{I}}}{\xi_L^{\mathrm{I}} + \xi_L^{\mathrm{II}}}.$$

The two-phase multicomponent model (2.77) then boils down to

$$\boxed{\begin{aligned} Y\xi_G^{\mathrm{I}} + (1-Y)\xi_L^{\mathrm{I}} - c &= 0, && \text{(2.83a)}\\ \xi_G^{\mathrm{I}}\,\Phi_G^{\mathrm{I}}(x_G) - \xi_L^{\mathrm{I}}\,\Phi_L^{\mathrm{I}}(x_L) &= 0, && \text{(2.83b)}\\ \xi_G^{\mathrm{II}}\,\Phi_G^{\mathrm{II}}(x_G) - \xi_L^{\mathrm{II}}\,\Phi_L^{\mathrm{II}}(x_L) &= 0, && \text{(2.83c)}\\ \min(Y; 1 - \xi_G^{\mathrm{I}} - \xi_G^{\mathrm{II}}) &= 0, && \text{(2.83d)}\\ \min(1-Y; 1 - \xi_L^{\mathrm{I}} - \xi_L^{\mathrm{II}}) &= 0. && \text{(2.83e)} \end{aligned}}$$

There are five equations in the unknowns $(Y, \xi_G^{\mathrm{I}}, \xi_G^{\mathrm{II}}, \xi_L^{\mathrm{I}}, \xi_L^{\mathrm{II}}) \in \mathbb{R}^5$. Admissibility of the fugacity coefficients $\Phi_\alpha^{\mathrm{I}}, \Phi_\alpha^{\mathrm{II}}$ for $\alpha \in \{G, L\}$ imply that they derive from the molar Gibbs energy functions $g_\alpha : [0, 1] \to \mathbb{R}$ defined as

$$g_\alpha(x) = x \ln(x\,\Phi_\alpha^{\mathrm{I}}(x)) + (1-x)\ln((1-x)\,\Phi_\alpha^{\mathrm{II}}(x)),$$

and in particular that they meet the Gibbs-Duhem condition

$$x\,(\ln \Phi_\alpha^{\mathrm{I}})'(x) + (1-x)\,(\ln \Phi_\alpha^{\mathrm{II}})'(x) = 0$$

for all $x \in (0, 1)$, where $'$ denotes the derivative with respect to x. In §2.3.3, Hypotheses 2.2 were set out in an attempt to guarantee existence and uniqueness in most situations. Here, we wish to strengthen these hypotheses in order to include the extreme cases $c = 0$ and $c = 1$ in the upcoming analytical solution for (2.83).

[3] In chemical engineering, the phases can no longer be separated by distillation at an azeotropic composition.

Hypotheses 2.3. The molar Gibbs energy g_α is strictly convex, that is, $g''_\alpha(x) > 0$ for all $x \in (0,1)$. Moreover, the gradient $g'_\alpha : (0,1) \to \mathbb{R}$ can be extended to be a surjective map from $[0,1]$ to $\overline{\mathbb{R}} = \{-\infty\} \cup \mathbb{R} \cup \{+\infty\}$, with $g'_\alpha(0) = -\infty$ and $g'_\alpha(1) = +\infty$.

Hypotheses 2.3 enable us to extend the inverse map $[g'_\alpha]^{-1} : \overline{\mathbb{R}} \to [0,1]$, with $[g'_\alpha]^{-1}(-\infty) = 0$ and $[g'_\alpha]^{-1}(+\infty) = 1$. Note that for $\mathrm{K} \geqslant 3$, it is no longer possible to include $\pm\infty$ in the range of the components of $\nabla_{\boldsymbol{x}} g_\alpha$. This is testified by the ideal gas law (2.71), for which the difficulty lies on the hyperplane $1 - x^{\mathrm{I}} - \ldots - x^{\mathrm{K}-1} = 0$.

2.4.2.2 Gibbs' geometric construction

There is a famous geometric construction due to Gibbs [39, 95] for the phase equilibrium of a two-phase binary mixture. Let us sketch out this construction, depicted in Figure 2.2. The first task is to look for a common tangent between the graphs $\mathcal{G}_G$ and $\mathcal{G}_L$.

- ▷ If there is no common tangent, then the mixture is single-phase. The present phase α is the one whose graph lies below the other. In phase α, the partial fraction is $\bar{x}_\alpha = c$.
- ▷ If there is a common tangent, let $\breve{x}_G$ and $\breve{x}_L$ be the abscissae of the contact points.
  - If c lies outside the open interval defined by $\breve{x}_G$ and $\breve{x}_L$, then the mixture is single-phase. The present phase α is the one whose graph lies below the other at the abscissa c. In phase α, the partial fraction is $\bar{x}_\alpha = c$.
  - If c lies inside the open interval defined by $\breve{x}_G$ and $\breve{x}_L$, then the mixture is two-phase. The partial fractions are then $\bar{x}_G = \breve{x}_G$ and $\bar{x}_L = \breve{x}_L$. The phasic fraction $\bar{Y}$ is given by the *lever rule*, which amounts to calculating the weights by which c can be expressed as a convex combination of $\bar{x}_G$ and $\bar{x}_L$.

Note that Gibbs' geometric construction is concerned with the natural-variable or variable-switching formulation (2.34)–(2.35). In a pure liquid regime ($\bar{Y} = 0$), for instance, it does not make sense to speak about $\bar{x}_G$ because phase G does not exist. Using the unified formulation, however, we can assign a well-defined value to $\bar{x}_G$. As represented in the lower panel of Figure 2.2, this value is $\bar{x}_G = [g'_G]^{-1}(g'_L(\bar{x}_L)) = [g'_G]^{-1}(g'_L(c))$.

2.4.2.3 Analytical derivation

In the same spirit as in §2.3.1, we seek to rigorously derive the Gibbs geometric construction from model (2.83). Our contribution is summarized in the Theorem below. To our knowledge, this is the first existence and uniqueness result for a non-azeotropic phase equilibrium problem using the unified formulation.

Theorem 2.5. *Assume that Hypotheses 2.3 hold and that the given composition $c \in [0,1]$ is not azeotropic. Then, system* (2.83) *has a unique solution $(\bar{Y}, \bar{\xi}^{\mathrm{I}}_G, \bar{\xi}^{\mathrm{II}}_G, \bar{\xi}^{\mathrm{I}}_L, \bar{\xi}^{\mathrm{II}}_L) \in [0,1] \times \mathbb{R}^4_+$, given by the following procedure. Let $\breve{g}$ be the lower convex envelope of $\min(g_G; g_L)$ over $[0,1]$, that is,*

$$\breve{g}(x) = \sup\{g(x) \mid g \text{ is convex and } g \leqslant \min(g_G; g_L) \text{ over } [0,1]\}.$$

- *If $\breve{g}(c) < \min(g_G(c); g_L(c))$, then in the neighborhood of $(c, \breve{g}(c))$ the graph of $\breve{g}(\cdot)$ is a straightline. This straightline is a common tangent to the graphs of $\mathcal{G}_G$ and $\mathcal{G}_L$. Let*

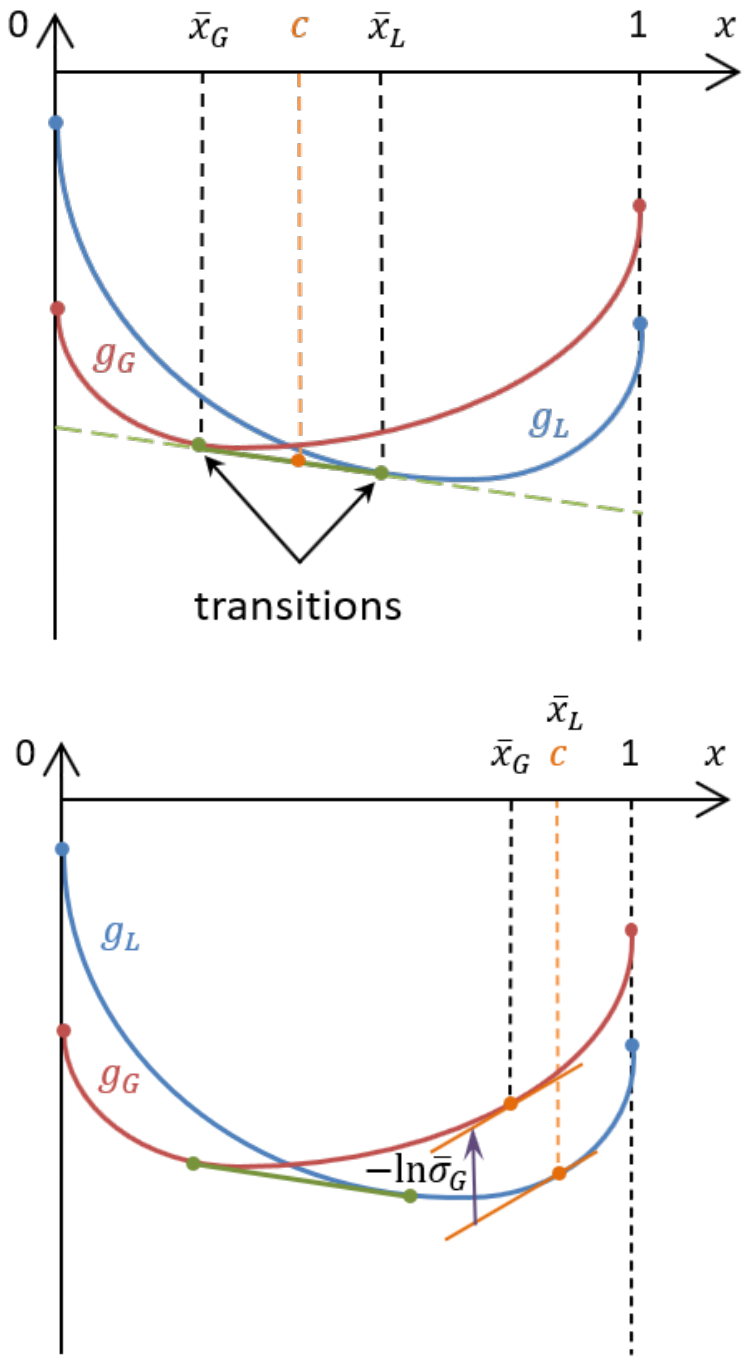


Figure 2.2: Gibbs' geometric construction for the phase equilibrium of a two-phase binary mixture. Top: two-phase solution. Bottom: single-phase solution.

$(\bar{x}_G, g_G(\bar{x}_G))$ and $(\bar{x}_L, g_L(\bar{x}_L))$ be the distinct contact points. The abscissae of these contact points are necessarily from distinct sides of c, one on the left and the other on the right. The solution is then in the two-phase regime, with

$$\bar{Y} = \frac{c - \bar{x}_L}{\bar{x}_G - \bar{x}_L}, \quad \bar{\xi}_G^{\mathrm{I}} = \bar{x}_G, \quad \bar{\xi}_G^{\mathrm{II}} = 1 - \bar{x}_G, \quad \bar{\xi}_L^{\mathrm{I}} = \bar{x}_L, \quad \bar{\xi}_L^{\mathrm{II}} = 1 - \bar{x}_L. \tag{2.84}$$

- *If $\breve{g}(c) = g_G(c)$, then at least in a half-neighborhood of $(c, \breve{g}(c))$ the graph of $\breve{g}(\cdot)$ coincides with $\mathcal{G}_G$. The solution is then in the G single-phase regime, with*

$$\bar{Y} = 1, \quad \bar{x}_G = c, \quad \bar{\xi}_G^{\mathrm{I}} = c, \quad \bar{\xi}_G^{\mathrm{II}} = 1 - c, \quad \bar{x}_L = [g_L']^{-1}(g_G'(c)), \tag{2.85a}$$

and, recalling the definition (2.46) *of tangent map $T_{\boldsymbol{x}_\alpha} g_\alpha$,*

$$\bar{\xi}_L^{\mathrm{I}} = \exp\big[T_c g_G(\bar{x}_L) - g_L(\bar{x}_L)\big]\, \bar{x}_L, \tag{2.85b}$$

$$\bar{\xi}_L^{\mathrm{II}} = \exp\big[T_c g_G(\bar{x}_L) - g_L(\bar{x}_L)\big](1 - \bar{x}_L). \tag{2.85c}$$

- *If $\breve{g}(c) = g_L(c)$, then at least in a half-neighborhood of $(c, \breve{g}(c))$ the graph of $\breve{g}(\cdot)$ coincides with $\mathcal{G}_L$. The solution is then in the L single-phase regime, with*

$$\bar{Y} = 0, \quad \bar{x}_L = c, \quad \bar{\xi}_L^{\mathrm{I}} = c, \quad \bar{\xi}_L^{\mathrm{II}} = 1 - c, \quad \bar{x}_G = [g_G']^{-1}(g_L'(c)), \tag{2.86a}$$

and, recalling the definition (2.46) *of tangent map $T_{\boldsymbol{x}_\alpha} g_\alpha$,*

$$\bar{\xi}_G^{\mathrm{I}} = \exp\big[T_c g_L(\bar{x}_G) - g_G(\bar{x}_G)\big]\, \bar{x}_G, \tag{2.86b}$$

$$\bar{\xi}_G^{\mathrm{II}} = \exp\big[T_c g_L(\bar{x}_G) - g_G(\bar{x}_G)\big](1 - \bar{x}_G). \tag{2.86c}$$

Chứng minh. Before proving Theorem 2.5, we remark that thanks to the non-azeotropy assumption, the above procedure is non-ambiguous: we cannot have $\breve{g}(c) = g_G(c) = g_L(c)$.

EXISTENCE. Let us check that the procedure described leads to a valid solution of (2.83). It is well-known that the graph of the lower convex envelope of a continuous function is made up of successive parts, where the envelope either exactly matches the graph of the function or is a straightline that lies strictly below the initial function but that is eventually tangent to the latter at the ends of that part.

Two-phase regime. If $\breve{g}(c) < \min(g_G(c); g_L(c))$, then the part of the envelope containing $(c, \breve{g}(c))$ is necessarily a segment of a straightline that is tangent to the graph of $\min(g_G(c); g_L(c))$ at two points $(x_-, \breve{g}(x_-))$ and $(x_+, \breve{g}(x_+))$, with $x_- < c < x_+$. We claim that

$$\text{either} \quad (\breve{g}(x_-), \breve{g}(x_+)) = (g_G(x_-), g_L(x_+)) \quad \text{or} \quad (\breve{g}(x_-), \breve{g}(x_+)) = (g_L(x_-), g_G(x_+)). \tag{2.87}$$

Suppose by contradiction that $(\breve{g}(x_-), \breve{g}(x_+)) = (g_G(x_-), g_G(x_+))$. Then, the part of the graph of $\breve{g}(\cdot)$ passing through $(c, \breve{g}(c))$ is a straightline tangent to the graph of $g_G(\cdot)$ at abscissae x_- and x_+. Hence, $g_G'(x_-) = g_G'(x_+)$. But this violates the strict convexity of g_G, according to which g_G' should be strictly increasing. Similarly, we get a contradiction by supposing that $(\breve{g}(x_-), \breve{g}(x_+)) = (g_L(x_-), g_L(x_+))$. This proves the claim (2.87). As a consequence, up to relabelling and reordering, the two contact points can be designated as $(\bar{x}_G, g_G(\bar{x}_G))$ and $(\bar{x}_L, g_L(\bar{x}_L))$, which geometrically determines the real fractions $\bar{x}_G$ and $\bar{x}_L$. The part of the envelope containing $(c, \breve{g}(c))$ is therefore a common tangent to $g_G(\cdot)$ and $g_L(\cdot)$.

Let us verify that the set of values (2.84) has correct range and solves indeed (2.83). Since c lies strictly between $\bar{x}_G$ and $\bar{x}_L$, the quantity $\bar{Y} = (c - \bar{x}_L)/(\bar{x}_G - \bar{x}_L)$ lies in $(0, 1)$. Since $\bar{x}_G \in [0, 1]$ and $\bar{x}_L \in [0, 1]$ by construction, the quantities $\bar{\xi}_G^{\mathrm{I}}$, $\bar{\xi}_G^{\mathrm{II}}$, $\bar{\xi}_L^{\mathrm{I}}$ and $\bar{\xi}_L^{\mathrm{II}}$ computed by (2.84) also belong to $[0, 1]$. These values plainly satisfy equations (2.83a) and (2.83d)–(2.83e). It remains to check (2.83b)–(2.83c). But we know that these are equivalent to

$$g_G'(\bar{x}_G) = g_L'(\bar{x}_L), \tag{2.88a}$$
$$g_G(\bar{x}_G) - \bar{x}_G g_G'(\bar{x}_G) = g_L(\bar{x}_L) - \bar{x}_L g_L'(\bar{x}_L), \tag{2.88b}$$

on the ground of earlier calculations [Theorem 2.1]. The first equality holds thanks to common tangency. Now, we observe that $g_G(\bar{x}_G) - \bar{x}_G g_G'(\bar{x}_G)$ is the ordinate of the intersection between the tangent line $T_{\bar{x}_G} g_G$ and the axis $x = 0$. Likewise, $g_L(\bar{x}_L) - \bar{x}_L g_L'(\bar{x}_L)$ is the ordinate of the intersection between the tangent line $T_{\bar{x}_L} g_L$ and the axis $x = 0$. Again, $T_{\bar{x}_G} g_G = T_{\bar{x}_L} g_L$ entails the second equality.

Single-phase regime. If $\breve{g}(c) = g_G(c)$, then the part of the envelope containing $(c, g_G(c))$ coincides with $\mathcal{G}_G$ in a neigborhood or at least in a half-neighborhood of c. Since $\breve{g}$ is convex, its graph lies above its tangent line at c, i.e.,

$$T_c g_G(x) = g_G(c) + (x - c) g_G'(c) = \breve{g}(c) + (x - c)\breve{g}'(c) \leqslant \breve{g}(x) \tag{2.89}$$

for all $x \in [0, 1]$. The quantities computed by (2.85) are all non-negative and obviously satisfy (2.83a) and (2.83d). However, the fact that $\bar{Y} = 1$ is not enough to infer (2.83e). We still have to check the inequality $1 - \bar{\xi}_L^{\mathrm{I}} - \bar{\xi}_L^{\mathrm{II}} \geqslant 0$. But

$$\bar{\sigma}_L = \bar{\xi}_L^{\mathrm{I}} + \bar{\xi}_L^{\mathrm{II}} = \exp\big[T_c g_G(\bar{x}_L) - g_L(\bar{x}_L)\big] \leqslant \exp\big[\breve{g}(\bar{x}_L) - g_L(\bar{x}_L)\big] \leqslant \exp(0) = 1,$$

where the last two inequalities result from (2.89) and from the definition of $\breve{g}$. Thus, $\bar{\sigma}_L \leqslant 1$ and (2.83e) is satisfied. It remains to check (2.83b)–(2.83c). On the ground of earlier calcuations [Theorem 2.1], these are equivalent to

$$g_G'(c) = g_L'(\bar{x}_L), \tag{2.90a}$$
$$g_G(c) - c g_G'(c) = g_L(\bar{x}_L) - \bar{x}_L g_L'(\bar{x}_L) + \ln \bar{\sigma}_L. \tag{2.90b}$$

The first equation is already satisfied. The second one stems from $\ln \bar{\sigma}_L = T_c g_G(\bar{x}_L) - g_L(\bar{x}_L)$.

If $\breve{g}(c) = g_L(c)$, the proof goes along the same lines.

UNIQUENESS. By virtue of Theorem 2.1, any solution $(Y, \xi_G^{\mathrm{I}}, \xi_G^{\mathrm{II}}, \xi_L^{\mathrm{I}}, \xi_L^{\mathrm{II}})$ of (2.83) satisfies $g_G'(x_G) = g_L'(x_L)$ regardless of its phase regime. For $\wp \in \mathbb{R}$, we define

$$\breve{x}_G(\wp) = [g_G']^{-1}(\wp), \qquad \breve{x}_L(\wp) = [g_L']^{-1}(\wp) \tag{2.91}$$

and consider the Legendre transforms

$$\mathfrak{L}_{g_G}(\wp) = g_G(\breve{x}_G(\wp)) - \breve{x}_G(\wp)\wp, \qquad \mathfrak{L}_{g_L}(\wp) = g_L(\breve{x}_L(\wp)) - \breve{x}_L(\wp)\wp$$

of g_G and g_L. A straightforward calculation shows that

$$\frac{\mathrm{d}\mathfrak{L}_{g_G}}{\mathrm{d}\wp}(\wp) = -\breve{x}_G(\wp), \qquad \frac{\mathrm{d}\mathfrak{L}_{g_L}}{\mathrm{d}\wp}(\wp) = -\breve{x}_L(\wp). \tag{2.92}$$

If $(\bar{Y}, \bar{\xi}_G^{\mathrm{I}}, \bar{\xi}_G^{\mathrm{II}}, \bar{\xi}_L^{\mathrm{I}}, \bar{\xi}_L^{\mathrm{II}})$ and $(\widetilde{Y}, \widetilde{\xi}_G^{\mathrm{I}}, \widetilde{\xi}_G^{\mathrm{II}}, \widetilde{\xi}_L^{\mathrm{I}}, \widetilde{\xi}_L^{\mathrm{II}})$ are two solutions of (2.83), we let

$$\bar{\wp} = g_G'(\bar{x}_G) = g_L'(\bar{x}_L), \qquad \widetilde{\wp} = g_G'(\widetilde{x}_G) = g_L'(\widetilde{x}_L). \tag{2.93}$$

Of a single-phase solution. Suppose $(\bar{Y}, \bar{\xi}_G^{\mathrm{I}}, \bar{\xi}_G^{\mathrm{II}}, \bar{\xi}_L^{\mathrm{I}}, \bar{\xi}_L^{\mathrm{II}})$ is a G single-phase solution. Then,

$$\bar{Y} = 1, \quad \bar{x}_G = c, \quad \bar{\xi}_G^{\mathrm{I}} = c, \quad \bar{\xi}_G^{\mathrm{II}} = 1 - c, \quad g_G'(\bar{x}_G) = g_L'(\bar{x}_L),$$

and

$$\ln \bar{\sigma}_L = \ln(\bar{\xi}_L^{\mathrm{I}} + \bar{\xi}_L^{\mathrm{II}}) = g_G(c) + (\bar{x}_L - c) g_G'(c) - g_L(\bar{x}_L) = \mathfrak{L}_{g_G}(\bar{\wp}) - \mathfrak{L}_{g_L}(\bar{\wp}).$$

by usual transformations. Thus, a G single-phase solution is necessarily given by formulas (2.85). Furthermore, in order to ensure $\bar{\sigma}_L \leqslant 1$, we must have

$$\mathfrak{L}_{g_G}(\bar{\wp}) - \mathfrak{L}_{g_L}(\bar{\wp}) \leqslant 0. \tag{2.94}$$

If the other solution $(\widetilde{Y}, \widetilde{\xi}_G^{\mathrm{I}}, \widetilde{\xi}_G^{\mathrm{II}}, \widetilde{\xi}_L^{\mathrm{I}}, \widetilde{\xi}_L^{\mathrm{II}})$ were in the same G single-phase regime, it would also be explicitly given formulas (2.85) and would therefore coincide with $(\bar{Y}, \bar{\xi}_G^{\mathrm{I}}, \bar{\xi}_G^{\mathrm{II}}, \bar{\xi}_L^{\mathrm{I}}, \bar{\xi}_L^{\mathrm{II}})$. For the second solution to be distinct from the first one, it has to be either in the L single-phase or in the two-phase regime.

If $(\widetilde{Y}, \widetilde{\xi}_G^{\mathrm{I}}, \widetilde{\xi}_G^{\mathrm{II}}, \widetilde{\xi}_L^{\mathrm{I}}, \widetilde{\xi}_L^{\mathrm{II}})$ were in the L single-phase regime, a similar analysis would reveal that it is necessarily given by formulas (2.86) and that

$$\mathfrak{L}_{g_G}(\widetilde{\wp}) - \mathfrak{L}_{g_L}(\widetilde{\wp}) \geqslant 0. \tag{2.95}$$

Let us investigate three subcases.

(i) If $\widetilde{\wp} > \bar{\wp}$, then by inverting the increasing functions in (2.93) we have $\bar{x}_L < c < \widetilde{x}_G$, and even $\bar{x}_L < \breve{x}_L(\wp) < c < \breve{x}_G(\wp) < \widetilde{x}_G$ for all $\wp \in (\bar{\wp}, \widetilde{\wp})$, using (2.91). Therefore, after (2.92),

$$\frac{\mathrm{d}}{\mathrm{d}\wp}(\mathfrak{L}_{g_G} - \mathfrak{L}_{g_L})(\wp) = \breve{x}_L(\wp) - \breve{x}_G(\wp) < 0$$

for all $\wp \in (\bar{\wp}, \widetilde{\wp})$. The difference $\mathfrak{L}_{g_G} - \mathfrak{L}_{g_L}$ is decreasing over $(\bar{\wp}, \widetilde{\wp})$, whence

$$\mathfrak{L}_{g_G}(\widetilde{\wp}) - \mathfrak{L}_{g_L}(\widetilde{\wp}) < \mathfrak{L}_{g_G}(\bar{\wp}) - \mathfrak{L}_{g_L}(\bar{\wp}).$$

But this obviously contradicts (2.94)–(2.95).

(ii) If $\widetilde{\wp} < \bar{\wp}$, then by inverting the increasing functions in (2.93), we have $\widetilde{x}_G < c < \bar{x}_L$, and even $\widetilde{x}_G < \breve{x}_G(\wp) < c < \breve{x}_L(\wp) < \bar{x}_L$ for all $\wp \in (\widetilde{\wp}, \bar{\wp})$, using (2.91). Therefore, after (2.92),

$$\frac{\mathrm{d}}{\mathrm{d}\wp}(\mathfrak{L}_{g_G} - \mathfrak{L}_{g_L})(\wp) = \breve{x}_L(\wp) - \breve{x}_G(\wp) > 0$$

for all $\wp \in (\widetilde{\wp}, \bar{\wp})$. The difference $\mathfrak{L}_{g_G} - \mathfrak{L}_{g_L}$ is increasing over $[\widetilde{\wp}, \bar{\wp}]$, whence

$$\mathfrak{L}_{g_G}(\widetilde{\wp}) - \mathfrak{L}_{g_L}(\widetilde{\wp}) < \mathfrak{L}_{g_G}(\bar{\wp}) - \mathfrak{L}_{g_L}(\bar{\wp}).$$

This is the same contradiction as in (i).

(iii) If $\widetilde{\wp} = \bar{\wp}$, then $g'_G(c) = g'_L(c)$ by (2.93). On the other hand, (2.94)–(2.95) imply that $\mathfrak{L}_{g_G}(\bar{\wp}) = \mathfrak{L}_{g_L}(\bar{\wp})$. In other words,

$$g_G(c) - cg'_G(c) = g_L(c) - cg'_L(c),$$

from which we infer that $g_G(c) = g_L(c)$. This means that c is an azeotropic composition, which is excluded by the assumptions of the Theorem.

If $(\widetilde{Y}, \widetilde{\xi}^{\mathrm{I}}_G, \widetilde{\xi}^{\mathrm{II}}_G, \widetilde{\xi}^{\mathrm{I}}_L, \widetilde{\xi}^{\mathrm{II}}_L)$ were in a strict two-phase regime, then c should lie strictly between $\widetilde{x}_G$ and $\widetilde{x}_L$. Furthermore, the common tangency implies

$$\mathfrak{L}_{g_G}(\widetilde{\wp}) - \mathfrak{L}_{g_L}(\widetilde{\wp}) = 0. \tag{2.96}$$

Let us investigate three subcases.

(i) If $\widetilde{\wp} > \bar{\wp}$, then by convexity $\bar{x}_L < \widetilde{x}_L < c < \widetilde{x}_G$. For all $\wp \in (\bar{\wp}, \widetilde{\wp})$, we have $\bar{x}_L < \breve{x}_L(\wp) < \widetilde{x}_L < c < \breve{x}_G(\wp) < \widetilde{x}_G$. Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}\wp}(\mathfrak{L}_{g_G} - \mathfrak{L}_{g_L})(\wp) = \breve{x}_L(\wp) - \breve{x}_G(\wp) < 0$$

for all $\wp \in (\bar{\wp}, \widetilde{\wp})$. As before, we obtain a contradiction by comparing the values of $\mathfrak{L}_{g_G} - \mathfrak{L}_{g_L}$ at $\bar{\wp}$ and $\widetilde{\wp}$.

(ii) If $\widetilde{\wp} < \bar{\wp}$, then by convexity $\widetilde{x}_G < \widetilde{x}_L < c < \bar{x}_L$. For all $\wp \in (\widetilde{\wp}, \bar{\wp})$, we have $\widetilde{x}_G < \breve{x}_G(\wp) < c < \widetilde{x}_L < \breve{x}_L(\wp) < \bar{x}_L$. Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}\wp}(\mathfrak{L}_{g_G} - \mathfrak{L}_{g_L})(\wp) = \breve{x}_L(\wp) - \breve{x}_G(\wp) > 0$$

for all $\wp \in (\widetilde{\wp}, \bar{\wp})$. Again, we obtain a contradiction by comparing the values of $\mathfrak{L}_{g_G} - \mathfrak{L}_{g_L}$ at $\widetilde{\wp}$ and $\bar{\wp}$.

(iii) If $\widetilde{\wp} = \bar{\wp}$, then $\widetilde{x}_G = c$ by applying $[g'_G]^{-1}$. This entails $\widetilde{Y} = 1$, which means that we are at a transition point. The second solution is not in a strict two-phase regime.

Of a two-phase solution. Suppose $(\bar{Y}, \bar{\xi}^{\mathrm{I}}_G, \bar{\xi}^{\mathrm{II}}_G, \bar{\xi}^{\mathrm{I}}_L, \bar{\xi}^{\mathrm{II}}_L)$ is a strict two-phase solution. By algebraic manipulations similar to the previous ones, it is not difficult to show that this two-phase solution is necessarily given by formulas (2.84). This prevents any other solution $(\widetilde{Y}, \widetilde{\xi}^{\mathrm{I}}_G, \widetilde{\xi}^{\mathrm{II}}_G, \widetilde{\xi}^{\mathrm{I}}_L, \bar{\xi}^{\mathrm{II}}_L)$ to be in the strict two-phase regime. The only possibility for a second solution to exist is that it is in a single-phase regime. However, we have just proven that a single-phase solution of (2.83) cannot co-exist with any other solution, unless it is the same. □

Chapter 3

Convexity analysis and extension of Gibbs energy functions

Contents

Après avoir formulé au chapitre précédent le problème de l'équilibre des phases d'un mélange compositionnel de manière générale, nous nous intéressons à présent à l'expression de quelques lois physiques spécifiques habituellement utilisées pour la fonction d'énergie de Gibbs.

La première famille de fonctions de Gibbs que nous examinons en §3.1 provient de lois physiques assez simples. Il s'agit de la loi des coefficients constants pour un gaz multiconstituant et des modèles d'activité de Margules et de Van Laar pour un liquide binaire. Pour chacune d'entre elles, nous étudions dans l'espace de ses paramètres les régions assurant les Hypothèses 2.2, en particulier la stricte convexité.

La seconde famille est celle des fonctions de Gibbs associées à une équation d'état cubique. Tout en rappelant en §3.2 leur construction, nous nous livrons à une analyse des zones d'existence d'une ou de trois racines réelles. Cette analyse est effectuée directement dans le plan des paramètres adimensionnés, ce qui consitue une originalité et fournit une expression analytique utile des frontières.

L'analyse révèle également des problèmes concernant le domaine de définition des fonctions de Gibbs, fort nuisibles au bon fonctionnement de la formulation unifiée. Deux remèdes sont proposés en §3.3 pour étendre les domaines de définition, le plus prometteur étant la méthode indirecte qui de par sa généralité n'est pas restreinte au cas binaire.

In chapter §2, we formulated the phase equilibrium problem for a multicomponent mixture in a quite general way, with an abstract molar Gibbs energy function per phase. Our goal is now to review some widely used expressions of these Gibbs functions. As introduced in (2.31)–(2.32), the molar Gibbs energy function takes the form

$$g_\alpha(\boldsymbol{x}) = \sum_{i=\mathrm{I}}^{\mathrm{K}} x^i \ln x^i + \Psi_\alpha(\boldsymbol{x}) \tag{3.1}$$

for each phase $\alpha \in \mathscr{P}$, where Ψ_α denotes the *excess* function. To alleviate notations, we have dropped the phase subscript α for the dummy argument $\boldsymbol{x} \in \overline{\Omega} \subset \mathbb{R}^{\mathrm{K}-1}$. Specifying g_α amounts therefore to specifying Ψ_α, from which the fugacity coefficients are deduced by means of (2.33a) [Lemma 2.2], which we rewrite as

$$\ln \Phi_\alpha^j(\boldsymbol{x}) = \Psi_\alpha(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} \Psi_\alpha(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^j - \boldsymbol{x}), \qquad \text{for all } \ j \in \mathcal{K}. \tag{3.2}$$

For each law of Ψ_α presented, we endeavour whenever possible to study its adequacy with the Hypotheses 2.2, in particular the issue of strict convexity of g_α.

3.1 Convexity analysis for simple Gibbs functions

In this section, the subscript α stands for the phase to which the physical law under consideration applies. Let us start with some simple laws. Perhaps the simplest one is that of an ideal gas

$$\Psi_\alpha \equiv 0.$$

In §2.3.3 [Proposition 2.4], we proved that an ideal gas fulfills Hypotheses 2.2.

3.1.1 Henry's law

Next in the level of complexity is Henry's law [62, 115]

$$\Psi_\alpha(\boldsymbol{x}) = \sum_{i=1}^{\mathrm{K}} x^i \ln k^i \tag{3.3}$$

where $\{k^i\}_{i \in \mathcal{K}}$ are positive constants, each of them embodying a property of the corresponding species. The fugacity coefficients are then

$$\ln \Phi_\alpha^j(\boldsymbol{x}) = \ln k^i, \qquad \text{for all } \ j \in \mathcal{K}. \tag{3.4}$$

This is why this law is also referred to as the *constant coefficients* law.

Proposition 3.1. *For all $(k^{\mathrm{I}}, \ldots, k^{\mathrm{K}}) \in (\mathbb{R}_+^*)^{\mathrm{K}}$, the molar Gibbs energy function g_α associated with Henry's law fulfills Hypotheses 2.2.*

Chứng minh. Since Ψ_α is affine with respect to $\boldsymbol{x} = (x^{\mathrm{I}}, \ldots, x^{\mathrm{K}-1})$, its second derivatives all vanish. Therefore, the Hessian matrix $\nabla^2_{\boldsymbol{x}\boldsymbol{x}} g_\alpha$ coincides with that of the Gibbs function of the ideal gas. But this matrix was shown to be definite positive in Proposition 2.4. We still have to check that the range of the gradient map

$$\nabla_{\boldsymbol{x}} g_\alpha(\boldsymbol{x}) = (\ln(k^{\mathrm{I}} x^{\mathrm{I}}) - \ln(k^{\mathrm{K}} x^{\mathrm{K}}), \ldots, \ln(k^{\mathrm{I}} x^{\mathrm{K}-1}) - \ln(k^{\mathrm{K}} x^{\mathrm{K}})).$$

is equal to $\mathbb{R}^{\mathrm{K}-1}$. For a given $\boldsymbol{u} = (u^{\mathrm{I}}, \dots, u^{\mathrm{K}-1}) \in \mathbb{R}^{\mathrm{K}-1}$, the nonlinear system $\nabla_{\boldsymbol{x}} g_\alpha(\boldsymbol{x}) = \boldsymbol{u}$ can be cast into the $\mathrm{K} \times \mathrm{K}$ linear system

$$\begin{aligned} k^{\mathrm{I}} x^{\mathrm{I}} - \exp(u^{\mathrm{I}}) k^{\mathrm{K}} x^{\mathrm{K}} &= 0, \\ &\vdots \\ k^{\mathrm{K}-1} x^{\mathrm{K}-1} - \exp(u^{\mathrm{K}-1}) k^{\mathrm{K}} x^{\mathrm{K}} &= 0, \\ x^{\mathrm{I}} + \ldots + x^{\mathrm{K}-1} + x^{\mathrm{K}} &= 1. \end{aligned}$$

The first $\mathrm{K}-1$ components of the solution are

$$x^{\mathrm{I}} = \frac{\exp(u^{\mathrm{I}})/k^{\mathrm{I}}}{1/k^{\mathrm{K}} + \sum_{i=\mathrm{I}}^{\mathrm{K}-1} \exp(u^i)/k^i}, \quad \dots, \quad x^{\mathrm{K}-1} = \frac{\exp(u^{\mathrm{K}-1})/k^{\mathrm{K}-1}}{1/k^{\mathrm{K}} + \sum_{i=\mathrm{I}}^{\mathrm{K}-1} \exp(u^i)/k^i}.$$

This defines a unique continuous inverse map $[\nabla_{\boldsymbol{x}} g_\alpha]^{-1} : \mathbb{R}^{\mathrm{K}-1} \to \Omega$. □

3.1.2 Margules' law

We now consider two laws dedicated to binary mixtures, namely, Margules' and Van Laar's [100]. From the viewpoint of physics, these are in reality models for *activity* coefficients of liquid instead fugacity coefficients. However, the mathematical structure of thermodynamic equilibria *via* activity coefficients remains the same [96].

Since $\mathrm{K} - 1 = 1$, we simply write x instead of x^{I} and $\boldsymbol{x} = (x^{\mathrm{I}})$, as was done in §2.4.2. The excess function associated with Margules' law is

$$\Psi_\alpha(x) = x(1-x)\left[A_{12}(1-x) + A_{21}x\right], \tag{3.5}$$

where $(A_{12},\, A_{21}) \in (\mathbb{R}^*)^2$ are two nonzero constants. By (3.2), the fugacity coefficients are

$$\ln \Phi_\alpha^{\mathrm{I}}(x) = \left[A_{12} + 2(A_{21} - A_{12})x\right](1-x)^2, \tag{3.6a}$$
$$\ln \Phi_\alpha^{\mathrm{II}}(x) = \left[A_{21} + 2(A_{12} - A_{21})(1-x)\right]x^2, \tag{3.6b}$$

For a binary mixture, Hypotheses 2.3 are more appropriate than Hypotheses 2.2, as explained in §2.4.2. To meet these requirements, the pair of parameters (A_{12}, A_{21}) must be restricted to some region of $\mathbb{R}^2$. The following result was obtained by Lai Nguyen [76] in his Master's internship at INSA Rennes, during which he joined the PhD team.

Proposition 3.2. *Let $S = A_{12} + A_{21}$ and $D = A_{12} - A_{21}$. Then, the molar Gibbs energy function g_α associated with Margules' law fulfills Hypotheses 2.3 if and only if*

$$S < 4 \quad \textit{and} \quad |D| < \frac{1}{3}\Big[S^2 - 18S + 54 + 2(9-2S)^{3/2}\Big]^{1/2}. \tag{3.7}$$

The "good" region indicated by (3.7) is colored in striped green in Figure 3.2. Its right-most point is located at $(S, D) = (4, 0)$, where it has a vertical tangent.

Chứng minh. We give an abridged version of the proof in [76]. The first derivative of g_α is

$$g'_\alpha(x) = \ln x - \ln(1-x) + A_{12} + (2A_{21} - 4A_{12})x + 3(A_{12} - A_{21})x^2.$$

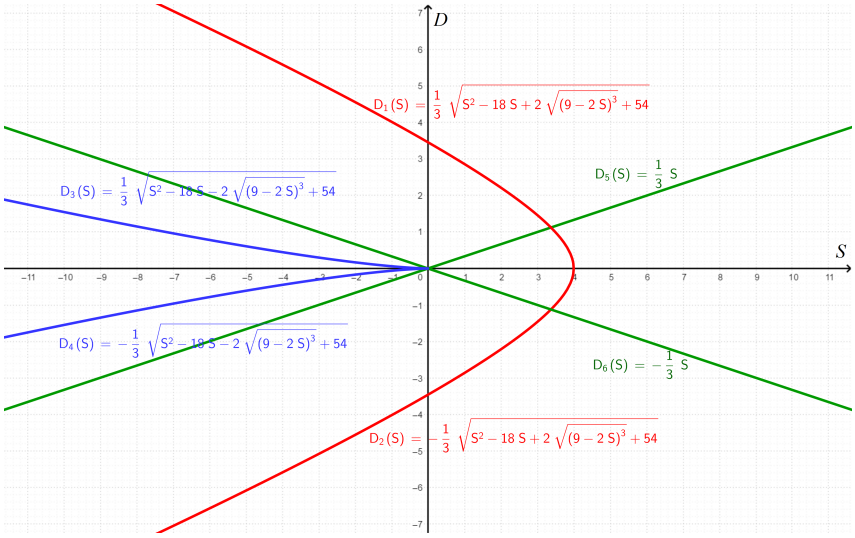


Figure 3.1: Plot of various curves involved in the proof of Proposition 3.2.

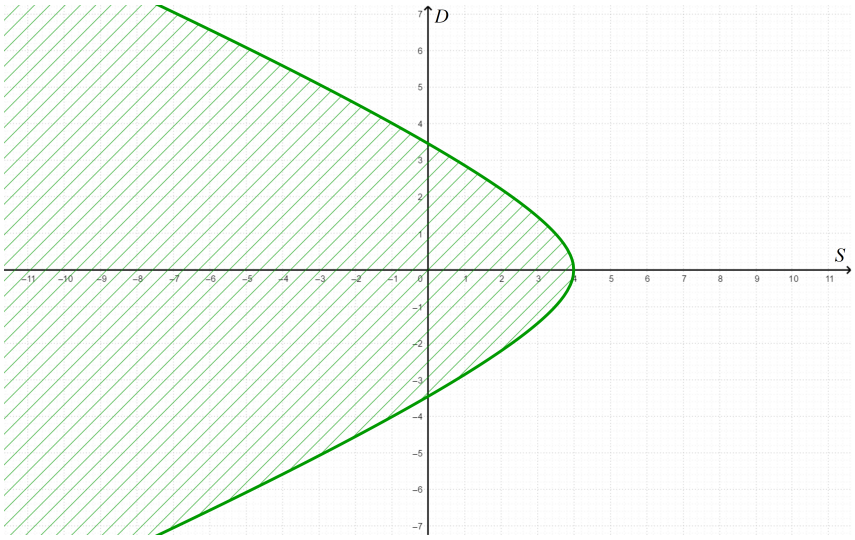


Figure 3.2: Region of strict convexity for the parameters of Margules' law in the (S, D)-plane.

This is a continuous function over $(0,1)$, with

$$\lim_{x\downarrow 0} g'_\alpha(x) = -\infty, \qquad \lim_{x\uparrow 1} g'_\alpha(x) = +\infty.$$

Thus, g'_α has range in $\mathbb{R}$ and can be extended to a surjection from $[0,1]$ to $\{-\infty\} \cup \mathbb{R} \cup \{+\infty\}$.

The second derivative of g_α, multiplied by $x(1-x)$ to remove singularities, is equal to

$$h(x) := x(1-x)g''_\alpha(x) = 1 + (x - x^2)\big[2A_{21} - 4A_{12} + 6(A_{12} - A_{21})x\big].$$

Let us change the variable to $y = x - \frac{1}{2} \in \big(-\frac{1}{2}, \frac{1}{2}\big)$ to work with the more symmetric function

$$H(y) := h\bigg(x - \frac{1}{2}\bigg) = 1 + \bigg(\frac{1}{4} - y^2\bigg)\big[6(A_{12} - A_{21})y - (A_{12} + A_{21})\big].$$

Introducing the sum $S = A_{12} + A_{21}$ and the difference $D = A_{12} - A_{21}$, the above function reads

$$H_{S,D}(y) = 1 + \bigg(\frac{1}{4} - y^2\bigg)(6Dy - S).$$

Our purpose is to look for the region

$$\mathcal{R} = \Big\{\ (S,D) \in \mathbb{R}^2 \mid \min_{[-1/2,1/2]} H_{S,D} > 0\ \Big\}.$$

Note that since $H_{S,-D}(y) = H_{S,D}(-y)$, this region is symmetric with respect to the axis $D = 0$. Therefore, we restrict ourselves to seeking (S,D) such that $D \geqslant 0$. For $D = 0$, if $S > 0$, the function

$$H_{S,0}(y) = 1 - S\bigg(\frac{1}{4} - y^2\bigg)$$

reaches its minimum value at $y = 0$, for which $H_{S,0}(0) = 1 - S/4$; if $S \leqslant 0$, the minimum is achieved on the boundary $y = \pm 1/2$, where $H_{S,0}(\pm 1/2) = 1 > 0$. Therefore, $(S,0) \in \mathcal{R}$ if and only if $S < 4$. Assume now $D > 0$. The derivative

$$H'_{S,D}(y) = D\left(\frac{3}{2} - 18y^2\right) + 2Sy = -18Dy^2 + 2Sy + \frac{3}{2}D$$

is cancelled at the two points

$$y_\pm = \frac{S \pm \sqrt{S^2 + 27D^2}}{18D}.$$

At least one of the two values $y_\pm$ must belong to $(-1/2, 1/2)$, since $H_{S,D}(-1/2) = H_{S,D}(1/2) = 1$ and by Rolle's theorem. More accurately, it is easily proven that: (a) in the subregion $0 < D < -S/3$, only $y_+ \in (-1/2, 1/2)$; (b) in the subregion $D \geqslant |S|/3$, both y_- and y_+ belong to $(-1/2, 1/2)$; (c) in the subregion $0 < D < S/3$, only $y_- \in (-1/2, 1/2)$.

Case (a) can be settled quickly, without calculating $H_{S,D}(y_+)$. Thanks to $0 < D < -S/3$, we have $6Dy - S > 0$ for all $y \in [-1/2, 1/2]$, hence $H'_{S,D} > 0$ on this interval. This entails $H_{S,D}(y) \geqslant H_{S,D}(-1/2) = 1 > 0$. Thus, the subregion $0 < D < -S/3$ is contained inside $\mathcal{R}$. In cases (b) and (c), a more careful inspection involving $H'_{S,D}(1/2) = -3D - S$ and $H'_{S,D}(1/2) = -3D + S$ shows that the minimum of $H_{S,D}$ is achieved at y_-. Let us compute $H_{S,D}(y_-)$ by using not only its value but also the identity

$$y_-^2 = \frac{S}{9D}y_- + \frac{1}{12},$$

which comes from $H'_{S,D}(y_-) = 0$. After simplification, we end up with

$$H_{S,D}(y_-) = 1 - \frac{2}{9}S + \left(D + \frac{S^2}{27D}\right)\frac{S - \sqrt{S^2 + 27D^2}}{18D}.$$

After multiplication by $486D^2$, the requirement $H_{S,D}(y_-) > 0$ is equivalent to

$$(S^2 + 27D^2)^{3/2} < 486D^2 - 81D^2S + S^3. \tag{3.8}$$

The right-hand side can be shown to be positive in cases (b) et (c) and under the additional condition $S < 4$. Indeed, for $S > 0$, $486D^2 - 81D^2S + S^3 = 81D^2(6 - S) + S^3 \geqslant 162D^2 + S^3 > 0$; for $S < 0$, $81D^2(6 - S) + S^3 > 9S^2(6 - S) + S^3 = 2S^2(27 - 4S) > 0$. Note, however, that the condition $S < 4$ is necessary for all points in $\mathcal{R}$. This is because $H_{S,D}(0) = 1 - S/4$ must be positive. Therefore, $S < 4$ can be taken for granted and the inequality (3.8) becomes equivalent to its squared version, i.e., $(S^2 + 27D^2)^3 < (486D^2 - 81D^2S + S^3)^2$. Expanding both sides and simplifying, we obtain

$$(3D)^4 - 2[S^2 - 18S + 54](3D)^2 + S^3(S - 4) < 0.$$

The reduced discriminant of this quadratic inequation in $(3D)^2$ is equal to $4(9 - 2S)^3$. It is positive, since $S < 4$. Then, the solution is given by

$$S^2 - 18S + 54 - 2(9 - 2S)^{3/2} < (3D)^2 < S^2 - 18S + 54 + 2(9 - 2S)^{3/2}. \tag{3.9}$$

From the observation that $S^2 - 18S + 54 + 2(9 - 2S)^{3/2} > 0$ for $S < 4$, we conclude that the second inequality of (3.9) is equivalent to

$$D < \frac{1}{3}\big[S^2 - 18S + 54 + 2(9 - 2S)^{3/2}\big]^{1/2}. \tag{3.10}$$

The upperbound is plotted as the red curve in Figure 3.1. Regarding the first inequality of (3.9), it is equivalent to $S \geqslant 0$ or $S < 0$ and

$$D > \frac{1}{3}\big[S^2 - 18S + 54 - 2(9 - 2S)^{3/2}\big]^{1/2}.$$

The lowerbound is plotted as the blue curve in Figure 3.1. It can be shown that this curve lies inside the region $0 < D < -S/3$ of case (a). Therefore, the previous inequality is trivially satisfied. The last detail to be checked is that the half-line $D = -S/3$ for $S < 0$ is included in the area bounded by the left part of the red curve, so that case (a) is algebraically contained in the desired inequality (3.10). This is left to the reader. □

3.1.3 Van Laar's law

Van Laar's law is also a model for activity coefficients of a liquid [100]. The excess Gibbs function associated with it is

$$\Psi_\alpha(x) = \frac{A_{12}A_{21}x(1 - x)}{A_{12}x + A_{21}(1 - x)}, \tag{3.11}$$

where $(A_{12}, A_{21}) \in (\mathbb{R}^*)^2$ are two nonzero constants. By (3.2), the fugacity coefficients are

$$\ln \Phi^{\mathrm{I}}_\alpha(x) = A_{12}\left[\frac{A_{21}(1 - x)}{A_{12}x + A_{21}(1 - x)}\right]^2, \tag{3.12a}$$

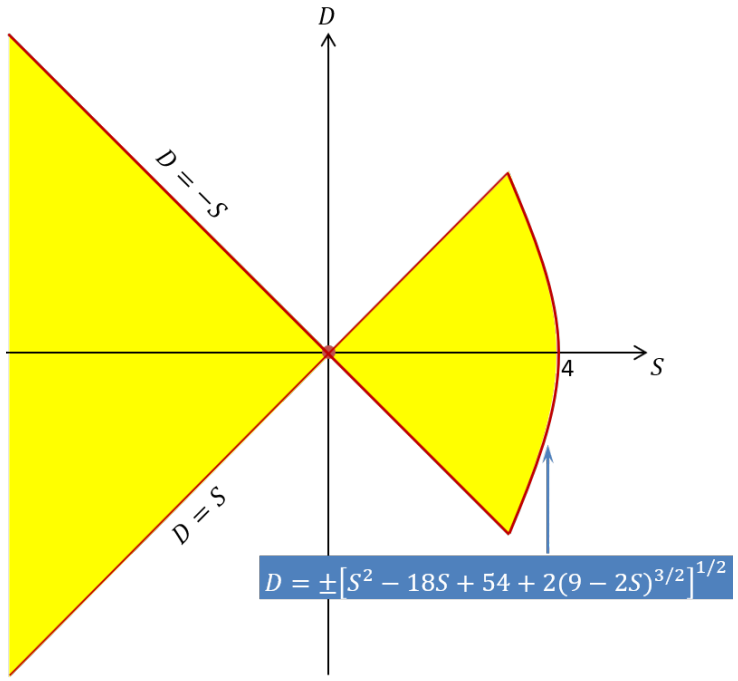


Figure 3.3: Region of strict convexity for the parameters of Van Laar's law.

$$\ln \Phi_\alpha^{\mathrm{II}}(x) = A_{21}\left[\frac{A_{12}x}{A_{12}x + A_{21}(1-x)}\right]^2. \tag{3.12b}$$

To make sure that formulas (3.11)–(3.12) are well-defined over $x \in (0,1)$, the denominator $A_{12}x + A_{21}(1-x)$ must keep the same sign. This amounts to requiring that

$$A_{12}A_{21} > 0. \tag{3.13}$$

In addition to (3.13), the pair of parameters (A_{12}, A_{21}) must be further restricted in order to comply with Hypotheses 2.3. Again, Lai Nguyen [76] obtained the following result in his Master's internship within the PhD team.

Proposition 3.3. *Let $S = A_{12}+A_{21}$ and $D = A_{12}-A_{21}$. Then, the molar Gibbs energy function g_α associated with Van Laar's law fulfills Hypotheses 2.3 if and only if*

$$(S, D) \in \mathcal{R}_- \cup \mathcal{R}_+, \tag{3.14a}$$

where

$$\mathcal{R}_- = \big\{S < 0 \ \ \textit{and} \ \ |D| < -S\big\}, \tag{3.14b}$$

$$\mathcal{R}_+ = \big\{0 < S < 4 \ \ \textit{and} \ \ |D| < \min(S;\, [S^2 - 18S + 54 + 2(9-2S)^{3/2}]^{1/2})\big\}. \tag{3.14c}$$

The "good" region indicated by (3.14) is colored in yellow in Figure 3.3. It lies inside the cone $D^2 < S^2$ that corresponds to condition (3.13). The origin $(0,0)$ must be excluded.

Chứng minh. Although the proof is supplied in [76], we summarize it here, for this part to be self-contained. The first derivative of g_α is

$$g'_\alpha(x) = \ln x - \ln(1-x) + A_{12}A_{21}\frac{A_{21}(1-x)^2 - A_{12}x^2}{[A_{12}x + A_{21}(1-x)]^2}.$$

Under assumption (3.13), this is a continuous function over $(0,1)$, with

$$\lim_{x\downarrow 0} g'_\alpha(x) = -\infty, \qquad \lim_{x\uparrow 1} g'_\alpha(x) = +\infty.$$

Thus, g'_α has range in $\mathbb{R}$ and can be extended to a surjection from $[0,1]$ to $\{-\infty\} \cup \mathbb{R} \cup \{+\infty\}$.

The second derivative of g_α, multiplied by $x(1-x)$ to get rid of singularities, is equal to

$$h(x) := x(1-x)g''_\alpha(x) = 1 - 2A_{12}^2A_{21}^2\frac{x(1-x)}{(A_{12}x + A_{21}(1-x))^3}.$$

Let us change the variable to $y = x - \frac{1}{2} \in \left(-\frac{1}{2}, \frac{1}{2}\right)$ to work with the more symmetric function

$$H(y) := h\left(x - \frac{1}{2}\right) = 1 - 2A_{12}^2A_{21}^2\frac{\frac{1}{4} - y^2}{\left[\frac{1}{2}(A_{12}+A_{21}) + (A_{12}-A_{21})y\right]^3}.$$

Introducing the sum $S = A_{12} + A_{21}$ and the difference $D = A_{12} - A_{21}$, the above function reads

$$H_{S,D}(y) = 1 - (S^2 - D^2)^2\frac{1/4 - y^2}{(S + 2Dy)^3}.$$

Our purpose is to look for the region

$$\mathcal{R} = \left\{ (S,D) \in \mathbb{R}^2 \mid D^2 < S^2 \text{ and } \min_{[-1/2,1/2]} H_{S,D} > 0 \right\},$$

where $D^2 < S^2$ is the expression of (3.13) in terms of (S,D). Note that since $H_{S,-D}(y) = H_{S,D}(-y)$, this region is symmetric with respect to the axis $D = 0$. Therefore, we restrict ourselves to seeking (S,D) such that $D \geqslant 0$. For $D = 0$, if $S > 0$, the function

$$H_{S,0}(y) = 1 - S\left(\frac{1}{4} - y^2\right)$$

reaches its minimum value at $y = 0$, for which $H_{S,0}(0) = 1 - S/4$; if $S \leqslant 0$, the minimum is achieved on the boundary $y = \pm 1/2$, where $H_{S,0}(\pm 1/2) = 1 > 0$. Therefore, $(S,0) \in \mathcal{R}$ if and only if $S \neq 0$ and $S < 4$. Assume now $D > 0$. Divide the upper half-plane $D > 0$ into 3 subregions: (a) $0 < D < -S$; (b) $|S| \leqslant D$; (c) $0 < D < S$. Subregion (b) is ruled out by (3.13). In subregion (a), $S + 2Dy \leqslant S + D < 0$ for all $y \in [-1/2, 1/2]$, so that $H_{S,D}(y) \geqslant 1$ for all $y \in [-1/2, 1/2]$. Thus, the subregion $0 < D < -S$ is a subset of $\mathcal{R}$.

It remains to see what happens in region (c). The derivative

$$H'_{S,D}(y) = (S^2 - D^2)^2\,\frac{-4Dy^2 + 4Sy + 3D}{2(S + 2Dy)^4}$$

is cancelled at the two points

$$y_\pm = \frac{1}{2}\left\{\frac{S}{D} \pm \left[\left(\frac{S}{D}\right)^2 + 3\right]^{1/2}\right\}.$$

At least one of the two values $y_\pm$ must belong to $(-1/2, 1/2)$, since $H_{S,D}(-1/2) = H_{S,D}(1/2) = 1$ and by Rolle's theorem. More accurately, it can be proven that in region (c) where $0 < D < S$, only y_- belongs to $(-1/2, 1/2)$ and this is where $H_{S,D}$ attains its minimum. Let us compute $H_{S,D}(y_-)$ by using not only its value but also the identities

$$y_-^2 = \frac{S}{D}y_- + \frac{3}{4}, \qquad \frac{D + 2Sy_-}{S + 2Dy_-} = \frac{2y_-}{3}, \qquad \frac{1}{4} - y_-^2 = -\frac{Sy_-}{D} - \frac{1}{2} = -\frac{D + 2Sy_-}{2D},$$

which all come from $H'_{S,D}(y_-) = 0$. After simplification, we end up with

$$H_{S,D}(y_-) = 1 + \frac{(S - \sqrt{S^2 + 3D^2})(2S + \sqrt{S^2 + 3D^2})^2}{54D^2}.$$

Then, $H_{S,D}(y_-) > 0$ is equivalent to $54D^2 + (S - \sqrt{S^2 + 3D^2})(2S + \sqrt{S^2 + 3D^2})^2 > 0$. After expansion and cancellations, the inequality can be reduced to

$$9D^2(6 - S) + S^3 > (S^2 + 3D^2)^{3/2}. \tag{3.15}$$

Contrary to the proof of Proposition 3.2 for Margules' law, at this point we are not sure that the left-hand side of (3.15) is positive in region (c), since the additional condition $S < 4$ cannot be proven *a priori* (here $H_{S,D}(0)$ is not as simple as before). However, the positivity of the left-hand side can be checked *a posteriori*, after squaring (3.15) to obtain

$$D^4 - 2(S^2 - 18S + 54)D^2 + S^4 - 4S^3 < 0. \tag{3.16}$$

Arguing in the same fashion as for Margules, with now D instead of $3D$, the above inequation can be turned into

$$D^2 < S^2 - 18S + 54 + 2(9 - 2S)^{3/2}.$$

The right-hand side vanishes for $S = 4$ and is negative for $S > 4$. This implies $S < 4$. The corresponding curve is plotted in red in Figure 3.3. □

3.2 Cubic equations of state from a numerical perspective

The fugacity laws investigated in §3.1 are simple and apply to a selected phase α, regardless of the remaining ones. We are now going to revisit a prominent category of laws for a two-phase (gas and liquid) mixture, in which the fugacity coefficients for both phases are computed in a "simultaneous" way. The coupling between the two phases is achieved through a third-degree polynomial equation, called *cubic equation of state* (EOS). Although there are many interesting physical aspects underlying the design of such laws, our presentation will rather focus on their computational structures and the mathematical issues that arise from their construction.

3.2.1 General principle

An equation of state is a formula that relates the state variables of a system under a given set of conditions, such as pressure, volume and temperature. To understand the mechanism that goes from an EOS to the fugacity coefficients *via* the excess Gibbs energy function, it is useful to first consider the case of a pure component system.

3.2.1.1 For a pure component

The most popular equation of state is probably Boyle-Mariotte's law for an ideal gas

$$\mathsf{P} = \frac{\mathsf{RT}}{\mathsf{V}}, \tag{3.17}$$

where P denotes the pressure, V the molar volume, T the temperature and R the universal gas constant[1]. Several corrections to (3.17) have been attempted in order to better reflect the behavior of a real gas. Let us enumerate a few of them in historical order:

- Van der Waals [114]
$$\mathsf{P} = \frac{\mathsf{RT}}{\mathsf{V}-b} - \frac{a}{\mathsf{V}^2}; \tag{3.18}$$
- Redlich-Kwong-Soave [108, 111]
$$\mathsf{P} = \frac{\mathsf{RT}}{\mathsf{V}-b} - \frac{a}{\mathsf{V}(\mathsf{V}+b)}; \tag{3.19}$$
- Peng-Robinson [99]
$$\mathsf{P} = \frac{\mathsf{RT}}{\mathsf{V}-b} - \frac{a}{\mathsf{V}^2+2\mathsf{V}b-b^2}. \tag{3.20}$$

Each of the relations (3.18)–(3.20) involves a pair of parameters $(a,b) \in (\mathbb{R}_+^*)^2$ that characterize some physical properties of the pure component under study. However, depending on the type of law, these are not the same! For each approximation, a long sequence of additional empirical formulas are usually provided to compute a and b from other quantities such as viscosity, acentricity... In this work, the parameters (a,b) will be considered as fixed constants.

For later purposes, it is convenient to cast (3.17)–(3.20) in a dimensionless form. To this end, let us introduce the dimensionless quantities

$$Z = \frac{\mathsf{PV}}{\mathsf{RT}}, \qquad A = \frac{\mathsf{P}a}{(\mathsf{RT})^2}, \qquad B = \frac{\mathsf{P}b}{\mathsf{RT}}. \tag{3.21}$$

The first quantity Z, called *compressibility factor*, will play a major role in the sequel. The last two quantities $(A,B) \in (\mathbb{R}_+^*)^2$ can be thought of as two dimensionless parameters that characterize the pure component under study at fixed pressure and temperature. In the same spirit as in chapter §2, we shall never write down explicitly the dependency of (A,B) on (P,T). Then, a straightforward calculation shows that equations (3.17)–(3.20) are respectively equivalent to:

- Boyle-Mariotte
$$Z - 1 = 0; \tag{3.22}$$
- Van der Waals
$$Z^3 - (B+1)Z^2 + AZ - AB = 0; \tag{3.23}$$
- Redlich-Kwong-Soave
$$Z^3 - Z^2 + (A - B - B^2)Z - AB = 0; \tag{3.24}$$

[1] $\mathsf{R} = 8.314462618$ S.I.

- Peng-Robinson

$$Z^3 + (B-1)Z^2 + (A - 2B - 3B^2)Z + (B^2 + B^3 - AB) = 0. \tag{3.25}$$

Except for the first equation (3.22), the last three equations (3.23)–(3.25) are cubic polynomials in Z. This is the rationale for the name "cubic EOS."

Given a law and a pair $(A, B) \in (\mathbb{R}_+^*)^2$, let us suppose that the corresponding cubic equation has three distinct real roots, all greater than B. These are then named

$$B < Z_L < Z_I < Z_G. \tag{3.26}$$

In other words, the smallest root is associated with the liquid phase L, while the largest one is associated with the gas phase G. From the physics point of view, at the same pressure and temperature, the gas phase occupies a larger volume the liquid phase, which by (3.21) implies that $Z_G > Z_L$. As for the intermediate root Z_I, it does not have any physical meaning[2]. Like (A, B), the physically significant roots (Z_G, Z_L) can also be viewed as functions of (P, T). This allows us to define

$$\Psi_\alpha = \int_0^{\mathsf{P}} \frac{Z_\alpha(\wp, \mathsf{T}) - 1}{\wp}\, \mathrm{d}\wp, \qquad \alpha \in \{G, L\}, \tag{3.27}$$

which also depend on (P, T). The Ψ_α's are called *excess* molar Gibbs energies, insofar as they measure an integrated amount of non-ideality represented by $Z_\alpha - 1$.

Lemma 3.1. *Under assumption* (3.26) *of three real roots greater than B for the cubic equation of the law considered, the excess molar Gibbs energies Ψ_α, $\alpha \in \{G, L\}$, are given by:*

- *Van der Waals*

$$\Psi_\alpha = Z_\alpha - 1 - \ln\big[Z_\alpha - B\big] - \frac{A}{Z_\alpha}; \tag{3.28}$$

- *Redlich-Kwong-Soave*

$$\Psi_\alpha = Z_\alpha - 1 - \ln\big[Z_\alpha - B\big] - \frac{A}{B}\ln\left[1 + \frac{B}{Z_\alpha}\right]; \tag{3.29}$$

- *Peng-Robinson*

$$\Psi_\alpha = Z_\alpha - 1 - \ln\big[Z_\alpha - B\big] - \frac{A}{2\sqrt{2}B}\ln\left[\frac{Z_\alpha + (\sqrt{2}+1)B}{Z_\alpha - (\sqrt{2}-1)B}\right]. \tag{3.30}$$

Chứng minh. The evaluation of integral (3.27) for the cubic EOS laws (3.23)–(3.25) can be found in standard textbooks such as [104, 115]. □

Let us suppose now that the cubic equation has only one real root greater that B. In this situation, two subcases have to be envisaged. If we manage to assign a "natural" phase label $\alpha = G$ or L to the real root, then the corresponding excess Gibbs energy Ψ_α is defined by (3.27), leaving its counterpart in the other phase undefined. If we do not succeed in attributing a "logical" phase label to the real root, then Ψ_α is undefined in both phases. This process is intuitive enough to describe with words, but raises many serious mathematical questions:

[2] A real root below B is not acceptable either, since b is meant to be the lower limit of the molar volume.

1. When does the cubic equation has three real roots greater than B and when does it have only one real root greater than B?
2. When can a "natural" phase label be assigned to the unique real root greater than B and when is it impossible?

These questions will be answered in §3.2.2 for Van der Waals' law and in §3.2.3 for Peng-Robinson's law. For the moment, let us go on to see how the definition of the excess Gibbs energies Ψ_α carries over to a multicomponent mixture.

3.2.1.2 For a multicomponent mixture

Let us go back to a multicomponent mixture whose set of species is $\mathcal{K} = \{\mathrm{I}, \mathrm{II}, \ldots, \mathrm{K}\}$, with $\mathrm{K} \geqslant 2$. Each component $i \in \mathcal{K}$ is characterized by a pair of parameters $(a^i, b^i) \in (\mathbb{R}_+^*)^2$ within a selected cubic EOS law. At fixed pressure and temperature (P, T), this gives rise to a pair of dimensionless parameters $(A^i, B^i) \in (\mathbb{R}_+^*)^2$ by the last two relations of (3.21).

The basic tenet here is that the multicomponent mixture behaves approximately as a fictitious pure component endowed with an averaged value for the pair (A, B). The latter is computed from the (A^i, B^i)'s and the current partial fractions by means of a *mixing rule*. More specifically, let $\boldsymbol{x} = (x^{\mathrm{I}}, \ldots, x^{\mathrm{K}-1}) \in \overline{\Omega}$ be the partial fractions of the mixture in some phase. We deliberately omit the phase subscript because $\boldsymbol{x}$ is a dummy argument at which we want to compute both $\Psi_G(\boldsymbol{x})$ and $\Psi_L(\boldsymbol{x})$. There can be found a wide variety of mixing rules [74, 96]. The most commonly used one is

$$A(\boldsymbol{x}) = \sum_{i\in\mathcal{K}} \sum_{j\in\mathcal{K}} x^i x^i \sqrt{A^i A^j}\,\big(1 - \kappa^{ij}\big), \tag{3.31a}$$

$$B(\boldsymbol{x}) = \sum_{j\in\mathcal{K}} x^j B^j, \tag{3.31b}$$

where the coefficients $\kappa^{ij} \in [0, 1)$ are coupling parameters and where we remind that x^{K} must be seen as $1 - x^{\mathrm{I}} - \ldots - x^{\mathrm{K}-1}$. In this manuscript, we shall consider the even simpler version where $\kappa^{ij} = 0$, which implies

$$A(\boldsymbol{x}) = \left(\sum_{j\in\mathcal{K}} x^j \sqrt{A^j} \right)^2. \tag{3.32}$$

Whichever the user's favorite mixing rule is, the idea is to plug $A(\boldsymbol{x})$, $B(\boldsymbol{x})$ into the cubic equations (3.23)–(3.25) to get the real roots $Z_\alpha(\boldsymbol{x})$, $\alpha \in \{G, L\}$, should one of these exist and be greater than $B(\boldsymbol{x})$. Then, insert $Z_\alpha(\boldsymbol{x})$ into definition (3.27) in order to obtain $\Psi_\alpha(\boldsymbol{x})$. This amounts, in practice, to directly substituting $Z_\alpha(\boldsymbol{x})$, $A(\boldsymbol{x})$, $B(\boldsymbol{x})$ into formulas (3.28)–(3.30). Finally, apply (3.2) to deduce the fugacity coefficients $\Phi_\alpha^i(\boldsymbol{x})$. In the upcoming subsections §3.2.2 and §3.2.3, we write down the explicit formulas for $\Psi_\alpha(\boldsymbol{x})$ and $\ln \Phi_\alpha^i(\boldsymbol{x})$ and address the two questions asked earlier.

3.2.2 Van der Waals' law

Historically, the Van der Waals equation of state was a major breakthrough compared to the perfect gas equation. Although it is nowadays no longer used in realistic simulations demanding a great physical accuracy, it remains a valuable reference as a "toy model," thanks to its relative mathematical simplicity.

3.2.2.1 Expression of fugacity coefficients

Given a smooth mixing rule that computes the parameters $(A(\boldsymbol{x}),\, B(\boldsymbol{x})) \in (\mathbb{R}_+^*)^2$ from the partial fractions $\boldsymbol{x} \in \overline{\Omega}$, we consider the cubic equation

$$Z^3(\boldsymbol{x}) - [B(\boldsymbol{x}) + 1]Z^2(\boldsymbol{x}) + A(\boldsymbol{x})Z(\boldsymbol{x}) - A(\boldsymbol{x})B(\boldsymbol{x}) = 0, \tag{3.33}$$

which is the multicomponent counterpart of (3.23). Under the same caveats as in the pure component case, let $Z_G(\boldsymbol{x})$ be the greatest real root and $Z_L(\boldsymbol{x})$ the smallest one, should there exist three real roots greater than $B(\boldsymbol{x})$. If there is only one real root greater than $B(\boldsymbol{x})$, let $\alpha \in \{G, L\}$ be the phase possibly assigned to it. The excess molar Gibbs energy is

$$\Psi_\alpha(\boldsymbol{x}) = Z_\alpha(\boldsymbol{x}) - 1 - \ln\left[Z_\alpha(\boldsymbol{x}) - B(\boldsymbol{x})\right] - \frac{A(\boldsymbol{x})}{Z_\alpha(\boldsymbol{x})}. \tag{3.34}$$

Theorem 3.1. *The Van der Waals fugacity coefficients are given by*

$$\begin{aligned}\ln \Phi_\alpha^i(\boldsymbol{x}) = {} & \frac{B(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} B(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B(\boldsymbol{x})} \left[Z_\alpha(\boldsymbol{x}) - 1\right] - \ln\left[Z_\alpha(\boldsymbol{x}) - B(\boldsymbol{x})\right] \\ & + \left[\frac{B(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} B(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B(\boldsymbol{x})} - \frac{2A(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} A(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{A(\boldsymbol{x})}\right] \frac{A(\boldsymbol{x})}{Z_\alpha(\boldsymbol{x})},\end{aligned} \tag{3.35}$$

for all $i \in \mathcal{K}$ and for any phase $\alpha \in \{G, L\}$ in which $Z_\alpha(\boldsymbol{x}) > B(\boldsymbol{x})$ is well-defined.

We recall that the components of $\boldsymbol{\delta}^i = (\delta_{i,1},\, \ldots,\, \delta_{i,\mathrm{K}-1}) \in \mathbb{R}^{\mathrm{K}-1}$ are Kronecker symbols and we stress out that this result is valid for all smooth mixing rules.

Chứng minh. Taking the gradient of (3.34), we have

$$\nabla \Psi_\alpha = \left[1 - \frac{1}{Z_\alpha - B} + \frac{A}{Z_\alpha^2}\right] \nabla Z_\alpha + \frac{1}{Z_\alpha - B} \nabla B - \frac{1}{Z_\alpha} \nabla A,$$

in which we dropped the variable $\boldsymbol{x}$ for clarity. By virtue of the cubic equation (3.33),

$$1 - \frac{1}{Z_\alpha - B} + \frac{A}{Z_\alpha^2} = 0, \qquad \frac{1}{Z_\alpha - B} = \frac{Z_\alpha - 1}{B} + \frac{A}{B Z_\alpha}.$$

Thus,

$$\nabla \Psi_\alpha = \frac{Z_\alpha - 1}{B} \nabla B + \frac{A}{Z_\alpha}\left[\frac{1}{B}\nabla B - \frac{1}{A}\nabla A\right].$$

Applying (3.2) and using (3.34), we arrive at the desired result. □

For the mixing rule (3.31b)–(3.32), let us define the "matrix-vector" product

$$A^i(\boldsymbol{x}) = \sqrt{A^i}\left(\sum_{j \in \mathcal{K}} x^j \sqrt{A^j}\right) \tag{3.36}$$

for all $i \in \mathcal{K}$, in order to state the following result.

Corollary 3.1. *For the mixing rule* (3.31b)–(3.32)*, the Van der Waals fugacity coefficients are given by*

$$\ln \Phi_\alpha^i(\boldsymbol{x}) = \frac{B^i}{B(\boldsymbol{x})}\left[Z_\alpha(\boldsymbol{x}) - 1\right] - \ln\left[Z_\alpha(\boldsymbol{x}) - B(\boldsymbol{x})\right] + \left[\frac{B^i}{B(\boldsymbol{x})} - \frac{2A^i(\boldsymbol{x})}{A(\boldsymbol{x})}\right]\frac{A(\boldsymbol{x})}{Z_\alpha(\boldsymbol{x})}, \tag{3.37}$$

for all $i \in \mathcal{K}$ and for any phase $\alpha \in \{G, L\}$ in which $Z_\alpha(\boldsymbol{x}) > B(\boldsymbol{x})$ is well-defined.

Chứng minh. From (3.31b), we infer that $\nabla B = (B^{\mathrm{I}} - B^{\mathrm{K}}, \ldots, B^{\mathrm{K}-1} - B^{\mathrm{K}})$, so that

$$B(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} B(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^i - \boldsymbol{x}) = \sum_{j=\mathrm{I}}^{\mathrm{K}-1} (B^j - B^{\mathrm{K}})x^j + B^{\mathrm{K}} + \sum_{j=\mathrm{I}}^{\mathrm{K}-1} (B^j - B^{\mathrm{K}})(\delta_{i,j} - x^j) = B^i,$$

the last equality being due to $\sum_{j=\mathrm{I}}^{\mathrm{K}-1} \delta_{i,j} = 1 - \delta_{i,\mathrm{K}}$. By the chain rule, we can check that

$$\frac{\partial A}{\partial x^j}(\boldsymbol{x}) = 2\sum_{i=\mathrm{I}}^{\mathrm{K}-1} [A^i(\boldsymbol{x}) - A^{\mathrm{K}}(\boldsymbol{x})]\delta_{j,i},$$

with $A^i(\boldsymbol{x})$ defined in (3.36). It then follows that

$$\nabla_{\boldsymbol{x}} A(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^i - \boldsymbol{x}) = 2\sum_{j=1}^{\mathrm{K}-1}\sum_{k=1}^{\mathrm{K}-1} [A^i(\boldsymbol{x}) - A^{\mathrm{K}}(\boldsymbol{x})]\delta_{j,k}(\delta_{i,j} - x^j) = 2\sum_{k=1}^{\mathrm{K}-1} [A^k(\boldsymbol{x}) - A^{\mathrm{K}}(\boldsymbol{x})](\delta_{i,k} - x^k).$$

On the other hand,

$$2A(\boldsymbol{x}) = 2\sum_{k=1}^{\mathrm{K}} A^k(\boldsymbol{x})x^k = 2\sum_{k=1}^{\mathrm{K}-1} [A^k(\boldsymbol{x}) - A^{\mathrm{K}}(\boldsymbol{x})]x^k + 2A^{\mathrm{K}}(\boldsymbol{x}).$$

Combining the last two equalities, we end up with

$$2A(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} A(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^i - \boldsymbol{x}) = 2\sum_{k=1}^{\mathrm{K}-1} [A^k(\boldsymbol{x}) - A^{\mathrm{K}}(\boldsymbol{x})]\delta_{i,k} + 2A^{\mathrm{K}}(\boldsymbol{x}) = 2A^i(\boldsymbol{x}).$$

In view of (3.35) [Theorem 3.1], the proof is completed. □

3.2.2.2 Critical parameters, subcritical and supercritical regimes

We now tackle the two questions formulated earlier about the number of real roots greater than B and the assignability of a phase label to a real root, should it be the only one greater than B. Part of these issues is already addressed in [80]. The available answers are always expressed in terms of (P, T). What we wish, however, is to prove results in terms of (A, B), since these dimensionless parameters are more useful to our numerical simulations.

Let $(A, B) \in (\mathbb{R}_+^*)^2$ be a fixed pair of dimensionless parameters. Instead of working with the polynomial

$$\Upsilon_{A,B}(Z) = Z^3 - (B+1)Z^2 + AZ - AB, \tag{3.38}$$

it is more convenient to work with the rational function

$$\Pi_{A,B}(Z) = \frac{1}{Z - B} - \frac{A}{Z^2} - 1 \tag{3.39}$$

over $(B, +\infty)$. As $\Upsilon_{A,B}(Z) = -Z^2(Z-B)\Pi_{A,B}(Z)$, $\Pi_{A,B}$ and $\Upsilon_{A,B}$ have the same roots over $(B, +\infty)$. Since

$$\lim_{Z \downarrow B} \Upsilon_{A,B}(Z) = +\infty, \qquad \lim_{Z \to +\infty} \Upsilon_{A,B}(Z) = -1, \tag{3.40}$$

there is at least one root larger than B. In order to study $\Pi_{A,B}$ more carefully, the following notion will be most helpful.

Definition 3.1 (Critical point)**.** A triplet $(Z_c,\, A_c,\, B_c) \in (B, +\infty) \times (\mathbb{R}_+^*)^2$ is said to be a *critical point* if

$$\Pi_{A_c,B_c}(Z_c) = 0, \qquad \Pi'_{A_c,B_c}(Z_c) = 0, \qquad \Pi''_{A_c,B_c}(Z_c) = 0. \tag{3.41}$$

Conditions (3.41), which are required on $\Pi_{A,B}$ and not $\Upsilon_{A,B}$, mean that the graph of Π_{A_c,B_c} has an inflection point at Z_c, as examplified in Figure 3.4. From the critical triplet (Z_c, A_c, B_c) and from (3.21), it can be deduced the critical pressure, molar volume and temperature

$$\mathsf{P}_c = \frac{a}{b^2}\frac{B_c^2}{A_c}, \qquad \mathsf{V}_c = b\,\frac{Z_c}{B_c}, \qquad \mathsf{T}_c = \frac{a}{b\mathsf{R}}\frac{B_c}{A_c}. \tag{3.42}$$

Lemma 3.2. *For Van der Waals' law, there is a unique critical point given by*

$$Z_c = \frac{3}{8}, \qquad A_c = \frac{27}{64}, \qquad B_c = \frac{1}{8}. \tag{3.43}$$

Chứng minh. The last two conditions of (3.41), i.e., $\Pi'_{A_c,B_c}(Z_c) = \Pi''_{A_c,B_c}(Z_c) = 0$, are equivalent to

$$\frac{1}{(Z_c - B_c)^2} = \frac{2A_c}{Z_c^3}, \qquad \frac{2}{(Z_c - B_c)^3} = \frac{6A_c}{Z_c^4},$$

from which we draw $\frac{1}{2}(Z_c - B_c) = \frac{1}{3}Z_c$ and

$$Z_c = 3B_c. \tag{3.44}$$

Therefore, $A_c = Z_c^3/[2(Z_c - B_c)^2] = 27B_c^3/8B_c^2 = 27B_c/8$ and

$$\frac{B_c}{A_c} = \frac{8}{27}. \tag{3.45}$$

Combining (3.44)–(3.45) with the first condition of (3.41), that is, $\Pi_{A_c,B_c}(Z_c) = 0$, we find $B_c = 1/8$. This results in the critical values given by (3.43). □

Using the critical values, the next statement is a first step in clarifying the behavior of $\Pi_{A,B}$.

