# Méthodes numériques avancées pour les problèmes à forte raideur en transport réactif

Maxime Jonval

**Université de Lille**
**IFP Energies nouvelles**

École doctorale **MADIS**

Unité de recherche **Centre Inria de l'Université de Lille**

Thèse présentée par

**Maxime Jonval**

Soutenue le **18 novembre 2024**

En vue de l'obtention du grade de docteur de l'Université de Lille

Discipline **Mathématiques**

# Méthodes numériques avancées pour les problèmes à forte raideur en transport réactif

| | | |
|---|---|---|
| **Équipe d'encadrement** | Clément Cancès | directeur |
| | Quang-Huy Tran | co-directeur |
| | Thibault Faney | co-encadrant |
| | Ibithel Ben Gharbia | co-encadrante |

**Composition du jury**

| | | | |
|---|---|---|---|
| *Rapporteurs* | Brahim Amaziane | MCF HDR à l'Université de Pau et des Pays de l'Adour | |
| | Pascal Omnes | directeur de recherche au CEA Saclay | |
| *Examinateurs* | Carole Rosier | professeure à l'Université du Littoral Côte d'Opale | présidente du jury |
| | Roland Masson | professeur à l'Université Côte d'Azur | |
| | Irina Sin | enseignante-chercheuse aux Mines Paris - PSL | |
| *Invités* | Thibault Faney | ingénieur de recherche à IFPEN | |
| | Ibithel Ben Gharbia | ingénieure de recherche à IFPEN | |
| *Directeurs de thèse* | Clément Cancès | directeur de recherche à Inria | |
| | Quang-Huy Tran | ingénieur de recherche HDR à IFPEN | |

**Université de Lille**
**IFP Energies nouvelles**

École doctorale **MADIS**

Unité de recherche **Centre Inria de l'Université de Lille**

Thèse présentée par

**Maxime Jonval**

Soutenue le **18 novembre 2024**

En vue de l'obtention du grade de docteur de l'Université de Lille

Discipline **Mathématiques**

# Méthodes numériques avancées pour les problèmes à forte raideur en transport réactif

| | | |
|---|---|---|
| **Équipe d'encadrement** | Clément Cancès | directeur |
| | Quang-Huy Tran | co-directeur |
| | Thibault Faney | co-encadrant |
| | Ibithel Ben Gharbia | co-encadrante |

**Composition du jury**

| | | | |
|---|---|---|---|
| *Rapporteurs* | Brahim Amaziane | MCF HDR à l'Université de Pau et des Pays de l'Adour | |
| | Pascal Omnes | directeur de recherche au CEA Saclay | |
| *Examinateurs* | Carole Rosier | professeure à l'Université du Littoral Côte d'Opale | présidente du jury |
| | Roland Masson | professeur à l'Université Côte d'Azur | |
| | Irina Sin | enseignante-chercheuse aux Mines Paris - PSL | |
| *Invités* | Thibault Faney | ingénieur de recherche à IFPEN | |
| | Ibithel Ben Gharbia | ingénieure de recherche à IFPEN | |
| *Directeurs de thèse* | Clément Cancès | directeur de recherche à Inria | |
| | Quang-Huy Tran | ingénieur de recherche HDR à IFPEN | |

**Université de Lille**
**IFP Energies nouvelles**

Doctoral School **MADIS**

Research Unit **Centre Inria de l'Université de Lille**

Thesis defended by

**Maxime Jonval**

Defended on **November 18, 2024**

In partial fulfillment of the requirements for degree of Doctor of Philosophy at Université de Lille

Academic Field **Mathematics**

# Advanced numerical methods for stiff problems in the context of reactive transport

| | | |
|---|---|---|
| **Prepared under the guidance of** | Clément Cancès | Supervisor |
| | Quang-Huy Tran | Co-Supervisor |
| | Thibault Faney | Co-Advisor |
| | Ibithel Ben Gharbia | Co-Advisor |

**Committee members**

| | | | |
|---|---|---|---|
| *Referees* | Brahim Amaziane | HDR Associate Professor at Université de Pau et des Pays de l'Adour | |
| | Pascal Omnes | Senior Researcher at CEA Saclay | |
| *Examiners* | Carole Rosier | Professor at Université du Littoral Côte d'Opale | Committee President |
| | Roland Masson | Professor at Université Côte d'Azur | |
| | Irina Sin | Assistant Professor at Mines Paris - PSL | |
| *Guests* | Thibault Faney | Research Engineer at IFPEN | |
| | Ibithel Ben Gharbia | Research Engineer at IFPEN | |
| *Supervisors* | Clément Cancès | Senior Researcher at Inria | |
| | Quang-Huy Tran | HDR Research Engineer at IFPEN | |

**Méthodes numériques avancées pour les problèmes à forte raideur en transport réactif**

**Résumé**

La simulation du transport réactif en milieu poreux est un enjeu majeur pour la transition énergétique, avec des applications pour la séquestration du CO2, la géothermie et le stockage d'hydrogène. La performance des codes de transport réactif est aujourd'hui fortement limitée par les difficultés numériques liées à la modélisation chimique. Ces difficultés se rattachent à un problème de raideur des équations à résoudre. Dans cette thèse, on s'intéresse aux réactions d'équilibre dans le cas d'une seule phase aqueuse et dans celui d'un mélange multiphasique pouvant contenir des phases aqueuses, gazeuses et minérales. Ces deux problèmes mènent à la résolution d'équations algébriques non linéaires par la méthode de Newton.

Les réactions d'équilibre monophasiques considérées recouvrent une large gamme de valeurs du domaine de définition des inconnues, avec des plages de fonctionnement et des ordres de grandeurs tout à fait différents. Cela pose des problèmes lors de la résolution par la méthode de Newton. Tantôt il est préférable de choisir comme inconnues les nombres de moles des espèces, tantôt il est souhaitable de prendre leurs logarithmes (ou potentiels chimiques), sous peine d'accroître le nombre d'itérations nécessaires voire de faire diverger la méthode de Newton. Les réactions d'équilibre multiphasiques, en plus de contenir les difficultés précédentes, peuvent contenir un nombre important de phase potentiellement présentes. La présence ou l'absence d'une phase est modélisée par un problème de complémentarité. La non différentiabilité des conditions de complémentarité met en défaut la méthode de Newton dans des cas difficiles mais réalistes.

Dans cette thèse, nous avons amélioré la robustesse de la méthode de Newton pour le cas monophasique grâce à une reformulation du système basé sur les fractions molaires et à l'utilisation des méthodes de paramétrage et de représentation Cartésienne. La méthode de paramétrage permet, grâce à une variable fictive, de rendre automatique le choix d'une résolution en fraction molaires ou en potentiels chimiques. La représentation Cartésienne, quant à elle, considère un système élargi où fractions molaires et potentiels chimiques sont des inconnues et où leur relation est relâchée et intégrée aux équations sous la forme d'une fonction non linéaire vérifiant ce lien à convergence. Nous avons démontré que la méthode de Newton appliquée à ces formulations vérifie la propriété de convergence quadratique locale. Les résultats numériques obtenus démontrent une robustesse accrue de nos méthodes comparées à la littérature.

Pour le cas multiphasique, nous avons établi une nouvelle modélisation du problème d'équilibres chimiques. Le système obtenu permet de considérer l'absence ou la présence des phases dans un cadre rigoureux et unifié. Cette unification fait référence à la notion de fractions molaires étendues, il s'agit d'une extension de la notion de fractions molaires aux phases absentes et permet de traiter indifféremment les phases grâce à une condition de complémentarité. Nous avons appliqué les méthodes de paramétrage et de représentation Cartésienne à ce problème ainsi qu'une nouvelle méthode de paramétrage de la complémentarité. Cette dernière a été comparée à différentes approches de la littérature pour le traitement de la complémentarité. Les expériences numériques obtenues ont montré une nette amélioration en termes de robustesse et de rapidité par rapport à l'état de l'art.

**Mots clés :** équilibres chimiques, méthodes de Newton, paramétrage, représentation cartésienne

**Centre Inria de l'Université de Lille**
Parc scientifique de la Haute-Borne – 40, avenue Halley – 59650 Villeneuve d'Ascq – France

**Advanced numerical methods for stiff problems in the context of reactive transport**

**Abstract**

The simulation of reactive transport in porous media is a major challenge for the energy transition, with applications in CO2 sequestration, geothermal energy and hydrogen storage. Today, the performance of reactive transport codes is severely limited by the numerical difficulties associated with chemical modeling. These difficulties are related to the stiffness of the equations to be solved. In this thesis, we focus on equilibrium reactions in the case of a single aqueous phase and in the case of a multiphase mixture that may contain aqueous, gaseous and mineral phases. Both problems lead to the solution of nonlinear algebraic equations using Newton's method.

The single-phase equilibrium reactions considered cover a wide range of values in the domain of definition of the unknowns, with different operating ranges and orders of magnitude. This poses problems when solving by Newton's method. Sometimes it is preferable to choose the mole numbers of the species as unknowns, sometimes it is desirable to take their logarithms (or chemical potentials), which may increase the number of iterations or cause Newton's method to diverge. Multiphase equilibrium reactions, in addition to containing the above difficulties, may also include a significant number of potentially present phases. The presence or absence of a phase is modeled by a complementarity problem. The non-differentiability of complementarity conditions causes Newton's method to fail in difficult but realistic cases.

In this thesis, we have improved the robustness of Newton's method for the single-phase case by reformulating the system based on mole fractions and using the parameterization and Cartesian representation methods. The parametrization method uses a fictitious variable to automatically select a resolution in mole fractions or chemical potentials. The Cartesian representation, on the other hand, considers an extended system in which mole fractions and chemical potentials are unknowns and their relationship is relaxed and integrated into the equations in the form of a non-linear function verifying this link only at convergence. We have demonstrated that Newton's method applied to these formulations verifies the property of local quadratic convergence. The numerical results obtained demonstrate the increased robustness of our methods compared with the literature.

For the multiphase case, we have established a new model of the chemical equilibrium problem. The resulting system allows us to consider the absence or presence of phases within a rigorous, unified framework. This unification refers to the notion of extended mole fractions, which is an extension of the notion of mole fractions to absent phases, and enables phases to be treated indifferently thanks to a complementarity condition. We have applied Cartesian parametrization and representation methods to this problem, as well as a new approach called the complementarity parametrization method. This method has been compared with various approaches to complementarity treatment from the literature. The numerical experiments obtained has shown a clear improvement in terms of robustness and speed compared with the state of the art.

**Keywords:** chemical equilibria, Newton's method, parametrization, Cartesian representation

# Remerciements

Je remercie chaleureusement Brahim Amaziane et Pascal Omnes d'avoir rapporté ma thèse. Votre retour sur mon travail est très précieux et j'ai apprécié chacune de vos remarques. Merci à Carole Rosier, Roland Masson et Irina Sin d'avoir accepté d'être examinateurs.

Un immense merci Clément pour m'avoir guidé tout au long de cette thèse. Ton expertise et ton expérience m'ont beaucoup apporté. Un grand merci Huy pour ton excellente co-direction, tes remarques et suggestions ainsi que ta clairvoyance ont été fort utiles au bon déroulement de cette aventure. Merci Thibault pour ta bienveillance et ton expertise. Merci Ibithel pour ta précieuse aide.

Je remercie Aurore Dalle, Isabelle Aslani, Sylvie Hoguet, Anne Danre, Aurore Smets et toutes autres personnes ayant agit dans l'ombre et qui ont veillé au bon déroulement des nombreuses missions et tâches administratives inhérentes à l'activité de recherche. Je remercie également Angélique Cosaert et toute l'équipe de l'agos pour les nombreuses activités ludiques proposées.

Merci aux nombreux enseignants-chercheurs qui, sans le savoir, m'ont donné goût aux maths et à la recherche : Bernhard Beckermann, Caterina Calgaro, Gwenaëlle Castellan, Claire Chainais, Emmanuel Creusé, André De Laire, Sandra Delaunay, Raphael Freitas, Ana Matos, Thomas Rey et Antoine Touzé.

Merci à tous les membres de l'équipe Rapsodi et Paradyse d'Inria. Merci Julien pour ces deux années de convivialité dans ce bureau devenu réserve alimentaire en cas de coup dur. Merci Amélie pour ces nombreuses parties de démineur, puisse mon record tenir encore longtemps. Merci Tino, Jules et Quentin d'avoir égayer la vie du village. J'aimerais également remercier tous les doctorants d'IFPEN avec qui j'ai pu sympathiser durant ces trois années et en particulier Julie à qui je souhaite bon courage pour la suite de la thèse.

Je remercie toute ma famille et belle-famille pour leur soutien et leurs nombreux encouragements. Merci à mon père, Patricia, Christelle et Robin d'être venus à ma soutenance. Merci à ma mère d'avoir toujours cru en moi et de m'avoir laissé tracer mon chemin. Merci à ma mamie pour sa gentillesse à toute épreuve. Une pensée pour mon papy qui, je l'espère, est fier de moi.

Merci à chacun de mes amis de la licence de maths pour tous ces moments de partage. J'espère que nos vacances annuelles continueront d'être une tradition. Merci Matthieu pour ces aventures à la montagne toujours suivies d'un bon BK. Merci Marion pour ces nombreux footings à la citadelle. Merci Louis, Julie, Franco, Sandro et Alexandre de vous être libérés pour venir à ma soutenance.

Merci Valentin pour tes encouragements et ton authentique amitié. Merci Maydine pour ton soutien, tes conseils et ton amitié sans faille. J'ai hâte qu'on se retrouve tous les trois sur le départ de la Saintélyon.

Un merci tout particulier à notre compagnon à quatre pattes, Ulysse, qui est toujours partant pour s'amuser ou manger une pâtée.

J'ai souvent entendu dire que la thèse n'est pas un sprint, mais un marathon. Ayant pratiqué les deux et plus encore, je dirai qu'il ne s'agit ni de l'un, ni de l'autre. La thèse est un ultra trail, avec beaucoup de dénivelé. Il y a parfois de longues et pénibles montées, menant cependant à de superbes paysages. Il y a ensuite d'appréciables descentes, très techniques, et où le moindre faux pas nous fait trébucher. Il y a des moments de doute, de perte de lucidité, d'euphorie, de douleur et de joie. Des ravitos pour reprendre des forces et repartir de plus belle. Une arrivée magnifique à laquelle on a tant rêvé. Mais surtout il y a toi, Claire, ma femme, qui a été présente sur tout le parcours, qui m'a relevé, qui a pansé mes blessures, qui m'a encouragé et soutenu avec un amour inconditionnel. Merci d'avoir rendu cette aventure possible.

# Sommaire

Chapitre 1

# Introduction

**Sommaire du présent chapitre**



## 1.1 Contexte général

### 1.1.1 Intérêt du transport réactif en milieux poreux

La simulation des phénomènes de transport réactif en milieux poreux représente un enjeu majeur pour de nombreux projets liés aux problématiques environnementales, tels que le stockage du dioxyde de carbone ($CO_2$), la géothermie ou le stockage de l'hydrogène. En particulier, la capture et le stockage du $CO_2$ font partie intégrante des stratégies de réduction des émissions de gaz à effet de serre présentées par le GIEC (Groupe d'experts intergouvernemental sur l'évolution du climat) dans leurs nombreux rapports [45, 67]. Ces technologies visent à capter le $CO_2$ émis par les industries émettrices, comme les centrales électriques au charbon et au gaz ainsi que les industries lourdes (sidérurgie, cimenteries), puis à l'injecter dans des réservoirs géologiques profonds pour un stockage

permanent. Plusieurs projets pilotes ont vu le jour ces dernières années, tel que le projet Northern Lights en Norvège [50, 65], démontrant l'intérêt de cette solution. Cependant, son déploiement à grande échelle représente encore un défi technique, économique et politique. Pour une analyse grand public du sujet, se référer à [46].

D'un point de vue purement technique, le pilotage de l'injection de $CO_2$ dans le sous-sol ne peut se passer de simulations numériques. Ces modèles permettent de simuler des scénarios complexes d'évolution du sous-sol en prédisant l'évolution des processus physiques et chimiques qui se déroulent dans le réservoir de stockage. Ces processus intègrent les interactions thermiques, hydrauliques, mécaniques et chimiques (THMC) qui se produisent lors de l'injection du $CO_2$. Ces phénomènes interconnectés sont fortement non linéaires et nécessitent des méthodes numériques avancées pour fournir des prévisions fiables de l'évolution du sous-sol sur le long terme.

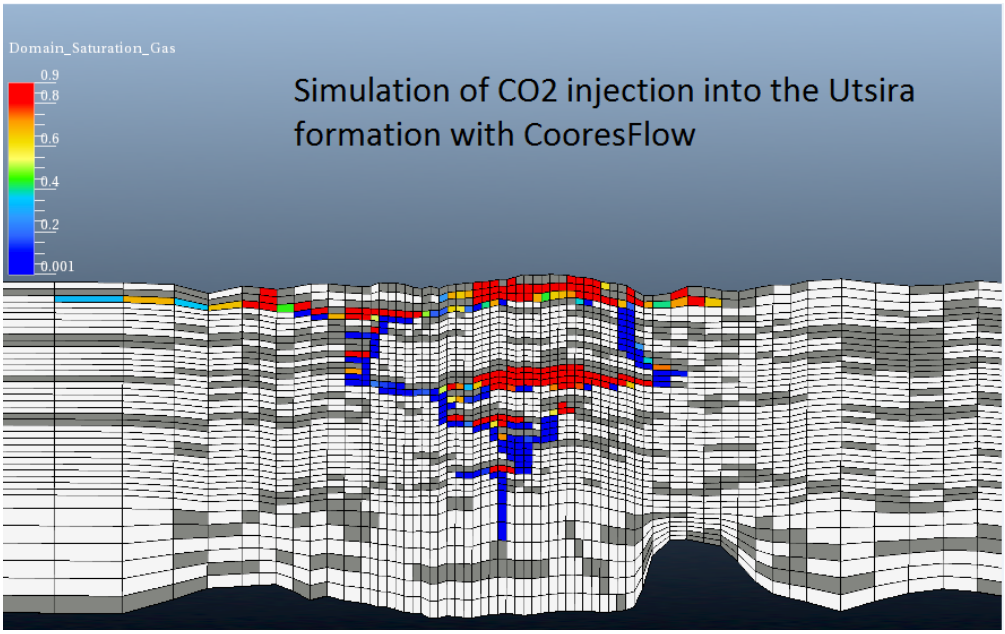


**Figure 1.1** – Simulation de l'injection de $CO_2$ avec CooresFlow.

Il existe de nombreux outils numériques permettant de simuler ces phénomènes, tels que CooresFlow, DuMuX [37], DARTS [72], HYTEC [42], MIN3P [48] et bien d'autres. En particulier, CooresFlow est un simulateur développé par IFPEN, composé de deux calculateurs distincts : GeoXim pour la simulation des phénomènes de transport et ArXim pour la modélisation chimique. Les performances des simulateurs de transport réactif actuels sont souvent limitées par la modélisation chimique du problème considéré. La raideur des systèmes modélisés peut entraîner des coûts de calcul prohibitifs, rendant

difficile l'exploration de scénarios variés dans des délais raisonnables. Améliorer la résolution des équations non linéaires pour le calcul des équilibres chimiques aurait un impact direct sur les performances, car ces calculs sont nécessaires à chaque pas de temps et pour chaque élément du maillage, conduisant à des milliers de systèmes d'équations à résoudre.