Theorem 3.2 (Supercritical and subcritical regimes)**.**

1. *If $B/A > B_c/A_c = 8/27$, the function $\Pi_{A,B}$ is decreasing over $(B, +\infty)$ and has only one zero greater than B.*

2. *If $B/A < B_c/A_c = 8/27$, the function $\Pi_{A,B}$ has two disctinct local extrema. In other words, there exist two distinct values $\zeta_L < \zeta_G$ in $(B, +\infty)$ such that*

$$\Pi'_{A,B}(\zeta_L) = \Pi'_{A,B}(\zeta_G) = 0.$$

Then, $\Pi_{A,B}$ is decreasing on $(B,\, \zeta_L)$, increasing on $(\zeta_L,\, \zeta_G)$ and decreasing on $(\zeta_G, +\infty)$. It may have one or three distinct zeroes over $(B, +\infty)$.

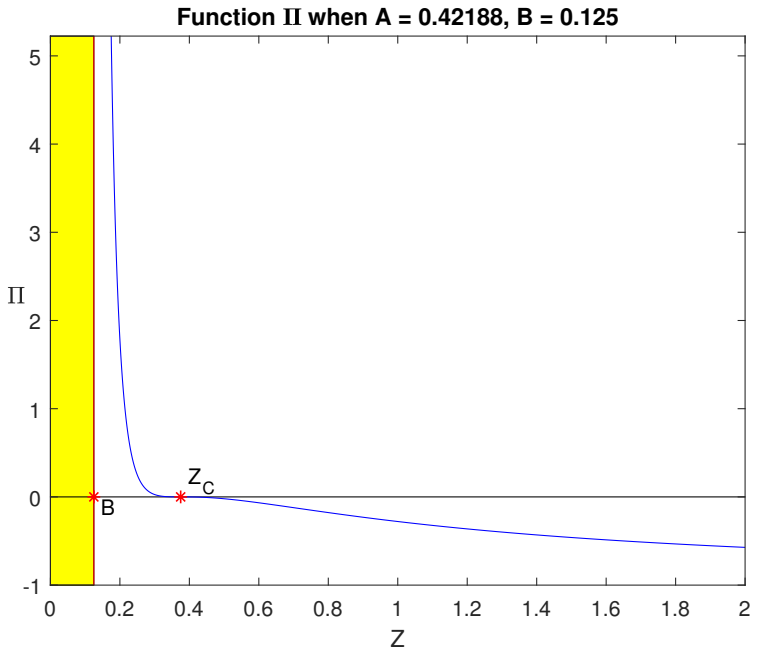


Figure 3.4: At critical values, Π_{A_c,B_c} has an inflection point at Z_c.

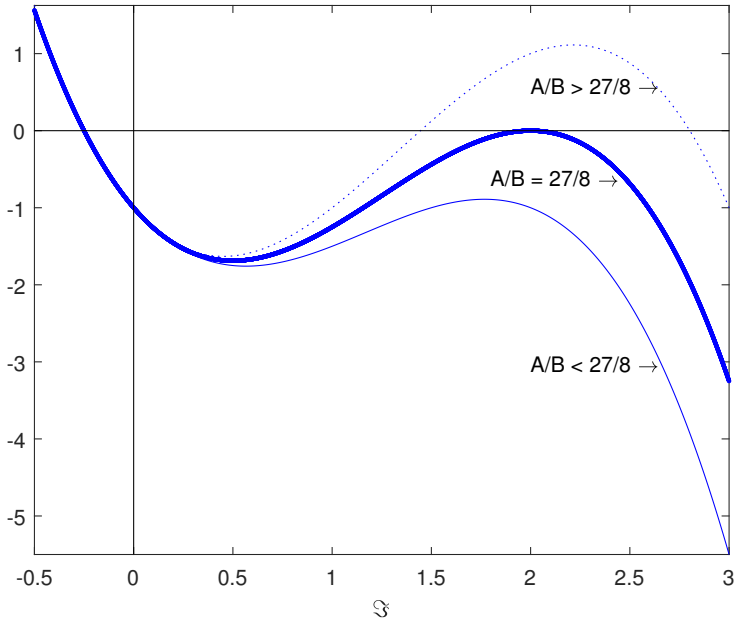


Figure 3.5: Plot of the function $\mathfrak{T} \mapsto q_{A,B}(\mathfrak{T})$ for various values of A/B.

The practical and fundamental interest of Theorem 3.2 lies in the following phase assignment procedure for a root, depending to its location.

Definition 3.2 (Phase label assignment)**.** The region $0 < B < (B_c/A_c)A = (8/27)A$ is said to be *subcritical*. In the subcritical region, a root $Z > B$ of the cubic equation (3.23) is said to be *associated with the liquid phase* L if $Z < \zeta_L$; a root $Z > B$ of the cubic equation (3.23) is said to be *associated with the gas phase* G if $Z > \zeta_G$.

Let us elaborate on this Definition before proving Theorem 3.2. If there is only one root $Z > B$, this root cannot belong to (ζ_L, ζ_G). Therefore, either $Z \in (B, \zeta_L)$ as in Figure 3.7, or $Z \in (\zeta_G, +\infty)$ as in Figure 3.8. This way of assigning a phase label to Z is most natural, since it extends by continuity the "topological" pattern observed in the case of three roots.

The region $B > (B_c/A_c)A = (8/27)$ is said to be *supercritical*. The graph of $\Pi_{A,B}$ no longer has two discernable branches, as shown in Figure 3.9. In this configuration, there is no natural way to associate Z with a phase. We shall not venture into supercritical fluids in this thesis. Physically speaking, the critical threshold B_c/A_c corresponds to a critical temperature T_c by (3.42). Above the critical temperature, the distinction between gas and liquid phases no longer holds [39] and it does not make sense to talk about phase transition.

Chứng minh. (of Theorem 3.2) To find the local extrema of $\Pi_{A,B}$ on $(B, +\infty)$, we search for the zeros on $(B, +\infty)$ of its dervivative

$$\Pi'_{A,B}(Z) = -\frac{1}{(Z-B)^2} + \frac{2A}{Z^3},$$

or equivalently, of the polynomial

$$Q_{A,B}(Z) := Z^3(Z-B)^2\Pi'_{A,B}(Z) = -Z^3 + 2A\,(Z-B)^2\,.$$

An even more convenient choice is to set $\mathfrak{T} = (Z-B)/B \in (0, +\infty)$ and to study

$$q_{A,B}(\mathfrak{T}) := \frac{1}{B^3}\,Q_{A,B}(B\mathfrak{T}+B) = -(\mathfrak{T}+1)^3 + 2\,\frac{A}{B}\,\mathfrak{T}^2.$$

By inserting A_c/B_c, the latter function can be recast as

$$q_{A,B}(\mathfrak{T}) = \left[-(\mathfrak{T}+1)^3 + 2\,\frac{A_c}{B_c}\,\mathfrak{T}^2 \right] + 2\left(\frac{A}{B} - \frac{A_c}{B_c}\right)\mathfrak{T}^2$$

The polynomial in the bracket of the right-hand side, equal to q_{A_c,B_c}, can be factored by $(\mathfrak{T}-2)^2$. This follows from the definition of the critical values, according to which $\mathfrak{T}_c = Z_c/B_c - 1 = 2$ is a double zero of the q_{A_c,B_c}. After using $A_c/B_c = 27/8$ and factoring the bracket, we have

$$q_{A,B}(\mathfrak{T}) = -(\mathfrak{T}-2)^2\left(\mathfrak{T}+\frac{1}{4}\right) + 2\left(\frac{A}{B} - \frac{A_c}{B_c}\right)\mathfrak{T}^2 = q_{A_c,B_c}(\mathfrak{T}) + 2\left(\frac{A}{B} - \frac{A_c}{B_c}\right)\mathfrak{T}^2. \tag{3.46}$$

Note that $q_{A,B}(0) = -1$ and $\lim_{\mathfrak{T}\to+\infty} q_{A,B}(\mathfrak{T}) = -\infty$ for all $(A,B) \in (\mathbb{R}_+^*)^2$. For (A_c, B_c), the graph of q_{A_c,B_c} is tangent to the $\mathfrak{T}$-axis at $\mathfrak{T} = 2$ while taking nonnegative values $q_{A_c,B_c}(\mathfrak{T}) \leqslant 0$ for $\mathfrak{T} \geqslant 0$, as shown in Figure 3.5.

If $A/B > A_c/B_c$, then $q_{A,B}(2) > 0$ and $q_{A,B}$ vanhishes twice on $(0, +\infty)$. If $A/B < A_c/B_c$, then $q_{A,B}(\mathfrak{T}) < q_{A_c,B_c}(\mathfrak{T})$ for all $\mathfrak{T} > 0$ (3.46) and $q_{A,B}$ does not vanish on $(0, +\infty)$. These two cases are also depicted in Figure 3.5. This completes the proof. □

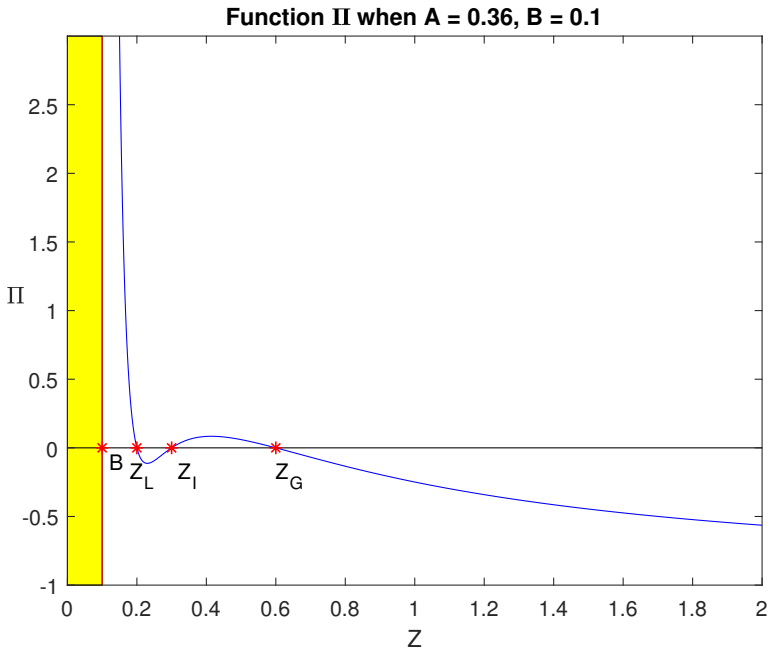


Figure 3.6: 3 roots $B < Z_L < Z_I < Z_G$.

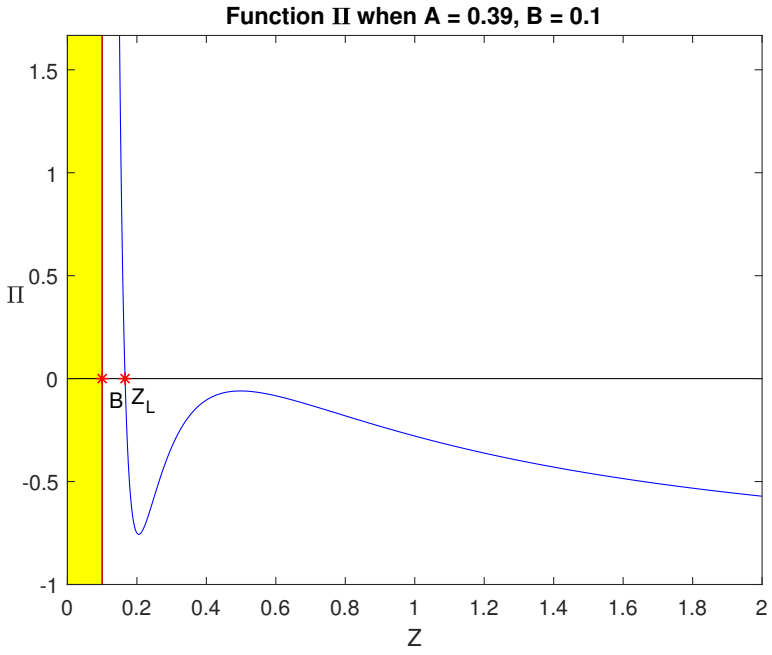


Figure 3.7: 1 subcritical root, assigned to phase L.

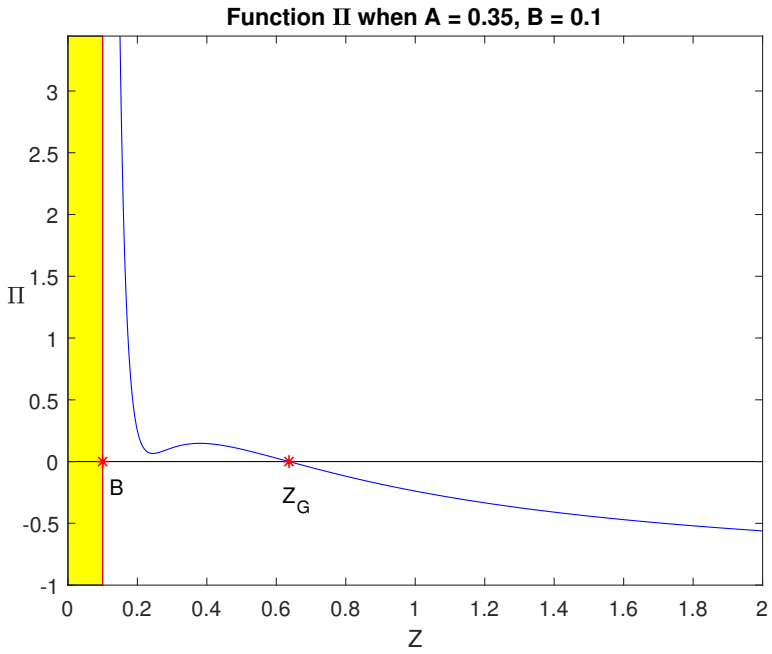


Figure 3.8: 1 subcritical root, assigned to phase G.

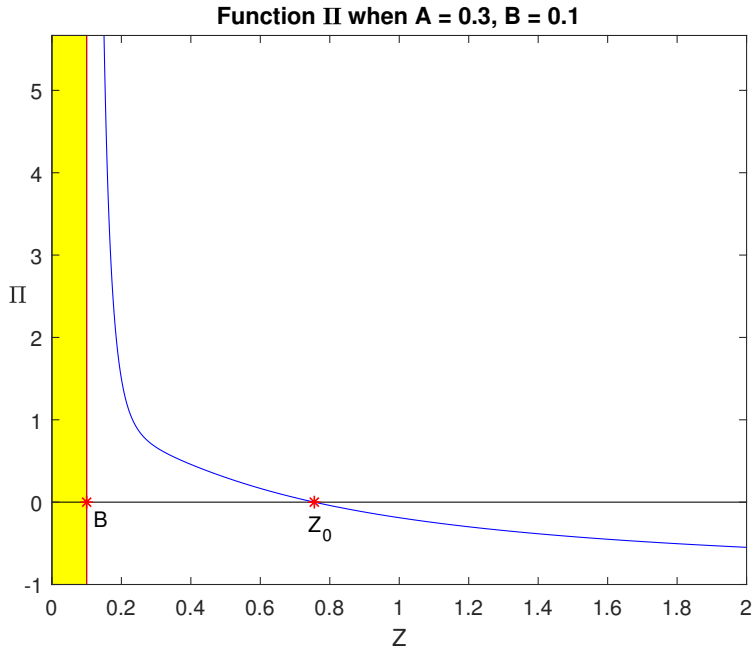


Figure 3.9: 1 supercritical root, not assignable to any phase.

3.2.2.3 Three-root and one-root regions

In terms of (A, B), we are also able to derive a necessary and sufficient condition for the existence of three real roots greater than B. This result does not seem to be well-known in the literature.

Theorem 3.3. *In the quarter-plane $(A, B) \in (\mathbb{R}_+^*)^2$, the region for which Van der Waals' cubic equation (3.23) has three real roots, all greater than B, is determined by*

$$\big\{0 < B < 1/8, \quad A_G(B) < A < A_L(B)\big\}, \tag{3.47a}$$

where

$$A_G(B) = -B^2 + \frac{5}{2}B + \frac{1}{8} - \Big(\frac{1}{4} - 2B\Big)^{3/2}, \tag{3.47b}$$

$$A_L(B) = -B^2 + \frac{5}{2}B + \frac{1}{8} + \Big(\frac{1}{4} - 2B\Big)^{3/2}. \tag{3.47c}$$

This three-root region lies entirely inside the subcritical domain $0 < (27/8)B < A$. Moreover,

- *for $\big\{0 < B < 1/8,\ (27/8)B < A < A_G(B)\big\}$, the only real root is associated with the gas phase G, in the sense of Definition 3.2;*
- *for $\big\{0 < B < 1/8,\ A_L(B) < A\big\}$ or $\big\{1/8 < B,\ (27/8)B < A\big\}$, the only real root is associated with the liquid phase L, in the sense of Definition 3.2.*

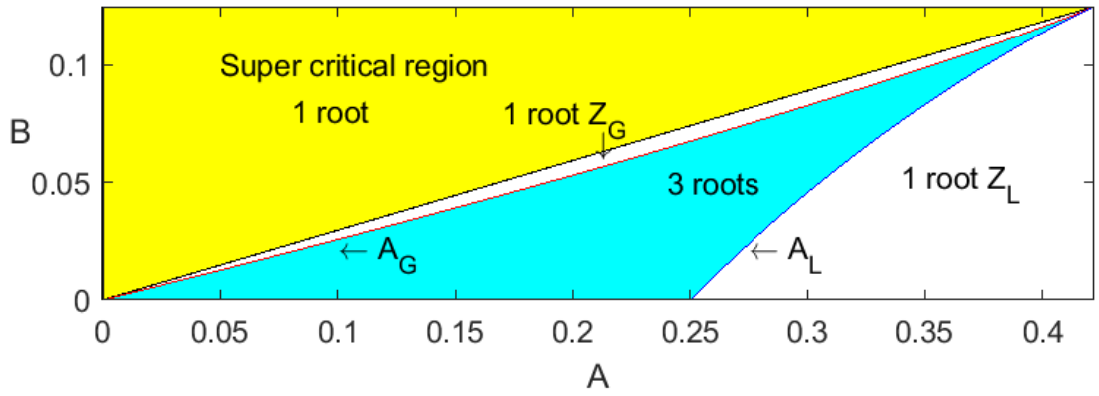


Figure 3.10: Number of roots for Van der Waal's law in the (A, B)-quarter plane.

The three-root region characterized by (3.47) is colored in cyan in Figure 3.10. The first branch $A_G(\cdot)$ starts at $(A, B) = (0, 0)$ with slope $A_G'(B = 0) = 4$. The second branch $A_L(\cdot)$ starts at $(A, B) = (1/4, 0)$ with slope $A_L'(B = 0) = 1$. Both branches end at $(A, B) = (27/64, 1/8)$, with the common slope $A_G'(B = 1/8) = A_L'(B = 1/8) = 9/4$.

Chứng minh. The discriminant of the cubic equation (3.23) is[3]

$$\Delta(A, B) = (B + 1)^2 A^2 - 4A^3 - 4(B + 1)^3 AB - 27A^2B^2 + 18(B + 1)A^2B.$$

Since $A > 0$, we can consider Δ/A and arrange it as a second-degree polynomial in A, that is,

$$\Delta(A, B)/A = -4A^2 + (1 + 20B - 8B^2)A - 4B(B + 1)^3. \tag{3.48}$$

[3]The discriminant of the cubic equation $aX^3 + bX^2 + cX + d = 0$ is $\Delta = b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd$.

For the cubic equation (3.23) to have three distinct real roots, $\Delta(A,B)/A$ must be positive. For this to happen, since its leading coefficient -4 is negative, the quadratic polynmial (3.48) must have two distinct real roots and A must lie between these two roots. But the discriminant of (3.48) with respect to A is

$$\Delta_A(B) = (1 + 20B - 8B^2)^2 - 64B(B+1)^3 = -512B^3 + 192B^2 - 24B + 1 = (1-8B)^3.$$

A necessary and sufficient condition for the quadratic polynomial (3.48) to have two distinct real roots is $0 < B < 1/8$. When this occurs, the two roots of (3.48) are precisely $A_G(B)$ and $A_L(B)$ defined by (3.47b)–(3.47c). Therefore, the region defined in (3.47) characterizes those $(A,B) \in (\mathbb{R}_+^*)^2$ for which Van der Waals' cubic equation (3.23) has three distinct real roots.

Nevertheless, we still have to verify that these three real roots are all greater than B. We already know that at least one of them, say Z_0, is greater than $B > 0$. Since the product of the roots are equal to $AB > 0$, the two remaing roots $Z_1 < Z_2$ must have the same sign. We claim that this common sign cannot be negative. Indeed, let $\Upsilon_{A,B}$ be the Van der Waals polynomial defined in (3.38). Since $\Upsilon_{A,B}(Z_1) = \Upsilon_{A,B}(Z_2) = 0$, there exists by Rolle's theorem $\zeta \in (Z_1, Z_2)$ such that $\Upsilon'_{A,B}(\zeta) = 0$. Assume that $Z_1 < 0$ and $Z_2 < 0$. Then $\zeta < 0$. But then it is obvious that $\Upsilon'_{A,B}(\zeta) = 3\zeta^2 - 2(B+1)\zeta + A > 0$. This is a contradiction.

Next, we claim that the commun sign shared by Z_1 and Z_2 cannot be positive either. For one, we observe that it is not possible to have $Z_1 < B$ and $Z_2 > B$: otherwise, there will be exactly two roots on $(B, +\infty)$, we contradicts what we already know. For another, assume that both Z_1 and Z_2 belong to $(0, B)$. As before, there exists $\zeta \in (Z_1, Z_2) \subset (0, B)$ such that

$$\Upsilon'_{A,B}(\zeta) = 3\zeta^2 - 2(B+1)\zeta + A = 0. \tag{3.49}$$

Since $\Upsilon_{A,B}(0) = -AB < 0$ and $\Upsilon_{A,B}(B) = -B^2 < 0$, we must have $\Upsilon_{A,B}(\zeta) > 0$. Using repeatedly $\zeta^2 = \frac{2}{3}(B+1)\zeta - \frac{1}{3}A$, we have $\zeta^3 = \frac{2}{3}\zeta^2 - \frac{1}{3}A\zeta = [\frac{4}{9}(B+1)^2 - \frac{1}{3}A]\zeta - \frac{2}{3}(B+1)A$, and finally (after some tedious algebra)

$$\Upsilon_{A,B}(\zeta) = -\frac{2}{9}\big[(B+1)^2 - 3A\big]\zeta - 2AB - A.$$

On the other hand, solving the quadratic equation (3.49), we find

$$\zeta = \frac{B + 1 - \sqrt{(B+1)^2 - 3A}}{3}. \tag{3.50}$$

Note that $(B+1)^2 - 3A \geqslant 0$ in the region defined by (3.47) and that we have to select the minus sign in (3.50), as the plus sign is for the other root of $\Upsilon'_{A,B}$ that lies between $Z_2 < B$ and $Z_0 > B$. Pluggin (3.50) into (3.49), we end up with

$$\Upsilon_{A,B}(\zeta) = -\frac{2}{27}\big[(B+1)^2 - 3A\big]\Big\{(B+1) - \big[(B+1)^2 - 3A\big]^{1/2}\Big\} - 2AB - A.$$

The right-hand side is negative, since $B + 1 > [(B+1)^2 - 3A]^{1/2}$. Again, this is a contradiction.

A study of the function $B \mapsto A_G(B) - (27/8)B$ shows that it is positive for $B \in (0, 1/8)$. Hence, the graph of A_G lies inside the subsonic domain. The same is true for $A_L > A_G$. We leave the statements regarding the phase labels of the one-root regions to the readers. □

3.2.3 Peng-Robinson's law

By today's standard, Peng-Robinson's law is the most advanced EOS in terms of accuracy. It is very widely used in industrial codes, including those of IFPEN.

3.2.3.1 Expression of fugacity coefficients

Given a smooth mixing rule that computes the parameters $(A(\boldsymbol{x}),\, B(\boldsymbol{x})) \in (\mathbb{R}_+^*)^2$ from the partial fractions $\boldsymbol{x} \in \overline{\Omega}$, we consider the cubic equation

$$\begin{aligned} &Z^3(\boldsymbol{x}) + (B(\boldsymbol{x}) - 1)Z^2(\boldsymbol{x}) \\ &\quad + [A(\boldsymbol{x}) - 2B(\boldsymbol{x}) - 3B^2(\boldsymbol{x})]Z(\boldsymbol{x}) + [B^2(\boldsymbol{x}) + B^3(\boldsymbol{x}) - A(\boldsymbol{x})B(\boldsymbol{x})] = 0, \end{aligned} \tag{3.51}$$

which is the multicomponent counterpart of (3.23). Let $Z_G(\boldsymbol{x})$ be the greatest real root and $Z_L(\boldsymbol{x})$ the smallest one, should there exist three real roots greater than $B(\boldsymbol{x})$. If there is only one real root greater than $B(\boldsymbol{x})$, let $\alpha \in \{G, L\}$ be the phase possibly assigned to it. The excess molar Gibbs energy is

$$\Psi_\alpha(\boldsymbol{x}) = Z_\alpha(\boldsymbol{x}) - 1 - \ln\left[Z_\alpha(\boldsymbol{x}) - B(\boldsymbol{x})\right] - \frac{A(\boldsymbol{x})}{2\sqrt{2}B(\boldsymbol{x})} \ln\left[\frac{Z_\alpha(\boldsymbol{x}) + (1+\sqrt{2})B(\boldsymbol{x})}{Z_\alpha(\boldsymbol{x}) - (\sqrt{2}-1)B(\boldsymbol{x})}\right]. \tag{3.52}$$

Theorem 3.4. *The Peng-Robinson fugacity coefficients are given by*

$$\begin{aligned} \ln \Phi_\alpha^i(\boldsymbol{x}) = \;& \frac{B(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} B(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B(\boldsymbol{x})} \left[Z_\alpha(\boldsymbol{x}) - 1\right] - \ln\left[Z_\alpha(\boldsymbol{x}) - B(\boldsymbol{x})\right] \\ &+ \left[\frac{B(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} B(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B(\boldsymbol{x})} - \frac{2A(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} A(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{A(\boldsymbol{x})}\right] \\ &\cdot \frac{A(\boldsymbol{x})}{2\sqrt{2}B(\boldsymbol{x})} \ln\left[\frac{Z_\alpha(\boldsymbol{x}) + (1+\sqrt{2})B(\boldsymbol{x})}{Z_\alpha(\boldsymbol{x}) - (\sqrt{2}-1)B(\boldsymbol{x})}\right], \end{aligned} \tag{3.53}$$

for all $i \in \mathcal{K}$ and for any phase $\alpha \in \{G, L\}$ in which $Z_\alpha(\boldsymbol{x}) > B(\boldsymbol{x})$ is well-defined.

Chứng minh. Taking the gradient of (3.52), we have

$$\begin{aligned} \nabla \Psi_\alpha = \;& \left\{1 - \frac{1}{Z_\alpha - B} - \frac{A}{2\sqrt{2}B[Z + (\sqrt{2}+1)B]} + \frac{A}{2\sqrt{2}B[Z - (\sqrt{2}+1)B]}\right\} \nabla Z_\alpha \\ &+ \left\{\frac{1}{Z_\alpha - B} - \frac{A(\sqrt{2}+1)}{2\sqrt{2}[Z + (\sqrt{2}+1)B]} - \frac{A(\sqrt{2}-1)}{2\sqrt{2}[Z - (\sqrt{2}+1)B]} \right. \\ &\qquad \left. + \frac{A}{2\sqrt{2}B^2} \ln\left[\frac{Z_\alpha(\boldsymbol{x}) + (1+\sqrt{2})B(\boldsymbol{x})}{Z_\alpha(\boldsymbol{x}) - (\sqrt{2}-1)B(\boldsymbol{x})}\right]\right\} \nabla B \\ &- \frac{1}{2\sqrt{2}B} \ln\left[\frac{Z_\alpha(\boldsymbol{x}) + (1+\sqrt{2})B(\boldsymbol{x})}{Z_\alpha(\boldsymbol{x}) - (\sqrt{2}-1)B(\boldsymbol{x})}\right] \nabla A, \end{aligned}$$

in which we dropped the variable $\boldsymbol{x}$ for clarity. By virtue of the cubic equation (3.51),

$$1 - \frac{1}{Z_\alpha - B} - \frac{A}{2\sqrt{2}B[Z + (\sqrt{2}+1)B]} + \frac{A}{2\sqrt{2}B[Z - (\sqrt{2}+1)B]} = 0$$

and

$$\frac{1}{Z_\alpha - B} - \frac{A(\sqrt{2}+1)}{2\sqrt{2}[Z + (\sqrt{2}+1)B]} - \frac{A(\sqrt{2}-1)}{2\sqrt{2}[Z - (\sqrt{2}+1)B]} = \frac{Z_\alpha - 1}{B}.$$

Thus,

$$\nabla \Psi_\alpha = \frac{Z_\alpha - 1}{B} \nabla B + \frac{A}{2\sqrt{2}B} \ln\left[\frac{Z_\alpha(\boldsymbol{x}) + (1+\sqrt{2})B(\boldsymbol{x})}{Z_\alpha(\boldsymbol{x}) - (\sqrt{2}-1)B(\boldsymbol{x})}\right] \left\{\frac{1}{B} \nabla B - \frac{1}{A} \nabla A\right\}.$$

Applying (3.2) and using (3.52), we arrive at the desired result. □

Theorem 3.4 is valid for all smooth mixing rules. For the mixing rule (3.31b)–(3.32), and using the notation $A^i(\boldsymbol{x})$ defined in (3.36), we have the following result.

Corollary 3.2. *For the mixing rule* (3.31b)–(3.32)*, the Peng-Robinson fugacity coefficients are given by*

$$\begin{aligned}\ln \Phi^i_\alpha(\boldsymbol{x}) = \; & \frac{B^i}{B(\boldsymbol{x})}\left[Z_\alpha(\boldsymbol{x}) - 1\right] - \ln\left[Z_\alpha(\boldsymbol{x}) - B(\boldsymbol{x})\right] \\ & + \left[\frac{B^i}{B(\boldsymbol{x})} - \frac{2A^i(\boldsymbol{x})}{A(\boldsymbol{x})}\right]\frac{A(\boldsymbol{x})}{2\sqrt{2}B(\boldsymbol{x})}\ln\left[\frac{Z_\alpha(\boldsymbol{x}) + (1+\sqrt{2})B(\boldsymbol{x})}{Z_\alpha(\boldsymbol{x}) - (\sqrt{2}-1)B(\boldsymbol{x})}\right],\end{aligned} \tag{3.54}$$

for all $i \in \mathcal{K}$ and for any phase $\alpha \in \{G, L\}$ in which $Z_\alpha(\boldsymbol{x}) > B(\boldsymbol{x})$ is well-defined.

Chứng minh. Identical to Corollary 3.1. □

3.2.3.2 Critical point, supersonic and subsonic regimes

The questions about the number of roots of Peng-Robinson's cubic equation (3.25) and the assignability of a phase label to a root can be dealt with in the same fashion as in the Van der Waals case, even though the calculations are slightly more technical. Let

$$\Upsilon_{A,B}(Z) = Z^3 + (B-1)Z^2 + (A - 2B - 3B^2)Z + (B^2 + B^3 - AB) \tag{3.55}$$

be the Peng-Robinson polynomial for a fixed pair $(A, B) \in (\mathbb{R}^*_+)^2$. Introduce the rational function

$$\Pi_{A,B}(Z) = \frac{1}{Z-B} - \frac{A}{Z^2 + 2BZ - B^2} - 1, \tag{3.56}$$

obtained from $\Upsilon_{A,B}$ through division by $-(Z-B)(Z^2+2BZ-B^2)$. Insofar as the roots of $Z^2 + 2BZ - B^2$, namely, $-B(\sqrt{2}+1)$ and $B(\sqrt{2}-1)$, are both lesser than B, $\Pi_{A,B}$ and $\Upsilon_{A,B}$ have the same roots over $(B, +\infty)$. Since

$$\lim_{Z \downarrow B} \Upsilon_{A,B}(Z) = +\infty, \qquad \lim_{Z \to +\infty} \Upsilon_{A,B}(Z) = -1, \tag{3.57}$$

there is at least one root larger than B. As in Definition 3.1, a triplet $(Z_c, A_c, B_c) \in (B, +\infty) \times (\mathbb{R}^*_+)^2$ is said to be a *critical point* if

$$\Pi_{A_c,B_c}(Z_c) = 0, \qquad \Pi'_{A_c,B_c}(Z_c) = 0, \qquad \Pi''_{A_c,B_c}(Z_c) = 0. \tag{3.58}$$

Lemma 3.3. *For Peng-Robinson' law, there is a unique critical point given by*

$$Z_c = \frac{1}{32}\left[11 + \sqrt[3]{16\sqrt{2} - 13} - \sqrt[3]{16\sqrt{2} + 13}\right], \tag{3.59a}$$

$$A_c = \frac{1}{512}\left[-59 + 3\sqrt[3]{276831 - 192512\sqrt{2}} + 3\sqrt[3]{276231 + 192512\sqrt{2}}\right], \tag{3.59b}$$

$$B_c = \frac{1}{32}\left[-1 - 3\sqrt[3]{16\sqrt{2} - 13} + 3\sqrt[3]{16\sqrt{2} + 13}\right]. \tag{3.59c}$$

Approximately,

$$Z_c \approx 0.307401308, \qquad A_c \approx 0.457235529, \qquad B_c \approx 0.077796073. \tag{3.59d}$$

Chứng minh. The last two conditions of (3.58), i.e., $\Pi'_{A_c,B_c}(Z_c) = \Pi''_{A_c,B_c}(Z_c) = 0$, are equivalent to

$$(Z_c^2 + 2B_cZ_c - B_c^2)^2 = 2A_c(Z_c + B_c)(Z_c - B_c)^2, \tag{3.60a}$$

$$4(Z_c + B_c)(Z_c^2 + 2B_cZ_c + B_c^2) = 2A_c(Z_c - B_c)(3Z_c + B_c). \tag{3.60b}$$

By eliminating A_c from (3.60), we have

$$4(Z_c - B_c)(Z_c + B_c)^2 = (3Z_c + B_c)^2(Z_c^2 + 2B_cZ_c - B_c^2).$$

Setting $z_c = Z_c/B_c$, the above equation becomes $4(z_c - 1)(z_c + 1)^2 = (3z_c + 1)(z_c^2 + 2z_c - 1)$ and reduces to $z_c^3 - 3z_c^2 - 3z_c - 3 = 0$. The only real root is

$$z_c = 1 + \sqrt[3]{4 - 2\sqrt{2}} + \sqrt[3]{4 + 2\sqrt{2}} \approx 3.951373036. \tag{3.61}$$

Dividing (3.60b) by B_c^3 yields

$$\frac{A_c}{B_c} = \frac{2(z_c + 1)(z_c^2 + 2z_c - 1)}{(z_c - 1)(3z_c + 1)}.$$

Plugging the value (3.61) of z_c into this expression, we obtain

$$\frac{A_c}{B_c} = \frac{1}{16}\left[41 + 3\sqrt[3]{827 - 384\sqrt{2}} + 3\sqrt[3]{827 + 384\sqrt{2}}\right] \approx 5.877359949 \tag{3.62}$$

The first condition of (3.58), i.e., $\Pi_{A_c,B_c}(Z_c) = 0$, reads

$$\frac{1}{B_c(z_c - 1)} - \frac{A_c/B_c}{B_c(z_c^2 + 2z_c - 1)} = 1.$$

Knowing z_c and A_c/B_c from (3.61)–(3.62), we can infer B_c from the previous equation. Once this is done, we can compute $Z_c = z_cB_c$ and $A_c = (A_c/B_c)B_c$ to retrieve (3.59). □

The behavior of $\Pi_{A,B}$ for Peng-Robinson's law is similar to that of Van der Waals' law. Before stating the corresponding theorem, let us remark that by taking the inverse of (3.62), we have

$$\frac{B_c}{A_c} = \frac{1}{16}\left[8 - 3\sqrt[3]{8 + 6\sqrt{2}} + 3\sqrt[3]{-8 + 6\sqrt{2}}\right] \approx 0.170144420 \tag{3.63}$$

Theorem 3.5 (Supercritical and subcritical regimes)**.**

1. *If $B/A > B_c/A_c \approx 0.170144420$, the function $\Pi_{A,B}$ is decreasing over $(B, +\infty)$ and has only one zero greater than B.*

2. *If $B/A < B_c/A_c \approx 0.170144420$, the function $\Pi_{A,B}$ has two disctinct local extrema. In other words, there exist two distinct values $\zeta_L < \zeta_G$ in $(B, +\infty)$ such that*

$$\Pi'_{A,B}(\zeta_L) = \Pi'_{A,B}(\zeta_G) = 0.$$

Then, $\Pi_{A,B}$ is decreasing on (B, ζ_L), increasing on (ζ_L, ζ_G) and decreasing on $(\zeta_G, +\infty)$. It may have one or three distinct zeros over $(B, +\infty)$.

As was the case for Theorem 3.2, Theorem 3.5 paves the way to a natural association of a root with a phase in the subcritical regime.

Definition 3.3 (Phase label assignment)**.** The region $0 < B < (B_c/A_c)A$ is said to be *subcritical*. In the subcritical region, a root $Z > B$ of the cubic equation (3.25) is said to be *associated with the liquid phase* L if $Z < \zeta_L$; a root $Z > B$ of the cubic equation (3.25) is said to be *associated with the gas phase* G if $Z > \zeta_G$.

Let us now prove Theorem 3.5.

Chứng minh. To find the local extrema of $\Pi_{A,B}$ on $(B, +\infty)$, we search for the zeros on $(B, +\infty)$ of its dervivative

$$\Pi'_{A,B}(Z) = -\frac{1}{(Z-B)^2} + \frac{A(2Z+2B)}{(Z^2+2BZ-B^2)^2},$$

or equivalently, of the polynomial

$$\begin{aligned} Q_{A,B}(Z) &:= (Z-B)^2(Z^2+2BZ-B^2)^2\Pi'_{A,B}(Z) \\ &= -(Z^2+2BZ-B^2)^2 + 2A(Z+B)(Z-B)^2. \end{aligned}$$

An even more convenient choice is to set $\mathfrak{T} = (Z-B)/B \in (0, +\infty)$ and to study

$$q_{A,B}(\mathfrak{T}) := \frac{1}{B^4}\, Q_{A,B}(B\mathfrak{T}+B) = -(\mathfrak{T}^2+4\mathfrak{T}+2)^2 + 2\,\frac{A}{B}\,(\mathfrak{T}+2)\mathfrak{T}^2. \tag{3.64}$$

By inserting A_c/B_c, the latter function can be recast as

$$q_{A,B}(\mathfrak{T}) = \left[-(\mathfrak{T}^2+4\mathfrak{T}+2)^2 + 2\,\frac{A_c}{B_c}\,(\mathfrak{T}+2)\mathfrak{T}^2 \right] + 2\left(\frac{A}{B} - \frac{A_c}{B_c}\right)(\mathfrak{T}+2)\mathfrak{T}^2$$

The polynomial in the bracket of the right-hand side, equal to q_{A_c,B_c}, can be factored by $(\mathfrak{T}-\mathfrak{T}_c)^2$, where $\mathfrak{T}_c = z_c - 1$. This follows from the definition of the critical values, according to which $\mathfrak{T}_c = Z_c/B_c - 1 = 2$ is a double zero of the q_{A_c,B_c}. The difficulty here is that, contrary to the Van der Waals case, factorization is not easy to carry out by hand, because A_c/B_c is irrational. To circumvent this difficulty, let us use another technique. Since q_{A_c,B_c} is a fourth-degree polynomial, it is equal to its fourth-order Taylor expansion at $\mathfrak{T} = \mathfrak{T}_c$, that is,

$$\begin{aligned} q_{A_c,B_c}(\mathfrak{T}) &= q_{A_c,B_c}(\mathfrak{T}_c) + q'_{A_c,B_c}(\mathfrak{T}_c)(\mathfrak{T}-\mathfrak{T}_c) \\ &\quad + \tfrac{1}{2}q''_{A_c,B_c}(\mathfrak{T}_c)(\mathfrak{T}-\mathfrak{T}_c)^2 + \tfrac{1}{6}q^{(3)}_{A_c,B_c}(\mathfrak{T}_c)(\mathfrak{T}-\mathfrak{T}_c)^3 + \tfrac{1}{24}q^{(4)}_{A_c,B_c}(\mathfrak{T}_c)(\mathfrak{T}-\mathfrak{T}_C)^4. \end{aligned}$$

In view of $q_{A_c,B_c}(\mathfrak{T}_c) = q'_{A_c,B_c}(\mathfrak{T}_c) = 0$, the factorization sought for is

$$\begin{aligned} q_{A_c,B_c}(\mathfrak{T}) &= \tfrac{1}{24}(\mathfrak{T}-\mathfrak{T}_c)^2\left[12q''_{A_c,B_c}(\mathfrak{T}_c) + 4q^{(3)}_{A_c,B_c}(\mathfrak{T}_c)(\mathfrak{T}-\mathfrak{T}_c) + q^{(4)}_{A_c,B_c}(\mathfrak{T}_c)(\mathfrak{T}-\mathfrak{T}_c)^2\right] \\ &= \tfrac{1}{24}(\mathfrak{T}-\mathfrak{T}_c)^2\left[q_0 + q_1\mathfrak{T} + q_2\mathfrak{T}^2\right], \end{aligned}$$

where the coefficients of the rearrangement in the second line are

$$\begin{aligned} q_0 &= 12q''_{A_c,B_c}(\mathfrak{T}_c) - 4q^{(3)}_{A_c,B_c}(\mathfrak{T}_c)\mathfrak{T}_c + q^{(4)}_{A_c,B_c}(\mathfrak{T}_c)\mathfrak{T}_C^2 \\ q_1 &= 4q^{(3)}_{A_c,B_c}(\mathfrak{T}_c) - 2q^{(4)}_{A_c,B_c}(\mathfrak{T}_c)\mathfrak{T}_c, \\ q_2 &= q^{(4)}_{A_c,B_c}(\mathfrak{T}_c). \end{aligned}$$

If we could prove that the coefficients of the polynomial in the bracket are all negative, i.e., $q_0 < 0$, $q_1 < 0$ and $q_2 < 0$, then it would be plain that $q_{A_c,B_c}(\mathfrak{T}) < 0$ for all $\mathfrak{T} > 0$, except at the double zero $\mathfrak{T} = \mathfrak{T}_c$. Then, the end of the proof would be similar to that of Theorem 3.2. Upon differentiating (3.64) repeatedly, we have

$$\begin{aligned} q''_{A_c,B_c}(\mathfrak{T}_c) &= -4(3\mathfrak{T}_c^2 + 12\mathfrak{T}_c + 10) + 4\frac{A_c}{B_c}(3\mathfrak{T}_c - 2), \\ q^{(3)}_{A_c,B_c}(\mathfrak{T}_c) &= -24(\mathfrak{T}_c + 2) + 12\frac{A_c}{B_c}, \\ q^{(4)}_{A_c,B_c}(\mathfrak{T}_c) &= -24. \end{aligned}$$

By a brute-force calculation relying on the values (3.61)–(3.62) for z_c and A_c/B_c, we end up with $q_0 \approx -11.02105$, $q_1 \approx -437.98968$, $q_2 = -24$. This completes the proof. □

3.2.3.3 Three-root and one-root regions

In terms of (A, B), we are going to derive a necessary (and perhaps sufficient) condition for the existence of three real roots greater than B.

Theorem 3.6. *In the quarter-plane $(A, B) \in (\mathbb{R}_+^*)^2$, the region for which Peng-Robinson's cubic equation (3.25) has three real roots, all greater than B, is contained in the region*

$$\big\{0 < B < B_c, \quad A_G(B) < A < A_L(B)\big\}, \tag{3.65a}$$

where $A_G(B)$ and $A_L(B)$ are respectively the middle root and greatest roots of the cubic equation

$$\begin{aligned} -4A^3 - (8B^2 - 40B - 1)A^2 + (16B^4 - 112B^3 - 88B^2 - 8B)A \\ + (32B^6 + 128B^5 + 160B^4 + 64B^3 + 8B^2) = 0. \end{aligned} \tag{3.65b}$$

The region (3.65) lies itself inside the subcritical domain $0 < B < (B_c/A_c)A$. Moreover,

- *for $\big\{0 < B < B_c,\ (A_c/B_c)B < A < A_G(B)\big\}$, the only real root is associated with the gas phase G, in the sense of Definition 3.3;*
- *for $\big\{0 < B < B_c,\ A_L(B) < A\big\}$ or $\big\{B_c < B,\ (A_c/B_c)B < A\big\}$, the only real root is associated with the liquid phase L, in the sense of Definition 3.3.*

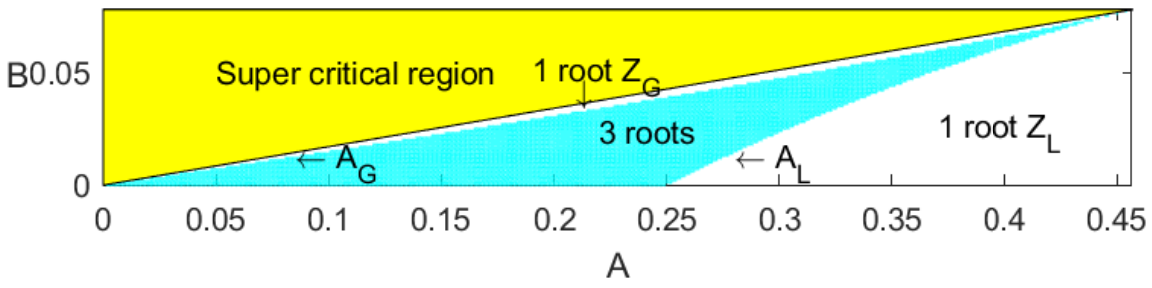


Figure 3.11: Number of roots for Peng-Robinson's law in the (A, B)-quarter plane.

The region characterized by (3.65) is colored in cyan in Figure 3.11. Inside it, Peng-Robinson's cubic equation (3.25) has three real roots. Nevertheless, we could not prove that all the roots

are greater than B, despite abundant numerical evidences supporting the validity of this claim. The first branch $A_G(\cdot)$ starts at $(A, B) = (0, 0)$ with slope $A'_G(B = 0) = 4 + 2\sqrt{2}$. The second branch $A_L(\cdot)$ starts at $(A, B) = (1/4, 0)$ with slope $A'_L(B = 0) = 2$. Both branches end at $(A, B) = (A_c, B_c)$, with the common slope $A'_G(B = B_c) = A'_L(B = B_c) \approx 2.95686087$.

Chứng minh. The discriminant of the cubic equation (3.25) is

$$\Delta(A, B) = -4A^3 - (8B^2 - 40B - 1)A^2 + (16B^4 - 112B^3 - 88B^2 - 8B)A + (32B^6 + 128B^5 + 160B^4 + 64B^3 + 8B^2). \quad (3.66)$$

For the cubic equation (3.25) to have three distinct real roots, $\Delta(A, B)$ must be positive. If the cubic polynomial (3.66) has only one real root $A_0(B)$, since the leading coefficient -4 is negative, we must have $A < A_0(B)$ to ensure $\Delta(A, B) > 0$. If the cubic polynomial (3.66) has three real roots $A_0(B) < A_G(B) < A_L(B)$, we must have $A < A_0(B)$ or $A \in (A_G(B), A_L(B))$. The discriminant of (3.66) with respect to A is equal to

$$\Delta_A(B) = -32B^2(64B^3 + 6B^2 + 12B - 1) \cdot (4096B^6 + 768B^5 + 1572B^4 + 16B^3 + 132B^2 - 24B + 1).$$

It can be shown that $\Delta_A(B) > 0$ for $B \in (0, B_c)$, $\Delta_A(B_c) = 0$[4] and $\Delta_A(B) < 0$ for $B > B_c$. Therefore, if $B > B_c$, only $A_0(B)$ exists. If $B \in (0, B_c)$, there exist $A_0(B) < A_G(B) < A_L(B)$.

Let us show that $A_0(B) > 0$. First, assume $B \in (0, B_c)$. Then, it is easily proven that

$$-8B^2 + 40B + 1 > 0,$$
$$16B^4 - 112B^3 - 88B^2 - 8B < 0,$$
$$32B^6 + 128B^5 + 160B^4 + 64B^3 + 8B^2 > 0.$$

As a consequence, $\Delta(A, B) > 0$ for all $A \leqslant 0$. This implies $A_0(B) > 0$. Next, assume $B > B_c$. Since $\Delta(A = 0, B) > 0$ and $\lim_{A\to+\infty} \Delta(A, B) = -\infty$, $\Delta(\cdot, B)$ has a positive root. But as said earlier, $\Delta_A(B) < 0$ and the only root of $\Delta(\cdot, B)$ must be $A_0(B)$. Hence, $A_0(B) > 0$.

A study of the function $B \mapsto \Delta((A_c/B_c)B, B)$ shows that is is negative for $B \in (0, B_c)$. As $\Delta(0, B) > 0$, this means that for $B \in (0, B_c)$, either $A_0(B)$ is the only root of $\Delta(\cdot, B)$ between 0 and $(A_c/B_c)B$, or the three roots $A_0(B)$, $A_G(B)$, $A_L(B)$ all belong to $(0, (A_c/B_c)B)$. Anyhow, for $A \in (0, A_0(B))$, the point (A, B) lies in the supercritical region where we know by Theorem 3.5 that there is only one real root greater than B for (3.25). Thus, the possibility $A < A_0(B)$ must be ruled out when $B \in (0, B_c)$.

The function $B \mapsto \Delta((A_c/B_c)B, B)$ vanishes at its double root B_c and remains negative for a while, until it vanishes again at $B = B_* \approx 2.435425$ and becomes positive. This means that $A_0(B) > (A_c/B_c)B$ for $B > B_*$ and the graph of $A_0(\cdot)$, now the only root of (3.66), enters the subcritical region. Let $A \in ((A_c/B_c)B, A_0(B))$. At (A, B), there are three real roots for (3.25) and at least one is greater than B. If all of the roots are greater than B, their sum is greater than $3B$. But for (3.25), this sum is equal to $1 - B$. The inequality $1 - B > 3B$ entails $B < 1/4$, which contradicts $B \geqslant B_* \approx 2.435425$.

To summarize, the only way for (3.25) to have three real roots, all greater than B, is that $B \in (0, B_c)$ and $A \in (A_G(B), A_L(B))$. It remains to show that this region is contained in the subcritical domain. Assume that $A_G(B) < (A_c/B_c)B$. In view of the previous discussion

[4] As a matter of fact, $64B^3 + 6B^2 + 12B - 1$ is the minimal polynomial of B_c.

on the number of roots for $\Delta(\cdot, B)$, we must also have $A_L(B) < (A_c/B_c)B$. Then $A_0(B) + A_G(B) + A_L(B) < 3(A_c/B_c)B$. But by (3.66), this sum is equal to $-8B^2 + 40B + 1$. Hence, $-8B^2 + [40 - 3(A_c/B_c)]B + 1 < 0$. But a study of the function $B \mapsto -8B^2 + [40 - 3(A_c/B_c)]B + 1$ reveals that it is positive for all $B \in (0, B_c)$. The statements regarding the phase labels of the one-root regions are left to the readers. □

3.3 Domain extension for cubic EOS-based Gibbs functions

In §3.2, we insisted on the fact that, for a pure component, the cubic equations (3.23)–(3.25) do not always have three real roots greater than B. This implies that, for a multicomponent mixture subject to a given mixing rule, the cubic equations (3.33), (3.51) do not always have three real roots greater than $B(\boldsymbol{x})$ for all $\boldsymbol{x} \in \overline{\Omega}$. As a consequence, the domain of definition for the functions Ψ_α, Φ_α^i for a given phase α does not always cover the whole simplex $\overline{\Omega}$. This physical feature turns out to be detrimental to the unified formulation introduced in §2.2.2.

3.3.1 Trouble ahead

In a nutshell, the molar Gibbs energy functions g_α associated with cubic equations of state grossly violate Hypotheses 2.2. To give a visual picture of the nature of the obstruction, let us consider the simplistic case of a two-phase binary mixture, governed by the mixing rule (3.31b)–(3.32), namely,

$$A(x) = \left[x\sqrt{A^{\mathrm{I}}} + (1-x)\sqrt{A^{\mathrm{II}}}\right]^2, \tag{3.67a}$$

$$B(x) = \; xB^{\mathrm{I}} + (1-x)B^{\mathrm{II}}. \tag{3.67b}$$

The mixture is assumed to obey Van der Waals' law. Thus, for $x \in [0,1]$ and $\alpha \in \{G, L\}$, the value of the excess molar Gibbs energy

$$\Psi_\alpha(x) = Z_\alpha(x) - 1 - \ln\left[Z_\alpha(x) - B(x)\right] - \frac{A(x)}{Z_\alpha(x)}$$

and that of the molar Gibbs energy

$$g_\alpha(x) = x\ln x + (1-x)\ln(1-x) + \Psi_\alpha(x)$$

are defined whenever there exists a real root $Z_\alpha(x)$ of the cubic equation

$$Z^3(x) - [B(x)+1]Z^2(x) + A(x)Z(x) - A(x)B(x) = 0$$

that is greater that $B(x)$ and that can be assigned to phase α. In such a case, we are able to define the fugacity coefficients by

$$\ln\Phi_\alpha^i(x) = \frac{B^i}{B(x)}\left[Z_\alpha(x) - 1\right] - \ln\left[Z_\alpha(x) - B(x)\right] + \left[\frac{B^i}{B(x)} - \frac{2A^i(x)}{A(x)}\right]\frac{A(x)}{Z_\alpha(x)}$$

for the components $i \in \{\mathrm{I}, \mathrm{II}\}$, with

$$A^{\mathrm{I}}(x) = xA^{\mathrm{I}} + (1-x)\sqrt{A^{\mathrm{I}}A^{\mathrm{II}}}, \qquad A^{\mathrm{II}}(x) = x\sqrt{A^{\mathrm{I}}A^{\mathrm{II}}} + (1-x)A^{\mathrm{II}}.$$

For an arbitrary choice of the two pairs $(A^{\mathrm{I}}, B^{\mathrm{I}})$ and $(A^{\mathrm{II}}, B^{\mathrm{II}})$ in the subcritical region $0 < B < (8/27)A$, the parametrized curve $\gamma : [0,1] \ni x \mapsto (A(x), B(x)) \in (\mathbb{R}_+^*)^2$ is an arc

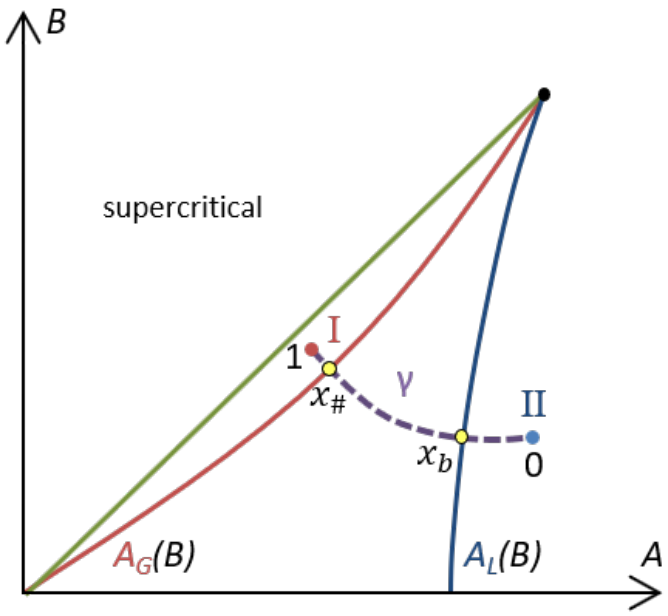


Figure 3.12: Curve γ defined by the mixing rule in the (A, B)-plane.

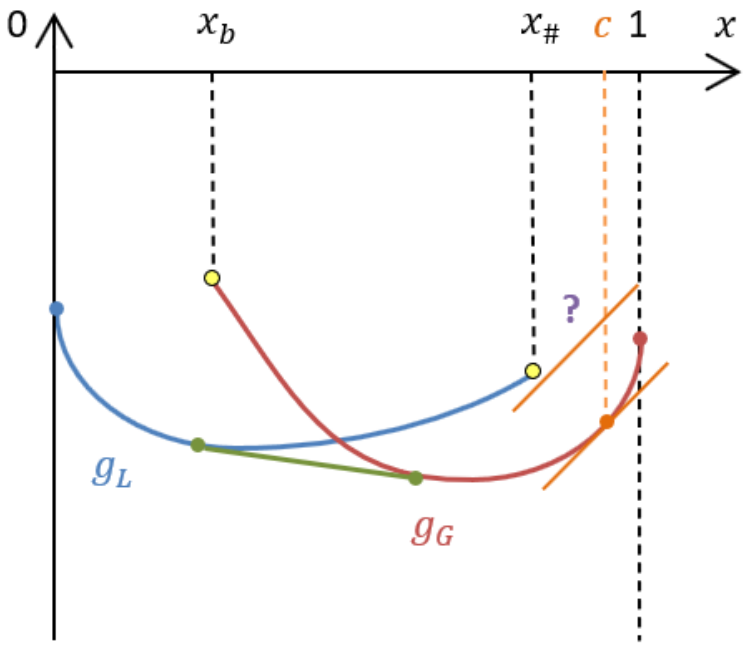


Figure 3.13: Typical situation where the fraction in the absent phase cannot be computed.

of parabola, as illustrated in Figure 3.12. We are not guaranteed that γ remains inside the subcritical region. Even if it does, because we have restricted ourselves to a choice of parameters that is meaningful to physicists, other unfavorable phenomena are likely to occur.

Assume that for $(A^{\mathrm{I}}, B^{\mathrm{I}})$, the Van der Waals cubic equation (3.23) has only one real root greater than B^{I}, associate with phase G. Assume that for $(A^{\mathrm{II}}, B^{\mathrm{II}})$, the Van der Waals cubic equation (3.23) has only one real root greater than B^{II}, associated with phase L. At $x = 0$, the curve γ starts from $(A^{\mathrm{II}}, B^{\mathrm{II}})$ in the L-root region. At some parameter value $x = x_\flat \in (0, 1)$, it enters the three-root region. At some furhter value $x = x_\sharp \in (x_\flat, 1)$, it exits the three-root region. At $x = 1$, it finally meets $(A^{\mathrm{I}}, B^{\mathrm{I}})$ in the G-root region. It is not difficult to realize that:

- the quantities $Z_L(x)$, $\Psi_L(x)$, $g_L(x)$ are well-defined only for $x \in [0, x_\sharp]$; $g_L(x_\sharp^-)$ and $g_L'(x_\sharp^-)$ remain bounded, while $g_L''(x_\sharp^-)$ and $Z_L'(x_\sharp^-)$ blow up; moreover, there is no guarantee that g_L is strictly convex over $[0, x_\sharp]$;
- the quantities $Z_G(x)$, $\Psi_G(x)$, $g_G(x)$ are well-defined only for $x \in [x_\flat, 1]$; $g_G(x_\flat^+)$ and $g_G'(x_\flat^+)$ remain bounded, while $g_G''(x_\flat^+)$ and $Z_G'(x_\flat^+)$ blow up; moreover, there is no guarantee that g_G is strictly convex over $[x_\flat, 1]$.

Since $g_G'(x_\flat^+)$ and $g_L'(x_\sharp^-)$ are finite, the image sets $g_G'([x_\flat, 1))$ and $g_L'((0, x_\sharp])$ are not equal to $\mathbb{R}$. This prevents us from assigning a correct value to the fractions of a vanishing phase. Indeed, according to Gibbs' geometric construction described in Theorem 2.5, when the global composition c is sufficiently close to 0, the solution of system (2.83) is in the single phase L, with $\bar{Y} = 0$, $\bar{\xi}_L = \bar{x}_L = c$. But as $\lim_{x \downarrow 0} g_L'(x) = -\infty$, it is expected that $g_L'(c) \notin g_G'([x_\flat, 1))$. In other words, it is impossible to find $\bar{x}_G \in [x_\flat, 1)$ such that $g_G'(\bar{x}_G) = g_L'(c)$. Likewise, when the global composition c is sufficiently close to 1, the solution of system (2.83) is in the single phase G, with $\bar{Y} = 1$, $\bar{\xi}_G = \bar{x}_G = c$. But as $\lim_{x \uparrow 1} g_G'(x) = +\infty$, it is expected that $g_G'(c) \notin g_L'((0, x_\sharp])$. In other words, it is impossible to find $\bar{x}_L \in (0, x_\sharp]$ such that $g_L'(\bar{x}_L) = g_G'(c)$. The latter situation is depicted in Figure 3.13.

It could be argued that the same flaws of cubic EOS laws should cause the same prejudice to the natural variable (or variable-switching) formulation of §2.2.1. Nothing could be further from the truth. In the variable-switching formulation, if the context is correctly guessed, we do not need to compute anything from the absent phase and the above problem is irrelevant. If the context is incorrectly alleged, the flash does not converge or may even crash, but there is an opportunity for us to make up for it by changing the context. The natural variable formulation does not seek to fathom the dark, invisible and uncharted side of the vanishing phases. The unified formulation has to do so, by its very vocation to treat all phases on an equal footing.

To give the unified formulation a fighting chance, it is essential that the domains of definition for the excess functions Ψ_α's be properly extended to $\overline{\Omega}$. By "properly," we mean that the corresponding extended Gibbs energy functions g_α fulfill Hypotheses (2.2). If strict convexity is too difficult to satisfy, at least we should require surjectivity of the extended gradient maps $\nabla_{\boldsymbol{x}} g_\alpha$ from Ω onto $\mathbb{R}^{\mathrm{K}} - 1$.

3.3.2 Direct method for binary mixture

For the two-phase binary mixture considered in §3.3.1, a natural workaround is to differentiably extend g_G over $[0, x_\flat)$ and g_L over $(x_\sharp, 1]$, in such a way that $g_G'([0, 1]) = g_L'([0, 1]) = \{-\infty\} \cup \mathbb{R} \cup \{+\infty\}$, together with strict convexity of g_G and g_L over $[0, 1]$. More accurately, let $\omega > 0$ be a small width such that

$$0 < x_\flat + \omega < x_\sharp - \omega < 1.$$

Over the domain $[0,1]$, we propose to extend the excess Gibbs functions by

$$\Psi_L[\omega](x) = \begin{cases} \Psi_L(x) & \text{if } x \in [0, x_\sharp - \omega], \\ \Psi_{L,\omega}(x) & \text{if } x \in [x_\sharp - \omega, 1], \end{cases} \tag{3.68a}$$

$$\Psi_G[\omega](x) = \begin{cases} \Psi_{G,\omega}(x) & \text{if } x \in [0, x_\flat + \omega], \\ \Psi_G(x) & \text{if } x \in [x_\flat + \omega, 1], \end{cases} \tag{3.68b}$$

in which the "artificial" parts are defined by the second-order Taylor expansions

$$\Psi_{L,\omega}(x) = \Psi_L(x_\sharp - \omega) + \Psi'_L(x_\sharp - \omega)[x - (x_\sharp - \omega)] + \frac{1}{2}\Psi''_L(x_\sharp - \omega)[x - (x_\sharp - \omega)]^2, \tag{3.69a}$$

$$\Psi_{G,\omega}(x) = \Psi_G(x_\flat + \omega) + \Psi'_G(x_\flat + \omega)[x - (x_\flat + \omega)] + \frac{1}{2}\Psi''_G(x_\flat + \omega)[x - (x_\flat + \omega)]^2. \tag{3.69b}$$

The reason why we cannot take $\omega = 0$ is that Ψ''_L blows up at $x_\sharp^-$ and Ψ''_G blows up at $x_\flat^+$. From the extended excess functions (3.68)–(3.69), we can deduce the extended Gibbs energies by applying (2.31), i.e.,

$$g_\alpha[\omega](x) = x \ln x + (1-x)\ln(1-x) + \Psi_\alpha[\omega](x),$$

for $\alpha \in \{G, L\}$. We can also infer the extended fugacity coefficients by applying (2.33a), i.e.,

$$\ln \Phi^{\mathrm{I}}_\alpha[\omega](x) = \Psi_\alpha[\omega](x) + (1-x)\Psi'_\alpha[\omega](x), \tag{3.70a}$$

$$\ln \Phi^{\mathrm{II}}_\alpha[\omega](x) = \Psi_\alpha[\omega](x) - x\Psi'_\alpha[\omega](x), \tag{3.70b}$$

for $\alpha \in \{G, L\}$. This direct approach enjoys the following property.

Proposition 3.4. *Assume that the original Gibbs energy functions g_L and g_G are strictly convex on their respective intervals of definition $[0, x_\sharp]$ and $[x_\flat, 1]$. Then, for all $\omega > 0$ small enough, their extended versions $g_L[\omega]$ and $g_G[\omega]$ fulfill Hypotheses 2.3.*

Chứng minh. The proof of this Proposition is very easy and is left to the readers. □

Figures 3.14–3.15 examplify the direct method of extension for two 4-tuple (A^{I}, B^{I}, A^{II}, B^{II}) and two parameters ω. Figures 3.16–3.17 provide a close-up comparison between the extended Gibbs functions and their derivatives for two width parameters ω.

3.3.3 Indirect method for multicomponent mixtures

The extension strategy developed in §3.3.2 is suitable for a binary mixture. For a multicomponent mixture, it appears to be most impractical: (i) we have to know in advance the boundary of the three-root region in Ω; (ii) we have to construct C^2-hypersurfaces starting this boundary and joining $\partial\Omega$. we need another extension for the model with more than two components.

Instead of working with $\boldsymbol{x} \in \Omega$, it is more judicious to work with $Z_\alpha(\boldsymbol{x}) \in \mathbb{R}$. When the cubic equation does not have three real roots greater than B, our idea is to use the arithmetic mean of the two other roots, which is also their common real part if these are complex conjugates. In place of the undefined $Z_\alpha(\boldsymbol{x})$, we plug this "surrogate" value into the expression of the excess Gibbs function for the missing phase α. From this ansatz, the fugacity coefficients can be derived accordingly.

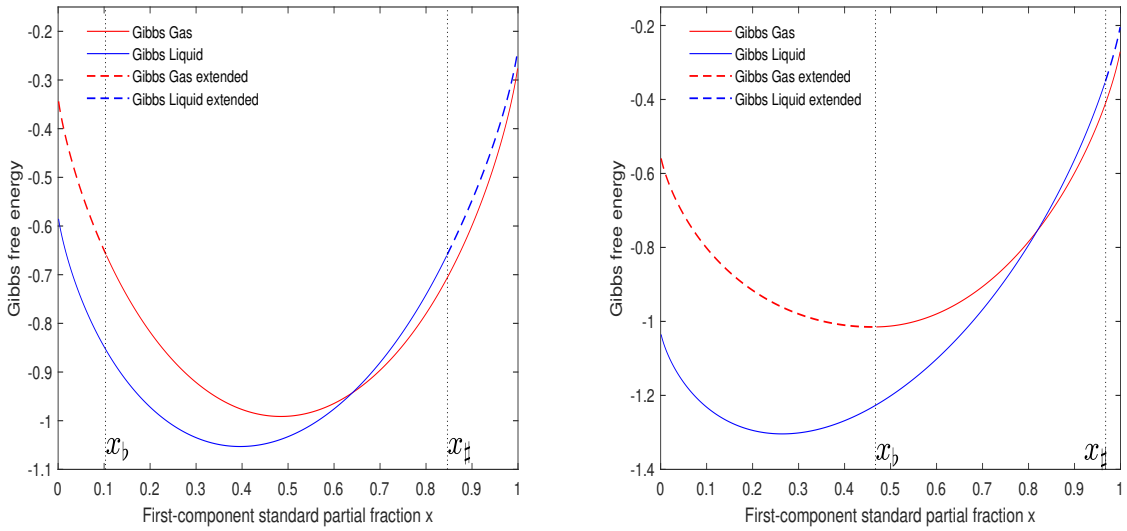


Figure 3.14: Extended Gibbs energy functions $g_L[\omega]$ (blue) and $g_G[\omega]$ (red) for Van der Waals' law by the direct method, with $\omega = 0.001$. Left panel: $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.33, 0.0955)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.35, 0.08)$. Right panel: $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.32, 0.09)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.37, 0.072)$.

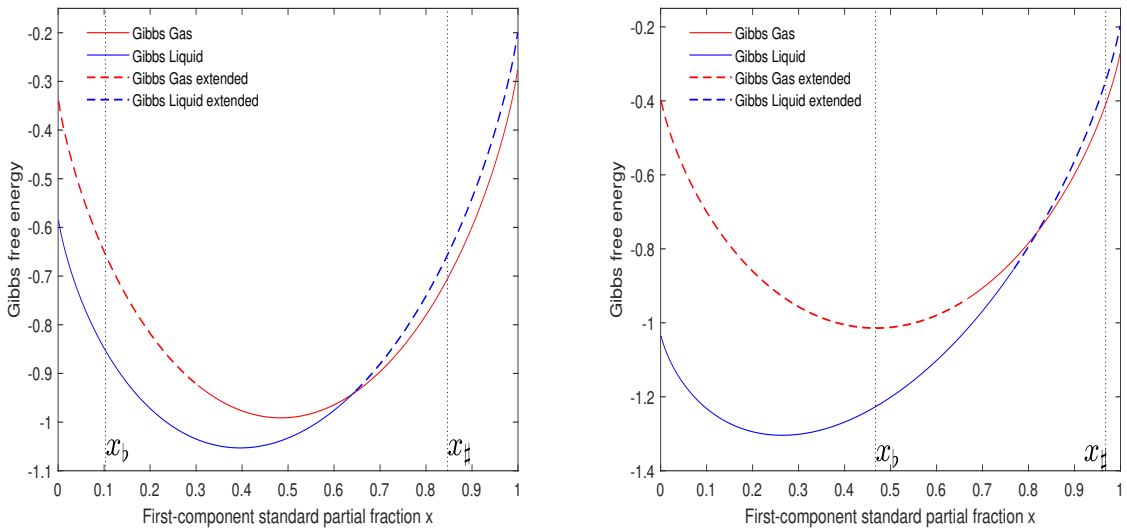
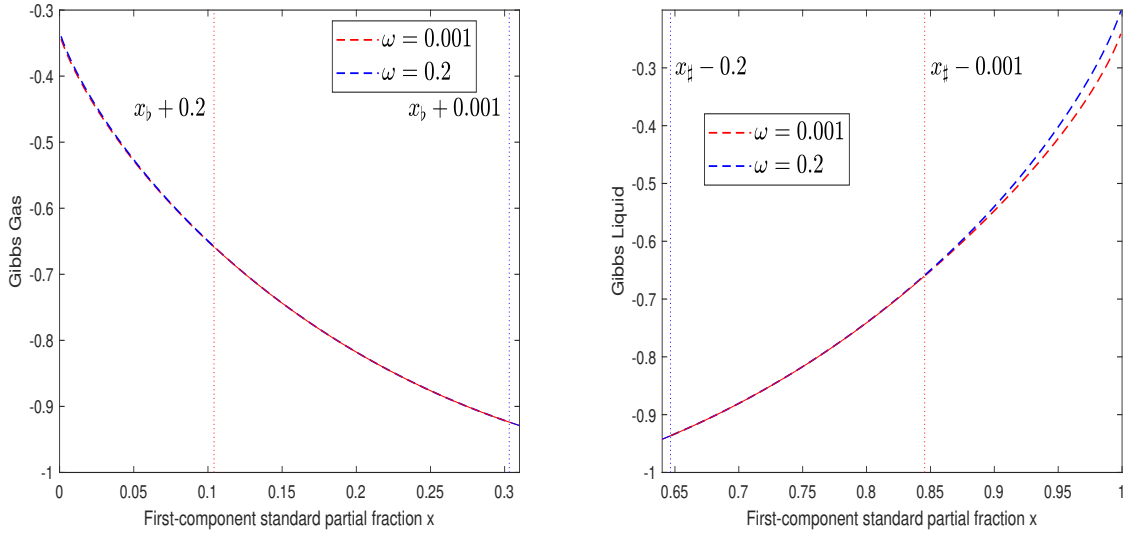
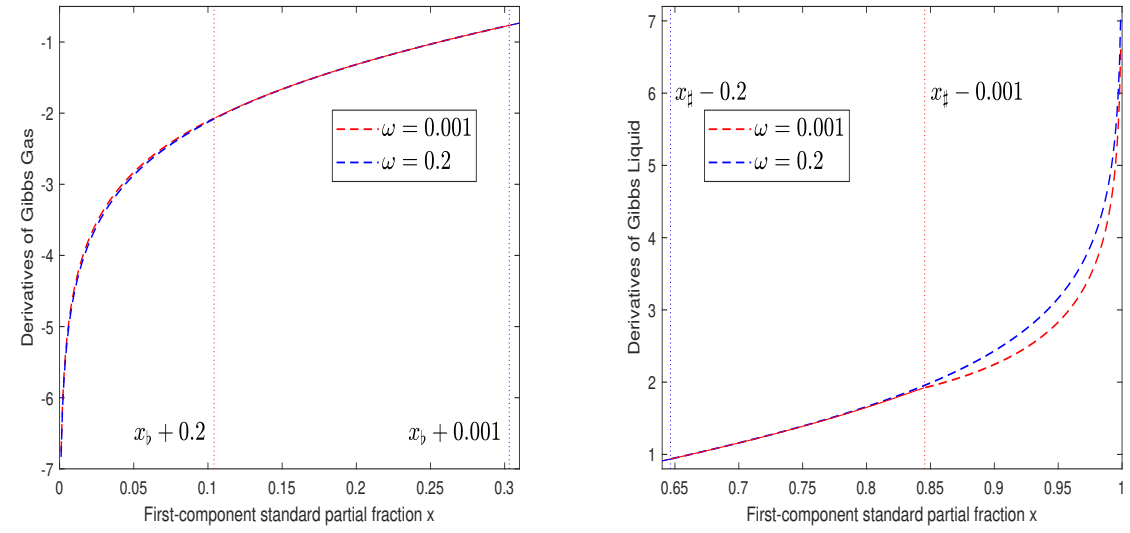


Figure 3.15: Extended Gibbs energy functions $g_L[\omega]$ (blue) and $g_G[\omega]$ (red) for Van der Waals' law by the direct method, with $\omega = 0.2$. Left panel: $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.33, 0.0955)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.35, 0.08)$. Right panel: $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.32, 0.09)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.37, 0.072)$.



(a) g_G (b) g_L

Figure 3.16: Close-up comparison of the extended Gibbs functions between $\omega = 0.001$ and $\omega = 0.2$ for Van der Waals' law with the direct method. $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.33, 0.0955)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.35, 0.08)$.



(a) g'_G (b) g'_L

Figure 3.17: Close-up comparison of the derivative of the extended Gibbs funcions between $\omega = 0.001$ and $\omega = 0.2$ for Van der Waals' law with the direct method. $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.33, 0.0955)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.35, 0.08)$.

3.3.3.1 For Van der Waals' law

Let us explain the idea on Van der Waals' cubic equation

$$Z^3 - (B+1)Z^2 + AZ - AB = 0.$$

For convenience, we do not explicitly indicate the dependency of A, B and Z on $\boldsymbol{x}$.

Construction in the one-root region. We assume that there is only one real root greater than B and that this root can be assigned a natural phase label $\alpha \in \{G, L\}$ in the sense of Definition 3.2, so that we can write it as Z_α. Let β be the other phase, that is, $\beta = L$ if $\alpha = G$ and $\beta = G$ if $\alpha = L$. If the two remaining roots of the cubic equation are complex conjugates, their common real part is

$$W_\beta = \frac{B + 1 - Z_\alpha}{2}, \tag{3.71}$$

since the sum of the three roots must be equal to $B+1$. In any case, W_β defined in (3.71) is the arithmetic mean of the two "bad" roots. The following favorable properties of W_β help convince us that it can be used as a substitute for Z_β, which does not exist.

Lemma 3.4. *Let (A, B) be a pair in the subcritical region $0 < B < (8/27)A$ and assume that Van der Waal's cubic equation has only one real root $Z_\alpha > B$ that corresponds to phase α.*

1. *If $B < \frac{1}{16}(3\sqrt{33} - 11) \approx 0.389605496$, then*

$$W_\beta > B. \tag{3.72a}$$

2. *If $B < \frac{1}{4}(9\sqrt{57} - 67) \approx 0.237127479$, then*

$$Z_\alpha < W_\beta \quad \text{if } \alpha = L, \qquad W_\beta < Z_\alpha \quad \text{if } \alpha = G. \tag{3.72b}$$

Chứng minh. In view of (3.71), the condition $W_\beta > B$ is tantamount to $1 - B > Z_\alpha$. This implies $1 - B > B$ and $B < 1/2$. Using the rational function

$$\Pi_{A,B}(Z) = \frac{1}{Z - B} - \frac{A}{Z^2} - 1$$

introduced in (3.39) and in light of Theorem 3.2 about its behavior, the condition $1 - B > Z_\alpha$ is itself equivalent to $\Pi_{A,B}(1 - B) < 0$. But

$$\Pi_{A,B}(1 - B) = \frac{1}{1 - 2B} - \frac{A}{(1 - B)^2} - 1 < 0$$

can be reduced after simplification to

$$A > \frac{2B(1 - B)^2}{1 - 2B}.$$

In the subsonic region, $A > (27/8)B$. The sufficient condition $(27/8)B > 2B(1 - B)^2/(1 - 2B)$ is satisfied for all $B \in (0, \frac{1}{16}(3\sqrt{33} - 11))$.

In view of (3.71), the condition $Z_\alpha \lesseqgtr W_\beta$ is tantamount to $Z_\alpha \lesseqgtr \frac{1}{3}(B+1)$. Assuming $(B+1)/3 > B$, that is, $B < 1/2$, the previous equality is also equivalent to $\Pi_{A,B}((B+1)/3) \lesseqgtr 0$. But

$$\Pi_{A,B}\left(\frac{B+1}{3}\right) = \frac{3}{1-2B} - \frac{9A}{(B+1)^2} - 1 \lesseqgtr 0$$

can be simplified to

$$A \gtreqless \frac{2(B+1)^3}{9(1-2B)}. \tag{3.73}$$

By studying the function defined in the right-hand side, we can show that for all $B \in (0, 1/8)$,

$$A_G(B) < \frac{2(B+1)^3}{9(1-2B)} < A_L(B). \tag{3.74}$$

The three curves meet at the critical point $(A_c, B_c) = (27/64, 1/8)$ where they have a common tangent.

Let us assume first that $\alpha = G$. This occurs only if $B \in (0, 1/8)$ and $(27/8)B < A < A_G(B)$. By (3.74), we have (3.73) with the "<" sign, which implies $Z_G > W_L$. Let us assume now that $\alpha = L$. This occurs only if: (i) $B \in (0, 1/8)$ and $A > A_L(B)$, or (ii) $B > 1/8$ and $A > (27/8)B$. In case (i), we have (3.73) with the ">" sign, which implies $Z_L < W_G$. In case (ii), notice that $2B(1-B)^2/(1-2B) < (27/8)B$ for all $B \in (1/8, \frac{1}{4}(9\sqrt{57}-67))$, so that for B in this range we still have (3.73) with the ">" sign and can reach the same conclusion. □

From now on, we shall restrict ourselves to $B < (9\sqrt{57}-67)/4$. Physically speaking, this is a reasonable assumption, since $B_c = 1/8$ is almost twice smaller. When Z_α is the only real root greater than B for Van der Waals' cubic equation, the excess Gibbs energy Ψ_α is defined for phase α by the usual formula (3.34). For the other phase β, we stipulate that

$$\Psi_\beta = W_\beta - 1 - \ln[W_\beta - B] - \frac{A}{W_\beta}, \tag{3.75}$$

which is well-defined thanks to Lemma 3.4. This is what we refer to as the "indirect" extension of the excess Gibbs energy Ψ_β when the root Z_β no longer exists. When applying (3.2) to (3.75) in order to derive the fugacity coefficients, we need to be careful.

Theorem 3.7. *When the indirect extension* (3.75) *is applied to phase β, the Van der Waals fugacity coefficients in this phase are given by*

$$\begin{aligned}\ln \Phi^i_\beta = {} & \frac{B + \nabla B \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B}[W_\beta - 1] - \ln[W_\beta - B] \\ & + \left[\frac{B + \nabla B \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B} - \frac{2A + \nabla_{\boldsymbol{x}} A \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{A}\right]\frac{A}{W_\beta} \\ & + \left[\frac{\nabla W_\beta \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{W_\beta} - \frac{\nabla B \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B}\right]\frac{\Upsilon_{A,B}(W_\beta)}{W_\beta(W_\beta - B)}\end{aligned} \tag{3.76}$$

for all $i \in \mathcal{K}$, with $\Upsilon_{A,B}(W) = W^3 - (B+1)W^2 + AW - AB$ as defined in (3.38).

Chứng minh. The proof is similar to that of Theorem 3.1, except for the fact that now

$$1 - \frac{1}{W_\beta - B} + \frac{A}{W_\beta^2} = \frac{\Upsilon_{A,B}(W_\beta)}{W_\beta^2(W_\beta - B)}, \qquad \frac{1}{W_\beta - B} = \frac{W_\beta - 1}{B} + \frac{A}{BW_\beta} - \frac{\Upsilon_{A,B}(W_\beta)}{W_\beta(W_\beta - B)},$$

instead of being 0. □

In (3.76), we need the gradient of W_β with respect to $\boldsymbol{x}$. After (3.71),

$$\nabla W_\beta = \frac{1}{2}(\nabla B - \nabla Z_\alpha).$$

The gradient of Z_α with respect to $\boldsymbol{x}$ can be obtained by differentiating Van der Waals' cubic equation. This operation yields

$$[3Z_\alpha^2 - 2(B+1)Z_\alpha + A]\nabla Z_\alpha = (B - Z_\alpha)\nabla A + (A + Z_\alpha^2)\nabla B,$$

from which ∇Z_α can be extracted, since Z_α is a simple root and $3Z_\alpha^2 - 2(B+1)Z_\alpha + A \neq 0$.

Alteration in the three-root region. From the one-root region, let us move towards the transition boundary where a new real root Z_β appears. In the one-root region, we only have the notion of the "generalized" root W_β, whose gradient ∇W_β remains well-defined. If we start from the three-root region and move towards the transition boundary where Z_β disappears, the gradient ∇Z_β does not remain bounded. Indeed, as

$$[3Z_\beta^2 - 2(B+1)Z_\beta + A]\nabla Z_\beta = (B - Z_\beta)\nabla A + (A + Z_\beta^2)\nabla B,$$

and as Z_β gets closer to being a double root, ∇Z_β blows up. However, we need a finite gradient ∇Z_β for the numerical resolution of system (2.77) by, say, the Newton method. Such a finite gradient is indeed required in the lines of the Jacobian matrix corresponding to the equalities of fugacity (2.77b). To achieve a smooth junction between the two regions, we accept to "sacrifice" a tiny portion of the three-root region. Let us assume that we are in the three-root region, with $B < Z_L < Z_I < Z_G$. We introduce

$$\vartheta = \frac{Z_I - Z_L}{Z_G - Z_L} \in [0, 1] \tag{3.77}$$

as an indicator of the closeness to the transition boundary. Indeed, the cubic equation has double roots when $\vartheta = 0$ or $\vartheta = 1$. Let $\varepsilon \in (0, 1/4)$ be a small threshold.

- If $\vartheta \in [2\varepsilon, 1 - 2\varepsilon]$, we apply the usual formulas for the case of three real-roots.
- If $\vartheta \in (1 - 2\varepsilon, 1]$, the two roots Z_I and Z_G are close to each other. We keep Z_L but progressively replace Z_G by

  $$W_G = \frac{B + 1 - Z_L}{2} = \frac{Z_I + Z_G}{2}, \tag{3.78}$$

  whose gradient is bounded. Instead of plugging Z_G into formula (3.34) for Ψ_G, we insert

  $$\widetilde{Z}_G = [1 - \nu_G(\vartheta)]Z_G + \nu_G(\theta)W_G, \tag{3.79}$$

  where

  $$\nu_G(\vartheta) = \begin{cases} 0 & \text{if } \vartheta \leqslant 1 - 2\varepsilon, \\ q\left(\dfrac{\vartheta - (1 - 2\varepsilon)}{\varepsilon}\right) & \text{if } \vartheta \in (1 - 2\varepsilon, 1 - \varepsilon), \\ 1 & \text{if } \vartheta \geqslant 1 - \varepsilon, \end{cases} \tag{3.80a}$$

  $$q(y) = y^2(3 - 2y). \tag{3.80b}$$

The rescaled function $y \mapsto q(y/\varepsilon)$ serves as a C^1 step function over the interval $[0, \varepsilon]$. We note that $q(0) = 0$, $q(1) = 1$ and $q'(0) = q'(1) = 0$. From the modified excess Gibbs energy

$$\Psi_G = \widetilde{Z}_G - 1 - \ln[\widetilde{Z}_G - B] - \frac{A}{\widetilde{Z}_G}, \tag{3.81a}$$

we can derive by (3.2) the fugacity coefficients

$$\begin{aligned} \ln \Phi_G^i = {} & \frac{B + \nabla B \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B} [\widetilde{Z}_G - 1] - \ln[\widetilde{Z}_G - B] \\ & + \left[\frac{B + \nabla B \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B} - \frac{2A + \nabla_{\boldsymbol{x}} A \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{A} \right] \frac{A}{\widetilde{Z}_G} \\ & + \left[\frac{\nabla \widetilde{Z}_G \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{\widetilde{Z}_G} - \frac{\nabla B \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B} \right] \frac{\Upsilon_{A,B}(\widetilde{Z}_G)}{\widetilde{Z}_G(\widetilde{Z}_G - B)}. \end{aligned} \tag{3.81b}$$

The gradient $\nabla \widetilde{Z}_G$ in the above formula can be approximated by

$$\nabla \widetilde{Z}_G = \begin{cases} \nabla Z_G & \text{if } \vartheta \leqslant 1 - 2\varepsilon, \\ [1 - \nu_G(\vartheta)] \nabla Z_G + \nu_G(\vartheta) \nabla W_G & \text{if } \vartheta \in (1 - 2\varepsilon, 1 - \varepsilon), \\ \nabla W_G & \text{if } \theta \geqslant 1 - \varepsilon, \end{cases} \tag{3.81c}$$

where $\nabla W_G = \frac{1}{2}(\nabla B - \nabla Z_L)$ and where the derivatives of ν_G are neglected.

- If $\vartheta \in [0, 2\varepsilon)$, we proceed in a similar and symmetric fashion to replace Z_L by $\widetilde{Z}_L = [1 - \nu_L(\vartheta)] Z_L + \nu_L(\vartheta) W_L$ in the expression of Ψ_L, while preserving Z_G.

Figures 3.18–3.19 display a few examples of the indirect method for the Van der Waals case. Figures 3.20–3.21 provide a close-up comparison between two choices of ε. It can be seen that ε has little influence on the extended Gibbs functions for the gas. For the liquid, this influence is more apparent.

3.3.3.2 For Peng-Robinson's law

We go through the same process as in the Van der Waals case. Assume that Z_α is the only real root greater than B of Peng-Robinson's cubic equation

$$Z^3 + (B - 1)Z^2 + (A - 2B - 3B^2)Z + (B^2 + B^3 - AB) = 0.$$

Construction in the one-root region. As Z_α is associated with phase α, let β be the other phase and let us introduce the arithmetic mean of the two remaining roots

$$W_\beta = \frac{1 - B - Z_\alpha}{2}, \tag{3.82}$$

which is their common real part when these are complex conjugates. We refer the readers to (3.59) [Lemma 3.3] and (3.62) for the critical values A_c, B_c for Peng-Robinson's law.

Lemma 3.5. *Let (A, B) be a pair in the subcritical region $0 < B < (B_c/A_c)A$ and assume that Peng-Robinson's cubic equation has only one real root $Z_\alpha > B$ that corresponds to phase α.*

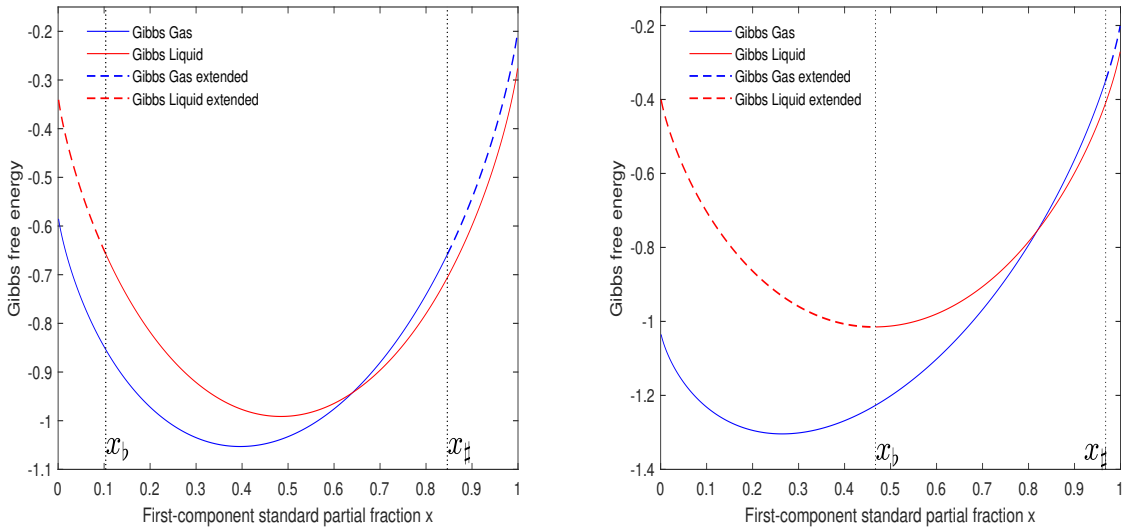


Figure 3.18: Extended Gibbs energy functions g_L (blue) and g_G (red) for Van der Waals' law by the indirect method, with $\varepsilon = 0.001$. Left panel: $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.33, 0.0955)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.35, 0.077)$. Right panel: $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.32, 0.09)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.37, 0.072)$.

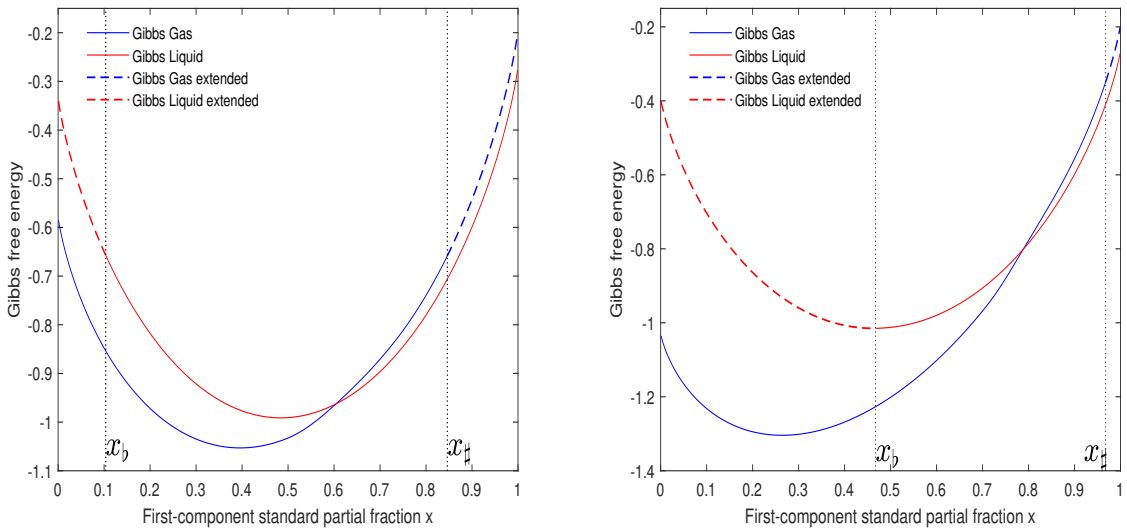
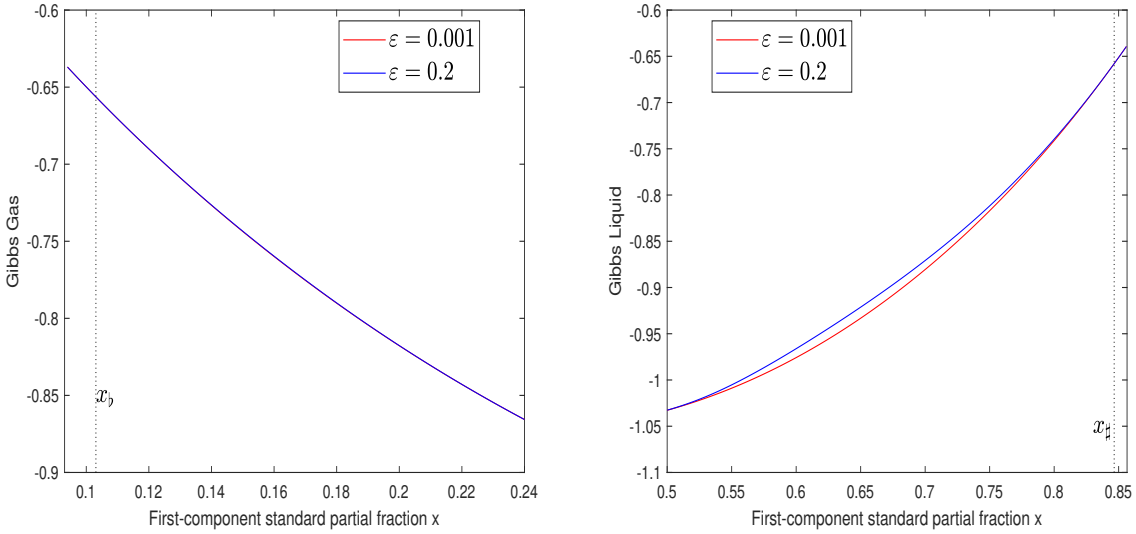
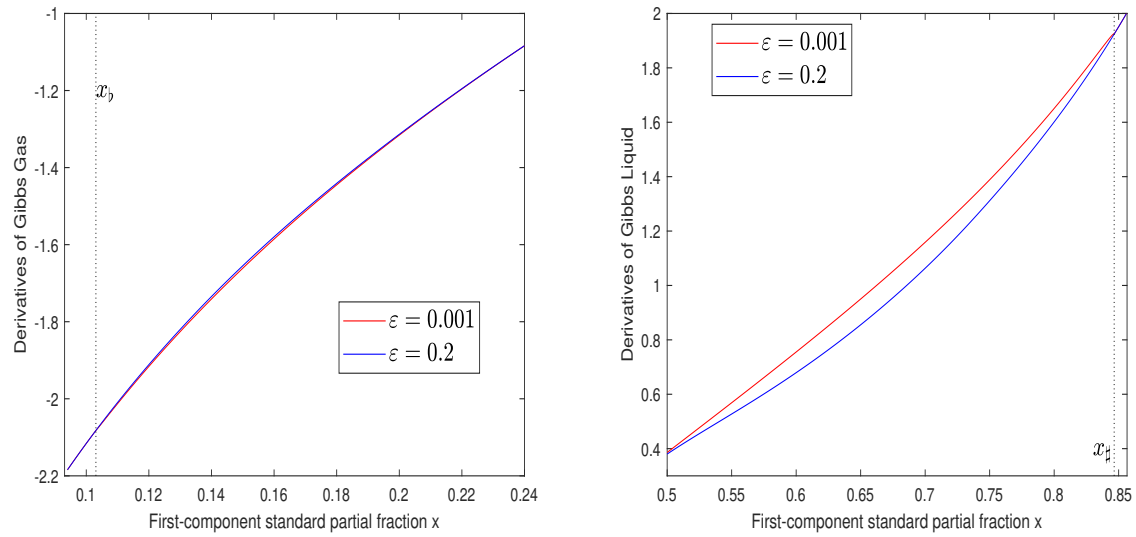


Figure 3.19: Extended Gibbs energy functions g_L (blue) and g_G (red) for Van der Waals' law by the indirect method, with $\varepsilon = 0.2$. Left panel: $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.33, 0.0955)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.35, 0.077)$. Right panel: $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.32, 0.09)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.37, 0.072)$.



(a) g_G (b) g_L

Figure 3.20: Close-up comparison of the extended Gibbs functions between $\varepsilon = 0.001$ and $\varepsilon = 0.2$ for Van der Waals' law with the indirect method. $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.33, 0.0955)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.35, 0.08)$.



(a) g'_G (b) g'_L

Figure 3.21: Close-up comparison of the derivative of the extended Gibbs functions between $\varepsilon = 0.001$ and $\varepsilon = 0.2$ for Van der Waals' law with the indirect method. $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.33, 0.0955)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.35, 0.08)$.

1. *If* $B < 0.206813$*, then*

$$W_\beta > B. \tag{3.83a}$$

2. *If* $B < 0.137072$*, then*

$$Z_\alpha < W_\beta \quad \textit{if} \ \ \alpha = L, \qquad\qquad W_\beta < Z_\alpha \quad \textit{if} \ \ \alpha = G. \tag{3.83b}$$

Chứng minh. The proof is similar to that of Lemma 3.4. □

By restricing ourselves to $B < 0.137072$, which is reasonable since $B_c \approx 0.077796$, we can rely on Lemme 3.5 to stipulate that

$$\Psi_\beta = W_\beta - 1 - \ln\big[W_\beta - B\big] - \frac{A}{2\sqrt{2}B}\,\ln\left[\frac{W_\beta + (\sqrt{2}+1)B}{W_\beta - (\sqrt{2}-1)B}\right]. \tag{3.84}$$

for the missing phase β. By virtue of (3.2), we can derive the corresponding fugacity coefficients.

Theorem 3.8. *When the indirect extension* (3.75) *is applied to phase* β*, the Peng-Robinson fugacity coefficients in this phase are given by*

$$\begin{aligned}\ln \Phi^i_\beta = \ & \frac{B + \nabla B\cdot(\boldsymbol{\delta}^i - \boldsymbol{x})}{B}\,[W_\beta - 1] - \ln\,[W_\beta - B] \\ & + \left[\frac{B + \nabla B\cdot(\boldsymbol{\delta}^i - \boldsymbol{x})}{B} - \frac{2A + \nabla_{\boldsymbol{x}} A\cdot(\boldsymbol{\delta}^i - \boldsymbol{x})}{A}\right]\frac{A}{2\sqrt{2}B}\,\ln\left[\frac{W_\beta + (\sqrt{2}+1)B}{W_\beta - (\sqrt{2}-1)B}\right] \\ & + \left[\frac{\nabla W_\beta\cdot(\boldsymbol{\delta}^i - \boldsymbol{x})}{W_\beta} - \frac{\nabla B\cdot(\boldsymbol{\delta}^i - \boldsymbol{x})}{B}\right]\frac{W_\beta\,\Upsilon_{A,B}(W_\beta)}{(W_\beta - B)(W_\beta^2 + 2BW_\beta - B^2)}\end{aligned} \tag{3.85}$$

for all $i \in \mathcal{K}$*, with* $\Upsilon_{A,B}(W) = W^3 + (B-1)W^2 + (A - 2B - 3B^2)W + (B^2 + B^3 - AB)$ *as defined in* (3.55)*.*

Chứng minh. The proof is similar to that of Theorem 3.7. □

The gradient of W_β with respect to $\boldsymbol{x}$ required by (3.85), can be computed by

$$\nabla W_\beta = -\frac{1}{2}(\nabla B + \nabla Z_\alpha),$$

in which ∇Z_α solves

$$\begin{aligned}[3Z_\alpha^2 + 2(B-1)Z_\alpha + (A - 2B - 3B^2)]\nabla Z_\alpha = \ & (B - Z_\alpha)\nabla A \\ & + (A - 2B - 3B^2 + 6BZ_\alpha + 2Z_\alpha - Z_\alpha^2)\nabla B.\end{aligned}$$

Alteration in the three-root region. For the same reasons as those mentioned in the Van der Waals case, the usual formulas need to be altered in the three-root region, where Z_β gets close to being a double root. The changes are aimed at circumventing the difficulty due to the blowing up of ∇Z_β and at enforcing a smooth junction between the two regions. We follow the same strategy as in the Van der Waals case. When there are three roots $B < Z_L < Z_I < Z_G$, we define the indicator ϑ as in (3.77). Let $\varepsilon \in (0, 1/4)$ be a small threshold.

- If $\vartheta \in [2\varepsilon, 1 - 2\varepsilon]$, no change is necessary.

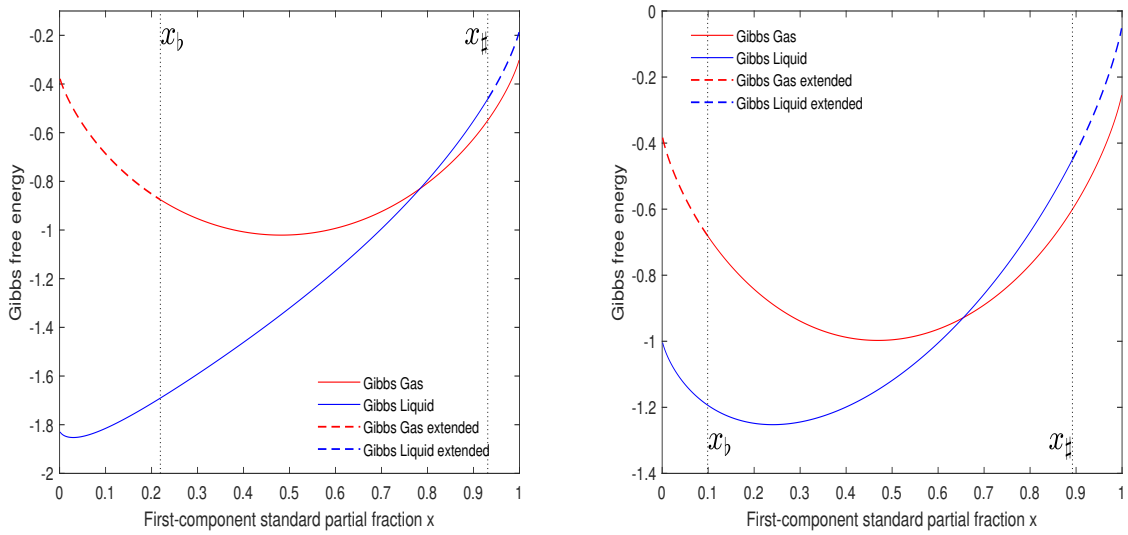


Figure 3.22: Extended Gibbs energy functions g_L (blue) and g_G (red) for Peng-Robinson's law by the indirect method, with $\varepsilon = 0.001$. Left panel: $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.322, 0.053)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.33, 0.03)$. Right panel: $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.275, 0.045)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.35, 0.04)$.

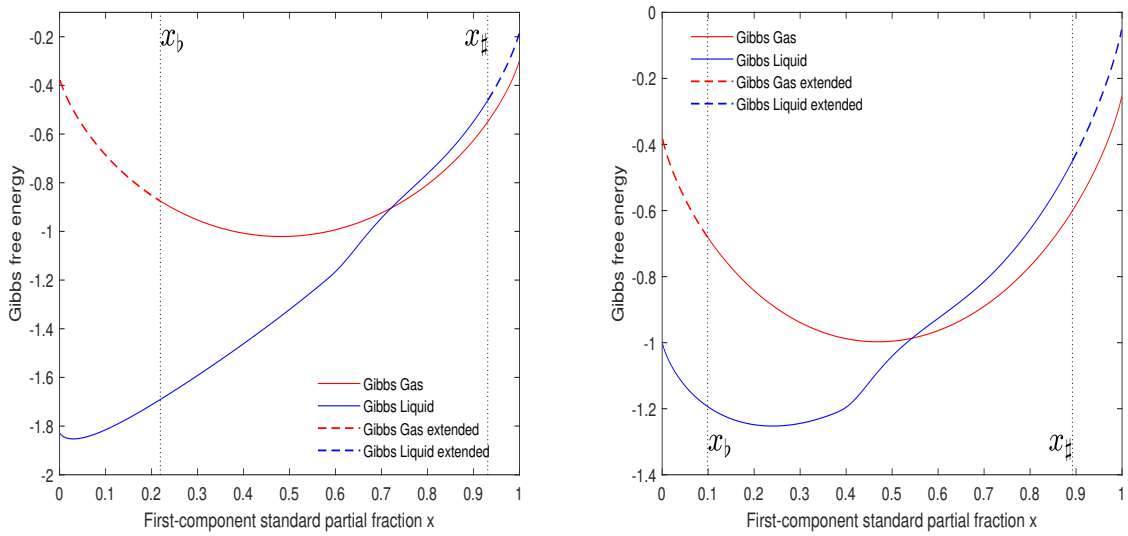
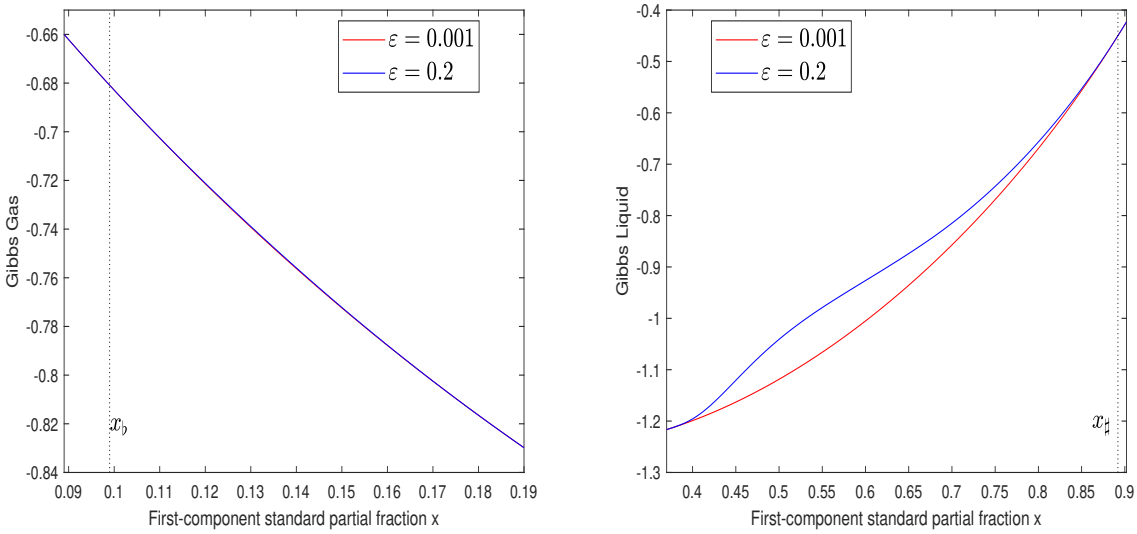
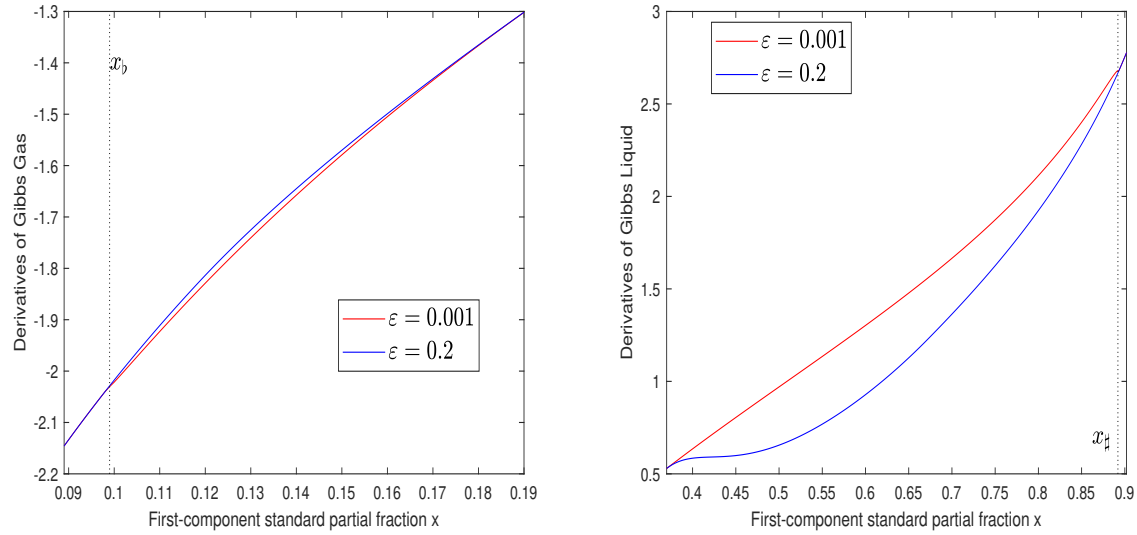


Figure 3.23: Extended Gibbs energy functions g_L (blue) and g_G (red) for Peng-Robinson's law by the indirect method, with $\varepsilon = 0.2$. Left panel: $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.322, 0.053)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.33, 0.03)$. Right panel: $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.275, 0.045)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.35, 0.04)$.



(a) g_G (b) g_L

Figure 3.24: Close-up comparison of the extended Gibbs functions between $\varepsilon = 0.001$ and $\varepsilon = 0.2$ for Peng-Robinson' law with the indirect method. $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.275, 0.045)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.35, 0.04)$.



(a) g'_G (b) g'_L

Figure 3.25: Close-up comparison of the derivative of the extended Gibbs functions between $\varepsilon = 0.001$ and $\varepsilon = 0.2$ for Peng-Robinson' law with the indirect method. $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.275, 0.045)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.35, 0.04)$.

- If $\vartheta \in (1-2\varepsilon, 1]$, we keep Z_L but progressively replace Z_G by

$$W_G = \frac{1 - B - Z_L}{2} = \frac{Z_I + Z_G}{2}. \tag{3.86}$$

Instead of plugging Z_G into formula (3.52) for Ψ_G, we insert

$$\widetilde{Z}_G = [1 - \nu_G(\vartheta)]Z_G + \nu_G(\theta)W_G, \tag{3.87a}$$

where ν_G is given by (3.80). From the modified excess Gibbs energy

$$\Psi_G = \widetilde{Z}_G - 1 - \ln\big[\widetilde{Z}_G - B\big] - \frac{A}{2\sqrt{2}B}\,\ln\left[\frac{\widetilde{Z}_G + (\sqrt{2}+1)B}{\widetilde{Z}_G - (\sqrt{2}-1)B}\right], \tag{3.87b}$$

the fugacity coefficients can be inferred by (3.2) as

$$\begin{aligned}\ln \Phi_G^i = {} & \frac{B + \nabla B \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B}\,[\widetilde{Z}_G - 1] - \ln[\widetilde{Z}_G - B] \\ & + \left[\frac{B + \nabla B \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B} - \frac{2A + \nabla_{\boldsymbol{x}} A \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{A}\right]\frac{A}{2\sqrt{2}B}\,\ln\left[\frac{\widetilde{Z}_G + (\sqrt{2}+1)B}{\widetilde{Z}_G - (\sqrt{2}-1)B}\right] \\ & + \left[\frac{\nabla \widetilde{Z}_G \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{\widetilde{Z}_G} - \frac{\nabla B \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B}\right]\frac{\widetilde{Z}_G\,\Upsilon_{A,B}(\widetilde{Z}_G)}{(\widetilde{Z}_G - B)(\widetilde{Z}_G^2 + 2B\widetilde{Z}_G - B^2)}\end{aligned} \tag{3.87c}$$

The gradient $\nabla \widetilde{Z}_G$ in the above formula can be approximated by (3.81c), in which $\nabla W_G = -\frac{1}{2}(\nabla B + \nabla Z_L)$.

- If $\vartheta \in [0, 2\varepsilon)$, we proceed in a similar and symmetric fashion.

Figures 3.22–3.23 display a few examples of the indirect method for the Peng-Robinson case. Figures 3.24–3.25 provide a close-up comparison between two choices of ε. In comparison with Van der Waals case, here the width parameter ε seems to have a slightly stronger influence on the extended Gibbs functions. Similarly to the Van der Waals case, this inflence is more visible for liquid phase.

Part II

Numerical methods and simulations

Chapter 4

Existing methods for sytems with complementarity conditions

Contents

Nous entamons cette seconde partie, consacrée au numérique, par un panorama des méthodes susceptibles de résoudre le problème thermodynamique posé dans la première partie. Pour cela, un survol des problèmes de complémentarité "purs" en §4.1 constitue une étape préliminaire indispensable pour connaître les principales classes de méthodes à explorer.

La non-différentiabilité du problème nous amène à examiner d'abord les méthodes non-lisses et semi-lisses en §4.2. Parmi celles-ci figure la méthode de Newton-min, qui est actuellement l'algorithme par défaut dans les prototypes d'IFPEN utilisant la formulation unifiée. Nous nous intéressons ensuite en §4.3 aux méthodes de régularisation, qui transforment le problème non-lisse de départ en une suite de problèmes lisses au moyen d'un paramètre destiné à tendre vers zéro. Après un retour sur la méthode de Newton classique et ses théorèmes de convergence locale, nous mettrons l'accent sur la technique de lissage par des θ-fonctions ainsi que la méthode des points intérieurs.

Pour terminer, mais aussi pour motiver la conception d'une méthode mieux adaptée, nous énumérons en §4.4 les problèmes de convergence des méthodes considérées sur des contre-exemples.

The phase equilibrium problem (2.42) or (2.77), studied in chapter §2, comes within the following abstract framework: find $X \in \mathcal{D}$, where $\mathcal{D} \subset \mathbb{R}^\ell$ is an open domain, such that

$$\Lambda(X) = 0, \qquad \in \mathbb{R}^{\ell-m}, \tag{4.1a}$$

$$\min(G(X), H(X)) = 0, \qquad \in \mathbb{R}^m. \tag{4.1b}$$

Here, the given functions $\Lambda : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^{\ell-m}$ and $G, H : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^m$, where $0 < m \leqslant \ell$, are assumed to be continuously differentiable on $\mathcal{D}$. The first $\ell - m$ equations (4.1a) are "ordinary" algebraic equations. By contrast, the last m equations (4.1b) are rather "special" in that they are nondifferentiable, because of the componentwise min function. They represent the so-called *complementarity conditions*, the exact significance of which is

$$0 \leqslant G(X) \perp H(X) \geqslant 0, \tag{4.2a}$$

or equivalently,

$$G(X) \geqslant 0, \qquad H(X) \geqslant 0, \qquad G(X)^T H(X) = 0. \tag{4.2b}$$

This name is justified by the observation that for each index $\alpha \in \{1, \ldots, m\}$, at least one of the two quantities $G_\alpha(X)$ and $H_\alpha(X)$ vanishes while the other remains nonnegative.

Our objective is to work out an efficient and robust numerical method to solve (4.1). The most severe difficulty that awaits us is the non-differentiability of the complementarity conditions. Therefore, before embarking on the quest for numerical methods, we have to fully understand the essence of this difficulty by stepping back to the simpler case of a "pure" complementarity problem.

4.1 Background on complementarity problems

A complementarity problem[1] is a specialized version of (4.1), in which

$$m = \ell, \qquad G(X) = X. \tag{4.3}$$

Over the last half-century, complementarity problems have grown into a vast discipline with many deep notions and rich results. A comprehensive survey can be found in the book of Acary and Brogliato [2] or the two-volume collection of Facchinei and Pang [46,47]. In this section, we just intend to provide some standard theoretical rudiments that will be useful in the sequel.

4.1.1 Classes of problems

We begin with a very basic notion, that of a cone. A subset $\mathfrak{K} \subset \mathbb{R}^\ell$ is said to be a *cone* if

$$\forall X \in \mathfrak{K}, \quad \forall t > 0, \quad tX \in \mathfrak{K}.$$

If $\mathfrak{K} \subset \mathbb{R}^\ell$ is a cone, its *dual cone* is defined as

$$\mathfrak{K}^\circ := \{d \in \mathbb{R}^\ell \mid \forall v \in \mathfrak{K}, \quad v^T d \geqslant 0\}.$$

These notions are actually defined in analysis, independently of complementarity problems. They enable us to properly introduce the general complementarity problem (GCP) associated with a cone.

[1] We sometimes add the adjective "pure" to mark the difference with the original "mixed" problem (4.1).

4.1.1.1 GCP and VIP

Definition 4.1 (GCP)**.** Given a cone $\mathfrak{K} \subset \mathbb{R}^\ell$ and a mapping $H : \mathfrak{K} \to \mathbb{R}^\ell$, the *general complementarity problem* $C(\mathfrak{K}, H)$ consists in finding a vector $X \in \mathbb{R}^\ell$ that satisfies the conditions

$$\mathfrak{K} \ni X \perp H(X) \in \mathfrak{K}^\circ, \tag{4.4}$$

where the notation "$\perp$" means "perpendicular", i.e., $X^T H(X) = 0$ in the matrix language.

This formulation of (GCP) includes a wide range of problems encountered in mathematical programming. It can be further extended to the infinite-dimensional setting by replacing $\mathbb{R}^\ell$ a pair of locally convex Hausdorff spaces related to each other by real-valued bilinear form [66]. Beside the world of mathematical programming, there is also another community in applied mathematics whose primary interest is focused on the *unilateral conditions* for nonlinear partial differential equations arising from mechanics, especially in elasticity and plasticity. The theoretical tool to study this type of free boundary problems is the variational inequality problem. We refer the readers to the monographs of Kinderlehrer and Stampacchia [69] and Glowinski et al. [53] for a broad review of this realm. Below we formulate the variational inequality problem (VIP) associated with a subset of $\mathbb{R}^\ell$ which is not necessarily a cone.

Definition 4.2 (VIP)**.** Given a subset $\mathfrak{K} \subset \mathbb{R}^\ell$ and a mapping $H : \mathfrak{K} \to \mathbb{R}^\ell$, the *variational inequality problem* $V(\mathfrak{K}, H)$ consists in finding a vector $X \in \mathfrak{K}$ such that

$$\forall Y \in \mathfrak{K}, \qquad (Y - X)^T H(X) \geqslant 0. \tag{4.5}$$

This formulation of (VIP) includes a wide range of problems encountered in mechanical engineering. It can be further extended to the infinite-dimensional setting, where it actually comes from [60]. Originally, the relationship between (VIP) and (GCP) has been noted by many authors. However, it was Karamardian [66] who proved that if the set $\mathfrak{K}$ involved in Definition 4.2 is a cone, then the two problems are equivalent.

Proposition 4.1. *Let $\mathfrak{K} \subset \mathbb{R}^\ell$ be a cone and H be a mapping from $\mathfrak{K}$ to $\mathbb{R}^N$. A vector $X \in \mathfrak{K}$ solves $V(\mathfrak{K}, H)$ if and only if it solves $C(\mathfrak{K}, H)$.*

Chứng minh. See [66] or [46, §1.1.3]. □

4.1.1.2 GCP, NCP and LCP

Many special cases of (GCP) are worth considering for their role in practical problems. When $\mathfrak{K}$ is the nonnegative orthant of $\mathbb{R}^\ell$, the general complementarity problem (GCP) gives rise to the nonlinear complementarity problem (NCP).

Definition 4.3 (NCP)**.** Given a mapping $H : \mathbb{R}^\ell_+ \to \mathbb{R}^\ell$, the *nonlinear complementarity problem* associated with H consists in finding a vector $X \in \mathbb{R}^\ell$ such that

$$0 \leqslant X \perp H(X) \geqslant 0, \tag{NCP}$$

which means

$$X \geqslant 0, \qquad H(X) \geqslant 0, \qquad X^T H(X) = 0. \tag{4.6a}$$

The nonlinear complementarity problem was introduced by Cottle [33], at about the same time as (VIP). Among the class of (NCP), it is customary to consider those for which H is

- a P_0-*function*, that is,

$$\forall X \neq Y \in \mathbb{R}^\ell_+, \qquad \max_{\substack{1\leqslant\alpha\leqslant\ell \\ X_\alpha\neq Y_\alpha}} (X_\alpha - Y_\alpha)(H_\alpha(X) - H_\alpha(Y)) \geqslant 0; \tag{4.7a}$$

- or a P-*function*, that is,

$$\forall X \neq Y \in \mathbb{R}^\ell_+, \qquad \max_{1\leqslant\alpha\leqslant\ell} (X_\alpha - Y_\alpha)(H_\alpha(X) - H_\alpha(Y)) > 0. \tag{4.7b}$$

Indeed, uniqueness can be proven for the latter case.

Theorem 4.1. *If H is a P-function in the sense of* (4.7b)*, then (NCP) has at most one solution.*

Chứng minh. See [46, §3.5.10] □

When H is an affine function, that is, $H(X) \equiv MX + q$ for some matrix $M \in \mathbb{R}^{\ell\times\ell}$ and vector $q \in \mathbb{R}^\ell$, the problem has a dedicated name.

Definition 4.4 (LCP)**.** Given $M \in \mathbb{R}^{\ell\times\ell}$ and $q \in \mathbb{R}^\ell$, the *linear complementarity problem* $LC(M, q)$ consists in finding a vector $X \in \mathbb{R}^\ell$ such that

$$0 \leqslant X \perp MX + q \geqslant 0, \tag{4.8a}$$

which means

$$X \geqslant 0, \qquad MX + q \geqslant 0, \qquad X^T(MX + q) = 0. \tag{4.8b}$$

As a matter of fact, (LCP) was the first type of complementarity problem to have been formalized in the literature. The motivation for this comes from the observation that KKT optimality conditions for linear and quadratic programs constitute an (LCP). After Lemke and Howson [82] showed that the problem of computing a Nash equilibrium point of a bimatrix game can be posed as an (LCP), Cottle and Dantzig [34] unified linear and quadratic programs and bimatrix games under the (LCP). Since then, (LCP) has gained considerable momentum. The history of the development of (LCP) is available in Cottle et al. [35].

In the case of (LCP), the P_0 and P properties of H can be detected at the level of the matrix M. A matrix $M \in \mathbb{R}^{\ell\times\ell}$ is said to be:

- a P_0-matrix if for all $X \neq 0$, there exists an index $\alpha \in \{1, \ldots, \ell\}$ such that $X_\alpha \neq 0$ and $(MX)_\alpha \geqslant 0$;
- a P-matrix if the inequality $X^T MX \leqslant 0$ implies $X = 0$.

A P_0-matrix generalizes a positive semi-definite (symmetric) matrix. There are many equivalent characterizations to the above definition, as enumerated in [11, §2.2.4] and [49]. A P-matrix generalizes a positive definite (symmetric) matrix and there are also many equivalent characterizations [11, §2.2.5]. Determining whether or not a given matrix is a P-matrix is an expensive task. In fact, this is a co-NP-complete problem [36].

Going back to the (NCP), let us assume that the mapping H is continuously differentiable. Then, the P_0 (resp. P) property of H is equivalent to that of its Jacobian matrix ∇H for all X in the domain [46, §3.5.9].

4.1.2 Classes of methods

Once the main classes of complementarity problems have been identified, we are now concerned with the numerical methods that can be used to solve them. Again, this will be a brief glimpse, but the overview we will have had in the case of complementarity problems will serve as a guiding outline for the more sophisticated case of (4.1) in later sections.

4.1.2.1 Early approaches

The first difficulty to point out with complementarity conditions in the general case (4.1) is their combinatorial nature. Anecdotally, some instances of (LCP) have been proved [29, 70] to be NP-complete in the strong sense.

For each index $\alpha \in \{1, \dots, m\}$, the equation $0 \leqslant G_\alpha(X) \perp H_\alpha(X) \geqslant 0$ expresses two possible operating regimes, depending on either $G_\alpha(X) \geqslant 0$ or $H_\alpha(X) \geqslant 0$. In the phase equilibrium system (2.77), for instance, the two regimes correspond to whether or not phase α is present in the mixture. Since there are m complementarity conditions, the total number of possible configurations for the physical system is 2^m. In the model problem (2.77), where $m = P$, the total number of possible contexts is $2^P - 1$ (the difference of one unit comes from the fact that the phases cannot be all absent). In realistic reservoir simulations, since the phase equilibrium problems of the cells are coupled to each other, m is equal to the product of the number of possible phases P by the number of cells in the mesh, which could reach ten million. Thus, any method by which it is proposed to explore all possible configurations is doomed to failure when m is large.

Pivotal methods. The situation described above naturally reminds us of KKT conditions for constrained optimization and of linear programming, for which the active-set methods and the simplex algorithm enable us to update the guessed configuration in a "smart" way, instead of visiting them all. The class of conceptually equivalent methods for (LCP) is known as *pivotal methods*. These are essentially variants of the so-called *complementarity pivot method* by Lemke and Howson [82]. The most well-known methods among them are the Lemke algorithm [81] and the *criss-cross* algorithm [40]. The common feature of all pivotal methods is that the worst-case complexity is exponential. We refer the readers to Billups and Murty [20] and Cottle et al. [35] for a more thorough review.

Nonsmooth methods. Pang [98] is credited for having developed the first globally convergent and locally superlinearly convergent B-differentiable Newton method with line search. It was followed by the path search method of Ralph [107] and a method for PC^1-functions by Kojima and Shindo [71], while Kummer [73] studied this method for general nondifferentiable functions. In §4.2.1, we will supply some elements of the general theory of nonsmooth Newton.

4.1.2.2 Recent approaches

Semismooth methods. Semismooth functions are an important special case of nonsmooth functions. The theory of semismooth functions was developed by Miflin [92] in the scalar case and extended to the vector case by Qi and Sun [105]. This class of methods involves reformulating the problem as a system of nonlinear equations by means of C-functions (C stands for complementarity). A function ψ is said to be a *C-function* if

$$\psi(a, b) = 0 \quad \Leftrightarrow \quad 0 \leqslant a \perp b \geqslant 0. \tag{4.9}$$

Using a C-function, the complementarity problem (NCP) can be stated as the system of equations

$$F(X) = 0, \tag{4.10a}$$

where $F : \mathbb{R}^\ell \to \mathbb{R}^\ell$ is defined component-wise by

$$F_\alpha(X) = \psi(X_\alpha, H_\alpha(X)). \tag{4.10b}$$

System (4.10) remains to be solved by a semi-smooth Newton-type method. Below is a non-exhaustive list of the most frequently used C-functions.

- Fischer-Burmeister function:

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

  This C-function is differentiable everywhere except at $(0, 0)$. In addition, its square $\psi^2_{FB}(a, b)$ is continuously differentiable on the entire plane. Introduced in [50], the Fischer-Burmeister function soon attracted the attention of many researchers [38,48] and played a central role in the development of efficient algorithms. The corresponding semi-smooth method to solve (4.10) is called *Newton-FB*.

- Minimum function:

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

  This C-function is a Lipschitz function, but not differentiable when $a = b$. The earliest use of the min function in complementarity problems dates back to Aganagić [3]. The corresponding semi-smooth method to solve (4.10) is called *Newton-min*. In the context of (LCP) and (NCP), its convergence properties were analyzed by [51, 59]. According to the numerical tests of [37,65], Newton-min gives better results than Newton-FB. Ben Gharbia and her co-authors [14–17] used it extensively in the context of mixed systems.

- Mangasarian function:

$$\psi_{\mathrm{M}}(a, b) = \zeta(|a - b|) - \zeta(b) - \zeta(a)$$

  where $\zeta : \mathbb{R} \to \mathbb{R}$ is a strictly increasing function and $\zeta(0) = 0$. It can be made differentiable everywhere by an appropriate choice of ζ, for instance $\zeta(t) = t^3$. Mangasarian [85] introduced this family of C-functions with the intention of solving (NCP), but the corresponding *Newton-M* method does not seem to be very popular. This is probably due to the fact that all smooth C-functions share the same deficiency: $\nabla\psi(0, 0) = (0, 0)$. This implies that for every index $\alpha \in \{1, \ldots, \ell\}$ for which $X_\alpha = H_\alpha(X) = 0$, we have $\nabla F_\alpha(X) = 0$ and the α-th row of the Jacobian matrix consists of zero entries, which makes it singular.

In §4.2.2, we will provide some basic notions on semismooth methods, with a focus on the Newton-min method in §4.2.3.

Smoothing methods. A complementarity condition can also be regularized by a smooth function, which introduces a regularization parameter. The idea is to apply smooth methods to the smoothed equations and to gradually drive the regularization parameter to zero. Chen and Mangasarian [27] were probably the first to come up with this strategy, that we will present in §4.3. Smoothing methods also include the large family of interior-point methods, a brief survey of which will be given in §4.3.3.

4.2 Nonsmooth approach to generalized equations

Let us return to the original mixed problem (4.1). We want to numerically solve

$$F(X) = 0, \tag{4.11}$$

where the function $F : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^\ell$ is not necessarily *smooth*, that is, not necessarily continuously Fréchet-differentiable everywhere its domain. We recall that Fréchet-differentiability[2] at $X \in \mathcal{D}$ means that there exists a linear map $\nabla F(X) : \mathbb{R}^\ell \to \mathbb{R}^\ell$, or equivalently a matrix $\nabla F(X) \in \mathbb{R}^{\ell\times\ell}$ in the canonical basis, such that

$$\lim_{d\to 0} \frac{\|F(X+d) - F(X) - \nabla F(X)d\,\|}{\|\,d\,\|} = 0.$$

For system (4.1), we have $F(X) = [\Lambda(X), \min(G(X), H(X))]^T$, but let us work with a general *nonsmooth* function F.

In the smooth case, the Newton method is based on the idea of replacing F by successive local models that are easier to solve. These local models rest upon the first-order Taylor expansion. More specifically, given some $X^k \in \mathcal{D}$, we consider the local model

$$\bar{X} \mapsto F(X^k) + \nabla F(X^k)(\bar{X} - X^k) \tag{4.12}$$

as an approximation of $F(\bar{X})$ when $\bar{X}$ is close to X^k, and search for X^{k+1} as the zero of (4.12) instead of (4.11). In the nonsmooth case, the philosophy of the nonsmooth approach is to attempt some generalization of the above process. We have to face many challenges. On the one hand, it is highly unlikely that we would be able to design a method for all nonsmooth functions. Reasonably, additional assumptions on F will have to be made. On the other hand, it is not clear what alternate local model could be used as a nonsmooth analog for the first-order Taylor expansion.

In this section, we are going to present a theory developed for nonsmooth functions that are locally Lipschitz-continuous. We recall that F is locally Lipschitz-continuous at $X \in \mathcal{D}$ if there exists a neighborhood $B(X, \epsilon_X)$ of X and a constant L_X such that

$$\|F(\check{X}) - F(\tilde{X})\| \leqslant L_X \|\check{X} - \tilde{X}\|, \qquad \forall (\check{X}, \tilde{X}) \in B(X, \epsilon_X) \times B(X, \epsilon_X).$$

In §4.2.1, an abstract framework for the local model is introduced, which gives rise to an abstract nonsmooth Newton method. In §4.2.2, at the price of further restricting ourselves to the subclass of semismooth functions, a concrete instance of this theory is provided, which gives rise to the semismooth Newton method.

4.2.1 Nonsmooth Newton method

4.2.1.1 Local model and algorithm

The generalization of the smooth local model (4.12) takes the form

$$\bar{X} \mapsto F(X^k) + T(X^k, \bar{X} - X^k), \tag{4.13}$$

[2] In a finite-dimensional space, Fréchet-differentiability is equivalent to the usual notion of differentiability. This is why we shall simply speak about "differentiability" throughout the remainder of the manuscirpt.

where $T(X^k, \cdot)$ represents some abstract function. To account for the dependency of this approximation on the current point X^k, we need to consider a family of functions $\mathcal{T}(X)$ to which each possible $T(X, \cdot)$ belongs. This is clarified in the following Definition, where we designate by $\mathscr{T}(\mathbb{R}^\ell)$ the set of functions from $\mathbb{R}^\ell$ to $\mathbb{R}^\ell$. No further property is required on $\mathscr{T}(\mathbb{R}^\ell)$.

Definition 4.5 (Newton approximation scheme)**.** Let $F : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^\ell$ be a locally Lipschitz-continuous function.

1. A *Newton approximation scheme* of F is a set-valued mapping $\mathcal{T} : \mathcal{D} \rightrightarrows \mathscr{T}(\mathbb{R}^\ell)$ such that

$$T(X, 0) = 0, \qquad \text{for all } \; T(X, \cdot) \in \mathcal{T}(X), \tag{4.14a}$$

and

$$\limsup_{\substack{X \to \bar{X} \\ T(X,\cdot) \in \mathcal{T}(X)}} \frac{\| F(X) + T(X, \bar{X} - X) - F(\bar{X}) \|}{\| X - \bar{X} \|} = 0, \qquad \text{for all } \bar{X} \in \mathcal{D}. \tag{4.14b}$$

2. A *strong Newton approximation scheme* of F is a Newton approximation of F strengthened by the condition

$$\limsup_{\substack{X \to \bar{X} \\ T(X,\cdot) \in \mathcal{T}(X)}} \frac{\| F(X) + T(X, \bar{X} - X) - F(\bar{X}) \|}{\| X - \bar{X} \|^2} < \infty, \qquad \text{for all } \bar{X} \in \mathcal{D}. \tag{4.14c}$$

3. A (strong) *nonsingular Newton approximation* of F is a (strong) Newton approximation of F strengthened by the condition that $\mathcal{T}$ is a family of uniformly Lipschitz homeomorphisms on $\mathcal{D}$, by which we mean that there exist positive constants $L_\mathcal{T}$ and $\varepsilon_\mathcal{T}$ such that for each $X \in \mathcal{D}$ and for each $T(X, \cdot) \in \mathcal{T}(X)$, there are two open sets U_X and V_X, both containing $B(0, \varepsilon_\mathcal{T})$, such that $T(X, \cdot)$ is a Lipschitz homeomorphism mapping U_X onto V_X with $L_\mathcal{T}$ being the Lipschitz modulus of the inverse of the restricted map $T(X, \cdot)|_{U_X}$.

Condition (4.14a) means that the local model

$$d \mapsto F(X) + T(X, d), \tag{4.15}$$

aimed at approximating $F(X + d)$ around X, must return the exact value $F(X)$ for $d = 0$. This is quite natural. Condition (4.14b)–(4.14c) expresses that the local model must possess good aproximation properties for $d \neq 0$ small enough. As for the notion of singular Newton approximation in the third item, it postulates that the local model must be invertible with respect to d, at least locally. This is where the locally Lipschitz-continuous assumption on F is really needed.

With the above definition, a natural extension of the Newton method is described in Algorithm 4.1 for nonsmooth equations. This algorithm is very abstract. We do not know what $T(X, \cdot)$ looks like. It is not even required to be linear. Our only hope is that in Step 3, solving for d^k in (4.16) is easier than coping with the original problem. Otherwise, the local model is irrelevant. Notice, however, that there may not be a unique solution d^k in Step 3. For one, we may pick another element $T(X^k, \cdot) \in \mathcal{T}(X^k)$ if $\mathcal{T}(X^k)$ is not a singleton. For another, equation (4.16) may have several solutions d^k for the same $T(X^k, \cdot)$. Some authors [47, §7.2.4] recommend looking for $d^k \in B(0, \epsilon)$ instead of $\mathbb{R}^\ell$, where ϵ is a user-prescribed maximal radius, in order to not get out of the "good" neighborhood. But the problem is then that equation (4.16) may not have any solution.

Algorithm 4.1 Nonsmooth Newton algorithm

1. Choose $X^0 \in \mathcal{D} \subset \mathbb{R}^\ell$. Set $k = 0$.
2. If $F(X^k) = 0$, stop.
3. Select an element $T(X^k, \cdot) \in \mathcal{T}(X^k)$. Find a direction $d^k \in \mathbb{R}^\ell$ such that
$$F(X^k) + T(X^k, d^k) = 0. \tag{4.16}$$
4. Set $X^{k+1} = X^k + d^k$ and $k \leftarrow k + 1$. Go to step 2.

4.2.1.2 Well-definedness and convergence

Nonsingularity of the Newton approximation scheme is crucial for the sequence of iterates $\{X^k\}_{k\in\mathbb{N}}$ in Algorithm 4.1 to be well-defined. This turns out to be the hardest point to verify in practice. The following Theorem provides a sufficient condition for nonsingularity based on an assumption of pointwise singularity at the solution $\bar{X}$.

Theorem 4.2. *Let $F : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^\ell$ be a locally Lipschitz-continuous function and let $\bar{X} \in \mathcal{D}$ be a solution of $F(\bar{X}) = 0$. Assume that $\mathcal{T}$ is a Newton approximation scheme of F for which there exist three positive constants $\varepsilon_1, \varepsilon_2$ and L satisfying*

(A) *for each $T(\bar{X}, \cdot) \in \mathcal{T}(\bar{X})$ there are two sets U and V containing $B(0, \varepsilon_1)$ and $B(0, \varepsilon_2)$ respectively, such that $T(\bar{X}, \cdot)$ is a Lipschitz homeomorphism from U to V and $T^{-1}(\bar{X}, \cdot)$ has Lipschitz modulus L.*

Assume further that a function $L : \mathbb{R}^*_+ \to \mathbb{R}_+$ with $\lim_{t\downarrow 0} L(t) = 0$ and a neighborhood $\mathcal{N}$ of $\bar{X}$ exist such that either one of the following two conditions holds:

(a) *for every $X \in \mathcal{N}$ and every $T(X, \cdot) \in \mathcal{T}(X)$, there exists a member $T(\bar{X}, \cdot)$ in $\mathcal{T}(\bar{X})$ such that $T(X, \cdot) - T(\bar{X}, \cdot)$ is Lipschitz-continuous with modulus $L(\|X - \bar{X}\|)$ on U; or*

(b) *for every $X \in \mathcal{N}, \mathcal{T}(X) = \{T(X, \cdot)\}$ is single valued and $\widetilde{T}(X, \cdot) - T(\bar{X}, \cdot)$ is Lipschitz-continuous with modulus $L(\|X - \bar{X}\|)$ on U, where $\widetilde{T}(X, d) \equiv T(X, \bar{X} - X + d)$.*

Then the Newton approximation scheme $\mathcal{T}$ is nonsingular.

Chứng minh. See [47, §7.2.12–§7.2.13]. □

Once the sequence of iterates is well-defined, the next question is about its convergence. Before stating the main result, we recall the following defintions regarding convergence rates that will be useful for other methods as well.

Definition 4.6 (Rates of convergence)**.** Let $\{X^k\}_{k\in\mathbb{N}^*} \subset \mathbb{R}^\ell$ be a sequence converging to $\bar{X} \in \mathbb{R}^\ell$, with $X^k \neq \bar{X}$ for all $k \geqslant 0$. We say that $\{X^k\}_{k\in\mathbb{N}^*}$ converges to $\bar{X}$:

1. *Q-linearly* if
$$0 < \limsup_{k\to\infty} \frac{\|X^{k+1} - \bar{X}\|}{\|X^k - \bar{X}\|} < 1. \tag{4.17a}$$

2. *Q-superlinearly* if

$$\limsup_{k\to\infty} \frac{\|X^{k+1} - \bar{X}\|}{\|X^k - \bar{X}\|} = 0. \tag{4.17b}$$

3. *Q-quadratically* if

$$0 < \limsup_{k\to\infty} \frac{\|X^{k+1} - \bar{X}\|}{\|X^k - \bar{X}\|^2} < \infty. \tag{4.17c}$$

The upcoming theorem recapitulates the key properties of Algorithm 4.1.

Theorem 4.3. *Let $F : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^\ell$ be a locally Lipschitz-continuous function and let $\bar{X} \in \mathcal{D}$ be a solution of $F(\bar{X}) = 0$. Assume that F admits a nonsingular Newton approximation $\mathcal{T}$.*

Then, for every $\varepsilon \in (0, \varepsilon_{\mathcal{T}}]$, there exists $\delta > 0$ such that if $X^0 \in B(\bar{X}, \delta)$, then Algorithm 4.1 generates a unique sequence $\{X^k\}_{k\in\mathbb{N}}$ that converges Q-superlinearly *to $\bar{X}$. Furthermore, if the Newton approximation scheme $\mathcal{T}$ is strong, the rate of convergence is* Q-quadratic.

Chứng minh. See [47, §7.2.5]. □

4.2.2 Semismooth Newton method

As said earlier, although the nonsmooth Newton method of §4.2.1 is a convenient theoretical tool, it is too generic a construction. To be of any practical use, the Newton approximation scheme $\mathcal{T}$ must be specified in a more substantial way. This can be achieved for semismooth functions [92, 105], the definition of which requires some preliminary notions on subdifferentials.

4.2.2.1 Local model and algorithm

By Rademacher's theorem [31, §3.4.1], every locally Lipschitz-continuous function is continuously differentiable almost everywhere. Put another way, the set $\mathcal{C}_F$ of points $X \in \mathcal{D}$ where $\nabla F(X)$ exists in the classical sense is non-empty and its complement $\mathcal{D}\backslash\mathcal{C}_F$ has measure zero. This property lies at the foundation of the following definitions.

Definition 4.7 (Bouligand and Clarke subdifferentials)**.** Let $F : \mathcal{D} \subseteq \mathbb{R}^\ell \to \mathbb{R}^\ell$ be a locally Lipschitz-continuous function and $\mathcal{C}_F \subset \mathcal{D}$ be the set of points at which F is differentiable.

1. The *B-subdifferential* or the *limiting Jacobian* of F at X is the set-valued mapping $\partial_B F : \mathcal{D} \rightrightarrows \mathbb{R}^{\ell\times\ell}$ defined as

$$\partial_B F(X) = \big\{M \in \mathbb{R}^{\ell\times\ell} \mid \exists (X^k)_{k\in\mathbb{N}} \subset \mathcal{C}_F,\ X^k \to X,\ \nabla F(X^k) \to M\big\}. \tag{4.18a}$$

   In other words, the Bouligand subdifferential $\partial_B F(X)$ is the set of all matrices M are the limits of the Frechet differentials $\nabla F(X^k)$ for a sequence X^k converging to X.

2. The *C-subdifferential* or the *generalized Jacobian* of F at X is the set-valued mapping $\partial F : \mathcal{D} \rightrightarrows \mathbb{R}^{\ell\times\ell}$ given by

$$\partial F(X) = \operatorname{conv}(\partial_B F(X)). \tag{4.18b}$$

   In other words, the Clarke subdifferential $\partial F(X)$ is the convex hull of the Bouligand subdifferential $\partial_B F(X)$.

As a classical example, let us consider $f(x) = |x|$ for $x \in \mathbb{R}$. Then, $\partial_B f(0) = \{-1, 1\}$ and $\partial f(0) = [-1, 1]$. The generalized Jacobian ∂F latter allows many classical results valid for smooth functions to be extended to locally Lipschitz-continuous functions. Regarding the Newton method, if the function F at hand is locally Lipschitz-continuous, it is of course tempting to associate each $M \in \partial F(X)$ with the function $T_M(X, \cdot) : \mathbb{R}^\ell \to \mathbb{R}^\ell$ defined by

$$T_M(X, d) = Md, \qquad \forall d \in \mathbb{R}^\ell, \tag{4.19a}$$

and to create the family

$$\mathcal{T}(X) = \big\{T_M(X, \cdot), \quad M \in \partial F(X)\big\} \tag{4.19b}$$

in order to obtain a Newton approximation scheme $\mathcal{T} :: \mathcal{D} \rightrightarrows \mathscr{T}(\mathbb{R}^\ell)$ in the sense of Definition 4.5. This Newton approximation scheme would then have the huge advantage the local model

$$d \mapsto F(X) + Md \tag{4.20}$$

is linear! Unfortunately, in general the linear model (4.19) does not satisfy the limit conditions (4.14b)–(4.14c). This is why we have to restrict ourselves to a subclass of locally Lipschitz-continuous functions. Semismooth functions are precisely the class of locally Lipschitz-continuous functions for which the generalized Jacobian furnishes a *bona fide* first-order approximation.

Definition 4.8 (Semismooth function)**.** Let $F : \mathcal{D} \subseteq \mathbb{R}^\ell \to \mathbb{R}^\ell$ be a locally Lipschitz-continuous function. We say that:

- F is *semismooth* at $\bar{X} \in \mathcal{D}$ if

$$\limsup_{\substack{X \to \bar{X} \\ M \in \partial F(X)}} \frac{\|F(X) + M(\bar{X} - X) - F(\bar{X})\|}{\|X - \bar{X}\|} = 0. \tag{4.21a}$$

- F is *strongly semismooth* at $\bar{X}$ if the above requirement is strengthened to

$$\limsup_{\substack{X \to \bar{X} \\ M \in \partial F(X)}} \frac{\|F(X) + M(\bar{X} - X) - F(\bar{X})\|}{\|X - \bar{X}\|^2} < \infty, \tag{4.21b}$$

The original definition of semismooth functions given in [92] and adopted in [47, §7.4.2] require F to be directionally differentiable at X. Here, following [110] we employ the equivalent definition (4.21a) in order to condense the narrative. For semismooth functions, the identification $\mathcal{T} \equiv \partial F$ by means of (4.19) is legitimate. Definition 4.8 may seem to rule out a lot of locally Lipschitz-continuous functions, but in fact the subclass of semismooth mappings is rich enough to include many functions of interest in real applications.

The semismooth Newton algorithm is described in Algorithm 4.2. In Step 3, we select a matrix M^k in $\partial F(X^k)$. As $\partial_B F(X^k) \subset \partial F(X^k)$, some authors [63] advocate picking M^k in $\partial_B F(X^k)$ instead, when it is difficult to identify the generic element of $\partial F(X^k)$. We will encounter an instance of this situation in §4.2.3 for the Newton-min method. If the matrix M^k is nonsingular, there is a unique solution d^k to the linear system (4.22). But there may be many choices for M^k if $\partial_B F(X^k)$ or $\partial F(X^k)$ is not a singleton. Note that the generalized Jacobian $\partial F(X^k)$ is a singleton if and only if F is differentiable at X^k. In this case $\partial F_B(X^k) = \partial F(X^k) = \{\nabla F(X^k)\}$ and we recover the smooth Newton method, at least for the current iteration.

Algorithm 4.2 Semismooth Newton algorithm

1. Choose $X^0 \in \mathcal{D} \subset \mathbb{R}^\ell$. Set $k = 0$.
2. If $F(X^k) = 0$, stop.
3. Select an element $M^k \in \partial F(X^k)$. Find a direction $d^k \in \mathbb{R}^\ell$ such that
$$F(X^k) + M^k d^k = 0. \tag{4.22}$$
4. Set $X^{k+1} = X^k + d^k$ and $k \leftarrow k + 1$. Go to step 2.

4.2.2.2 Well-definedness and convergence

For Algorithm 4.2 to be well-defined, the linear system (4.22) in the unknown d^k must be nonsingular at each iteration. The next Lemma guarantees that M^k is nonsingular provided that all the elements of the generalized Jacobians $\partial F(\bar{X})$ are nonsingular and that X^k is sufficiently close to $\bar{X}$.

Theorem 4.4. *Let $F : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^\ell$ be a semismooth function. Suppose that at a point $\bar{X} \in \mathcal{D}$, all the matrices of $\partial F(\bar{X})$ are nonsingular. Then, there exists a constant $\delta > 0$ such that, for any $X \in B(\bar{X}, \delta)$, all the matrices of $\partial F(X)$ are nonsingular.*

Chứng minh. This follows from the fact that the generalized Jacobian mapping $X \mapsto \partial F(X)$ is compact-valued and upper semicontinuous [47, §7.1.4], and from the technical result of [47, §7.5.2]. □

Now, we state a local convergence theorem with convergence rates for the semismooth Newton method. We recall that the notions of Q-superlinear and Q-quadratic convergence have been introduced in (4.17b)–(4.17c) [Definition 4.6].

Theorem 4.5. *Let $F : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^\ell$ be a semismooth function and let $\bar{X} \in \mathcal{D}$ be a solution of $F(\bar{X}) = 0$. If all the matrices of $\partial F(\bar{X})$ are nonsingular, then there exists a $\delta > 0$ such that, if $X^0 \in B(\bar{X}, \delta)$, the sequence $\{X^k\}_{k\in\mathbb{N}}$ generated by Algorithm 4.2 is well-defined and converges* Q-superlinearly *to $\bar{X}$. Furthermore, if F is strongly semismooth at $\bar{X}$, then the convergence rate is* Q-quadratic.

Chứng minh. See [47, §7.5.3]. □

4.2.3 Newton-min method

As an application of the previous theory, let us consider the mixed problem (4.1), in which the complementarity conditions are expressed by the min function. The system to be solved is $F(X) = 0$, with
$$F(X) = \begin{bmatrix} \Lambda(X) \\ \min(G(X), H(X)) \end{bmatrix}. \tag{4.23}$$

Proposition 4.2. *If Λ, G, $H : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^\ell$ are continuously differentiable, then F is semismooth.*

When F corresponds to the phase equilibrium problems (2.42) *or* (2.77)*, its B-subdifferential consists of all matrices $M \in \mathbb{R}^{\ell\times\ell}$ of the form*

$$\partial_B F(X) = \left\{ M = \begin{bmatrix} \nabla\Lambda(X) \\ \nabla\!\!\!\!\!\!\nabla \end{bmatrix}, \quad \nabla\!\!\!\!\!\!\nabla \in \mathbb{R}^{m\times\ell} \right\}, \tag{4.24a}$$

where the α-th row of $\nabla\!\!\!\!\!\!\nabla$ for $\alpha \in \{1,\dots,m\}$ is

$$\nabla\!\!\!\!\!\!\nabla_\alpha = \begin{cases} \nabla G_\alpha(X) & \text{if } G_\alpha(X) < H_\alpha(X), \\ \nabla G_\alpha(X) \text{ or } \nabla H_\alpha(X) & \text{if } G_\alpha(X) = H_\alpha(X), \\ \nabla H_\alpha(X) & \text{if } G_\alpha(X) > H_\alpha(X). \end{cases} \tag{4.24b}$$

Chứng minh. A smooth (i.e., continuously differentiable) function is also a semismooth function. The compononentwise minimum of two semismooth functions is a semismooth function [63, §1.75]. The second part of the Proposition can readily be proven by verifying Definition 4.7 or by applying more general results on the B-subdifferential of a vector-valued function [63, §1.54] and of the componentwise minimum mapping [63, §1.55].

For the latter result on the B-subdifferential of the min function, a technical condition is required: if $G_\alpha(X) = H_\alpha(X)$ for some $\alpha \in \{1,\dots,m\}$ and $X \in \mathcal{D}$, there must exist two sequences $\{\check{X}^k\}_{k\in\mathbb{N}*} \subset \mathcal{D}$ and $\{\hat{X}^k\}_{k\in\mathbb{N}*} \subset \mathcal{D}$ both converging to X such that $G_\alpha(\check{X}^k) < H_\alpha(\check{X}^k)$ and $G_\alpha(\hat{X}^k) > H_\alpha(\hat{X}^k)$ for all $k \in \mathbb{N}^*$. In the case of (2.42) or (2.77), this can be checked by a direct inspection of the equations. □

The corresponding semismooth Newton method, in which a matrix $M^k \in \partial_B F(X^k)$ is chosen to define the local model, is called the *Newton-min* algorithm and described in Algorithm 4.3. Note that, in this problem, it is not easy to work out an explicit form for the generic matrix of the Clarke subdifferential $\partial F(X^k)$.

Algorithm 4.3 Newton-min algorithm

1. Choose $X^0 \in \mathcal{D} \subset \mathbb{R}^\ell$. Set $k = 0$.
2. If $F(X^k) = 0$, stop.
3. Select an element $M^k \in \partial_B F(X^k)$ as in (4.24). Find a direction $d^k \in \mathbb{R}^\ell$ such that
$$F(X^k) + M^k d^k = 0. \tag{4.25}$$
4. Set $X^{k+1} = X^k + d^k$ and $k \leftarrow k+1$. Go to step 2.

By virtue of Theorem 4.5, the Newton-min algorithm converges if the initial iterate is close enough to a solution $\bar{X}$ of $F(X) = 0$, for which all the elements of $\partial F(\bar{X})$ are nonsingular. It is tempting to resort to a line search technique [22, 94] in an effort to ensure a *globally convergent* behavior, by which we mean that the sequence of iterates always converges to some limit (which is not necessarily the sought-after solution). The idea of line search is to apply a *damping factor* $\varsigma^k \in (0,1)$ to the Newton-min direction d^k determined in (4.25), so that the updated state in Step 4 is now

$$X^{k+1} = X^k + \varsigma^k d^k,$$

along with the guarantee that $\Theta(X^{k+1}) < \Theta(X^k)$ for some merit function whose minimum value is achieved precisely at the zero $\bar{X}$. More on this can be found in §4.3.1.3 for the smooth equations and in [47, §8.3 and §9.2] for nonsmooth equations.

Regarding the Newton-min method, the utmost difficulty is that the direction d^k computed by (4.25) is not always a descent direction for the least-squares merit function

$$\Theta(X) := \frac{1}{2}\|F(X)\|^2, \tag{4.26}$$

as observed by Ben Gharbia [11]. Globalization of Newton-min remains therefore a delicated issue. In this respect, a recent work by Dussault et al. [44] is worth mentioning, where the authors proposed a variant called the *polyhedral Newton-min* algorithm and for which some globalization process becomes possible.

4.3 Smoothing methods for nonsmooth equations

Instead of deploying a nonsmooth Newton method to solve nonsmooth equations, an alternative would be to approximate the nonsmooth system by a smooth one, to which a smooth Newton method with enhanced properties can be applied.

The privileged tool for producing a smooth approximation of a nonsmooth function is *regularization*, which usually introduces a small *regularization parameter*. Informally, let $F : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^\ell$ be the nonsmooth function for which we look for a zero $\bar{X} \in \mathcal{D}$ such that $F(\bar{X}) = 0$. A regularization of F is a family of functions

$$\left\{ \widetilde{F}(\cdot;\nu) : \ \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^\ell, \ \ \nu > 0 \right\} \tag{4.27}$$

such that

- $\widetilde{F}(\cdot;\nu)$ is a smooth (continuously differentiable) function of X, for all $\nu > 0$;
- $\widetilde{F}(\cdot;\nu)$ is continuous with respect to ν, in some functional sense;
- $\lim_{\nu \downarrow 0} \widetilde{F}(\cdot;\nu) = F(\cdot)$, in some functional sense.

Starting from a current pair of values (X^k, ν^k), the overall strategy of a smoothing method is to

1. Solve $\widetilde{F}(X^{k+1};\nu^k) = 0$ in the unknown X^{k+1} by means of the smooth Newton method, using X^k as the initial point.
2. Decrease the regularization parameter from ν^k to ν^{k+1} by some "rule of thumb." Start over the process until $F(X^{k+1}) = 0$.

If the nonlinear system in Step 1 is solved "exactly" by letting the smooth Newton algorithm go until convergence, the smoothing method is said to be *full Newton*. A full Newton resolution is in perfect agreement with the smoothing philosophy, which is to replace the original "difficult" problem by a sequence of "easier" problems and to gradually push the easy problem towards the difficult one. However, the price to be paid for the full Newton resolution is very expensive, since the full Newton method must be executed for each parameter ν. The computational cost can be lowered if the nonlinear system in Step 1 is solved "approximately" by letting the smooth Newton algorithm do just one iteration. In this case, the method is said to be *diagonal Newton*.

The diagonal Newton resolution naturally induces more approximation error, but it is obviously of great practical interest.

Although we shall not consider full Newton smoothing methods in this work, we take this opportunity to briefly survey the smooth Newton method and the numerous convergence theorems associated with it in §4.3.1. Then, in §4.3.2, we review a family of smoothing functions called *θ-functions* for complementarity conditions. Finally, in §4.3.3, we turn our attention to interior-point methods, from which a new method will be designed in chapter §5.

4.3.1 Newton's method

For conciseness, we shall be using the notation F instead of $\widetilde{F}(\cdot, \nu)$ for the smoothed out function at some fixed parameter $\nu > 0$.

4.3.1.1 Algorithm

The idea of Newton's method is to construct a sequence $\{X^k\}_{k\in\mathbb{N}*}$ by successively linearizing the equation $F(X) = 0$ at the current iterate by invoking the first-order local model

$$X \mapsto F(X^k) + \nabla F(X^k)(X - X^k) \tag{4.28a}$$

to approximate $F(X)$ when X is close to X^k. The local model can equivalently be thought of as the mapping

$$d \mapsto F(X^k) + \nabla F(X^k)d, \tag{4.28b}$$

meant to approximate $F(X^k + d)$ for $\|d\|$ small. Our purpose is then shifted to looking for the zero of the local model (4.28b). If the Jacobian matrix $\nabla F(X^k)$ is invertible, the unique zero of (4.28b) can be seen to be

$$d^k = -[\nabla F(X^k)]^{-1}F(X^k), \tag{4.29a}$$

so that the new iterate is

$$X^{k+1} = X^k - [\nabla F(X^k)]^{-1}F(X^k). \tag{4.29b}$$

The sequence (4.29b) is said to be *well-defined* if at each iteration k, the matrix $\nabla F(X^k)$ is invertible and the updated state X^{k+1} remains in the domain $\mathcal{D}$ of F. For later analysis, it is convenient to introduce another concept.

Definition 4.9 (Newton direction)**.** At any point $X \in \mathcal{D}$ where the Jacobian matrix $\nabla F(X)$ is invertible, the vector

$$d(X) = -[\nabla F(X)]^{-1}F(X) \tag{4.30}$$

is called the *Newton direction* for F at X.

Using this notation, the Newton method can be written as

$$X^{k+1} = X^k + d(X^k). \tag{4.31}$$

The two issues to be addressed now relate to the well-definedness of the sequence $\{X^k\}_{k\in\mathbb{N}*}$ and its (local and global) convergence. With respect to local convergence, several classical theorems are at our disposal. Below we go through some of them, emphasizing their differences.

4.3.1.2 Local convergence analysis

The first theorem is what we qualify as the *regular-zero Newton* theorem. To state this theorem, we need the following definition.