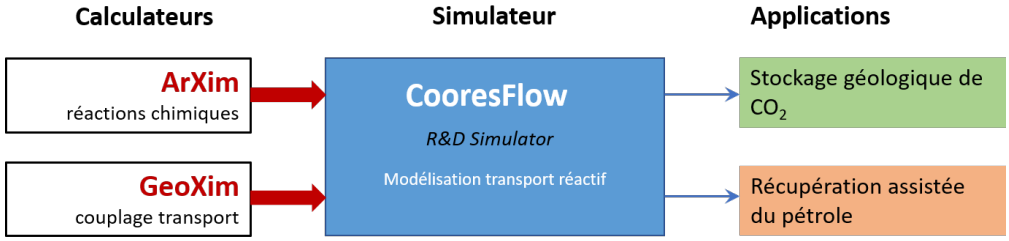


**Figure 1.2 –** Le logiciel CooresFlow et ses applications.

En modélisation chimique, il existe principalement deux types de réactions : les réactions d'équilibre et les réactions cinétiques. Le travail de cette thèse s'inscrit dans l'amélioration des méthodes numériques de résolution des équilibres chimiques.

### 1.1.2 Historique du calcul d'équilibre chimique

L'équilibre chimique a été conceptualisé pour la première fois par Berthollet en 1803 [9]. En 1864, Guldberg et Waage [75] ont défini la loi d'action de masse, qui permet de calculer l'équilibre chimique. En 1873, Gibbs a montré [29] que la minimisation d'une fonctionnelle, connue aujourd'hui sous le nom d'énergie libre de Gibbs, permettait également ce calcul. Il a démontré que le minimum global de cette fonction d'état est atteint pour une composition d'espèces chimiques à l'équilibre. Jusque dans les années 1940, les calculs d'équilibre chimique ne portaient que sur quelques espèces et étaient effectués de manière analytique [36]. Après la Seconde Guerre mondiale, Brinkley [12] a proposé un algorithme pour le calcul par ordinateur. Storey et Van Zeggeren [80] notent que le développement des méthodes de résolution des équilibres chimiques à cette époque a été principalement motivé par le calcul des propriétés des propergols (produits de propulsion) et des moteurs de fusées [66], des explosifs [19, 81], des applications dans le traitement chimique et dans le comportement des systèmes cellulaires biologiques multiphasés [20]. Smith a passé en revue les méthodes de cette période [63], et plus tard Smith et Missen [64] ont proposé une classification des différentes approches en deux catégories : les méthodes stœchiométriques et les méthodes non stœchiométriques. Les méthodes stœchiométriques sont basées sur les équations d'action de masse, tandis que

les méthodes non stœchiométriques sont basées sur la minimisation de l'énergie libre de Gibbs. Bien que notre approche soit basée sur la minimisation de l'énergie de Gibbs, nous reformulons nos équations de manière à utiliser la loi d'action de masse, ce qui en fait une méthode stœchiométrique. Comme l'indiquent Leal et al. [41], de nombreux codes géochimiques utilisent ce type de méthode, notamment EQ3/6 [79], PHREEQC [53], WATEQ [69], MINEQL [77], CHESS [43], CHIM-XPT [57] et SOLVED-XPT [56]. Pour les méthodes non stœchiométriques, une liste non exhaustive de codes utilisant cette méthode comprend ChemSage [24], THERIAK [13], HCh [61], FactSage [2], PERPLEX [17, 18], GEM-Selektor [38] et Reaktoro [40]. Les développements ultérieurs dans ce domaine ont été examinés par Tsanas et al. [71, 70], et Coatléven et Michel [15].

## 1.2 Problématiques traitées

Les problématiques abordées au cours de cette thèse concernent l'équilibre chimique d'un mélange idéal qui est défini comme suit. Étant donné des quantités d'éléments chimiques, une pression et une température, un calcul d'équilibre chimique consiste à trouver les quantités d'espèces chimiques $\mathbf{n}$ minimisant une fonction d'état $G$, appelée énergie libre de Gibbs, et satisfaisant la conservation des éléments chimiques. Cette minimisation s'écrit

$$\mathbf{n} \in \arg\min\{G(\mathbf{n}) \mid \mathbf{An} = \mathbf{b}, \mathbf{n} \geq \mathbf{0}\}, \tag{1.1}$$

où $G(\mathbf{n})$ représente l'énergie de Gibbs pour une quantité de matière $\mathbf{n}$ d'un système chimique et $\mathbf{An} = \mathbf{b}$ représente la conservation des éléments chimiques. Si le système chimique considéré peut ne contenir qu'une seule phase, il est qualifié de monophasique ; s'il peut contenir un ensemble de phase $\alpha = 1, \ldots, N_{Ph}$, dans ce cas on parle d'équilibre multiphasique.

### 1.2.1 Équilibre monophasique

Dans le cas d'un équilibre chimique monophasique. Le problème (1.1) implique la résolution d'équations linéaires exprimant la conservation de la masse ainsi que d'équations non linéaires liées aux réactions chimiques impliquées. Ces équations s'expriment classiquement selon le nombre de moles $\mathbf{n} = (n_i)_{1,\ldots,N}$ de chaque espèce chimique. Cependant, nous avons conduit une étude dans la Section 2.2.3 sur les performances numériques de différentes formulations. Nous avons conclu qu'une formulation utilisant les fractions molaires $\mathbf{x}$ des espèces chimiques présente de meilleures performances. Ce problème s'écrit :

Trouver $(\mathbf{x}, \omega)$ tels que

$$\begin{aligned} \mathbf{A}\mathbf{x} - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T\mathbf{y}(\mathbf{x}) &= \mathbf{d}, \\ \langle \mathbf{x}, \mathbf{1} \rangle &= 1, \end{aligned}$$

où

$$(\mathbf{y}(\mathbf{x}))_i = y(x_i) := \ln x_i,\ i = 1, \dots, N \quad \text{et} \quad \mathbf{d} := -\mathbf{S}^T\boldsymbol{\mu}^\circ/(\mathrm{RT}).$$

Dans ces équations, les symboles $\mu_i^\circ$, R et T représentent des constantes, tandis que $\omega$ est une inconnue qui correspond à l'inverse du nombre total de moles. Cette variable permet d'établir un lien entre les fractions molaires et les quantités de matière. De plus, $\mathbf{S}$ est appelée matrice de stœchiométrie. L'utilisation de l'algorithme de Newton [35, 34, 21, 22] pour la résolution par linéarisation successive de ces équations rencontre un certain nombre de difficultés :

— Les itérés peuvent prendre des valeurs négatives, ce qui est incompatible avec le log.

— Les valeurs de la solution couvrent une large gamme de valeurs du domaine de définition des inconnues, ce qui conduit à des problèmes de préconditionnement.

— La convergence de l'algorithme n'est pas assurée si l'on démarre loin de la solution.

### Approche classique

Parmi les solutions mises en œuvre dans Arxim, l'utilisation du logarithme des nombres de moles comme inconnues permet de gérer la contrainte de positivité et de réduire les ordres de grandeur entre espèces. Cette méthode, appelée log-trick [78], est très populaire dans les codes de transport réactif. Pour notre formulation, l'inconnue devient $\mathbf{y} = \ln \mathbf{x}$ et le système à résoudre devient :

Trouver $(\mathbf{y}, \omega)$ tels que

$$\begin{aligned} \mathbf{A}\mathbf{x}(\mathbf{y}) - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T\mathbf{y} - \mathbf{d} &= \mathbf{0}, \\ \langle \mathbf{x}(\mathbf{y}), \mathbf{1} \rangle - 1 &= 0, \end{aligned}$$

où $\mathbf{x}(\mathbf{y}) = (x(y_i))_{i=1,\dots,N}$ avec $x(y_i) := y^{-1}(y_i) = \exp y_i$,

Bien que cette approche évite les deux problèmes mentionnés précédemment, une valeur élevée d'un $y_i$ au cours des itérations peut mener à des difficultés dans la résolution du système à cause de l'exponentiel.

**Paramétrage**

Dans le cadre de la thèse, la technique de paramétrage développée par Brenner et Cancès [11] est utilisée pour basculer automatiquement entre les deux formulations tout en s'assurant que les dérivées partielles de la jacobienne restent bornées. Cette technique a déjà été appliquée avec succès par Bassetto [4, 3] pour un problème de même nature. La résolution du système se fait selon un paramètre $\tau_i \in \mathbb{R}$ pour chaque espèce et le problème devient : trouver $(\tau, \omega)$ tels que :

$$\begin{aligned} \mathbf{A}\mathbf{X}(\boldsymbol{\tau}) - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T\mathbf{Y}(\boldsymbol{\tau}) - \mathbf{d} &= \mathbf{0}, \\ \langle \mathbf{X}(\boldsymbol{\tau}), \mathbf{1} \rangle - 1 &= 0, \end{aligned}$$

où $X$ et $Y$ sont deux fonctions tels que

$$Y(\tau_i) = \ln X(\tau_i).$$

Il est essentiel que ces fonctions possèdent des dérivées bornées afin d'éviter des coefficients de jacobienne qui pourraient atteindre des valeurs très élevées, voire exploser, dans certains régimes de fonctionnement. Il existe plusieurs choix pour la définition de ces fonctions. Cependant, celui présenté dans la Figure 1.3 permet d'obtenir de meilleurs performances d'un point de vue numérique.

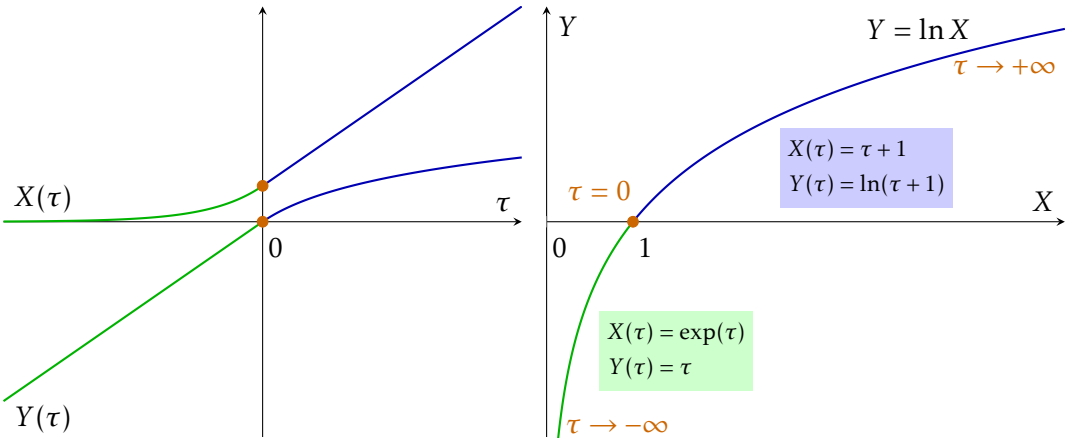


**Figure 1.3 –** Fonction $X$ et $Y$ et paramétrage du graphe $Y(\tau) = \ln X(\tau)$.

**Représentation cartésienne**

Une deuxième approche développée dans la thèse est la représentation cartésienne équilibrée. Cette méthode consiste à choisir les fractions molaires et leurs logarithmes comme inconnues, une fonction établissant la relation entre ces deux quantités est alors introduite. Cette fonction possède des propriétés qui permettent de surmonter les problèmes mentionnés ci-dessus et de contrôler les coefficients de la jacobienne. La résolution se fait sur un système agrandi où à la fois $\mathbf{x}$ et $\mathbf{y}$ sont des inconnues, le problème devient :

Trouver $(\mathbf{x}, \mathbf{y}, \omega)$ tels que

$$\begin{aligned} \mathbf{A}\mathbf{x} - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T\mathbf{y} - \mathbf{d} &= \mathbf{0}, \\ \langle \mathbf{x}, \mathbf{1} \rangle - 1 &= 0, \\ f(\mathbf{x}, \mathbf{y}) &= \mathbf{0}, \end{aligned} \tag{1.2}$$

avec $\boldsymbol{f}(\mathbf{x}, \mathbf{y}) = (f(x_i, y_i))_{i=1,\dots,N}$ une fonction liant $x_i$ et $y_i$ et de tel sorte que

$$f(x_i, y_i) = 0 \Leftrightarrow y_i = \ln x_i.$$

Comme pour le paramétrage, il existe un grand nombre de possibilités pour définir la fonction $f$ avec pour principale condition d'avoir des dérivées partielles bornées pour contrôler les coefficients de la jacobienne. La fonction présentée dans la Figure 1.4 est celle ayant le meilleur comportement numérique. Cette fonction est définie selon quatre régions dans le plan $(x, y)$ avec un raccord $\mathcal{C}^1$ entre chaque zone.

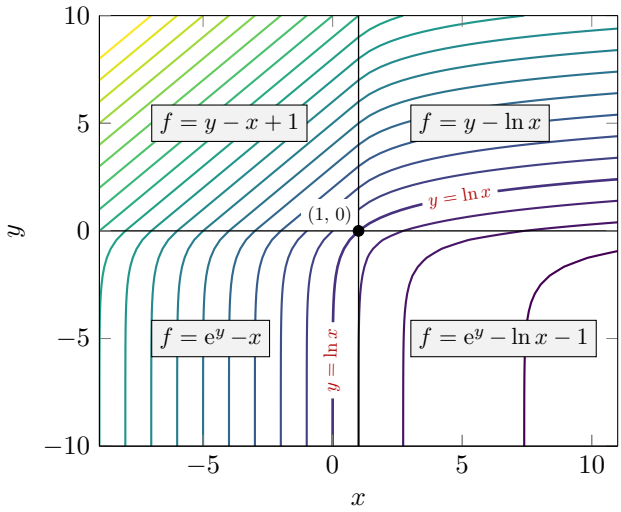


**Figure 1.4 –** Lignes de niveau de la fonction $f$ pour la représentation cartésienne

### 1.2.2 Équilibre multiphasique

Dans le cadre d'un équilibre multiphasique, aux problématiques précédentes s'ajoute le fait de devoir traiter la présence ou l'absence de phases. En effet, il n'est pas aisé de connaître en amont les phases qui seront présentes à l'équilibre. Il est alors nécessaire d'intégrer une stratégie de gestion des phases présentes et absentes. Une approche couramment employée consiste à gérer de manière dynamique ces phases. Le calcul d'équilibre chimique est effectué pour un ensemble de phases potentiellement présentes, puis on teste l'ajout d'une nouvelle phase en faisant à nouveau les calculs. Cette stratégie est très coûteuse car elle nécessite plusieurs calculs d'équilibres chimiques pour diverses combinaisons de phases.

Dans le cadre du calcul d'équilibre de phases (équilibre ne prenant pas en compte les réactions chimiques), une approche récemment développée par Lauzer *et al.* [39] consiste à considérer des fractions molaires étendues aux phases absentes. Cette approche permet le traitement des phases présentes et absentes dans un cadre unifié et rigoureusement posé. Cette formulation a notamment été étudié par Ben Gharbia [6, 8], Masson *et al.*[44] ainsi que dans diverses travaux à IFPEN [7, 5]. En particulier, les travaux doctoraux de Vu [73] ont permit de montrer une équivalence entre la formulation unifiée et la minimisation d'une énergie de Gibbs modifiée.

Dans cette thèse, nous proposons une formulation unifiée pour le calcul d'équilibre chimique multiphasique avec réactions. Cette formulation s'écrit comme suit :

Trouver $(\boldsymbol{\xi}^\alpha, s_\alpha, r_\alpha)_{\alpha=1,\dots,N_{Ph}}$ tels que

$$\sum_\alpha s_\alpha \mathbf{A}^\alpha \boldsymbol{\xi}^\alpha - b = 0, \tag{1.3}$$

$$\mathbf{S}^T \boldsymbol{\mu}(\boldsymbol{\xi}) = \mathbf{0}, \tag{1.4}$$

$$\langle \boldsymbol{\xi}^\alpha, \mathbf{1} \rangle + r_\alpha - 1 = 0, \quad (\alpha = 1, \dots, N_{Ph}), \tag{1.5}$$

$$s_\alpha r_\alpha = 0, \quad (\alpha = 1, \dots, N_{Ph}), \tag{1.6}$$

$$s_\alpha \geq 0, r_\alpha \geq 0, \quad (\alpha = 1, \dots, N_{Ph}), \tag{1.7}$$

avec

$$\mu_i(\xi_i) = \mu_i^\circ + \mathrm{RT} \ln \xi_i$$

Dans cette formulation, $\boldsymbol{\xi}^\alpha$ représente les fractions molaires étendues de la phase $\alpha$, tandis que $s_\alpha$ représente la quantité total d'atomes de la phase $\alpha$. La présence ou l'absence de cette phase est gérée par l'équation (1.6) qui, couplée à la condition (1.7), représente ce que l'on nomme un problème de complémentarité. Par conséquent, si une phase $\alpha$ est présente à l'équilibre, alors $s_\alpha > 0$, $r_\alpha = 0$ et $\boldsymbol{\xi}^\alpha$ représente exactement les fractions molaires des différentes espèces au sein de la phase. Réciproquement, si une phase $\alpha$ est absente à l'équilibre, alors $s_\alpha = 0$, $r_\alpha \geq 0$ et le vecteur $\boldsymbol{\xi}^\alpha$ ne somme pas nécessairement à 1. Dans ce cas, il en résulte que

$$\langle \boldsymbol{\xi}^\alpha, \mathbf{1} \rangle \leq 1.$$

Bien que présentant des similarités avec l'approche de Coatléven et Michel [15], notre formulation diffère de par un établissement plus naturel des équations à partir du problème de minimisation grâce à l'utilisation du concept de sous-différentiel d'une fonction convexe. D'un point de vue théorique, nous avons montré sous quelles conditions la solution de ce système est unique. Pour le numérique, les deux principales difficultés rencontrées sont les suivantes :

— La non-linéarité du logarithme : comme pour le cas monophasique, la présence du logarithme introduit une certaine raideur dans le système, compliquant la résolution du système.

— Le problème de complémentarité : la présence ou l'absence de phases introduisent un problème de complémentarité, nécessitant des approches spécialisés pour son traitement.

Afin de traiter le problème du log, nous avons appliqué nos méthodes de paramétrage et de représentation cartésienne. Pour la complémentarité, nous avons appliqué quelques méthodes classiques et proposons une nouvelle approche basée sur la paramétrage.