Definition 4.10 (Regular zero)**.** Let $\bar{X} \in \mathcal{D} \subset \mathbb{R}^\ell$ be a zero of F, that is, $F(\bar{X}) = 0$. If the Jacobian matrix $\nabla F(\bar{X})$ is nonsingular, $\bar{X}$ is said to be a *regular zero* of F.

In the scalar case $\ell = 1$, a regular zero means a simple zero. The regular-zero Newton theorem assumes that a regular zero $\bar{X}$ exists, together with the Lipschitz-continuity of the Jacobian mapping $X \mapsto \nabla F(X)$ in a neighborhood of $\bar{X}$. The conclusion is that if the initial point X^0 is close enough to the solution $\bar{X}$, then the iterates are well-defined and converge Q-quadratically. The constant involved in this Q-quadractic convergence is the product of the norm β of the inverse $[\nabla F(\bar{X})]^{-1}$ with the Lipschitz modulus γ of ∇F.

Theorem 4.6 (regular-zero Newton)**.** *Let $F : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^\ell$ be continuously differentiable on the open convex domain $\mathcal{D}$. Assume that there exists a regular zero $\bar{X} \in \mathcal{D}$, i.e.,*

$$F(\bar{X}) = 0, \qquad \det \nabla F(\bar{X}) \neq 0,$$

and that there exist $\overline{r}$, β, $\gamma > 0$ such that

$$B(\bar{X}, \overline{r}) \subset \mathcal{D}, \qquad \| \, [\nabla F(\bar{X})]^{-1} \, \| \leqslant \beta, \qquad \nabla F \in \mathrm{Lip}_\gamma(B(\bar{X}, \overline{r})). \tag{4.32}$$

Then, there exists $\bar{\varepsilon} > 0$ such that, for all $X^0 \in B(\bar{X}, \bar{\varepsilon})$, the sequence $\{X^k\}_{k \in \mathbb{N}^*}$ generated by (4.29b) *is well-defined, converges to $\bar{X}$ and obeys*

$$\|X^{k+1} - \bar{X}\| \leqslant \beta\gamma \, \|X^k - \bar{X}\|^2. \tag{4.33}$$

Chứng minh. See [41, §5.2]. □

There is another famous convergence theorem for Newton's method, due to Kantorovich. Contrary to the regular-zero theorem, the Newton-Kantorovich theorem does not make any requirement about the existence of a zero $\bar{X}$. Its assumptions are rather focused on the initial point X^0. It asserts that if $\nabla F(X^0)$ is nonsingular, ∇F is Lipschitz-continuous in a neighborhood of X^0, and the first Newton step is small enough relative to the nonlinearity of F, then there must be a root $\bar{X}$ in this region, and furthermore it is unique. In exchange for these broader hypotheses, the rate of convergence is slightly weaker: it is only *R-quadratic* instead of Q-quadratic. This means that the error sequence can be bounded by $\|X^k - \bar{X}\| \leqslant \rho^k$, where $\{\rho^k\}_{k \in \mathbb{N}^*}$ converges Q-quadratically to zero.

Theorem 4.7 (Newton-Kantorovich)**.** *Let $F : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^\ell$ be continuously differentiable on the open convex domain $\mathcal{D}$. Assume that $\nabla F(X^0)$ is nonsignular and that there exist r^0, β, $\gamma > 0$ and $\eta \geqslant 0$ such that*

$$\beta\gamma\eta \leqslant \frac{1}{2}, \qquad r^0 \geqslant r_- := \frac{1 - \sqrt{1 - 2\beta\gamma\eta}}{\beta\gamma}, \qquad B(X^0, r^0) \subset \mathcal{D}, \tag{4.34a}$$

and

$$\|[\nabla F(X^0)]^{-1}\| \leqslant \beta, \qquad \|[\nabla F(X^0)]^{-1} F(X^0)\| \leqslant \eta, \qquad \nabla F \in \mathrm{Lip}_\gamma(B(X^0, r^0)). \tag{4.34b}$$

Then, there is a unique root $\bar{X}$ *of* F *in* $\overline{B}(X^0, r_-)$. *The sequence* $\{X^k\}_{k\in\mathbb{N}*}$ *generated by* (4.29b) *is well-defined and converges to* $\bar{X}$. *If* $\beta\gamma\eta < 1/2$, *then* $\bar{X}$ *is also the unique zero of* F *in* $\overline{B}(X^0, \min(r^0, r_+))$, *where*

$$r_+ := \frac{1 - \sqrt{1 + 2\beta\gamma\eta}}{\beta\gamma},$$

and the sequence of iterates obeys

$$\|X^k - \bar{X}\| \leqslant \frac{(2\beta\gamma\eta)^{2^k}}{2^k\beta\gamma}. \tag{4.35}$$

Chứng minh. See [41, §5.3] and [68, §5.5]. □

The Newton-Mysovskikh[3] theorem, a third one, resembles the Newton-Kantorovich theorem in that: (i) it does not make any requirement about the existence of a solution, and (ii) it assumes that the first Newton step is sufficiently small. However, it differs from the Newton-Kantorovich in three aspects. Firstly, it explicitly makes the stronger assumption on the invertibility of $\nabla F(X)$ in a neighborhood of the initial point X^0. Secondly, it ensures the existence of a zero $\bar{X}$ but does not make any claim about uniqueness. Thirdly, it supplies us not only with a nearly quadractic rate of convergence (4.37a), but also with an *a posteriori* error estimate (4.37b) for the current iterate. Indeed, the upperbound in (4.37b) can be computed even if the exact solution $\bar{X}$ is not known, but provided that the various constants are known.

Theorem 4.8 (Newton-Mysovskikh)**.** *Let* $F : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^\ell$ *be continuously differentiable on the open convex domain* $\mathcal{D}$. *Assume that there exist* r^0, β, $\gamma > 0$ *and* $\eta \geqslant 0$ *such that*

$$\beta\gamma\eta < 2, \qquad r^0 \geqslant \frac{\eta}{1 - \frac{1}{2}\beta\gamma\eta}, \qquad B(X^0, r^0) \subset \mathcal{D}, \tag{4.36a}$$

and

$$\|[\nabla F(X)]^{-1}\| \leqslant \beta, \qquad \|[\nabla F(X^0)]^{-1}F(X^0)\| \leqslant \eta, \qquad \nabla F \in \mathrm{Lip}_\gamma(B(X^0, r^0)), \tag{4.36b}$$

for all $X \in B(X^0, r^0)$. *Then, the sequence* $\{X^k\}_{k\in\mathbb{N}*}$ *generated by* (4.29b) *is well-defined and converges to a zero* $\bar{X} \in \overline{B}(X^0, r^0)$ *of* F. *Moreover, the sequence of iterates obeys*

$$\|X^k - \bar{X}\| \leqslant \frac{\eta\left(\frac{1}{2}\beta\gamma\right)^{2^k-1}}{1 - \left(\frac{1}{2}\beta\gamma\eta\right)^{2^k}}, \tag{4.37a}$$

$$\|X^k - \bar{X}\| \leqslant \frac{\frac{1}{2}\beta\gamma}{1 - \left(\frac{1}{2}\beta\gamma\eta\right)^{2^k}}\|X^k - X^{k-1}\|^2. \tag{4.37b}$$

Chứng minh. See [97, §12.4.6] □

There are many possible improvements and other theorems born out of other motivations. A summary can be found in [52, 120], along with the historical developments of the convergence theory of smooth Newton's method. Another series of results, due to Deuflhard [42], is worth mentioning. Deuflhard investigated the question of *affine invariance* for the convergence theorems of Newton's method. This questions comes from the observation that Newton's method is

[3] also spelled "Mysovskii"

invariant with respect to affine transformation of the variables and of the equations. In other words, let M_X and M_F are two invertible matrices of $\mathbb{R}^{\ell\times\ell}$. If we perform the change of variables $X = M_X\widehat{X}$ (e.g., by adopting other units of measurements) and the change of equations $\widehat{F} = M_F F$ (e.g., by rescaling the laws of physics), the Newton iterates for the system

$$\widehat{F}(\widehat{X}) = M_F F(M_X\widehat{X}) = 0 \tag{4.38}$$

are

$$\begin{aligned}\widehat{X}^{k+1} &= \widehat{X}^k - [\nabla_{\widehat{X}}\widehat{F}(\widehat{X}^k)]^{-1}\widehat{F}(\widehat{X}^k)\\ &= \widehat{X}^k - [M_F\nabla_X F(M_X\widehat{X}^k)M_X]^{-1}M_F F(M_X\widehat{X}^k)\\ &= \widehat{X}^k - M_X^{-1}[\nabla_X F(M_X\widehat{X^k})]^{-1}F(M_X\widehat{X}^k),\end{aligned}$$

and exactly match those of the original system $F(X) = 0$, up to the same transformation of variable M_X. Affine invariance of the Newton method is fundamental for industrial codes to be robust with respect to a change of units in the quantities computed and to a rescaling of the equations.

However, the convergence theorems such as Newton-Kantorovich and Newton-Mysovskikh are not affine invariant, in that they are not automatically preserved by affine transformations. In fact, the inequaltities still hold but with a different set of constants which can be much more defavorable than the original ones. It is therefore a serious research topic to find an affine-invariant formulation for the classical theorems. Below we write down the *affine covariant* versions of the last two theorems. By "affine covariant," we mean invariance with respect to an arbitrary rescaling M_F of the equations (but no change of variables is considered, i.e., $M_X = I$).

Theorem 4.9 (affine covariant Newton-Kantorovich)**.** *Let $F : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^\ell$ be a continuously differentiable on the open convex domain $\mathcal{D}$. Assume that $\nabla F(X^0)$ is nonsignular and that there exist r^0, $\omega > 0$ and $\eta \geqslant 0$ such that*

$$\omega\eta \leqslant \frac{1}{2}, \qquad r^0 \geqslant r_- := \frac{1-\sqrt{1-2\omega\eta}}{\omega}, \qquad B(X^0, r^0) \subset \mathcal{D}, \tag{4.39a}$$

and

$$\|[\nabla F(X^0)]^{-1}F(X^0)\| \leqslant \eta, \qquad \frac{\|[\nabla F(X^0)]^{-1}(\nabla F(\check{X}) - \nabla F(X))\|}{\|\check{X} - X\|} \leqslant \omega, \tag{4.39b}$$

for all $X \neq \check{X} \in B(X^0, r^0)$. Then, the sequence $\{X^k\}_{k\in\mathbb{N}^*}$ generated by (4.29b) *is well-defined and converges to a zero $\overline{X} \in \overline{B}(X^0, r_-)$ of F. If $\omega\eta < 1/2$, the convergence is R-quadratic.*

Chứng minh. See [42, Theorem 2.1]. □

In comparison with the formulation of Theorem 4.8, the two conditions $\|[\nabla F(X^0)]^{-1}\| \leqslant \beta$ and $\nabla F \in \mathrm{Lip}_\gamma(B(X^0, r^0))$ have been merged into the single covariant condition

$$\|[\nabla F(X^0)]^{-1}(\nabla F(\check{X}) - \nabla F(X))\| \leqslant \omega\|\check{X} - X\|.$$

Likewise, the constants β and γ have been telescoped into the single constant ω. In the same spirit, we have the following reformulation of Theorem 4.8.

Theorem 4.10 (affine covariant Newton-Mysovskikh)**.** *Let $F : \mathcal{D} \subset \mathbb{R}^\ell \to \mathbb{R}^\ell$ be continuously differentiable on the open convex domain $\mathcal{D}$. Assume that there exist r^0, $\omega > 0$ and $\eta \geqslant 0$ such that*

$$\omega\eta < 2, \qquad r^0 \geqslant \frac{\eta}{1 - \frac{1}{2}\omega\eta}, \qquad B(X^0, r^0) \subset \mathcal{D}, \tag{4.40a}$$

and

$$\|[\nabla F(X^0)]^{-1} F(X^0)\| \leqslant \eta, \qquad \frac{\|[\nabla F(\widetilde{X})]^{-1}(\nabla F(\check{X}) - \nabla F(X))(\check{X} - X)\|}{\|\check{X} - X\|^2} \leqslant \omega, \tag{4.40b}$$

for all collinear $X \neq \check{X}$, $\widetilde{X} \in B(X^0, r^0)$. Then, the sequence $\{X^k\}_{k \in \mathbb{N}*}$ generated by (4.29b) is well-defined and converges to a zero $\bar{X} \in \overline{B}(X^0, r^0)$ of F. Moreover, the sequence of iterates obeys

$$\|X^{k+1} - X^k\| \leqslant \tfrac{1}{2}\omega\|X^k - X^{k-1}\|^2, \tag{4.41a}$$

$$\|X^k - \bar{X}\| \leqslant \frac{\|X^{k+1} - X^k\|}{1 - \frac{1}{2}\omega\|X^{k+1} - X^k\|}. \tag{4.41b}$$

Chứng minh. See [42, Theorem 2.2]. □

4.3.1.3 Globalization with line search

The above Theorems testify to an excellent theoretical rate of convergence in a neighborhood of a zero of F. However, Newton's method may fail to converge if the starting point is too far from the desired zero. In an effort to improve its behavior, we can use a *globalization* strategy such as line search or trust region [32]. In this thesis, we will focus on the line search technique.

To this end, let us consider the least-squares potential

$$\Theta(X) := \frac{1}{2}\|F(X)\|^2 \tag{4.42}$$

as the *merit* function. If there exists a zero $\bar{X} \in \mathcal{D}$ of F, then $\min_{X \in \mathcal{D}} \Theta(X) = 0$ and $\bar{X}$ is a solution of the minimization problem whose objective function is Θ. The next Lemma provides the gradient $\nabla\Theta$, represented as a row vector, and defines a descent direction for Θ.

Lemma 4.1. *If F is smooth, then the function Θ defined by (4.42) is also smooth and*

$$\nabla\Theta(X)\, d = \langle F(X), \nabla F(X) d \rangle$$

for all $d \in \mathbb{R}^\ell$, where $\langle \cdot, \cdot \rangle$ denotes the dot product in $\mathbb{R}^\ell$. In particular, if $\nabla F(X)$ is invertible, then the Newton direction (4.30) exists and

$$\nabla\Theta(X)\, d(X) = -2\Theta(X). \tag{4.43}$$

We say that $d \in \mathbb{R}^\ell$ is a descent direction for Θ at X if $\nabla\Theta(X)d < 0$. Lemma 4.1 implies that Newton direction, whenever it is well-defined, is a descent direction for the least-squares potential. The idea is then to replace the full increment $d(X^k)$ by $\varsigma^k d(X^k)$, where $\varsigma^k \in (0, 1)$, in the update formula (4.31), which yields the *damped* Newton iteration

$$X^{k+1} = X^k + \varsigma^k d(X^k).$$

The flexibility of being able to choose $\varsigma^k \in (0,1)$ is vital for global convergence. Usually, this damping parameter is selected so as to decrease the potential, i.e.,

$$\Theta(X^k + \varsigma^k d(X^k)) \leqslant \Theta(X^k) + \{\text{some negative term}\}.$$

This will always be possible for $\varsigma^k > 0$ small enough, because the Newton direction is a descent direction. Nevertheless, it is interesting to take ς^k as close to 1 as possible, in order to benefit from superlinear convergence.

Algorithm 4.4 sketches out the Newton method with a line search technique due to Armijo [5]. This technique rests upon a backtracking procedure described in Step 4, the purpose of which is to meet the Armijo condition

$$\Theta(X^k + \varsigma^k d^k) \leqslant \Theta(X^k) + \kappa\varsigma^k \left[\nabla\Theta(X^k)\, d^k\right] \tag{4.44}$$

for some constant $\kappa \in (0, 1/2)$. For the least-squares potential (4.42), the Armijo condition is equivalent to (4.46).

Algorithm 4.4 Newton algorithm with Armijo line search

1. Choose $X^0 \in \mathcal{D} \subset \mathbb{R}^\ell$, $\kappa \in (0, 1/2)$, $\varrho \in (0,1)$. Set $k = 0$.
2. If $F(X^k) = 0$, stop.
3. Find a direction $d^k \in \mathbb{R}^\ell$ such that
$$F(X^k) + \nabla F(X^k) d^k = 0. \tag{4.45}$$
4. Choose $\varsigma^k = \varrho^{j_k} \in (0,1)$, where $j_k \in \mathbb{N}$ is the smallest integer such that
$$\Theta(X^k + \varrho^{j_k} d^k) \leqslant \Theta(X^k) + \kappa\varsigma^k \left[\nabla\Theta(X^k)\, d^k\right] = (1 - 2\kappa\varrho^{j_k})\, \Theta(X^k). \tag{4.46}$$
5. Set $X^{k+1} = X^k + \varsigma^k d^k$ and $k \leftarrow k+1$. Go to step 2.

It has to be pointed out that there are many other possible conditions [94, §3.1] for line search, such as

- the Wolfe conditions:
$$\begin{aligned} \Theta(X^k + \varsigma^k d^k) &\leqslant \Theta(X^k) + \kappa\varsigma^k [\nabla\Theta(X^k)\, d^k], \\ \nabla\Theta(X^k + \varsigma^k d^k)\, d^k &\geqslant \lambda\, [\nabla\Theta(X^k)\, d^k], \end{aligned}$$
where $0 < \kappa < \lambda < 1$.
- the Goldstein conditions:
$$\Theta(X^k) + (1-\kappa)\varsigma^k [\nabla\Theta(X^k)\, d^k] \leqslant \Theta(X^k + \varsigma^k d^k) \leqslant \Theta(X^k) + \kappa\varsigma^k [\nabla\Theta(X^k)\, d^k],$$
where $0 < \kappa < 1/2$.

The theoretical advantage of the Wolfe or Goldstein conditions is that by Zoutendijk's theorem [94, Theorem 3.2], it can be guaranteed that, for any initial point X^0, the sequence of iterates is *globally convergent* in the sense that

$$\lim_{k\to+\infty} \|\nabla\Theta(X^k)\| = 0.$$

It is important to be aware that this does not mean that the iterates converge to a minimizer $\bar{X}$, but only that they are attracted by stationary points. For stronger global results, we need stronger assumptions. In practice, however, our preference goes to Armijo's condition and the associated backtracking procedure for its greater efficiency, despite the lack of theoretical results.

4.3.2 Smoothing functions for complementarity conditions

In the previous section, we revisisted the smooth Newton method for computing a root of the regularized function $\widetilde{F}(\cdot;\nu)$ hypothetically set up in (4.27). In this section, we elaborate on how such a regularized function can be actually built up from the initial function (4.23). Our smoothing technique is based on the continuous approximation of a more elementary object, namely, the step function.

4.3.2.1 θ-smoothing of the step function

The *step function* is understood to be the function $\mathfrak{S} : \mathbb{R}_+ \to \{0,1\}$ defined as

$$\mathfrak{S}(t) = \begin{cases} 0 & \text{if } t = 0, \\ 1 & \text{if } t > 0. \end{cases} \tag{4.47}$$

As an indicator of positive arguments $t > 0$ over $\mathbb{R}_+$, the step function $\mathfrak{S}$ "discriminates" the argument $t = 0$ by assigning a zero value to it. The price to be paid for this sharp detection is the discontinuity of $\mathfrak{S}$ at $t = 0$. Analogously to (4.27), we wish to have a regularization of $\mathfrak{S}$, that is, a family of functions

$$\big\{\, \widetilde{\mathfrak{S}}(\cdot;\nu) : \mathbb{R}_+ \to [0,1), \quad \nu > 0 \,\big\}, \tag{4.48}$$

such that

- $\widetilde{\mathfrak{S}}(\cdot;\nu)$ is a smooth function of $t \geqslant 0$, for all $\nu > 0$;
- $\widetilde{\mathfrak{S}}(\cdot;\nu)$ is continuous with respect to ν, in some functional sense;
- $\lim_{\nu\downarrow 0} \widetilde{\mathfrak{S}}(\cdot;\nu) = \mathfrak{S}(\cdot)$, in some functional sense.

To obtain such a family, we follow the methodology developed by Haddou and his coauthors [7, 55], the key ingredient of which is a smoothing function. This notion turned out to be a versatile tool in a wide variety of pure and applied mathematical problems [19, 56, 57, 93]. We begin with a "father" function, from which all other regularized functions will be generated.

Definition 4.11 (θ-smoothing function)**.** A function $\theta : \mathbb{R}_+ \to [0,1)$ is said to be a *θ-smoothing function* if it is continuous, nondecreasing, concave, and

$$\theta(0) = 0, \tag{4.49a}$$

$$\lim_{t\to+\infty} \theta(t) = 1. \tag{4.49b}$$

Furthermore, if θ can be defined for negative arguments $t \in (-T, 0)$, with $T > 0$, while remaining continuous, nondecreasing and concave, it is required that

$$\theta(t) < 0 \quad \text{for } t \in (-T, 0). \tag{4.49c}$$

The two most common examples of smoothing functions are:

1. the rational function $\theta^1 : (-1, +\infty) \to (-\infty, 1)$ defined by

$$\theta^1(t) = \frac{t}{t+1}. \tag{4.50a}$$

2. the exponential function $\theta^2 : \mathbb{R} \to (-\infty, 1)$ defined by

$$\theta^2(t) = 1 - \exp(-t). \tag{4.50b}$$

A more general "recipe" to build θ-smoothing functions is to consider nondecreasing probability density functions $f : \mathbb{R}_+ \to \mathbb{R}_+$ and then take the corresponding cumulative distribution function, i.e.,

$$\theta(t) = \int_0^t f(y)\,\mathrm{d}y, \qquad t \geqslant 0, \tag{4.51}$$

to get a continuous, nondecreasing function. The nonincreasing assumption on f gives the concavity of θ. Finally, it is straightforward to check that conditions (4.49) are satisfied by the function (4.51). Once a favorite θ-smoothing has been selected, the next step is to dilate and to compress it in order to produce a family of regularized functions for the step function $\mathfrak{S}$.

Definition 4.12 (θ-smoothing family)**.** Let θ be a θ-smoothing function. The family of functions

$$\left\{ \theta_\nu(t) := \theta\left(\frac{t}{\nu}\right), \;\; \nu > 0 \right\} \tag{4.52}$$

is said to be the θ-smoothing family associated with θ.

Obviously, θ_ν is a smooth function of $t \geqslant 0$ for all $\nu > 0$. It is also continuous with respect to ν at each fixed $t \geqslant 0$. From the defining properties (4.49), it can be readily shown that

$$\lim_{\nu\downarrow 0} \theta_\nu(t) = \mathfrak{S}(t), \quad \forall t \geqslant 0. \tag{4.53}$$

In other words, $\mathfrak{S}$ is the limit of θ_ν in the sense of pointwise convergence. Thus, $\{\mathfrak{S}(\cdot, \nu) = \theta_\nu,\ \nu > 0\}$ is a good family of regularized functions in the sense of (4.48). Associated with the two examples (4.50) are:

1. the rational family $\theta^1_\nu : (-\nu, +\infty) \to (-\infty, 1)$ defined by

$$\theta^1_\nu(t) = \frac{t}{t+\nu}. \tag{4.54a}$$

2. the exponential family $\theta^2 : \mathbb{R} \to (-\infty, 1)$ defined by

$$\theta^2_\nu(t) = 1 - \exp(-t/\nu). \tag{4.54b}$$

Figures 4.1–4.2 display the two families (4.54) for a few parameters ν. We can see that the smaller ν is, the steeper is the slope at $t = 0$ and the closer to $\mathfrak{S}$ the function is.

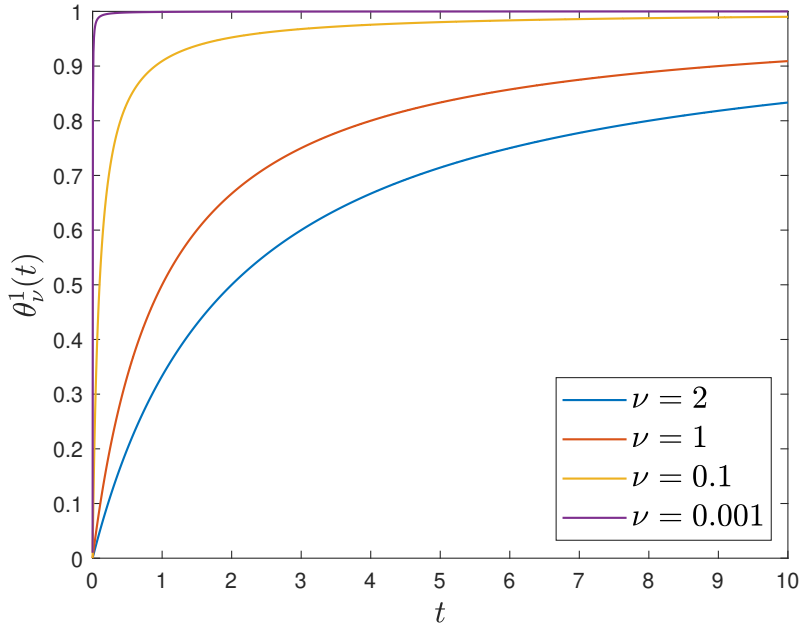


Figure 4.1: Function θ_ν^1 for a few values of ν.

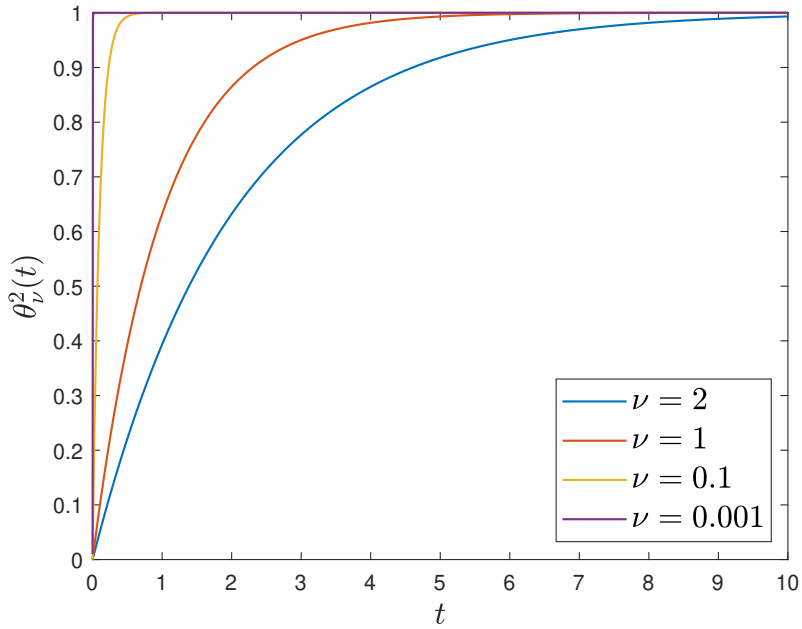


Figure 4.2: Function θ_ν^2 for a few values of ν.

4.3.2.2 θ-smoothing of a complementarity condition

A θ-smoothing family of the step function paves the way for a smooth approximation of a complementarity condition. Let $(v, w) \in \mathbb{R}^2$ be two scalars such that

$$0 \leqslant v \perp w \geqslant 0, \tag{4.55a}$$

that is,

$$v \geqslant 0, \qquad w \geqslant 0, \qquad vw = 0. \tag{4.55b}$$

In the (v, w)-plane, the set of points obeying (4.55) is the union of the two semi-axes $\{v \geqslant 0,\ w = 0\}$ and $\{v = 0,\ w \geqslant 0\}$. Visually, the nonsmoothness of (4.55) is manifested by the "kink" at the corner $(v, w) = (0, 0)$. We consider two possible smooth approximations of (4.55), depending how it is rewritten in terms of the step function $\mathfrak{S}$.

"Sum-to-one." The first approximation, to which we give the name *sum-to-one*, comes from the following observation.

Lemma 4.2. *Assuming $v \geqslant 0$ and $w \geqslant 0$, we have the equivalence*

$$vw = 0 \ \Leftrightarrow\ \mathfrak{S}(v) + \mathfrak{S}(w) \leqslant 1. \tag{4.56}$$

Chứng minh. If $v = 0$ for instance, then $\mathfrak{S}(v) + \mathfrak{S}(w) = \mathfrak{S}(w) \in \{0, 1\}$ because $\mathfrak{S}(0) = 0$. This proves "$\Rightarrow$". Conversely, the inequality $\mathfrak{S}(v) + \mathfrak{S}(w) \leqslant 1$ forbids $\mathfrak{S}(v) + \mathfrak{S}(w)$ to take the value 2. But this is precisely the value reached by the sum when $v > 0$ and $w > 0$. This proves "$\Leftarrow$". □

The equivalence (4.56) suggests us to impose

$$v \geqslant 0, \qquad w \geqslant 0, \qquad \theta_\nu(v) + \theta_\nu(w) = 1 \tag{4.57}$$

for $\nu > 0$, as a smooth approximation of (4.55). Replacing $\mathfrak{S}$ by θ_ν in (4.56) is logical. Replacing "$\leqslant$" by "$=$" in (4.56) seems to be a bold move, but this is motivated by the fact that we want an equality to be mounted into the system of equations. Let us examine the impact of this "sum-to-one" approach on the examples (4.54).

1. For the rational family (4.54a), we have the remarkable equivalence

$$\theta^1_\nu(v) + \theta^1_\nu(w) = 1 \ \Leftrightarrow\ vw = \nu^2, \tag{4.58a}$$

   as can be shown by a straightforward calculation. The equality $vw = \nu^2$ appears to be a natural relaxation of $vw = 0$, which could have been worked out directly, without resorting any θ-smoothing function! As will be seen later, this is the smoothing paradigm used in interior-point methods, with ν in the right-hand side instead of ν^2 though.

2. For the exponential family (4.54b), the equality $\theta^2_\nu(v) + \theta^2_\nu(w) = 1$ leads to the equivalence

$$\theta^2_\nu(v) + \theta^2_\nu(w) = 1 \ \Leftrightarrow\ -\nu \ln[\exp(-v/\nu) + \exp(-w/\nu)] = 0, \tag{4.58b}$$

   as can be shown by a straightforward calculation. In the left-hand side, the function

$$\min\nolimits_\nu(v, w) := -\nu \ln[\exp(-v/\nu) + \exp(-w/\nu)]$$

can be interpreted as a smooth approximation of $\min(v, w)$. Indeed, assuming $0 \leqslant v < w$, we can see that $\exp(-v/\nu)$ prevails over the other term when $\nu \downarrow 0$. Therefore, this is a natural relaxation of $\min(v, w) = 0$. It could have been worked out without the help of any θ-smoothing function, since the smooth approximation

$$\max_\nu(v, w) := \nu \ln[\exp(v/\nu) + \exp(w/\nu)]$$

of $\max(v, w)$ is rather well-known in the literature [10, 18].

"Sum-to-theta." The second smooth approximation, which we refer to as *sum-to-theta*, comes from another observation.

Lemma 4.3. *Assuming $v \geqslant 0$ and $w \geqslant 0$, we have the equivalence*

$$vw = 0 \;\Leftrightarrow\; \mathfrak{S}(v) + \mathfrak{S}(w) = \mathfrak{S}(v + w). \tag{4.59}$$

Chứng minh. If $v = 0$ for instance, the left-hand side is equal to $\mathfrak{S}(0) + \mathfrak{S}(w) = \mathfrak{S}(w)$, while the right-hand is also equal to $\mathfrak{S}(w)$. This proves "⇒". Conversely, the equality $\mathfrak{S}(v) + \mathfrak{S}(w) = \mathfrak{S}(v + w)$ prevents the sum $\mathfrak{S}(v) + \mathfrak{S}(w)$ from being equal to 2. But this is precisely the value reached by the sum when $v > 0$ and $w > 0$. This proves "⇐". □

The equivalence (4.59) suggests us to impose

$$v \geqslant 0, \qquad w \geqslant 0, \qquad \theta_\nu(v) + \theta_\nu(w) = \theta_\nu(v + w) \tag{4.60}$$

for $\nu > 0$, as a smooth approximation of (4.55). But this time, the approximation turns out to be exact, as demonstrated by the following Proposition.

Proposition 4.3. *For all $\nu > 0$ and for all $v, w \geqslant 0$, we have*

$$\theta_\nu(v) + \theta_\nu(w) \geqslant \theta_\nu(v + w). \tag{4.61}$$

Equality holds if and only if $vw = 0$.

Chứng minh. By the concavity of θ_ν which follows from that of θ, we have

$$\theta_\nu(\gamma t) = \theta_\nu(\gamma t + (1 - \gamma)0) \geqslant \gamma\, \theta_\nu(v) + (1 - \gamma)\, \theta_\nu(0) = \gamma\, \theta_\nu(t) \tag{4.62}$$

for all $\gamma \in [0, 1]$ and $t \in \mathbb{R}_+$, with equality if and only if $\gamma \in \{0, 1\}$ or $t = 0$. If $v = w = 0$, inequality (4.61) is obvious, since $\theta_\nu(0) = 0$. Assume that at least one of the two quantities v, w is positive, so that $v + w > 0$. Then, owing to (4.62),

$$\begin{aligned}
\theta_\nu(v) + \theta_\nu(w) &= \theta_\nu\Big(\frac{v}{v + w}(v + w)\Big) + \theta_\nu\Big(\frac{w}{v + w}(v + w)\Big) \\
&\geqslant \frac{v}{v + w}\, \theta_\nu(v + w) + \frac{w}{v + w}\, \theta_\nu(v + w) \\
&= \theta_\nu(v + w).
\end{aligned}$$

Equality holds if and and only if $v = 0$ or $w = 0$. This completes the proof. □

The exactness of the sum-to-theta smoothing (4.60) looks attractive at first sight. A close inspection reveals, however, that this comes at the price of a singularity at $(v, w) = (0, 0)$. Indeed, let

$$\psi_\nu(v, w) = \theta_\nu(v) + \theta_\nu(w) - \theta_\nu(v + w).$$

Then, $\nabla\psi_\nu(0, 0) = (0, 0)$, where the gradient is taken with respect to v, w. This phenomenon is similar to what was already pointed out for the Mangasarian C-function in §4.1.2. Let us contemplate the impact of this "sum-to-theta" approach on the examples (4.54).

1. For the rational family (4.54a), we have the equivalence

$$\theta^1_\nu(v) + \theta^1_\nu(w) = \theta^1_\nu(v + w) \;\Leftrightarrow\; vw(v + w + 2\nu) = 0, \tag{4.63}$$

   which follows from a simple but tedious calculation. The latter condition is also equivalent to the exact condition $vw = 0$.

2. For the exponential family (4.54b), we have the remarkable equivalence

$$\theta^2_\nu(v) + \theta^2_\nu(w) = \theta^2_\nu(v + w) \;\Leftrightarrow\; [1 - \exp(-v/\nu)][1 - \exp(-w/\nu)] = 0, \tag{4.64}$$

   as can be shown by a factorization procedure. From the latter, we can see the exactness of the smoothing in this particular case.

4.3.2.3 Integration into the system of equations

The "sum-to-theta" smoothing approach will not be used in this thesis due to this singularity. Restricting ourselves to the "sum-to-one" approximation, we consider the family $\{\widetilde{F}(\cdot, \nu),\ \nu > 0\}$, where

$$\widetilde{F}(X\,, \nu) = \begin{bmatrix} \Lambda(X) \\ \nu\,(\theta_\nu(G(X)) + \theta_\nu(H(X)) - \mathbf{1}) \end{bmatrix} \tag{4.65}$$

is a regularized function of F defined in (4.23). Here, it is understood that θ_ν operates componentwise on $G(X)$ and $H(X)$, while $\mathbf{1} \in \mathbb{R}^m$ is the vector whose entries are all equal to 1. It is highly recommended that the smoothed complementarity equations in (4.65) be premultiplied by ν, so as to control the magnitude of their partial derivatives. Indeed, for all $t \geqslant 0$,

$$\theta'_\nu(t) = \frac{1}{\nu}\theta'\left(\frac{t}{\nu}\right)$$

can be seen to blow up when $\nu \downarrow 0$, while $\nu\theta'_\nu(t)$ tends to the finite limit $\theta'(0)$.

4.3.3 Standard and modified interior-point methods

The general philosophy of smoothing, presented at the beginning of §4.3, also lies at the heart of an important category of algorithms for constrainted optimization called *interior-point methods*. Despite some pioneering work in 1967 by Dikin [43], which remained unknown for a long time, the field really took off in 1984 with the publication by Karmakar [67] of an algorithm with polynomial-time complexity for linear programming, capable of outperforming even the simplex method. For a historical review of interior-point methods, see [54, 118].

Karmakar's algorithm is *primal*, i.e., it is crafted only in terms of primal variables, without any reference to the dual problem. It was not long before theoreticians realized the power and

the superiority of *primal-dual* methods [119], in which the primal variables (original unknowns) and the dual ones (Lagrange multipliers) are put on an equal footing. A primal-dual method is then none other than a "clever" way to solve the system of equations made up by the KKT optimality conditions of the minimization problem. The idea is therefore natural to draw inspiration from existing primal-dual methods in order to solve nonlinear algebraic systems containing complementarity equations that do not necessarily come from any minimization problem.

4.3.3.1 General principle

Let us consider the family of regularized problems

$$\Lambda(X) = 0, \tag{4.66a}$$
$$G(X) \odot H(X) = \nu \mathbf{1}, \tag{4.66b}$$

where $\nu \geqslant 0$ is the smoothing parameter, $\mathbf{1} \in \mathbb{R}^m$ is the vector whose components are all equal to 1, and $\odot$ denotes Hadamard's componentwise product. System (4.66) takes the abstract form

$$\widetilde{F}(X;\nu) = 0, \tag{4.67a}$$

with

$$\widetilde{F}(X;\nu) = \begin{bmatrix} \Lambda(X) \\ G(X) \odot H(X) - \nu \mathbf{1} \end{bmatrix} \in \mathbb{R}^\ell. \tag{4.67b}$$

Equation (4.66b), which can be explicitly written as

$$G_\alpha(X) H_\alpha(X) = \nu, \qquad \forall \alpha \in \{1, \ldots, m\},$$

means that we are using the same parameter ν for all the complementarity equations. This common practice corresponds to what is known as the *central path* in the theory of interior-point methods. In considering (4.66), we have somehow "forgotten" the positivity conditions

$$G(X) \geqslant 0, \qquad H(X) \geqslant 0.$$

In reality, these conditions will be specifically taken into account in the algorithm. We will go back to this later.

To unfold the mechanism of interior-point methods to our system, it is more convenient to reformulate system (4.66) as

$$\Lambda(X) = 0, \tag{4.68a}$$
$$G(X) - V = 0, \tag{4.68b}$$
$$H(X) - W = 0, \tag{4.68c}$$
$$V \odot W = \nu \mathbf{1}, \tag{4.68d}$$

where $(V, W) \in \mathbb{R}^m \times \mathbb{R}^m$ are called *slack variables*. These are of course subject to the componentwise positivity conditions

$$V \geqslant 0, \qquad W \geqslant 0, \tag{4.69}$$

which must be constantly "remembered" during the algorithm. System (4.68) can be given the abstract form

$$\mathbf{F}(\mathbf{X};\nu) = 0, \tag{4.70a}$$

where

$$\mathbf{X} = \begin{bmatrix} X \\ V \\ W \end{bmatrix} \in \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^m \subset \mathbb{R}^{\ell+2m}, \qquad \mathbf{F}(\mathbf{X}\,;\nu) = \begin{bmatrix} \Lambda(X) \\ G(X) - V \\ H(X) - W \\ V \odot W - \nu \mathbf{1} \end{bmatrix} \in \mathbb{R}^{\ell+2m}. \tag{4.70b}$$

Enlarging the size of the system and the number of unknowns does not change the determinant of the Jacobian matrix at the corresponding solution. Let us formally state this result, since it will be useful later. Due to definitions (4.67b) and (4.70b), the Jacobian matrices $\nabla_X \widetilde{F}(X;\nu)$ and $\nabla_{\mathbf{X}} \mathbf{F}(\mathbf{X};\nu)$ do not depend on ν. For short, they will be denoted by $\nabla \widetilde{F}(X)$ and $\nabla \mathbf{F}(\mathbf{X})$.

Lemma 4.4. *Let $X \in \mathcal{D}$ and $\mathbf{X} = (X, V, W) \in \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^m$ such that $V = G(X)$ and $W = H(X)$. Then,*

$$\det \nabla \mathbf{F}(\mathbf{X}) = \det \nabla \widetilde{F}(X). \tag{4.71}$$

In particular, the two Jacobian matrices are singular or nonsigular at the same time.

Chứng minh. The determinant of the Jacobian matrix of $\widetilde{F}(\cdot;\nu)$ is equal to

$$\det \nabla \widetilde{F}(X) = \begin{vmatrix} \nabla \Lambda(X) \\ \nabla G(X) \odot H(X) + \nabla H(X) \odot G(X) \end{vmatrix}, \tag{4.72}$$

where the Hadamard product $\nabla G(X) \odot H(X)$ between the $m \times \ell$-matrix $\nabla G(X)$ and the m-vector $H(X)$ is defined as the $m \times \ell$-matrix whose each column is the Hadamard product of a column of $\nabla G(X)$ and $H(X)$, and similarly for $\nabla H(X) \odot G(X)$. The determinant of the Jacobian matrix of $\mathbf{F}(\mathbf{X}\,;\nu)$ is equal to

$$\det \nabla \mathbf{F}(\mathbf{X}) = \begin{vmatrix} \nabla \Lambda(X) & 0 & 0 \\ \nabla G(X) & -I_m & 0 \\ \nabla H(X) & 0 & -I_m \\ 0 & I_m \odot W & I_m \odot V \end{vmatrix},$$

with the same definition for $I_m \odot V$ and $I_m \odot W$. By linear combination of the last (block)-row with the second and third (block)-rows, this can be shown to be equal to

$$\det \nabla \mathbf{F}(\mathbf{X}) = \begin{vmatrix} \nabla \Lambda(X) & 0 & 0 \\ \nabla G(X) & -I_m & 0 \\ \nabla H(X) & 0 & -I_m \\ \nabla G(X) \odot W + \nabla H(X) \odot V & 0 & 0 \end{vmatrix}.$$

By means of $2m$ row permutations, we end up with

$$\det \nabla \mathbf{F}(\mathbf{X}) = \begin{vmatrix} \nabla \Lambda(X) & 0 & 0 \\ \nabla G(X) \odot W + \nabla H(X) \odot V & 0 & 0 \\ \nabla G(X) & -I_m & 0 \\ \nabla H(X) & 0 & -I_m \end{vmatrix} = \begin{vmatrix} \nabla \Lambda(X) \\ \nabla G(X) \odot W + \nabla H(X) \odot V \end{vmatrix}.$$

Invoking $V = G(X)$, $W = H(X)$ and comparing with (4.72), we have the desired conclusion. Note that the Lemma does not require X to be a solution of (4.67). □

A primal-dual interior-point method strives to generate a sequence $\{\mathbf{X}^k = (X^k, V^k, W^k)\}_{k\in\mathbb{N}} \subset \mathbb{R}^{\ell+2m}$ and an auxiliary sequence $\{\nu^k\}_{k\in\mathbb{N}} \subset \mathbb{R}_+^*$ such that

$$(X^k,\, V^k,\, W^k) \to (\bar{X},\, G(\bar{X}),\, H(\bar{X})), \qquad \nu^k \to 0,$$

where $\bar{X} \in \mathcal{D}$ is a zero of $F = \widetilde{F}(\cdot\,;0)$. This sequence must satisfy the componentwise *strict* positivity condition

$$V^k > 0, \qquad W^k > 0,$$

for all $k \geqslant 0$. Another way to express this strict positivity condition is to define the *interior region*

$$\mathfrak{I} = \big\{\mathbf{X} = (X,\, V,\, W) \in \mathbb{R}^{\ell+2m} \;\mid\; V > 0,\;\; W > 0\big\}, \tag{4.73a}$$

and to require

$$\mathbf{X}^k = (X^k,\, V^k,\, W^k) \in \mathfrak{I} \tag{4.73b}$$

for all $k \geqslant 0$. The interest of enforcing these strict bounds is to avoid spurious solutions, which satisfy $\mathbf{F}(X, V, W) = 0$ but not $V \geqslant 0$ and $W \geqslant 0$. Some interior-point methods require the iterates to be *strictly feasible*, that is, $(V^k, W^k) = (G(X^k), H(X^k))$ for all $k \geqslant 0$. We shall not request feasibility.

To go from current iterate $\mathbf{X}^k$ to the next one $\mathbf{X}^{k+1}$, primal-dual interior-point methods modify the Newton algorithm in some judicious way to compute a search direction

$$\mathbf{d}^k = (\mathrm{d}X^k,\, \mathrm{d}V^k,\, \mathrm{d}W^k)$$

as the solution of a linear system of the form

$$\nabla\mathbf{F}(\mathbf{X}^k)\,\mathbf{d}^k + \{\text{something homogeneous to } \mathbf{F}\} = 0.$$

Usually, the full step along this direction is not acceptable, since the corresponding update would violate (4.73). To circumvent this difficulty, a *truncation* is performed so that

$$\mathbf{X}^k + \varsigma^k\mathbf{d}^k \in \mathfrak{I}$$

for some $\varsigma^k \in (0, 1]$, as close to 1 as possible. This operation can also be viewed as a *damped* Newton iteration, as in §4.3.1.3 where the purpose was not positivity but global convergence.

We are now going to scrutinize two embodiments of this general principle: a simplistic version called *single-stage* method and a highly sophisticated version known as *Mehrotra's predictor-corrector* method. From now on, we use the notation

$$\langle V,\, W\rangle = \sum_{\alpha=1}^{m} V_\alpha W_\alpha \tag{4.74}$$

to designate the dot product in $\mathbb{R}^m$.

4.3.3.2 Single-stage interior-point method

We first consider a very simple interior-point method, perhaps the simplest that could possibly be imagined. It is described in Algorithm 4.5 and consists primarily of one Newton iteration (Step 3) followed by a truncation (Step 4) and an update for the regularization parameter (Step 6). Let us comment the steps in more details.

Algorithm 4.5 Single-stage interior-point algorithm

1. Choose $\mathbf{X}^0 \in \mathfrak{I}$. Set $k = 0$, $\nu^0 = \langle V^0, W^0 \rangle / m$, $\gamma = 0.99$.
2. If $\mathbf{F}(\mathbf{X}^k ; 0) = 0$, stop.
3. Find a direction $\mathbf{d}^k = (\mathrm{d}X^k, \mathrm{d}V^k, \mathrm{d}W^k) \in \mathbb{R}^{\ell+2m}$ such that
$$\mathbf{F}(\mathbf{X}^k ; \nu^k) + \nabla \mathbf{F}(\mathbf{X}^k)\, \mathbf{d}^k = 0. \tag{4.75}$$
4. Compute $\varsigma^k \in (0,1)$ such that $\mathbf{X}^k + \varsigma^k \mathbf{d}^k \in \mathfrak{I}$ by
$$\varsigma^k = \gamma \cdot \arg\max \big\{ \varsigma \in [0,1] \mid V^k + \varsigma\, \mathrm{d}V^k \geqslant 0,\ W^k + \varsigma\, \mathrm{d}W^k \geqslant 0 \big\}. \tag{4.76}$$
5. Set $\mathbf{X}^{k+1} = \mathbf{X}^k + \varsigma^k \mathbf{d}^k$.
6. Set $\nu^{k+1} =$ one of the heuristic strategies (4.77).
7. Set $k \leftarrow k+1$. Go to step 2.

The name of the method comes from the fact that there is only one linear system to be solved at each iteration. This single Newton iteration (4.75), in Step 3, is aimed at finding an approximate solution to $\mathbf{F}(\mathbf{X} ; \nu^k) = 0$, starting from $\mathbf{X}^k$. The full Newton step $\mathbf{d}^k$ is truncated in Step 4, where we look for the largest possible reduction factor ς^k in the interval $(0,1)$. The initial value ν^0 of the regularization parameter, set to $\langle V^0, W^0 \rangle / m$, has the flavor of *centrality*. But this will be soon forgotten in the course of the iterations, where ν^k is no longer required to be $\langle V^k, W^k \rangle / m$. Instead, the parameter ν^k "lives its own life" according to an *a priori* procedure. Below are a few common heuristic ways to progressively drive ν^k to 0:

- A geometric sequence
$$\nu^{k+1} = 0.5\, \nu^k, \tag{4.77a}$$
the advantage of which is to go slowly to zero, which is useful when ν^k is still large;
- A power sequence
$$\nu^{k+1} = (\nu^k)^2, \tag{4.77b}$$
the advantage of which is to go quickly to zero, which is useful when ν^k is already small;
- A hybrid geometric-power sequence
$$\nu^{k+1} = \min\big(0.5\, \nu^k,\ (\nu^k)^2\big), \tag{4.77c}$$
which combines the advantages of the first two strategies.
- A hybrid geometric-power sequence compared to a duality measure
$$\nu^{k+1} = \min\big(0.5\, \nu^k,\ (\nu^k)^2,\ \langle V^{k+1}, W^{k+1} \rangle / m\big), \tag{4.77d}$$
the interest of which is to reconnect the sequence to some current "reality" [56, §5].

Unfortunately, there is no universal magic formula to monitor the sequence of regularization parameter $\{\nu^k\}$. A heuristic strategy that works fine with one problem may fail miserably with another. We have to try several sequences $\{\nu^k\}$ before knowing which one is best suited to the problem at hand. This seems to be the Achilles heel of such smoothing methods, for which we will propose a novel approach in chapter §5.

4.3.3.3 Mehrotra's predictor–corrector method

Most of today's interior-point general-purpose softwares for linear programming are based on Mehrotra's predictor-corrector algorithm [88]. In Algorithm 4.6, we have extended it to our equation-solving problem. Unlike the previous single-stage method, each iteration of Mehrotra's method consists of two stages, namely, the *prediction* stage (from Step 3 to Step 7) and the *correction* stage (from Step 8 to Step 11). In each stage, there is a linear system to be solved. Nevertheless, the two linear systems (4.78) and (4.80) share the same matrix, which allows for some cost-savings by means of factorization.

Mehrotra's original algorithm for linear programming combines several key ingredients that had been separately suggested and implemented by other authors before [119, §10]. But the subtle order in which these ingredients are assembled and organized, as well as some ingenious heuristics of his own for the adaptive centering parameter and the step lengths, are the real assets that have contributed to its outstanding success. These ingredients are still present in the extended version. Let us comment on Algorithm 4.6 step-by-step.

Prediction stage. The algorithm starts by computing an *affine-scaling* direction $\mathbf{d}^k_{\text{aff}}$. By "affine-scaling," it is meant that we leave the current parameter ν^k aside and set our sights on the ultimate goal $\nu = 0$ right away. This accounts for equation (4.78) in Step 3, which is the first-order linearization of $\mathbf{F}(\mathbf{X}^k + \mathbf{d}^k_{\text{aff}}\,; 0) = 0$ around $\mathbf{X}^k$. We recall that here the Jacobian matrix $\nabla\mathbf{F}(\mathbf{X}^k)$ does not depend on ν and thus can be viewed as being located at $\nu = 0$. After truncation in Step 4, we need to assess the payback of this audacious attempt. This is done in Step 5 and Step 6, where we compute the centrality measure ν^k_{aff} of the state $\mathbf{X}^k_{\text{aff}}$. The prediction stage culminates in Step 7, where we compute Mehrotra's ratio

$$\sigma^k = \left(\frac{\nu^k_{\text{aff}}}{\nu^k}\right)^3. \tag{4.82}$$

This heuristic ratio, called *adaptive centering factor* and found by trial and error on a wide range of problems, has proved its remarkable effectiveness. It is a special feature of Mehrotra's algorithm. A value $\sigma^k \ll 1$ means that our ambitious affine-scaling venture has been rewarded with success. The predicted state is significantly closer to the boundary than at the beginning of the iteration. A value $\sigma^k \gg 1$ implies that we took too many risks and the odds have been against us. In the former case, we must follow the predictor's advice by reducing ν significantly. In the latter case, we must return inside the interior region in hope for a better update at the next iteration.

Correction stage. Thanks to the centering factor (4.82), the second stage can deal with both cases in a "unified" fashion, by simply targeting $\sigma^k\nu^k$ as the new parameter value at which an approximate zero of $\mathbf{F}$ must be searched for. Indeed, equation (4.80) in Step 8 can be seen as a local model for $\mathbf{F}(\mathbf{X}^k + \mathbf{d}^k_{\text{cor}}\,; \sigma^k\nu^k)$ around $\mathbf{X}^k$. In spite of appearances, this local model is not quite linear. In fact, it is quadratic but in a special way. For one, the second-order terms are

Algorithm 4.6 Mehrotra predictor–corrector algorithm

1. Choose $\mathbf{X}^0 \in \mathfrak{I}$. Set $k = 0$, $\nu^0 = \langle V^0, W^0 \rangle / m$, $\gamma = 0.99$.
2. If $\mathbf{F}(\mathbf{X}^k ; 0) = 0$, stop.
3. Find a direction $\mathbf{d}^k_{\text{aff}} = (\mathrm{d}X^k_{\text{aff}}, \mathrm{d}V^k_{\text{aff}}, \mathrm{d}W^k_{\text{aff}}) \in \mathbb{R}^{\ell+2m}$ such that
$$\mathbf{F}(\mathbf{X}^k ; 0) + \nabla \mathbf{F}(\mathbf{X}^k)\, \mathbf{d}^k_{\text{aff}} = 0. \tag{4.78}$$
4. Compute $\varsigma^k_{\text{aff}} \in (0, 1)$ such that $\mathbf{X}^k + \varsigma^k_{\text{aff}}\, \mathbf{d}^k_{\text{aff}} \in \mathfrak{I}$ by
$$\varsigma^k_{\text{aff}} = \gamma \cdot \arg\max \big\{ \varsigma \in [0, 1] \mid V^k + \varsigma\, \mathrm{d}V^k_{\text{aff}} \geqslant 0,\ W^k + \varsigma\, \mathrm{d}W^k_{\text{aff}} \geqslant 0 \big\}. \tag{4.79}$$
5. Set $\mathbf{X}^k_{\text{aff}} = \mathbf{X}^k + \varsigma^k_{\text{aff}}\, \mathbf{d}^k_{\text{aff}}$.
6. Set $\nu^k_{\text{aff}} = \langle V^k_{\text{aff}}, W^k_{\text{aff}} \rangle / m$.
7. Set $\sigma^k = (\nu^k_{\text{aff}} / \nu^k)^3$.
8. Find a direction $\mathbf{d}^k_{\text{cor}} \in \mathbb{R}^{\ell+2m}$ such that
$$\mathbf{F}(\mathbf{X}^k ; \sigma^k \nu^k) + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \mathrm{d}V^k_{\text{aff}} \odot \mathrm{d}W^k_{\text{aff}} \end{bmatrix} + \nabla \mathbf{F}(\mathbf{X}^k)\, \mathbf{d}^k_{\text{cor}} = 0. \tag{4.80}$$
9. Compute $\varsigma^k_{\text{cor}} \in (0, 1)$ such that $\mathbf{X}^k + \varsigma^k_{\text{cor}}\, \mathbf{d}^k_{\text{cor}} \in \mathfrak{I}$ by
$$\varsigma^k_{\text{cor}} = \gamma \cdot \arg\max \big\{ \varsigma \in [0, 1] \mid V^k + \varsigma\, \mathrm{d}V^k_{\text{cor}} \geqslant 0,\ W^k + \varsigma\, \mathrm{d}W^k_{\text{cor}} \geqslant 0 \big\}. \tag{4.81}$$
10. Set $\mathbf{X}^{k+1} = \mathbf{X}^k + \varsigma^k_{\text{cor}}\, \mathbf{d}^k_{\text{cor}}$.
11. Set $\nu^{k+1} = \langle V^{k+1}, W^{k+1} \rangle / m$.
12. Set $k \leftarrow k + 1$. Go to step 2.

present only for the last block of equations containing $V \odot W$. For another, the corresponding increments $\mathrm{d}V^k$ and $\mathrm{d}W^k$ have been "freezed" at the predicted affine-scaling values, instead of being considered as unknowns, which would have given rise to an intricate quadratic equation with respect to the direction. Anyhow, Mehrotra's algorithm demonstrates an effort to take into account *curvature* information in order to speed up convergence. To our knowledge, there is no theoretical results on the exact rate of convergence and on the polynomial complexity of Mehrotra's algorithm for linear programming, although such results are available for some variants [123, 124], the analysis of which is easier.

In Mehrotra's algorithm, the regularization parameter ν^k is always equal the duality measure $\langle V^k, W^k \rangle / m$. Paradoxically, it never appears as the target of the linearized Newton iterations: the predictor sets out to achieve $\nu = 0$, while the corrector aims to reach $\nu = \sigma^k \nu^k$. Finally, it is worth noticing that

$$\mathbf{F}(\mathbf{X}^k \,; \sigma^k \nu^k) + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \mathrm{d}V^k_{\mathrm{aff}} \odot \mathrm{d}W^k_{\mathrm{aff}} \end{bmatrix} = \mathbf{F}(\mathbf{X}^k \,; 0) + \begin{bmatrix} 0 \\ 0 \\ 0 \\ -\sigma^k \nu^k \mathbf{1} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \mathrm{d}V^k_{\mathrm{aff}} \odot \mathrm{d}W^k_{\mathrm{aff}} \end{bmatrix},$$

so that the direction $\mathbf{d}^k_{\mathrm{cor}}$ can be regarded as the aggregation of three increments, namely,

$$\mathbf{d}^k_{\mathrm{cor}} = \mathbf{d}^k_{\mathrm{aff}} + \mathbf{d}^k_{\mathrm{cen}} + \mathbf{d}^k_{\mathrm{qua}}, \tag{4.83}$$

where $\mathbf{d}^k_{\mathrm{aff}}$ is the affine-scaling direction (4.78), $\mathbf{d}^k_{\mathrm{aff}}$ is the centering direction defined by

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ -\sigma^k \nu^k \mathbf{1} \end{bmatrix} + \nabla \mathbf{F}(\mathbf{X}^k)\, \mathbf{d}^k_{\mathrm{cen}} = 0,$$

and $\mathbf{d}^k_{\mathrm{qua}}$ is the quadratic correction defined by

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ \mathrm{d}V^k_{\mathrm{aff}} \odot \mathrm{d}W^k_{\mathrm{aff}} \end{bmatrix} + \nabla \mathbf{F}(\mathbf{X}^k)\, \mathbf{d}^k_{\mathrm{qua}} = 0.$$

Therefore, if no damping occurred, that is, if $\varsigma^k_{\mathrm{aff}} = \varsigma^k_{\mathrm{cor}} = 1$, then the final state

$$\mathbf{X}^{k+1} = \mathbf{X}^k_{\mathrm{aff}} + (\mathbf{d}^k_{\mathrm{cen}} + \mathbf{d}^k_{\mathrm{qua}}),$$

could be thought of as a correction brought to the predicted state $\mathbf{X}^k_{\mathrm{aff}}$. This justifies the name of the second stage. For linear programming, Mehrotra's algorithm can also be insightfully reinterpreted as a perturbed composite damped Newton method [113].

4.4 What may go wrong?

The numerical methods described so far all "look good" on the paper. To gain insight into the actual difficulties at the practical level, let us run them on the toy model

$$u + \tau q - u_\flat = 0, \tag{4.84a}$$

$$\min(1-q,\, u^2-q)=0 \tag{4.84b}$$

in the unknown $X=(u,q)\in\mathbb{R}^2$. Here, $(u_\flat,\tau)\in\mathbb{R}_+^*\times\mathbb{R}_+^*$ are the parameters of the problem. System (4.84) comes from the Euler implicit discretization of the ordinary differential equation

$$\frac{\mathrm{d}u}{\mathrm{d}t}=-q,\qquad q=\min(1,u^2),$$

using $\tau>0$ as the time-step and $u_\flat$ as the current state. This system is an extremely reduced model for stratigraphy[4]. More on the above continuous model will be said in §6.1.1.

System (4.84) is the model of interest to us. It is made up of a linear equation (4.84a), i.e., $\ell-m=1$, and a nonlinear complementarity equation (4.84b), i.e., $m=1$. We will use (4.84) as a benchmark test for various numerical methods, insofar as we know its solution.

Theorem 4.11. *For all $(u_\flat,\,\tau)\in\mathbb{R}_+^*\times\mathbb{R}_+^*$, system* (4.84) *has a unique solution $(\overline{u},\,\overline{q})\in\mathbb{R}_+^*\times\mathbb{R}_+^*$, called* reference solution *and given by*

$$(\overline{u},\,\overline{q})=\begin{cases}(u_\flat-\tau,\,1) & \text{if } \tau\leqslant u_\flat-1\\ \left(\dfrac{2u_\flat}{1+\sqrt{1+4\tau u_\flat}},\,\dfrac{4u_\flat^2}{(1+\sqrt{1+4\tau u_\flat})^2}\right) & \text{otherwise.}\end{cases} \tag{4.85}$$

If $\tau<u_\flat+1$, then $(\overline{u},\,\overline{g})$ is also the unique solution of (4.84) *over $\mathbb{R}^2$. If $\tau\geqslant u_\flat+1$, then system* (4.84) *has two other solutions with negative values for u.*

Chứng minh. It is more convenient to carry out the analysis for the scalar equation $\varphi(u)=0$, where the graph of the continuous function

$$\varphi(u)=u+\tau\min(u^2,\,1)-u_\flat$$

consists of three parts. Over $(-\infty,-1]$ and $[1,+\infty)$, it coincides with two half-lines belonging to the straight line $\varphi=u+\tau-u_\flat$. Over $[-1,1]$, it coincides with an arc of the convex parabola $\varphi=u+\tau u^2-u_\flat$. This arc always lie below the segment $\varphi=u+\tau-u_\flat$. Moreover,

$$\varphi(-1)=-1+\tau-u_\flat,\qquad \varphi(0)=-u_\flat,\qquad \varphi(1)=1+\tau-u_\flat.$$

It then appears that:

1. If $\varphi(1)\leqslant 0$, i.e., $\tau\leqslant u_\flat-1$, there is a unique solution over $\mathbb{R}$ which is given by the intersection of the right half-line and the axis of abscissae. This solution is in the saturated regime ($\overline{q}=1$), given by $\overline{u}+\tau-u_\flat=0$.

2. If $\varphi(1)>0$, since $\varphi(0)<0$ and because $\varphi'(u)=1+2\tau u>0$ over $[0,1]$, there is only one $\overline{u}\in(0,1)$ for which $\varphi(\overline{u})=0$. By solving the quadratic equation $\tau u^2+u-u_\flat=0$ and by choosing the positive root, we end up with the unsaturated solution $\overline{u}=2u_\flat/(1+\sqrt{1+4\tau u_\flat})$, $\overline{q}=\overline{u}^2$. But there are two subcases to be discussed.

   (a) If $\varphi(-1)<0$, i.e., $\tau<u_\flat+1$, there is obviously no other root over $\mathbb{R}$.

   (b) If $\varphi(-1)\geqslant 0$, i.e., $\tau\geqslant u_\flat+1$, then there are two spurious roots, one saturated solution in $(-\infty,-1]$ and one unsaturated solution in $[-1,0)$.

This completes the proof. □

[4] a branch of geology concerned with the study of sedimentary rock layers (strata) and layering (stratification)

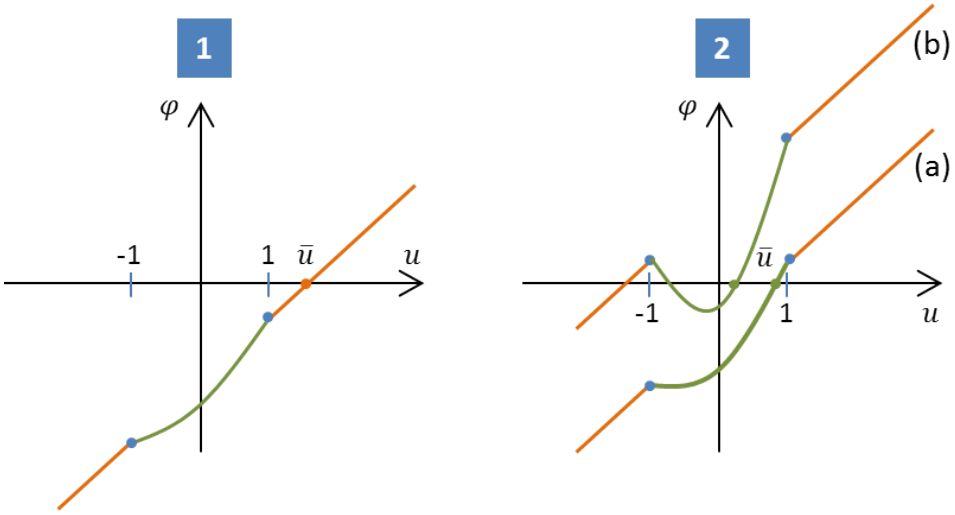


Figure 4.3: Analysis of solutions for the stratigraphic system (4.84).

4.4.1 Issues with nonsmooth methods

The existence of non-physical solutions when the time-step is large enough ($\tau \geqslant u_\flat + 1$) may cause an iterative solver to converge toward a wrong solution. This indeed occurs to the Newton-min method. Let

$$X = \begin{bmatrix} u \\ q \end{bmatrix} \in \mathbb{R}^2, \qquad F = \begin{bmatrix} u + \tau q - u_\flat \\ \min(1 - q,\, u^2 - q) \end{bmatrix},$$

so that system (4.84) reads $F(X) = 0$. Let us apply the Newton-min method described in the previous chapter, using the initial point

$$X^0 = \begin{bmatrix} u_\flat \\ \min(1,\, u_\flat^2) \end{bmatrix} \tag{4.86}$$

This initial point is the most "natural" one, insofar as $u_\flat$ represents the value of u at the previous discrete time. Thanks to the extreme simplicity of the model, it is possible to predict the behavior of this Newton-min algorithm. The following statement should be read in conjunction with Figure 4.4.

Theorem 4.12. *Let $(u_\flat, \tau) \in \mathbb{R}_+^* \times \mathbb{R}_+^*$. The Newton-min method applied to system (4.84) using the starting point (4.86)*

- *converges to the reference solution $\bar{X} = (\overline{u}, \overline{g})$ if and only if $u_\flat \leqslant 1$ or $u_\flat \geqslant \tau + 1 - 1/\tau$;*
- *exhibits a cyclic behavior, namely, oscillates between two iterates, if and only if $u_\flat > \max(1; \tau - 1)$ and $u_\flat < \tau + 1 - 1/\tau$;*
- *converges to a wrong solution if and only if $u_\flat > 1$ and $u_\flat \leqslant \tau - 1$.*

Chứng minh. The basic idea is to do the Newton iterations by hands. See Hamani [58] for more details. □

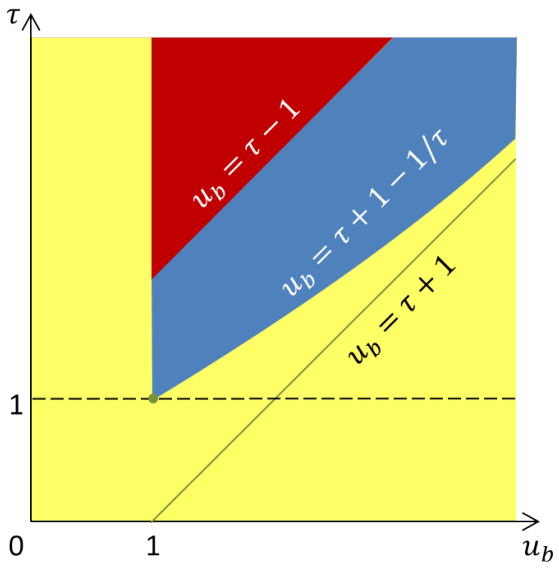


Figure 4.4: Behavior of the Newton-min algorithm for (4.84) with starting point (4.86). Yellow: convergence toward the correct solution; blue: periodic oscillation between two iterates; red: convergence toward a wrong solution.

4.4.2 Issues with smoothing methods

When working with smoothing methods, the issue may occur when we update the values of variables for the next iteration. In the single-stage interior-point Algorithm 4.5, this is in Step 4 where we compute ς^k. In the Mehrotra predictor–corrector Algorithm 4.6, this is in Step 9 where we compute ς^k_{cor}. Similarly, with θ-smoothing techniques, we also need all variables of θ-functions to remain nonnegative during iterations. We will also use stratigraphic model as an illustration.

At the k-th iteration, we obtain Δu^k, Δq^k and we need to find $\varsigma \in [0,1]$ such that $u^{k+1} = u^k + \varsigma \Delta u^k$ and $q^{k+1} = q^k + \varsigma \Delta q^k$ satisfy

$$\begin{aligned} 1 - q^{k+1} &\geqslant 0, \\ (u^{k+1})^2 - q^{k+1} &\geqslant 0. \end{aligned}$$

Rewriting these conditions in terms of the current iterates at k, we have

$$\begin{aligned} 1 - q^k - \varsigma \Delta q^k &\geqslant 0, \\ (\Delta u^k)^2 \varsigma^2 + (2u^k \Delta u^k - \Delta q^k)\varsigma + (u^k)^2 - q^k &\geqslant 0. \end{aligned}$$

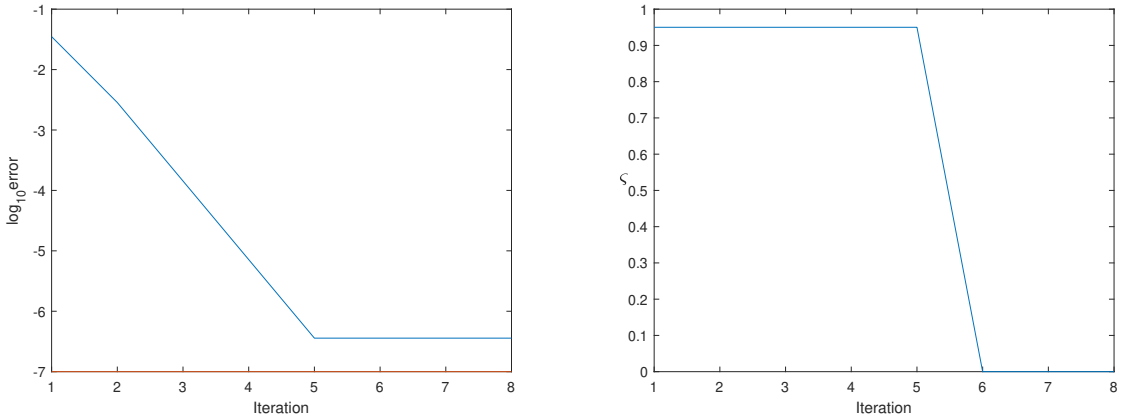
We analyze the second inequality as a quadratic inequation with respect to ς. In the neighbordhood of a solution (but the iterations have not finished yet), $(u^k)^2 - q^k$ may be very small or even zero. Hence, the quadratic inequality becomes

$$(\Delta u^k)^2 \varsigma^2 + (2u^k \Delta u^k - \Delta q^k)\varsigma \geqslant 0.$$

Unfortunately, in some cases,

$$\frac{-(2u^k \Delta u^k - \Delta q^k)}{(\Delta u^k)^2} > 1.$$

Therefore, the unique solution we obtain is $\varsigma = 0$. This means that values of u and q do not change from this iteration, even though we are close to a solution. Another difficulty coming from the single-stage interior-point algorithm is that it is not easy to find a good strategy to define values of ν^k during iterations.



(a) Errors during iterations

(b) ς during iterations

Figure 4.5: Stratigraphy model, Predictor-Corrector Mehrotra, $u_b = 0.9, \tau = 0.1, u_0 = 0.54, g_0 = 0.07$.

The smoothing methods require all the terms of complementarity equations positive during all iterations. However, it may make the algorithm stay at a point without progress. From these disadvantages, we believe a new approach could be one that does not require positivity on the arguments of complementarity equations during the iterations, but still ensures positivity at the end, when the algorithm converges.

Chapter 5

A new nonparametric interior-point method

Contents

Les difficultés numériques signalées à la fin du chapitre précédent, ainsi que celles qui seront exposées en détails au chapitre suivant, indiquent que la généralisation à nos problèmes des méthodes existantes pour l'optimisation et pour les problèmes de complémentarité purs ne débouche pas nécessairement sur une méthode de résolution adaptée. Nous avons toutefois pu constater une relative supériorité des méthodes de points intérieurs du point de vue de la robustesse et souhaitons poursuivre dans cette voie.

L'absence d'une stratégie systématique pour piloter le paramètre de régularisation étant la principale faiblesse des méthodes de points intérieurs, nous avons entrepris de chercher une manière plus automatique de faire tendre ce paramètre vers zéro, laquelle préserverait les avantages des méthodes par points intérieurs sans en subir les inconvénients. La section §5.1 est consacrée à notre nouvelle méthode, appelée nonparametric interior-point method *(NPIPM). L'idée clé est de traiter le paramètre de régularisation comme une inconnue à part entière en introduisant une nouvelle équation dans le système. On est ainsi ramené à l'application de la méthode de Newton lisse à un problème lisse, ce qui permet de dérouler une analyse de convergence locale et globale reposant sur la régularité du zéro en question.*

La régularité, c'est-à-dire la non-singularité de la matrice jacobienne évaluée en un point solution, devient ainsi un critère essentiel pour le bon fonctionnement de la nouvelle méthode. Nous la vérifions en §5.2 sur le modèle diphasique compositionnel introduit dans la première partie. Par un enchaînement de calculs non-triviaux, notre montrons que sous l'hypothèse de stricte convexité des fonctions d'énergie molaire de Gibbs, la solution est régulière dès qu'elle n'est ni transitionnelle ni azéotropique.

The numerical issues mentioned at the end of the previous chapter, as well as those that will be illustrated in the next chapter, show that the existing methods are not well suited to our problems. For the models considered in our numerical tests, however, the interior-point methods turned out to be far more robust than the others, at least when the sequence of regularization parameters is properly adjusted. We also said that the lack of a systematic strategy to steer this sequence toward zero is the main weakness of interior-point methods. Therefore, we undertook to look for a more automatic way to decrease this parameter, which preserves the advantages of interior-point methods without suffering from their drawbacks.

Section §5.1 is devoted to our new method, called the *nonparametric interior-point method* (NPIPM). The key idea is to treat the regularization parameter as a full-fledged unknown by introducing a new equation into the system. We are thus brought back to applying the smooth Newton method to a smooth problem, which allows for local and global convergence analysis based on the regularity of the zero at hand. In section §5.2, we verify the regularity condition on the zeros of the two-phase multicomponent model introduced in Part I.

5.1 Design principle and properties of NPIPM

We recall that the interior-point methods considered in §4.3.3 have replaced the original nonsmooth problem (4.1) by a sequence of regularized problems

$$\mathsf{F}(\mathsf{X}\,;\nu) = 0, \tag{5.1a}$$

where

$$\mathsf{X} = \begin{bmatrix} X \\ V \\ W \end{bmatrix} \in \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^m \subset \mathbb{R}^{\ell+2m}, \qquad \mathsf{F}(\mathsf{X}\,;\nu) = \begin{bmatrix} \Lambda(X) \\ G(X) - V \\ H(X) - W \\ V \odot W - \nu\mathbf{1} \end{bmatrix} \in \mathbb{R}^{\ell+2m}, \tag{5.1b}$$

where $\nu \geqslant 0$ is the smoothing parameter, $\mathbf{1} \in \mathbb{R}^m$ is the vector whose components are all equal to 1 and $(V, W) \in \mathbb{R}^m \times \mathbb{R}^m$ are the *slack variables*, subject to

$$V \geqslant 0, \qquad W \geqslant 0. \tag{5.1c}$$

5.1.1 When the parameter becomes a variable

In system (5.1), the status of the parameter ν is very distinct from that of the variable X. While X is computed "automatically" by a Newton iteration, ν has to be updated "manually" in an *ad hoc* manner. On two occasions, we witnessed that progress occurs when two objects of ostensibly different natures are put on an equal footing and given a unified treatment: the present and absent phases in the phase equilibrium problem (chapter §2), the primal and dual variables in interior-point methods (chapter §4). From this experience, we feel that it would be judicious to incorporate the parameter ν into the variables X.

5.1.1.1 Enlarged equivalent systems

Let us therefore consider the enlarged vector of unknowns

$$\mathbb{X} = \begin{bmatrix} \mathsf{X} \\ \nu \end{bmatrix} \in \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}_+ \subset \mathbb{R}^{\ell+2m+1}, \tag{5.2}$$

and let us try to find a system of $\ell + 2m + 1$ equations

$$\mathbb{F}(\mathbb{X}) = 0 \tag{5.3}$$

to be prescribed on $\mathbb{X}$. To this end, let us remind ourselves that our ultimate goal is to solve $\mathbf{F}(\mathbf{X}, 0) = 0$, together with the inequalities (5.1c). Thus, it is really natural to first consider

$$\mathbb{F}(\mathbb{X}) = \begin{bmatrix} \mathbf{F}(\mathbf{X}; \nu) \\ \nu \end{bmatrix}. \tag{5.4}$$

This construction turns out to be too naive. Indeed, if we start from some $\nu^0 > 0$ and solve the smooth system (5.3)–(5.4) by the smooth Newton method, since the last equation is linear, we end up with $\nu^1 = 0$ at the first iteration. Once the boundary of the interior region is reached, we are "stuck" there.

To prevent ν from rushing to zero in just one iteration, we could set

$$\mathbb{F}(\mathbb{X}) = \begin{bmatrix} \mathbf{F}(\mathbf{X}; \nu) \\ \nu^2 \end{bmatrix}, \tag{5.5}$$

which is equivalent at the continuous level. At the level of Newton iterates, there is still a deficiency: since $\nu = 0$ is now a double root of the last equation, quadratic convergence will be lost when ν^k approaches 0! A remedy to this is to add a small linear term, that is,

$$\mathbb{F}(\mathbb{X}) = \begin{bmatrix} \mathbf{F}(\mathbf{X}; \nu) \\ \eta\nu + \nu^2 \end{bmatrix}, \tag{5.6}$$

where $\eta > 0$ is a small parameter. The price to be paid for recovering quadratic convergence is that there is now a spurious negative solution $\nu = -\eta < 0$. This should not be a problem, however, if we start from a positive value for ν.

At this stage, system (5.6) is not yet fully adequate. Indeed, the last equation is totally decoupled from the others. Everything happens as if ν follows a prefixed sequence, generated by the Newton iterates of the scalar equation $\eta\nu + \nu^2 = 0$, regardless of $\mathbf{X}$. It is desirable to couple ν and $\mathbf{X}$ in a tighter way. In this respect, we advocate

$$\mathbb{F}(\mathbb{X}) = \begin{bmatrix} \mathbf{F}(\mathbf{X}; \nu) \\ \frac{1}{2}\|V^-\|^2 + \frac{1}{2}\|W^-\|^2 + \eta\nu + \nu^2 \end{bmatrix}, \tag{5.7a}$$

where

$$\|V^-\|^2 = \sum_{\alpha=1}^{m} (\min(V_\alpha, 0))^2, \qquad \|W^-\|^2 = \sum_{\alpha=1}^{m} (\min(W_\alpha, 0))^2. \tag{5.7b}$$

This choice has the benefit of taking into account the nonnegativity condition (5.1c). Indeed, the last equation of (5.7a) implies that, as long as $\nu \geqslant 0$, we are ascertained that $V^- = W^- = 0$. This amounts to saying that $V \geqslant 0$ and $W \geqslant 0$. Should a component of V or W become negative during the iteration, this equation would contribute to "penalize" it.

5.1.1.2 Globalized algorithm

From now on, the enlarged equations (5.7) are selected as the reference system in the design of our new algorithm. The idea is simply to apply the standard Newton method to the smooth system (5.3), (5.7). To enforce a globally convergent behavior, we also recommend using Armijo's

line search, as in Algorithm 4.4. Before writing down the new algorithm, let us investigate the new Jacobian matrix.