**Nouvelle approche pour le problème de complémentarité**

Le problème de complémentarité est défini par

$$s_\alpha r_\alpha = 0 \quad \text{et} \quad s_\alpha \geq 0, r_\alpha \geq 0$$

a été largement étudié en optimisation [1, 25, 26] et de nombreuses techniques existent pour son traitement. Graphiquement, l'ensemble des $(s_\alpha, r_\alpha)$ admissible est représenté par les deux demi-axes $\{s_\alpha \in \mathbb{R} \mid s_\alpha \geq 0\}$ et $\{r_\alpha \in \mathbb{R} \mid r_\alpha \geq 0\}$. La Figure 1.5 représente la région de faisabilité du problème de complémentarité. Les difficultés principales

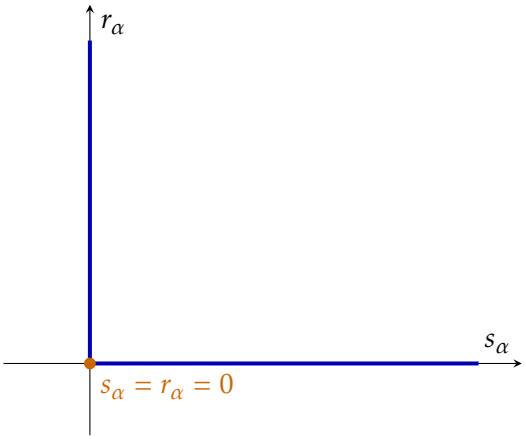


**Figure 1.5 –** Le problème de complémentarité

rencontrées par l'algorithme de Newton pour le traitement de la complémentarité sont la condition de positivité des deux variables de complémentarité et la non-différentiabilité de la contrainte au point $(0, 0)$. Les méthodes classiquement utilisées pour ce problème se répartissent en deux groupes :

— Les méthodes semi-lisses : le problème de complémentarité est remplacé par une fonction de complémentarité. Cette fonction est lipschitzienne mais n'est pas dérivable en tout point. Cependant, le concept de sous-différentiel permet d'utiliser un algorithme de Newton adapté aux problèmes semi-lisses.

— Les méthodes de régularisation : les équations de complémentarité sont régularisées grâce à un paramètre de régularisation. La méthode de Newton classique peut alors s'appliquer sur ce problème régularisé. Au cours des itérations, on fait tendre progressivement ce paramètre vers zéro selon une certaine stratégie afin de retrouver le problème de complémentarité initial quand les itérés sont proches de la solution.

Dans cette thèse, nous proposons une méthode de paramétrage de la complémentarité grâce à l'introduction d'une variable $\eta_\alpha$ pour chaque phase et de deux fonctions, $S$ et $R$ de telles sorte que

$$S(\eta_\alpha)R(\eta_\alpha) = 0.$$

Ainsi, la condition de complémentarité (1.6)–(1.7) est directement intégrée dans les équations restantes et le système à résoudre devient :

Trouver $(\xi^\alpha, \eta_\alpha)_{\alpha=1,\dots,N_{Ph}}$ tels que

$$\begin{aligned}
\sum_\alpha S(\eta_\alpha)\mathbf{A}^\alpha \xi^\alpha - b &= 0,\\
\mathbf{S}^T \boldsymbol{\mu}(\xi) &= \mathbf{0},\\
\langle \xi^\alpha, \mathbf{1}\rangle + R(\eta_\alpha) - 1 &= 0, \quad (\alpha = 1, \dots, N_{Ph}).
\end{aligned}$$

Les fonctions $S$ et $R$ sont présentées dans la Figure 1.6.

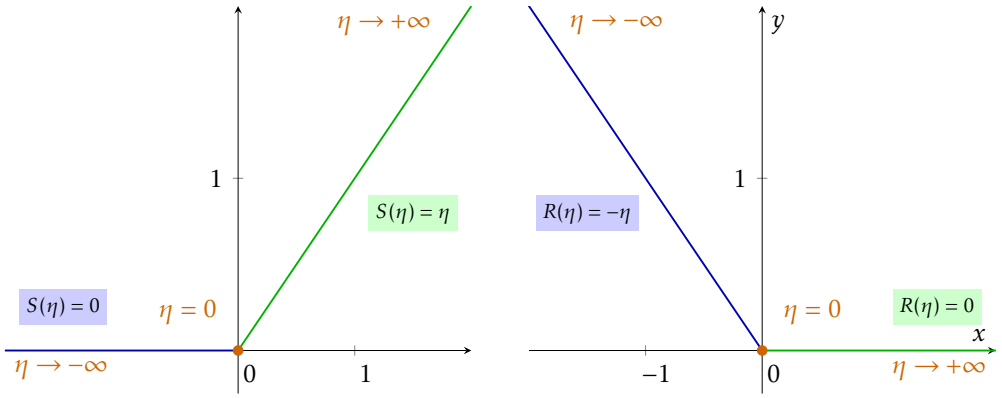


**Figure 1.6 –** Fonctions $S$ et $R$ pour le paramétrage de la complémentarité.

## 1.3 Contenu du manuscrit

Les algorithmes présentés dans cette thèse ont été implémentés dans un code écrit avec le langage de programmation Julia [10]. Ce solveur, baptisé JuXim (fusion de Julia et ArXim), contient un solveur générique de Newton pour les équilibres chimiques monophasiques et multiphasiques. Ce code utilise notamment le package ForwardDiff [58] pour calculer efficacement les dérivées nécessaires à la méthode de Newton. JuXim lit la base de données thermochimiques SUPCRT92 [32] afin de générer le système chimique à partir d'une pression, température et d'un ensemble d'espèces chimiques

primaires spécifiés par l'utilisateur. À terme, il est prévu de rendre le solveur JuXim accessible publiquement afin qu'il puisse bénéficier à la communauté scientifique travaillant sur la modélisation des équilibres chimiques. De plus, toutes les spécificités des cas de test présentés dans cette thèse sont disponibles en annexe A afin de garantir une reproductibilité des résultats.

Le contenu du manuscript chapitre par chapitre est le suivant.

**Description mathématique du problème d'équilibre chimique**

Le Chapitre 2 se consacre à la description mathématique du problème d'équilibre chimique, structuré en trois sections principales.

La Section 2.1 aborde la construction du système chimique, en introduisant les équations de conservation de la masse et la définition de l'énergie libre de Gibbs. Ces éléments sont essentiels pour formuler le problème d'équilibre chimique comme un problème de minimisation sous contraintes.

La Section 2.2 traite du cas simplifié de l'équilibre chimique monophasique dans une solution diluée, également connu sous le nom de spéciation. Elle détaille la formulation du système ainsi que les conditions d'existence et d'unicité des solutions, en s'appuyant sur les travaux de Shapiro et Shapley [60]. Ces derniers ont démontré qu'un système idéal monophasique admet une solution unique, à condition que l'ensemble des contraintes soit compact. Cette section étend ces résultats aux mélanges non idéaux avec une énergie de Gibbs convexe, garantissant ainsi l'existence et l'unicité des solutions pour une définition plus large des potentiels chimiques. Un système d'équations algébriques équivalent au problème de minimisation est également présenté.

La Section 2.3 établit les équations pour le cas multiphasique idéal et donne les conditions nécessaires pour l'existence et l'unicité des solutions. Bien que Shapiro et Shapley aient exploré rigoureusement l'équilibre chimique multiphasique idéal [60], l'unicité n'est pas toujours garantie si certaines phases peuvent disparaître. Cette section introduit une nouvelle condition qui assure l'unicité d'un minimiseur même en l'absence de certaines phases, enrichissant ainsi le cadre analytique. Dans la littérature, la complexité des calculs d'équilibre multiphasique, due à la contrainte de non-négativité, est abordée par deux approches principales : une méthode combinatoire qui ajoute les phases une par une pour tester leur présence [71] et une formulation unifiée utilisant des fractions molaires étendues [15]. Notre travail prolonge la formulation unifiée en établissant une équivalence entre le problème de minimisation et un nouvel ensemble d'équations algébriques, en s'appuyant sur le concept de sous-différentiel de l'énergie

de Gibbs.

**Nouveaux algorithmes pour l'équilibre chimique multiphasique**

Le Chapitre 3 se concentre sur la résolution numérique du problème d'équilibre chimique monophasique, structuré en trois sections. Les résultats présentés dans ce chapitre constituent une grande partie du preprint [33].

La Section 3.1 aborde les méthodes de paramétrisation et de représentation cartésienne, conçues pour surmonter les défis liés à la résolution du système d'équations non linéaires du problème d'équilibre chimique à l'aide de la méthode de Newton.

La Section 3.2 fournit une preuve théorique de la convergence quadratique locale de la méthode de Newton pour les deux approches discutées.

La Section 3.3 présente des expériences numériques qui illustrent l'exactitude des deux techniques, permettant le calcul des équilibres pour des espèces chimiques à très faibles concentrations. De plus, nous comparons nos résultats avec ceux obtenus en utilisant un équivalent du solveur Arxim dans notre cadre, en considérant différentes initialisations.

**Méthodes numériques innovantes pour l'équilibre chimique monophasique**

Le Chapitre 4 se concentre sur le développement d'algorithmes numériques destinés à résoudre les complexités des problèmes d'équilibre chimique multiphasique.

La Section 4.1 rappelle les techniques de paramétrisation et de représentation cartésienne introduites dans le Chapitre 2, qui ont été appliquées à l'équilibre chimique monophasique.

La Section 4.2 explore les méthodes classiques pour traiter le problème de complémentarité. Nous examinons les fonctions de complémentarité min et Fischer-Burmeister, ainsi que la méthode des points intérieurs. En outre, nous proposons une approche innovante qui intègre directement les équations de complémentarité dans le système via des techniques de paramétrisation. Cette approche offre un nouveau cadre pour résoudre ces problèmes tout en maintenant les conditions essentielles durant les itérations.

Dans la Section 4.3, nous validons nos approches proposées à l'aide d'un cas de test simple, suivi d'un cas de test plus complexe visant à évaluer la robustesse des méthodes. Ce dernier cas implique 22 phases, fournissant un test rigoureux pour les performances de nos algorithmes. À travers ces expériences, nous cherchons à démontrer l'efficacité et la fiabilité de nos algorithmes dans la résolution de problèmes complexes d'équilibre chimique multiphasique. Nous fournissons en particulier des preuves de la

robustesse et de l'efficacité de l'approche de paramétrisation pour résoudre les conditions de complémentarité.

Chapter 2

# Mathematical description of the chemical equilibrium problem

**Outline of the current chapter**



This chapter contains the mathematical description of the chemical equilibrium problem and is divided into three parts.

The first section is dedicated to the construction of the chemical system, including the mass conservation equations and the definition of the Gibbs free energy, needed to construct the chemical equilibrium problem as a constrained minimization problem.

The second section deals with the simplified case of single-phase equilibrium in dilute solution, also known as speciation. The formulation of the system and the conditions of existence and uniqueness of solutions are detailed. The existence and uniqueness of a solution to this problem have been studied by Shapiro and Shapley [60]. They proved that for a single-phase ideal system, the problem admits a unique solution, provided that the set of constraints is compact. We show that this compactness condition is satisfied in the case discussed. Additionally, Smith and Missen [64] proved the strict convexity of the Gibbs energy on the set of constraints for a single-phase ideal mixture, further supporting the uniqueness of the solution. In this part, we will extend these results to the more general case of mixtures with a convex Gibbs energy that may be non-ideal, which has not been studied in the literature. This extension ensures the existence and uniqueness of solutions for a broader definition of the chemical potentials and shows that the case of ideal mixtures covered by Shapiro and Shapley is a special case included within the more general result. Finally, a system of algebraic equations equivalent to the minimization problem is presented.

The third and final section establishes the equations for the multiphase case and gives the conditions for the existence and uniqueness of solutions to the problem. Shapiro and Shapley [60] rigorously explored the existence and uniqueness of minimizers for ideal multiphase chemical equilibrium. They characterized the minimizers set and established that when all phases are required at equilibrium, the minimizer is unique. However, this uniqueness is not guaranteed if some phases may vanish. Despite this, they proved that the mole fractions of the present phases remain uniquely defined. This section extends their work by introducing a new condition that guarantees the uniqueness of a minimizer even when certain phases are missing, offering a deeper analytical framework.

Multiphase equilibrium computations are inherently more complex than single-phase systems due to the non-negativity constraint, which complicates the determination of present phases at equilibrium. To address this complexity, two main approaches are used in the literature: a combinatorial method that compute the chemical equilibrium for all possible phase subsets, and a unified formulation that uses extended mole fractions and complementarity equations to consider all potentially present phases at once. The unified formulation, initially introduced by Lauser et al. [39] and adapted by Vu et al. [28] for nonreactive phase equilibrium, has been applied by Coatléven and Michel [15] to the multiphase chemical equilibrium case. Our work extends this framework by establishing an equivalence between the minimization problem and a new set of algebraic equations based on extended mole fractions, utilizing the concept

of the subdifferential of Gibbs energy.

## 2.1 Chemical system

A chemical species is a a group of linked atoms characterizes by its molecular formula and the phase to which its belongs. The considered phases are the following: aqueous (aq), gaseous (g) and solid (s). For instance $\mathrm{H_2O(aq)}$, $\mathrm{H^+(aq)}$, $\mathrm{NaCl(aq)}$, $\mathrm{CO_2(g)}$, $\mathrm{H_2O(g)}$, $\mathrm{CaCO_3(s)}$ are all different chemical species. It is worth noting that $\mathrm{H_2O(aq)}$ and $\mathrm{H_2O(g)}$ are considered as two different species. For a given temperature T and pressure P, a chemical system $\mathcal{S}_{\mathrm{P,T}} = \{\mathcal{C}, \mathcal{E}, \mathcal{R}, \mathcal{P}\}$ is a collection of four sets:

- a set of $N$ chemical species $\mathcal{C} = \{C_1, \ldots, C_N\}$;
- a set of $M$ chemical elements, typically atoms, $\mathcal{E} = \{E_1, \ldots, E_M\}$, $M < N$;
- a set of $N - M$ chemical reactions $\mathcal{R} = \{R_1, \ldots, R_{N-M}\}$;
- and a set of $N_{Ph}$ phases $\mathcal{P} = \{P_1, \ldots, P_{N_{Ph}}\}$.

The set $\mathcal{E}$ contains all the elements that compose the species of the set $\mathcal{C}$ and the reactions in $\mathcal{R}$ describe how these species interact with each other. A chemical reaction $R_j \in \mathcal{R}$ can be written as

$$\sum_{i=1}^{N} s_{ij} C_i = 0,$$

where the $s_{ij}$ are the stoichiometric coefficients that represent the number of molecules of the species $C_i$ involved in the reaction $R_j$.

The set $\mathcal{P}$ contains the different phases which compose the system (aqueous, gaseous or pure mineral phases). To link a species $C_i$ to its phase $\alpha$, we introduce the following map:

$$\sigma : i \in \{1, \ldots, N\} \mapsto \alpha \in \{1, \ldots, N_{Ph}\}. \tag{2.1}$$

In this way, $\sigma^{-1}(\alpha)$ represents the set of all species in the phase $\alpha$.

### 2.1.1 Mass conservation

The systems we are studying are closed, so the principle of conservation of mass holds and leads to the conservation of the quantities $\mathbf{b} = (\mathrm{b}_1, \ldots, \mathrm{b}_M)$ of each elements of $\mathcal{E}$. To express this conservation, let $\mathbf{a}_i$ be the formula vector of $C_i \in \mathcal{C}$ in the element basis $\mathcal{E}$ – meaning that if $\mathcal{E}$=(H, C, O) and $C_i$ =$\mathrm{HCO_3^-}$, then $\mathbf{a}_i = (1, 1, 3)^T$ – then the set of species $\mathcal{C}$ can be subdivided into two particular sets $\mathcal{C}_{Pr}$ and $\mathcal{C}_{Sd}$ such that:

- $\mathcal{C}_{Pr} = \{C_1, \ldots, C_M\}$ is the primary species set composed of species which have linearly independent formula vectors $(\mathbf{a}_1, \ldots, \mathbf{a}_M)$. This set is the primary basis for the system and its size is equal to $M$ which is also the number of element in the system;
- $\mathcal{C}_{Sd} = \{C_{M+1}, \ldots, C_N\}$ is the secondary species set containing species which formula vectors can be obtained by linear combinations of primary species and its size is equal to $N - M$ which corresponds to the $N - M$ chemical reactions of $\mathcal{R}$.

Note that the choice of the primary species is not unique. Since the primary species are linearly independent, it is useful to have an ordered set of species with the primary species first followed by the secondary species. The *formula matrix* $\mathbf{A}$ is the matrix composed of the formula vectors. Its first $M$ columns correspond to the formula vectors of the primary species and the last $N - M$ columns to the secondary species. This matrix is then written as

$$\mathbf{A} = [\mathbf{A}_{Pr}, \mathbf{A}_{Sd}],$$

where $\mathbf{A}_{Pr}$ is a $M \times M$ invertible matrix and $\mathbf{A}_{Sd}$ is a $M \times (N - M)$ rectangular matrix. A simple example of such a problem is the case of the dissociation of water which is composed of one aqueous phase and the following sets:

$$\mathcal{C} = (\mathrm{H^+}, \mathrm{OH^-}, \mathrm{H_2O},), \quad \mathcal{E} = (\mathrm{H}, \mathrm{O}), \quad \mathcal{R} = (\mathrm{H_2O} = \mathrm{H^+} + \mathrm{OH^-}), \quad \mathcal{P} = \{\text{aqueous}\}.$$

The corresponding formula matrix is

$$\mathbf{A} = \begin{array}{c} \begin{matrix} \mathrm{H^+} & \mathrm{OH^-} & \mathrm{H_2O} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \end{bmatrix} \begin{matrix} \mathrm{H} \\ \mathrm{O} \end{matrix} \end{array}.$$

Let $\mathbf{n} = (n_1, \ldots, n_N)$ be the vector of quantities of mole of each species of $\mathcal{C}$, the conservation of elements can then be written as

$$\mathbf{A}\mathbf{n} = \mathbf{b}. \tag{2.2}$$

In the case of a multiphase equilibrium, it is interesting to partition the vector $\mathbf{n}$ and the matrix $\mathbf{A}$ into sub-vectors $\mathbf{n}^\alpha$ and sub-matrices $\mathbf{A}^\alpha$ for each phase $\alpha$. Equation (2.2) thus becomes

$$\sum_{\alpha=1}^{N_{Ph}} \mathbf{A}^{\alpha}\mathbf{n}^{\alpha} = \mathbf{b}.$$

The matrix $\mathbf{A}$ has interesting properties and allows to define the stoichiometry matrix $\mathbf{S}$, which is very useful to simplify the formulation of the chemical equilibrium problem. This matrix can be defined as

$$\mathbf{S} := \begin{bmatrix} \mathbf{A}_{Pr}^{-1}\mathbf{A}_{Sd} \\ -\mathbf{I}_{Sd} \end{bmatrix}, \tag{2.3}$$

in order to satisfy the properties

$$\mathbf{AS} = \mathbf{0} \quad \text{and} \quad \operatorname{rank}\mathbf{S} = N - M.$$

Furthermore, the matrix $\mathbf{S}$ is composed of the stoichiometry coefficients involved in the chemical reactions of $\mathcal{R}$ with $\mathbf{S}_{ij} = s_{ij}$. The stoichiometry matrix for the example of dissociation of water is

$$\begin{matrix} \\ \mathbf{S} = \end{matrix} \begin{matrix} R_1 \\ \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} \end{matrix} \begin{matrix} \\ \mathrm{H}^+ \\ \mathrm{OH}^- \\ \mathrm{H_2O} \end{matrix}.$$

The following lemma formalizes the fundamental link between the matrix $\mathbf{A}$ and $\mathbf{S}$.

**Lemma 2.1.1.** *One has the following result:*

$$\ker\mathbf{S}^T = (\ker\mathbf{A})^{\perp} = \operatorname{Im}\mathbf{A}^T.$$

*Proof.* Let $\mathbf{n} = (\mathbf{n}_{Pr}, \mathbf{n}_{Sd}) \in \ker\mathbf{A}$ where $\mathbf{n}_{Pr}$ and $\mathbf{n}_{Sd}$ are respectively the vector of quantities of the primary and the secondary species. We have the following link between $\mathbf{A}$ and $\mathbf{S}$:

$$\mathbf{An} = 0 \quad \Leftrightarrow \quad \mathbf{n}_{Pr} = -\mathbf{A}_{Pr}^{-1}\mathbf{A}_{Sd}\mathbf{n}_{Sd} \quad \Leftrightarrow \quad \mathbf{n} = -\begin{bmatrix} \mathbf{A}_{Pr}^{-1}\mathbf{A}_{Sd} \\ -\mathbf{I}_{Sd} \end{bmatrix}\mathbf{n}_{Sd} = -\mathbf{S}\mathbf{n}_{Sd}.$$

It follows that $\operatorname{Im}\mathbf{S} = \ker\mathbf{A}$, then $(\operatorname{Im}\mathbf{S})^{\perp} = (\ker\mathbf{A})^{\perp}$. The result is obtained using the property $\ker\mathbf{S}^T = (\operatorname{Im}\mathbf{S})^{\perp}$ and $(\ker\mathbf{A})^{\perp} = \operatorname{Im}\mathbf{A}^T$ from linear algebra. □

More details on the stoichiometry matrix and its link with the formula matrix can be found in the book of Smith and Missen [64].

Our second lemma characterizes the kernel of the formula matrix $\mathbf{A}$.

**Lemma 2.1.2.** *The components of an element in* $\ker \mathbf{A} \setminus \{\mathbf{0}\}$ *do not all have the same sign, in particular*

$$\ker \mathbf{A} \cap \mathbb{R}^N_+ = \{\mathbf{0}\}.$$

*Proof.* Let $\mathbf{n} \in \ker \mathbf{A} \cap \mathbb{R}^N_+$, then for each $k \in \{1, \dots, M\}$, $\sum_{i=1}^N a_{ki} n_i = 0$. Since $\mathbf{A}$ is composed of formula vectors, all its components are positive and so the previous sum is a sum of positive terms. It follows that $a_{ki} n_i = 0$, for each $i$ and $k$. Moreover, each species is composed of at least one element, hence for each $i$ there exists $k$ such that $a_{ki}$ is non-zero. Therefore $n_i = 0$, for all $i$. □

### 2.1.2 Gibbs free energy and chemical potentials

The state of a closed system $\mathcal{S}_{\mathbf{P},\mathbf{T}}$ at constant pressure and temperature can be described by the Gibbs free energy function $G : \mathbb{R}^N_+ \to \mathbb{R}$, also known as the Gibbs energy or sometimes free enthalpy. This function is extensive with respect to the quantities, meaning that it is a homogeneous function of degree 1 and can be expressed using its partial derivatives. It is a sum of Gibbs energy functions per phase $G_\alpha : \mathbb{R}_+^{\#\sigma^{-1}(\alpha)} \to \mathbb{R}$ which only depend on the species in the phase $\alpha$ and are also extensive. Its standard expression for the study of chemical equilibrium is as follows:

$$G(\mathbf{n}) = \sum_{\alpha=1}^{N_{Ph}} G_\alpha(\mathbf{n}^\alpha) \quad \text{where} \quad G_\alpha(\mathbf{n}^\alpha) = \sum_{i \in \sigma^{-1}(\alpha)} n_i \frac{\partial G_\alpha(\mathbf{n}^\alpha)}{\partial n_i} = \sum_{i \in \sigma^{-1}(\alpha)} n_i \mu_i(\mathbf{n}^\alpha), \tag{2.4}$$

where $\mu_i(\mathbf{n}^\alpha) = \partial G_\alpha(\mathbf{n}^\alpha)/\partial n_i = \partial G(\mathbf{n})/\partial n_i$ is the chemical potential of the species $C_i$ expressing the variation of energy induced by a variation of the quantity $n_i$. There are a variety of different analytical expressions for chemical potentials that depend on the physics of the problem under study. Here, for any species $C_i$ in a phase $\alpha$, we consider a chemical potential of the form

$$\mu_i := \mu_i(\mathbf{n}^\alpha) = \mu_i^\circ(\mathrm{P}, \mathrm{T}) + \mathrm{RT} \ln a_i(\mathbf{n}^\alpha). \tag{2.5}$$

In (2.5), $\mu_i^\circ(\mathrm{P}, \mathrm{T})$ is the chemical potential of the species $C_i$ in phase $\alpha$ in its standard state at pressure P and temperature T, to be computed from thermodynamic tables, whereas $a_i$ is the activity of species $C_i$ that depends in general on the concentration of all the species in phase $\alpha$.

The activity of a species $C_i$ is generically written as $a_i = \gamma_i x_i$, where $\gamma_i$ is refereed to in the literature as the activity coefficient and $x_i$ stands for the mole fraction of $C_i$ in

phase $\alpha$ defined by

$$x_i := x_i(\mathbf{n}^\alpha) = n_i / \sum_{j \in \sigma^{-1}(\alpha)} n_j = n_i / \langle \mathbf{n}^\alpha, \mathbf{1} \rangle, \quad \text{where} \quad \mathbf{1} := (1, \ldots, 1)^T.$$

There are several, increasingly complex activity models for $\gamma_i$ in the scientific literature [49, 79], the most simple of which being the ideal activity model $\gamma_i = 1$. This ideal model corresponds to a theoretical ideal mixture where the mean strength of inter-molecular interactions are the same between all the molecules of the system. The activity in (2.5) is then reduced to the mole fraction. The resulting ideal Gibbs energy in (2.4) is a convex function on $\mathbb{R}^N_+$ (see [60, Theorem 8.13]).

#### The Gibbs-Duhem relation

An important property to mention is the Gibbs-Duhem relation. It is a fundamental equation in thermodynamics which states that for any state $\mathbf{n}$ of the system:

$$\sum_{i=1}^{N} n_i \mathrm{d}\mu_i(\mathbf{n}) = 0, \tag{2.6}$$

with $\mathrm{d}\mu_i$ the differential of $\mu_i$ defined as

$$\mathrm{d}\mu_i(\mathbf{n}) := \sum_{j=1}^{N} \frac{\partial \mu_i(\mathbf{n})}{\partial n_j} \mathrm{d}n_j,$$

where $\mathrm{d}n_j$ is an infinitesimal change in the value of $n_j$.

## 2.2 Single-phase chemical equilibrium problem

The type of system considered in this section involves diluted solutions of aqueous species belonging to a single aqueous phase. These solutions are composed of a predominant species called the solvent, typically water. Additionally, there are diluted aqueous species present in very small quantities. The Gibbs energy of such a system writes

$$G(\mathbf{n}) = \sum_{i=1}^{N} n_i \mu_i(\mathbf{n}).$$

In a closed system at constant pressure and temperature, chemical reactions occur spontaneously by decreasing the Gibbs free energy. A chemical equilibrium computation

consists in finding the quantities $\mathbf{n}$ of mole for each species of $\mathcal{C}$ in a system $\mathcal{S}_{\mathrm{P,T}}$ which minimizes, for a fixed temperature T, pressure P and element quantities $\mathbf{b}$, the function $G$, under constraints of element conservation and non-negativity. To describe this calculation as a constrained minimization problem, let

$$\Omega := \left\{\mathbf{n} \in \mathbb{R}^N \;\middle|\; n_i > 0,\ i = 1, \ldots, N\right\} \quad \text{and} \quad \overline{\Omega} := \left\{\mathbf{n} \in \mathbb{R}^N \;\middle|\; n_i \geq 0,\ i = 1, \ldots, N\right\},$$

be the sets of positive and non-negative vectors of $\mathbb{R}^N$ respectively, one defines the set of vectors verifying the constraints of conservation of elements and positivity or non-negativity by

$$\mathcal{M}_{\mathbf{A},\mathbf{b}} := \{\mathbf{n} \in \Omega \mid \mathbf{A}\mathbf{n} = \mathbf{b}\} \quad \text{and} \quad \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}} := \left\{\mathbf{n} \in \overline{\Omega} \;\middle|\; \mathbf{A}\mathbf{n} = \mathbf{b}\right\}. \tag{2.7}$$

We make the following assumption:

(H1) $\mathcal{M}_{\mathbf{A},\mathbf{b}}$ is nonempty.

Therefore, the single-phase chemical equilibrium problem can be expressed as

$$\mathbf{n} \in \left\{\arg\min G(\mathbf{n}) \;\middle|\; \mathbf{n} \in \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}\right\}. \tag{2.8}$$

### 2.2.1 Existence and uniqueness of a minimizer

The existence and uniqueness of a solution to the problem (2.8) for an ideal mixture have been thoroughly investigated by Shapiro and Shapley [60]. In their seminal work, they proved in Theorem 9.9 and Corollary 12.3 that for a single-phase ideal system, the problem (2.8) admits a unique solution, provided that the set $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ is compact. In our case, this compactness condition is satisfied, as will be demonstrated by Lemma 2.2.1. Furthermore, they proved in Theorem 9.2 that the inequality constraint is never saturated for this point, meaning that (2.8) can be rewritten as

$$\mathbf{n} \in \{\arg\min G(\mathbf{n}) \mid \mathbf{n} \in \mathcal{M}_{\mathbf{A},\mathbf{b}}\}. \tag{2.9}$$

Another important result worth mentioning, which proves the uniqueness of the solution, is the strict convexity of $G$ on $\mathcal{M}_{\mathbf{A},\mathbf{b}}$ for a single-phase ideal mixture, as proven by Smith and Missen in [64, Section 3.9].

In this section, we will extend the previous results to the more general case of mixtures which have a convex Gibbs energy but which may be non-ideal. To our knowledge, this more general case has not been studied in the literature. This extension

ensures the existence and uniqueness of solutions for a broader definition of the chemical potentials. Next, we will prove that the case of ideal mixtures covered by Shapiro and Shapley is a special case included within our more general result.

**Single-phase mixture with convex Gibbs energy**

Let $G : \Omega \to \mathbb{R}$ defined as $G(\mathbf{n}) = \sum_{i=1}^{N} n_i \mu_i(\mathbf{n})$ where the chemical potentials $\mu_i(\mathbf{n})$ do not have explicit formulas but where the function $G$ is still convex. Let us also take into account the set

$$\mathbb{X} := \left\{ \mathbf{x} \in \mathbb{R}^{N-1} \;\middle|\; x_i > 0,\ \forall i = 1, \dots, N-1, \text{ and } \sum_{i=1}^{N-1} x_i < 1 \right\}.$$

and define the function $\mathcal{G} : \mathbb{X} \to \mathbb{R}$ as

$$\mathcal{G}(\mathbf{x}) = \mathcal{G}(x_1, \dots, x_{N-1}) = G\left( x_1, \dots, x_{N-1}, 1 - \sum_{i=1}^{N-1} x_i \right). \tag{2.10}$$

To establish the existence and uniqueness of a solution to the minimization problem of $G$ on the set $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$, we will make the following assumptions:

(H2) $G \in \mathcal{C}^0(\overline{\Omega}, \mathbb{R}) \cap \mathcal{C}^2(\Omega, \mathbb{R})$ is a homogeneous function of degree 1;

(H3) $G$ is convex on $\overline{\Omega}$;

(H4) $\mu_i(\mathbf{n}) \to -\infty$ when $n_i \to 0$, $\forall i = 1, \dots, N$, $\forall \mathbf{n} \in \mathcal{M}_{\mathbf{A},\mathbf{b}}$;

(H5) the Hessian of $\mathcal{G}$, denoted $\mathcal{Q}$, is symmetric positive-definite on $\mathbb{X}$.

Using these assumptions, we obtain the following result.

**Theorem 2.2.1.** *Assuming that conditions (H1), (H2), (H3), (H4) and (H5) holds, then $G$ has a unique minimum in $\mathcal{M}_{\mathbf{A},\mathbf{b}}$.*

Before proving this theorem, we will first establish the existence of a minimizer of $G$ on $\mathcal{M}_{\mathbf{A},\mathbf{b}}$ by using hypotheses (H1), (H2), (H3) and (H4). Then, we will prove the uniqueness of this minimizer using hypothesis (H5).

**Lemma 2.2.1.** *The set $\mathcal{M}_{\mathbf{A},\mathbf{b}}$ is convex and its closure $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ is compact.*

*Proof.* The set $\mathcal{M}_{\mathbf{A},\mathbf{b}}$ is convex as the intersection of the two convex sets $\Omega$ and $\{\mathbf{n} \in \mathbb{R}^N \mid \mathbf{An} = \mathbf{b}\}$. We will now demonstrate that the set $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ is bounded. Indeed one has

$$\|\mathbf{b}\|_1 = \|\mathbf{An}\|_1 = \sum_{i=1}^{M} \left| \sum_{j=1}^{N} A_{ij} n_j \right| = \sum_{i=1}^{M} \sum_{j=1}^{N} A_{ij} n_j,$$

since $A_{ij} \geq 0$ and $n_j \geq 0$ for all $i \in \{1, \dots, M\}$ and $j \in \{1, \dots, N\}$. Moreover, none of the columns of the matrix $\mathbf{A}$ is zero, so

$$\underbrace{\min_{j\in\{1,\dots,N\}} \left( \sum_{i=1}^{M} A_{ij} \right)}_{>0} \|\mathbf{n}\|_1 \leq \sum_{j=1}^{N} \left( \sum_{i=1}^{M} A_{ij} \right) n_j = \|\mathbf{b}\|_1.$$

Therefore $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ is bounded. It follows that $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ is compact, as it is a closed and bounded subset of $\mathbb{R}^N$. □

**Lemma 2.2.2.** *Under assumptions (H1), (H2), (H3), (H4), there exists a unique minimum of problem 2.8 and this minimum is in $\mathcal{M}_{\mathbf{A},\mathbf{b}}$.*

*Proof.* From Lemma 2.2.1, the set $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ is compact. According to the Weierstrass theorem, since $G$ is a continuous function on a compact set, there exists at least one minimum value of $G$ in $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$. Let us prove that it belongs to $\mathcal{M}_{\mathbf{A},\mathbf{b}}$. Let $\mathbf{n}^\star \in \{\arg\min G(\mathbf{n}) \mid \mathbf{n} \in \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}\}$ be this minimizer. Let us assume that $\mathbf{n}^\star \in \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}} \setminus \mathcal{M}_{\mathbf{A},\mathbf{b}}$, meaning that there exists $\mathcal{J} \subsetneq \{1, \dots, N\}$ such that $n_j^\star = 0$ for all $j \in \mathcal{J}$. Note that the case $\mathcal{J} = \{1, \dots, N\}$ is excluded since $\mathbf{b} > 0$. Let $\mathbf{n} \in \mathcal{M}_{\mathbf{A},\mathbf{b}}$ which is assumed to be nonempty and $\varepsilon \in (0,1)$, then one defines $\mathbf{n}^0 := \mathbf{n} - \mathbf{n}^\star$ and $\mathbf{n}^\varepsilon := \mathbf{n}^\star + \varepsilon\mathbf{n}^0 = \varepsilon\mathbf{n} + (1-\varepsilon)\mathbf{n}^\star$. The vector $\mathbf{n}^\varepsilon$ is a convex linear combination of vectors of $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ which is a convex set according to Lemma 2.2.1, hence $\mathbf{n}^\varepsilon \in \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$. Furthermore, $\mathbf{n}^\varepsilon \in \mathcal{M}_{\mathbf{A},\mathbf{b}}$ since $\mathbf{n}^\varepsilon = \varepsilon\mathbf{n} + (1-\varepsilon)\mathbf{n}^\star \geq \varepsilon\mathbf{n} > 0$. By convexity of $G$ on $\overline{\Omega}$ from (H3),

$$G(\mathbf{n}^\star) \geq G(\mathbf{n}^\varepsilon) + \langle \boldsymbol{\mu}(\mathbf{n}^\varepsilon), \mathbf{n}^\star - \mathbf{n}^\varepsilon \rangle \Leftrightarrow \frac{G(\mathbf{n}^\star) - G(\mathbf{n}^\varepsilon)}{\varepsilon} \geq -\langle \boldsymbol{\mu}(\mathbf{n}^\varepsilon), \mathbf{n}^0 \rangle, \tag{2.11}$$

where $\boldsymbol{\mu} = \boldsymbol{\nabla} G$.

We will now take the limit when $\varepsilon$ tends to 0 in inequality (2.11). In the right-hand side one has

$$\lim_{\varepsilon\to 0} -\langle \boldsymbol{\mu}(\mathbf{n}^\varepsilon), \mathbf{n}^0 \rangle = -\sum_{j\in\mathcal{J}} n_j^0 \lim_{\varepsilon\to 0} \mu_j(\mathbf{n}^\varepsilon) - \sum_{i=1,\, i\notin\mathcal{J}}^{N} n_i^0 \lim_{\varepsilon\to 0} \mu_i(\mathbf{n}^\varepsilon). \tag{2.12}$$

Noting that $\lim_{\epsilon\to 0} \mathbf{n}^\varepsilon = \mathbf{n}^\star$ and in particular that $\lim_{\varepsilon\to 0} n_j^\varepsilon = n_j^\star = 0$, for all $j \in \mathcal{J}$, it follows from (H4) and the continuity of $\boldsymbol{\mu}$ that

$$\lim_{\varepsilon\to 0} \mu_j(\mathbf{n}^\varepsilon) = -\infty, \forall j \in \mathcal{J} \quad \text{and} \quad \lim_{\varepsilon\to 0} \mu_i(\mathbf{n}^\varepsilon) = \mu_i(\mathbf{n}^\star) \in \mathbb{R}, \forall i \in \{1, \dots, N\} \setminus \mathcal{J}. \tag{2.13}$$

By combining (2.12) and (2.13) with $n_j^0 = n_j > 0$, for all $j \in \mathcal{J}$, one finds that the right-hand side of (2.11) tends to $+\infty$. However if $\mathbf{n}^\star$ minimizes $G$ on $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$, then the left-hand side of (2.11) is non-positive which is a contradiction. Therefore $\mathcal{J} = \emptyset$ and $\mathbf{n}^\star \in \mathcal{M}_{\mathbf{A},\mathbf{b}}$. □

Having established the existence of a minimizer, we will now prove its uniqueness by demonstrating that $G$ is strictly convex on $\mathcal{M}_{\mathbf{A},\mathbf{b}}$. To accomplish this, we will establish an equality between $\ker \mathbf{Q}(\mathbf{n})$, where $\mathbf{Q}$ is the Hessian of $G$, and $\mathrm{Vect}\{\mathbf{n}\}$, the vector space generated by $\mathbf{n}$. This should be thought as a strong version of the Gibbs-Duhem relation (2.6).