We saw in §4.3.3.1 that $\nabla_{\mathbf{X}}\mathbf{F}(\mathbf{X};\nu)$, the Jacobian matrix of $\mathbf{F}$ with respect to $\mathbf{X}$, does not depend on ν and can be denoted by $\nabla\mathbf{F}(\mathbf{X})$. It is useful to decompose it in (block)-columns as

$$\nabla\mathbf{F}(\mathbf{X}) = \begin{bmatrix}\nabla_X\mathbf{F}(\mathbf{X}) & \nabla_V\mathbf{F}(\mathbf{X}) & \nabla_W\mathbf{F}(\mathbf{X})\end{bmatrix}.$$

Since ν is now considered as a variable, it makes sense to define the partial derivatives $\partial_\nu\mathbf{F}(\mathbf{X})$. From (5.1), we deduce that

$$\partial_\nu\mathbf{F}(\mathbf{X}) = \begin{bmatrix}0\\0\\0\\-\mathbf{1}\end{bmatrix} \in \mathbb{R}^{\ell+2m}$$

does not depend on $\mathbf{X}$ and therefore can be safely written as $\partial_\nu\mathbf{F}$. On the other hand, the scalar function $x \mapsto \frac{1}{2}|\min(x,0)|^2$ is differentiable and its derivative is equal to $\min(x,0)$. From this observation, it follows that

$$\nabla\mathbb{F}(\mathbb{X}) = \begin{bmatrix}\nabla_X\mathbf{F}(\mathbf{X}) & \nabla_V\mathbf{F}(\mathbf{X}) & \nabla_W\mathbf{F}(\mathbf{X}) & \partial_\nu\mathbf{F}\\ 0 & (V^-)^T & (W^-)^T & \eta+2\nu\end{bmatrix} \tag{5.8}$$

where V^- is the vector of components $V_\alpha^- = \min(V_\alpha, 0)$ and similarly for W^-. Below is a result about this Jacobian matrix, which is in the same vein as Lemma 4.4 and which will be useful for later purposes.

Lemma 5.1. *Let $\mathbf{X} \in \overline{\mathfrak{I}}$, where $\mathfrak{I}$ is the interior region defined in* (4.73)*. Let $\nu \in \mathbb{R}$ and $\mathbb{X} = [\mathbf{X}^T;\nu]^T$. Then,*

$$\det\nabla\mathbb{F}(\mathbb{X}) = (\eta+2\nu)\det\nabla\mathbf{F}(\mathbf{X}). \tag{5.9}$$

If $\nu > -\eta/2$, the two Jacobian matrices are singular or nonsigular at the same time.

Chứng minh. Thanks to the assumption $\mathbf{X} \in \overline{\mathfrak{I}}$, we have $V \geqslant 0$ and $W \geqslant 0$, so that $V^- = W^- = 0$. Expanding the determinant of (5.8) with respect to the last row yields the desired result. Note that the Lemma does not require $(\mathbf{X};\nu)$ to solve (5.1a)–(5.1b) or $\mathbb{X}$ to solve (5.3), (5.7). □

Introduce the least-squares potential

$$\Theta(\mathbb{X}) = \frac{1}{2}\|\mathbb{F}(\mathbb{X})\|^2.$$

A detailed description of NPIPM is given in Algorithm 5.1. A few comments are in order:

- The initial point $\mathbb{X}^0 = (\mathbf{X}^0, \nu^0)$ must be an interior point, namely, $\mathbf{X}^0 \in \mathfrak{I}$. Furthermore, it is often taken at equilibrium, that is, $V^0 = G(X^0)$ and $W^0 = H(X^0)$, so that the initial parameter $\nu^0 = \langle V^0, W^0\rangle/m$ has the correct order of magnitude.
- If $\mathbf{X}^k \in \mathfrak{I}$, then $(V^k)^- = (W^k)^- = 0$ and

$$\mathrm{d}^k = \begin{bmatrix}\mathrm{d}\mathbf{X}^k\\ \mathrm{d}\nu^k\end{bmatrix} = -\begin{bmatrix}\nabla\mathbf{F}(\mathbf{X}^k) & -\partial_\nu\mathbf{F}\\ 0 & \eta+2\nu^k\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{F}(\mathbf{X}^k;\nu^k)\\ \eta\nu^k+(\nu^k)^2\end{bmatrix}$$

  provided that the Jacobian matrix is invertible. The increment for the parameter is then

$$d\nu^k = -\frac{\eta\nu^k+(\nu^k)^2}{\eta+2\nu^k}.$$

Algorithm 5.1 Nonparametric interior point algorithm with Armijo line search

1. Choose $\mathbb{X}^0 = (\mathbf{X}^0, \nu^0)$, $\mathbf{X}^0 \in \mathfrak{I}$, $\nu^0 = \langle V^0, W^0 \rangle / m$, $\kappa \in (0, 1/2)$, $\varrho \in (0, 1)$. Set $k = 0$.
2. If $\mathbb{F}(\mathbb{X}^k) = 0$, stop.
3. Find a direction $\mathrm{d}^k \in \mathbb{R}^{\ell+2m+1}$ such that
$$\mathbb{F}(\mathbb{X}^k) + \nabla \mathbb{F}(\mathbb{X}^k) \mathrm{d}^k = 0. \tag{5.10}$$
4. Choose $\varsigma^k = \varrho^{j_k} \in (0, 1)$, where $j_k \in \mathbb{N}$ is the smallest integer such that
$$\Theta(\mathbb{X}^k + \varrho^{j_k} \mathrm{d}^k) \leqslant (1 - 2\kappa\varrho^{j_k})\, \Theta(\mathbb{X}^k). \tag{5.11}$$
5. Set $\mathbb{X}^{k+1} = \mathbb{X}^k + \varsigma^k \mathrm{d}^k$ and $k \leftarrow k + 1$. Go to step 2.

- There is no need to truncate the Newton direction d^k to preserve positivity for V^{k+1} and W^{k+1}, since nonnegativity is "guaranteed" at convergence. However, if we wish all the iterates to belong to the interior region $\mathfrak{I}$, then we are free to carry out an additional damping after Step 4 (Armijo's line search).

A final remark concerns the qualification of the method as *nonparametric*. It can be rightly objected that the method still involves a small positive parameter η. Nevertheless, this parameter is chosen once and for all and does not need to be driven to zero. It is in this sense that the term *nonparametric* is to be understood.

5.1.2 Global convergence analysis

The main interest of Algorithm 5.1 lies in the prospect of global convergence, as envisioned by the theory that we are developing now. This global convergence theory, due to Bonnans [21, §6], is primarily based on the regularity of zeros [Definition 4.10], an assumption that we will be able to check for each model under consideration. We reproduce most of Bonnans' theory here, in view of its importance to our algorithm.

We first need to define the continuity modulus of a function and some useful lemmas. We denote $c_{\mathbb{F}}$ the continuity modulus of $\mathbb{F}$ at $\bar{\mathbb{X}}$ defined by
$$c_{\mathbb{F}}(\gamma) := \sup_{\|\mathbb{X} - \bar{\mathbb{X}}\| \leqslant \gamma} \|\mathbb{F}(\mathbb{X}) - \mathbb{F}(\bar{\mathbb{X}})\|, \tag{5.12}$$
and $c_{\nabla\mathbb{F}}, c_{\nabla\mathbb{F}^{-1}}$ are defined similarly. The first Lemma establishes that near a regular zero $\bar{\mathbb{X}}$, the quantities $\|\mathbb{F}(\mathbb{X})\|$ and $\|\mathbb{X} - \bar{\mathbb{X}}\|$ are of the same order.

Lemma 5.2 (Lemma 6.5, [21]). *Let $\bar{\mathbb{X}}$ be a regular zero of $\mathbb{F}$ and assume that $c_{\mathbb{F}}(\gamma)$ is well-defined. There exist $\gamma_1 > 0, c_1 > 0$ and $c_2 > 0$ such that*
$$c_{\mathbb{F}}(\gamma) \leqslant c_1 \gamma, \tag{5.13a}$$
$$\|\mathbb{X} - \bar{\mathbb{X}}\| \leqslant c_2 \|\mathbb{F}(\mathbb{X})\|, \tag{5.13b}$$
for all $\mathbb{X}$ satisfying $\|\mathbb{X} - \bar{\mathbb{X}}\| \leqslant \gamma \leqslant \gamma_1$.

Chứng minh. We consider all $\gamma_1 > 0$ such that $c_{\nabla\mathbb{F}}(\gamma_1) < \|\nabla\mathbb{F}(\bar{\mathbb{X}})^{-1}\|^{-1}$. Let $\mathbb{X}$ such that $\|\mathbb{X} - \bar{\mathbb{X}}\| \leqslant \gamma \leqslant \gamma_1$. Since $\mathbb{F}(\bar{\mathbb{X}}) = 0$, then

$$\mathbb{F}(\mathbb{X}) = \mathbb{F}(\mathbb{X}) - \mathbb{F}(\bar{\mathbb{X}}) = \int_0^1 \nabla\mathbb{F}(\bar{\mathbb{X}} + t(\mathbb{X} - \bar{\mathbb{X}}))(\mathbb{X} - \bar{\mathbb{X}})\,\mathrm{d}t,$$

and

$$\|\mathbb{F}(\mathbb{X})\| \leqslant \sup_{t\in[0,1]} \|\nabla\mathbb{F}(\bar{\mathbb{X}} + t(\mathbb{X} - \bar{\mathbb{X}}))\|\|\mathbb{X} - \bar{\mathbb{X}}\|$$

Let $t \in [0,1]$ and $\mathbb{X}_t = \bar{\mathbb{X}} + t(\mathbb{X} - \bar{\mathbb{X}})$, then

$$\|\mathbb{X}_t - \bar{\mathbb{X}}\| = \|t(\mathbb{X} - \bar{\mathbb{X}})\| = t\|\mathbb{X} - \bar{\mathbb{X}}\| \leqslant \gamma_1.$$

This means that

$$\|\nabla\mathbb{F}(\mathbb{X}_t)\| - \|\nabla\mathbb{F}(\bar{\mathbb{X}})\| \leqslant \|\nabla\mathbb{F}(\mathbb{X}_t) - \nabla\mathbb{F}(\bar{\mathbb{X}})\| \leqslant c_{\nabla\mathbb{F}}(\gamma_1),$$

or

$$\|\nabla\mathbb{F}(\mathbb{X}_t)\| \leqslant \|\nabla\mathbb{F}(\bar{\mathbb{X}})\| + c_{\nabla\mathbb{F}}(\gamma_1) \text{ for all } t \in [0,1].$$

So, we have

$$\sup_{t\in[0,1]} \|\nabla\mathbb{F}(\bar{\mathbb{X}} + t(\mathbb{X} - \bar{\mathbb{X}}))\| \leqslant \|\nabla\mathbb{F}(\bar{\mathbb{X}})\| + c_{\nabla\mathbb{F}}(\gamma_1)$$

and then

$$\mathbb{F}(\mathbb{X}) \leqslant \left(\|\nabla\mathbb{F}(\bar{\mathbb{X}})\| + c_{\nabla\mathbb{F}}(\gamma_1)\right)\gamma. \tag{5.14}$$

Thus, (5.13a) holds with $c_1 := \|\nabla\mathbb{F}(\bar{\mathbb{X}})\| + c_{\nabla\mathbb{F}}(\gamma_1)$. To prove (5.13b), we start from

$$\mathbb{F}(\mathbb{X}) = \nabla\mathbb{F}(\bar{\mathbb{X}})(\mathbb{X} - \bar{\mathbb{X}}) + \int_0^1 \left[\nabla\mathbb{F}(\bar{\mathbb{X}} + t(\mathbb{X} - \bar{\mathbb{X}})) - \nabla\mathbb{F}(\bar{\mathbb{X}})\right](\mathbb{X} - \bar{\mathbb{X}})\,\mathrm{d}t. \tag{5.15}$$

Hence,

$$\|\mathbb{F}(\mathbb{X})\| \geqslant \|\nabla\mathbb{F}(\bar{\mathbb{X}})(\mathbb{X} - \bar{\mathbb{X}})\| - c_{\nabla\mathbb{F}}(\gamma_1)\|\mathbb{X} - \bar{\mathbb{X}}\|. \tag{5.16}$$

We also have

$$\|\mathbb{X} - \bar{\mathbb{X}}\| = \|\nabla\mathbb{F}(\bar{\mathbb{X}})^{-1}\nabla\mathbb{F}(\bar{\mathbb{X}})(\mathbb{X} - \bar{\mathbb{X}})\| \leqslant \|\nabla\mathbb{F}(\bar{\mathbb{X}})^{-1}\|\|\nabla\mathbb{F}(\bar{\mathbb{X}})(\mathbb{X} - \bar{\mathbb{X}})\|. \tag{5.17}$$

Combining (5.16) and (5.17), we get

$$\|\mathbb{F}(\mathbb{X})\| \geqslant [\|\nabla\mathbb{F}(\bar{\mathbb{X}})^{-1}\|^{-1} - c_{\nabla\mathbb{F}}(\gamma_1)]\|\mathbb{X} - \bar{\mathbb{X}}\|,$$

so that (5.13b) holds true with

$$c_2 := [\|\nabla\mathbb{F}(\bar{\mathbb{X}})^{-1}\|^{-1} - c_{\nabla\mathbb{F}}(\gamma_1)]^{-1} > 0,$$

which completes the proof. □

The second Lemma gives an estimation of the distance to a regular zero after a Newton step.

Lemma 5.3 (Lemma 6.6, [21]). *Let γ_1 be constant of by Lemma 5.2. There exist $\gamma_2 \in (0, \gamma_1), c_3 > 0, c_4 > 0$ and $c_5 > 0$ such that*

$$\|d(\mathbb{X})\| \leqslant c_3\gamma, \tag{5.18a}$$
$$\|\mathbb{X} + d(\mathbb{X}) - \bar{\mathbb{X}}\| \leqslant c_4 c_{\nabla\mathbb{F}}(c_5\gamma)\gamma, \tag{5.18b}$$

for all $\mathbb{X}$ satisfying $\|\mathbb{X} - \bar{\mathbb{X}}\| = \gamma \leqslant \gamma_2$.

Chứng minh. Let $\gamma_2 \in (0, \gamma_1)$ such that $c_{\nabla\mathbb{F}^{-1}}(\gamma_2) \leqslant 1$. Owing to (5.13a),

$$\|d(\mathbb{X})\| \leqslant \|\nabla\mathbb{F}(\mathbb{X})^{-1}\|\|\mathbb{F}(\mathbb{X})\| \leqslant [\|\nabla\mathbb{F}(\bar{\mathbb{X}})^{-1}\| + c_{\nabla\mathbb{F}^{-1}}(\gamma)]c_1\gamma. \tag{5.19}$$

In other words, we obtain (5.18a) with

$$c_3 := c_1[\|\nabla\mathbb{F}(\bar{\mathbb{X}})^{-1}\| + c_{\nabla\mathbb{F}^{-1}}(\gamma_2)].$$

Defining $c_5 := c_3 + 1$, we have the upper-bound

$$\|\mathbb{X} + \mathrm{d}(\mathbb{X}) - \bar{\mathbb{X}}\| \leqslant \|\mathbb{X} - \bar{\mathbb{X}}\| + \|\mathrm{d}(\mathbb{X})\| \leqslant c_5\|\mathbb{X} - \bar{\mathbb{X}}\|. \tag{5.20}$$

As $\mathrm{d}(\mathbb{X})$ is the Newton direction of $\mathbb{F}$ at $\mathbb{X}$, we can write

$$\begin{aligned}\mathbb{F}(\mathbb{X} + \mathrm{d}(\mathbb{X})) &= \mathbb{F}(\mathbb{X}) + \int_0^1 \nabla\mathbb{F}(\mathbb{X} + t\mathrm{d}(\mathbb{X}))\mathrm{d}(\mathbb{X})\,\mathrm{d}t \\ &= \int_0^1 [\nabla\mathbb{F}(\mathbb{X} + t\mathrm{d}(\mathbb{X})) - \nabla\mathbb{F}(\mathbb{X})]\,\mathrm{d}(\mathbb{X})\,\mathrm{d}t,\end{aligned}$$

from which we infer that

$$\|\mathbb{F}(\mathbb{X} + \mathrm{d}(\mathbb{X}))\| \leqslant \sup_{t\in[0,1]} \|\nabla\mathbb{F}(\mathbb{X} + t\mathrm{d}(\mathbb{X})) - \nabla\mathbb{F}(\mathbb{X})\|\|\mathrm{d}(\mathbb{X})\|. \tag{5.21}$$

Since $c_5 > 1$ and $c_{\nabla\mathbb{F}}$ is a nondecreasing function of its argument,

$$\begin{aligned}\|\nabla\mathbb{F}(\mathbb{X} + t\mathrm{d}(\mathbb{X})) - \nabla\mathbb{F}(\mathbb{X})\| &\leqslant \|\nabla\mathbb{F}(\mathbb{X} + t\mathrm{d}(\mathbb{X})) - \nabla\mathbb{F}(\bar{\mathbb{X}})\| + \|\nabla\mathbb{F}(\bar{\mathbb{X}}) - \nabla\mathbb{F}(\mathbb{X})\| \\ &\leqslant c_{\nabla\mathbb{F}}(c_5\gamma) + c_{\nabla\mathbb{F}}(\gamma) \leqslant 2c_{\nabla\mathbb{F}}(c_5\gamma).\end{aligned}$$

Combining with (5.21) and (5.18a), we obtain

$$\|\mathbb{F}(\mathbb{X} + \mathrm{d}(\mathbb{X}))\| \leqslant 2c_3 c_{\nabla\mathbb{F}}(c_5\gamma)\gamma.$$

Using (5.13b), we obtain (5.18b) with $c_4 := 2c_2c_3$. □

Theorem 5.1 (Theorem 6.4, [21]). *Let $\bar{\mathbb{X}}$ be a regular zero of $\mathbb{F}$.*

(i) If $\mathbb{X}^0$ is close enough to $\bar{\mathbb{X}}$, the sequence $\{\mathbb{X}^k\}$ in (4.28a) is well-defined and converges superlinearly to $\bar{\mathbb{X}}$.

(ii) Moreover, if

$$\|\nabla\mathbb{F}(\mathbb{X}) - \nabla\mathbb{F}(\bar{\mathbb{X}})\| = O(\|\mathbb{X} - \bar{\mathbb{X}}\|), \tag{5.22}$$

then the convergence is quadratic.

Chứng minh. Let $\alpha \in (0,1)$. By Lemma 5.3, if $\|\mathbb{X}^k - \bar{\mathbb{X}}\| = \gamma \leqslant \gamma_2$, then

$$\|\mathbb{X}^{k+1} - \bar{\mathbb{X}}\| \leqslant c_4 c_{\nabla \mathbb{F}}(c_5\gamma)\|\mathbb{X}^k - \bar{\mathbb{X}}\|. \tag{5.23}$$

When $\gamma_\alpha \in (0, \gamma_1)$ is small enough, $c_4 c_{\nabla \mathbb{F}}(c_5\gamma) \leqslant \alpha$. For a certain $k_\alpha \in \mathbb{N}$ such that $\|\mathbb{X}^{k_\alpha} - \bar{\mathbb{X}}\| \leqslant \gamma_\alpha$, we have $\|\mathbb{X}^k - \bar{\mathbb{X}}\| \leqslant \alpha^{k-k_\alpha}\|\mathbb{X}^{k_\alpha} - \bar{\mathbb{X}}\|$ for all $k > k_\alpha$. If $\mathbb{X}^0$ is close enough to $\bar{\mathbb{X}}$, then the sequence $\{\mathbb{X}^k\}$ is well-defined and converges linearly to $\bar{\mathbb{X}}$. Furthermore, for all $\alpha \in (0,1)$, there exists an integer $k_\alpha \leqslant \gamma_\alpha$ and the convergence is linear with rate α. This implies that the convergence is superlinear. If (5.22) is satisfied, then $c_{\nabla \mathbb{F}}(\gamma) = O(\gamma)$. Combining with (5.23), we deduce that the convergence is quadratic. □

Theorem 5.2 (Theorem 6.9, [21]). *Let $\mathbb{F} : \mathbb{R}^{\ell+2m+1} \to \mathbb{R}^{\ell+2m+1}$ be a continuously-differentiable function.*

(i) [Local analysis] *Let $\bar{\mathbb{X}}$ be a regular zero of $\mathbb{F}$. If $\mathbb{X}^0$ is close enough to $\bar{\mathbb{X}}$, then $\varsigma_k = 1$ for all k, and $\mathbb{X}^k \to \bar{\mathbb{X}}$ superlinearly (and we recover the standard Newton method).*

(ii) [Limit point] *Let $\tilde{\mathbb{X}}$ be a limit point of sequence $\{\mathbb{X}^k\}$. If $\nabla\mathbb{F}(\tilde{\mathbb{X}})$ is invertible, then $\tilde{\mathbb{X}}$ is a regular zero of $\mathbb{F}$. If $\tilde{\mathbb{X}}$ is a regular zero of F, then $\varsigma_k = 1$ for k big enough and $\mathbb{X}^k \to \tilde{\mathbb{X}}$ superlinearly.*

(iii) [General behavior] *At least one of three possibilities below holds:*

  (a) $\mathbb{F}(\mathbb{X}^k) \to 0$.

  (b) $\|\mathbb{d}(\mathbb{X}^k)\|$ *is unbounded.*

  (c) *The sequence $\{\mathbb{X}^k\}$ converges to $\tilde{\mathbb{X}}$ where $\nabla\mathbb{F}(\tilde{\mathbb{X}})$ is not invertible.*

The three items of the Theorem illustrate the conditions and the qualities of convergence of the algorithm. Item (i) corresponds to the behavior of the algorithm near a regular zero. Item (ii) states the rate of convergence in some particular situations. Item (iii) summarizes all of the possible scenarios when running the algorithm. In particular, if $\nabla\mathbb{F}(\mathbb{X})$ is invertible everywhere (or at least during the iterations of the algorithm) and $\|\mathbb{F}(\mathbb{X})\| \to \infty$ as $\|\mathbb{X}\| \to \infty$, then only the possibility (a) of (iii) can occur; conditions of (ii) are satisfied so that if the algorithm converges, it will converge superlinearly to a regular zero.

Chứng minh. (i) If $\mathbb{X}^0$ is close enough to $\bar{\mathbb{X}}$, then sequence $\{\mathbb{X}^k\}$ is generated by the standard Newton method, i.e., $\mathbb{X}^{k+1} = \mathbb{X}^k + \mathbb{d}(\mathbb{X}^k)$. By (5.13b) [Lemma 5.2], we have $\|\mathbb{X}^k - \bar{\mathbb{X}}\| \leqslant c_2\|\mathbb{F}(\mathbb{X}^k)\|$. By the proof of Lemma 5.3, we have

$$\|\mathbb{F}(\mathbb{X}^k + d(\mathbb{X}^k))\| \leqslant 2c_3 c_{\nabla \mathbb{F}}(c_5\gamma)\|\mathbb{X}^k - \bar{\mathbb{X}}\|.$$

Hence,

$$\|\mathbb{F}(\mathbb{X}^{k+1})\| \leqslant 2c_2 c_3 c_{\nabla \mathbb{F}}(c_5\gamma)\|\mathbb{F}(\mathbb{X}^k)\|.$$

Since $c_{\nabla \mathbb{F}}(c_5\gamma) \to 0$ as $\gamma \to 0$, we can choose γ such that

$$\|\mathbb{F}(\mathbb{X}^{k+1})\| \leqslant (1 - 2\kappa)\|\mathbb{F}(\mathbb{X}^k)\|.$$

It satisfies (4.46) with $j = 0$ or $\varsigma^k = 1$ for all k. By Theorem 5.1, $\mathbb{X}^k \to \bar{\mathbb{X}}$ superlinear.

(ii) Let $\widetilde{\mathbb{X}}$ be a limit point of $\{\mathbb{X}^k\}$ and $\nabla\mathbb{F}(\widetilde{\mathbb{X}})$ is invertible. Suppose $\widetilde{\mathbb{X}}$ is not a zero of $\mathbb{F}$. We have

$$\begin{aligned}\Theta(\mathbb{X}+\varsigma d(\mathbb{X})) &= \Theta(\mathbb{X}) + \varsigma\int_0^1 \nabla\Theta(\mathbb{X}+t\varsigma d(\mathbb{X}))d(\mathbb{X})\,\mathrm{d}t\\ &= \Theta(\mathbb{X}) + \varsigma\nabla\Theta(\mathbb{X})d(\mathbb{X}) + \varsigma A(\mathbb{X},\varsigma)d(\mathbb{X}),\end{aligned}$$

where

$$A(\mathbb{X},\varsigma) := \int_0^1 \left[\nabla\Theta(\mathbb{X}+t\varsigma d(\mathbb{X})) - \nabla\Theta(\mathbb{X})\right]\,\mathrm{d}t.$$

When $\mathbb{X}\to\widetilde{X}$ and $\varsigma\to 0$, we get $A(\mathbb{X},\varsigma)\to 0$ and $d(\mathbb{X})$ is bounded. There exist $\gamma>0$ and $\widetilde{\varsigma}>0$ such that, if $\|\mathbb{X}-\widetilde{\mathbb{X}}\|\leqslant\gamma$ and $\varsigma\leqslant\widetilde{\varsigma}$, we have $|A(\mathbb{X},\varsigma)d(\mathbb{X})|\leqslant\Theta(\mathbb{X})$, and then

$$\begin{aligned}\Theta(\mathbb{X}+\varsigma d(\mathbb{X})) &\leqslant \Theta(\mathbb{X}) + \varsigma\nabla\Theta(\mathbb{X})^T d(\mathbb{X}) + \varsigma\Theta(\mathbb{X}) = (1-\varsigma)\Theta(\mathbb{X})\\ &\leqslant (1-2\varsigma\kappa)\Theta(\mathbb{X}).\end{aligned}$$

This proves that (4.46) is satisfied when $\|\mathbb{X}^k-\widetilde{\mathbb{X}}\|\leqslant\gamma$ and $\varrho^{j_k}\leqslant\varrho\widetilde{\varsigma}$, or $j_k\leqslant J$. If $\widetilde{\mathbb{X}}$ is a limit point of $\{\mathbb{X}^k\}$, using $\Theta(\widetilde{\mathbb{X}})\geqslant\frac{1}{2}\Theta(\mathbb{X}^k)$, for k large enough, we obtain

$$\Theta(\mathbb{X}^{k+1})-\Theta(\mathbb{X}^k)\leqslant -2\kappa\varrho^J\Theta(\mathbb{X}^k)\leqslant -\kappa\varrho^J\Theta(\widetilde{\mathbb{X}})$$

for the corresponding subsequence. Since $\Theta(\mathbb{X}^k)$ decreases, this implies that $\Theta(\mathbb{X}^k)$ tends to $-\infty$, which is impossible. Therefore, if a limit point of $\mathbb{X}^k$ is not a zero of $\mathbb{F}$, then $\nabla\mathbb{F}$ is not invertible at this point. Since this leads to a contradiction, $\widetilde{\mathbb{X}}$ must be a zero of $\mathbb{F}$. If $\widetilde{\mathbb{X}}$ is a regular zero of $\mathbb{F}$, then $\mathbb{X}^k$ is close enough to $\widetilde{\mathbb{X}}$ for k big enough. Apply point (i), we conclude that $\varsigma^k=1$ for k big enough and $\mathbb{X}^k\to\widetilde{\mathbb{X}}$ superlinear.

(iii) We consider the point (c). We assume that $\lim_{k\to+\infty}\|\mathbb{F}(\mathbb{X}^k)\|>0$ and $\|d(\mathbb{X}^k)\|$ is bounded. Then, with the inequality (4.46)

$$-\Theta(\mathbb{X}^0)\leqslant\lim_k\Theta(\mathbb{X}^k)-\Theta(\mathbb{X}^0)=\sum_{k\geqslant 0}(\Theta(\mathbb{X}^{k+1})-\Theta(\mathbb{X}^k))\leqslant -2\kappa\sum_{k\geqslant 0}\varsigma^k\Theta(\mathbb{X}^k).$$

Put $l:=\lim_k\Theta(\mathbb{X}^k)>0$, it implies $\sum_{k\geqslant 0}\varsigma^k\leqslant\dfrac{\Theta(\mathbb{X}^0)}{2\kappa l}$. If $d(\mathbb{X}^k)$ is bounded, then

$$\sum_{k\geqslant 0}\|\mathbb{X}^{k+1}-\mathbb{X}^k\|=\sum_{k\geqslant 0}\varsigma^k\|d(\mathbb{X}^k)\|<\infty.$$

Hence, $\{\mathbb{X}^k\}$ is a Cauchy sequence. It converges to a certain $\widetilde{\mathbb{X}}$. But $\widetilde{\mathbb{X}}$ is not a regular zero of $\mathbb{F}$. We conclude that $\nabla\mathbb{F}(\widetilde{\mathbb{X}})$ is not invertible by point (ii).

This completes the proof. □

5.2 Regularity of zeros for the two-phase multicomponent model

According to Theorem 5.2, the promise of global convergence for the NPIPM algorithm hinges on the regularity of the zeros of the system at hand. Put another way, if we could prove that the Jacobian matrix $\nabla\mathbb{F}(\bar{\mathbb{X}})$ at a solution $\bar{\mathbb{X}}$ is nonsingular, this would be an auspicious sign of the adequacy of the NPIPM algorithm to the problem. In this section, we derive the necessary and sufficient conditions for a zero of the two-phase multicomponent model (2.77) to be regular.

5.2.1 A general proof for strictly convex laws

We begin with a proof of the regularity of all nondegenerate zeros of (2.77), i.e., solutions that are neither a transition point [Definition 2.1] nor an azeotropic point [Definition 2.2], in the most general case. By "general," we mean that the Gibbs functions g_G and g_L satisfy Hypotheses 2.2.

5.2.1.1 Preliminary lemmas

We first need a few technicalities to transform the determinant to be computed into a simpler one.

Lemma 5.4. *Let $\bar{\mathbb{X}} = (\bar{\mathbf{X}}, \bar{\nu}) \in \mathbb{R}^{\ell+2m} \times \mathbb{R}$ be a solution of (5.3), (5.7), with $\bar{\mathbf{X}} = (\bar{X}, \bar{V}, \overline{W}) \in \overline{\mathfrak{J}} \subset \mathbb{R}^\ell \times \mathbb{R}^m \times \mathbb{R}^m$, where $\mathfrak{J}$ is the interior region (4.73). Then,*

$$\det \nabla \mathbb{F}(\bar{\mathbb{X}}) = (\eta + 2\bar{\nu}) \begin{vmatrix} \nabla \Lambda(\bar{X}) \\ \nabla G(\bar{X}) \odot H(\bar{X}) + \nabla H(\bar{X}) \odot G(\bar{X}) \end{vmatrix}. \tag{5.24}$$

In particular, if $\bar{\nu} > -\eta/2$, the two determinants above are singular or nonsigular at the same time.

Chứng minh. By virtue of Lemma 5.1 and from $\bar{\mathbf{X}} \in \overline{\mathfrak{J}}$, we have $\det \nabla \mathbb{F}(\bar{\mathbb{X}}) = (\eta + 2\bar{\nu}) \det \nabla \mathbf{F}(\bar{\mathbf{X}})$. Since $\bar{\mathbb{X}}$ is a solution, $\bar{V} = G(\bar{X})$ and $\overline{W} = H(\bar{X})$. We are thus in a position to apply Lemma 4.4 and to obtain $\det \nabla \mathbb{F}(\bar{\mathbb{X}}) = (\eta + 2\bar{\nu}) \det \nabla \widetilde{F}(\bar{X})$. The latter determinant is given by (4.72), which leads to the desired result. □

The matrix in the left-hand side of (4.72) is of order $\ell + 2m + 1$, while that in the right-hand side of (4.72) is of order ℓ. In addition to this reduction in size, the following transformation will be helpful. Assume that for some $i \in \{1, \dots, \ell - m\}$, the i-th component of Λ takes the form

$$\Lambda_i(X) = \varphi_{i,G}(X) - \varphi_{i,L}(X),$$

where $\varphi_{i,G}$ is associated with the G-phase and $\varphi_{i,L}$ is associated with the L-phase. Typically, this can be an equality of extended fugacites (2.77b) for some species. Let us consider Λ^f the vector-valued function in which Λ_i has been replaced by

$$\Lambda_i^f(X) = f(\varphi_{i,G}(X)) - f(\varphi_{i,L}(X)),$$

where f is an increasing and differentiable scalar function. Typically, f is the logarithm function, by which an extended fugacity is mapped to an extended chemical potential. It is obvious that since $\Lambda_i(X) = 0$ is equivalent to $\Lambda_i^f(X) = 0$, $\Lambda(X) = 0$ is equivalent to $\Lambda^f(X) = 0$. But what can be said about the determinant of the Jacobian matrix at a solution when we write $\Lambda^f(X) = 0$ instead of $\Lambda(X) = 0$?

Lemma 5.5. *Let $\bar{\mathbb{X}} = (\bar{X}, \bar{V}, \overline{W}, \bar{\nu}) \in \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}$ be a solution of (5.3), (5.7). Then,*

$$\begin{vmatrix} \nabla \Lambda^f(\bar{X}) \\ \nabla G(\bar{X}) \odot H(\bar{X}) + \nabla H(\bar{X}) \odot G(\bar{X}) \end{vmatrix} = f'(\overline{\varphi}_i) \begin{vmatrix} \nabla \Lambda(\bar{X}) \\ \nabla G(\bar{X}) \odot H(\bar{X}) + \nabla H(\bar{X}) \odot G(\bar{X}) \end{vmatrix}, \tag{5.25}$$

where $\overline{\varphi}_i = \varphi_{i,G}(\bar{X}) = \varphi_{i,L}(\bar{X})$ is the common value of $\varphi_{i,G}$ and $\varphi_{i,L}$ at the solution. In particular, for an increasing function f, the two determinants above are singular or nonsingular at the same time.

Chứng minh. The gradient (row) of Λ_i^f with respect to X reads

$$\nabla\Lambda_i^f(X) = f'(\varphi_{i,G}(X))\nabla\varphi_{i,G}(X) - f'(\varphi_{i,L}(X))\nabla\varphi_{i,L}(X).$$

At a solution, we have $\varphi_{i,G}(\bar{X}) = \varphi_{i,L}(\bar{X}) =: \overline{\varphi}_i$. Hence, $f'(\varphi_{i,G}(X)) = f'(\varphi_{i,L}(X)) = f'(\overline{\varphi}_i)$ can be factorized, so that

$$\nabla\Lambda_i^f(\bar{X}) = f'(\overline{\varphi}_i)\,\nabla\Lambda_i(\bar{X}).$$

The proof is completed by taking this factor out of the i-th row of the Jacobian matrix. □

Our last preparatory Lemma is concerned with with positive definite symmetric matrices, which will be needed at the end of the proof.

Lemma 5.6. *Let* A *and* B *be two positive definite symmetric matrices. Then,* $\mathsf{C} = \mathsf{A}^{-1/2}\mathsf{B}\mathsf{A}^{-1/2}$ *and* $\mathsf{D} = I - (I + \mathsf{C})^{-1}$ *are also positive definite symmetric matrices.*

Chứng minh. It easy to see that C is symmetric. Besides,

$$z^T\mathsf{C}z = z^T\mathsf{A}^{-1/2}\mathsf{B}\mathsf{A}^{-1/2}z = (\mathsf{A}^{-1/2}z)^T\mathsf{B}(\mathsf{A}^{-1/2}z) \geqslant 0,$$

where equality holds if and only if $\mathsf{A}^{-1/2}z = 0$, that is, if and only if $z = 0$. Thus, C is positive definite. Therefore, its eigenvalues are all positive. In the a basis that diagonalizes C, the matrix $\mathsf{D} = I - (I + \mathsf{C})^{-1}$ is also transformed into a diagonal form. If $\lambda > 0$ is one of the eigenvalues of C, the corresponding eigenvalue of D is

$$1 - (1 + \lambda)^{-1} = \frac{\lambda}{1 + \lambda} > 0.$$

Therefore, D is positive definite. □

5.2.1.2 Criterion for a regular zero

The two-phase multicomponent system (2.77) corresponds to

$$\ell = 2\mathrm{K} + 1, \qquad m = 2.$$

Let $X = (Y, \xi_G^{\mathrm{I}}, \ldots, \xi_G^{\mathrm{K}}, \xi_L^{\mathrm{I}}, \ldots, \xi_L^{\mathrm{K}}) \in \mathbb{R}^{2\mathrm{K}+1}$ be the vector of unknowns. The functions Λ, G and H associated with (2.77) are

$$\Lambda(X) = \begin{bmatrix} Y\xi_G^{\mathrm{I}} + (1-Y)\xi_L^{\mathrm{I}} - c^{\mathrm{I}} \\ \vdots \\ Y\xi_G^{\mathrm{K}-1} + (1-Y)\xi_G^{\mathrm{K}-1} - c^{\mathrm{K}-1} \\ \xi_G^{\mathrm{I}}\Phi_G^{\mathrm{I}}(\boldsymbol{x}_G) - \xi_L^{\mathrm{I}}\Phi_L^{\mathrm{I}}(\boldsymbol{x}_L) \\ \vdots \\ \xi_G^{\mathrm{K}}\Phi_G^{\mathrm{K}}(\boldsymbol{x}_G) - \xi_L^{\mathrm{K}}\Phi_L^{\mathrm{K}}(\boldsymbol{x}_L) \end{bmatrix} \in \mathbb{R}^{2\mathrm{K}-1} \tag{5.26a}$$

and

$$G(X) = \begin{bmatrix} Y \\ 1-Y \end{bmatrix} \in \mathbb{R}^2, \qquad H(X) = \begin{bmatrix} 1 - \xi_G^1 - \ldots - \xi_G^{\mathrm{K}} \\ 1 - \xi_L^1 - \ldots - \xi_L^{\mathrm{K}} \end{bmatrix} \in \mathbb{R}^2, \tag{5.26b}$$

where $\boldsymbol{x}_G = (x_G^{\mathrm{I}}, \ldots, x_G^{\mathrm{K}-1})$ and $\boldsymbol{x}_L = (x_L^{\mathrm{I}}, \ldots, x_L^{\mathrm{K}-1})$ are defined in (2.38b) as functions of X.

Theorem 5.3. *Let $\bar{\mathbb{X}} = (\bar{X}, \bar{V}, \overline{W}, \bar{\nu}) \in \mathbb{R}^{2\mathrm{K}+6}$ be a solution of (5.3), (5.7) using the functions (5.26). Assume that $\bar{\nu} = 0$ and that the Gibbs energy functions g_G and g_L meet Hypotheses 2.2.*

Then, $\bar{\mathbb{X}}$ is a regular zero if and only if $\bar{X}$ is neither a transition point (in the sense of Definition 2.1) nor an azeotropic point (in the sense of Definition 2.2).

Chứng minh. If $\bar{\nu} = 0$, then the last equation of (5.7) implies $\bar{V}^- = \overline{W}^- = 0$, that is $\bar{\mathbf{X}} \in \bar{\mathfrak{I}}$. By Lemma 5.4, we have $\det \nabla \mathbb{F}(\bar{\mathbb{X}}) = \eta \bar{\mathfrak{d}}$, where

$$\bar{\mathfrak{d}} = \begin{vmatrix} \nabla \Lambda(\bar{X}) \\ \nabla G(\bar{X}) \odot H(\bar{X}) + \nabla H(\bar{X}) \odot G(\bar{X}) \end{vmatrix}.$$

By Lemma 5.5, we know that $\bar{\mathfrak{d}}$ is zero or nonzero simultaneously with

$$\bar{\mathfrak{d}}^\bullet = \begin{vmatrix} \nabla \Lambda^\bullet(\bar{X}) \\ \nabla G(\bar{X}) \odot H(\bar{X}) + \nabla H(\bar{X}) \odot G(\bar{X}) \end{vmatrix},$$

where

$$\Lambda^\bullet(X) = \begin{bmatrix} Y\xi_G^{\mathrm{I}} + (1-Y)\xi_L^{\mathrm{I}} - c^{\mathrm{I}} \\ \vdots \\ Y\xi_G^{\mathrm{K}-1} + (1-Y)\xi_G^{\mathrm{K}-1} - c^{\mathrm{K}-1} \\ \ln(\xi_G^{\mathrm{I}}\Phi_G^{\mathrm{I}}(\boldsymbol{x}_G)) - \ln(\xi_L^{\mathrm{I}}\Phi_L^{\mathrm{I}}(\boldsymbol{x}_L)) \\ \vdots \\ \ln(\xi_G^{\mathrm{K}}\Phi_G^{\mathrm{K}}(\boldsymbol{x}_G)) - \ln(\xi_L^{\mathrm{K}}\Phi_L^{\mathrm{K}}(\boldsymbol{x}_L)) \end{bmatrix}.$$

Henceforth, we shall be studying $\bar{\mathfrak{d}}^\bullet$. Each of the last K components of $\Lambda^\bullet(X)$ can be rewritten as

$$\ln(\sigma_G) + \mu_G^i - \ln(\sigma_L) - \mu_L^i,$$

for $i \in \{\mathrm{I}, \mathrm{II}, \ldots, \mathrm{K}\}$, with

$$\begin{aligned} \sigma_G &= \xi_G^{\mathrm{I}} + \ldots + \xi_G^{\mathrm{K}}, & \sigma_L &= \xi_L^{\mathrm{I}} + \ldots + \xi_L^{\mathrm{K}}, \\ \mu_G^i &= \ln(x_G^i\, \Phi_G^i(\boldsymbol{x}_G)), & \mu_L^i &= \ln(x_L^i\, \Phi_L^i(\boldsymbol{x}_L)). \end{aligned}$$

After this transformation, $\bar{\mathfrak{d}}^\bullet$ has the structure

$$\bar{\mathfrak{d}}^\bullet = \begin{vmatrix} \Delta\bar{\xi}^{\mathrm{I}} & \bar{Y} & \ldots & 0 & 0 & 1-\bar{Y} & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \Delta\bar{\xi}^{\mathrm{K}-1} & 0 & \ldots & \bar{Y} & 0 & 0 & \ldots & 1-\bar{Y} & 0 \\ 0 & \overline{M}_{G,\mathrm{I}}^{\mathrm{I}} & \ldots & \overline{M}_{G,\mathrm{K}-1}^{\mathrm{I}} & \overline{M}_{G,\mathrm{K}}^{\mathrm{I}} & -\overline{M}_{L,\mathrm{I}}^{\mathrm{I}} & \ldots & -\overline{M}_{L,\mathrm{K}-1}^{\mathrm{I}} & -\overline{M}_{L,\mathrm{K}}^{\mathrm{I}} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & \overline{M}_{G,\mathrm{I}}^{\mathrm{K}-1} & \ldots & \overline{M}_{G,\mathrm{K}-1}^{\mathrm{K}-1} & \overline{M}_{G,\mathrm{K}}^{\mathrm{K}-1} & -\overline{M}_{L,\mathrm{I}}^{\mathrm{K}-1} & \ldots & -\overline{M}_{L,\mathrm{K}-1}^{\mathrm{K}-1} & -\overline{M}_{L,\mathrm{K}}^{\mathrm{K}-1} \\ 0 & \overline{M}_{G,\mathrm{I}}^{\mathrm{K}} & \ldots & \overline{M}_{G,\mathrm{K}-1}^{\mathrm{K}} & \overline{M}_{G,\mathrm{K}}^{\mathrm{K}} & -\overline{M}_{L,\mathrm{I}}^{\mathrm{K}} & \ldots & -\overline{M}_{L,\mathrm{K}-1}^{\mathrm{K}} & -\overline{M}_{L,\mathrm{K}}^{\mathrm{K}} \\ 1-\bar{\sigma}_G & -\bar{Y} & \ldots & -\bar{Y} & -\bar{Y} & 0 & \ldots & 0 & 0 \\ -1+\bar{\sigma}_L & 0 & \ldots & 0 & 0 & \bar{Y}-1 & \ldots & \bar{Y}-1 & \bar{Y}-1 \end{vmatrix},$$

where

$$\Delta\bar{\xi}^i = \bar{\xi}_G^i - \bar{\xi}_L^i, \qquad \overline{M}_{\alpha,j}^i = \frac{1}{\bar{\sigma}_\alpha} + \frac{\partial \mu_\alpha^i}{\partial \xi_\alpha^j}(\bar{\boldsymbol{\xi}}_\alpha), \tag{5.27}$$

for $\alpha \in \{G, L\}$, $(i, j) \in \{\mathrm{I}, \ldots, \mathrm{K}\}^2$. We subtract the $(\mathrm{K}+1)$-th column to each of the columns from the 2-nd to the K-th. Likewise, we subtract the $(2\mathrm{K}+1)$-th to each of the columns from the $(\mathrm{K}+2)$-th to the 2K-th. This yields

$$\overline{\mathfrak{d}}^{\bullet} = \begin{vmatrix} \Delta\bar{\xi}^{\mathrm{I}} & \bar{Y} & \ldots & 0 & 0 & 1-\bar{Y} & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \Delta\bar{\xi}^{\mathrm{K}-1} & 0 & \ldots & \bar{Y} & 0 & 0 & \ldots & 1-\bar{Y} & 0 \\ 0 & \widetilde{M}^{\mathrm{I}}_{G,\mathrm{I}} & \ldots & \widetilde{M}^{\mathrm{I}}_{G,\mathrm{K}-1} & \overline{M}^{\mathrm{I}}_{G,\mathrm{K}} & -\widetilde{M}^{\mathrm{I}}_{L,\mathrm{I}} & \ldots & -\widetilde{M}^{\mathrm{I}}_{L,\mathrm{K}-1} & -\overline{M}^{\mathrm{I}}_{L,\mathrm{K}} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & \widetilde{M}^{\mathrm{K}-1}_{G,\mathrm{I}} & \ldots & \widetilde{M}^{\mathrm{K}-1}_{G,\mathrm{K}-1} & \overline{M}^{\mathrm{K}-1}_{G,\mathrm{K}} & -\widetilde{M}^{\mathrm{K}-1}_{L,\mathrm{I}} & \ldots & -\widetilde{M}^{\mathrm{K}-1}_{L,\mathrm{K}-1} & -\overline{M}^{\mathrm{K}-1}_{L,\mathrm{K}} \\ 0 & \widetilde{M}^{\mathrm{K}}_{G,\mathrm{I}} & \ldots & \widetilde{M}^{\mathrm{K}}_{G,\mathrm{K}-1} & \overline{M}^{\mathrm{K}}_{G,\mathrm{K}} & -\widetilde{M}^{\mathrm{K}}_{L,\mathrm{I}} & \ldots & -\widetilde{M}^{\mathrm{K}}_{L,\mathrm{K}-1} & -\overline{M}^{\mathrm{K}}_{L,\mathrm{K}} \\ 1-\bar{\sigma}_G & 0 & \ldots & 0 & -\bar{Y} & 0 & \ldots & 0 & 0 \\ -1+\bar{\sigma}_L & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & \bar{Y}-1 \end{vmatrix},$$

with

$$\widetilde{M}^i_{\alpha,j} = \overline{M}^i_{\alpha,j} - \overline{M}^i_{\alpha,\mathrm{K}} = \left[\frac{\partial \mu^i_G}{\partial \xi^j_\alpha} - \frac{\partial \mu^i_\alpha}{\partial \xi^{\mathrm{K}}_\alpha}\right](\bar{\boldsymbol{\xi}}_\alpha) = \frac{1}{\bar{\sigma}_\alpha}\frac{\partial \mu^i_\alpha}{\partial x^j_\alpha}(\bar{\boldsymbol{x}}_\alpha), \tag{5.28}$$

for $\alpha \in \{G, L\}$, $i \in \{\mathrm{I}, \ldots, \mathrm{K}\}$, $j \in \{\mathrm{I}, \ldots, \mathrm{K}-1\}$. The last equality follows from the chain rule

$$\frac{\partial \mu^i_G}{\partial \xi^j_\alpha} - \frac{\partial \mu^i_\alpha}{\partial \xi^{\mathrm{K}}_\alpha} = \sum_{k=\mathrm{I}}^{\mathrm{K}-1}\left(\frac{\partial x^k_\alpha}{\partial \xi^j_\alpha} - \frac{\partial x^k_\alpha}{\partial \xi^{\mathrm{K}}_\alpha}\right)\frac{\partial \mu^i_G}{\partial x^k_\alpha}$$

and from

$$\frac{\partial x^k_\alpha}{\partial \xi^j_\alpha} = \frac{\delta_{j,k}\sigma_\alpha - \xi^k_\alpha}{\sigma^2_\alpha}, \qquad \frac{\partial x^k_\alpha}{\partial \xi^j_\alpha} = -\frac{\xi^k_\alpha}{\sigma^2_\alpha}.$$

Now, we subtract the $(2\mathrm{K}-1)$-th row to each of the rows from the K-th to the $(2\mathrm{K}-2)$-th. This gives

$$\overline{\mathfrak{d}}^{\bullet} = \begin{vmatrix} \Delta\bar{\xi}^{\mathrm{I}} & \bar{Y} & \ldots & 0 & 0 & 1-\bar{Y} & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \Delta\bar{\xi}^{\mathrm{K}-1} & 0 & \ldots & \bar{Y} & 0 & 0 & \ldots & 1-\bar{Y} & 0 \\ 0 & \widetilde{M}^{\mathrm{I}}_{G,\mathrm{I}} & \ldots & \widetilde{M}^{\mathrm{I}}_{G,\mathrm{K}-1} & \overline{\overline{M}}^{\mathrm{I}}_{G,\mathrm{K}} & -\widetilde{M}^{\mathrm{I}}_{L,\mathrm{I}} & \ldots & -\widetilde{M}^{\mathrm{I}}_{L,\mathrm{K}-1} & -\overline{\overline{M}}^{\mathrm{I}}_{L,\mathrm{K}} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & \widetilde{M}^{\mathrm{K}-1}_{G,\mathrm{I}} & \ldots & \widetilde{M}^{\mathrm{K}-1}_{G,\mathrm{K}-1} & \overline{\overline{M}}^{\mathrm{K}-1}_{G,\mathrm{K}} & -\widetilde{M}^{\mathrm{K}-1}_{L,\mathrm{I}} & \ldots & -\widetilde{M}^{\mathrm{K}-1}_{L,\mathrm{K}-1} & -\overline{\overline{M}}^{\mathrm{K}-1}_{L,\mathrm{K}} \\ 0 & \widetilde{M}^{\mathrm{K}}_{G,\mathrm{I}} & \ldots & \widetilde{M}^{\mathrm{K}}_{G,\mathrm{K}-1} & \overline{M}^{\mathrm{K}}_{G,\mathrm{K}} & -\widetilde{M}^{\mathrm{K}}_{L,\mathrm{I}} & \ldots & -\widetilde{M}^{\mathrm{K}}_{L,\mathrm{K}-1} & -\overline{M}^{\mathrm{K}}_{L,\mathrm{K}} \\ 1-\bar{\sigma}_G & 0 & \ldots & 0 & -\bar{Y} & 0 & \ldots & 0 & 0 \\ -1+\bar{\sigma}_L & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & \bar{Y}-1 \end{vmatrix},$$

with

$$\widetilde{M}^i_{\alpha,j} = \widetilde{M}^i_{\alpha,j} - \widetilde{M}^{\mathrm{K}}_{\alpha,j} = \frac{1}{\bar{\sigma}_\alpha}\left[\frac{\partial \mu^i_\alpha}{\partial x^j_\alpha} - \frac{\partial \mu^{\mathrm{K}}_\alpha}{\partial x^j_\alpha}\right](\bar{\boldsymbol{x}}_\alpha) = \frac{1}{\bar{\sigma}_\alpha}\frac{\partial^2 g_\alpha}{\partial x^i_\alpha \partial x^j_\alpha}(\bar{\boldsymbol{x}}_\alpha), \tag{5.29}$$

for $\alpha \in \{G, L\}$, $(i, j) \in \{\mathrm{I}, \ldots, \mathrm{K}-1\}^2$ in view of (2.24c), and

$$\overline{\overline{M}}^i_{\alpha,\mathrm{K}} = \overline{M}^i_{\alpha,\mathrm{K}} - \overline{M}^{\mathrm{K}}_{\alpha,\mathrm{K}} = \frac{\partial}{\partial \xi^{\mathrm{K}}_\alpha}(\mu^i_\alpha - \mu^{\mathrm{K}}_\alpha)(\bar{\boldsymbol{\xi}}_\alpha) = \frac{\partial}{\partial \xi^{\mathrm{K}}_\alpha}\left(\frac{\partial g_\alpha}{\partial x^i_\alpha}\right)(\bar{\boldsymbol{\xi}}_\alpha).$$

By the chain rule,

$$\frac{\partial}{\partial \xi_\alpha^{\mathrm{K}}} = \sum_{j=\mathrm{I}}^{\mathrm{K}-1} \frac{\partial x_\alpha^j}{\partial \xi_\alpha^{\mathrm{K}}} \frac{\partial}{\partial x_\alpha^j} = -\frac{1}{\sigma_\alpha} \sum_{j=\mathrm{I}}^{\mathrm{K}-1} x_\alpha^j \frac{\partial}{\partial x_\alpha^j}. \tag{5.30}$$

Applying this to $\partial g_\alpha / \partial x_\alpha^i$, we obtain

$$\overline{\overline{M}}{}_{\alpha,\mathrm{K}}^i = -\frac{1}{\bar{\sigma}_\alpha} \sum_{j=\mathrm{I}}^{\mathrm{K}-1} \bar{x}_\alpha^j \frac{\partial^2 g_\alpha}{\partial x_\alpha^i \partial x_\alpha^j}(\bar{\boldsymbol{x}}_\alpha).$$

This can be further transformed by observing that the Gibbs-Duhem condition (2.25) can be recast as

$$\begin{aligned} 0 &= \sum_{j=\mathrm{I}}^{\mathrm{K}-1} x_\alpha^j \frac{\partial \mu_\alpha^j}{\partial x_\alpha^i} + (1 - x_\alpha^{\mathrm{I}} - \ldots - x_\alpha^{\mathrm{K}-1}) \frac{\partial \mu_\alpha^{\mathrm{K}}}{\partial x_\alpha^i} \\ &= \sum_{j=\mathrm{I}}^{\mathrm{K}-1} x_\alpha^j \frac{\partial}{\partial x_\alpha^i}(\mu_\alpha^j - \mu_\alpha^{\mathrm{K}}) + \frac{\partial \mu_\alpha^{\mathrm{K}}}{\partial x_\alpha^i} \\ &= \sum_{j=\mathrm{I}}^{\mathrm{K}-1} x_\alpha^j \frac{\partial^2 g_\alpha}{\partial x_\alpha^i \partial x_\alpha^j} + \frac{\partial \mu_\alpha^{\mathrm{K}}}{\partial x_\alpha^i}. \end{aligned} \tag{5.31}$$

Hence,

$$\overline{\overline{M}}{}_{\alpha,\mathrm{K}}^i = \frac{1}{\bar{\sigma}_\alpha} \frac{\partial \mu_\alpha^{\mathrm{K}}}{\partial x_\alpha^i}(\bar{\boldsymbol{x}}_\alpha). \tag{5.32}$$

Single-phase solution. Assume $\bar{Y} = 1$, i.e., the solution is in the gas phase. Because $\bar{\nu} = 0$, we must have $\bar{\sigma}_G = 1$. Then,

$$\bar{\mathfrak{d}}^\bullet = \begin{vmatrix} \Delta\bar{\xi}^{\mathrm{I}} & 1 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \Delta\bar{\xi}^{\mathrm{K}-1} & 0 & \ldots & 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & \widetilde{M}_{G,\mathrm{I}}^{\mathrm{I}} & \ldots & \widetilde{M}_{G,\mathrm{K}-1}^{\mathrm{I}} & \overline{\overline{M}}{}_{G,\mathrm{K}}^{\mathrm{I}} & -\widetilde{M}_{L,\mathrm{I}}^{\mathrm{I}} & \ldots & -\widetilde{M}_{L,\mathrm{K}-1}^{\mathrm{I}} & -\overline{\overline{M}}{}_{L,\mathrm{K}}^{\mathrm{I}} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & \widetilde{M}_{G,\mathrm{I}}^{\mathrm{K}-1} & \ldots & \widetilde{M}_{G,\mathrm{K}-1}^{\mathrm{K}-1} & \overline{\overline{M}}{}_{G,\mathrm{K}}^{\mathrm{K}-1} & -\widetilde{M}_{L,\mathrm{I}}^{\mathrm{K}-1} & \ldots & -\widetilde{M}_{L,\mathrm{K}-1}^{\mathrm{K}-1} & -\overline{\overline{M}}{}_{L,\mathrm{K}}^{\mathrm{K}-1} \\ 0 & \widetilde{M}_{G,\mathrm{I}}^{\mathrm{K}} & \ldots & \widetilde{M}_{G,\mathrm{K}-1}^{\mathrm{K}} & \overline{M}_{G,\mathrm{K}}^{\mathrm{K}} & -\widetilde{M}_{L,\mathrm{I}}^{\mathrm{K}} & \ldots & -\widetilde{M}_{L,\mathrm{K}-1}^{\mathrm{K}} & -\overline{M}_{L,\mathrm{K}}^{\mathrm{K}} \\ 0 & 0 & \ldots & 0 & -1 & 0 & \ldots & 0 & 0 \\ -1 + \bar{\sigma}_L & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 \end{vmatrix},$$

Expanding the determinant with respect to the last two rows, we get

$$\bar{\mathfrak{d}}^\bullet = (-1)^{\mathrm{K}}(1 - \bar{\sigma}_L) \begin{vmatrix} 1 & \ldots & 0 & 0 & \ldots & 0 & 0 \\ \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & \ldots & 1 & 0 & \ldots & 0 & 0 \\ \widetilde{M}_{G,\mathrm{I}}^{\mathrm{I}} & \ldots & \widetilde{M}_{G,\mathrm{K}-1}^{\mathrm{I}} & -\widetilde{M}_{L,\mathrm{I}}^{\mathrm{I}} & \ldots & -\widetilde{M}_{L,\mathrm{K}-1}^{\mathrm{I}} & -\overline{\overline{M}}{}_{L,\mathrm{K}}^{\mathrm{I}} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots \\ \widetilde{M}_{G,\mathrm{I}}^{\mathrm{K}-1} & \ldots & \widetilde{M}_{G,\mathrm{K}-1}^{\mathrm{K}-1} & -\widetilde{M}_{L,\mathrm{I}}^{\mathrm{K}-1} & \ldots & -\widetilde{M}_{L,\mathrm{K}-1}^{\mathrm{K}-1} & -\overline{\overline{M}}{}_{L,\mathrm{K}}^{\mathrm{K}-1} \\ \widetilde{M}_{G,\mathrm{I}}^{\mathrm{K}} & \ldots & \widetilde{M}_{G,\mathrm{K}-1}^{\mathrm{K}} & -\widetilde{M}_{L,\mathrm{I}}^{\mathrm{K}} & \ldots & -\widetilde{M}_{L,\mathrm{K}-1}^{\mathrm{K}} & -\overline{M}_{L,\mathrm{K}}^{\mathrm{K}} \end{vmatrix}$$

Taking advantage of the block-triangular structure and using (5.27)–(5.29), (5.32),

$$\overline{\mathfrak{d}}^{\bullet} = (1-\bar\sigma_L)\begin{vmatrix} \widetilde{M}^{\mathrm{I}}_{L,\mathrm{I}} & \dots & \widetilde{M}^{\mathrm{I}}_{L,\mathrm{K}-1} & \overline{\overline{M}}^{\mathrm{I}}_{L,\mathrm{K}} \\ \vdots & & \vdots & \vdots \\ \widetilde{M}^{\mathrm{K}-1}_{L,\mathrm{I}} & \dots & \widetilde{M}^{\mathrm{K}-1}_{L,\mathrm{K}-1} & \overline{\overline{M}}^{\mathrm{K}-1}_{L,\mathrm{K}} \\ \widetilde{M}^{\mathrm{K}}_{L,\mathrm{I}} & \dots & \widetilde{M}^{\mathrm{K}}_{L,\mathrm{K}-1} & \overline{M}^{\mathrm{K}}_{L,\mathrm{K}} \end{vmatrix} = \frac{1-\bar\sigma_L}{(\bar\sigma_L)^{\mathrm{K}}}\begin{vmatrix} \nabla^2 g_L(\bar{\boldsymbol{x}}_L) & (\nabla\mu^{\mathrm{K}}_L)^T(\bar{\boldsymbol{x}}_L) \\ \nabla\mu^{\mathrm{K}}_L(\bar{\boldsymbol{x}}_L) & 1+\bar\sigma_L\dfrac{\partial\mu^{\mathrm{K}}_L}{\partial\xi^{\mathrm{K}}_L}(\bar{\boldsymbol{\xi}}_L) \end{vmatrix}.$$

Let C_j denote the j-th column of the latter $\mathrm{K}\times\mathrm{K}$-matrix. We perform the column substitution $C_{\mathrm{K}} \leftarrow C_{\mathrm{K}} + \sum_{j=1}^{\mathrm{K}-1}\bar x^j_L C_j$ and invoke (5.30)–(5.31) to end up with

$$\overline{\mathfrak{d}}^{\bullet} = \frac{1-\bar\sigma_L}{(\bar\sigma_L)^{\mathrm{K}}}\begin{vmatrix} \nabla^2 g_L(\bar{\boldsymbol{x}}_L) & 0 \\ \nabla\mu^{\mathrm{K}}_L(\bar{\boldsymbol{x}}_L) & 1 \end{vmatrix} = \frac{1-\bar\sigma_L}{(\bar\sigma_L)^{\mathrm{K}}}\det\nabla^2 g_L(\bar{\boldsymbol{x}}_L).$$

Thanks to the strict convexity assumption, $\det\nabla^2 g_L(\bar{\boldsymbol{x}}_L) > 0$. Because of the complementarity condition $0 \leqslant 1-\bar Y \perp 1-\bar\sigma_L \geqslant 0$, we have $\overline{\mathfrak{d}}^{\bullet} \geqslant 0$. If $1-\bar\sigma_L > 0$, that is, if the solution is not a transition point, then $\overline{\mathfrak{d}}^{\bullet} > 0$ and we have a regular zero. Otherwise, the zero is singular. The other single-phase case $\bar Y = 0$ can be dealt with analogously. We obtain $\overline{\mathfrak{d}}^{\bullet} = [(1-\bar\sigma_G)/(\bar\sigma_G)^{\mathrm{K}}]\det\nabla^2 g_G(\bar{\boldsymbol{x}}_G)$, from which a similar conclusion can be drawn.

Two-phase solution. Assume $\bar Y \in (0,1)$. Then, $\bar\sigma_G = \bar\sigma_L = 0$ and $\bar{\boldsymbol{\xi}}_\alpha = \bar{\boldsymbol{x}}_\alpha$. We shall therefore write $\Delta\bar x^i$ instead of $\Delta\bar\xi^i$. Expanding

$$\overline{\mathfrak{d}}^{\bullet} = \begin{vmatrix} \Delta\bar x^{\mathrm{I}} & \bar Y & \dots & 0 & 0 & 1-\bar Y & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \Delta\bar x^{\mathrm{K}-1} & 0 & \dots & \bar Y & 0 & 0 & \dots & 1-\bar Y & 0 \\ 0 & \widetilde{M}^{\mathrm{I}}_{G,\mathrm{I}} & \dots & \widetilde{M}^{\mathrm{I}}_{G,\mathrm{K}-1} & \overline{\overline{M}}^{\mathrm{I}}_{G,\mathrm{K}} & -\widetilde{M}^{\mathrm{I}}_{L,\mathrm{I}} & \dots & -\widetilde{M}^{\mathrm{I}}_{L,\mathrm{K}-1} & -\overline{\overline{M}}^{\mathrm{I}}_{L,\mathrm{K}} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & \widetilde{M}^{\mathrm{K}-1}_{G,\mathrm{I}} & \dots & \widetilde{M}^{\mathrm{K}-1}_{G,\mathrm{K}-1} & \overline{\overline{M}}^{\mathrm{K}-1}_{G,\mathrm{K}} & -\widetilde{M}^{\mathrm{K}-1}_{L,\mathrm{I}} & \dots & -\widetilde{M}^{\mathrm{K}-1}_{L,\mathrm{K}-1} & -\overline{\overline{M}}^{\mathrm{K}-1}_{L,\mathrm{K}} \\ 0 & \widetilde{M}^{\mathrm{K}}_{G,\mathrm{I}} & \dots & \widetilde{M}^{\mathrm{K}}_{G,\mathrm{K}-1} & \overline{M}^{\mathrm{K}}_{G,\mathrm{K}} & -\widetilde{M}^{\mathrm{K}}_{L,\mathrm{I}} & \dots & -\widetilde{M}^{\mathrm{K}}_{L,\mathrm{K}-1} & -\overline{M}^{\mathrm{K}}_{L,\mathrm{K}} \\ 0 & 0 & \dots & 0 & -\bar Y & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & \bar Y-1 \end{vmatrix}$$

with respect to the last two rows, we arrive at

$$\begin{aligned} \overline{\mathfrak{d}}^{\bullet} &= \bar Y(1-\bar Y)\begin{vmatrix} \Delta\bar x^{\mathrm{I}} & \bar Y & \dots & 0 & \bar Y-1 & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \Delta\bar x^{\mathrm{K}-1} & 0 & \dots & \bar Y & 0 & \dots & \bar Y-1 \\ 0 & \widetilde{M}^{\mathrm{I}}_{G,\mathrm{I}} & \dots & \widetilde{M}^{\mathrm{I}}_{G,\mathrm{K}-1} & \widetilde{M}^{\mathrm{I}}_{L,\mathrm{I}} & \dots & \widetilde{M}^{\mathrm{I}}_{L,\mathrm{K}-1} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & \widetilde{M}^{\mathrm{K}-1}_{G,\mathrm{I}} & \dots & \widetilde{M}^{\mathrm{K}-1}_{G,\mathrm{K}-1} & \widetilde{M}^{\mathrm{K}-1}_{L,\mathrm{I}} & \dots & \widetilde{M}^{\mathrm{K}-1}_{L,\mathrm{K}-1} \\ 0 & \widetilde{M}^{\mathrm{K}}_{G,\mathrm{I}} & \dots & \widetilde{M}^{\mathrm{K}}_{G,\mathrm{K}-1} & \widetilde{M}^{\mathrm{K}}_{L,\mathrm{I}} & \dots & \widetilde{M}^{\mathrm{K}}_{L,\mathrm{K}-1} \end{vmatrix} \\ &= \bar Y(1-\bar Y)\begin{vmatrix} \Delta\bar{\boldsymbol{x}} & \bar Y I_{\mathrm{K}-1} & (\bar Y-1)I_{\mathrm{K}-1} \\ 0 & \nabla^2 g_G(\bar{\boldsymbol{x}}_G) & \nabla^2 g_L(\bar{\boldsymbol{x}}_L) \\ 0 & \nabla\mu^{\mathrm{K}}_G(\bar{\boldsymbol{x}}_G) & \nabla\mu^{\mathrm{K}}_L(\bar{\boldsymbol{x}}_L) \end{vmatrix}. \end{aligned}$$

For $j \in \{2, \ldots, \mathrm{K}\}$, we perform the column substitution $C_j \leftarrow C_j + \frac{\bar{Y}}{1-\bar{Y}} C_{j+\mathrm{K}-1}$ to obtain

$$\overline{\mathfrak{d}}^{\bullet} = \bar{Y}(1-\bar{Y}) \begin{vmatrix} \Delta\bar{\boldsymbol{x}} & 0 & (\bar{Y}-1) I_{\mathrm{K}-1} \\ 0 & \nabla^2 g_G(\bar{\boldsymbol{x}}_G) + \dfrac{\bar{Y}}{1-\bar{Y}} \nabla^2 g_L(\bar{\boldsymbol{x}}_L) & \nabla^2 g_L(\bar{\boldsymbol{x}}_L) \\ 0 & \nabla \mu_G^{\mathrm{K}}(\bar{\boldsymbol{x}}_G) + \dfrac{\bar{Y}}{1-\bar{Y}} \nabla \mu_L^{\mathrm{K}}(\bar{\boldsymbol{x}}_L) & \nabla \mu_L^{\mathrm{K}}(\bar{\boldsymbol{x}}_L) \end{vmatrix}.$$

Expanding with respect to the first column, we have

$$\overline{\mathfrak{d}}^{\bullet} = \bar{Y}(1-\bar{Y}) \sum_{i=\mathrm{I}}^{\mathrm{K}-1} (-1)^{i-1} \Delta x^i M_i,$$

where M_i is the $\mathrm{K} \times \mathrm{K}$-matrix obtained by removing the i-th line and the first column of $\overline{\mathfrak{d}}^{\bullet}$. To compute M_i, we expand it with respect to its first $\mathrm{K}-2$ rows, each of which contains exactly one nonzero entry, equal to $\bar{Y}-1$. When doing the expansion, we must pay a lot of attention to the sign of the various minors involved. At the end of the algebra, we obtain

$$(-1)^{i-1} M_i = (1-\bar{Y})^{\mathrm{K}-2} \begin{vmatrix} \nabla^2 g_G(\bar{\boldsymbol{x}}_G) + \dfrac{\bar{Y}}{1-\bar{Y}} \nabla^2 g_L(\bar{\boldsymbol{x}}_L) & \dfrac{\partial (\nabla g_L)^T}{\partial x_L^i}(\bar{\boldsymbol{x}}_L) \\ \nabla \mu_G^{\mathrm{K}}(\bar{\boldsymbol{x}}_G) + \dfrac{\bar{Y}}{1-\bar{Y}} \nabla \mu_L^{\mathrm{K}}(\bar{\boldsymbol{x}}_L) & \dfrac{\partial \mu_L^{\mathrm{K}}}{\partial x_L^i}(\bar{\boldsymbol{x}}_L) \end{vmatrix},$$

where $\partial (\nabla g_L)^T / \partial x_L^i (\bar{\boldsymbol{x}}_L)$ is the i-th column of $\nabla^2 g_L(\bar{\boldsymbol{x}}_L)$. Multiplying by Δx_L^i, summing over i and redistributing $(1-\bar{Y})^{\mathrm{K}-1}$ to the first $\mathrm{K}-1$ columns result in

$$\overline{\mathfrak{d}}^{\bullet} = \bar{Y} \begin{vmatrix} (1-\bar{Y}) \nabla^2 g_G(\bar{\boldsymbol{x}}_G) + \bar{Y} \nabla^2 g_L(\bar{\boldsymbol{x}}_L) & \nabla^2 g_L(\bar{\boldsymbol{x}}) \Delta\bar{\boldsymbol{x}} \\ (1-\bar{Y}) \nabla \mu_G^{\mathrm{K}}(\bar{\boldsymbol{x}}_G) + \bar{Y} \nabla \mu_L^{\mathrm{K}}(\bar{\boldsymbol{x}}_L) & \nabla \mu_L^{\mathrm{K}}(\bar{\boldsymbol{x}}_L) \Delta\bar{\boldsymbol{x}} \end{vmatrix}.$$

Let R_i denote the i-th row of the latter $\mathrm{K} \times \mathrm{K}$-matrix. We perform the row substitution $R_{\mathrm{K}} \leftarrow R_{\mathrm{K}} + \sum_{i=\mathrm{I}}^{\mathrm{K}-1} \bar{x}_G^i R_i$. After (5.31),

$$\bar{\boldsymbol{x}}_G^T \nabla^2 g_G(\bar{\boldsymbol{x}}_G) + \nabla \mu_G^{\mathrm{K}}(\bar{\boldsymbol{x}}_G) = 0.$$

To compute $\bar{\boldsymbol{x}}_G^T \nabla^2 g_L(\bar{\boldsymbol{x}}_L) + \nabla \mu_L^{\mathrm{K}}(\bar{\boldsymbol{x}}_L)$, we start from

$$\bar{\boldsymbol{x}}_L^T \nabla^2 g_L(\bar{\boldsymbol{x}}_L) + \nabla \mu_L^{\mathrm{K}}(\bar{\boldsymbol{x}}_L) = 0,$$

which is also due to (5.31). Since $\bar{\boldsymbol{x}}_G = \bar{\boldsymbol{x}}_L + \Delta\bar{\boldsymbol{x}}$, we have

$$\bar{\boldsymbol{x}}_G^T \nabla^2 g_L(\bar{\boldsymbol{x}}_L) + \nabla \mu_L^{\mathrm{K}}(\bar{\boldsymbol{x}}_L) = \Delta\bar{\boldsymbol{x}}^T \nabla^2 g_L(\bar{\boldsymbol{x}}_L).$$

As a result,

$$\overline{\mathfrak{d}}^{\bullet} = \bar{Y} \begin{vmatrix} (1-\bar{Y}) \nabla^2 g_G(\bar{\boldsymbol{x}}_G) + \bar{Y} \nabla^2 g_L(\bar{\boldsymbol{x}}_L) & \nabla^2 g_L(\bar{\boldsymbol{x}}) \Delta\bar{\boldsymbol{x}} \\ \bar{Y} \Delta\bar{\boldsymbol{x}}^T \nabla^2 g_L(\bar{\boldsymbol{x}}_L) & \Delta\bar{\boldsymbol{x}}^T \nabla^2 g_L(\bar{\boldsymbol{x}}_L) \Delta\bar{\boldsymbol{x}} \end{vmatrix}.$$

Now, we expand this determinant with respect to the last row. In doing so, we see that each entry of the row vector $\bar{Y}(\Delta\bar{\boldsymbol{x}})^T \nabla^2 g_L(\bar{\boldsymbol{x}}_L)$ will be multiplied by the determinant of a matrix in which the corresponding column of $(1-\bar{Y}) \nabla^2 g_G(\bar{\boldsymbol{x}}_G) + \bar{Y} \nabla^2 g_L(\bar{\boldsymbol{x}}_L)$ has been replaced by the

column vector $\nabla^2 g_L(\bar{\boldsymbol{x}})\Delta\bar{\boldsymbol{x}}$, up to a permutation. This is reminiscent of Cramer's rule for solving a linear system, except for the fact that the determinant of $(1-\bar{Y})\nabla^2 g_G(\bar{\boldsymbol{x}}_G) + \bar{Y}\nabla^2 g_L(\bar{\boldsymbol{x}}_L)$ is missing here. Guided by this intuition, we can readily check that

$$\overline{\mathfrak{d}}^{\bullet} = \bar{Y}\det\big[(1-\bar{Y})\nabla^2 g_G + \bar{Y}\nabla^2 g_L\big]\,\Delta\bar{\boldsymbol{x}}^T \mathsf{M}_L \Delta\bar{\boldsymbol{x}}, \tag{5.33}$$

with

$$\mathsf{M}_L = \nabla^2 g_L - \bar{Y}\,\nabla^2 g_L\big[(1-\bar{Y})\nabla^2 g_G + \bar{Y}\nabla^2 g_L\big]^{-1}\nabla^2 g_L,$$

where we have dropped the arguments $\bar{\boldsymbol{x}}_G$ and $\bar{\boldsymbol{x}}_L$ for short. This matrix can be rearranged as

$$\begin{aligned}\mathsf{M}_L &= \nabla^2 g_L\Big\{\big[\nabla^2 g_L\big]^{-1} - \Big[\nabla^2 g_L + \frac{1-\bar{Y}}{\bar{Y}}\nabla^2 g_G\Big]^{-1}\Big\}\nabla^2 g_L \\ &= (\nabla^2 g_L)^{1/2}\Big\{I_{\mathrm{K}-1} - \Big[I_{\mathrm{K}-1} + \frac{1-\bar{Y}}{\bar{Y}}(\nabla^2 g_L)^{-1/2}\,\nabla^2 g_G\,(\nabla^2 g_L)^{-1/2}\Big]^{-1}\Big\}(\nabla^2 g_L)^{1/2}.\end{aligned}$$

We could have done calculations the other way around and this would have given us

$$\overline{\mathfrak{d}}^{\bullet} = (1-\bar{Y})\det\big[(1-\bar{Y})\nabla^2 g_G + \bar{Y}\nabla^2 g_L\big]\,\Delta\bar{\boldsymbol{x}}^T \mathsf{M}_G \Delta\bar{\boldsymbol{x}}, \tag{5.34}$$

with

$$\mathsf{M}_G = (\nabla^2 g_G)^{1/2}\Big\{I_{\mathrm{K}-1} - \Big[I_{\mathrm{K}-1} + \frac{1-\bar{Y}}{\bar{Y}}(\nabla^2 g_G)^{-1/2}\,\nabla^2 g_L\,(\nabla^2 g_G)^{-1/2}\Big]^{-1}\Big\}(\nabla^2 g_G)^{1/2}.$$

To restore symmetry, we consider the combination $(1-\bar{Y})\cdot$ (5.33) $+\bar{Y}\cdot$ (5.34) and thus obtain

$$\overline{\mathfrak{d}}^{\bullet} = \bar{Y}(1-\bar{Y})\det\big[(1-\bar{Y})\nabla^2 g_G + \bar{Y}\nabla^2 g_L\big]\,\Delta\bar{\boldsymbol{x}}^T(\mathsf{M}_G + \mathsf{M}_L)\Delta\bar{\boldsymbol{x}}.$$

By Lemma 5.6 and the strict convexity assumption, the symmetric matrix $\mathsf{M}_G + \mathsf{M}_L$ is positive definite. Hence, $\overline{\mathfrak{d}}^{\bullet} \geqslant 0$ and equality $\overline{\mathfrak{d}}^{\bullet} = 0$ occurs if and only if $\Delta\bar{\boldsymbol{x}} = 0$, that is $\bar{\boldsymbol{x}}_G = \bar{\boldsymbol{x}}_L$. This is precisely the characterization of an azeotropic solution. □

5.2.2 A special proof for Henry's law

In the special case where the Gibbs functions g_G and g_L are derived from the ideal or Henry's law (see §3.1.1), there is a shorter proof of regularity for the nondegenerate zeros of (2.77). We present it below for mathematical completeness. Let us consider the fugacity laws

$$\Phi_G^i \equiv 1, \qquad \Phi_L^i \equiv k^i, \qquad i \in \{\mathrm{I}, \ldots, \mathrm{K}\},$$

where the constants k^i's are positive. The equality of extended fugacities (2.77b) becomes $\xi_G^i = k^i\xi_L^i$ for $i \in \{\mathrm{I}, \ldots, \mathrm{K}\}$. By eliminating the liquid fractions ξ_L^i in (2.77), we obtain an equivalent system of $\mathrm{K}+1$ equations

$$Y\xi_G^{\mathrm{I}} + (1-Y)\frac{\xi_G^{\mathrm{I}}}{k^{\mathrm{I}}} - c^{\mathrm{I}} = 0, \tag{5.35a}$$

$$\vdots$$

$$Y\xi_G^{\mathrm{K}-1} + (1-Y)\frac{\xi_G^{\mathrm{K}-1}}{k^{\mathrm{K}-1}} - c^{\mathrm{K}-1} = 0, \tag{5.35b}$$

$$\min\big(Y, 1 - \textstyle\sum_{j=1}^{\mathrm{K}}\xi_G^j\big) = 0, \tag{5.35c}$$

$$\min\left(1-Y, 1-\textstyle\sum_{j=1}^{\mathrm{K}} \xi_G^j/k^j\right) = 0. \tag{5.35d}$$

in the unknowns $X = (Y, \xi_G^{\mathrm{I}}, \dots \xi_G^{\mathrm{K}}) \in \mathbb{R}^{\mathrm{K}+1}$. With respect to the abstract framework, this model corresponds to

$$\ell = \mathrm{K}-1, \qquad m = 2,$$

with the continuous-differentiable functions

$$\Lambda(X) = \begin{bmatrix} Y\xi_G^{\mathrm{I}} + (1-Y)\dfrac{\xi_G^{\mathrm{I}}}{k^{\mathrm{I}}} - c^{\mathrm{I}} \\ \vdots \\ Y\xi_G^{\mathrm{K}-1} + (1-Y)\dfrac{\xi_G^{\mathrm{K}-1}}{k^{\mathrm{K}-1}} - c^{\mathrm{K}-1} \end{bmatrix} \in \mathbb{R}^{\mathrm{K}-1}$$

and

$$G(X) = \begin{bmatrix} Y \\ 1-Y \end{bmatrix} \in \mathbb{R}^2, \qquad H(X) = \begin{bmatrix} 1-\xi_G^{\mathrm{I}} - \dots - \xi_G^{\mathrm{K}} \\ 1-\xi_G^{\mathrm{I}}/k^{\mathrm{I}} - \dots - \xi_G^{\mathrm{K}}/k^{\mathrm{K}} \end{bmatrix} \in \mathbb{R}^2.$$

The determinant to be computed reads

$$\begin{aligned} \bar{\mathfrak{d}} &= \begin{vmatrix} \nabla\Lambda(\bar{X}) \\ \nabla G(\bar{X}) \odot H(\bar{X}) + \nabla H(\bar{X}) \odot G(\bar{X}) \end{vmatrix} \\ &= \begin{vmatrix} \Delta\bar{\xi}^{\mathrm{I}} & \bar{Y} + (1-\bar{Y})/k^{\mathrm{I}} & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ \Delta\bar{\xi}^{\mathrm{K}-1} & 0 & \dots & \bar{Y} + (1-\bar{Y})/k^{\mathrm{K}-1} & 0 \\ 1-\bar{\sigma}_G & -\bar{Y} & \dots & -\bar{Y} & -\bar{Y} \\ -1+\bar{\sigma}_L & (\bar{Y}-1)/k^{\mathrm{I}} & \dots & (\bar{Y}-1)/k^{\mathrm{K}-1} & (\bar{Y}-1)/k^{\mathrm{K}} \end{vmatrix}, \end{aligned}$$

where

$$\Delta\bar{\xi}^i = \bar{\xi}_G^i - \bar{\xi}_L^i = \bar{\xi}_G^i(1-1/k^i), \qquad \bar{\sigma}_G = \xi_G^{\mathrm{I}} + \dots + \xi_G^{\mathrm{K}}, \qquad \bar{\sigma}_L = \xi_G^{\mathrm{I}}/k^{\mathrm{I}} + \dots + \xi_G^{\mathrm{K}}/k^{\mathrm{K}}.$$

Single-phase solution. Assume $\bar{Y} = 1$, i.e., the solution is in the gas phase. Because $\bar{\nu} = 0$, we must have $\bar{\sigma}_G = 1$. Then, the last row of the above matrix is zero except for $1-\bar{\sigma}_L$. By expanding the determinant with respect to this row, we have

$$\bar{\mathfrak{d}} = (-1)^{\mathrm{K}+2}(-1+\bar{\sigma}_L)\begin{vmatrix} 1 & \dots & 0 & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \dots & 1 & 0 \\ -1 & \dots & -1 & -1 \end{vmatrix} = (-1)^{\mathrm{K}}(1-\bar{\sigma}_L).$$

This can only vanish if $1-\bar{\sigma}_L = 0$. Combined with $1-\bar{Y} = 0$, this implies that the solution is a transition point. The case $\bar{Y} = 0$ can be dealt with analogously, with

$$\bar{\mathfrak{d}} = (-1)^{\mathrm{K}}(1-\bar{\sigma}_G)/(k^{\mathrm{I}}\cdots k^{\mathrm{K}-1}k^{\mathrm{K}}),$$

and the conclusion is the same.

Two-phase solution. Assume $\bar{Y} \in (0,1)$. Then, $\bar{\sigma}_G = \bar{\sigma}_L = 1$ and $\xi_\alpha^j = x_\alpha^j$ for $\alpha \in \{G, L\}$,

$j \in \{\mathrm{I}, \ldots, \mathrm{K}\}$. Let R_i be the i-th row of the matrix defining $\overline{\mathfrak{d}}$. We perform the row substitution $R_{\mathrm{K}+1} \leftarrow R_{\mathrm{K}+1} + R_{\mathrm{K}} + \ldots + R_{\mathrm{I}}$ to obtain

$$\overline{\mathfrak{d}} = \begin{vmatrix} \Delta\bar{x}^{\mathrm{I}} & \bar{Y} + (1-\bar{Y})/k^{\mathrm{I}} & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ \Delta\bar{x}^{\mathrm{K}-1} & 0 & \ldots & \bar{Y} + (1-\bar{Y})/k^{\mathrm{K}-1} & 0 \\ 0 & -\bar{Y} & \ldots & -\bar{Y} & -\bar{Y} \\ -\Delta\bar{x}^{\mathrm{K}} & 0 & \ldots & 0 & -\bar{Y} - (1-\bar{Y})/k^{\mathrm{K}} \end{vmatrix},$$

the last entry of the first column being due to

$$\sum_{j=1}^{\mathrm{K}-1} \Delta\bar{x}^j = \sum_{j=1}^{\mathrm{K}-1} (\bar{x}_G^j - \bar{x}_L^j) = (1 - \bar{x}_G^{\mathrm{K}}) - (1 - \bar{x}_L^{\mathrm{K}}) = -(\bar{x}_G^{\mathrm{K}} - \bar{x}_L^{\mathrm{K}}).$$

By swapping the last two rows, changing signs in the penultimate row and factorizing by $-\bar{Y}$ from the last one, we get

$$\overline{\mathfrak{d}} = -\bar{Y} \begin{vmatrix} \Delta\bar{x}^{\mathrm{I}} & \bar{Y} + (1-\bar{Y})/k^{\mathrm{I}} & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ \Delta\bar{x}^{\mathrm{K}-1} & 0 & \ldots & \bar{Y} + (1-\bar{Y})/k^{\mathrm{K}-1} & 0 \\ \Delta\bar{x}^{\mathrm{K}} & 0 & \ldots & 0 & \bar{Y} + (1-\bar{Y})/k^{\mathrm{K}} \\ 0 & 1 & \ldots & 1 & 1 \end{vmatrix}. \tag{5.36}$$

Starting from the original matrix, if we had performed the row substitution $R_{\mathrm{K}} \leftarrow R_{\mathrm{K}+1} + R_{\mathrm{K}} + \ldots + R_{\mathrm{I}}$ instead, we would have ended up with

$$\overline{\mathfrak{d}} = \begin{vmatrix} \Delta\bar{x}^{\mathrm{I}} & \bar{Y} + (1-\bar{Y})/k^{\mathrm{I}} & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ \Delta\bar{x}^{\mathrm{K}-1} & 0 & \ldots & \bar{Y} + (1-\bar{Y})/k^{\mathrm{K}-1} & 0 \\ -\Delta\bar{x}^{\mathrm{K}} & 0 & \ldots & 0 & -\bar{Y} - (1-\bar{Y})/k^{\mathrm{K}} \\ 0 & (\bar{Y}-1)/k^{\mathrm{I}} & \ldots & (\bar{Y}-1)/k^{\mathrm{K}-1} & (\bar{Y}-1)/k^{\mathrm{K}} \end{vmatrix}.$$

Changing signs in the last two rows and factorizing by $-(1-\bar{Y})$ in the last row, we obtain $R_{\mathrm{K}} + \ldots + R_{\mathrm{I}}$ instead, we would have ended up with

$$\overline{\mathfrak{d}} = -(1-\bar{Y}) \begin{vmatrix} \Delta\bar{x}^{\mathrm{I}} & \bar{Y} + (1-\bar{Y})/k^{\mathrm{I}} & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ \Delta\bar{x}^{\mathrm{K}-1} & 0 & \ldots & \bar{Y} + (1-\bar{Y})/k^{\mathrm{K}-1} & 0 \\ \Delta\bar{x}^{\mathrm{K}} & 0 & \ldots & 0 & \bar{Y} + (1-\bar{Y})/k^{\mathrm{K}} \\ 0 & -1/k^{\mathrm{I}} & \ldots & -1/k^{\mathrm{K}-1} & -1/k^{\mathrm{K}} \end{vmatrix}. \tag{5.37}$$

To recover symmetry, we consider $(1-\bar{Y})\cdot$ (5.36) $+\bar{Y}\cdot$ (5.37). This yields

$$\overline{\mathfrak{d}} = -\bar{Y}(1-\bar{Y}) \begin{vmatrix} \Delta\bar{x}^{\mathrm{I}} & \bar{Y} + (1-\bar{Y})/k^{\mathrm{I}} & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ \Delta\bar{x}^{\mathrm{K}-1} & 0 & \ldots & \bar{Y} + (1-\bar{Y})/k^{\mathrm{K}-1} & 0 \\ \Delta\bar{x}^{\mathrm{K}} & 0 & \ldots & 0 & \bar{Y} + (1-\bar{Y})/k^{\mathrm{K}} \\ 0 & 1 - 1/k^{\mathrm{I}} & \ldots & 1 - 1/k^{\mathrm{K}-1} & 1 - 1/k^{\mathrm{K}} \end{vmatrix}$$

After K column permutations, we can put the determinant under the form

$$\overline{\mathfrak{d}} = (-1)^{\mathrm{K}+1}\bar{Y}(1-\bar{Y})\begin{vmatrix} \bar{Y}+(1-\bar{Y})/k^{\mathrm{I}} & \dots & 0 & 0 & \Delta\bar{x}^{\mathrm{I}} \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \dots & \bar{Y}+(1-\bar{Y})/k^{\mathrm{K}-1} & 0 & \Delta\bar{x}^{\mathrm{K}-1} \\ 0 & \dots & 0 & \bar{Y}+(1-\bar{Y})/k^{\mathrm{K}} & \Delta\bar{x}^{\mathrm{K}} \\ 1-1/k^{\mathrm{I}} & \dots & 1-1/k^{\mathrm{K}-1} & 1-1/k^{\mathrm{K}} & 0 \end{vmatrix}.$$

Arguing that $1-1/k^j = (\bar{x}^j_G - \bar{x}^j_G/k^j)/\bar{x}^j_G = \Delta\bar{x}^j/\bar{x}^j_G$, we can write

$$\overline{\mathfrak{d}} = (-1)^{\mathrm{K}+1}\bar{Y}(1-\bar{Y})\begin{vmatrix} \bar{Y}+(1-\bar{Y})/k^{\mathrm{I}} & \dots & 0 & 0 & \Delta\bar{x}^{\mathrm{I}} \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \dots & \bar{Y}+(1-\bar{Y})/k^{\mathrm{K}-1} & 0 & \Delta\bar{x}^{\mathrm{K}-1} \\ 0 & \dots & 0 & \bar{Y}+(1-\bar{Y})/k^{\mathrm{K}} & \Delta\bar{x}^{\mathrm{K}} \\ \Delta\bar{x}^{\mathrm{I}}/\bar{x}^{\mathrm{I}}_G & \dots & \Delta\bar{x}^{\mathrm{K}-1}/\bar{x}^{\mathrm{K}-1}_G & \Delta\bar{x}^{\mathrm{K}}/\bar{x}^{\mathrm{K}}_G & 0 \end{vmatrix}.$$

To make the determinant even more symmetric, let us multiply each column j by $(\bar{x}^j_G)^{1/2}$ and divide each row i by $(\bar{x}^i_G)^{1/2}$. Overall, after sweeping over all columns and all rows, we do not change the determinant. Setting

$$\bar{z}^i = \frac{\Delta\bar{x}^i_G}{(\bar{x}^i_G)^{1/2}}, \qquad i \in \{1, \dots, \mathrm{K}\},$$

we can now write the determinant as

$$\overline{\mathfrak{d}} = (-1)^{\mathrm{K}+1}\bar{Y}(1-\bar{Y})\begin{vmatrix} \bar{Y}+(1-\bar{Y})/k^{\mathrm{I}} & \dots & 0 & 0 & \bar{z}^{\mathrm{I}} \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \dots & \bar{Y}+(1-\bar{Y})/k^{\mathrm{K}-1} & 0 & \bar{z}^{\mathrm{K}-1} \\ 0 & \dots & 0 & \bar{Y}+(1-\bar{Y})/k^{\mathrm{K}} & \bar{z}^{\mathrm{K}} \\ \bar{z}^{\mathrm{I}} & \dots & \bar{z}^{\mathrm{K}-1} & \bar{z}^{\mathrm{K}} & 0 \end{vmatrix}.$$

We expand this new form of the determinant with respect to the last row, using the same technique as in the previous section for the general proof: each entry of the row vector $\bar{\boldsymbol{z}}^T = (\bar{z}_{\mathrm{I}}, \dots, \bar{z}_{\mathrm{K}})$ will be multiplied by the determinant of a matrix in which the corresponding column of

$$\mathsf{D} = \operatorname{diag}\left(\bar{Y}+(1-\bar{Y})/k^{\mathrm{I}}, \dots, \bar{Y}+(1-\bar{Y})/k^{\mathrm{K}}\right)$$

has been replaced by the column vector $\bar{\boldsymbol{z}} \in \mathbb{R}^{\mathrm{K}}$, up to a permutation. This somehow reminds us of Cramer's rule for solving a linear system. Exploring this path, it can be then proven that

$$\overline{\mathfrak{d}} = (-1)^{\mathrm{K}}\ \bar{Y}(1-\bar{Y})\ \ \bar{\boldsymbol{z}}^T \operatorname{adj}(\mathsf{D})\bar{\boldsymbol{z}},$$

where $\operatorname{adj}(\mathsf{D}) = \det(\mathsf{D})\,\mathsf{D}^{-1}$ denotes the adjugate matrix of D. This adjugate matrix is easily seen to be symmetric and positive definite. Therefore, $\overline{\mathfrak{d}} = 0$ if and only if $\bar{\boldsymbol{z}} = 0$, which is equivalent to $\Delta\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}_G - \bar{\boldsymbol{x}}_L = 0$. In other words, the solution is azeotropic.

Chapter 6

Numerical experiments on various models

Contents

Ce chapitre rend compte des essais numériques que nous avons effectués avec plusieurs algorithmes sur cinq modèles représentatifs des problèmes avec conditions de complémentarité qui intéressent les chercheurs d'IFPEN.

Nous qualifions de "simplifiés" les quatre premiers modèles, présentés en §6.1, en raison de leurs petites tailles (moins d'une dizaine de variables). Deux sont de nature intrinsèquement stationnaire, deux proviennent de la discrétisation d'un problème d'évolution. Classés par ordre de difficulté croissante, ils permettent de trier, par élimination progressive des plus mauvais, les algorithmes en compétition et de faire émerger le meilleur d'entre eux, NPIPM, ainsi que la méthode de référence pour la famille semi-lisse, Newton-min.

Ceux-ci sont ensuite appliqués en §6.2 à un modèle d'écoulement diphasique (partiellement triphasique) compositionnel en deux dimensions d'espace, qui n'est certes pas aussi complexe qu'un modèle de réservoir usuel mais dont les lois thermodynamiques sont complètes et réalistes. Nous décrirons le modèle, mais pas la discrétisation en temps et en espace. Deux tests d'injection de CO_2 *seront considérés et mettront en évidence les lacunes actuelles de NPIPM.*

We apply the numerical methods of chapters §4–§5 to various physical models of interest, presented here in the order of increasing complexity. The competing algorithms are gradually left out, based on their performance. At the end of this process, only two of them remain NPIPM and Newton-min. The latter is the default method in many industrial codes and serves as the reference semismooth algorithm for our comparison.

6.1 Simplified models

6.1.1 Stratigraphic model

Continuous model. We consider the differential equation

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -\min(u^2,\, 1), \tag{6.1a}$$

$$u(t=0) = u_0. \tag{6.1b}$$

The unknown u represents the height $u(t) \in \mathbb{R}^+$ of sediments in a basin as a function of time t. Since the right-hand side of (6.1a) is always nonpositive, u is a nonincreasing function of t. In other words, the basin is always eroding. However, this erosion can occur at two different regimes: (i) if $u^2 < 1$, then the erosion rate is equal to $-u^2$; this is the "unsaturated" regime; (ii) if $u^2 > 1$, then the erosion rate is equal to -1, a maximal erosion rate prescribed by geologists; this is the "saturated" regime.

It is very easy to show that the solution of system (6.1) is given by

$$u(t) = \begin{cases} u_0 - t & \text{for } 0 \leqslant t \leqslant t_*, \\ \dfrac{u(t_*)}{1 + u(t_*)(t - t_*)} & \text{for } t > t_*, \end{cases} \tag{6.2}$$

where $t_* = \max(0,\, u_0 - 1)$ is the instant when the regime switches from saturated to unsaturated. But the exact solution at the continuous level is not our center of interest.

Discretized system. Our center of interest is what happens when (6.2) is numerically solved (6.1) by the Euler backward scheme

$$\frac{u^{n+1} - u^n}{\Delta t} = -\min((u^{n+1})^2,\, 1), \tag{6.3}$$

where $\Delta t > 0$ is the time-step. Scheme (6.3) results in the nonlinear scalar equation

$$u^{n+1} + \Delta t \min((u^{n+1})^2,\, 1) - u^n = 0 \tag{6.4}$$

in the unknown u^{n+1}. Changing the notations from u^{n+1} to u, u^n to $u_\flat$, Δt to τ and introducing the auxiliary variable $q = \min(u^2,\, 1)$, we can cast (6.4) under the equivalent system

$$\boxed{u + \tau q - u_\flat = 0,} \tag{6.5a}$$

$$\boxed{\min(1 - q,\, u^2 - q) = 0,} \tag{6.5b}$$

in the two unknowns $(u, q) \in \mathbb{R}^2$, given the parameters $(u_\flat, \tau) \in \mathbb{R}_+^* \times \mathbb{R}_+^*$.

In Theorem 4.11, we proved that

$$(\overline{u},\,\overline{q}) = \begin{cases} (u_\flat - \tau,\,1) & \text{if } \tau \leqslant u_\flat - 1 \\ \left(\dfrac{2u_\flat}{1+\sqrt{1+4\tau u_\flat}},\,\dfrac{4u_\flat^2}{(1+\sqrt{1+4\tau u_\flat})^2}\right) & \text{otherwise.} \end{cases} \tag{6.6}$$

is a solution of (6.5), called *reference solution*. This solution is unique if $\tau < u_\flat + 1$. If $\tau \geqslant u_\flat + 1$, there appear two spurious solutions.

Regularity of zeros. Let $X = (u, q)$, $\Lambda(X) = u + \tau q - u_\flat$, $G(X) = 1 - q$, $H(X) = u^2 - q$. We wish to known whether or not the solution $\bar{X} = (\overline{u},\,\overline{q})$ given by (6.6) gives rise to a regular zero $\bar{\mathbb{X}}$ of $\mathbb{F}$ for the NPIMP algorithm. To this end, we must check that $\det \nabla \mathbb{F}(\bar{\mathbb{X}})$ or $\det \nabla \mathbf{F}(\bar{\mathbf{X}})$ is nonzero. But by Lemma 5.4 and Lemma 4.4, we can also compute

$$\overline{\mathfrak{d}} = \begin{vmatrix} \nabla\Lambda(\bar{X}) \\ \nabla G(\bar{X}) \odot H(\bar{X}) + \nabla H(\bar{X}) \odot G(\bar{X}) \end{vmatrix} = \begin{vmatrix} 1 & \tau \\ 2(1-\overline{q})\overline{u} & -[(1-\overline{q}) + (\overline{u}^2 - \overline{q})] \end{vmatrix}.$$

Since $\overline{u} > 0$, and it follows from

$$\overline{\mathfrak{d}} = -(1 + 2\tau\overline{u})(1-\overline{q}) - (\overline{u}^2 - \overline{q})$$

that $\overline{\mathfrak{d}} \leqslant 0$. Equality holds if and only if $\overline{u}^2 = \overline{q} = 1$, that is, $\overline{u} = \overline{q} = 1$ because $\overline{u} > 0$. By inverting (6.6), we find that this is equivalent to

$$u_\flat = \tau + 1. \tag{6.7}$$

Therefore, the zeros are regular except for the singular situation (6.7).

Numerical results. We compare the NPIPM algorithm with four other methods: Newton-min, Newton-min with line search, Mehrotra predictor-corrector, and θ^1-smoothing. The stopping criterion is $\|\mathbb{F}(\mathbb{X})\| < 10^{-7}$. We set the maximum number of iterations to be 50. If the number of iterations of the algorithm exceeds this maximum number, the case will be considered as divergent. With NPIPM, the parameters for the line search are $\kappa = 0.4$ and $\varrho = 0.99$. In the last equation of the NPIPM system, we take $\eta = 10^{-6}$.

We sweep over the grid of parameters

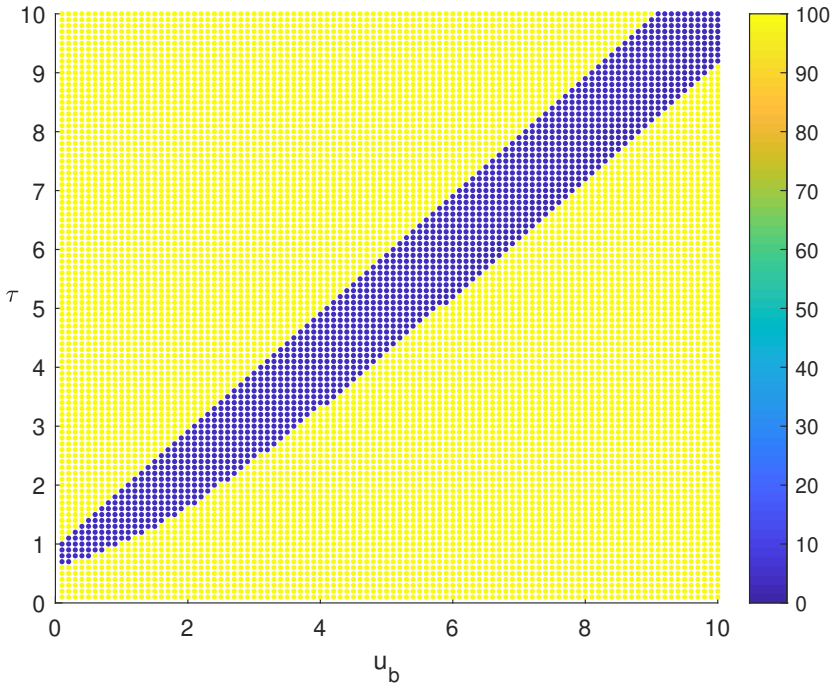
$$(u_\flat, \tau) \in \{0.1;\, 0.2;\, \ldots;\, 10\} \times \big\{0.1;\, 0.2;\, \ldots;\, 10\big\}.$$

and the set of initial points

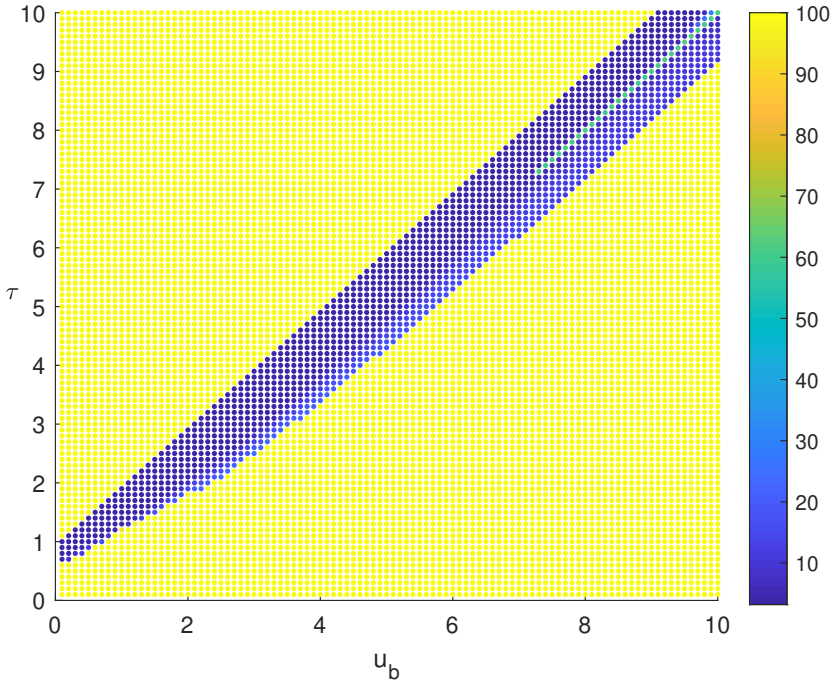
$$\mathcal{D}^0 = \big\{(u^0, q^0) \in \{0.1;\, 0.2;\, \ldots;\, 10\} \times \{0.1;\, 0.2;\, \ldots;\, 0.9\} \;\mid\; (u^0)^2 - q^0 > 0\big\}.$$

The number of initial points used for the tests is $|\mathcal{D}^0| = 843$. For each pair $(u_\flat, \tau)$, we count the number of initial points for which the method converges and then plot the percentage of success for each algorithm.

The results are displayed in Figures 6.1–6.3. It is clearly seen that Mehrotra, Theta-1 and NPIPM all give better results than Newton-min. More accurately, NPIPM and Theta-1 reach an impressive rate of 100% of initial points with convergence. Mehrotra seems to be as perfect as the other two in Figure 6.2(a), but in fact it diverges in a small region, as evidenced by the close-up in Figure 6.2(b).

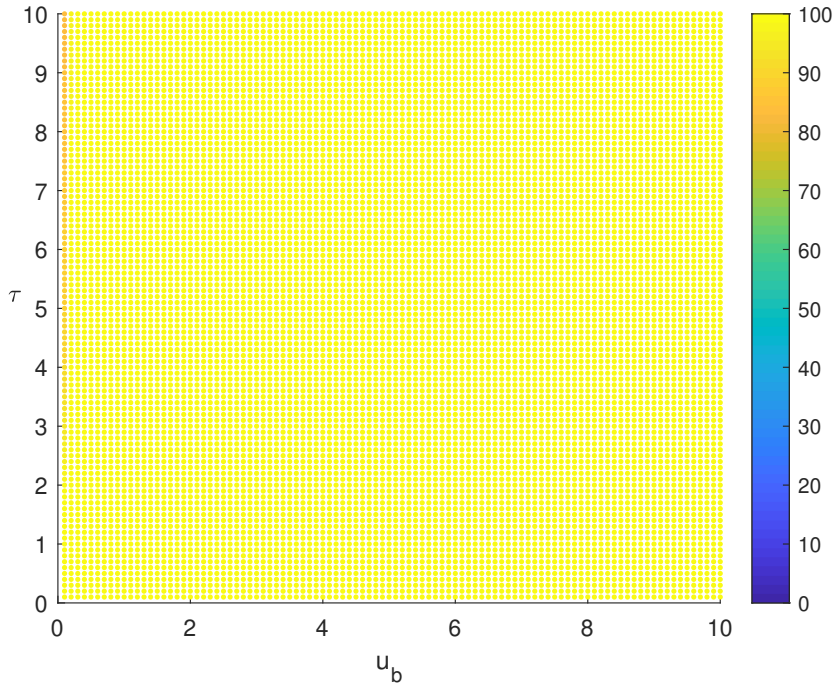


(a) Newton-min

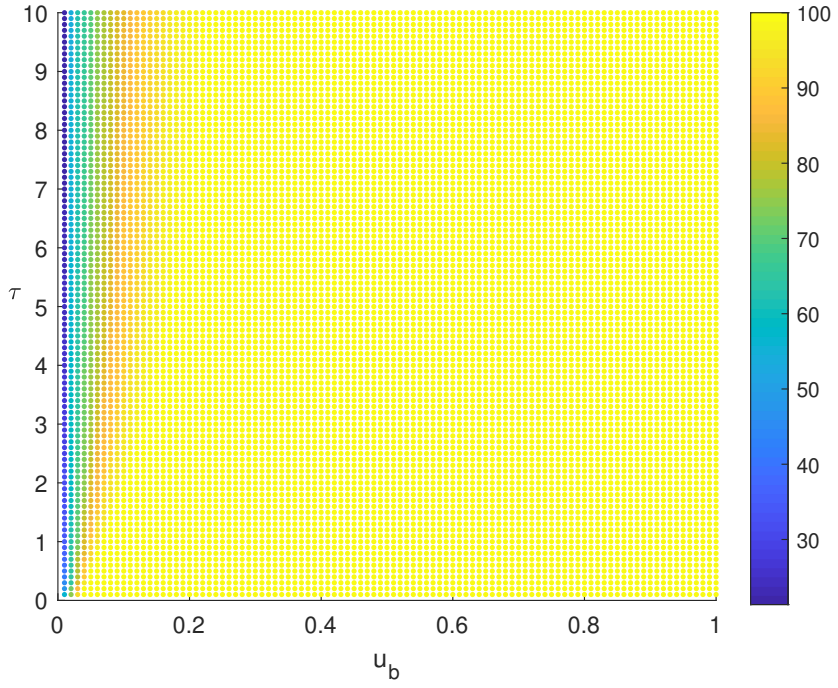


(b) Newton-min with line search

Figure 6.1: Stratigraphic model: Newton-min without and with line search.

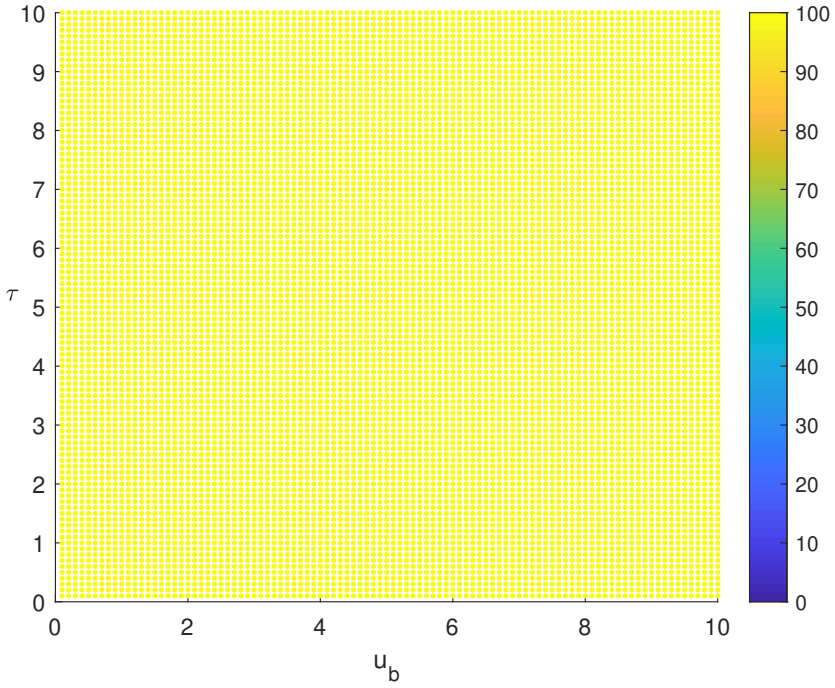


(a) Mehrotra's predictor-corrector algorithm

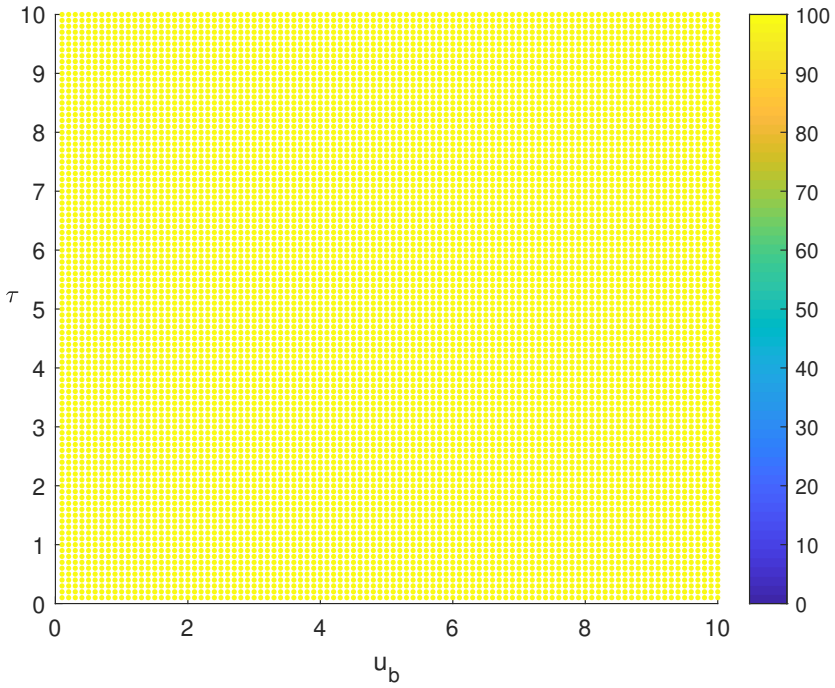


(b) Close-up of the divergence zone in Mehrotra's algorithm

Figure 6.2: Stratigraphic model: Mehrotra's algorithm.



(a) Theta 1



(b) NPIPM

Figure 6.3: Stratigraphic model: θ^1-smoothing and NPIPM.

6.1.2 Stationary binary model

After the stratigraphic model as an "appetizer," we return to thermodynamics with the two-phase binary model (2.83) of §2.4.2. For this phase equilibrium problem, we will consider four families of fugacity coefficients in the order of increasing complexity: Henry's law, Van Laar's law, Van der Waals' law and Peng-Robinson's law.

6.1.2.1 Henry's law

The gas phase is *ideal*, while the liquid phase has constant fugacity coefficents. In other words,

$$\Phi_G^{\mathrm{I}} \equiv 1, \qquad \Phi_G^{\mathrm{II}} \equiv 1, \qquad \Phi_L^{\mathrm{I}} \equiv k^{\mathrm{I}}, \qquad \Phi_L^{\mathrm{II}} \equiv k^{\mathrm{II}}. \tag{6.8}$$

To fix ideas, we assume that

$$k^{\mathrm{I}} > 1 > k^{\mathrm{II}} > 0. \tag{6.9}$$

Reference solution. Thanks to the simplicity of (6.8), we can eliminate ξ_L^{I}, ξ_L^{II} by substituting $\xi_G^{\mathrm{I}}/k^{\mathrm{I}}$, $\xi_G^{\mathrm{II}}/k^{\mathrm{II}}$ into (2.83). This leads to the three-equation system

$$\boxed{\begin{aligned} Y\xi_G^{\mathrm{I}} + (1-Y)\xi_G^{\mathrm{I}}/k^{\mathrm{I}} - c &= 0, && \text{(6.10a)}\\ \min(Y,\, 1-\xi_G^{\mathrm{I}} - \xi_G^{\mathrm{II}}) &= 0, && \text{(6.10b)}\\ \min(1-Y,\, 1-\xi_G^{\mathrm{I}}/k^{\mathrm{I}} - \xi_G^{\mathrm{II}}/k^{\mathrm{II}}) &= 0, && \text{(6.10c)} \end{aligned}}$$

in the unknowns $(Y,\, \xi_G^{\mathrm{I}},\, \xi_G^{\mathrm{II}}) \in [0,1] \times \mathbb{R}_+ \times \mathbb{R}_+$. The following Proposition provides the solution of (6.10), which we call *reference solution.*

Proposition 6.1. *For $k^{\mathrm{I}} > 1 > k^{\mathrm{II}} > 0$, the quantities*

$$K_G = \frac{k^{\mathrm{I}}(1-k^{\mathrm{II}})}{k^{\mathrm{I}} - k^{\mathrm{II}}}, \qquad K_L = \frac{1-k^{\mathrm{II}}}{k^{\mathrm{I}} - k^{\mathrm{II}}}, \tag{6.11}$$

are well-defined and satisfy $0 < K_L < K_G < 1$. Then, the solution of system (6.10) *is given by*

$$(\bar{Y},\, \bar{\xi}_G^{\mathrm{I}},\, \bar{\xi}_G^{\mathrm{II}}) = \begin{cases} (0,\, k^{\mathrm{I}}c,\, k^{\mathrm{II}}(1-c)) & \text{if } c \in [0, K_L], \\ \left(\dfrac{c-K_L}{K_G - K_L},\, K_G,\, k^{\mathrm{II}}(1-K_L)\right) & \text{if } c \in (K_L, K_G), \\ (1,\, c,\, 1-c) & \text{if } c \in [K_G, 1]. \end{cases} \tag{6.12}$$

Chứng minh. This follows from Gibbs' geometric construction described in Theorem 2.5. For $k^{\mathrm{I}} > 1 > k^{\mathrm{II}} > 0$, there is exactly one common tangent to the graphs of $g_G(\cdot)$ and $g_L(\cdot)$. A little algebra shows that the contact points are precisely $(K_L, g_L(K_L))$ and $(K_G, g_G(K_G))$, the former being on the left of the latter. The lower convex envelope $\breve{g}$ of $\min(g_G, g_L)$ coincides with $g_L(\cdot)$ over $[0, K_L]$, with the common tangent over (K_L, K_G), and with $g_G(\cdot)$ over $[K_G, 1]$. □

Regularity of zeros. As a consequence of Proposition 3.1 (strict convexity of the Gibbs functions) and the general Theorem 5.3 (a special proof for Henry's law was given §5.2.2), the reference solution $\bar{X} = (\bar{Y},\, \bar{\xi}_G^{\mathrm{I}},\, \bar{\xi}_G^{\mathrm{II}})$ gives rise to a regular zero $\bar{\mathbb{X}}$ of the NPIPM system $\mathbb{F}(\mathbb{X}) = 0$, provided that it is not a transitional or azeotropic point.

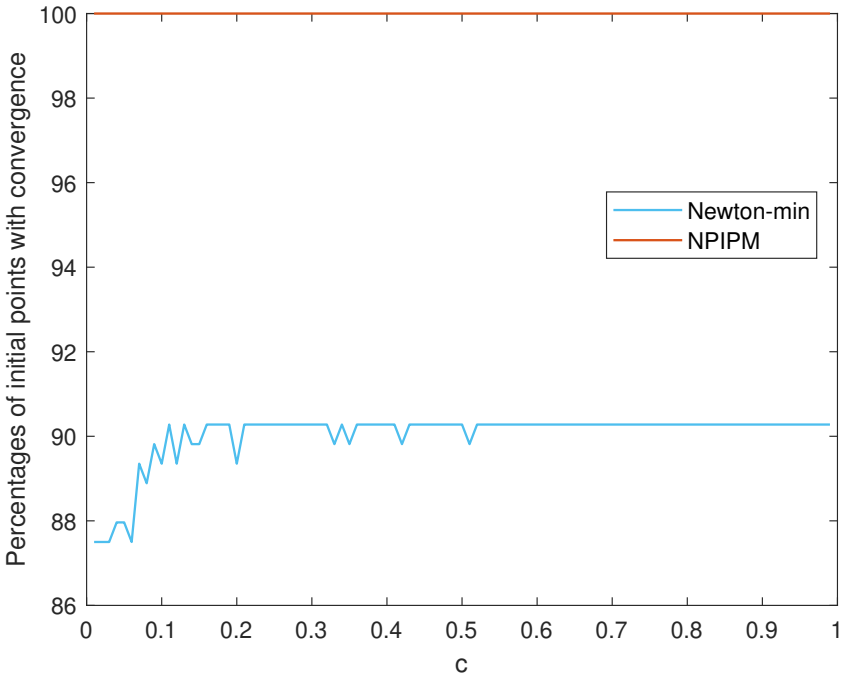


Figure 6.4: Henry's law: percentage of convergence over all initial points.

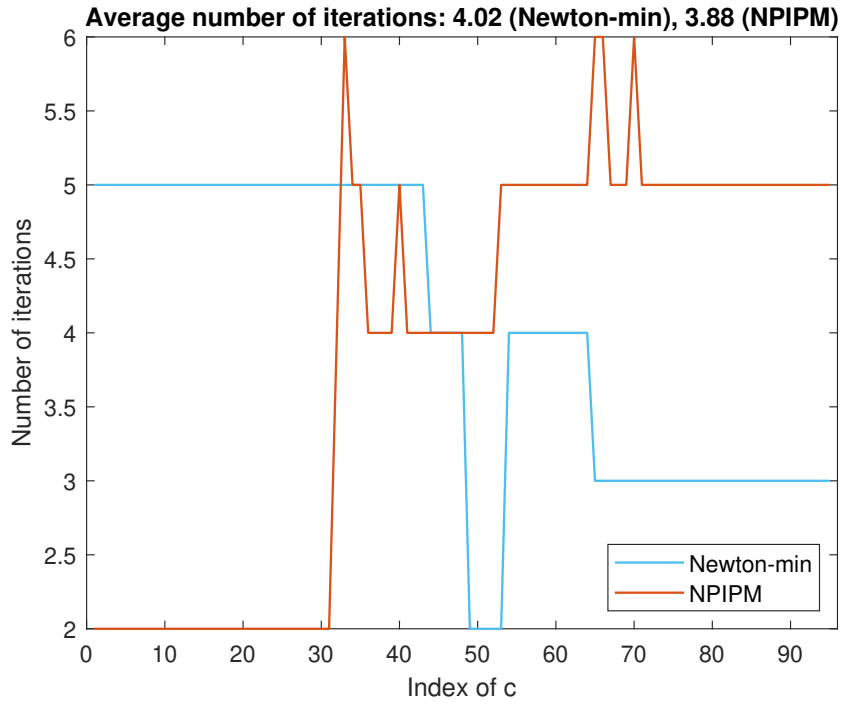
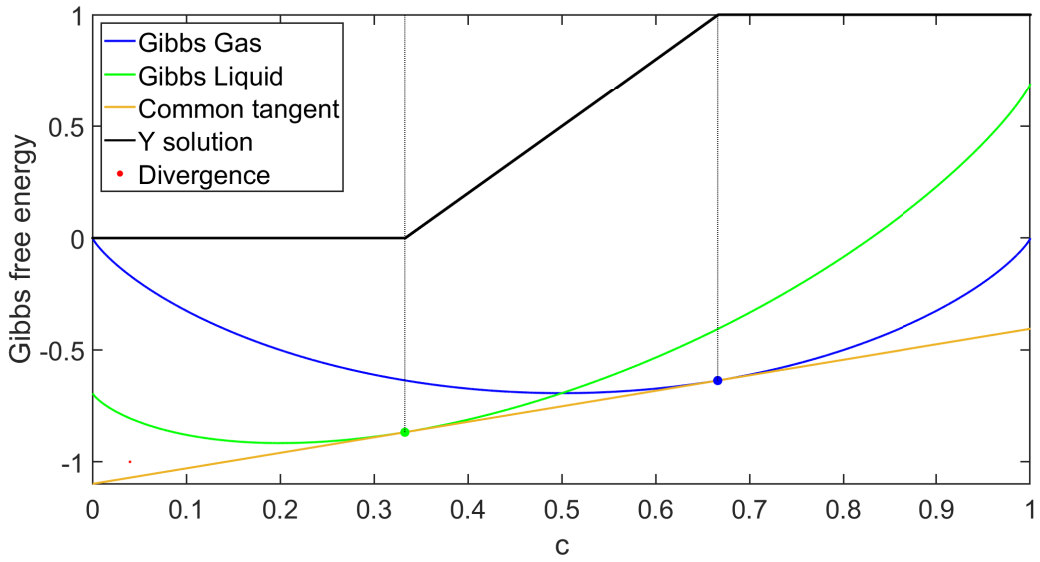
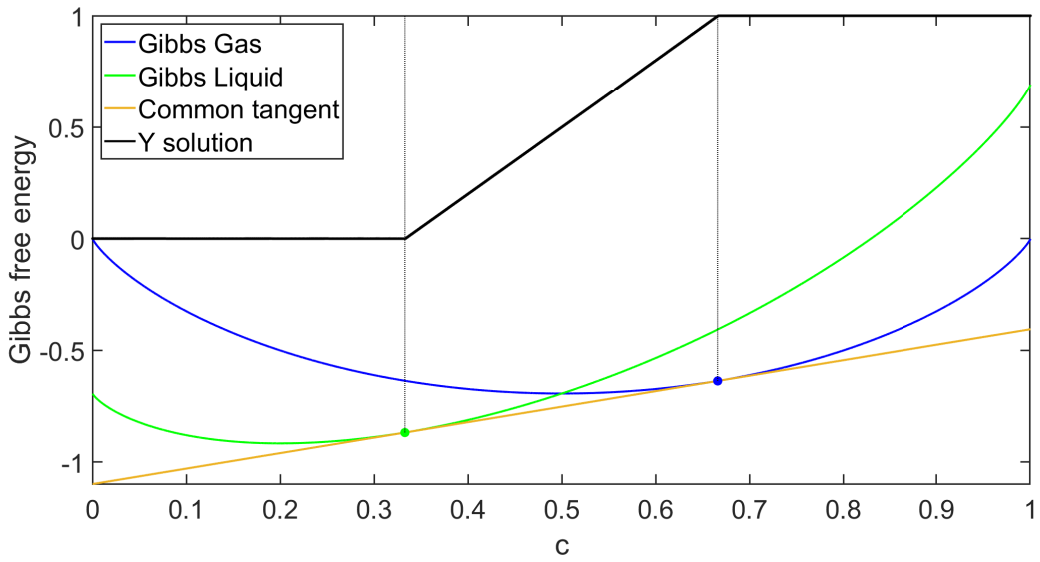


Figure 6.5: Henry's law: number of iterations with the same initial points.



(a) Newton-min



(b) NPIPM

Figure 6.6: Henry's law, tested with the same initial point.

Numerical results. We compare NPIPM with other methods as we did in stratigraphic model. We fix $k^{\mathrm{I}} = 2$, $k^{\mathrm{II}} = 0.5$. The stopping criteria is $\|\mathbb{F}(\mathbb{X})\| < 10^{-7}$. We set the maximum number of iterations to be 50. With NPIPM, the line search parameters are $\kappa = 0.4$ and $\varrho = 0.99$. In the last equation of the system, we take $\eta = 10^{-6}$.

We sweep over the grid of parameters

$$c \in \big\{0.01;\, 0.02;\, \ldots;\, 0.99\big\}.$$

and the set of initial points

$$\mathcal{D}^0 = \big\{(Y,\, \xi_G^{\mathrm{I}},\, \xi_G^{\mathrm{II}})^0 \in \mathcal{M}^3 \;\mid\; 1 - (\xi_G^{\mathrm{I}})^0 - (\xi_G^{\mathrm{II}})^0 > 0 \;\text{ and }\; 1 - (\xi_G^{\mathrm{I}})^0/k^{\mathrm{I}} - (\xi_G^{\mathrm{II}})^0/k^{\mathrm{II}} > 0\big\}.$$

where $\mathcal{M} = \{0.1;\, 0.2;\, \ldots;\, 0.9\}$. The number of initial points used for the tests is $|\mathcal{D}^0| = 216$. For each c, we count the number of initial points for which the method converges and then plot the percentage of success for each algorithm in Figure 6.4. Since the percentages for Newton-min with line search, Mehrotra and Theta-1 are **less than 10%**, we just show the results for Newton-min and NPIPM. Figure 6.4 testifies to the remarkable efficiency of NPIPM relatively to Newton-min, with 100% of convergence.

The next test takes place between NPIPM and Newton-min method. Starting from the same initial point $(Y,\, \xi_G^{\mathrm{I}},\, \xi_G^{\mathrm{II}}) = (0.2,\, 0.6,\, 0.3)$, we run the two algorithms for all values of $c \in \{0.0001,\, 0.0002,\, ...,\, 0.9999\}$. In each panel of Figure 6.6, we also plot the Gibbs energy functions g_G and g_L. The common tangent between the graphs of g_G and g_L is represented by the orange line. The tangency points represent transitional solutions, between a single-phase regime and a two-phase regime. The black line is the value of $\bar{Y}$ for each c when the algorithm converges. If the algorithm diverges at a value c, we assign the value -1. We observe that NPIPM (lower panel) converges with all values of c tested. Newton-min (upper panel) is also not bad either, with just one case of divergence.

The last test with Henry's law is the number of iterations if the algorithm converges. We still use the same parameters for the convergence test. However, in Figure 6.5, we display the number of iterations versus the "index" of c. This index is the rank of c within the subset of $\big\{0.01;\, 0.02;\, \ldots;\, 0.99\big\}$ containing those values of c for which convergence occurs for both methods.

6.1.2.2 Van Laar's law

In the two-phase binary model (2.83), we now assign Henry's law to the gas phase, that is,

$$\Phi_G^{\mathrm{I}} \equiv k^{\mathrm{I}}, \qquad \Phi_G^{\mathrm{II}} \equiv k^{\mathrm{II}}, \tag{6.13}$$

while the liquid phase obeys Van Laar's law (3.12), namely,

$$\ln \Phi_L^{\mathrm{I}}(x) = A_{12}\left[\frac{A_{21}(1-x)}{A_{12}x + A_{21}(1-x)}\right]^2, \tag{6.14a}$$

$$\ln \Phi_L^{\mathrm{II}}(x) = A_{21}\left[\frac{A_{12}x}{A_{12}x + A_{21}(1-x)}\right]^2. \tag{6.14b}$$

Regularity of zeros. By Proposition 3.1, the Gibbs function g_G of the gas phase satisfies Hypotheses 2.2 for $k^{\mathrm{I}},\, k^{\mathrm{II}} > 0$. By Proposition 3.3, if the pair (A_{12}, A_{21}) belongs to the "good" region (3.14), the Gibbs function g_L of the liquid phase satisfies Hypotheses 2.2. Then, owing to

Theorem 2.5 ensures existence and uniqueness of a solution for those c at which azeotropy does not occur. Thanks to Theorem 5.3, this solution gives rise to a regular zero $\bar{\mathbb{X}}$ of the NPIPM sytem $\mathbb{F}(\mathbb{X}) = 0$, provided that it is not a transition point.

Numerical results. We fix $k^{\mathrm{I}} = 2$, $k^{\mathrm{II}} = 0.5$ and select the pair

$$A_{12} = -0.8643, \qquad A_{21} = -0.5899,$$

which corresponds to acetone (species I) and chloroform (species II). It can be readily checked that this pair belongs indeed to the strict convexity region (3.14) of Van Laar's law.

The first test is between NPIPM and Newton-min method. We choose the same initial point $(Y, \xi_G^{\mathrm{I}}, \xi_G^{\mathrm{II}}, \xi_L^{\mathrm{I}}, \xi_L^{\mathrm{II}})^0 = (0.1, 0.3, 0.6, 0.2, 0.1)$ and run both algorithms for each value of $c \in \{0.0001, 0.0002, \ldots, 0.9999\}$. The stopping criteria is $\|\mathbb{F}(\mathbb{X})\| < 10^{-7}$. We set the maximum number of iterations to be 50. With NPIPM, we choose parameters for line search step: $\kappa = 0.4$ and $\varrho = 0.99$. In the last equation of the system, $\eta = 10^{-6}$. In each panel of Figure 6.7, we plot the Gibbs energy functions g_G and g_L. The black line is the value of $\bar{Y}$ for each c when the algorithm converges. If the algorithm diverges at a value c, we assign the flag value -1.

After this first test, we display in Figure 6.8 the number of iterations at convergence corresponding to the two methods. The last test with Van Laar's law involves many initial points. We compare the NPIPM algorithm and Newton-min algorithm for several values of $c \in \{0.01; 0.02; \ldots; 0.99\}$. For each value of c, we sweep over the set of initial points

$$\mathcal{D}^0 = \big\{(Y, \xi_G^{\mathrm{I}}, \xi_G^{\mathrm{II}}, \xi_L^{\mathrm{I}}, \xi_L^{\mathrm{II}})^0 \in \mathcal{M}^5 \;\mid\; 1 - (\xi_G^{\mathrm{I}})^0 - (\xi_G^{\mathrm{II}})^0 > 0 \text{ and } 1 - (\xi_L^{\mathrm{I}})^0 - (\xi_L^{\mathrm{II}})^0 > 0\big\},$$

where $\mathcal{M} = \{0.1; 0.2; \ldots; 0.9\}$. The number of initial points used is $|\mathcal{D}^0| = 11664$. We count the number of initial points for which the method converges and then plot the percentage on the figure for each algorithm. Again, Figure 6.9 demonstrates the outstanding efficiency of NPIPM, with a 100% rate of convergence. Nevertheless, it needs slightly more iterations than Newton-min when the latter converges.

6.1.2.3 Van der Waals' law

We consider the two-phase binary model (2.83) with Van der Waals' fugacity coefficients (3.37), namely,

$$\ln \Phi_\alpha^i(x) = \frac{B^i}{B(x)}\left[Z_\alpha(x) - 1\right] - \ln\left[Z_\alpha(x) - B(x)\right] + \left[\frac{B^i}{B(x)} - \frac{2A^i(x)}{A(x)}\right]\frac{A(x)}{Z_\alpha(x)}, \tag{6.15}$$

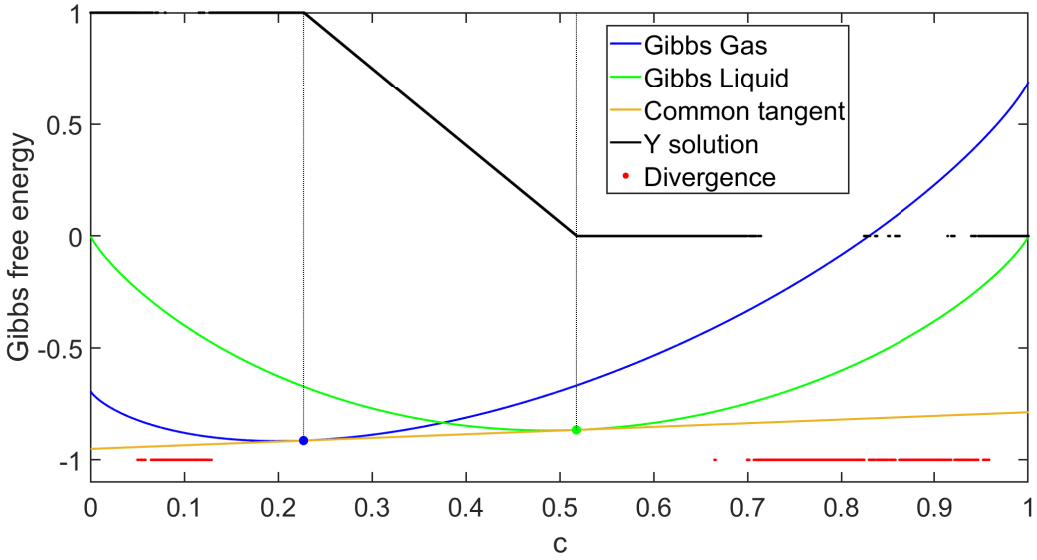
for $i \in \{\mathrm{I}, \mathrm{II}\}$, $\alpha \in \{G, L\}$, $x \in [0, 1]$, where $Z_\alpha(x)$ is a real root of the cubic equation (3.33), that is,

$$Z^3(x) - [B(x) + 1]Z^2(x) + A(x)Z(x) - A(x)B(x) = 0. \tag{6.16}$$

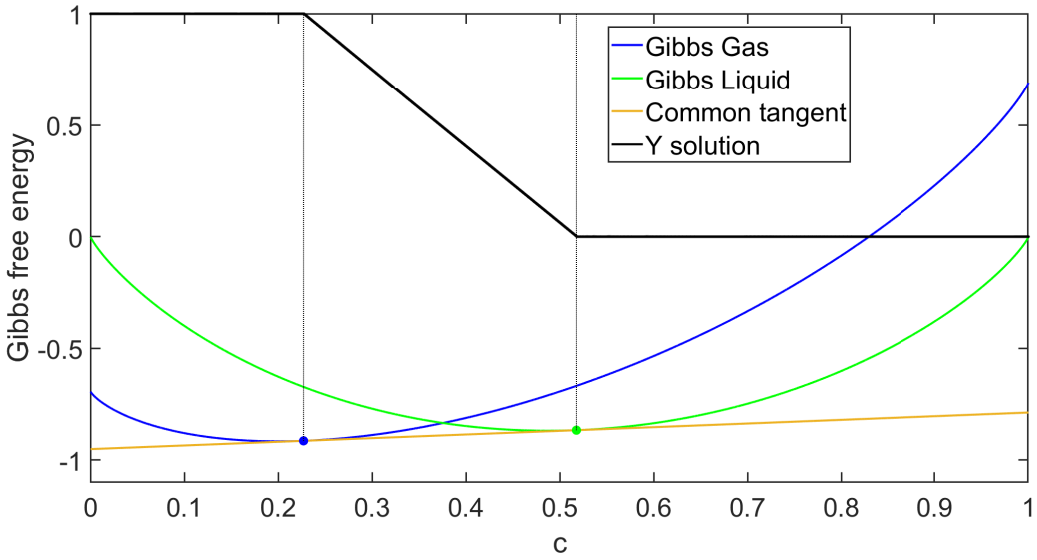
The mixing rules (3.31b)–(3.32) have been used, i.e.,

$$A(x) = \left(x\sqrt{A^{\mathrm{I}}} + (1 - x)\sqrt{A^{\mathrm{II}}}\right)^2, \tag{6.17a}$$

$$B(x) = xB^{\mathrm{I}} + (1 - x)B^{\mathrm{II}}. \tag{6.17b}$$



(a) Newton-min



(b) NPIPM

Figure 6.7: Van Laar's law, tested with the same initial point.

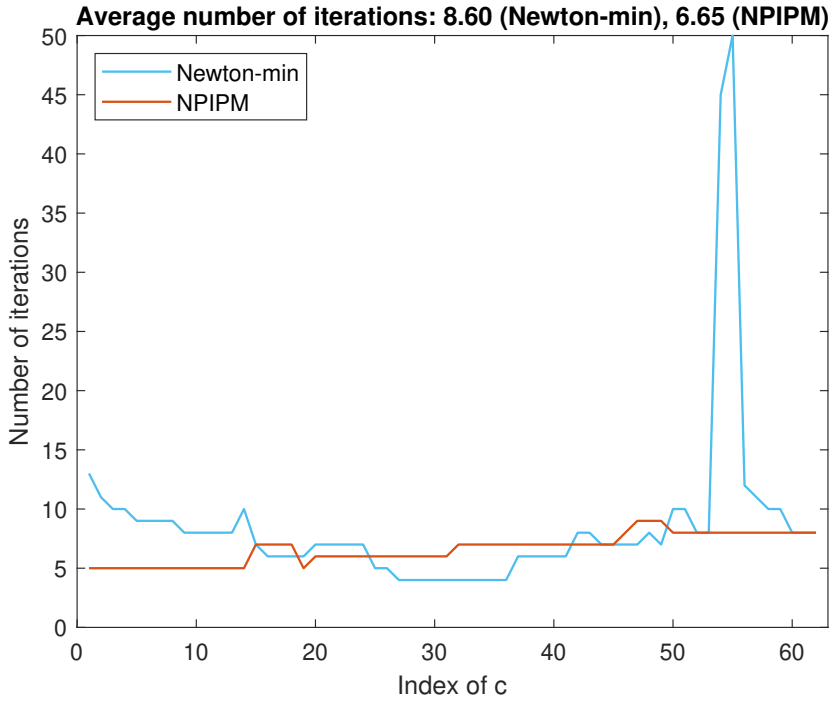


Figure 6.8: Van Laar's law: number of iterations with the same initial points.

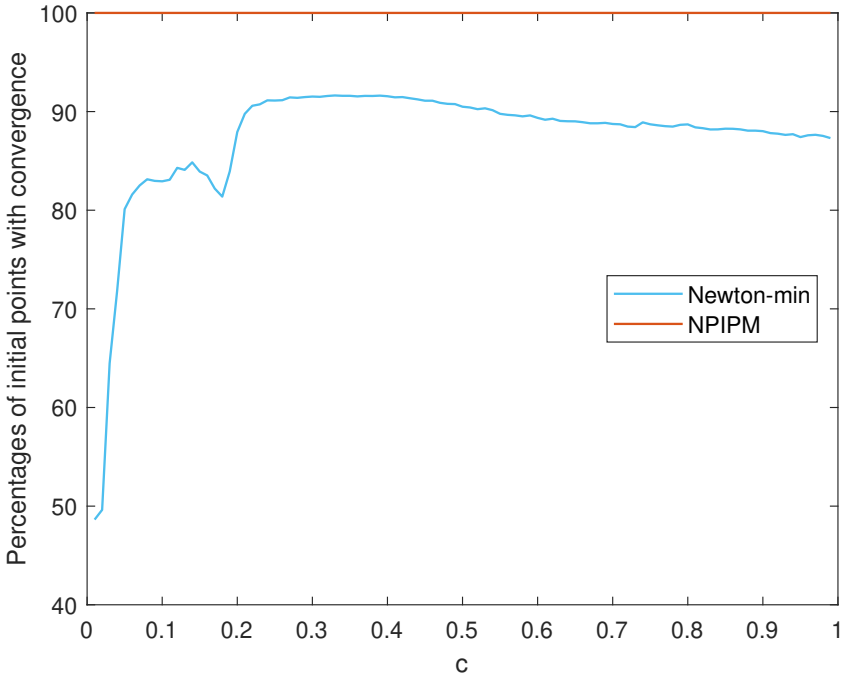


Figure 6.9: Van Laar's law: percentage of convergence over all initial points.

Existence, uniqueness and regularity? Due to the complexity of Van der Waals' law, it is difficult to tell anything about the strict convexity of the Gibbs functions g_G and g_L, the excess parts of which are given by (3.34). In the binary case, these Gibbs functions can be numerically plotted as functions of x. Extensive numerical investigations by Le Hénaff [79] have confirmed that, in general, we do not have strict convexity for two arbitrary pairs $(A^{\mathrm{I}}, B^{\mathrm{I}})$ and $(A^{\mathrm{II}}, B^{\mathrm{II}})$ in the subcritical region of the (A, B)-plane, although there are some "choices" for which strict convexity holds. As a consequence, there is nothing we can predict about the existence, uniqueness and regularity of a solution.

Need for domain extension. In general, g_G and g_L are not even defined on the whole interval $(0, 1)$, as explained at length in §3.3.1 and as corroborated by numerical studies. Here, we wish to illustrate this issue numerically. Let us choose

$$(A^{\mathrm{I}},\, B^{\mathrm{I}}) = (0.33,\, 0.0955), \qquad (A^{\mathrm{II}},\, B^{\mathrm{II}}) = (0.35,\, 0.08)$$

as depicted in Figure 6.10, so as to ensure that each of the Gibbs functions g_G and g_L is "visually" strictly convex on its domain of definition, which is not $(0, 1)$.

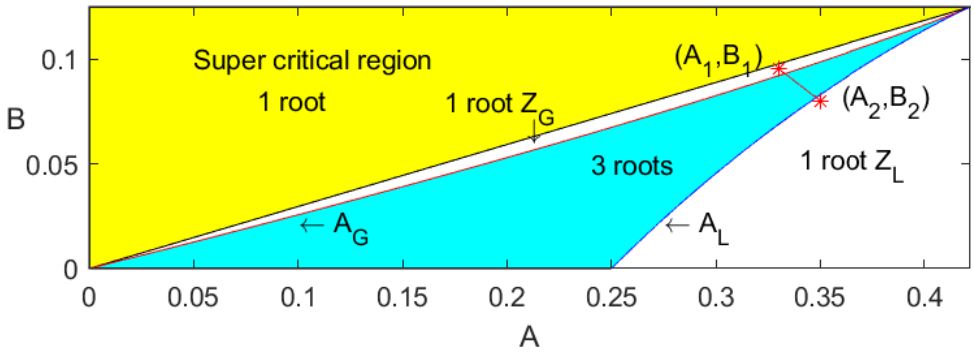


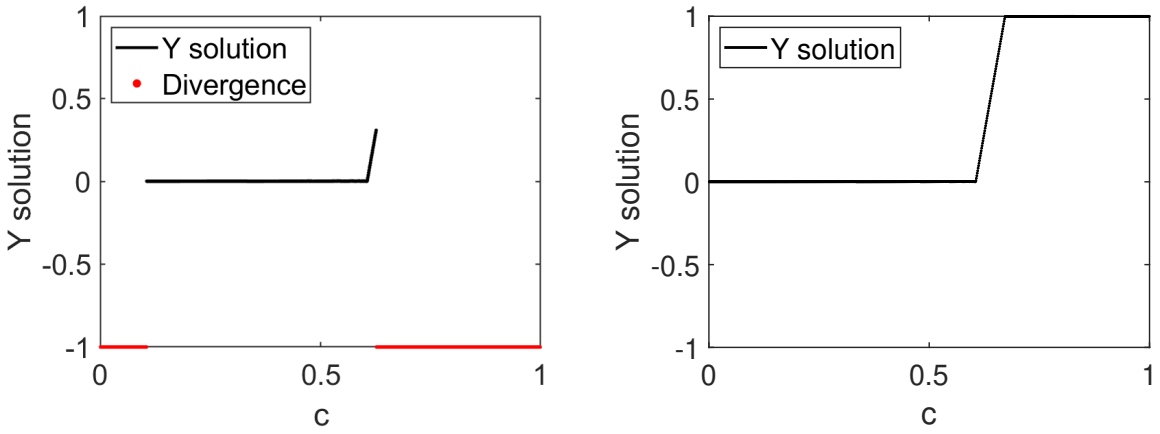
Figure 6.10: Van der Waals' law: $(A^{\mathrm{I}},\, B^{\mathrm{I}}) = (0.33,\, 0.0955)$ and $(A^{\mathrm{II}},\, B^{\mathrm{II}}) = (0.35,\, 0.08)$.

We run NPIPM without and with the extension procedures described in §3.3.2–§3.3.3 using the same initial point $(Y,\, \xi_G^{\mathrm{I}},\, \xi_G^{\mathrm{II}},\, \xi_L^{\mathrm{I}},\, \xi_L^{\mathrm{II}})^0 = (0.8,\, 0.4,\, 0.2,\, 0.2,\, 0.6)$ and sweeping over all $c \in \{0.001,\, 0.002,\, \ldots,\, 0.999\}$. Figure 6.11 represents $\bar{Y}$ at convergence, with the flag value -1 when NPIPM diverges or "crashes." Without extension, NPIPM abruptly stops when the cubic equation has a unique real root. With the direct extension of §3.3.2, the problem is fully avoided.

Numerical results. As we did with Van Laar's law, we first compare NPIPM and Newton-min method with the same initial point $(Y,\, \xi_G^{\mathrm{I}},\, \xi_G^{\mathrm{II}},\, \xi_L^{\mathrm{I}},\, \xi_L^{\mathrm{II}})^0 = (0.8,\, 0.4,\, 0.2,\, 0.2,\, 0.6)$. The results are provided in Figure 6.12 for the direct extension and in Figure 6.15 for the indirect extension. With NPIPM, the line search parameters are $\kappa = 0.4$ and $\varrho = 0.99$. In the last equation of the system, we take $\eta = 10^{-6}$. We run Newton-min and NPIPM for all values of $c \in \{0.001,\, 0.002,\, \ldots,\, 0.999\}$. The stopping criteria is $\|\mathbb{F}(\mathbb{X})\| < 10^{-7}$. We set the maximum number of iterations to be 50.

Next, we analyze the number of iterations at convergence. The corresponding results are given in Figure 6.13 for the direct extension and in Figure 6.16 for the indirect extension. The last test with Van der Waals' law aims at measuring the percentage of convergence over many initial points. To this end, we sweep over the set of parameter $c \in \{0.01;\, 0.02;\, \ldots;\, 0.99\}$. For each value of c, the set of initial points is

$$\mathcal{D}^0 = \big\{(Y,\, \xi_G^{\mathrm{I}},\, \xi_G^{\mathrm{II}},\, \xi_L^{\mathrm{I}},\, \xi_L^{\mathrm{II}})^0 \in \mathcal{M}^5 \;\mid\; 1 - (\xi_G^{\mathrm{I}})^0 - (\xi_G^{\mathrm{II}})^0 > 0 \;\text{ and }\; 1 - (\xi_L^{\mathrm{I}})^0 - (\xi_L^{\mathrm{II}})^0 > 0\big\}.$$



(a) NPIPM without extension (b) NPIPM with direct extension

Figure 6.11: Van der Waals' law without extension and with extension for the same initial point.

where $\mathcal{M} = \{0.2; 0.4; 0.6; 0.8\}$. The number of initial points used for the tests is $|\mathcal{D}^0| = 144$. We count the number of initial points for which the method converges and then display the percentage of success, in Figure 6.14 for the direct extension and in Figure 6.17 for the indirect extension.

For the direct extension, the width parameter is $\omega = 0.05$. For the indirect extension, the width parameter is $\varepsilon = 0.03$. One more time, we observe that the new algorithm converges for all initial points, despite the high complexity of Van der Waals' law. This behavior is promising.

6.1.2.4 Peng-Robinson's law

We consider the two-phase binary model (2.83) with Peng-Robinson's fugacity coefficients (3.53), namely,

$$\begin{aligned}\ln \Phi_\alpha^i(x) = \; & \frac{B^i}{B(x)}\left[Z_\alpha(x) - 1\right] - \ln\left[Z_\alpha(x) - B(x)\right] \\ & + \left[\frac{B^i}{B(x)} - \frac{2A^i(x)}{A(x)}\right]\frac{A(x)}{2\sqrt{2}B_(x)}\ln\left[\frac{Z_\alpha(x) + (1+\sqrt{2})B(x)}{Z_\alpha(x) - (\sqrt{2}-1)B(x)}\right],\end{aligned} \tag{6.18}$$

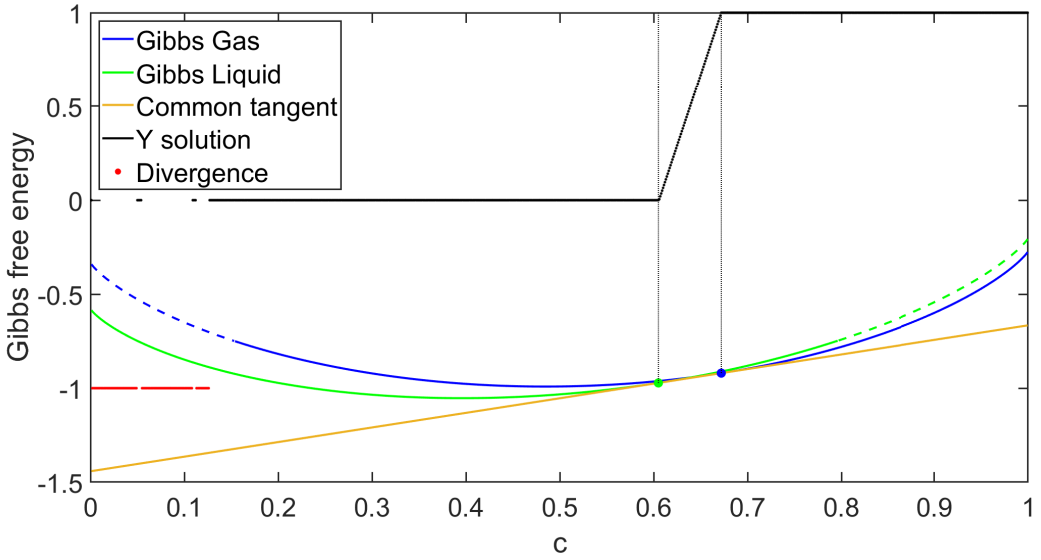
for $i \in \{\mathrm{I}, \mathrm{II}\}$, $\alpha \in \{G, L\}$, $x \in [0, 1]$, where $Z_\alpha(x)$ is a real root of the cubic equation (3.51), that is,

$$\begin{aligned}& Z^3(x) + (B(x) - 1)Z^2(x) \\ & \quad + [A(x) - 2B(x) - 3B^2(x)]Z(x) + [B^2(x) + B^3(x) - A(x)B(x)] = 0.\end{aligned} \tag{6.19}$$

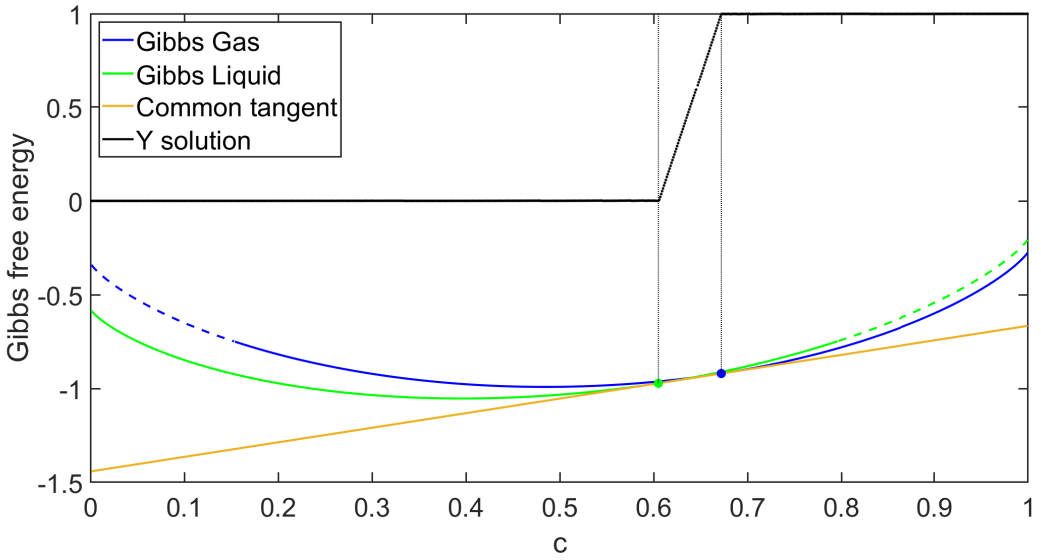
The mixing rules (3.31b)–(3.32) have been used, i.e.,

$$A(x) = \left(x\sqrt{A^{\mathrm{I}}} + (1-x)\sqrt{A^{\mathrm{II}}}\right)^2, \tag{6.20a}$$

$$B(x) = xB^{\mathrm{I}} + (1-x)B^{\mathrm{II}}. \tag{6.20b}$$



(a) Newton-min



(b) NPIPM

Figure 6.12: Van der Waals' law with direct extension, tested with the same initial point.

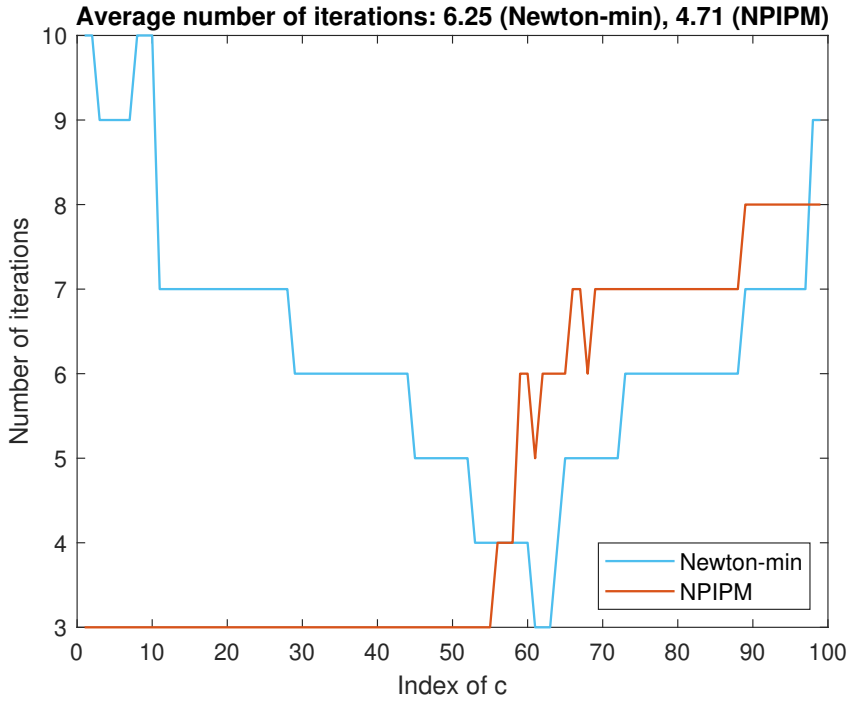


Figure 6.13: Van der Waals' law with direct extension: number of iterations with the same initial point.

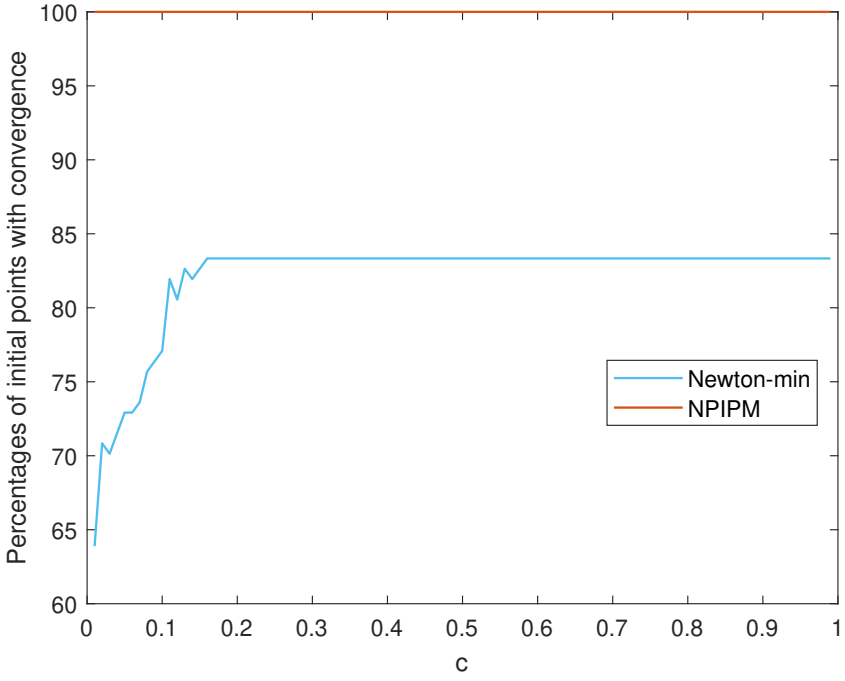
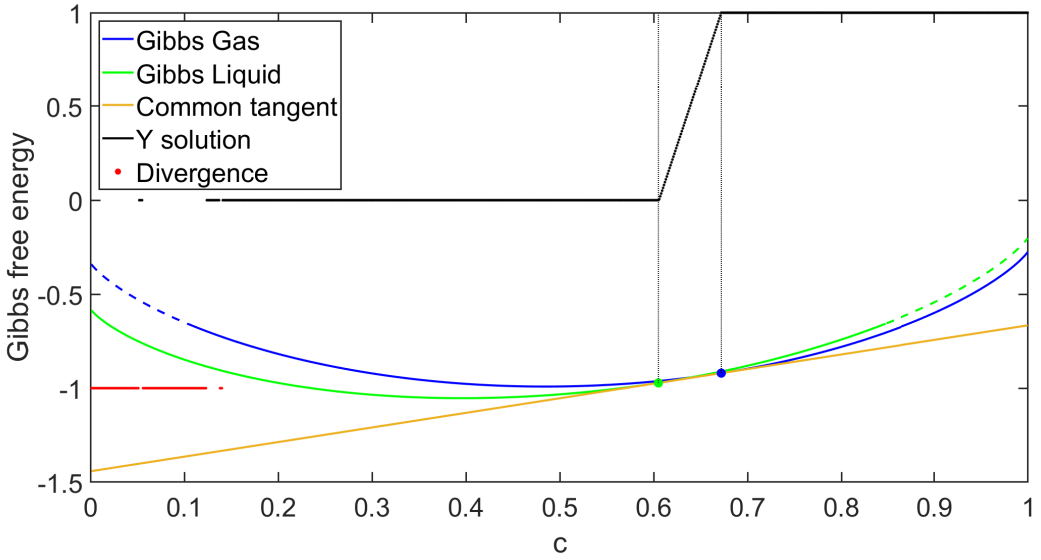
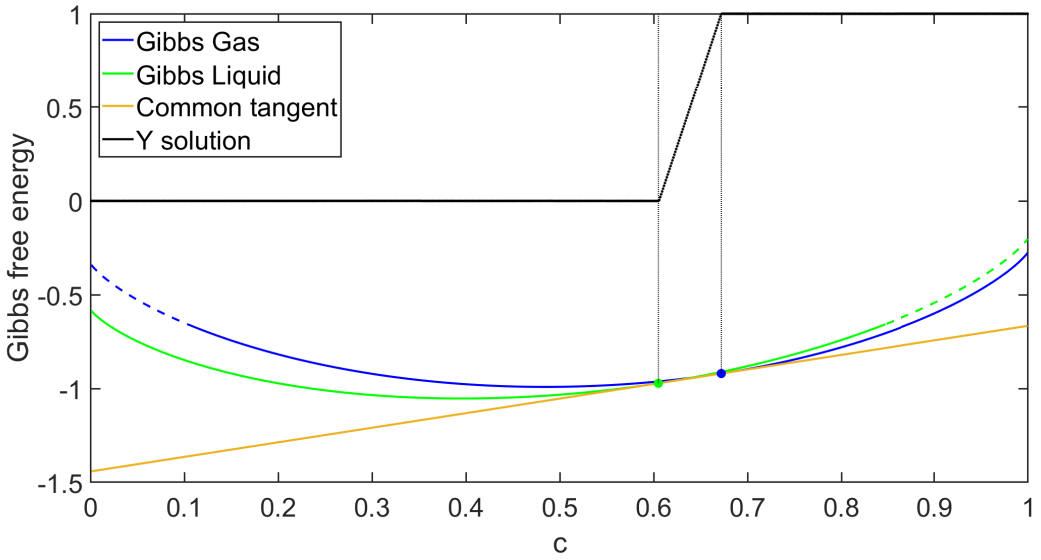


Figure 6.14: Van der Waals law with direct extension: percentage of convergence over all initial points.