**Proposition 2.2.1.** *Assume that hypotheses (H2), (H3) and (H5) are satisfied. Let $\mathbf{n} \in \Omega$ and let $\mathbf{Q}$ be the Hessian of $G$, then $\mathrm{Vect}\{\mathbf{n}\} = \ker \mathbf{Q}(\mathbf{n})$.*

*Proof.* Let us start by showing the inclusion $\mathrm{Vect}\{\mathbf{n}\} \subset \ker \mathbf{Q}(\mathbf{n})$, let us take $h > 0$ and use the 1-homogeneity of $G$ to write

$$G((1+h)\mathbf{n}) = G(\mathbf{n}) + hG(\mathbf{n}). \tag{2.14}$$

The left-hand side of (2.14) can also be written using a Taylor expansion:

$$G((1+h)\mathbf{n}) = G(\mathbf{n}) + h\langle \boldsymbol{\mu}(\mathbf{n}), \mathbf{n}\rangle + \frac{h^2}{2}\langle \mathbf{Q}(\mathbf{n})\cdot\mathbf{n}, \mathbf{n}\rangle + o(h^2). \tag{2.15}$$

By identifying the terms in (2.14) and (2.15), one finds that

$$G(\mathbf{n}) = \langle \boldsymbol{\mu}(\mathbf{n}), \mathbf{n}\rangle \quad \text{and} \quad \langle \mathbf{Q}(\mathbf{n})\cdot\mathbf{n}, \mathbf{n}\rangle = 0.$$

Moreover $\mathbf{Q}(\mathbf{n}) \in \mathcal{S}_N^+(\mathbb{R})$ since it is the Hessian of a convex function, it follows that $\mathbf{n} \in \ker \mathbf{Q}(\mathbf{n})$. Therefore $\mathrm{Vect}\{\mathbf{n}\} \subset \ker \mathbf{Q}(\mathbf{n})$, which is nothing but a rewriting of the Gibbs-Duhem relation (2.6).

To prove the inclusion $\ker \mathbf{Q}(\mathbf{n}) \subset \mathrm{Vect}\{\mathbf{n}\}$, let $\mathbf{w} \in \{\mathbf{1}\}^\perp := \{\mathbf{w} \in \mathbb{R}^N \mid \sum_{i=1}^N w_i = 0\}$, $\mathbf{w} \neq \mathbf{0}$, $\mathbf{n} \in \Omega$ and let $h > 0$ small enough to ensure that $\mathbf{n} + h\mathbf{w} \in \Omega$, then

$$\sum_{i=1}^N (n_i + hw_i) = \sum_{i=1}^N n_i = \langle \mathbf{n}, \mathbf{1}\rangle.$$

Hence the mole-fractions of $\mathbf{n} + h\mathbf{w}$ are written

$$x_i(\mathbf{n} + h\mathbf{w}) = \frac{n_i + hw_i}{\langle \mathbf{n}, \mathbf{1}\rangle} = x_i(\mathbf{n}) + h\frac{w_i}{\langle \mathbf{n}, \mathbf{1}\rangle},\ i = 1, \ldots, N$$

It follows from the 1-homogeneity of $G$ and the definition of $\mathcal{G}$, given by (2.10), that

$$G(\mathbf{n}+h\mathbf{w}) = \langle \mathbf{n}, \mathbf{1}\rangle G\left(\frac{\mathbf{n}+h\mathbf{w}}{\langle \mathbf{n}, \mathbf{1}\rangle}\right) = \langle \mathbf{n}, \mathbf{1}\rangle \mathcal{G}\left(\tilde{\mathbf{x}} + h\frac{\tilde{\mathbf{w}}}{\langle \mathbf{n}, \mathbf{1}\rangle}\right), \tag{2.16}$$

where $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{w}}$ are the mole fractions vector $\mathbf{x}$ and the vector $\mathbf{w}$ without their last coordinate. Therefore, a Taylor expansion of $\mathcal{G}$ in (2.16) yields:

$$\begin{aligned} G(\mathbf{n}+h\mathbf{w}) &= \langle \mathbf{n}, \mathbf{1}\rangle \left[\mathcal{G}(\tilde{\mathbf{x}}) + \frac{h}{\langle \mathbf{n}, \mathbf{1}\rangle}\langle \boldsymbol{\nabla}\mathcal{G}(\tilde{\mathbf{x}}), \tilde{\mathbf{w}}\rangle + \frac{1}{2}\left(\frac{h}{\langle \mathbf{n}, \mathbf{1}\rangle}\right)^2 \langle \boldsymbol{\mathcal{Q}}(\tilde{\mathbf{x}})\tilde{\mathbf{w}}, \tilde{\mathbf{w}}\rangle + o(h^2)\right] \\ &= \langle \mathbf{n}, \mathbf{1}\rangle \mathcal{G}(\tilde{\mathbf{x}}) + h\langle \boldsymbol{\nabla}\mathcal{G}(\tilde{\mathbf{x}}), \tilde{\mathbf{w}}\rangle + \frac{1}{2}\frac{h^2}{\langle \mathbf{n}, \mathbf{1}\rangle}\langle \boldsymbol{\mathcal{Q}}(\tilde{\mathbf{x}})\tilde{\mathbf{w}}, \tilde{\mathbf{w}}\rangle + o(h^2). \end{aligned}$$

But on the other hand a Taylor expansion of $G(\mathbf{n}+h\mathbf{w})$ yields:

$$G(\mathbf{n}+h\mathbf{w}) = G(\mathbf{n}) + h\langle \boldsymbol{\nabla}G(\mathbf{n}), \mathbf{w}\rangle + \frac{1}{2}h^2\langle \mathbf{Q}(\mathbf{n})\mathbf{w}, \mathbf{w}\rangle + o(h^2).$$

Then by identification:

$$\langle \mathbf{Q}(\mathbf{n})\mathbf{w}, \mathbf{w}\rangle = \frac{1}{\langle \mathbf{n}, \mathbf{1}\rangle}\langle \boldsymbol{\mathcal{Q}}(\tilde{\mathbf{x}})\tilde{\mathbf{w}}, \tilde{\mathbf{w}}\rangle \tag{2.17}$$

From (H5), the matrix $\boldsymbol{\mathcal{Q}}$ is positive-definite on $\mathbb{X}$, it follows from $\mathbf{w} \neq 0$ that the right-hand side of (2.17) is positive. Therefore

$$\langle \mathbf{Q}(\mathbf{n})\mathbf{w}, \mathbf{w}\rangle = 0 \Leftrightarrow \mathbf{w} = 0 \Leftrightarrow \{\mathbf{1}\}^{\perp} \cap \ker \mathbf{Q}(\mathbf{n}) = \{\mathbf{0}\}.$$

We deduce from this last relation that

$$\begin{aligned} N \geq \dim(\ker \mathbf{Q}(\mathbf{n}) + \{\mathbf{1}\}^{\perp}) &= \dim(\ker \mathbf{Q}(\mathbf{n})) + \dim(\{\mathbf{1}\}^{\perp}) - \dim(\{\mathbf{1}\}^{\perp} \cap \ker \mathbf{Q}(\mathbf{n})) \\ &= \dim(\ker \mathbf{Q}(\mathbf{n})) + (N-1) - 0. \end{aligned}$$

Therefore $\dim(\ker \mathbf{Q}(\mathbf{n})) \leq 1$ and since it has been proven that $\mathrm{Vect}\{\mathbf{n}\} \subset \ker \mathbf{Q}(\mathbf{n})$, it follows that $\mathrm{Vect}\{\mathbf{n}\} = \ker \mathbf{Q}(\mathbf{n})$. □

We can then prove the strict convexity of $G$ on $\mathcal{M}_{\mathbf{A},\mathbf{b}}$.

**Lemma 2.2.3.** *The function $G$ is strictly convex on $\mathcal{M}_{\mathbf{A},\mathbf{b}}$.*

*Proof.* Let $\mathbf{n}_1, \mathbf{n}_2 \in \mathcal{M}_{\mathbf{A},\mathbf{b}}$, $\mathbf{n}_1 \neq \mathbf{n}_2$, and let $\mathbf{u} := \mathbf{n}_1 - \mathbf{n}_2$. Due to the convexity of $G$ on $\mathcal{M}_{\mathbf{A},\mathbf{b}}$, one has

$$\langle \mathbf{u}, \mathbf{Q}(\mathbf{n}_1)\cdot\mathbf{u}\rangle \geq 0. \tag{2.18}$$

However $\mathbf{u} \in \ker \mathbf{A} \backslash \{\mathbf{0}\}$ and according to Lemma 2.1.2 the components of $\mathbf{u}$ do not have all the same sign, it follows that $\mathbf{u} \notin \mathrm{Vect}\{\mathbf{n}_1\}$ since $\mathbf{n}_1 \in \Omega$. Thus, based on Proposition 2.2.1, it can be concluded that $\mathbf{u} \notin \ker \mathbf{Q}(\mathbf{n}_1)$. Therefore, inequality (2.18) is strict and so $G$ is strictly convex on $\mathcal{M}_{\mathbf{A},\mathbf{b}}$. □

It is now possible to prove the uniqueness of the minimizer of the function $G$ on $\mathcal{M}_{\mathbf{A},\mathbf{b}}$.

*Proof of Theorem* 2.2.1. From Lemma 2.2.2, $G$ has a minimum on $\mathcal{M}_{\mathbf{A},\mathbf{b}}$. Furthermore, the strict convexity of $G$ on this set, as proven in Lemma 2.2.3, guarantees its uniqueness. □

**Single-phase ideal mixture**

In this section, attention is limited to the case of an ideal activity model based on mole fractions in which the function $G : \Omega \to \mathbb{R}$ is written

$$G(\mathbf{n}) = \sum_{i=1}^{N} n_i \mu_i(\mathbf{n}) \quad \text{with} \quad \mu_i(\mathbf{n}) := \mu_i^\circ + \mathrm{RT} \ln(n_i / \langle \mathbf{n}, \mathbf{1} \rangle). \tag{2.19}$$

Assuming (H1), we will prove that this function satisfies assumptions (H2)–(H5) in order to apply Theorem 2.2.1. First of all, this function is clearly continuous on $\Omega$. By decomposing it as

$$G(\mathbf{n}) = \sum_{i=1}^{N} n_i \mu_i^\circ + \mathrm{RT} \sum_{i=1}^{N} n_i \ln(n_i) - \mathrm{RT} \sum_{i=1}^{N} n_i \ln \sum_{j=1}^{N} n_j$$

and using the limit $\lim_{t \to 0^+} t \ln(t) = 0$, we can define a continuous extension of $G$ on $\overline{\Omega}$. Therefore $G : \overline{\Omega} \to \mathbb{R}$ is continuous. Let $\mathbf{n} \in \Omega$, the partial derivatives of $G$ obtained from (2.19) are

$$\frac{\partial G}{\partial n_i}(\mathbf{n}) = \mu_i(\mathbf{n}) + \sum_{j=1}^{N} n_j \frac{\partial \mu_j}{\partial n_i}(\mathbf{n}), \tag{2.20}$$

where the partial derivatives of the chemicals potentials are given by

$$\frac{\partial \mu_j}{\partial n_i}(\mathbf{n}) = \mathrm{RT} \left( \frac{\delta_{ij}}{n_i} - \frac{1}{\langle \mathbf{n}, \mathbf{1} \rangle} \right) \tag{2.21}$$

where $\delta_{ij}$ is the Kronecker delta function. By incorporating the formula (2.21) into the sum in (2.20), we obtain:

$$\sum_{j=1}^{N} n_j \frac{\partial \mu_j}{\partial n_i}(\mathbf{n}) = \mathrm{RT} \sum_{j=1}^{N} n_j \left( \frac{\delta_{ij}}{n_i} - \frac{1}{\langle \mathbf{n}, \mathbf{1} \rangle} \right) = \mathrm{RT} \left( 1 - \frac{\sum_{j=1}^{N} n_j}{\langle \mathbf{n}, \mathbf{1} \rangle} \right) = 0.$$

As a consequence, the Gibbs-Duhem relation (2.6) is satisfied and $G$ is a homogeneous function of degree 1. The partial derivatives of $G$ are then given by the chemical potentials which are continuous on $\Omega$. Furthermore the entries of its Hessian $\mathbf{Q}$, given by

$$\frac{\partial^2 G}{\partial n_i \partial n_j}(\mathbf{n}) = \frac{\partial \mu_j}{\partial n_i}(\mathbf{n}) = \mathrm{RT} \left( \frac{\delta_{ij}}{n_i} - \frac{1}{\langle \mathbf{n}, \mathbf{1} \rangle} \right), \; i, j = 1, \ldots, N,$$

are also continuous on $\Omega$. Therefore $G \in \mathcal{C}^0(\overline{\Omega}, \mathbb{R}) \cap \mathcal{C}^2(\Omega, \mathbb{R})$, meaning that the assumption (H2) is satisfied.

The assumption (H3) states that $G$ is convex on $\overline{\Omega}$. To prove this property, we can begin by showing that its Hessian $\mathbf{Q}$ is positive semi-definite on $\Omega$. Let $\mathbf{n}, \mathbf{m} \in \Omega$, $\mathbf{n} \neq \mathbf{m}$ and let $\mathbf{h} := \mathbf{n} - \mathbf{m}$, then

$$\frac{1}{\mathrm{RT}} \mathbf{h}^T \mathbf{Q}(\mathbf{n}) \mathbf{h} = \sum_{i=1}^{N} \frac{h_i^2}{n_i} - \frac{1}{\langle \mathbf{n}, \mathbf{1} \rangle} \left( \sum_{i=1}^{N} h_i \right)^2 = \sum_{i=1}^{N} n_i \left( \frac{h_i}{n_i} - \frac{1}{\langle \mathbf{n}, \mathbf{1} \rangle} \sum_{j=1}^{N} h_j \right)^2 \geq 0, \tag{2.22}$$

since $\mathbf{n} > 0$. It follows from the continuity of $G$ on $\overline{\Omega}$ that $G$ is convex on $\overline{\Omega}$. Therefore, the assumption (H3) is satisfied.

It is clear from the definition of the logarithm that for $\mathbf{n} \in \mathcal{M}_{\mathbf{A},\mathbf{b}}$, one has

$$\mu_i(\mathbf{n}) = \mu_i^\circ + \mathrm{RT} \ln(n_i / \langle \mathbf{n}, \mathbf{1} \rangle) \to -\infty \quad \text{when} \quad n_i \to 0.$$

Hence condition (H4) is also fulfilled.

The last remaining step is to show that the Hessian of $\mathcal{G}$, referred to as $\boldsymbol{\mathcal{Q}}$ in the following, is positive definite on $\mathbb{X}$ to satisfy assumption (H5). Let $\mathbf{x} \in \mathbb{X}$, the expression of $\mathcal{G}(\mathbf{x})$ is given by:

$$\mathcal{G}(\mathbf{x}) = \mu_N^\circ + \sum_{i=1}^{N-1} x_i (\mu_i^\circ - \mu_N^\circ) + \mathrm{RT} \left[ \sum_{i=1}^{N-1} x_i \ln(x_i) + \left( 1 - \sum_{j=1}^{N-1} x_j \right) \ln \left( 1 - \sum_{j=1}^{N-1} x_j \right) \right].$$

It follows that

$$\frac{\partial \mathcal{G}}{\partial x_i}(\mathbf{x}) = \mu_i^\circ - \mu_N^\circ + \mathrm{RT}\left[\ln(x_i) - \ln\left(1 - \sum_{j=1}^{N-1} x_j\right)\right],\ i = 1, \ldots, N-1,$$

and

$$\frac{\partial^2 \mathcal{G}}{\partial x_i \partial x_j}(\mathbf{x}) = \mathrm{RT}\left(\frac{\delta_{ij}}{x_i} + \frac{1}{1 - \sum_{k=1}^{N-1} x_k}\right),\ i, j = 1, \ldots, N-1.$$

Let $\mathbf{x}, \mathbf{y} \in \mathbb{X}$, $\mathbf{x} \neq \mathbf{y}$ and let $\mathbf{h} := \mathbf{x} - \mathbf{y}$, then

$$\frac{1}{\mathrm{RT}}\mathbf{h}^T \mathcal{Q}(\mathbf{n})\mathbf{h} = \sum_{i=1}^{N-1} \frac{h_i^2}{x_i} + \frac{1}{1 - \sum_{k=1}^{N-1} x_k}\left(\sum_{i=1}^{N-1} h_i\right)^2 \geq 0, \tag{2.23}$$

since $\mathbf{x} > 0$ and $1 - \sum_{k=1}^{N-1} x_k > 0$. Furthermore, since equality in (2.23) is attained only when $\mathbf{h} = 0$, the assumption (H5) is satisfied.

All conditions required to apply Theorem 2.2.1 are fulfilled. Hence, the function $G$ defined as (2.19) has a unique minimum on $\mathcal{M}_{\mathbf{A},\mathbf{b}}$.

### 2.2.2 From minimization problem to algebraic equations

In this section, we will establish an equivalent system that can be solved to find the unique solution of the ideal single-phase chemical equilibrium problem.

**Proposition 2.2.2.** *Assuming that $\mathcal{M}_{\mathbf{A},\mathbf{b}}$ is nonempty, then*

$$\mathbf{n} \in \{\arg\min G(\mathbf{n}) \mid \mathbf{n} \in \mathcal{M}_{\mathbf{A},\mathbf{b}}\} \quad \Leftrightarrow \quad \begin{aligned} \mathbf{A}\mathbf{n} &= \mathbf{b}, \\ \mathbf{S}^T \boldsymbol{\mu}(\mathbf{n}) &= \mathbf{0}, \end{aligned} \tag{2.24}$$

*where $\boldsymbol{\mu}$ is the vector of chemical potentials.*

*Proof.* The first order optimality conditions of the constrained minimization problem

$$\min_{\mathbf{A}\mathbf{n}=\mathbf{b}} G(\mathbf{n}) \tag{2.25}$$

are given by the Euler-Lagrange equations which state that if $\mathbf{n}^\star$ is the unique solution of (2.25), it must satisfy

$$\mathbf{A}\mathbf{n}^\star - \mathbf{b} = \mathbf{0}, \tag{2.26}$$

$$\nabla G(\mathbf{n}^\star) + \mathbf{A}^T \boldsymbol{\Lambda} = \mathbf{0}, \tag{2.27}$$

where $\nabla G$ is the gradient of $G$ and $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_M)^T$ is the Lagrange multipliers vector.