(a) Newton-min



(b) NPIPM

Figure 6.15: Van der Waals' law with indirect extension, tested with the same initial point.

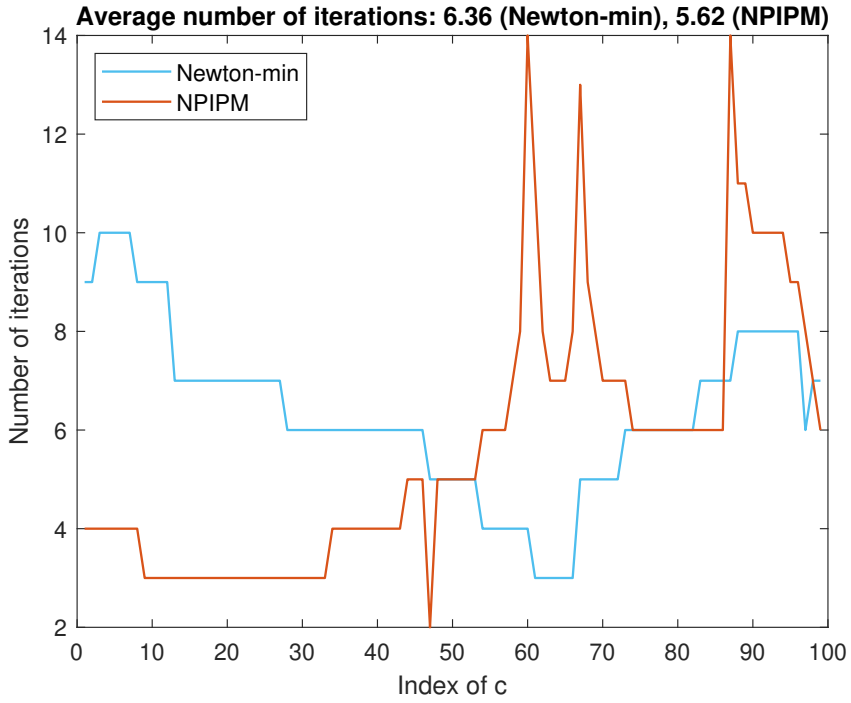


Figure 6.16: Van der Waals' law with indirect extension: number of iterations with the same initial point.

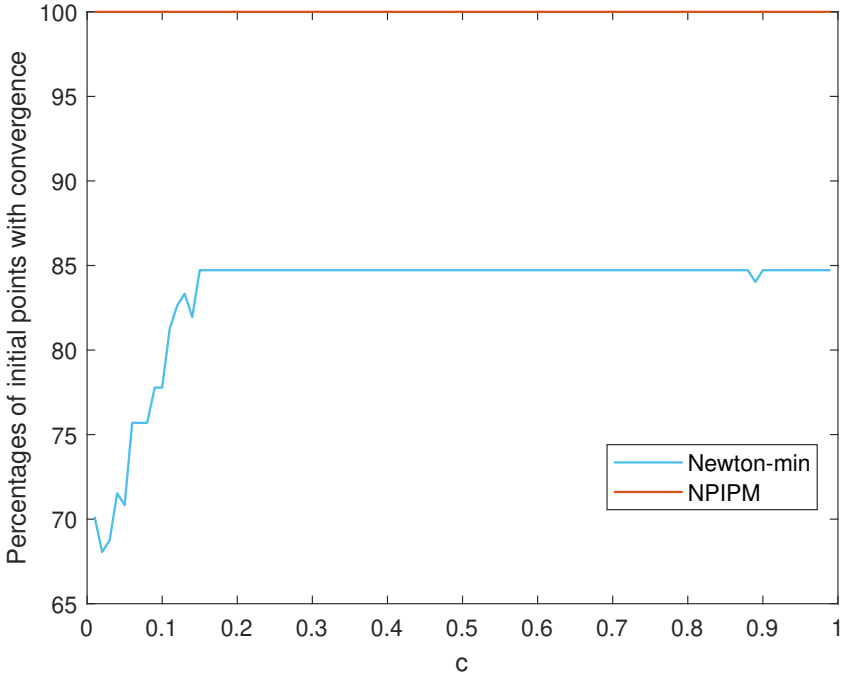


Figure 6.17: Van der Waals law with indirect extension: percentage of convergence over all initial points.

Existence, uniqueness and regularity? Due to the complexity of Peng-Robinson's law, it is difficult to tell anything about the strict convexity of the Gibbs functions g_G and g_L, the excess parts of which are given by (3.52). In the binary case, the Gibbs functions can be numerically plotted as functions of x, and we can try to select the pairs $(A^{\mathrm{I}}, B^{\mathrm{I}})$ and $(A^{\mathrm{II}}, B^{\mathrm{II}})$ in such a way that g_G and g_L are "visually" strictly convex on their respective domains of definition. An example of two such pairs is

$$(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.322,\, 0.053), \qquad (A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.33,\, 0.03)$$

as depicted in Figure 6.18. We carry out numerical simulations with these values.

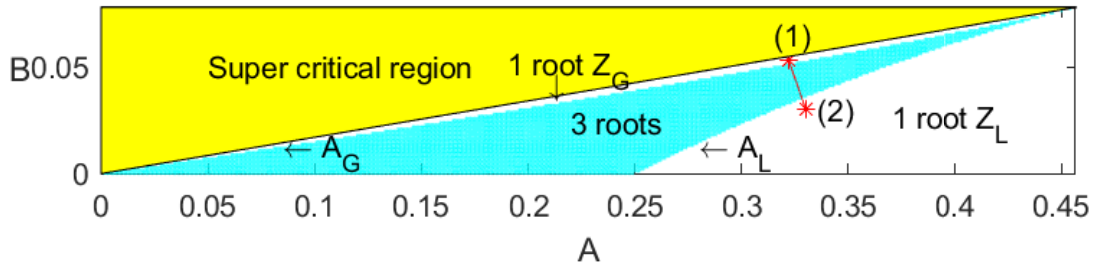


Figure 6.18: Peng-Robinson's law: $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.322, 0.053)$ and $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.33, 0.03)$

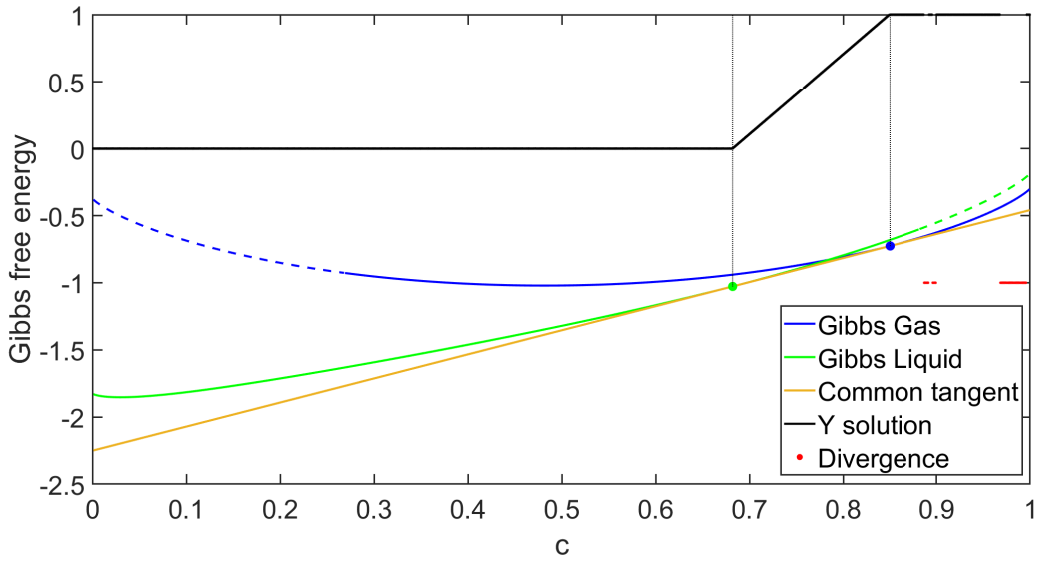
Numerical results. As we did with Van der Waals' law, we first compare NPIPM and Newton-min method with the same initial point $(Y, \xi_G^{\mathrm{I}}, \xi_G^{\mathrm{II}}, \xi_L^{\mathrm{I}}, \xi_L^{\mathrm{II}})^0 = (0.2,\, 0.2,\, 0.4,\, 0.4,\, 0.2)$. The results are provided in Figure 6.19 for the direct extension and in Figure 6.22 for the indirect extension. With NPIPM, the line search parameters are $\kappa = 0.4$ and $\varrho = 0.99$. In the last equation of the system, we take $\eta = 10^{-6}$. We run Newton-min and NPIPM for all values of $c \in \{0.001,\, 0.002,\, \ldots,\, 0.999\}$. The stopping criteria is $\|\mathbb{F}(\mathbb{X})\| < 10^{-7}$. We set the maximum number of iterations to be 50.

Next, we analyze the number of iterations at convergence. The corresponding results are given in Figure 6.20 for the direct extension and in Figure 6.23 for the indirect extension. The last test with Van der Waals' law aims at measuring the percentage of convergence over many initial points. To this end, we sweep over the set of parameter $c \in \{0.01;\, 0.02;\, \ldots;\, 0.99\}$. For each value of c, the set of initial points is

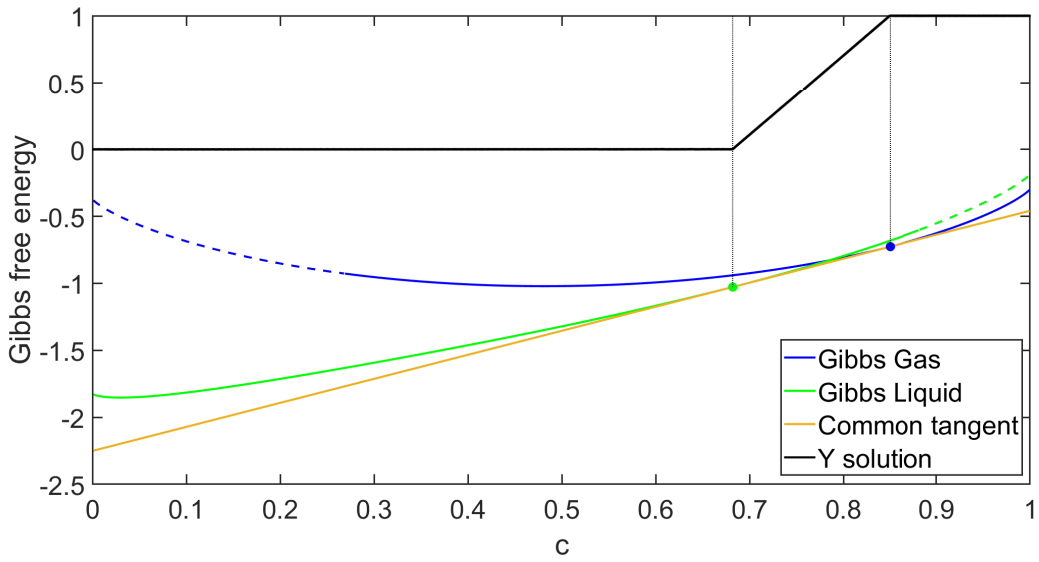
$$\mathcal{D}^0 = \left\{(Y, \xi_G^{\mathrm{I}}, \xi_G^{\mathrm{II}}, \xi_L^{\mathrm{I}}, \xi_L^{\mathrm{II}})^0 \in \mathcal{M}^5 \;\middle|\; 1 - (\xi_G^{\mathrm{I}})^0 - (\xi_G^{\mathrm{II}})^0 > 0 \text{ and } 1 - (\xi_L^{\mathrm{I}})^0 - (\xi_L^{\mathrm{II}})^0 > 0\right\}.$$

where $\mathcal{M} = \{0.2;\, 0.4;\, 0.6;\, 0.8\}$. The number of initial points used for the tests is $|\mathcal{D}^0| = 144$. We count the number of initial points for which the method converges and then display the percentage of success, in Figure 6.21 for the direct extension and in Figure 6.24 for the indirect extension.

For the direct extension, the width parameter is $\omega = 0.05$. For the indirect extension, the width parameter is $\varepsilon = 0.03$. One more time, we observe that the new algorithm converges for all initial points, despite the high complexity of Peng-Robinson's law.



(a) Newton-min



(b) NPIPM

Figure 6.19: Peng-Robinson's law with direct extension: one initial point.

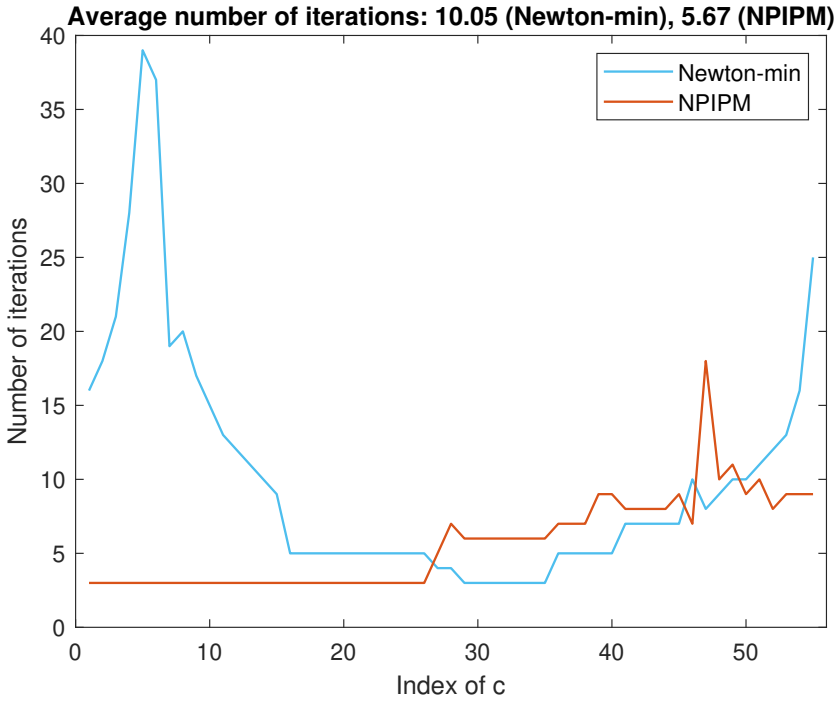


Figure 6.20: Peng-Robinson's law with direct extension: number of iterations with the same initial point.

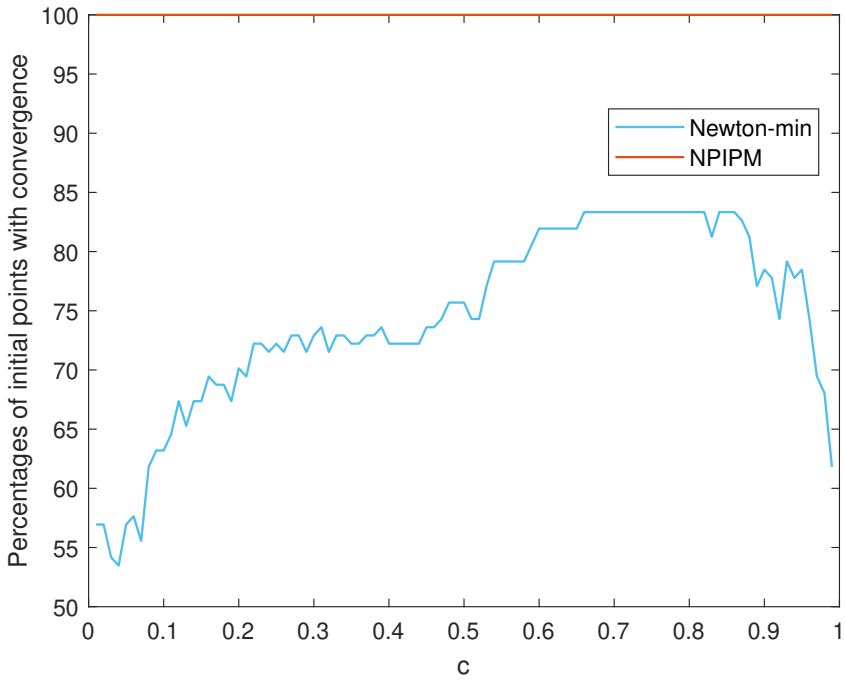
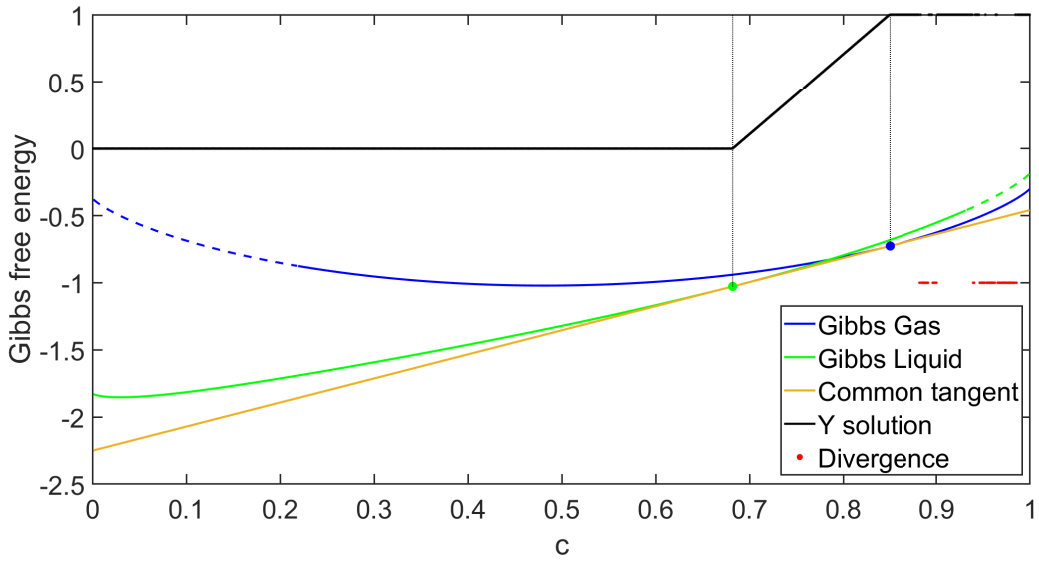
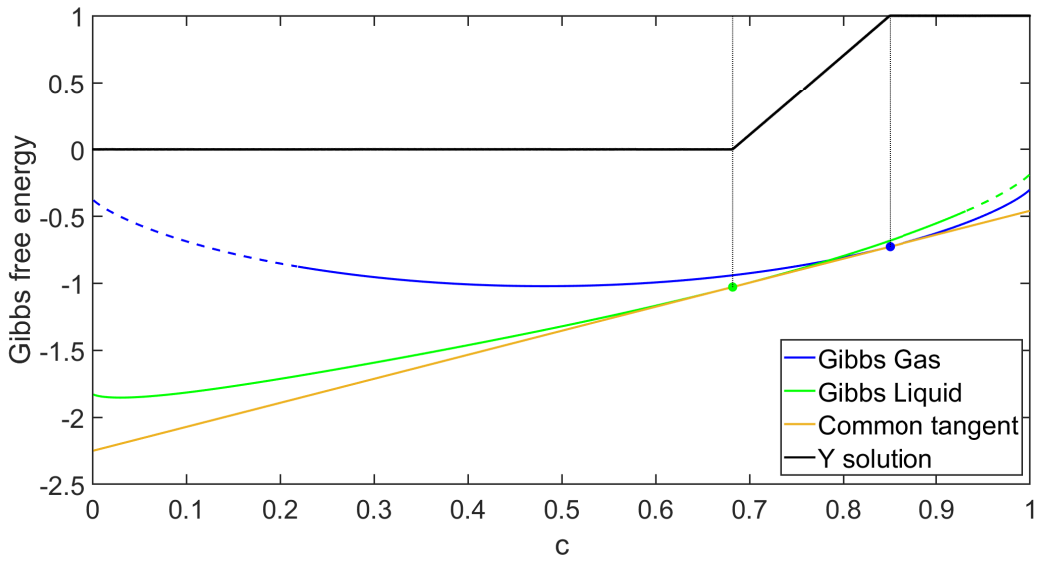


Figure 6.21: Peng-Robinson's law with direct extension: percentage of convergence over all initial points.



(a) Newton-min



(b) NPIPM

Figure 6.22: Peng-Robinson's law with indirect extension: one initial point.

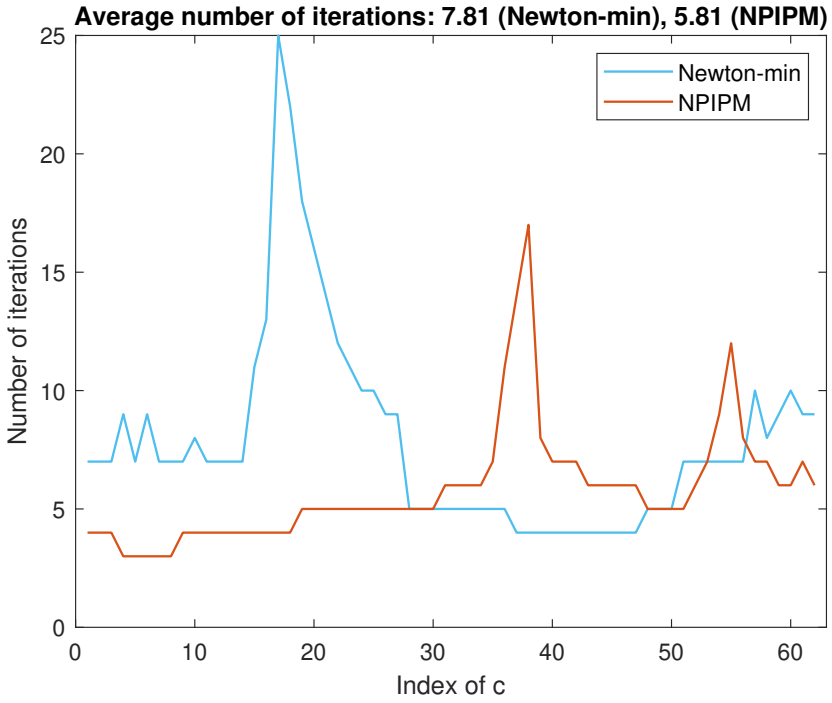


Figure 6.23: Peng-Robinson's law with indirect extension: number of iterations with the same initial points.

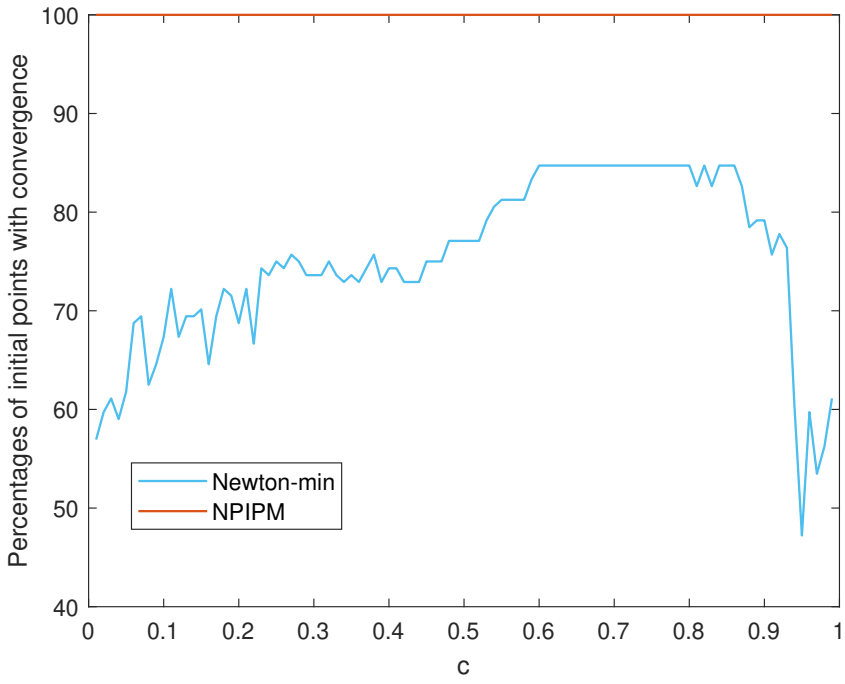


Figure 6.24: Peng-Robinson's law with indirect extension: percentage of convergence over all initial points.

6.1.3 Stationary ternary model

Let us now go to the more complex case of a two-phase ternary model. With K = 3, system (2.77) reads

$$Y\xi_G^{\mathrm{I}} + (1-Y)\xi_L^{\mathrm{I}} - c^{\mathrm{I}} = 0, \tag{6.21a}$$
$$Y\xi_G^{\mathrm{II}} + (1-Y)\xi_L^{\mathrm{II}} - c^{\mathrm{II}} = 0, \tag{6.21b}$$
$$\xi_G^{\mathrm{I}}\,\Phi_G^{\mathrm{I}}(x_G^{\mathrm{I}}, x_G^{\mathrm{II}}) - \xi_L^{\mathrm{I}}\,\Phi_L^{\mathrm{I}}(x_L^{\mathrm{I}}, x_L^{\mathrm{II}}) = 0, \tag{6.21c}$$
$$\xi_G^{\mathrm{II}}\,\Phi_G^{\mathrm{II}}(x_G^{\mathrm{I}}, x_G^{\mathrm{II}}) - \xi_L^{\mathrm{II}}\,\Phi_L^{\mathrm{II}}(x_L^{\mathrm{I}}, x_L^{\mathrm{II}}) = 0, \tag{6.21d}$$
$$\xi_G^{\mathrm{III}}\,\Phi_G^{\mathrm{III}}(x_G^{\mathrm{I}}, x_G^{\mathrm{II}}) - \xi_L^{\mathrm{III}}\,\Phi_L^{\mathrm{III}}(x_L^{\mathrm{I}}, x_L^{\mathrm{II}}) = 0, \tag{6.21e}$$
$$\min(Y,\, 1-\xi_G^{\mathrm{I}} - \xi_G^{\mathrm{II}} - \xi_G^{\mathrm{III}}) = 0, \tag{6.21f}$$
$$\min(1-Y,\, 1-\xi_L^{\mathrm{I}} - \xi_L^{\mathrm{II}} - \xi_L^{\mathrm{III}}) = 0. \tag{6.21g}$$

The ternary phase equilibrium problem (6.21) will be considered with three families of fugacity coefficients in the order of increasing complexity: Henry's law, Van der Waals' law and Peng-Robinson's law.

6.1.3.1 Henry's law

The gas phase is *ideal*, while the liquid phase has constant fugacity coefficents. In other words,

$$\Phi_G^{\mathrm{I}} \equiv 1, \qquad \Phi_G^{\mathrm{II}} \equiv 1, \qquad \Phi_G^{\mathrm{III}} \equiv 1, \tag{6.22a}$$
$$\Phi_L^{\mathrm{I}} \equiv k^{\mathrm{I}}, \qquad \Phi_L^{\mathrm{II}} \equiv k^{\mathrm{II}}, \qquad \Phi_L^{\mathrm{III}} \equiv k^{\mathrm{III}}. \tag{6.22b}$$

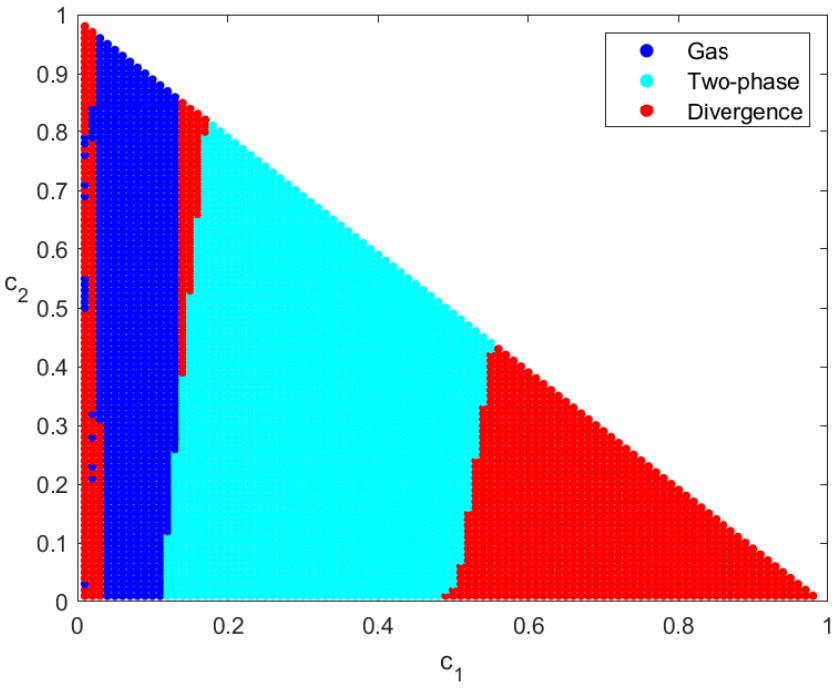
Thanks to the simplicity of (6.22), we can eliminate ξ_L^{I}, ξ_L^{II}, ξ_L^{III} by substituting $\xi_G^{\mathrm{I}}/k^{\mathrm{I}}$, $\xi_G^{\mathrm{II}}/k^{\mathrm{II}}$, $\xi_G^{\mathrm{III}}/k^{\mathrm{III}}$ into (6.21). This leads to the four-equation system

$$\boxed{\begin{aligned} Y\xi_G^{\mathrm{I}} + (1-Y)\xi_G^{\mathrm{I}}/k^{\mathrm{I}} - c^{\mathrm{I}} &= 0, && (6.23a)\\ Y\xi_G^{\mathrm{II}} + (1-Y)\xi_G^{\mathrm{II}}/k^{\mathrm{II}} - c^{\mathrm{II}} &= 0, && (6.23b)\\ \min(Y,\, 1-\xi_G^{\mathrm{I}} - \xi_G^{\mathrm{II}} - \xi_G^{\mathrm{III}}) &= 0, && (6.23c)\\ \min(1-Y,\, 1-\xi_G^{\mathrm{I}}/k^{\mathrm{I}} - \xi_G^{\mathrm{II}}/k^{\mathrm{II}} - \xi_G^{\mathrm{III}}/k^{\mathrm{III}}) &= 0, && (6.23d) \end{aligned}}$$

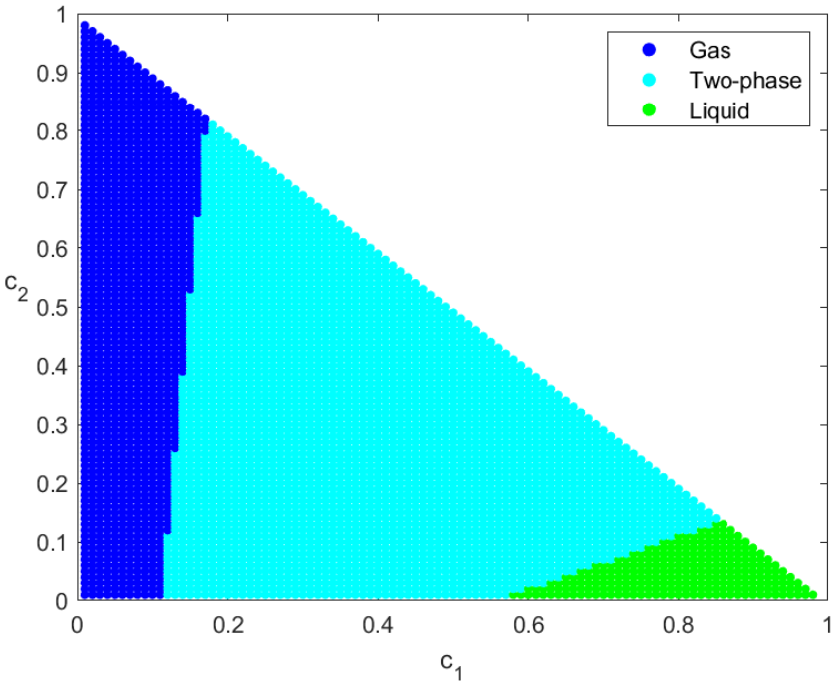
in the unknowns $(Y,\, \xi_G^{\mathrm{I}},\, \xi_G^{\mathrm{II}},\, ,\xi_G^{\mathrm{III}}) \in [0,1] \times \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}_+$.

Regularity of zeros. As a consequence of Proposition 3.1 (strict convexity of the Gibbs functions) and the general Theorem 5.3 (a special proof for Henry's law was given §5.2.2), the reference solution $\bar{X} = (\bar{Y},\, \bar{\xi}_G^{\mathrm{I}},\, \bar{\xi}_G^{\mathrm{II}},\, \bar{\xi}_G^{\mathrm{III}})$ gives rise to a regular zero $\bar{\mathbb{X}}$ of the NPIPM system $\mathbb{F}(\mathbb{X}) = 0$, provided that it is not a transitional or azeotropic point.

Numerical results. We compare NPIPM with other methods as we did in stratigraphic model. We fix $k^{\mathrm{I}} = 0.2$, $k^{\mathrm{II}} = 6$, $k^{\mathrm{III}} = 2$. The stopping criteria is $\|\mathbb{F}(\mathbb{X})\| < 10^{-12}$. We set the maximum number of iterations to be 50. With NPIPM, the line search parameters are $\kappa = 0.4$ and $\varrho = 0.99$. In the last equation of the system, we take $\eta = 10^{-6}$.



(a) Newton-min



(b) NPIPM

Figure 6.25: Henry's law: one initial point.

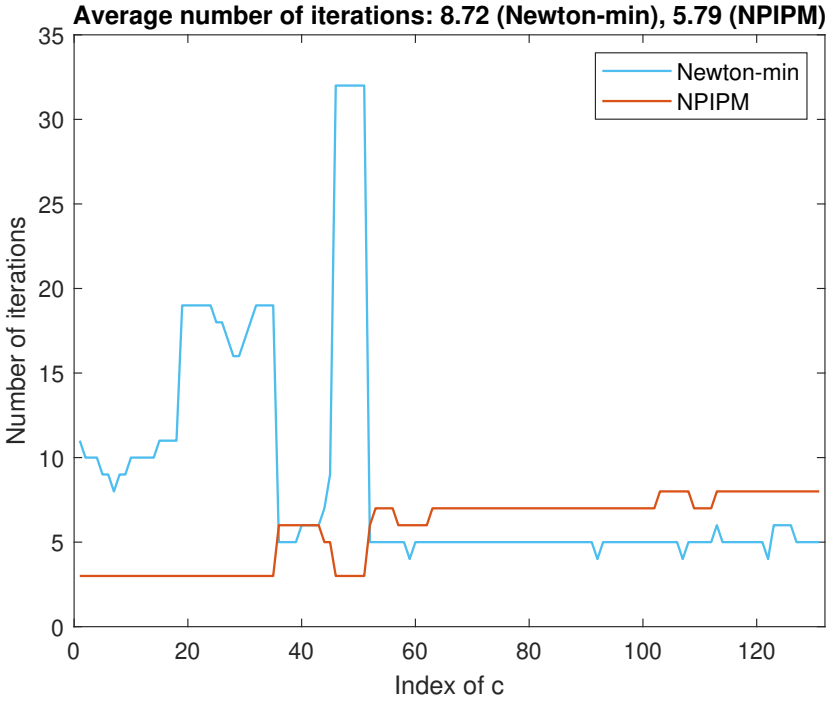


Figure 6.26: Henry's law: number of iterations with the same initial point.

The first test takes place between NPIPM and Newton-min method. Starting from the same initial point $(Y, \xi_G^{\mathrm{I}}, \xi_G^{\mathrm{II}}, \xi_G^{\mathrm{III}}) = (0.9, 0.1, 0.7, 0.1)$, we run the two algorithms for all values of $\boldsymbol{c} = (c^{\mathrm{I}}, c^{\mathrm{II}})$. In each panel of Figure 6.25, we plot the phase regime for each parameter

$$\boldsymbol{c} \in \mathcal{C} = \left\{ (c^{\mathrm{I}}, c^{\mathrm{II}}) \in \mathcal{P}^2 \;\middle|\; c^{\mathrm{I}} + c^{\mathrm{II}} < 1 \right\},$$

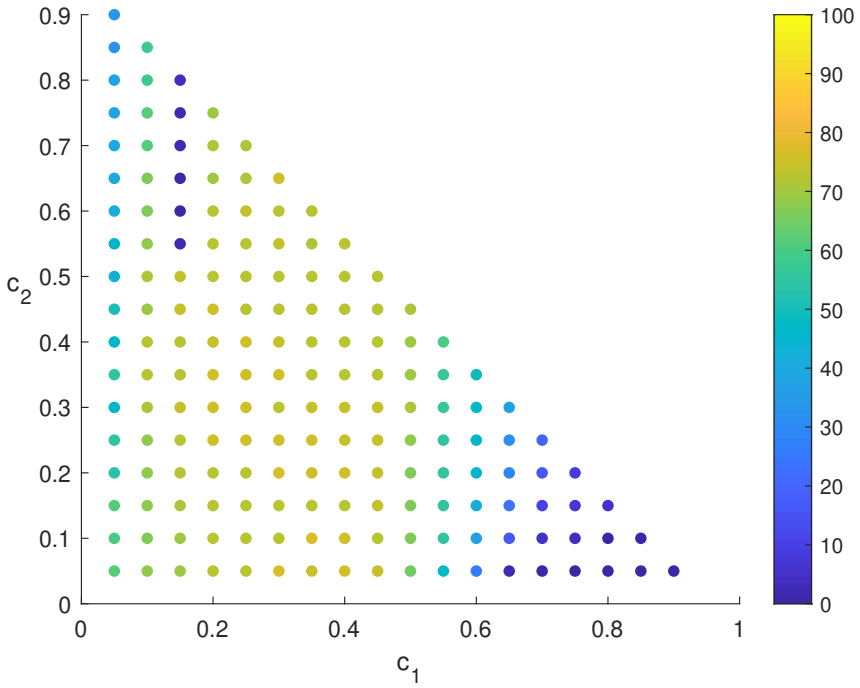
where $\mathcal{P} = \{0.01; 0.02; \ldots; 0.99\}$, when the algorithm converges. We assign the blue color to the gas single-phase regime, the cyan color to the two-phase, the green color to the liquid single-phase regime, and the red color to the case of divergence. NPIPM (lower panel) converges with all values of $\boldsymbol{c}$ tested, while Newton-min (upper panel) exhibits many cases of divergence.

The next test with Henry's law is the number of iterations if the algorithm converges. We still use the same parameters for the convergence test. However, in Figure 6.26, we display the number of iterations instead of values of $\bar{Y}$.

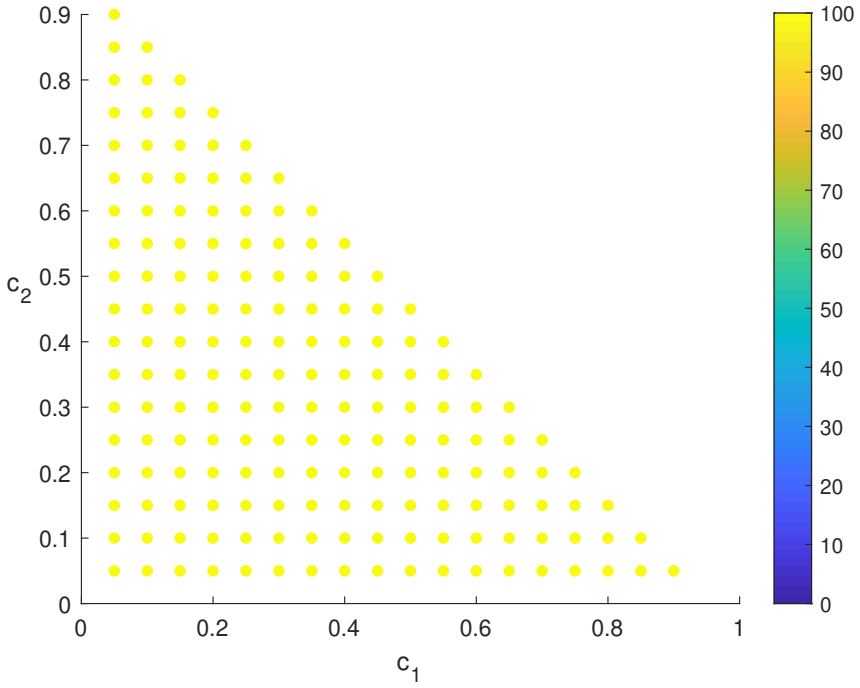
In the last test, we sweep over the grid of parameters $\boldsymbol{c} \in \mathcal{C}$ and the set of initial points

$$\begin{aligned}\mathcal{D}^0 = \big\{ (Y, \xi_G^{\mathrm{I}}, \xi_G^{\mathrm{II}}, \xi_G^{\mathrm{III}})^0 \in \mathcal{M}^4 \;\big|\; & 1 - (\xi_G^{\mathrm{I}})^0 - (\xi_G^{\mathrm{II}})^0 - (\xi_G^{\mathrm{III}})^0 > 0 \text{ and} \\ & 1 - (\xi_G^{\mathrm{I}})^0/k^{\mathrm{I}} - (\xi_G^{\mathrm{II}})^0/k^{\mathrm{II}} - (\xi_G^{\mathrm{III}})^0/k^{\mathrm{III}} > 0 \big\},\end{aligned}$$

where $\mathcal{M} = \{0.1; 0.2; \ldots; 0.9\}$. The number of initial points used for the tests is $|\mathcal{D}^0| = 252$. For each $\boldsymbol{c}$, we count the number of initial points for which the method converges and then plot the percentage of success for each algorithm in Figure 6.27. Figure 6.27 testifies to the remarkable efficiency of NPIPM relatively to Newton-min, with 100% of convergence.



(a) Newton-min



(b) NPIPM

Figure 6.27: Henry's law: percentage of convergence over all initial points.

6.1.3.2 Van der Waals' law

We consider the two-phase ternary model (6.21) with Van der Waals' fugacity coefficients (3.35), namely,

$$\begin{aligned}\ln \Phi_\alpha^i(\boldsymbol{x}) = \; & \frac{B(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} B(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B(\boldsymbol{x})} \left[Z_\alpha(\boldsymbol{x}) - 1\right] - \ln\left[Z_\alpha(\boldsymbol{x}) - B(\boldsymbol{x})\right] \\ & + \left[\frac{B(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} B(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{B(\boldsymbol{x})} - \frac{2A(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} A(\boldsymbol{x}) \cdot (\boldsymbol{\delta}^i - \boldsymbol{x})}{A(\boldsymbol{x})}\right] \frac{A(\boldsymbol{x})}{Z_\alpha(\boldsymbol{x})}, \end{aligned} \tag{6.24}$$

for $i \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}\}$, $\alpha \in \{G, L\}$, $\boldsymbol{x} \in \left\{(x^{\mathrm{I}}, x^{\mathrm{II}}) \in [0,1]^2 \;\middle|\; x^{\mathrm{I}} + x^{\mathrm{II}} \leqslant 1\right\}$, where $Z_\alpha(\boldsymbol{x})$ is a real root of the cubic equation (3.33), that is,

$$Z^3(\boldsymbol{x}) - [B(\boldsymbol{x}) + 1]Z^2(\boldsymbol{x}) + A(\boldsymbol{x})Z(\boldsymbol{x}) - A(\boldsymbol{x})B(\boldsymbol{x}) = 0. \tag{6.25}$$

The mixing rules (3.31b)–(3.32) have been used, i.e.,

$$A(\boldsymbol{x}) = \left(x^{\mathrm{I}}\sqrt{A^{\mathrm{I}}} + x^{\mathrm{II}}\sqrt{A^{\mathrm{II}}} + (1 - x^{\mathrm{I}} - x^{\mathrm{II}})\sqrt{A^{\mathrm{III}}}\right)^2, \tag{6.26a}$$

$$B(\boldsymbol{x}) = x^{\mathrm{I}}B^{\mathrm{I}} + x^{\mathrm{II}}B^{\mathrm{II}} + (1 - x^{\mathrm{I}} - x^{\mathrm{II}})B^{\mathrm{III}}. \tag{6.26b}$$

Existence, uniqueness and regularity? Due to the complexity of Van der Waals' law, it is difficult to tell anything about the strict convexity of the Gibbs functions g_G and g_L, the excess parts of which are given by (3.34). Thus, there is nothing we can predict about the existence, uniqueness and regularity of a solution. In the ternary case, the Gibbs functions can be numerically plotted as functions of $\boldsymbol{x} = (x^{\mathrm{I}}, x^{\mathrm{II}})$, and we can try to select the pairs $(A^{\mathrm{I}}, B^{\mathrm{I}})$, $(A^{\mathrm{II}}, B^{\mathrm{II}})$ and $(A^{\mathrm{III}}, B^{\mathrm{III}})$ in such a way that g_G and g_L are "visually" strictly convex on their respective domains of definition. An example of three such pairs is

$$(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.33, 0.0955), \qquad (A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.35, 0.08), \qquad (A^{\mathrm{III}}, B^{\mathrm{III}}) = (0.355, 0.0953)$$

as depicted in Figure 6.28. We apply the indirect extension procedure of §3.3.3 with $\varepsilon = 0.01$.

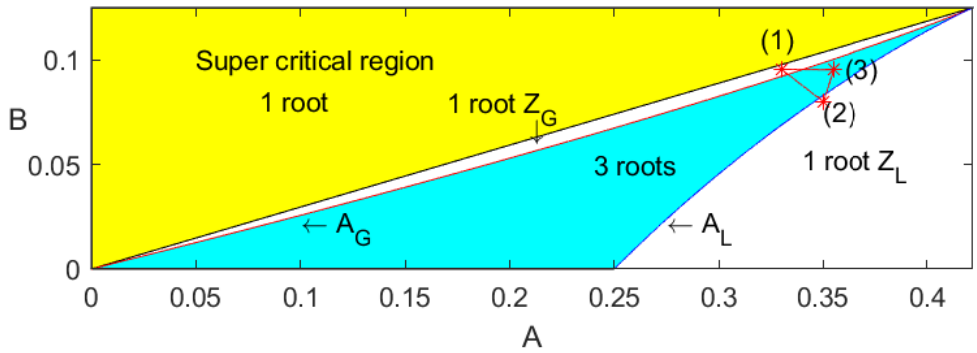
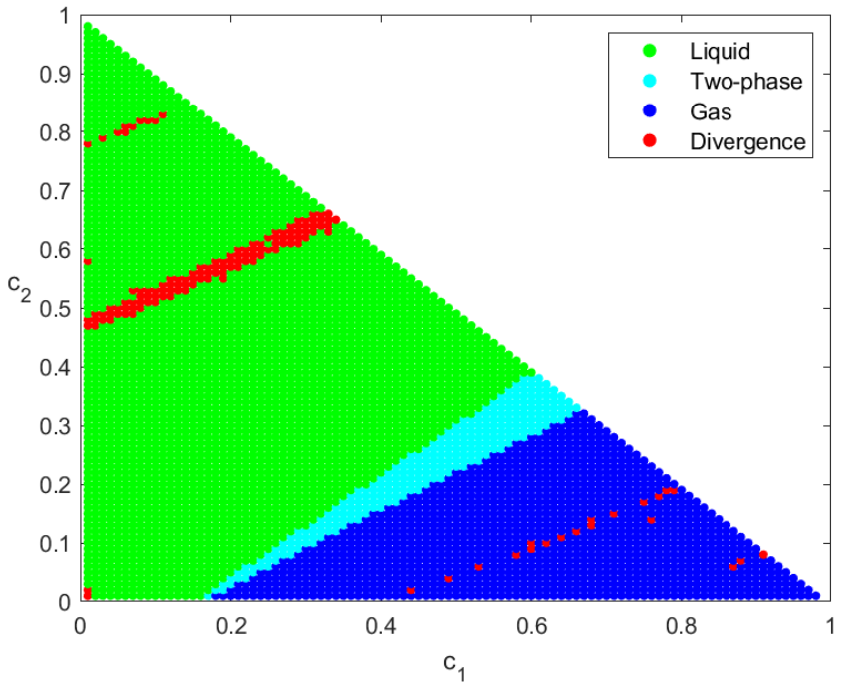
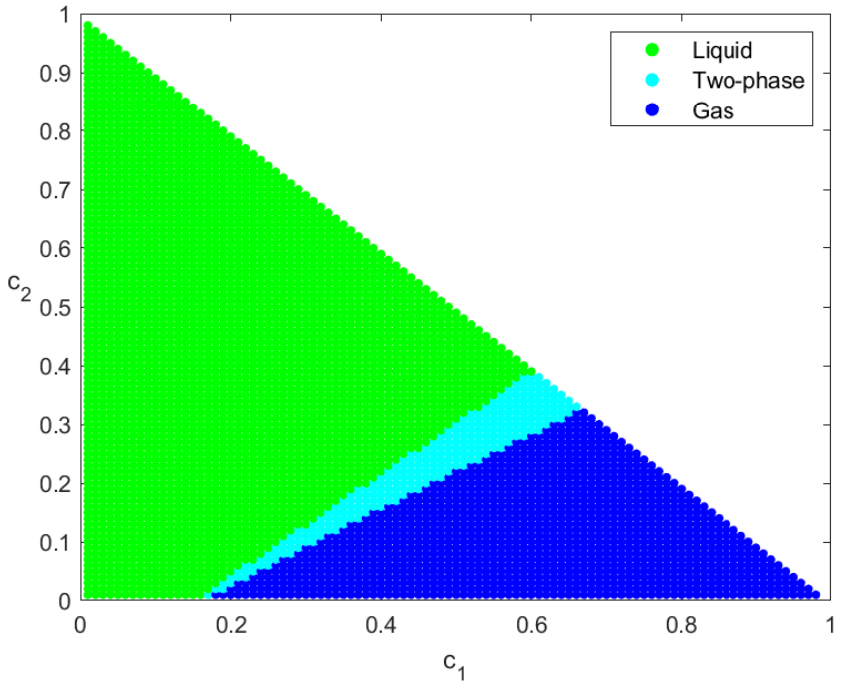


Figure 6.28: Van der Waals' law: $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.33, 0.0955)$, $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.35, 0.08)$ and $(A^{\mathrm{III}}, B^{\mathrm{III}}) = (0.355, 0.0953)$



(a) Newton-min



(b) NPIPM

Figure 6.29: Van der Waals' law: one initial point.

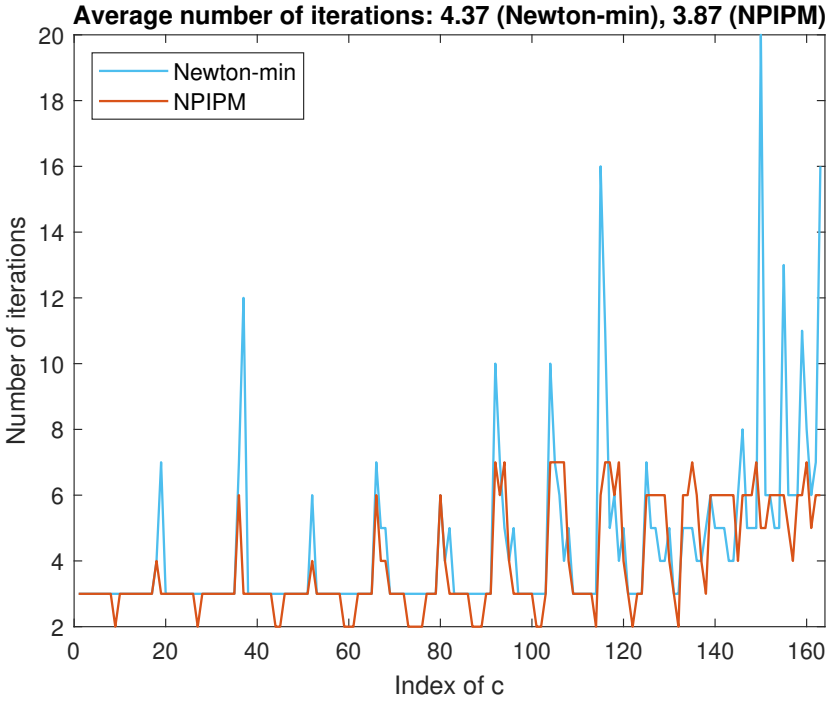


Figure 6.30: Van der Waals' law: number of iterations with the same initial point.

Numerical results. We first compare NPIPM and Newton-min method with the same initial point $(Y, \xi_G^{\mathrm{I}}, \xi_G^{\mathrm{II}}, \xi_G^{\mathrm{III}}, \xi_L^{\mathrm{I}}, \xi_L^{\mathrm{II}}, \xi_L^{\mathrm{III}})^0 = (0.4, 0.2, 0.2, 0.4, 0.4, 0.2, 0.2)$. The stopping criteria is $\|\mathbb{F}(\mathbb{X})\| < 10^{-10}$. We set the maximum number of iterations to be 50. With NPIPM, the line search parameters are $\kappa = 0.4$ and $\varrho = 0.99$. In the last equation of the NPIPM system, we take $\eta = 10^{-4}$. We run Newton-min and NPIPM for all parameters

$$\boldsymbol{c} \in \mathcal{C} = \left\{ (c^{\mathrm{I}}, c^{\mathrm{II}}) \in \mathcal{P}^2 \;\middle|\; c^{\mathrm{I}} + c^{\mathrm{II}} < 1 \right\},$$

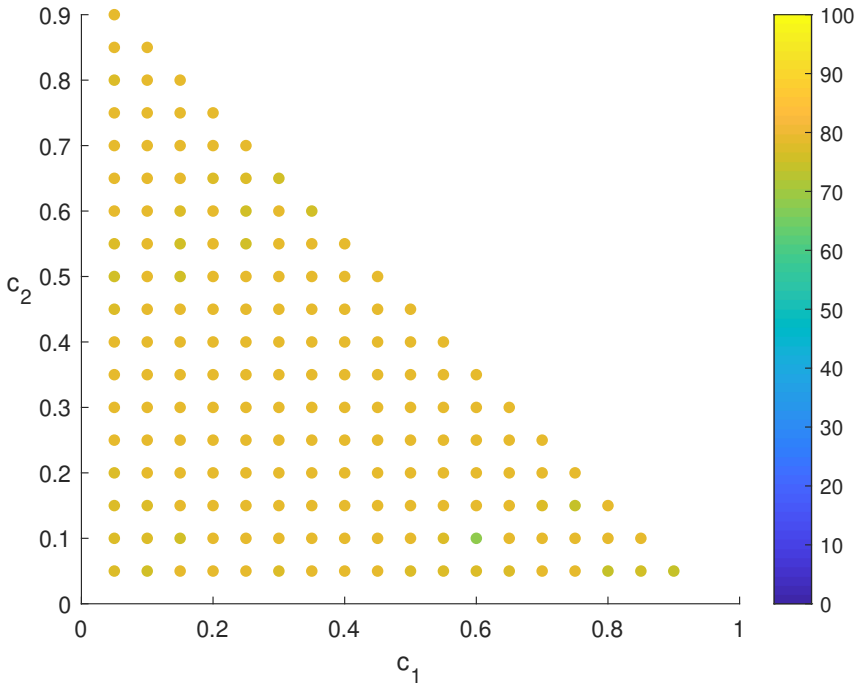
where $\mathcal{P} = \{0.01; 0.02; \ldots; 0.99\}$, when the algorithm converges. In Figure 6.29, we assign the blue color to the gas single-phase regime, the cyan color to the two-phase regime, the green color to the liquid single-phase regime, and the red color for divergence. NPIPM (lower panel) converges with all $\boldsymbol{c}$ tested, while Newton-min (upper panel) exhibits many cases of divergence.

The next test with Van der Waals' law is the number of iterations when the algorithms converge. We still use the same parameters for the convergence test. However, in Figure 6.30, we display the number of iterations instead of values of $\bar{Y}$. Figure 6.30 shows that when Newton-min algorithm converges, it converges in fewer iterations than NPIPM.

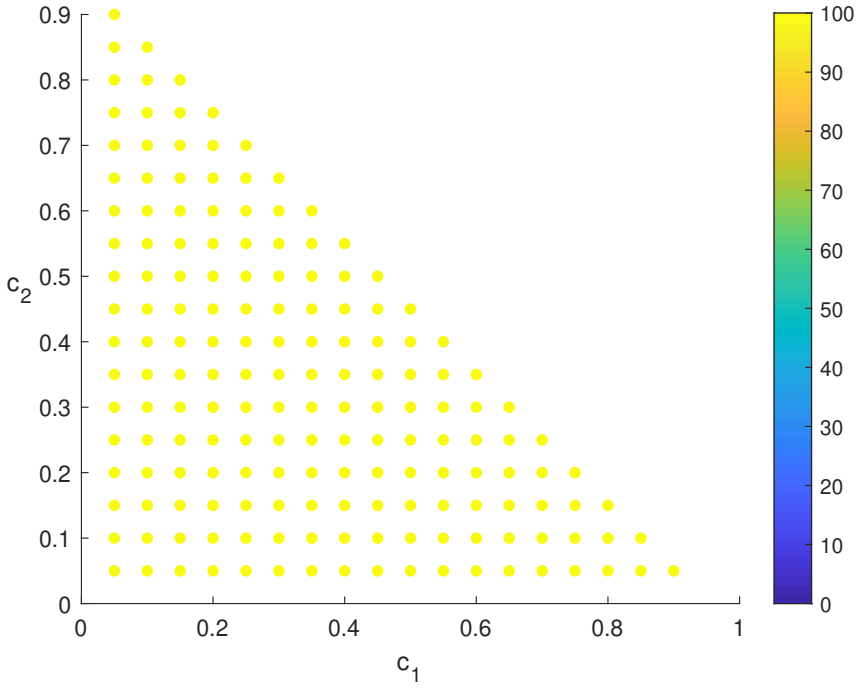
In the last test, we sweep over the grid of parameters

$$\boldsymbol{c} \in \mathcal{C} = \left\{ (c^{\mathrm{I}}, c^{\mathrm{II}}) \in \mathcal{P}^2 \;\middle|\; c^{\mathrm{I}} + c^{\mathrm{II}} < 1 \right\},$$

where $\mathcal{P} = \{0.05; 0.10; \ldots; 0.95\}$ and the set of initial points



(a) Newton-min



(b) NPIPM

Figure 6.31: Van der Waals' law: percentage of convergence over all initial points.

$$\mathcal{D}^0 = \big\{(Y,\, \xi_G^{\mathrm{I}},\, \xi_G^{\mathrm{II}},\, \xi_G^{\mathrm{III}},\, \xi_L^{\mathrm{I}},\, \xi_L^{\mathrm{II}},\, \xi_L^{\mathrm{III}})^0 \in \mathcal{M}^7 \;\mid\; 1-(\xi_G^{\mathrm{I}})^0-(\xi_G^{\mathrm{II}})^0-(\xi_G^{\mathrm{III}})^0>0 \text{ and } 1-(\xi_L^{\mathrm{I}})^0-(\xi_L^{\mathrm{II}})^0-(\xi_L^{\mathrm{III}})^0>0\big\},$$

where $\mathcal{M} = \{0.1;\, 0.2;\, \ldots;\, 0.9\}$. The number of initial points used for the tests is $|\mathcal{D}^0| = 64$. For each c, we count the number of initial points for which the method converges and then plot the percentage of success for each algorithm in Figure 6.31. Figure 6.31 confirms the great efficiency of NPIPM relatively to Newton-min, with 100% of convergence.

6.1.3.3 Peng-Robinson's law

We consider the two-phase ternary model (6.21) with Peng-Robinson' fugacity coefficients (3.53), namely,

$$\begin{aligned}\ln \Phi_\alpha^i(\boldsymbol{x}) = \;& \frac{B(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} B(\boldsymbol{x})\cdot(\boldsymbol{\delta}^i - \boldsymbol{x})}{B(\boldsymbol{x})}\,[Z_\alpha(\boldsymbol{x})-1] - \ln\,[Z_\alpha(\boldsymbol{x}) - B(\boldsymbol{x})] \\ & + \left[\frac{B(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} B(\boldsymbol{x})\cdot(\boldsymbol{\delta}^i - \boldsymbol{x})}{B(\boldsymbol{x})} - \frac{2A(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} A(\boldsymbol{x})\cdot(\boldsymbol{\delta}^i - \boldsymbol{x})}{A(\boldsymbol{x})}\right] \\ & \cdot\; \frac{A(\boldsymbol{x})}{2\sqrt{2}B(\boldsymbol{x})} \ln\left[\frac{Z_\alpha(\boldsymbol{x}) + (1+\sqrt{2})B(\boldsymbol{x})}{Z_\alpha(\boldsymbol{x}) - (\sqrt{2}-1)B(\boldsymbol{x})}\right],\end{aligned} \tag{6.27}$$

for $i \in \{\mathrm{I},\, \mathrm{II},\, \mathrm{III}\}$, $\alpha \in \{G, L\}$, $\boldsymbol{x} \in \big\{(x^{\mathrm{I}}, x^{\mathrm{II}}) \in [0,1]^2 \;\mid\; x^{\mathrm{I}} + x^{\mathrm{II}} \leqslant 1\big\}$, where $Z_\alpha(\boldsymbol{x})$ is a real root of the cubic equation (3.51), that is,

$$\begin{aligned}& Z^3(\boldsymbol{x}) + (B(\boldsymbol{x}) - 1)Z^2(\boldsymbol{x}) \\ & \quad + [A(\boldsymbol{x}) - 2B(\boldsymbol{x}) - 3B^2(\boldsymbol{x})]Z(\boldsymbol{x}) + [B^2(\boldsymbol{x}) + B^3(\boldsymbol{x}) - A(\boldsymbol{x})B(\boldsymbol{x})] = 0.\end{aligned} \tag{6.28}$$

The mixing rules (3.31b)–(3.32) have been used, i.e.,

$$A(\boldsymbol{x}) = \big(\, x^{\mathrm{I}}\sqrt{A^{\mathrm{I}}} + x^{\mathrm{II}}\sqrt{A^{\mathrm{II}}} + (1 - x^{\mathrm{I}} - x^{\mathrm{II}})\sqrt{A^{\mathrm{III}}}\,\big)^2, \tag{6.29a}$$

$$B(\boldsymbol{x}) = \; x^{\mathrm{I}}B^{\mathrm{I}} + x^{\mathrm{II}}B^{\mathrm{II}} + (1 - x^{\mathrm{I}} - x^{\mathrm{II}})B^{\mathrm{III}}. \tag{6.29b}$$

Existence, uniqueness and regularity? Due to the complexity of Peng-Robinson's law, it is difficult to tell anything about the strict convexity of the Gibbs functions g_G and g_L, the excess parts of which are given by (3.52). There is nothing we can predict about the existence, uniqueness and regularity of a solution. In the ternary case, the Gibbs functions can be numerically plotted as functions of $\boldsymbol{x} = (x^{\mathrm{I}}, x^{\mathrm{II}})$, and we can try to select the pairs $(A^{\mathrm{I}},\, B^{\mathrm{I}})$, $(A^{\mathrm{II}},\, B^{\mathrm{II}})$ and $(A^{\mathrm{III}},\, B^{\mathrm{III}})$ in such a way that g_G and g_L are "visually" strictly convex on their respective domains of definition. An example of three such pairs is

$$(A^{\mathrm{I}},\, B^{\mathrm{I}}) = (0.322,\, 0.053), \qquad (A^{\mathrm{II}},\, B^{\mathrm{II}}) = (0.33,\, 0.03), \qquad (A^{\mathrm{III}},\, B^{\mathrm{III}}) = (0.337,\, 0.048)$$

as depicted in Figure 6.32.

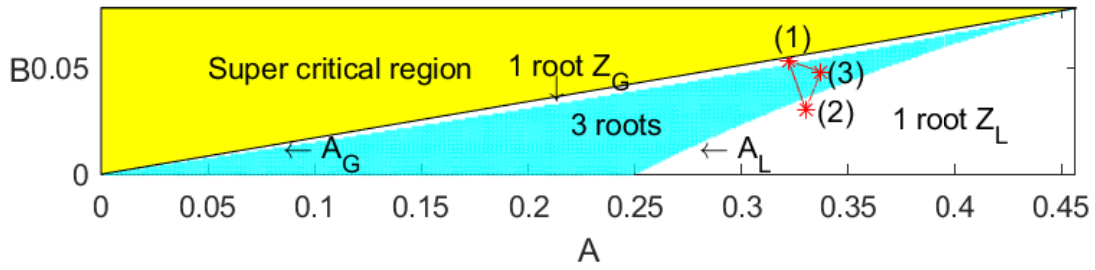


Figure 6.32: Peng-Robinson's law: $(A^{\mathrm{I}}, B^{\mathrm{I}}) = (0.322, 0.053)$, $(A^{\mathrm{II}}, B^{\mathrm{II}}) = (0.33, 0.03)$ and $(A^{\mathrm{III}}, B^{\mathrm{III}}) = (0.337, 0.048)$.

Numerical results. We first compare NPIPM and Newton-min method with the same initial point $(Y, \xi_G^{\mathrm{I}}, \xi_G^{\mathrm{II}}, \xi_G^{\mathrm{III}}, \xi_L^{\mathrm{I}}, \xi_L^{\mathrm{II}}, \xi_L^{\mathrm{III}})^0 = (0.4, 0.3, 0.5, 0.1, 0.325, 0.2, 0.17)$. The stopping criteria is $\|\mathbb{F}(\mathbb{X})\| < 10^{-10}$. We set the maximum number of iterations to be 50. With NPIPM, the line search parameters are $\kappa = 0.4$ and $\varrho = 0.99$. In the last equation of the system, we take $\eta = 10^{-4}$. We use indirect extension with $\varepsilon = 0.03$. We run Newton-min and NPIPM for all parameters

$$\boldsymbol{c} \in \mathcal{C} = \left\{ (c^{\mathrm{I}}, c^{\mathrm{II}}) \in \mathcal{P}^2 \;\middle|\; c^{\mathrm{I}} + c^{\mathrm{II}} < 1 \right\},$$

where $\mathcal{P} = \{0.01; 0.02; \ldots; 0.99\}$ when the algorithm converges. In Figure 6.33, we assign the blue color to the gas single-phase regime, the cyan color to the two-phase regime, the green color to the liquid single-phase regime, and the red color for divergence. We observe that NPIPM (lower panel) converges with all values of $\boldsymbol{c}$ tested, while Newton-min (upper panel) exhibits many cases of divergence.

The next test with Peng-Robinson's law is the number of iterations if the algorithm converges. We still use the same parameters for the convergence test. However, in Figure 6.35, we display the number of iterations instead of values of $\bar{Y}$. Figure 6.35 shows that when Newton-min algorithm converges, it converges in fewer iterations than NPIPM.

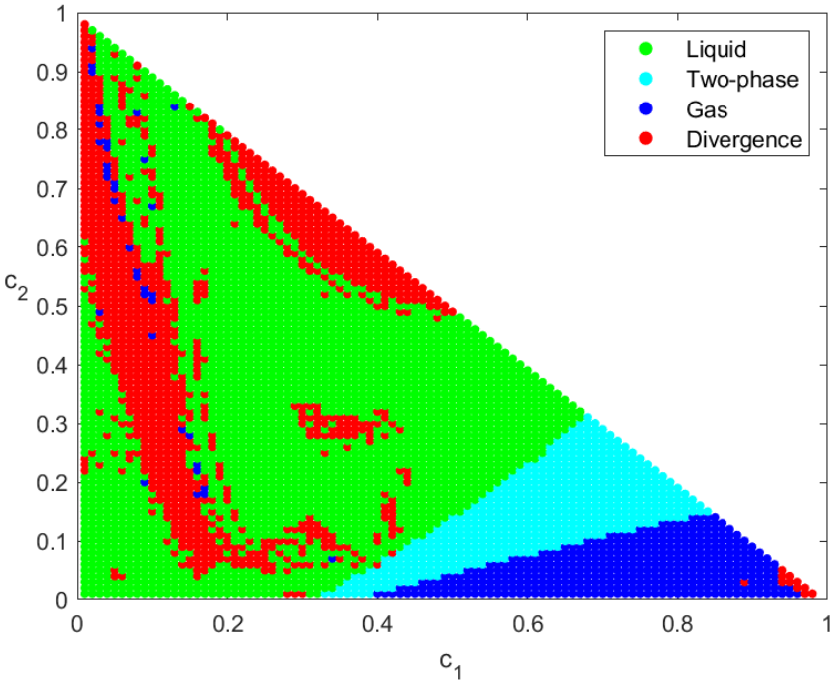
In the last test, we sweep over the grid of parameters

$$\boldsymbol{c} \in \mathcal{C} = \left\{ (c^{\mathrm{I}}, c^{\mathrm{II}}) \in \mathcal{P}^2 \;\middle|\; c^{\mathrm{I}} + c^{\mathrm{II}} < 1 \right\},$$

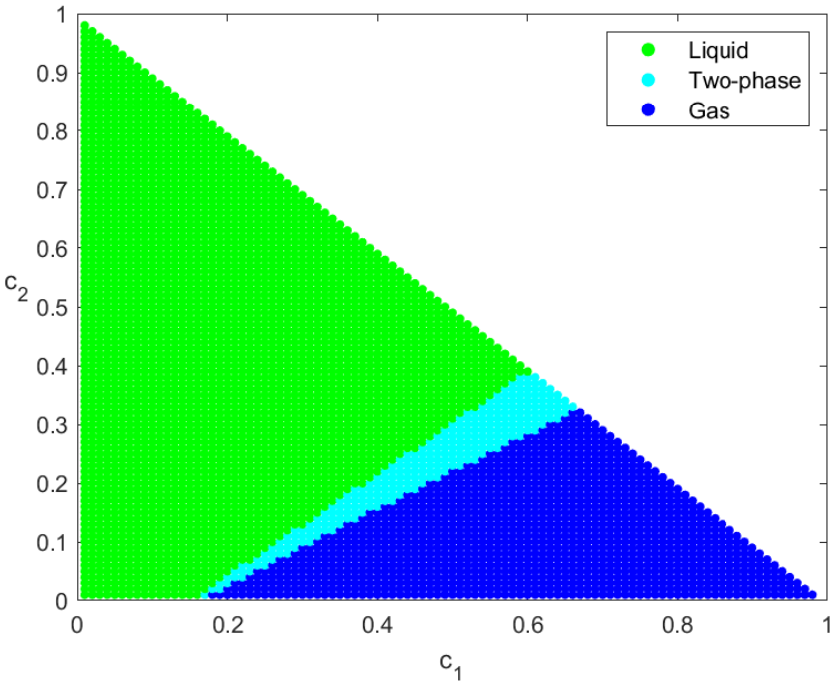
where $\mathcal{P} = \{0.05; 0.10; \ldots; 0.95\}$ and the set of initial points

$$\mathcal{D}^0 = \big\{(Y, \xi_G^{\mathrm{I}}, \xi_G^{\mathrm{II}}, \xi_G^{\mathrm{III}}, \xi_L^{\mathrm{I}}, \xi_L^{\mathrm{II}}, \xi_L^{\mathrm{III}})^0 \in \mathcal{M}^7 \;\big|\; 1 - (\xi_G^{\mathrm{I}})^0 - (\xi_G^{\mathrm{II}})^0 - (\xi_G^{\mathrm{III}})^0 > 0 \text{ and } 1 - (\xi_L^{\mathrm{I}})^0 - (\xi_L^{\mathrm{II}})^0 - (\xi_L^{\mathrm{III}})^0 > 0\big\},$$

where $\mathcal{M} = \{0.1; 0.2; \ldots; 0.9\}$. The number of initial points used for the tests is $|\mathcal{D}^0| = 64$. For each c, we count the number of initial points for which the method converges and then plot the percentage of success for each algorithm in Figure 6.34. Figure 6.34 testifies to the remarkable efficiency of NPIPM relatively to Newton-min, with 100% of convergence.

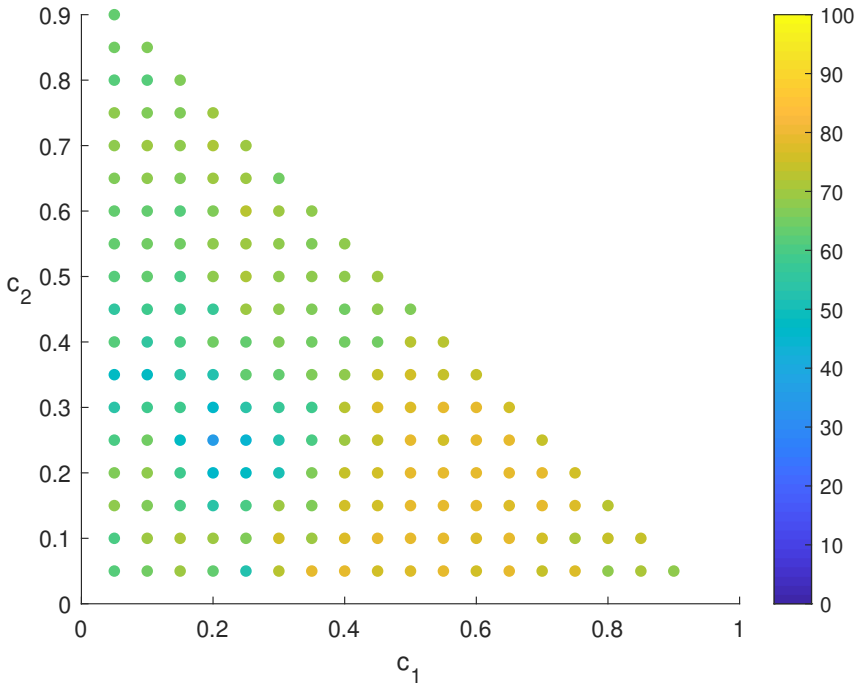


(a) Newton-min

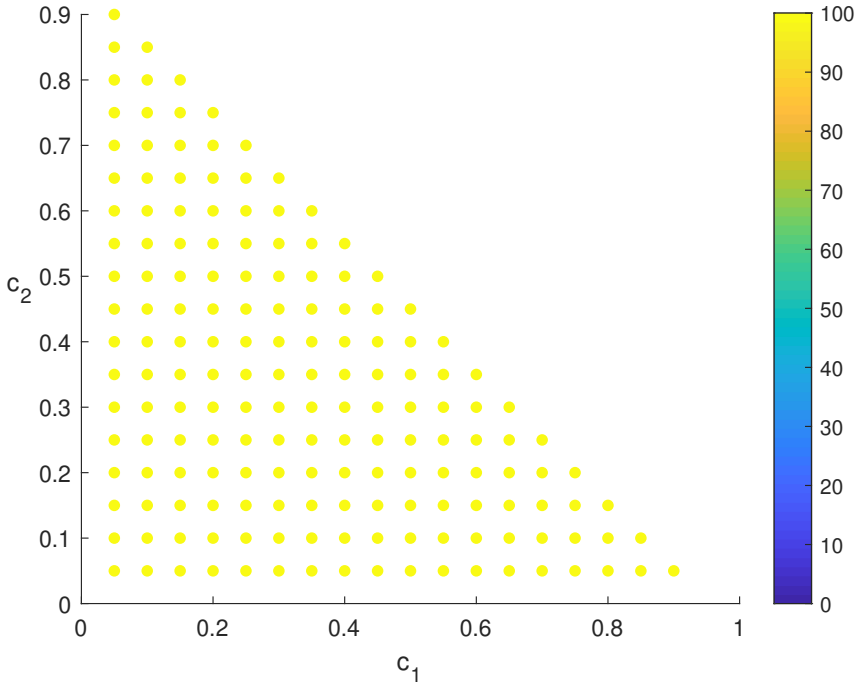


(b) NPIPM

Figure 6.33: Peng-Robinson's law: one initial point.



(a) Newton-min



(b) NPIPM

Figure 6.34: Peng-Robinson's law: percentage of convergence over all initial points.

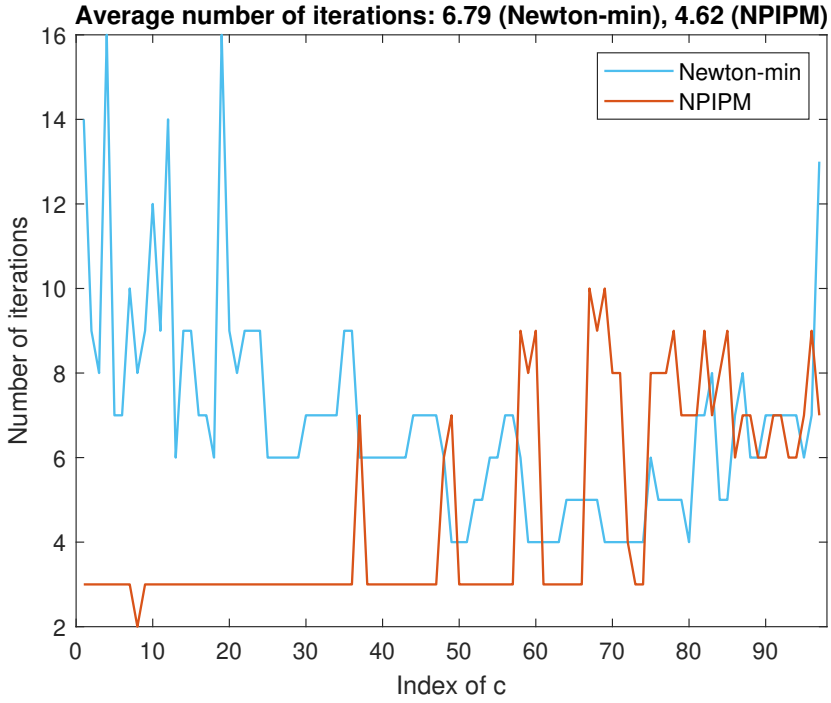


Figure 6.35: Peng-Robinson's law: number of iterations with the same initial point.