We can simplify equation (2.27) by eliminating the Lagrange multipliers. To achieve this, we multiply (2.27) by $\mathbf{S}^T$, and as shown by Lemma 2.1.1, the matrix product $\mathbf{S}^T\mathbf{A}^T$ vanishes. Consequently, (2.27) reduces to $\mathbf{S}^T\nabla G(\mathbf{n}^\star) = \mathbf{0}$. By denoting the vector of chemical potentials as $\boldsymbol{\mu} = \nabla G$, we obtain the system (2.24).

It remains to prove that $\mathbf{n}^\star$ is the unique solution of the system (2.24). To do so, we assume the existence of $\overline{\mathbf{n}} \in \Omega$ that satisfies (2.24). Then, one has $\mathbf{n}^\star - \overline{\mathbf{n}} \in \ker \mathbf{A}$ and $\boldsymbol{\mu}(\mathbf{n}^\star) - \boldsymbol{\mu}(\overline{\mathbf{n}}) \in \ker \mathbf{S}^T$. By Lemma 2.1.1, we know that $\ker \mathbf{S}^T = (\ker \mathbf{A})^\perp$, it follows that

$$\langle \mathbf{n}^\star - \overline{\mathbf{n}}, \boldsymbol{\mu}(\mathbf{n}^\star) - \boldsymbol{\mu}(\overline{\mathbf{n}}) \rangle = 0.$$

The strict convexity of the function $G$ on the set $\mathcal{M}_{\mathbf{A},\mathbf{b}}$, as established in Lemma 2.2.3, implies the strict monotonicity of its gradient $\boldsymbol{\mu}$. This leads to the conclusion that the optimal solution $\mathbf{n}^\star$ must be equal to the vector $\overline{\mathbf{n}}$. □

### 2.2.3 Reformulation of the system

The formulation of the system (2.24) contains a strong non-linearity in the definition of the activity in the chemical potential, which can lead to considerable difficulty in solving these equations. We have proposed two new equivalent formulations for this problem in order to reduce this non-linearity, one based on quantities of matter and the other on mole fractions. After presenting these formulations and demonstrating their equivalence, we will present numerical results to justify the choice of the mole fraction formulation.

#### Activity formulation

The activity formulation is the one presented in (2.24). The problem is to find $\mathbf{n}$ such that

$$\begin{aligned} \mathbf{A}\mathbf{n} &= \mathbf{b} \\ \mathbf{S}^T\boldsymbol{\mu}(\mathbf{n}) &= \mathbf{0} \end{aligned} \tag{2.28}$$

where

$$\mu_i(\mathbf{n}) = \mu_i^\circ/(\mathrm{RT}) + \mathrm{RT}\ln(n_i/\langle \mathbf{n}, \mathbf{1} \rangle),\ i = 1, \ldots, N. \tag{2.29}$$

**Quantity formulation**

We introduce two new variables $s$ and $l$ such that

$$s := \langle \mathbf{n}, \mathbf{1} \rangle \quad \text{and} \quad l := \ln s.$$

From the property of the logarithm, we can rewrite the logarithm in (2.29) as

$$\ln(n_i / \langle \mathbf{n}, \mathbf{1} \rangle) = \ln(n_i) - l. \tag{2.30}$$

Therefore the problem to solve is: find $(\mathbf{n}, s, l)$ such that

$$\begin{aligned} \mathbf{A}\mathbf{n} &= \mathbf{b}, \\ \mathbf{S}^T \boldsymbol{\mu}(\mathbf{n}, l) &= \mathbf{0}, \\ s - \langle \mathbf{n}, \mathbf{1} \rangle &= 0, \\ \ln(s) - l &= 0, \end{aligned} \tag{2.31}$$

where

$$\mu_i(n_i, l) = \mu_i^\circ / (\mathrm{RT}) + \ln n_i - l, \; i = 1, \ldots, N.$$

By construction, it is clear that this system is equivalent to the one of the activity formulation.

**Mole fraction formulation**

We introduce the new variable:

$$\omega := 1 / \langle \mathbf{n}, \mathbf{1} \rangle. \tag{2.32}$$

Then, multiplying the element conservation equations by $\omega$ leads to $\mathbf{A}\mathbf{x} = \omega\mathbf{b}$, where $\mathbf{x} = \omega\mathbf{n}$ is the vector of mole fractions. The unknowns become the $N + 1$ variables $\mathbf{x}$ and $\omega$. Furthermore, since there are only $N$ equations, the addition of one more equation is needed. A fundamental property of the mole fractions is that $\langle \mathbf{x}, \mathbf{1} \rangle = 1$ which can be the additional equation. Thus the problem to solve becomes: find $(\mathbf{x}, \omega)$ such that

$$\begin{aligned} \mathbf{A}\mathbf{x} - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T \mathbf{y}(\mathbf{x}) &= \mathbf{d}, \\ \langle \mathbf{x}, \mathbf{1} \rangle &= 1, \end{aligned} \tag{2.33}$$

where

$$(\mathbf{y}(\mathbf{x}))_i = y(x_i) := \ln x_i,\ i = 1, \dots, N \quad \text{and} \quad \mathbf{d} := -\mathbf{S}^T \boldsymbol{\mu}^\circ/(\mathrm{RT}).$$

**Proposition 2.2.3.** *The system* (2.33) *is equivalent to the system* (2.24).

*Proof.* Let $\mathbf{n}^\star$ be the unique solution of (2.24) and let $\omega = 1/\langle \mathbf{n}^\star, \mathbf{1} \rangle > 0$. Then by construction it is clear that $(\omega \mathbf{n}^\star, \omega)$ solves (2.33). In particular, this ensures the existence of a point $(\mathbf{x}, \omega)$ verifying (2.33). Moreover, if $(\mathbf{x}, \omega)$ solves (2.33), then $\omega \neq 0$. Indeed, if this were not the case, then $\mathbf{x}$ would belong to $\ker \mathbf{A}$, implying from Lemma 2.1.2 that its coefficients are not all of the same sign, which is not compatible with the logarithm. Thus $\mathbf{n} = \mathbf{x}/\omega$ verifies (2.24). □

### Choice of the formulation

We conducted a convergence study of Newton's algorithm for these different formulations in order to select the best one. Using the Seawater test case (see Appendix A.2.2), we studied the number of iterations of the algorithm for different difficult initializations. A reminder of Newton's method and details of the methods used (log trick, Param. $L^\infty$ and Cart. repr. $L^\infty$) can be found in Chapter 3. Convergence has been studied with and without Newton's increment limiter. Indeed, many of the cases of divergence come from the fact that Newton's iterates become very high, so it is interesting to limit the evolution of these iterates based on the physics of the system. The right-hand side vector $\mathbf{b}$ gives a limit on the maximum number of moles that the chemical species can contain, if a quantity of species is low it is useless that it varies more than necessary, on the contrary if it is too high compared to the total quantity of matter, it is necessary that it can decrease as much as necessary. Thus after each calculation of Newton's direction for iteration $k+1$, the evolution of each unknown of the system from step $k$ is carried out by limiting its variation: - with respect to the norm 1 of the vector $\mathbf{b}$ if this unknown is in a range of value in accordance with physics; - with respect to its own value if this one is higher than the norm 1 of $\mathbf{b}$. Details of the limiter used for each method are given in the Appendix B.1.

The results in Table 2.1 show that, without a limiter, the activity and quantity formulations are not very robust to initialization, whereas the mole fraction formulation converges in all cases except one. Adding a limiter allows this case to converge, and also makes the activity formulation more robust than before. However, the number of iterations for this formulation can reach 90, which is very high compared to the other cases.

| Initial guess (mol) | | | |
|---|---|---|---|
| $H_2O$ | 55 | 55 | $0_\epsilon$ |
| Other species | 55 | $0_\epsilon$ | $0_\epsilon$ |
| ***Activity formulation*** | | | |
| log trick | 34 | × | × |
| Param. $L^\infty$ | × | × | × |
| Cart. repr. $L^\infty$ | × | × | × |
| ***Quantity formulation*** | | | |
| log trick | × | × | × |
| Param. $L^\infty$ | × | 24 | 25 |
| Cart. repr. $L^\infty$ | **22** | **18** | × |
| ***Mole fraction formulation*** | | | |
| log trick | 38 | × | 38 |
| Param. $L^\infty$ | 23 | 29 | **21** |
| Cart. repr. $L^\infty$ | 23 | 22 | 23 |

**(a)** Newton's method without limiter.

| Initial guess (mol) | | | |
|---|---|---|---|
| $H_2O$ | 55 | 55 | $0_\epsilon$ |
| Other species | 55 | $0_\epsilon$ | $0_\epsilon$ |
| ***Activity formulation*** | | | |
| log trick | 26 | 27 | 37 |
| Param. $L^\infty$ | 22 | **23** | 22 |
| Cart. repr. $L^\infty$ | 90 | 27 | 26 |
| ***Quantity formulation*** | | | |
| log trick | 34 | 27 | × |
| Param. $L^\infty$ | 31 | × | × |
| Cart. repr. $L^\infty$ | **21** | 26 | 34 |
| ***Mole fraction formulation*** | | | |
| Param. log | 25 | 26 | 22 |
| Param. $L^\infty$ | 23 | 24 | **21** |
| Cart. repr. $L^\infty$ | 23 | 24 | 29 |

**(b)** Newton's method with limiter.

**Table 2.1 –** Number of iterations of the methods and formulations with the Seawater test case and various difficult initializations, for each experiment the best result obtained is in bold, the cross × indicates that there was a divergence of Newton's algorithm.

Based on these results, we chose to use the mole fraction formulation for its robustness and homogeneous number of iterations between the different methods and initializations.

### 2.2.4 Other types of constraint considered

System (2.33) is the simplest form of chemical equilibrium calculation that can be performed. It is possible to replace one or more of the constraints on element conservation with others. There is a wide choice of constraints [79], but the ones we will use are charge conservation and redox constraints.

#### Charge conservation constraint

When defining the matrix formula **A**, it is possible to consider the conservation of charge instead of the conservation of one of the elements. It is thus common to replace the hydrogen conservation line H by the charge Z. For instance, The matrix formula for

water dissociation becomes

$$\mathbf{A} = \begin{array}{c} \begin{array}{ccc} \mathrm{H^+} & \mathrm{OH^-} & \mathrm{H_2O} \end{array} \\ \begin{bmatrix} 0 & 1 & 1 \\ 1 & -1 & 0 \end{bmatrix} \end{array} \begin{array}{c} \mathrm{O} \\ \mathrm{Z} \end{array}.$$

It is also necessary to adapt the coefficient of the vector **b** corresponding to the charge.

**Redox constraint**

The concept of electrochemical potential pE (see [68]) allows for the description of the tendency of electron transfer between species in a chemical system. This potential is a dimensionless number defined as a measure of the activity of electrons $a_{e^-}$. it is written as

$$\mathrm{pE} = -\log_{10}(a_{e^-}). \tag{2.34}$$

In order to constrained the value of pE in system (2.33), we consider the electron chemical potential:

$$\mu_{e^-} = \mu^\circ_{e^-} + \mathrm{RT} \ln a_{e^-}, \tag{2.35}$$

where $\mu^\circ_{e^-}$ is a standard chemical potential for the electron to be computed from a thermodynamic database. By substituting equation (2.34) into equation (2.35), we can derive the equality

$$\mu_{e^-} = \mu^\circ_{e^-} - \mathrm{pE} \times \mathrm{RT} \ln 10. \tag{2.36}$$

In our framework, this constraint is linked to a primary species, resulting in fewer chemical elements than the number of primary species. To address this issue, we consider the charge Z as an element of the set of chemical elements $\mathcal{E}$. Furthermore, we need to treat the electron as a fictitious secondary species, implying that its quantity is not an unknown in the system, and introduce a half-reaction as the $N - M + 1$-th reaction:

$$-\mathrm{e^-(v)} + \sum_{i \in \mathcal{C}} s_{i,N-M+1} C_i = 0.$$

However, since the quantity of electron is not an unknown, we need to remove an equation from the system. Considering the matrix $\tilde{\mathbf{A}}$ obtained by deleting a row from **A**,

depending on the quantity we choose to keep, the system to solve becomes

$$\begin{aligned}
\tilde{\mathbf{A}}\mathbf{x} - \omega\mathbf{b} &= \mathbf{0}, \\
\mathbf{S}^T\mathbf{y}(\mathbf{x}) &= \mathbf{d}, \\
\mathbf{A}_{Pr}^{-1}\mathbf{a}_{\mathrm{e}^-}\begin{bmatrix}\mathbf{y}(\mathbf{x}_{Pr}) + \boldsymbol{\mu}_{Pr}^{\circ}/(\mathrm{RT}) \\ \mu_{e^-}\end{bmatrix} &= \mathbf{0}, \\
\langle \mathbf{x}, \mathbf{1}\rangle &= 1.
\end{aligned}$$

where $\mathbf{a}_{\mathrm{e}^-}$ is the formula vector of the electron. In this way, $\mathbf{A}_{Pr}^{-1}\mathbf{a}_{\mathrm{e}^-}$ corresponds to the stoichiometric coefficients of the half-reaction.

As an example, let us consider the following chemical system:

$$\begin{aligned}
\mathcal{C} &= \{\mathrm{H_2O}, \mathrm{H^+}, \mathrm{H_2(aq)}, \mathrm{HO_2^-}, \mathrm{O_2(aq)}, \mathrm{OH^-}, \mathrm{H_2O_2(aq)}\}, \\
\mathcal{E} &= \{\mathrm{O}, \mathrm{H}, \mathrm{Z}\}, \\
\mathcal{R} &= \{ \quad \mathrm{HO_2^-} = 2\mathrm{H_2O} - \mathrm{H^+} - \mathrm{H_2(aq)}, \\
&\qquad \mathrm{O_2(aq)} = 2\mathrm{H_2O} - 2\mathrm{H_2}, \\
&\qquad \mathrm{OH^-} = \mathrm{H_2O} - \mathrm{H^+}, \\
&\qquad \mathrm{H_2O_2(aq)} = 2\mathrm{H_2O} - \mathrm{H_2(aq)}\}.
\end{aligned}$$

The corresponding half-reaction is the proton reduction:

$$-\mathrm{H^+} + \frac{1}{2}\mathrm{H_2(aq)} = \mathrm{e^-(v)}.$$

The formula matrix is defined as

$$\mathbf{A} = \begin{array}{c} \begin{matrix} \mathrm{H_2O} & \mathrm{H^+} & \mathrm{H_2(aq)} & \mathrm{O_2(aq)} & \mathrm{HO_2^-} & \mathrm{OH^-} & \mathrm{H_2O_2(aq)} \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 2 & 2 & 1 & 2 \\ 2 & 1 & 2 & 0 & 1 & 1 & 2 \\ 0 & 1 & 0 & 0 & -1 & -1 & 0 \end{bmatrix} \begin{matrix} \mathrm{O} \\ \mathrm{H}, \\ \mathrm{Z} \end{matrix} \end{array}$$

and the formula vector as

$$\mathbf{a}_{\mathrm{e}^-} = \begin{array}{c} \begin{matrix} \mathrm{e^-(v)} \end{matrix} \\ \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} \begin{matrix} \mathrm{O} \\ \mathrm{H}. \\ \mathrm{Z} \end{matrix} \end{array}$$

Once the stoichiometry matrix $\mathbf{S}$ is constructed, one row in $\mathbf{A}$ must be removed to

ensure that the resulting system has an equal number of unknowns and equations.

## 2.3 Multiphase chemical equilibrium problem

In a multiphase equilibrium scenario, we must consider $N_{Ph}$ potentially present phases. Utilizing the map $\sigma$ defined in (2.1), we can define the vectors of quantities $\mathbf{n}^\alpha = (n_i)_{i\in\sigma^{-1}(\alpha)}$ for each phase $\alpha$. The chemical equilibrium problem is then defined as the search for the species quantities vector $\mathbf{n} = (\mathbf{n}^\alpha)_{\alpha=1,\dots,N_{Ph}}$ that minimizes the function $G$, subject to constraints of element conservation and non-negativity, at a fixed temperature T, pressure P, and element quantities $\mathbf{b}$. Utilizing the set of constraints $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ defined in equation (2.7) and assumed to be non-empty as in (H1), we can formulate the multiphase chemical equilibrium problem as follows:

$$\mathbf{n} \in \arg\min\left\{G(\mathbf{n}) \;\middle|\; \mathbf{n} \in \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}\right\}, \tag{2.37}$$

where the Gibbs energy function $G(\mathbf{n})$ is defined as the sum of the Gibbs energies of individual phases:

$$G(\mathbf{n}) = \sum_{i=1}^{N_{Ph}} G_\alpha(\mathbf{n}^\alpha) \quad \text{with} \quad G_\alpha(\mathbf{n}^\alpha) = \sum_{i\in\sigma^{-1}(\alpha)} n_i[\mu_i^\circ + \mathrm{RT}\ln(n_i/\langle\mathbf{n}^\alpha, \mathbf{1}\rangle)]. \tag{2.38}$$

### 2.3.1 Existence and uniqueness of a minimizer

The existence and uniqueness of a minimizer for an ideal multiphase chemical equilibrium have been rigorously examined by Shapiro and Shapley in [60]. In their fundamental contribution, they provide a comprehensive characterization of the minimizers set for (2.37) in Theorem 9.2. Furthermore, they prove in Theorem 12.1 that if the configuration of the chemical system necessitates the presence of all phases at equilibrium, then the minimizer is unique. In cases where one or more phases are absent, uniqueness does not hold for the overall solution. However, Shapiro and Shapley demonstrate in Theorem 12.4 that the mole fractions of the present phases are uniquely defined. A last important result is the Theorem 9.8, which states that if a species is present in a phase, then the entire phase must also be present.

In this section, we extend the work of Shapiro and Shapley by presenting a novel condition that ensures the uniqueness of a minimizer, even in cases where certain phases may be absent. This condition is derived from their characterization of the set of minimizers and provides a more comprehensive analysis framework.

**Condition for the uniqueness of a minimizer**

Shapiro and Shapley have proved in [60, Theorem 9.2] that if the minimizers set for the multiphase chemical equilibrium problem (2.37) is nonempty, which is the case since $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ is compact as shown in Lemma 2.2.1, then there exists a minimizer $\tilde{\mathbf{n}}$ such that:

$$\arg\min\left\{G(\mathbf{m}) \;\middle|\; \mathbf{m} \in \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}\right\} = \mathcal{C}(\tilde{\mathbf{n}}) \cap \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}, \tag{2.39}$$

where

$$\mathcal{C}(\tilde{\mathbf{n}}) = \left\{\mathbf{m} \in \mathbb{R}^N \;\middle|\; \mathbf{m}^\alpha = \lambda_\alpha \tilde{\mathbf{n}}^\alpha,\ \lambda_\alpha \in \mathbb{R},\ \forall\, \alpha = 1, \ldots, N_{Ph}\right\}. \tag{2.40}$$

It follows from the definition of $\mathcal{C}(\tilde{\mathbf{n}})$ that if $\tilde{\mathbf{n}}^\alpha = \mathbf{0}$ then $\mathbf{m}^\alpha = \mathbf{0}$, for each $\mathbf{m}$ belonging to $\mathcal{C}(\tilde{\mathbf{n}}) \cap \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$. We introduce the set of present phases in $\tilde{\mathbf{n}}$ by

$$\Gamma_{\tilde{\mathbf{n}}} := \{\alpha \in \{1, \ldots, N_{Ph}\} \mid \tilde{\mathbf{n}}^\alpha \neq \mathbf{0}\}.$$

We can then establish the necessary and sufficient condition for the uniqueness of $\tilde{\mathbf{n}}$ in the subsequent theorem.

**Theorem 2.3.1.** *Let $\tilde{\mathbf{n}}$ the minimizer defined in (2.39) and (2.40) and let $\mathbf{A}^\alpha$ be the $M \times \#\sigma^{-1}(\alpha)$ matrix corresponding to the part of $\mathbf{A}$ related to the phase $\alpha$, then the minimizer $\tilde{\mathbf{n}}$ is unique if and only if the set of vectors $\{\mathbf{A}^\alpha \tilde{\mathbf{n}}^\alpha\}_{\alpha \in \Gamma_{\tilde{\mathbf{n}}}}$ is linearly independent.*

*Proof.* Let $\mathbf{m} \in \mathcal{C}(\tilde{\mathbf{n}}) \cap \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ with $\mathbf{m} \neq \tilde{\mathbf{n}}$. Since both $\tilde{\mathbf{n}}$ and $\mathbf{m}$ belongs to $\in \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$, we have

$$\mathbf{A}(\tilde{\mathbf{n}} - \mathbf{m}) = \mathbf{0} \quad \Leftrightarrow \quad \sum_{\alpha \in \Gamma_{\tilde{\mathbf{n}}}} \mathbf{A}^\alpha (\tilde{\mathbf{n}}^\alpha - \mathbf{m}^\alpha) = \mathbf{0}.$$

Since $\mathbf{m} \in \mathcal{C}(\tilde{\mathbf{n}}) \cap \overline{\Omega}$, there exists $\lambda_\alpha \geq 0$ such that $\mathbf{m}^\alpha = \lambda_\alpha \tilde{\mathbf{n}}^\alpha$, $\forall \alpha = 1, \ldots, N_{Ph}$. Therefore

$$\sum_{\alpha \in \Gamma_{\tilde{\mathbf{n}}}} (1 - \lambda_\alpha) \mathbf{A}^\alpha \tilde{\mathbf{n}}^\alpha = \mathbf{0},$$

with $\{1 - \lambda_\alpha\}_{\alpha \in \Gamma_{\tilde{\mathbf{n}}}}$ not all zero.

Conversely, assume that $\{\mathbf{A}^\alpha \tilde{\mathbf{n}}^\alpha\}_{\alpha \in \Gamma_{\tilde{\mathbf{n}}}}$ is not linearly independent, then there exists a set of non-zero scalars $\{\gamma_\alpha\}_{\alpha \in \Gamma_{\tilde{\mathbf{n}}}}$ such that

$$\sum_{\alpha \in \Gamma_{\tilde{\mathbf{n}}}} \gamma_\alpha \mathbf{A}^\alpha \tilde{\mathbf{n}}^\alpha = \mathbf{0}.$$

Since $\tilde{\mathbf{n}} \geq \mathbf{0}$ and from Lemma 2.1.2, there exists $\beta \in \Gamma_{\tilde{\mathbf{n}}}$ such that $\max\{\gamma_\alpha\}_{\alpha \in \Gamma_\alpha} =: \gamma_\beta > 0$.

One can defines a set of non-negative scalars $\{\lambda_\alpha\}_{\alpha\in\Gamma_{\tilde{\mathbf{n}}}}$ as $\lambda_\alpha = 1 - \gamma_\alpha/\gamma_\beta$. Therefore one defines the vector $\mathbf{m}$ such that $\mathbf{m}^\alpha = \lambda^\alpha \tilde{\mathbf{n}}^\alpha$ if $\alpha \in \Gamma_{\tilde{\mathbf{n}}}$ and $\mathbf{m}^\alpha = \mathbf{0}$ otherwise. By construction, $\mathbf{m} \in \mathcal{C}(\tilde{\mathbf{n}}) \cap \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ and $\mathbf{m} \neq \tilde{\mathbf{n}}$. □

Theorem 2.3.1 extends the work of Shapiro and Shapley by guaranteeing the uniqueness of the total quantities in each phase, rather than merely the uniqueness of mole fractions. Specifically, it addresses a well-known issue of non-uniqueness related to azeotropic compositions. In such mixtures, the liquid and gaseous phases maintain the same composition at a fixed temperature corresponding to a constant boiling point, making it impossible to separate the components through simple distillation. This characteristic renders the two phases indistinguishable.

Moreover, it serves as a clear example of the equality in mole fractions of two minimizers. The condition of Theorem 2.3.1 ensures that the total quantities of each phase in such cases are the same, thereby making the solution unique.

### 2.3.2 From minimization problem to algebraic equations

Compared to single-phase systems, multiphase equilibrium computations are complicated by the non-negativity constraint, which allows for the possibility of absent phases at equilibrium. Consequently, the number of present phases is not known a priori. Two main approaches are typically used to address this challenge:

1. a combinatorial method that computes chemical equilibrium for all possible subsets of potentially present phases;
2. a unified formulation that uses extended mole fractions together with complementarity equations to handle all potentially present phases simultaneously.

The recent approach developed by Leal *et al.* [41] uses a formulation that uses species quantity vectors, denoted as $\mathbf{n}$, as the primary unknowns. This method incorporates one complementarity equation for each species. However, as point out by Shapiro and Shapley [60], if a species is present in a given phase, then the entire phase must also be present. In contrast, the formulation proposed by Smith *et al.* [62] adopts mole fractions as unknowns, which allows for the establishment of one complementarity problem per phase.

In this thesis, we build upon this concept by formulating Gibbs energy in terms of mole fractions. Utilizing the notion of subdifferential, we derive a formulation that uses a broader definition of these fractions, which we refer to as extended mole fractions.

The concept of extended mole fractions was initially developed by Lauser *et al.* [39] and later applied to non-reactive phase equilibrium problems by Vu *et al.* [28], who termed it the "unified formulation." Coatléven and Michel [15] expanded upon this approach for multiphase chemical equilibrium problems. They introduced extended mole fractions in their formulation by integrating the optimality criterion of Smith *et al.* [62] into the Karush-Kuhn-Tucker (KKT) optimality conditions.

Our formulation enables the direct derivation of Coatléven and Michel's formulation through the KKT optimality conditions of the minimization problem by utilizing the Gibbs energy subdifferential. This approach offers a rigorous mathematical framework for establishing the equivalence between the minimization problem and the algebraic equations based on extended mole fractions. Although Coatléven and Michel's formulation is similar to ours, it differs in the choice of variables used to solve the equations.

This new formulation is presented in the following proposition.

**Proposition 2.3.1.** *If the vector* $\mathbf{n} \in \mathbb{R}^N$ *is a minimizer for the problem* (2.37)*, then there exists* $(\boldsymbol{\xi}^\alpha, s_\alpha, r_\alpha)_{\alpha=1,\dots,N_{Ph}}$ *such that* $n_i = s_{\sigma(i)}\xi_i$, $\xi_i > 0$*, for all* $i = 1, \dots, N$*, and satisfying:*

$$\sum_{\alpha=1}^{N_{Ph}} s_\alpha \mathbf{A}^\alpha \boldsymbol{\xi}^\alpha - b = 0, \tag{2.41a}$$

$$\mathbf{S}^T [\mu_i^\circ + \mathrm{RT} \ln \xi_i]_{i=1,\dots,N} = \mathbf{0}, \tag{2.41b}$$

$$\langle \boldsymbol{\xi}^\alpha, \mathbf{1} \rangle + r_\alpha - 1 = 0, \qquad (\alpha = 1, \dots, N_{Ph}), \tag{2.41c}$$

$$s_\alpha r_\alpha = 0, \qquad (\alpha = 1, \dots, N_{Ph}), \tag{2.41d}$$

$$s_\alpha \geq 0, r_\alpha \geq 0, \qquad (\alpha = 1, \dots, N_{Ph}). \tag{2.41e}$$

In the formulation (2.41), $\boldsymbol{\xi}^\alpha$ are the extended mole fractions, while $s_\alpha$ represents the total quantities of phase $\alpha$. Consequently, if a phase is present at equilibrium, we have $s_\alpha > 0$, and $\boldsymbol{\xi}^\alpha$ corresponds to the mole fractions:

$$\xi_i = \frac{n_i}{s_\alpha} = x_i(\mathbf{n}^\alpha).$$

Conversely, if a phase is absent, $r_\alpha \geq 0$, and the vector $\boldsymbol{\xi}^\alpha$ do not necessarily sum to one:

$$\langle \boldsymbol{\xi}^\alpha, \mathbf{1} \rangle \leq 1.$$

In this context, $\boldsymbol{\xi}^\alpha$ serves as an extension of the mole fractions for vanishing phases, facilitating the computation of equilibrium in an unified framework.

It is important to highlight that the strict positivity of $\boldsymbol{\xi}^\alpha$ implies that any solution to

this system satisfies the result of [60, Theorem 9.8]. This theorem states that if a phase is present at equilibrium, then all species within that phase must also be present.

The rest of this section is dedicated to the proof of Proposition 2.3.1. We introduce the sets of positive and non-negative vectors in $\mathbb{R}^{\#\sigma^{-1}(\alpha)}$ by

$$\Omega_\alpha := \left\{ \mathbf{n}^\alpha \in \mathbb{R}^{\#\sigma^{-1}(\alpha)} \mid \mathbf{n}^\alpha > \mathbf{0} \right\} \quad \text{and} \quad \overline{\Omega_\alpha} := \left\{ \mathbf{n}^\alpha \in \mathbb{R}^{\#\sigma^{-1}(\alpha)} \mid \mathbf{n}^\alpha \geq \mathbf{0} \right\}$$

respectively. Using these sets, we rewrite the minimization problem (2.37) as

$$\mathbf{n} \in \arg\min \left\{ \sum_{\alpha=1}^{N_{Ph}} \mathcal{G}_\alpha(\mathbf{n}^\alpha) \,\middle|\, \mathbf{n} \in \mathcal{M}_{\mathbf{A},\mathbf{b}} \right\}, \tag{2.42}$$

where

$$\mathcal{G}_\alpha(\mathbf{n}^\alpha) := \begin{cases} G_\alpha(\mathbf{n}^\alpha) & \text{if } \mathbf{n}^\alpha \in \overline{\Omega_\alpha}, \\ +\infty & \text{if } \mathbf{n}^\alpha \notin \overline{\Omega_\alpha}. \end{cases} \tag{2.43}$$

In this way, non-negativity constraint are incorporated into the objective function.

The Gibbs function $\mathcal{G}_\alpha$ in (2.43) is not differentiable on the boundary of $\overline{\Omega_\alpha}$, but it is possible to define a set known as the subdifferential in order to define the subgradient of $\mathcal{G}_\alpha$. The subdifferential of the convex function $\mathcal{G}_\alpha : \mathbb{R}^{\#\sigma^{-1}(\alpha)} \to \mathbb{R}$ at a point $\mathbf{n}^\alpha \in \overline{\Omega_\alpha} \setminus \Omega_\alpha$ is denoted $\partial\mathcal{G}_\alpha(\mathbf{n}^\alpha)$ and is defined by [59, Section 23]:

$$\boldsymbol{\mu}^\alpha \in \partial\mathcal{G}_\alpha(\mathbf{n}^\alpha) \quad \Leftrightarrow \quad \mathcal{G}_\alpha(\mathbf{m}^\alpha) \geq \mathcal{G}_\alpha(\mathbf{n}^\alpha) + \langle \boldsymbol{\mu}^\alpha, \mathbf{m}^\alpha - \mathbf{n}^\alpha \rangle, \, \forall \mathbf{m}^\alpha \in \mathbb{R}^{\#\sigma^{-1}(\alpha)}.$$

The function $\boldsymbol{\mu}^\alpha$ is called the subgradient of $\mathcal{G}_\alpha$. For a point $\mathbf{n}^\alpha \in \Omega_\alpha$, the subdifferential is reduced to the gradient of $\mathcal{G}_\alpha$.

Therefore, if $\mathbf{n} = (\mathbf{n}^\alpha)_{\alpha=1,\dots,N_{Ph}}$ is a solution to the problem defined in (2.42), it must satisfy the first order KKT optimality conditions:

$$\mathbf{A}\mathbf{n} - \mathbf{b} = \mathbf{0}, \tag{2.44a}$$

$$\boldsymbol{\mu} + \mathbf{A}^T \boldsymbol{\Lambda} = \mathbf{0}, \tag{2.44b}$$

$$\boldsymbol{\mu} = (\boldsymbol{\mu}^\alpha)_{\alpha=1,\dots,N_{Ph}} \tag{2.44c}$$

$$\boldsymbol{\mu}^\alpha \in \partial\mathcal{G}_\alpha(\mathbf{n}^\alpha). \tag{2.44d}$$

In (2.44), $\partial\mathcal{G}_\alpha$ denotes the subdifferential of $\mathcal{G}_\alpha$, which includes the non-negativity constraint, while $\boldsymbol{\Lambda} = (\lambda_1, \dots, \lambda_M)^T$ represents the vector of Lagrange multipliers associated with the linear constraint (2.44a).

To establish our formulation, we need to characterize $\partial\mathcal{G}_\alpha(\mathbf{n}^\alpha)$. The case where $\mathbf{n}^\alpha < \mathbf{0}$

is excluded from the minimization problem (2.42) by the definition of $\mathcal{G}_\alpha$. The following lemma provides a characterization of $\partial\mathcal{G}_\alpha(\mathbf{n}^\alpha)$ for $\mathbf{n}^\alpha \geq \mathbf{0}$.

**Lemma 2.3.1.** *Let $\mathbf{n}^\alpha \geq 0$, then $\boldsymbol{\mu}^\alpha \in \partial\mathcal{G}_\alpha(\mathbf{n}^\alpha)$ if and only if there exists $s_\alpha$, $\boldsymbol{\xi}^\alpha = (\xi_i)_{i\in\sigma^{-1}(\alpha)}$ such that:*

$$s_\alpha \geq 0,\ \boldsymbol{\xi}^\alpha > 0, \quad \langle \boldsymbol{\xi}^\alpha, \mathbf{1}\rangle \leq 1, \quad s_\alpha\,(1 - \langle \boldsymbol{\xi}^\alpha, \mathbf{1}\rangle) = 0,$$

*and for all $i \in \sigma^{-1}(\alpha)$,*

$$\mu_i = \mu_i^0 + \mathrm{RT}\ln\xi_i, \quad n_i = s_\alpha \xi_i.$$

In order to prove this lemma, we will describe the subdifferential of the Legendre transform $\mathcal{G}_\alpha^\star$ of $\mathcal{G}_\alpha$. Indeed, since $\mathcal{G}_\alpha$ is a lower semi-continuous convex function, the following equivalence holds:

$$\boldsymbol{\mu}_\alpha \in \partial\mathcal{G}_\alpha(\mathbf{n}^\alpha) \Leftrightarrow \mathbf{n}^\alpha \in \partial\mathcal{G}_\alpha^\star(\boldsymbol{\mu}_\alpha), \tag{2.45}$$

where $\mathcal{G}_\alpha^\star$ is the Legendre transform of $\mathcal{G}_\alpha$. This transformation is widely used in thermodynamics to convert a function defined in one set of variables into another, effectively representing the function in terms of its tangents. It is defined as

$$\mathcal{G}_\alpha^\star(\boldsymbol{\mu}^\alpha) = \sup_{\mathbf{n}^\alpha\in\mathbb{R}^{\#\sigma^{-1}(\alpha)}} \sum_{i\in\sigma^{-1}(\alpha)} n_i\mu_i - \mathcal{G}_\alpha(\mathbf{n}^\alpha), \quad \boldsymbol{\mu}^\alpha \in \mathbb{R}^{\#\sigma^{-1}(\alpha)}. \tag{2.46}$$

For $\mathbf{n}^\alpha \notin \overline{\Omega_\alpha}$, we have $\mathcal{G}_\alpha(\mathbf{n}^\alpha) = +\infty$, we can then restrict the supremum in (2.46) to $\mathbf{n}^\alpha \geq 0$ and consider the function $G_\alpha$ instead of $\mathcal{G}_\alpha$. We thus obtain

$$\mathcal{G}_\alpha^\star(\boldsymbol{\mu}^\alpha) = \sup_{\mathbf{n}^\alpha\geq 0} \sum_{i\in\sigma^{-1}(\alpha)} n_i\mu_i - G_\alpha(\mathbf{n}^\alpha) = \sup_{\mathbf{n}^\alpha\geq 0} \sum_{i\in\sigma^{-1}(\alpha)} n_i[\mu_i - \mu_i^\circ - \mathrm{RT}\ln x_i(\mathbf{n}^\alpha)].$$

Shifting from the supremum to the infimum, we get

$$\mathcal{G}_\alpha^\star(\boldsymbol{\mu}^\alpha) = -\inf_{\mathbf{n}^\alpha\geq 0} \sum_{i\in\sigma^{-1}(\alpha)} n_i[\mu_i^\circ + \mathrm{RT}\ln x_i(\mathbf{n}^\alpha) - \mu_i]. \tag{2.47}$$

We then introduce mole fractions by means of the following change of variables:

$$x_i s_\alpha = n_i \quad \text{with} \quad s_\alpha = \langle \mathbf{n}^\alpha, \mathbf{1}\rangle.$$

The Legendre transform (2.47) is then written as

$$\mathcal{G}_\alpha^\star(\boldsymbol{\mu}^\alpha) = - \inf_{\substack{s_\alpha,\, \mathbf{x} \geq 0 \\ \langle \mathbf{x}^\alpha,\, \mathbf{1} \rangle = 1}} s_\alpha \sum_{i \in \sigma^{-1}(\alpha)} x_i [\mu_i^\circ + \mathrm{RT} \ln x_i - \mu_i]. \tag{2.48}$$

From this relation, we can extract a minimization problem by defining the convex function $g_\alpha(\cdot\,; \boldsymbol{\mu}^\alpha) : \Omega_\alpha \to \mathbb{R}$ defined as

$$g_\alpha(\mathbf{x}^\alpha; \boldsymbol{\mu}^\alpha) = \sum_{i \in \sigma^{-1}(\alpha)} x_i [\mu_i^\circ + \mathrm{RT} \ln x_i - \mu_i], \tag{2.49}$$

where $\boldsymbol{\mu}^\alpha$ is considered as a parameter. Then, introducing the convex sets

$$X_\alpha = \{\mathbf{x}^\alpha > 0 \mid \langle \mathbf{x}^\alpha, \mathbf{1} \rangle = 1\}, \quad \text{and} \quad \overline{X_\alpha} = \{\mathbf{x}^\alpha \geq 0 \mid \langle \mathbf{x}^\alpha, \mathbf{1} \rangle = 1\},$$

the relation (2.48) becomes

$$\mathcal{G}_\alpha^\star(\boldsymbol{\mu}^\alpha) = - \inf_{s_\alpha \geq 0} s_\alpha \min_{\mathbf{x}^\alpha \in \overline{X_\alpha}} g_\alpha(\mathbf{x}^\alpha; \boldsymbol{\mu}^\alpha). \tag{2.50}$$

We can then study the minimization problem:

$$\min_{\mathbf{x}^\alpha \in \overline{X_\alpha}} g_\alpha(\mathbf{x}^\alpha; \boldsymbol{\mu}^\alpha). \tag{2.51}$$

The following lemma states that the inequality constraint on $\mathbf{x}^\alpha$ is never saturated at the solution of (2.51).