6.1.4 Evolutionary binary model

Continuous model. In the two-phase binary model (2.83), the global fraction of the first component c is a given data. We now consider a more sophisticated model in which this composition depends on time. The model consists of an algebro-differential system

$$\frac{\mathrm{d}c}{\mathrm{d}t} - Y(1-Y)\xi_G^{\mathrm{I}}\left(1 - \frac{1}{k^{\mathrm{I}}}\right) = 0, \tag{6.30a}$$

$$Y\xi_G^{\mathrm{I}} + (1-Y)\xi_G^{\mathrm{I}}/k^{\mathrm{I}} - c = 0, \tag{6.30b}$$

$$\min(Y,\, 1 - \xi_G^{\mathrm{I}} - \xi_G^{\mathrm{II}}) = 0, \tag{6.30c}$$

$$\min(1-Y,\, 1 - \xi_G^{\mathrm{I}}/k^{\mathrm{I}} - \xi_G^{\mathrm{II}}/k^{\mathrm{II}}) = 0 \tag{6.30d}$$

in the four unknowns $(c,\, Y,\, \xi_G^{\mathrm{I}},\, \xi_G^{\mathrm{II}})$, equipped with the intial condition

$$(c,\, Y,\, \xi_G^{\mathrm{I}},\, \xi_G^{\mathrm{II}})(t=0) = (c_0, Y_0, (\xi_G^{\mathrm{I}})_0, (\xi_G^{\mathrm{II}})_0) \tag{6.31}$$

subject to the equilibrium relations

$$Y_0(\xi_G^{\mathrm{I}})_0 + (1-Y_0)(\xi_G^{\mathrm{II}})_0/k^{\mathrm{I}} - c_0 = 0, \tag{6.32a}$$

$$\min(Y_0,\, 1 - (\xi_G^{\mathrm{I}})_0 - (\xi_G^{\mathrm{II}})_0) = 0, \tag{6.32b}$$

$$\min(1-Y_0,\, 1 - (\xi_G^{\mathrm{I}})_0/k^{\mathrm{I}} - (\xi_G^{\mathrm{II}})_0/k^{\mathrm{II}}) = 0. \tag{6.32c}$$

Henry's law with $k^{\mathrm{I}},\, k^{\mathrm{II}} > 0$ have been implicitly used.

Let K_L, K_G be the constants defined by (6.11). It can then be easily proven that the exact solution of (6.30)–(6.32) is given by

$$c(t) = \begin{cases} c_0 & \text{if } c_0 \in [0, K_L], \\ \dfrac{K_G \gamma_0 \exp(t) + K_L}{\gamma_0 \exp(t) + 1} & \text{if } c_0 \in (K_L, K_G), \\ c_0 & \text{if } c_0 \in [K_G, 1], \end{cases} \tag{6.33}$$

where

$$\gamma_0 = \frac{c_0 - K_L}{K_G - c_0}. \tag{6.34}$$

The values of $Y(t)$, $\xi_G^{\mathrm{I}}(t)$ and $\xi_G^{\mathrm{II}}(t)$ are deduced from $c(t)$ by formulas (6.12) [Proposition 6.1].

Discretized system. But our primary interest is the algebraic system that arises when we apply the Euler backward scheme to (6.30) with a time-step $\Delta t > 0$. This system reads

$$\boxed{\begin{aligned} c - c_\flat - \tau\left(1 - \frac{1}{k^{\mathrm{I}}}\right)\xi_G^{\mathrm{I}} Y(1-Y) &= 0, && \text{(6.35a)} \\ Y\xi_G^{\mathrm{I}} + (1-Y)\xi_G^{\mathrm{I}}/k^{\mathrm{I}} - c &= 0, && \text{(6.35b)} \\ \min(Y,\, 1 - \xi_G^{\mathrm{I}} - \xi_G^{\mathrm{II}}) &= 0, && \text{(6.35c)} \\ \min(1-Y,\, 1 - \xi_G^{\mathrm{I}}/k^{\mathrm{I}} - \xi_G^{\mathrm{II}}/k^{\mathrm{II}}) &= 0, && \text{(6.35d)} \end{aligned}}$$

where the notations have been changed from Δt to τ, from c^n to $c_\flat$ and from $(c,\, Y,\, \xi_G^{\mathrm{I}},\, \xi_G^{\mathrm{II}})^{n+1}$ to $(c,\, Y,\, \xi_G^{\mathrm{I}},\, \xi_G^{\mathrm{II}})$. In (6.35), $c_\flat \in [0,1]$, $\tau > 0$ and $k^{\mathrm{I}} > 1 > k^{\mathrm{II}} > 0$ play the role of parameters. The upcoming Theorem addresses the question of its solutions.

Proposition 6.2. *Let K_L, K_G be the constants defined by* (6.11). *Except for the case 3(b) in the enumeration below, system* (6.35) *has a unique solution $(\bar{c}, \bar{Y}, \bar{\xi}_G^{\mathrm{I}}, \bar{\xi}_G^{\mathrm{II}}) \in [0,1] \times [0,1] \times \mathbb{R}_+ \times \mathbb{R}_+$ called* reference solution.

1. *If $c_\flat \in [K_G, 1]$, then the reference solution is in the G single-phase regime and given by*

$$\bar{c} = c_\flat, \quad \bar{Y} = 1, \quad \bar{\xi}_G^{\mathrm{I}} = c_\flat, \quad \bar{\xi}_G^{\mathrm{II}} = 1 - c_\flat. \tag{6.36}$$

2. *If $c_\flat \in (K_L, K_G)$, then the reference solution is in the two-phase regime and given by*

$$\bar{c} = \frac{K_G + K_L}{2} - \frac{K_G - K_L}{2\tau}\left\{1 - \left[1 - 2\frac{K_G + K_L - 2c_\flat}{K_G - K_L}\tau + \tau^2\right]^{1/2}\right\}. \tag{6.37}$$

   The values of $\bar{Y}$, $\bar{\xi}_G^{\mathrm{I}}$ and $\bar{\xi}_G^{\mathrm{II}}$ are deduced from $\bar{c}$ by formulas (6.12) *[Proposition 6.1].*

3. *If $c_\flat \in [0, K_L]$, then the number*

$$\tau_{\max} = \frac{K_G + K_L - 2c_\flat}{K_G - K_L} + \sqrt{\left(\frac{K_G + K_L - 2c_\flat}{K_G - K_L}\right)^2 - 1} \tag{6.38}$$

   is well-defined and greater than or equal to 1.

(a) For $\tau < \tau_{\max}$, the reference solution is in the L single-phase regime and given by

$$\bar{c} = c_\flat, \quad \bar{Y} = 0, \quad \bar{\xi}_G^{\mathrm{I}} = k^{\mathrm{I}} c_\flat, \quad \bar{\xi}_G^{\mathrm{II}} = k^{\mathrm{II}}(1 - c_\flat); \tag{6.39}$$

(b) For $\tau \geqslant \tau_{\max}$, in addition to (6.39) *that we declare to be the reference solution, there are two spurious solutions (counted with multiplicity).*

Chứng minh. The last three equations of model (6.35) are exactly the stationary binary model (6.10). Therefore, $(Y, \xi_G^{\mathrm{I}}, \xi_G^{\mathrm{II}})$ can be expressed as functions of c by means of (6.12). In particular,

$$Y = \frac{c - K_L}{K_G - K_L}\, \mathbf{1}_{(K_L, K_G)}(c)$$

for all phase regimes, using the characteristic function $\mathbf{1}$. Inserting this into the first equation (6.35a) and invoking $k^{\mathrm{I}} = K_G/K_L$, we obtain a scalar equation on c, namely,

$$c - c_\flat + \frac{\tau}{K_G - K_L}(c - K_L)(c - K_G)\, \mathbf{1}_{(K_L, K_G)}(c) = 0. \tag{6.40}$$

The rest of the proof relies on studying the function representing the left-hand side of the above equation. This part is not difficult and is left to the readers. □

Remark 6.1. The choice $\bar{c} = c_\flat$ for the reference solution in case 3(b) is really natural insofar as this is the continuous extension —with respect to τ— of the reference solution of case 3(a).

Regularity of zeros. The most significant result for this model is that the reference solution corresponds most of the time to a regular zero.

Theorem 6.1. *For all $\tau \geqslant 0$, the reference solution of* (6.35) *defined in Proposition 6.2 gives rise to a regular zero for the NPIPM system, except at transitional and azeotropic points.*

Chứng minh. Let $X = (c,\, Y,\, \xi_G^{\mathrm{I}},\, \xi_G^{\mathrm{II}})$. Define

$$\Lambda(X) = \begin{bmatrix} c - c_\flat - \tau(1 - 1/k^{\mathrm{I}})\, \xi_G^{\mathrm{I}}\, Y(1 - Y) \\ Y\xi_G^{\mathrm{I}} + (1 - Y)\, \xi_G^{\mathrm{I}}/k^{\mathrm{I}} - c \end{bmatrix},$$

and

$$G(X) = \begin{bmatrix} Y \\ 1 - Y \end{bmatrix}, \qquad H(X) = \begin{bmatrix} 1 - \xi_G^{\mathrm{I}} - \xi_G^{\mathrm{II}} \\ 1 - \xi_G^{\mathrm{I}}/k^{\mathrm{I}} - \xi_G^{\mathrm{II}}/k^{\mathrm{II}} \end{bmatrix}.$$

By Lemma 5.4 and Lemma 4.4, we can study the sign of

$$\bar{\mathfrak{d}} = \begin{vmatrix} \nabla\Lambda(\bar{X}) \\ \nabla G(\bar{X}) \odot H(\bar{X}) + \nabla H(\bar{X}) \odot G(\bar{X}) \end{vmatrix},$$

where $\bar{X} = (\bar{c},\, \bar{Y},\, \bar{\xi}_G^{\mathrm{I}},\, \bar{\xi}_G^{\mathrm{II}})$ is the reference solution, instead of the sign of $\det \nabla\mathbb{F}(\bar{\mathbb{X}})$ or $\det \nabla\mathbf{F}(\bar{\mathbf{X}})$. In this case, we have

$$\bar{\mathfrak{d}} = \begin{vmatrix} 1 & -\tau\Delta\bar{\xi}^{\mathrm{I}}(1 - 2\bar{Y}) & -\tau\Delta\bar{\xi}^{\mathrm{I}}\, \bar{Y}(1 - \bar{Y})/\bar{\xi}_G^{\mathrm{I}} & 0 \\ -1 & \Delta\bar{\xi}^{\mathrm{I}} & \bar{Y} + (1 - \bar{Y})/k^{\mathrm{I}} & 0 \\ 0 & 1 - \bar{\sigma}_G & -\bar{Y} & -\bar{Y} \\ 0 & -1 + \bar{\sigma}_L & (\bar{Y} - 1)/k^{\mathrm{I}} & (\bar{Y} - 1)/k^{\mathrm{II}} \end{vmatrix},$$

with

$$\Delta\bar{\xi}^{\mathrm{I}} = \bar{\xi}^{\mathrm{I}}_G - \bar{\xi}^{\mathrm{I}}_L = \bar{\xi}^{\mathrm{I}}_G(1 - 1/k^{\mathrm{I}}), \qquad \bar{\sigma}_G = \bar{\xi}^{\mathrm{I}}_G + \bar{\xi}^{\mathrm{II}}_G, \qquad \bar{\sigma}_L = \bar{\xi}^{\mathrm{I}}_L + \bar{\xi}^{\mathrm{II}}_L.$$

Expanding the determinant with respect to the first column, we find

$$\overline{\mathfrak{d}} = \overline{\mathfrak{d}}_0 - \tau\,\overline{\mathfrak{d}}_1, \tag{6.41}$$

where

$$\overline{\mathfrak{d}}_0 = \begin{vmatrix} \Delta\bar{\xi}^{\mathrm{I}} & \bar{Y} + (1-\bar{Y})/k^{\mathrm{I}} & 0 \\ 1-\bar{\sigma}_G & -\bar{Y} & -\bar{Y} \\ -1+\bar{\sigma}_L & (\bar{Y}-1)/k^{\mathrm{I}} & (\bar{Y}-1)/k^{\mathrm{II}} \end{vmatrix}$$

is the determinant of the stationary binary model and was already computed in §5.2.2, and

$$\overline{\mathfrak{d}}_1 = \begin{vmatrix} \Delta\bar{\xi}^{\mathrm{I}}(1-2\bar{Y}) & \Delta\bar{\xi}^{\mathrm{I}}\,\bar{Y}(1-\bar{Y})/\bar{\xi}^{\mathrm{I}}_G & 0 \\ 1-\bar{\sigma}_G & -\bar{Y} & -\bar{Y} \\ -1+\bar{\sigma}_L & (\bar{Y}-1)/k^{\mathrm{I}} & (\bar{Y}-1)/k^{\mathrm{II}} \end{vmatrix}.$$

Assume $\bar{Y} = 1$, i.e., the solution is in the gas phase. Then, $\bar{\sigma}_G = 1$. From §5.2.2, we know that $\overline{\mathfrak{d}}_0 = 1 - \bar{\sigma}_L$. By a direct computation, we have $\overline{\mathfrak{d}}_1 = 0$. Therefore, $\overline{\mathfrak{d}} = \overline{\mathfrak{d}}_0 = 1 - \bar{\sigma}_L \geqslant 0$. Equality holds at a transition point. The other single-phase case $\bar{Y} = 0$ is similar.

Assume now $\bar{Y} \in (0,1)$, i.e., the solution is in the two-phase regime. Then, $\bar{\sigma}_G = \bar{\sigma}_L = 1$, $\boldsymbol{\xi}_G = \boldsymbol{x}_G$ and $\boldsymbol{\xi}_L = \boldsymbol{x}_L$. From §5.2.2, we know that

$$\overline{\mathfrak{d}}_0 = \Delta\bar{x}^{\mathrm{I}} \begin{vmatrix} -\bar{Y} & -\bar{Y} \\ (\bar{Y}-1)/k^{\mathrm{I}} & (\bar{Y}-1)/k^{\mathrm{II}} \end{vmatrix}$$

can be expressed as a quadratic form and hence $\overline{\mathfrak{d}}_0 \geqslant 0$, with equality if and only if $\boldsymbol{x}_G = \boldsymbol{x}_L$, namely, at an azeotropic point. Let us compute $\overline{\mathfrak{d}}_1$. By expanding with respect to its first column and by noticing that the, we obtain

$$\overline{\mathfrak{d}}_1 = (1 - 2\bar{Y})\,\overline{\mathfrak{d}}_0.$$

Coming back to (6.41), we get

$$\overline{\mathfrak{d}} = \overline{\mathfrak{d}}_0\,[1 - \tau(1 - 2\bar{Y})].$$

From (6.37), we infer by (6.12) that

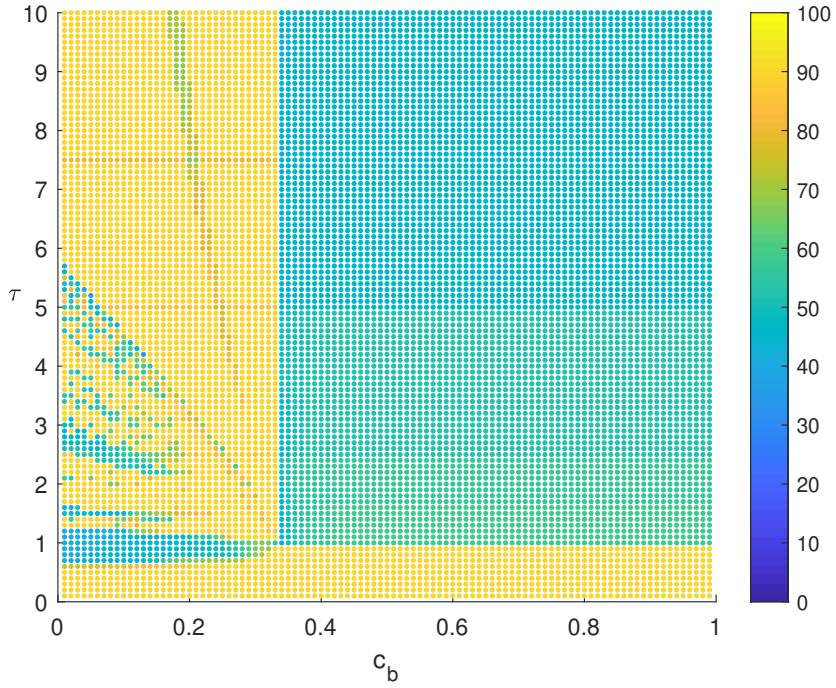
$$\tau(1-2\bar{Y}) = 1 - \left[1 - 2\frac{K_G + K_L - 2c_\flat}{K_G - K_L}\tau + \tau^2\right]^{1/2} < 1.$$

Consequently, $\overline{\mathfrak{d}}$ has the same sign behavior as $\overline{\mathfrak{d}}_0$. This completes the proof. □

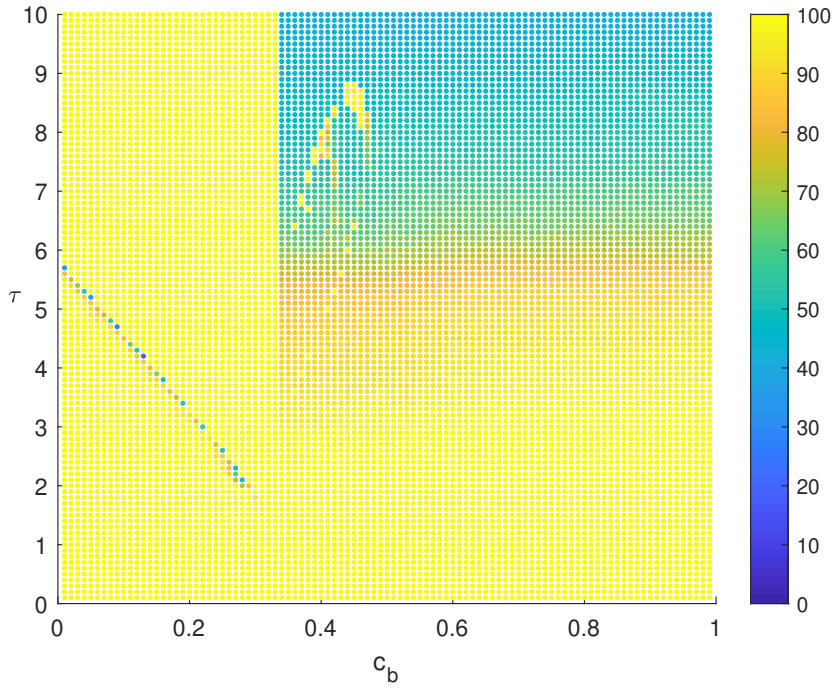
Numerical results. We compare NPIPM with Newton-min. We fix $k^{\mathrm{I}} = 2$, $k^{\mathrm{II}} = 0.5$. The stopping criterion is $\|\mathbb{F}(\mathbb{X})\| < 10^{-7}$. We set the maximum number of iterations to be 50. If the number of iterations of the algorithm exceeds this maximum number, the case will be considered as divergent. With NPIPM, the parameters for the line search are $\kappa = 0.4$ and $\varrho = 0.99$. In the last equation of the NPIPM system, we take $\eta = 10^{-6}$.

We sweep over the grid of parameters

$$(c_\flat, \tau) \in \{0.01;\, 0.02;\, \ldots;\, 0.99\} \times \big\{0.1;\, 0.2;\, \ldots;\, 10\big\}.$$



(a) Newton-min



(b) NPIPM

Figure 6.36: Evolutionary binary model: percentage of convergence over all initial points.

and the set of initial points

$$\mathcal{D}^0 = \left\{ (Y, \xi_G^{\mathrm{I}}, \xi_G^{\mathrm{II}}, c)^0 \in \mathcal{M}^4 \;\middle|\; 1 - (\xi_G^{\mathrm{I}})^0 - (\xi_G^{\mathrm{II}})^0 > 0 \text{ and } 1 - (\xi_G^{\mathrm{I}})^0/k^{\mathrm{I}} - (\xi_G^{\mathrm{II}})^0/k^{\mathrm{II}} > 0 \right\}$$

where $\mathcal{M} = \{0.1; 0.2; \ldots; 0.9\}$. The number of initial points used for the tests is $|\mathcal{D}^0| = 1944$. For each pair $(c_\flat, \tau)$, we count the number of initial points for which the method converges and then plot the percentage of success for each algorithm. The results are shown in Figure 6.36. The specific case $\tau = 1$ is highlighted in Figure 6.37.

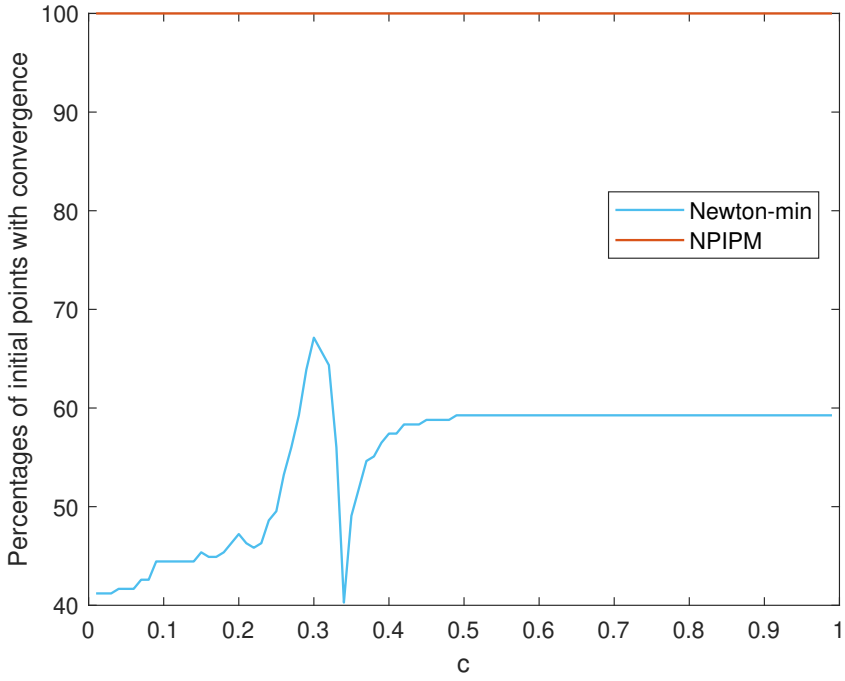


Figure 6.37: Evolutionary binary model with $\tau = 1$.

6.2 Multiphase compositional model

After the simple models of the previous section, we now consider a relatively realistic multiphase compositional fluid flow model used at IFPEN. Our purpose is to compare NPIPM to the Newton-min method on this model developed in Fortran 90.

6.2.1 Continuous model

In this model, there can be up to three phases, namely,

$$\mathscr{P} = \{W, O, G\},$$

where W stands for water, O stands for oil and G stands for gas. Moreover, the water phase is assumed to be *pure* and *immiscible*, that is, it contains only one component referred to as H_2O and this component does not appear in the two other phases.

$$\mathcal{K} = \{\mathrm{I}, \mathrm{II}, \ldots, \mathrm{K}\}$$

be the set of hydrocarbon components, with $\mathrm{K} \geqslant 2$. As a consequence of the previous assumption on the water phase, the hydrocarbons are present only in the oil and gas phases, which are mixable and compositional. Assuming that the medium is isotherm with fixed temperature T, we consider the following problem.

GIVEN

$$\phi,\ \rho_W^\circ,\ \{\rho_\alpha\}_{\alpha\in\mathscr{P}\backslash\{W\}},\ \{\Phi_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\mathscr{P}\backslash\{W\}},\ \{\lambda_\alpha\}_{\alpha\in\mathscr{P}},\ \{Q_\alpha\}_{\alpha\in\mathscr{P}},$$

FIND

$$\{\mathsf{S}_\alpha\}_{\alpha\in\mathscr{P}},\ \{\xi_\alpha^i\}_{(i,\alpha)\in\mathcal{K}\times\mathscr{P}\backslash\{W\}},\ \{\boldsymbol{u}_\alpha\}_{\alpha\in\mathscr{P}},\ \mathsf{P}$$

as functions of $(\boldsymbol{\chi}, t) \in \mathbf{D}_\chi \times \mathbb{R}_+$, where $\mathbf{D}_\chi \subset \mathbb{R}^2$ is a bounded domain, satisfying

- the mass conservation of H_2O and hydrocarbons

$$\phi\frac{\partial}{\partial t}(\rho_W^\circ \mathsf{S}_W) + \mathrm{div}_{\boldsymbol{\chi}}(\rho_W^\circ \boldsymbol{u}_W) = q_W, \tag{6.42a}$$

$$\phi\frac{\partial}{\partial t}(\rho_O \mathsf{S}_O \xi_O^i + \rho_G \mathsf{S}_G \xi_G^i) + \mathrm{div}_{\boldsymbol{\chi}}(\rho_O \xi_O^i \boldsymbol{u}_O + \rho_G \xi_G^i \boldsymbol{u}_G) = q^i, \tag{6.42b}$$

  for all $i \in \mathcal{K}$, where the source terms are given by

$$\begin{aligned} q_W &= \rho_W^\circ Q_W, \\ q^i &= \rho_O \xi_O^i Q_O + \rho_G \xi_G^i Q_G; \end{aligned}$$

- the conservation of volume

$$\sum_{\alpha\in\mathscr{P}} \mathsf{S}_\alpha - 1 = 0; \tag{6.42c}$$

- the extended fugacity equalities

$$\xi_O^i \Phi_O^i(\boldsymbol{x}_O, \mathsf{P}) = \xi_G^i \Phi_G^i(\boldsymbol{x}_G, \mathsf{P}), \qquad \forall i \in \mathcal{K}, \tag{6.42d}$$

  where the components of $\boldsymbol{x}_\alpha = (x_\alpha^{\mathrm{I}}, \ldots, x_\alpha^{\mathrm{K}-1}) \in \mathbb{R}^{\mathrm{K}-1}$ are defined as

$$x_\alpha^i = \frac{\xi_\alpha^i}{\sum_{j\in\mathcal{K}} \xi_\alpha^j};$$

- the complementarity conditions

$$\min\left(\mathsf{S}_\alpha,\ 1 - \sum_{i\in\mathcal{K}} \xi_\alpha^i\right) = 0, \qquad \forall \alpha \in \mathscr{P}\backslash\{W\}; \tag{6.42e}$$

- the Darcy-Muskat law

$$\boldsymbol{u}_\alpha = -\lambda_\alpha \nabla_{\boldsymbol{\chi}} \mathsf{P}, \qquad \forall \alpha \in \mathscr{P}. \tag{6.42f}$$

- homogeneous Neumann boundary conditions on $\partial\mathbf{D}_\chi$.

The fugacity coefficents Φ_α^i, $\alpha \in \{O, G\}$, are those of the Peng-Robinson cubic law, elaborated on in §3.2, where the liquid phase L has been replaced by the oil phasse O.

In comparison with the introductory model (1.4), there are two additional features. Firstly, the phase densities ρ_α for $\alpha \in \{O, G\}$ are no longer constant. Instead, they are now known functions of the pressure P and the extended composition $\boldsymbol{\xi}_\alpha$, in order to account for the *compressibility* of the flow. Secondly, the source terms q_W and q^i in (6.42a)–(6.42b) represent injection and production wells located in the domain. The functions Q_α are concentrated in space and depend on time by means of some given *scenarios*.

In practice, we do not really retain the velocity fields $\boldsymbol{u}_\alpha$ as unknowns. To reduce the size of the system, the velocities $\boldsymbol{u}_\alpha$ are eliminated by means of the last equation (6.42f). The number of remaining unknown scalar fields and equations is then equal to $2\mathrm{K} + 4$. The compositional multiphase model (6.42) is a PDE system, stated at the continuous level. It has to be discretized in space and in time.

6.2.2 Discretized system and resolution

We use the cell centered finite volume method with an upstream two point flux discretization as spatial discretization [4, 12, 84, 101] and the backward Euler method with variable time step for the time discretization. Let $\mathtt{M}_h$ be an admissible finite volume mesh of the reservoir $\mathbf{D}_\chi$, a generic control volume (or cell) of which are denoted by $\mathtt{V}$. Let $|\mathtt{M}_h|$ be the number of volumes. We also introduce an increasing sequence of discret times $\{t^n\}_{0 \leqslant n \leqslant N}$ such that $t^0 = 0$ and $t^N = T$. The vectors of the discrete unknowns in each finite volume $\mathtt{V}$ and at each time t^n are denoted by

$$X_\mathtt{V}^n = \left(\mathsf{P}_\mathtt{V}^n,\ (\mathsf{S}_W)_\mathtt{V}^n,\ (\mathsf{S}_G)_\mathtt{V}^n,\ (\xi_O^{\mathrm{I}})_\mathtt{V}^n,\ \ldots,\ (\xi_O^{\mathrm{K}})_\mathtt{V}^n,\ (\xi_G^{\mathrm{I}})_\mathtt{V}^n,\ \ldots,\ (\xi_G^{\mathrm{K}})_\mathtt{V}^n\right) \in \mathbb{R}^{2\mathrm{K}+3},$$

which gives rise to a global unknown vector

$$X_h^n = \{X_\mathtt{V}^n\}_{\mathtt{V} \in \mathtt{M}_h} \in \mathbb{R}^{(2\mathrm{K}+3)|\mathtt{M}_h|}.$$

Here, we have implicitly eliminated S_O from the set of unknowns by using $\mathsf{S}_O = 1 - \mathsf{S}_W - \mathsf{S}_G$ from equation (6.42c). Therefore, in order to go from t^n to t^{n+1}, we need to solve a nonlinear system of the form

$$\Lambda_h(X_h^{n+1}) = 0, \tag{6.43a}$$

$$\min(G_h(X_h^{n+1}),\ H_h(X_h^{n+1})) = 0. \tag{6.43b}$$

The vector $\Lambda_h(X_h^{n+1}) = \{\Lambda_\mathtt{V}(X_h^{n+1})\}_{\mathtt{V} \in \mathtt{M}_h} \in \mathbb{R}^{(2\mathrm{K}+1)|\mathtt{M}_h|}$ contains the discretized conservation laws (6.42a)–(6.42b) and the extended equilibrium equations (6.42d). Note that the argument of $\Lambda_\mathtt{V}$ is X_h^{n+1} and not $X_\mathtt{V}^{n+1}$, since the discretization of conservation laws (6.42a)–(6.42b) in a given control volume involves its neighbor cells. Meanwhile,

$$G_h(X_h^{n+1}) = \{G_\mathtt{V}(X_\mathtt{V}^{n+1})\}_{\mathtt{V} \in \mathtt{M}_h} \in \mathbb{R}^{2|\mathtt{M}_h|}, \qquad H_h(X_h^{n+1}) = \{H_\mathtt{V}(X_\mathtt{V}^{n+1})\}_{\mathtt{V} \in \mathtt{M}_h} \in \mathbb{R}^{2|\mathtt{M}_h|},$$

come from (6.42e) and are strictly local to each cell, with

$$G_\mathtt{V}(X_\mathtt{V}^{n+1}) = \begin{bmatrix} 1 - (\mathsf{S}_W)_\mathtt{V}^{n+1} - (\mathsf{S}_G)_\mathtt{V}^{n+1} \\ (\mathsf{S}_G)_\mathtt{V}^{n+1} \end{bmatrix}, \quad H_\mathtt{V}(X_\mathtt{V}^{n+1}) = \begin{bmatrix} 1 - \sum_{i \in \mathcal{K}} (\xi_G^i)_\mathtt{V}^{n+1} \\ 1 - \sum_{i \in \mathcal{K}} (\xi_O^i)_\mathtt{V}^{n+1} \end{bmatrix}.$$

Newton-min method. At each time-step $t^n \to t^{n+1}$, after combining all equations over all finite volumes $\mathtt{V} \in \mathtt{M}_h$, we have a system of $(2\mathrm{K}+3)|\mathtt{M}_h|$ equations and then apply Newton-min method to solve this system. In particular, if $\mathrm{K}=3$ the system has $9|\mathtt{M}_h|$. It is natural to choose the solution $X_h^n = \{X_\mathtt{V}^n\}_{\mathtt{V}\in\mathtt{M}_h}$ at time t^n as the initial point $(X_h^{n+1})^0$ when applying the Newton-min solver to (6.43). To alleviate notations, we shall henceforth omit the time label $n+1$ in all variables.

NPIPM. When applying NPIPM, we normally need to add three slack variables per cell. This introduces 3 extra variables per cell, as well as one extra global variable ν. The system to be solved will have $(2\mathrm{K}+7)|\mathtt{M}_h|+1$ equations. In particular, if $\mathrm{K}=3$, the system has $13|\mathtt{M}_h|+1$ equations. At each iteration, the Jacobian matrix must be inverted. In comparison to Newton-min, the complexity of this task has thus increased by the ratio $\left((13|\mathtt{M}_h|+1)\,/\,9|\mathtt{M}_h|\right)^2 \approx 2$. To avoid this waste of resource, we will add to our system just one extra variable ν and no explicit slack variable. With NPIPM, we need initial points which satisfy the positivity of the arguments in complementarity conditions. Since X_h^n is not a strictly interior point, we cannot use it as an intial point. Instead, we will have to modify it to obtain an appropriate value for $(X_h^{n+1})^0$.

6.2.3 Comparison of the results

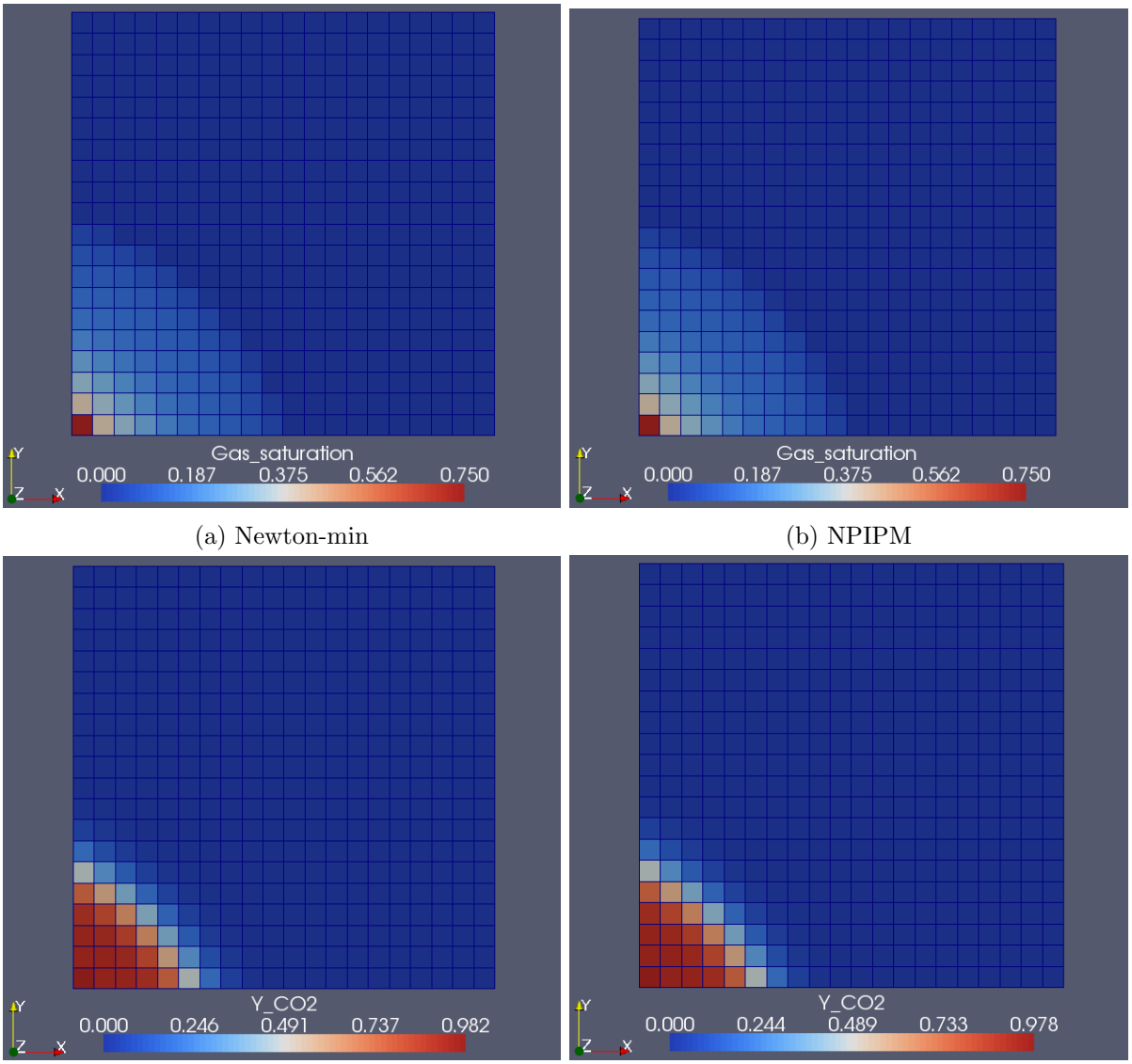
We will compare NPIPM to the Newton-min method, which is currently used in IFPEN's code for the multiphase compositional model by means of two selected test settings. In order to reflect the different physical phenomena present in these test cases, the following models are chosen. In the triphasic cases, in which water oil and gas are present, the Brooks and Corey model is used for the relative permeabilities. The relative permeability of the oil is computed from the previous model and from the Stone II model [112]. The other physical properties of oil and gas such as the fugacities and the densities are computed using the Peng-Robinson cubic law and for the computation of the viscosities the Lohrenz-Bray-Clark model [83] is used. The properties of water (density and viscosity) are computed using data from [45].

Test of CO_2 injection in a three-component system. The first case is a miscible gas (CO_2) injection in a two-dimensional quarter of five-spot saturated with oil. The domain has a size of 100m in both directions and it is discretized using $|\mathtt{M}_h| = 20\times 20$ regular grid blocks. The reservoir model is homogeneous: the permeability is equal to 500 mD and the porosity is 0.3. The gas, composed only of CO_2, is injected with a constant rate that is equal to 80 m^3/day and the pressure at the producer is fixed to 55 bar. The temperature is assumed to be constant at 80^oC and the initial pressure is equal to 95 bar.

The total simulation time is 30 days, the initial time step is 0.05 day and the minimum and maximum time step are respectively 10^{-5} day and 20 days. The initial water saturation is given by $\mathsf{S}_W = 0.25$ and the oil saturation is equal to $1-\mathsf{S}_W$. The oil and gas phases are a mixture of three components $\mathcal{K} = \{\mathrm{C}_1, \mathrm{C}_6, \mathrm{CO}_2\}$ and the initial oil composition is given by C_1 (20%), C_6 (80%) and CO_2 (0%).

	Time steps	Number of iterations	Restarts
Newton-min	34	164	1
NPIPM	33	199	0

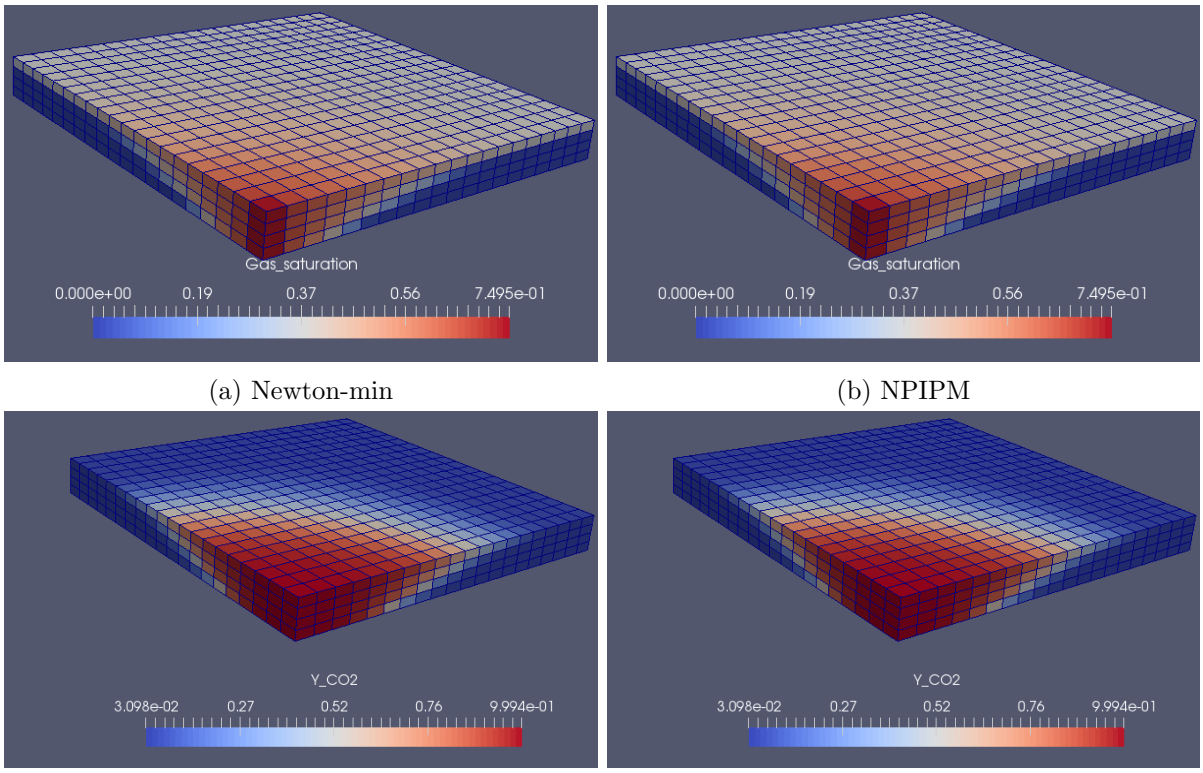
Table 6.1: Three-component system: numerical results of Newton-min method and NPIPM.



(a) Newton-min (b) NPIPM

(c) Newton-min (d) NPIPM

Figure 6.38: Gas saturation and partial fraction of component CO_2 in gas phase after 30 days: Newton-min method and NPIPM.



(a) Newton-min (b) NPIPM

(c) Newton-min (d) NPIPM

Figure 6.39: CO_2 injection in a seven-component system: gas saturation and CO_2 molar component in gas phase after 100 days.

Test of CO_2 injection in a seven-component system. The second case study still simulates a CO_2 injection in a three-dimensional quarter of five-spot saturated with oil. The reservoir size is $100 \times 100 \times 20$ m and we use $|\mathtt{M}_h| = 20 \times 20 \times 4$ grid blocks to discretize the reservoir model. The fluid is a seven-component mixture $\mathcal{K} = \{C_1N_2, C_2, CO_2, C_{46}, C_{712}, C_{1319}, C_{20}^+\}$, with the following initial composition : C_1N_2 (38.8209%), C_{23} (14.5821%), CO_2 (2.2685%), C_{46} (11.9334%), C_{712} (19.4598%), C_{1319} (8.7079%) and C_{20}^+ (4.2274%). The initial pressure and temperature are respectively 200 bar and 132.77°C. The CO_2 is injected with a fixed rate of 200 m^3/day and the production pressure is 150 bar.

	Time steps	Number of iterations	Restarts
Newton-min	43	175	0
NPIPM	43	179	0

Table 6.2: Seven-component system: numerical results of Newton-min method and NPIPM.

Results and discussions. Figures 6.38 and 6.39 display the spatial distribution of the gas saturation S_G and the partial fraction of CO_2 in the gas phase at the end of the simulation obtained by Newton-min and NPIPM algorithms. For each test, the two algorithms give the same physical results in terms of saturations, pressure and molar fractions. Tables 6.2 and 6.1 summarize the numerical results in terms of number of time steps, number of Newton iterations

and number of restarted time-steps for each case test. We observe that NPIPM converges at every time step and does not need to restart by dividing the time-step by 2. However, NPIPM takes a few more iterations. Further analysis shows that this is due to the choice of the initial point, since NPIPM needs to start at interior point whereas Newton-min method uses the state at the previous time-step as a starting point. For this realistic model, it was not easy to find a good strategy to go back inside this region without taking several iterations to converge. Other warm start strategies are under investigation.

While the four simplified models of §6.1 were simple enough to be implemented using Matlab, the multiphase compositional model (6.42) required partially existing subroutines for realistic physical closure laws and was therefore implemented in a heavier Fortran prototype. Due to a lack of time, we were unable to code the domain extension for Peng-Robinson's law in this prototype. This is the reason why we observed that when one of the phases (oil or gas) disappears, the two algorithms abruptly stopped because the cubic equation has a unique real root. Naturally, the extension procedure described in §3.3.3.2 should be added to overcome this issue.

Chapter 7

Conclusion and perspectives

Contents

7.1 Summary of key results

In response to the objectives stated in §1.1.3, we have conducted research works in two distinct but interrelated directions. The corresponding developments and contributions have given rise to several presentations at national and international conferences. We are now finalizing two scientific publications directly related to the thesis.

7.1.1 Theoretical aspects of the unified formulation

The first direction, presented in Part I (chapters §2–§3), is concerned with a better mathematical understanding of the unified formulation for the phase equilibrium problem. It was not initially planned as such, but became increasingly evident as our investigations progressed. Let us single out the most prominent results of this part.

When postulated as a founding model, the unified formulation is able to recover all the properties known to physicists on phase equilibrium. Indeed, the complementary equations do encapsulate the tangent plane criterion [Theorem 2.1], which cannot be derived from the natural variable formulation alone, without the help of some extra stability analysis. The unified formulation can also be regarded as a characterization of a constrained minimization problem [Theorems 2.3 and 2.4], in which the objective function is some modified Gibbs energy of the mixture. This characterization is slightly stronger than the usual KKT optimality conditions, insofar as it implies a choice (by a continuity principle) of one among an infinity of minimizers when a phase vanishes.

The possibility of assigning well-defined values to the extended fractions of an absent phase appears to be a theoretical strength of the unified formulation. Upon closer inspection, however, this possibility can only be achieved if the Gibbs functions meet some restrictive requirements

[Hypotheses 2.2]. In particular, they must be strictly convex over the whole domain of fractions. Shedding light on the favorable assumptions for the unified formulation in terms of Gibbs functions is perhaps the most consequential outcome of this part.

Unfortunately, Hypotheses 2.2 are not satisfied by all commonly used Gibbs functions. In these circumstances, the obligation of assigning well-defined values to the extended fractions of an absent phase becomes a weakness that dangerously jeopardizes the whole unified approach. This is especially true for Gibbs functions derived from cubic equations of states, for which they are not even defined on the whole domain of fractions. The extension procedures proposed in §3.3 is another substantial contribution, which is merely aimed at improving the "survival" chance of the unified formulation.

7.1.2 Practical algorithms for the numerical resolution

The second direction, presented in Part II (chapters §4–§6), deals with numerical methods. We reviewed and implemented several existing algorithms in the family of semismooth methods (Newton-min) and that of smoothing methods (θ-regularization, Mehrotra's interior-point), each having the line search option. These were tested on a hierarchy of models including not only stationary binary and ternary two-phase mixtures but also evolutionary models.

The need for a new method appeared very soon. On the ground of previous numerical results, we deemed interior-point methods to be a sound basis and built the NPIPM variant, in which the regularization parameter receives the same status as the unknowns. The superiority of NPIPM over Newton-min was overwhelming for small test cases. Therefore, it came as a disappointment that on the big test cases corresponding to a real flow model, NPIPM could not achieve the same astounding success. For some injection scenarios, it even performed rather poorly in comparison with Newton-min.

The difference with small test cases lies in the initial point. For these, it was easy to start with a good interior point. For the full flow model, the natural initial point is the state at the previous time-step. However, because of thermodynamic equilibrium, this state is always on the boundary of the interior region, and so far we do not have any good strategy to go back inside this region. Taking inspiration from existing work in optimization around warm start strategies, we have tried several perturbation techniques of the current state. However, we remain dissatisfied and believe that there is still even better to do.

7.2 Recommendations for future research

We outline two avenues to pursue this work. The first one is a technique that could lead to a good initial point for NPIPM on the full flow model. The second one is a further improvement that could be essential for large time-steps.

7.2.1 Warm start strategy

At least in the context of optimization problems, interior-point methods work by following some central path to an optimal solution. In practice, one needs to start the algorithm at a well-centered interior point. Any even small change in the objective function or in the constraints can cause the optimal solution from the previous version of the problem to move to the boundary and thus to be far from the central path for the new problem, so the algorithm takes several iterations to get back near the central path and to converge.

This well-known issue can be addressed by using smart perturbations of the current iterate. There are relatively few papers discussing these strategies for warm starting (see [64, 121], for instance). In some particular situations (linear programs), some of the strategies prove to be efficient and reduce significantly the number of iterations. In our problems, we still need to deeply understand what is a well-centered point and how to perturb the current state in order to get closer to such a point.

7.2.2 Continuation Newton for large time-steps

Another source of stiffness unfavorable to Newton convergence for evolutionary problems lies in a too large size of the time-step Δt, which is allowed by the backward Euler time-discretization. As a remedy, Younis et al. [122] suggested a continuation Newton procedure, in which the iterates match with intermediate times instead of being all targeted at $t^n + \Delta t$. This idea rests upon homotopic continuation, which is not easy not implement.

It turns out that we can use the same "trick" as in NPIPM to work out a more automatic version of this idea. We sketch out this prospect for the differential equation

$$\frac{\mathrm{d}X}{\mathrm{d}t} = -f(X), \tag{7.1}$$

where f is a continuously differentiable function. By the backward Euler scheme

$$\frac{X - X_b}{\Delta t} = -f(X), \tag{7.2}$$

where X is the value at the next time-step and X_b the state at the current time-step, we are led to solving the nonlinear system

$$X + \Delta t f(X) - X_b = 0. \tag{7.3}$$

When Δt is small, the nonlinearity is "mild." As Δt grows larger and larger, we may run into trouble. Instead of decreasing the time-step, we consider the equivalent system

$$X + (1 - \nu)\Delta t f(X) - X_b = 0, \tag{7.4a}$$
$$\nu = 0, \tag{7.4b}$$

where ν is considered as a new variable. Following the same lines as in §5.1.1, we arrive at another enlarged system, i.e.,

$$X + (1 - \nu)\Delta t f(X) - X_b = 0, \tag{7.5a}$$
$$\{\text{coupling terms}\} + \eta\nu + \nu^2 = 0, \tag{7.5b}$$

where $\eta > 0$ is a small parameter. The last system can be solved by the classical Newton method in the unknown (X, ν), starting from the initial point $(X_b, 1)$.

Bibliography

[1] A. Abadpour and M. Panfilov, *Method of negative saturations for modelling two-phase compositional flows with oversaturated zones*, Transp. Porous Media, 79 (2009), pp. 197–214, https://doi.org/10.1007/s11242-008-9310-0.

[2] V. Acary and B. Brogliato, *Numerical Methods for Nonsmooth Dynamical Systems: Applications in Mechanics and Electronics*, vol. 35 of Lecture Notes in Applied and Computational Mechanics, Springer, Berlin, 2008.

[3] M. Aganagić, *Newton's method for linear complementarity problems*, Math. Program., 28 (1984), pp. 349–362, https://doi.org/10.1007/BF02612339.

[4] O. Angelini, C. Chavant, E. Chénier, R. Eymard, and S. Granet, *Finite volume approximation of a diffusion-dissolution model and application to nuclear waste storage*, Math. Comput. Simul., 81 (2011), pp. 2001–2017, https://doi.org/10.1016/j.matcom.2010.12.016. MAMERN 2009: 3rd International Conference on Approximation Methods and Numerical Modeling in Environment and Natural Resources.

[5] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacif. J. Math., 16 (1966), pp. 1–3, https://doi.org/10.2140/pjm.1966.16.1.

[6] L. Asselineau, G. Bogdanic, and J. Vidal, *A versatile algorithm for calculating vapour-liquid equilibria*, Fluid Phase Equilibria, 3 (1979), pp. 273–290.

[7] A. Auslender, R. Cominetti, and M. Haddou, *Asymptotic analysis for penalty and barrier methods in convex and linear programming*, Math. Oper. Res., 22 (1997), pp. 43–62, https://doi.org/10.1287/moor.22.1.43.

[8] K. Aziz and A. Settari, *Petroleum Reservoir Simulation*, Applied Science Publishers, London, 1979.

[9] L. Beaude, K. Brenner, S. Lopez, R. Masson, and F. Smai, *Non-isothermal compositional liquid gas Darcy flow: formulation, soil-atmosphere boundary condition and application to high-energy geothermal simulations*, Comput. Geosci., 23 (2019), pp. 443–470, https://doi.org/10.1007/978-3-319-57394-6_34.

[10] A. Beck and M. Teboulle, *Smoothing and first order methods: A unified framework*, SIAM J. Optim., 22 (2012), pp. 557–580, https://doi.org/10.1137/100818327.

[11] I. Ben Gharbia, *Résolution de problèmes de complémentarité. : Application à un écoulement diphasique dans un milieu poreux*, PhD thesis, Université Paris Dauphine (Paris IX), December 2012, http://tel.archives-ouvertes.fr/tel-00776617.

[12] I. Ben Gharbia and É. Flauraud, *Study of compositional multiphase flow formulation using complementarity conditions*, Oil Gas Sci. Technol., 74 (2019), p. 43, https://doi.org/10.2516/ogst/2019012.

[13] I. Ben Gharbia, É. Flauraud, and A. Michel, *Study of compositional multi-phase flow formulations with cubic EOS*, in SPE Reservoir Simulation Symposium, 23-25 February, Houston, Texas, USA, vol. 2, 01 2015, pp. 1015–1025, https://doi.org/10.2118/173249-MS.

[14] I. Ben Gharbia and J. C. Gilbert, *Nonconvergence of the plain Newton-min algorithm for linear complementarity problems with a P-matrix*, Math. Prog., 134 (2012), pp. 349–364, https://doi.org/10.1007/s10107-010-0439-6.

[15] I. Ben Gharbia and J. C. Gilbert, *An algorithmic characterization of P-matricity*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 904–916, https://doi.org/10.1137/120883025.

[16] I. Ben Gharbia and J. C. Gilbert, *An algorithmic characterization of P-matricity II: adjustments, refinements, and validation*, SIAM J. Matrix Anal. Appl., 40 (2019), pp. 800–813, https://doi.org/10.1137/18M1168522.

[17] I. Ben Gharbia and J. Jaffré, *Gas phase appearance and disappearance as a problem with complementarity constraints*, Math. Comput. Simul., 99 (2014), pp. 28–36, https://doi.org/10.1016/j.matcom.2013.04.021.

[18] A. Ben-Tal and M. Teboulle, *A smoothing technique for nondifferentiable optimization problems*, in Optimization, S. Dolecki, ed., Berlin, Heidelberg, 1989, Springer Berlin Heidelberg, pp. 1–11, https://doi.org/10.1007/BFb0083582.

[19] M. Bergounioux and M. Haddou, *A new relaxation method for a discrete image restoration problem*, J. Convex Anal., 17 (2010), pp. 861–883, http://www.heldermann.de/JCA/JCA17/JCA173/jca17055.htm.

[20] S. C. Billups and K. G. Murty, *Complementarity problems*, J. Comput. Appl. Math., 124 (2000), pp. 303–318, https://doi.org/10.1016/S0377-0427(00)00432-5. Numerical Analysis 2000. Vol. IV: Optimization and Nonlinear Equations.

[21] F. Bonnans, *Optimisation continue: cours et problèmes corrigés*, Mathématiques appliquées pour le Master, Dunod, 2006.

[22] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical Optimization: Theoretical and Practical Aspects*, Universitext, Springer Verlag, Berlin, 2006.

[23] V. Bouvier, *Algorithmes de résolution compositionnnelle dans TACITE*, tech. report, Institut Français du Pétrole, 1995. 42485.

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*, Berichte über verteilte messysteme, Cambridge University Press, Cambridge, UK, 2004.

[25] H. Cao, *Development of techniques for general purpose simulators*, PhD thesis, Stanford University, 2002, https://pangea.stanford.edu/ERE/pdf/pereports/PhD/Cao02.pdf.

[26] G. Chavent and J. Jaffré, *Mathematical Models and Finite Elements for Reservoir Simulation: Single Phase, Multiphase and Multicomponent Flows through Porous Media*, vol. 17 of Studies in Mathematics and its Applications, North-Holland, Amsterdam, 1986.

[27] C. Chen and O. L. Mangasarian, *Smoothing methods for convex inequalities and linear complementarity problems*, Math. Program., 71 (1995), pp. 51–69, https://doi.org/10.1007/BF01592244.

[28] Z. Chen, G. Huan, and Y. Ma, *Computational methods for multiphase flows in porous media*, vol. 2 of Computational Science & Enginering, SIAM, Philadelphia, 2006.

[29] S.-J. Chung, *NP-completeness of the linear complementarity problem*, J. Optim. Theory Appl., 60 (1989), pp. 393–399, https://doi.org/10.1007/BF00940344.

[30] K. H. Coats, *An equation of state compositional model*, SPE Journal, 20 (1980), pp. 363–376, https://doi.org/10.2118/8284-PA. SPE-8284-PA.

[31] Ş. Cobzaş, R. Miculescu, and A. Nicolae, *Lipschitz Functions*, vol. 2241 of Lecture Notes in Mathematics, Springer, Cham, Switzerland, 2019, https://doi.org/10.1007/978-3-030-16489-8.

[32] A. Conn, N. Gould, and P. Toint, *Trust-Region Methods*, MPS-SIAM Series on Optimization 1, SIAM and MPS, Philadelphia, 2000.

[33] R. W. Cottle, *Nonlinear programs with positively bounded Jacobians*, SIAM J. Appl. Math., 14 (1966), pp. 147–158, https://doi.org/10.1137/0114012.

[34] R. W. Cottle and G. B. Dantzig, *Complementary pivot theory of mathematical programming*, Linear Alg. Appl., 1 (1968), pp. 103–125, https://doi.org/10.1016/0024-3795(68)90052-9.

[35] R. W. Cottle, J.-S. Pang, and R. E. Stone, *The linear complementarity problem*, vol. 60 of Classics in Applied Mathematics, SIAM, Philadelphia, 2009.

[36] G. E. Coxson, *The P-matrix problem is co-NP-complete*, Math. Program., 64 (1994), pp. 173–178, https://doi.org/10.1007/BF01582570.

[37] J. C. 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, https://doi.org/10.1007/s10092-004-0092-7.

[38] T. De Luca, F. Facchinei, and C. Kanzow, *A theoretical and numerical comparison of some semismooth algorithms for complementarity problems*, Comput. Optim. Appl., 16 (2000), pp. 173–205, https://doi.org/10.1023/A:1008705425484.

[39] U. K. Deiters and T. Kraska, *High-pressure Fluid Phase Equilibria: Phenomenology and Computation*, vol. 2 of Supercritical Fluid Science and Technology, Elsevier, Amsterdam, 2012, http://store.elsevier.com/High-Pressure-Fluid-Phase-Equilibria/isbn-9780444563545/.

[40] D. den Hertog, C. Roos, and T. Terlaky, *The linear complimentarity problem, sufficient matrices, and the criss-cross method*, Linear Alg. Appl., 187 (1993), pp. 1–14, https://doi.org/10.1016/0024-3795(93)90124-7.

[41] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, vol. 16 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[42] P. Deuflhard, *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*, vol. 35 of Springer Series in Computational Mathematics, Springer, Berlin, 2011.

[43] I. I. Dikin, *Iterative solution of linear and quadratic programming*, Dokl. Akad. Nauk SSSR, 174 (1967), pp. 747–748, http://mi.mathnet.ru/eng/dan33112.

[44] J.-P. Dussault, M. Frappier, and J. C. Gilbert, *Polyhedral Newton-min algorithms for complementarity problems*, research report, Inria Paris ; Université de Sherbrooke (Québec, Canada), Oct 2019, https://hal.archives-ouvertes.fr/hal-02306526.

[45] S. Ernst, *Properties of water and steam in SI-units*, Springer-Verlag, Berlin, 1969.

[46] F. Facchinei and J. S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems, I*, Springer Series in Operations Research, Springer, New York, 2003.

[47] F. Facchinei and J. S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems, II*, Springer Series in Operations Research, Springer, New York, 2003.

[48] F. Facchinei and J. Soares, *A new merit function for nonlinear complementarity problems and a related algorithm*, SIAM J. Optim., 7 (1997), pp. 225–247, https://doi.org/10.1137/S1052623494279110.

[49] M. Fiedler and V. Pták, *Some generalizations of positive definiteness and monotonicity*, Numer. Math., 9 (1966), pp. 163–172, https://doi.org/10.1007/BF02166034.

[50] A. Fischer, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284, https://doi.org/10.1080/02331939208843795.

[51] A. Fischer and C. Kanzow, *On finite termination of an iterative method for linear complementarity problems*, Math. Program., 74 (1996), pp. 279–292, https://doi.org/10.1007/BF02592200.

[52] A. Galántai, *The theory of Newton's method*, J. Comput. Appl. Math., 124 (2000), pp. 25–44, https://doi.org/10.1016/S0377-0427(00)00435-0. Numerical Analysis 2000. Vol. IV: Optimization and Nonlinear Equations.

[53] R. Glowinski, J. L. Lions, and R. Trémolières, *Numerical Analysis of Variational Inequalities*, vol. 8 of Studies in Mathematics and its Applications, North-Holland Publishing Company, Amsterdam, 1981.

[54] J. Gondzio, *Interior point methods 25 years later*, Eur. J. Oper. Res., 218 (2012), pp. 587–601, https://doi.org/10.1016/j.ejor.2011.09.017.

[55] M. Haddou, *A new class of smoothing methods for mathematical programs with equilibrium constraints*, Pacif. J. Optim., 5 (2009), pp. 86–96.

[56] M. Haddou and P. Maheux, *Smoothing methods for nonlinear complementarity problems*, J. Optim. Theory Appl., 160 (2014), pp. 711–729, https://doi.org/10.1007/s10957-013-0398-1.

[57] M. Haddou, T. Migot, and J. Omer, *A generalized direction in interior point method for monotone linear complementarity problems*, Optim. Lett., (2018), https://doi.org/10.1007/s11590-018-1241-2.

[58] M. Hamani, *Méthodes numériques pour la résolution de systèmes d'équations algébriques contenant des équations de complémentarité*, master's thesis, Sup Galilée, 2017.

[59] P. Harker and J.-S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications.*, Math. Program., 48 (1990), pp. 161–220, https://doi.org/10.1007/BF01582255.

[60] P. Hartman and G. Stampacchia, *On some non-linear elliptic differential-functional equations*, Acta Mathematica, 115 (1966), pp. 271–310, https://doi.org/10.1007/BF02392210.

[61] R. A. Heidemann, *Computation of high pressure phase equilibria*, Fluid Phase Equilibria, 14 (1983), pp. 55–78, https://doi.org/10.1016/0378-3812(83)80115-0.

[62] W. Henry and J. Banks, *III. Experiments on the quantity of gases absorbed by water, at different temperatures, and under different pressures*, Phil. Trans. Royal Soc. London, 93 (1803), pp. 29–274, https://doi.org/10.1098/rstl.1803.0004.

[63] A. F. Izmailov and M. V. Solodov, *Newton-Type Methods for Optimization and Variational Problems*, Springer Series in Operations Research and Financial Engineering, Springer, Cham, Switzerland, 2014, https://doi.org/10.1007/978-3-319-04247-3.

[64] E. John and E. A. Yildirim, *Implementation of warm-start strategies in interior-point methods for linear programming in fixed dimension*, Comput. Optim. Appl., 41 (2008), pp. 151–183, https://doi.org/10.1007/s10589-007-9096-y.

[65] C. Kanzow, *Inexact semismooth Newton methods for large-scale complementarity problems*, Optim. Meth. Software, 19 (2004), pp. 309–325, https://doi.org/10.1080/10556780310001636369.

[66] S. Karamardian, *Generalized complementarity problem*, J. Optim. Theory Appl., 8 (1971), pp. 161–168, https://doi.org/10.1007/BF00932464.

[67] N. Karmarkar, *A new polynomial-time algorithm for linear programming*, in Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing, STOC '84, New York, USA, 1984, Association for Computing Machinery, pp. 302–311, https://doi.org/10.1145/800057.808695.

[68] C. T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, vol. 16 of Frontiers in Applied Mathematics, SIAM, Philadelphia, 1995.

[69] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, vol. 31 of Classics in Applied Mathematics, SIAM, Philadelphia, 2000.

[70] M. Kojima, N. Megiddo, T. Noma, and A. Yoshise, *A unified approach to interior point algorithms for linear complementarity problems: A summary*, Operations Research Letters, 10 (1991), pp. 247–254, https://doi.org/10.1016/0167-6377(91)90010-M.

[71] M. Kojima and S. Shindo, *Extension of Newton and quasi-Newton methods to systems of PC^1 equations*, J. Oper. Res. Soc. Japan, 29 (1986), pp. 352–375, https://doi.org/10.15807/jorsj.29.352.

[72] S. Kräutle, *The semismooth Newton method for multicomponent reactive transport with minerals*, Adv. Water Res., 34 (2011), pp. 137–151, https://doi.org/10.1016/j.advwatres.2010.10.004.

[73] B. Kummer, *Newton's method for non-differentiable functions*, Adv. Math. Optim., 45 (1988), pp. 114–125.

[74] T. Y. Kwak and G. A. Mansoori, *Van der Waals mixing rules for cubic equations of state. Applications for supercritical fluid extraction modelling*, Chem. Eng. Sci., 41 (1986), pp. 1303–1309, https://doi.org/10.1016/0009-2509(86)87103-2.

[75] V. Lachet and V. Ruffier-Meray, *Documentation du module thermodynamique de TACITE*, tech. report, Institut Français du Pétrole, 1999. 45311.

[76] T. C. Lai Nguyen, *Analysis of a nonlinear algebraic system arising in phase equilibria problems*, master's thesis, INSA Rennes, 2018.

[77] A. Lauser, *Theory and Numerical Applications of Compositional Multi-Phase Flow in Porous Media*, PhD thesis, Universität Stuttgart, 2013.

[78] A. Lauser, C. Hager, R. Helmig, and B. Wohlmuth, *A new approach for phase transitions in miscible multi-phase flow in porous media*, Adv. Water Res., 34 (2011), pp. 957–966, https://doi.org/10.1016/j.advwatres.2011.04.021.

[79] Y. Le Hénaff, *Convexity of Gibbs function*, master's thesis, INSA Rennes, 2018.

[80] S. Le Vent, *A summary of the properties of van der Waals fluids*, Int. J. Mech. Engrg Edu., 29 (2001), pp. 257–277, https://doi.org/10.7227/IJMEE.29.3.8.

[81] C. E. Lemke, *Bimatrix equilibrium points and mathematical programming*, Manage. Sci., 11 (1965), pp. 681–689, https://doi.org/10.1287/mnsc.11.7.681.

[82] C. E. Lemke and J. T. Howson, *Equilibrium points of bimatrix games*, SIAM J. Appl. Math., 12 (1964), pp. 413–423, https://doi.org/10.1137/0112033.

[83] J. Lohrenz, B. G. Bray, and C. R. Clark, *Calculating viscosities of reservoir fluids from their compositions*, J. Petrol. Technology, (1964), https://doi.org/10.2118/915-PA.

[84] I. Lusetti, *Numerical methods for compositional multiphase flow models with cubic EOS*, tech. report, IFPEN, 2016.

[85] O. L. Mangasarian, *Equivalence of the complementarity problem to a system of nonlinear equations*, SIAM J. Appl. Math., 31 (1976), pp. 89–92, https://doi.org/10.1137/0131009.

[86] R. Masson, L. Trenty, and Y. Zhang, *Formulations of two phase liquid gas compositional Darcy flows with phase transitions*, Int. J. Finite Vol., 11 (2014), pp. 1–34, http://ijfv.math.cnrs.fr/IMG/pdf/gazliqcomp-ijfv-1.pdf.

[87] R. Masson, L. Trenty, and Y. Zhang, *Coupling compositional liquid gas Darcy and free gas flows at porous and free-flow domains interface*, J. Comput. Phys., 321 (2016), pp. 708–728, https://doi.org/10.1016/j.jcp.2016.06.003.

[88] S. Mehrotra, *On the implementation of a primal-dual interior point method*, SIAM J. Optim., 2 (1992), pp. 575–601, https://doi.org/10.1137/0802028.

[89] M. L. Michelsen, *The isothermal flash problem. Part I. Stability*, Fluid Phase Equilibria, 9 (1982), pp. 1–19, https://doi.org/10.1016/0378-3812(82)85001-2.

[90] M. L. Michelsen, *The isothermal flash problem. Part II. Phase-split calculation*, Fluid Phase Equilibria, 9 (1982), pp. 21–40, https://doi.org/10.1016/0378-3812(82)85002-4.

[91] M. L. Michelsen and J. M. Mollerup, *Thermodynamic Models: Fundamentals & Computational Aspects*, Tie-Line Publications, Holte, 2007.

[92] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972, https://doi.org/10.1137/0315061.

[93] T. Migot, *Contributions aux méthodes numériques pour les problèmes de complémentarité et problèmes d'optimisation sous contraintes de complémentarité*, PhD thesis, INSA Rennes, 2017.

[94] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer, New York, 2006.

[95] M. d. M. Olaya, J. A. Reyes-Labarta, M. D. Serrano, and A. Marcilla, *Vapor-liquid equilibria using the Gibbs energy and the common tangent plane criterion*, Chem. Eng. Edu., 44 (2010), p. 236, https://eric.ed.gov/?id=EJ935045.

[96] H. Orbey and S. I. Sandler, *Modeling Vapor-Liquid Equilibria: Cubic Equations of State and Their Mixing Rules*, Cambridge Series in Chemical Engineering, Cambridge University Press, 1998.

[97] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, vol. 30 of Classics in Applied Mathematics, SIAM, Philadelphia, 2000.

[98] J.-S. Pang, *Newton's method for B-differentiable equations*, Math. Oper. Res., 15 (1990), pp. 311–341, https://doi.org/10.1287/moor.15.2.311.

[99] D.-Y. Peng and D. B. Robinson, *A new two-constant equation of state*, Ind. Eng. Chem. Fundam., 15 (1976), pp. 59–64, https://doi.org/10.1021/i160057a011.

[100] R. H. Perry and D. W. Green, *Perry's Chemical Engineers' Handbook*, McGraw-Hill chemical engineering series, McGraw-Hill, 1999.

[101] N. Peton, *Comparaison de plusieurs formulations pour les écoulements multiphasiques et compositionnels en milieu poreux*, tech. report, IFPEN, 2015.

[102] N. Peton, *Étude et simulation d'un modèle stratigraphique advecto-diffusif non-linéaire avec frontières mobiles*, PhD thesis, Université Paris-Saclay, 2018, http://www.theses.fr/2018SACLC058.

[103] N. Peton, C. Cancès, D. Granjeon, Q.-H. Tran, and S. Wolf, *Numerical scheme for a water flow-driven forward stratigraphic model*, Comput. Geosci., 24 (2020), pp. 37–60, https://doi.org/10.1007/s10596-019-09893-w.

[104] J. M. Prausnitz, R. N. Lichtenthaler, and E. G. de Azevedo, *Molecular Thermodynamics of Fluid-Phase Equilibria*, Prentice-Hall International Series in the Physical and Chemical Engineering Sciences, Pearson Education, 1998.

[105] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Program., 58 (1993), pp. 353–367, https://doi.org//10.1007/BF01581275.

[106] H. H. Rachford and J. D. Rice, *Procedure for use of electronic digital computers in calculating flash vaporization hydrocarbon equilibrium*, J. Petrol. Technol., 4 (1952), p. 19, https://doi.org/10.2118/952327-G.

[107] D. Ralph, *Global convergence of damped Newton's method for nonsmooth equations via the path search*, Math. Oper. Res., 19 (1994), pp. 352–389, https://doi.org/10.1287/moor.19.2.352.

[108] O. Redlich and J. N. S. Kwong, *On the thermodynamics of solutions. v. an equation of state. fugacities of gaseous solutions*, Chem. Rev., 44 (1949), pp. 233–244, https://doi.org/10.1021/cr60137a013. PMID: 18125401.

[109] R. T. Rockafellar, *Convex Analysis*, vol. 28 of Princeton Landmarks in Mathematics and Physics, Princeton University Press, Princeton, New Jersey, 1970.

[110] M. Seetharama Gowda, *Inverse and implicit function theorems for H-differentiable and semismooth functions*, Optim. Meth. Software, 19 (2004), pp. 443–461, https://doi.org/10.1080/10556780410001697668.

[111] G. Soave, *Equilibrium constants from a modified Redlich-Kwong equation of state*, Chem. Eng. Sci., 27 (1972), pp. 1197–1203.

[112] H. L. Stone, *Estimation of three-phase relative permeability and residual oil data*, J. Canad. Petrol. Technology, (1973), https://doi.org/10.2118/73-04-06.

[113] R. A. Tapia, Y. Zhang, M. Saltzman, and A. Weiser, *The Mehrotra predictor-corrector interior-point method as a perturbed composite Newton method*, SIAM J. Optim., 6 (1996), pp. 47–56, https://doi.org/10.1137/0806004, https://arxiv.org/abs/http://dx.doi.org/10.1137/0806004.

[114] J. D. van der Waals, *On the continuity of the gas and liquid state*, PhD thesis, Universiteit Leiden, 1873.

[115] J. Vidal, *Thermodynamics. Applications in Chemical Engineering and The Petroleum Industry*, Institut Français du Pétrole Publications, Technip, Paris, 2003.

[116] D. V. Voskov and H. A. Tchelepi, *Comparison of nonlinear formulations for two-phase multi-component EOS based simulation*, J. Petrol. Sci. Engrg, 82–83 (2012), pp. 101–111, https://doi.org/10.1016/j.petrol.2011.10.012.

[117] C. H. Whitson and M. L. Michelsen, *The negative flash*, Fluid Phase Equilibria, 53 (1989), pp. 51–71, https://doi.org/10.1016/0378-3812(89)80072-X.

[118] M. H. Wright, *The interior-point revolution in optimization: history, recent developments, and lasting consequences*, Bull. Amer. Math. Soc., 42 (2005), pp. 39–56, https://doi.org/10.1090/S0273-0979-04-01040-7.

[119] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.

[120] T. Yamamoto, *Historical developments in convergence analysis for Newton's and Newton-like methods*, J. Comput. Appl. Math., 124 (2000), pp. 1–23, https://doi.org/10.1016/S0377-0427(00)00417-9. Numerical Analysis 2000. Vol. IV: Optimization and Nonlinear Equations.

[121] E. A. Yildirim and S. J. Wright, *Warm-start strategies in interior-point methods for linear programming*, SIAM J. Optim., 12 (2002), pp. 782–810, https://doi.org/10.1137/S1052623400369235.

[122] R. Younis, H. A. Tchelepi, and K. Aziz, *Adaptively localized continuation-Newton method–nonlinear solvers that converge all the time*, SPE Journal, 15 (2010), pp. 526–544, https://doi.org/10.2118/119147-PA. SPE-119147-PA.

[123] D. Zhang and Y. Zhang, *A Mehrotra-type predictor-corrector algorithm with polynomiality and Q-subquadratic convergence*, Ann. Oper. Res., 62 (1996), pp. 131–150, https://doi.org/10.1007/BF02206814.

[124] Y. Zhang and D. Zhang, *On polynomiality of the Mehrotra-type predictor-corrector interior-point algorithms*, Math. Program., 68 (1995), pp. 303–318, https://doi.org/10.1007/BF01585769.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Résolution numérique des systèmes algébriques contenant des équations de complémentarité. Application à la thermodynamique des mélanges polyphasiques compositionnels

Mots clés : condition de complémentarité, méthodes de Newton et Newton-min, méthode des points intérieurs, problème de l'équilibre des phases, formulation unifiée, écoulement polyphasique compositionnel

Résumé : Dans les simulateurs de réservoir, la prise en compte des lois d'équilibre thermodynamique pour les mélanges polyphasiques d'hydrocarbures est une partie délicate. La difficulté réside dans la gestion de l'apparition et de la disparition des phases pour différents constituants. L'approche dynamique traditionnelle, dite de *variable switching*, consiste à ne garder que les inconnues des phases présentes et les équations relatives à celles-ci. Elle est lourde et coûteuse, dans la mesure où le « switching » se produit constamment, même d'une itération de Newton à l'autre.
Une approche alternative, appelée *formulation unifiée*, permet de maintenir au cours des calculs un jeu fixe d'inconnues et d'équations. Sur le plan théorique, c'est un progrès important. Sur le plan pratique, comme la nouvelle formulation fait intervenir des équations de *complémentarité* qui sont non-lisses, on est obligé après discrétisation d'avoir recours à la méthode semi-lisse Newton-min, au comportement souvent pathologique.
Pour aller au bout de l'intérêt de la démarche unifiée, cette thèse a pour objectif de lever cet obstacle numérique en élaborant des algorithmes de résolution mieux adaptés, avec une meilleure convergence. Notre méthodologie consiste à s'inspirer des méthodes qui ont fait leur preuve en optimisation sous contraintes et à les transposer aux systèmes généraux. Cela conduit aux méthodes de points intérieurs, dont nous proposons une version *non-paramétrique* appelée NPIPM, avec des résultats supérieurs à Newton-min.
Une autre contribution de ce travail doctoral est la compréhension et la résolution (partielle) d'une autre obstruction au bon fonctionnement de la formulation unifiée, jusque-là non identifiée dans la littérature. Il s'agit de la limitation du domaine de définition des fonctions de Gibbs associées aux lois d'état cubiques. Pour remédier à l'éventuelle non-existence de solution du système, nous préconisons un prolongement naturel des fonctions de Gibbs.

Title: Numerical resolution of algebraic systems with complementarity conditions. Application to the thermodynamics of compositional multiphase mixtures

Keywords: complementarity condition, Newton's and Newton-min method, interior-point method, phase equilibrium problem, unified formulation, multiphase multicomponent flows

Abstract: In reservoir simulators, it is usually delicate to take into account the laws of thermodynamic equilibrium for multiphase hydrocarbon mixtures. The difficulty lies in handling the appearance and disappearance of phases for different species. The traditional dynamic approach, known as *variable switching*, consists in considering only the unknowns and equations of the present phases. It is cumbersome and costly, insofar as "switching" occurs constantly, even from one Newton iteration to another.
An alternative approach, called *unified formulation*, allows a fixed set of unknowns and equations to be maintained during the calculations. From a theoretical point of view, this is an major advance. On the practical level, because of the nonsmoothness of the *complementarity* conditions involved in the new formulation, the discretized equations have to be solved by the semi-smooth Newton-min method, whose behavior is often pathological.
In order to fully exploit the interest of the unified approach, this thesis aims at circumventing this numerical obstacle by means of more robust resolution algorithms, with a better convergence. To this end, we draw inspiration from the methods that have proven their worth in constrained optimization and we try to transpose them to general systems. This gives rise to interior-point methods, of which we propose a *nonparametric* version called NPIPM. The results appear to be superior to those of Newton-min.
Another contribution of this doctoral work is the understanding and (partial) resolution of another obstruction to the proper functioning of the unified formulation, hitherto unidentified in the literature. This is the limitation of the domain of definition of Gibbs' functions associated with cubic equations of state. To remedy the possible non-existence of a system solution, we advocate a natural extension of Gibbs' functions.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France