**Lemma 2.3.2.** *If $\mathbf{x}^\star \in \overline{X_\alpha}$ minimizes $g_\alpha$ on $\overline{X_\alpha}$, then $\mathbf{x}^\star \in X_\alpha$.*

*Proof.* The proof uses analogous reasoning to the arguments presented in the demonstration of Lemma 2.2.2. The function $g_\alpha$ is continuous on the compact set $\overline{X_\alpha}$, then from the Weierstrass theorem, there exists $\mathbf{x}^\star \in \{\arg\min g_\alpha(\mathbf{x}^\alpha; \boldsymbol{\mu}^\alpha) \mid \mathbf{x}^\alpha \in \overline{X_\alpha}\}$. Let us assume that $\mathbf{x}^\star \in \overline{X_\alpha} \setminus X_\alpha$, meaning that there exists a subset $\mathcal{J} \subsetneq \sigma^{-1}(\alpha)$ such that $x_j^\star = 0$, for all $j \in \mathcal{J}$. Note that the case $\mathcal{J} = \sigma^{-1}(\alpha)$ is excluded since $\langle \mathbf{x}^\alpha, \mathbf{1} \rangle = 1$. Let $\mathbf{x} \in X_\alpha$ and $\varepsilon \in (0,1)$, then one defines $\mathbf{x}^0 := \mathbf{x} - \mathbf{x}^\star$ and $\mathbf{x}^\varepsilon := \mathbf{x}^\star + \varepsilon \mathbf{x}^0 = \mathbf{x}^\varepsilon = \varepsilon \mathbf{x} + (1-\varepsilon)\mathbf{x}^\star$. The vector $\mathbf{x}^\varepsilon$ is a convex linear combination of vectors of $\overline{X_\alpha}$ which is a convex set, hence $\mathbf{x}^\varepsilon \in \overline{X_\alpha}$. Furthermore, $\mathbf{x}^\varepsilon \in X_\alpha$ since $\mathbf{x}^\varepsilon = \varepsilon \mathbf{x} + (1-\varepsilon)\mathbf{x}^\star \geq \varepsilon \mathbf{x} > 0$. By convexity of $g_\alpha$ on $\overline{\Omega_\alpha}$,

$$g_\alpha(\mathbf{x}^\star; \boldsymbol{\mu}^\alpha) \geq g_\alpha(\mathbf{x}^\varepsilon; \boldsymbol{\mu}^\alpha) + \langle \nabla g_\alpha(\mathbf{x}^\varepsilon), \mathbf{x}^\star - \mathbf{x}^\varepsilon \rangle \tag{2.52}$$

$$\Leftrightarrow \frac{g_\alpha(\mathbf{x}^\star;\boldsymbol{\mu}^\alpha)-g_\alpha(\mathbf{x}^\varepsilon;\boldsymbol{\mu}^\alpha)}{\varepsilon} \geq -\langle \nabla g_\alpha(\mathbf{x}^\varepsilon), \mathbf{x}^0\rangle, \tag{2.53}$$

where $\nabla g_\alpha(\mathbf{x}^\varepsilon) := (\mu_i^\circ + \mathrm{RT}\ln x_i^\varepsilon + \mathrm{RT})_{i\in\sigma^{-1}(\alpha)}$.

We will now take the limit when $\varepsilon$ tends to 0 in inequality (2.53). In the right-hand side one has

$$\lim_{\varepsilon\to 0} -\langle \nabla g_\alpha(\mathbf{x}^\varepsilon), \mathbf{x}^0\rangle = -\sum_{j\in\mathcal{J}} x_j^0 \lim_{\varepsilon\to 0} \partial_{x_j} g_\alpha(\mathbf{x}^\varepsilon) - \sum_{i\in\sigma^{-1}(\alpha)\setminus\mathcal{J}} x_i^0 \lim_{\varepsilon\to 0} \partial_{x_i} g_\alpha(\mathbf{x}^\varepsilon). \tag{2.54}$$

Noting that $\lim_{\epsilon\to 0}\mathbf{x}^\varepsilon = \mathbf{x}^\star$ and in particular that $\lim_{\varepsilon\to 0} x_j^\varepsilon = x_j^\star = 0$, for all $j\in\mathcal{J}$, it follows from the continuity of $\nabla g_\alpha$ that

$$\lim_{\varepsilon\to 0} \partial_{x_j} g_\alpha(\mathbf{x}^\varepsilon) = -\infty, \forall j\in\mathcal{J} \quad \text{and} \quad \lim_{\varepsilon\to 0} \partial_{x_i} g_\alpha(\mathbf{x}^\varepsilon) = \partial_{x_i} g_\alpha(\mathbf{x}^\star) \in \mathbb{R}, \forall i\in\sigma^{-1}(\alpha)\setminus\mathcal{J}. \tag{2.55}$$

By combining (2.54) and (2.55) with $x_j^0 = x_j > 0$, for all $j\in\mathcal{J}$, one finds that the right-hand side of (2.53) tends to $+\infty$. However if $\mathbf{x}^\star$ minimizes $g_\alpha$ on $\overline{X_\alpha}$, then the left-hand side of (2.53) is non-positive which is a contradiction. Therefore $\mathcal{J}=\emptyset$ and $\mathbf{x}^\star\in X_\alpha$. □

Based on Lemma 2.3.2, the problem (2.51) can be simplified to

$$\min_{\langle \mathbf{x}^\alpha,\mathbf{1}\rangle-1=0} \sum_{i\in\sigma^{-1}(\alpha)} x_i[\mu_i^\circ + \mathrm{RT}\ln x_i - \mu_i]. \tag{2.56}$$

The Euler-Lagrange equations associated to the problem (2.56) are

$$\mu_i^\circ + \mathrm{RT}\ln x_i - \mu_i + \mathrm{RT} - \lambda = 0, \quad i\in\sigma^{-1}(\alpha), \tag{2.57}$$

$$\langle \mathbf{x}^\alpha, \mathbf{1}\rangle - 1 = 0, \tag{2.58}$$

where $\lambda$ is the Lagrange multiplier for the constraint (2.58). Applying equation (2.57) and then equation (2.58) to the problem (2.56) yields the following minimum:

$$\min_{\langle \mathbf{x}^\alpha,\mathbf{1}\rangle-1=0} \sum_{i\in\sigma^{-1}(\alpha)} x_i[\mu_i^\circ + \mathrm{RT}\ln x_i - \mu_i] = \min_{\langle \mathbf{x}^\alpha,\mathbf{1}\rangle-1=0} \sum_{i\in\sigma^{-1}(\alpha)} x_i[\lambda - \mathrm{RT}] = \lambda - \mathrm{RT}.$$

In order to introduce this minimum into the Legendre transform (2.50), we will rewrite it in terms of $\boldsymbol{\mu}^\alpha$. To do so, we define the functions $\xi_i : \mathbb{R}\to\mathbb{R}_+^\star$, $i\in\sigma^{-1}(\alpha)$, by

$$\xi_i(\mu_i) := \exp\left(\frac{\mu_i-\mu_i^\circ}{\mathrm{RT}}\right). \tag{2.59}$$

Then using (2.57) and (2.58), we get that:

$$\xi_i(\mu_i) = x_i \exp\left(\frac{\mathrm{RT} - \lambda}{\mathrm{RT}}\right) \quad \text{and} \quad \langle \boldsymbol{\xi}^\alpha(\boldsymbol{\mu}^\alpha), \mathbf{1}\rangle = \sum_{i \in \sigma^{-1}(\alpha)} x_i \exp\left(\frac{\mathrm{RT} - \lambda}{\mathrm{RT}}\right) = \exp\left(\frac{\mathrm{RT} - \lambda}{\mathrm{RT}}\right),$$

where $\boldsymbol{\xi}^\alpha(\boldsymbol{\mu}^\alpha) = (\xi_i(\mu_i))_{i \in \sigma^{-1}(\alpha)}$. The solution of problem (2.51) is then given by:

$$\min_{\mathbf{x}^\alpha \in \overline{X_\alpha}} g_\alpha(\mathbf{x}^\alpha) = \lambda - \mathrm{RT} = -\mathrm{RT} \ln\langle \boldsymbol{\xi}^\alpha(\boldsymbol{\mu}^\alpha), \mathbf{1}\rangle. \tag{2.60}$$

Using the minimum (2.60) into (2.50), we obtain:

$$\mathcal{G}_\alpha^\star(\boldsymbol{\mu}^\alpha) = -\inf_{s_\alpha \geq 0} -s_\alpha \mathrm{RT} \ln\langle \boldsymbol{\xi}^\alpha(\boldsymbol{\mu}^\alpha), \mathbf{1}\rangle = \sup_{s_\alpha \geq 0} s_\alpha \mathrm{RT} \ln\langle \boldsymbol{\xi}^\alpha(\boldsymbol{\mu}^\alpha), \mathbf{1}\rangle.$$

Therefore, since $s_\alpha \mathrm{RT} \geq 0$, we can consider two cases based on the sign of $\ln\langle \boldsymbol{\xi}^\alpha(\boldsymbol{\mu}^\alpha), \mathbf{1}\rangle$:

- $\langle \boldsymbol{\xi}^\alpha(\boldsymbol{\mu}^\alpha), \mathbf{1}\rangle \leq 1 \quad \Rightarrow \quad \ln\langle \boldsymbol{\xi}^\alpha(\boldsymbol{\mu}^\alpha), \mathbf{1}\rangle \leq 0 \quad \Rightarrow \quad \mathcal{G}_\alpha^\star(\boldsymbol{\mu}^\alpha) = 0;$
- $\langle \boldsymbol{\xi}^\alpha(\boldsymbol{\mu}^\alpha), \mathbf{1}\rangle < 1 \quad \Rightarrow \quad \ln\langle \boldsymbol{\xi}^\alpha(\boldsymbol{\mu}^\alpha), \mathbf{1}\rangle < 0 \quad \Rightarrow \quad \mathcal{G}_\alpha^\star(\boldsymbol{\mu}^\alpha) = +\infty.$

We define the set $K_\alpha$ as follows

$$K_\alpha := \{\boldsymbol{\mu}^\alpha \mid \langle \boldsymbol{\xi}^\alpha(\boldsymbol{\mu}^\alpha), \mathbf{1}\rangle \leq 1\}.$$

This set is a convex set as it is the sublevel set of the convex function $\Psi : \boldsymbol{\mu}^\alpha \mapsto \langle \boldsymbol{\xi}^\alpha(\boldsymbol{\mu}^\alpha), \mathbf{1}\rangle$. Alongside $K_\alpha$, we introduce the associated characteristic function:

$$\chi_{K_\alpha}(\boldsymbol{\mu}^\alpha) = \begin{cases} 0 & \text{if } \boldsymbol{\mu}^\alpha \in K_\alpha, \\ +\infty & \text{if } \boldsymbol{\mu}^\alpha \notin K_\alpha. \end{cases}$$

With these definitions, we can express the Legendre transform as follows:

$$\mathcal{G}_\alpha^\star(\boldsymbol{\mu}^\alpha) = \chi_{K_\alpha}(\boldsymbol{\mu}^\alpha).$$

We can now easily calculate the subdifferential of $\mathcal{G}_\alpha^\star$ at a point $\boldsymbol{\mu}^\alpha \in K_\alpha$. If $\boldsymbol{\mu}^\alpha$ lies in the interior of $K_\alpha$, the subdifferential is reduced to the gradient of $\mathcal{G}_\alpha^\star$, which is equal to zero. Otherwise, if $\boldsymbol{\mu}^\alpha$ is on the boundary of $K_\alpha$, the definition of the subdifferential leads to the condition:

$$\mathbf{n}^\alpha \in \partial \mathcal{G}_\alpha^\star(\boldsymbol{\mu}^\alpha) \quad \Leftrightarrow \quad 0 \geq \langle \mathbf{n}^\alpha, \boldsymbol{\eta}^\alpha - \boldsymbol{\mu}^\alpha\rangle, \forall \boldsymbol{\eta}^\alpha \in \mathbb{R}^{\#\sigma^{-1}(\alpha)},$$

which is known as the normal cone of the convex set $K_\alpha$. The boundary of $K_\alpha$ is smooth because it is defined by $\Psi(\boldsymbol{\mu}^\alpha) = 1$. Consequently, this cone can be simplified to the normal direction, which is determined by the gradient of $\Psi$. Therefore, the subdifferential of $\mathcal{G}_\alpha^\star$ is:

$$\partial \mathcal{G}_\alpha^\star(\boldsymbol{\mu}^\alpha) = \begin{cases} 0 & \text{if } \langle \boldsymbol{\xi}^\alpha(\boldsymbol{\mu}^\alpha), \mathbf{1} \rangle < 1, \\ \gamma(\xi_i(\mu_i))_{i \in \sigma^{-1}(\alpha)}, \gamma \geq 0 & \text{if } \langle \boldsymbol{\xi}^\alpha(\boldsymbol{\mu}^\alpha), \mathbf{1} \rangle = 1. \end{cases}$$

We can now conclude the proof of Lemma 2.3.1.

Let $\mathbf{n}^\alpha \geq 0$, then

$$\begin{aligned} \boldsymbol{\mu}^\alpha \in \partial \mathcal{G}_\alpha(\mathbf{n}^\alpha) \Leftrightarrow \mathbf{n}^\alpha \in \partial \mathcal{G}_\alpha^\star(\boldsymbol{\mu}^\alpha) &\Leftrightarrow \begin{cases} n_i = \gamma \xi_i(\mu_i), \gamma \geq 0, i \in \sigma^{-1}(\alpha), \\ \langle \boldsymbol{\xi}^\alpha(\boldsymbol{\mu}^\alpha), \mathbf{1} \rangle \leq 1, \end{cases} \\ &\Leftrightarrow \begin{cases} n_i = \gamma \xi_i, \gamma \geq 0, i \in \sigma^{-1}(\alpha), \\ \mu_i = \mu_i^\circ + \mathrm{RT} \ln \xi_i, \xi_i > 0, \\ \langle \boldsymbol{\xi}^\alpha, \mathbf{1} \rangle \leq 1. \end{cases} \end{aligned}$$

We already know that if $\mathbf{n}^\alpha > 0$, the derivative of $g_\alpha$ is given by $\mu_i(\mathbf{n}^\alpha) = \mu_i^\circ + \mathrm{RT} \ln n_i / s_\alpha$, thus $\xi = n_i / s_\alpha$, $\gamma = s_\alpha > 0$ and $\langle \boldsymbol{\xi}^\alpha, \mathbf{1} \rangle = 1$. In the case where $\mathbf{n}^\alpha = 0$, it is clear that $\gamma = 0 = s_\alpha$ and that $\langle \boldsymbol{\xi}^\alpha, \mathbf{1} \rangle \leq 1$. Therefore, in each of these cases, one has

$$s_\alpha \geq 0, \boldsymbol{\xi}^\alpha > 0, \quad \langle \boldsymbol{\xi}^\alpha, \mathbf{1} \rangle \leq 1, \quad s_\alpha (1 - \langle \boldsymbol{\xi}^\alpha, \mathbf{1} \rangle) = 0, n_i = s_\alpha \xi_i, i \in \sigma^{-1}(\alpha).$$

Using this result, we can rewrite the KKT optimality conditions (2.44) as

$$\begin{aligned} \sum_{\alpha=1}^{N_{Ph}} s_\alpha \mathbf{A}^\alpha \boldsymbol{\xi}^\alpha - b = 0, & \\ \mathbf{S}^T [\mu_i^\circ + \mathrm{RT} \ln \xi_i]_{i=1,\dots,N} = \mathbf{0}, & \\ s_\alpha (1 - \langle \boldsymbol{\xi}^\alpha, \mathbf{1} \rangle) = 0, & \qquad (\alpha = 1, \dots, N_{Ph}), \\ s_\alpha \geq 0, 1 - \langle \boldsymbol{\xi}^\alpha, \mathbf{1} \rangle \geq 0, & \qquad (\alpha = 1, \dots, N_{Ph}). \end{aligned}$$

where $\boldsymbol{\mu}(\boldsymbol{\xi}) := (\mu_i^\circ + \mathrm{RT} \ln \xi_i)_{i=1,\dots,N}$. Therefore, based on the relation $\mathbf{S}^T \mathbf{A}^T = \mathbf{0}$ from Lemma 2.1.1, and defining

$$r_\alpha := 1 - \langle \boldsymbol{\xi}^\alpha, \mathbf{1} \rangle,$$

we finally obtain:

$$\sum_{\alpha=1}^{N_{Ph}} s_\alpha \mathbf{A}^\alpha \boldsymbol{\xi}^\alpha - b = 0, \tag{2.41a}$$

$$\mathbf{S}^T[\mu_i^\circ + \mathrm{RT}\ln \xi_i]_{i=1,\dots,N} = \mathbf{0}, \tag{2.41b}$$

$$\langle \boldsymbol{\xi}^\alpha, \mathbf{1}\rangle + r_\alpha - 1 = 0, \qquad (\alpha = 1,\dots,N_{Ph}), \tag{2.41c}$$

$$s_\alpha r_\alpha = 0, \qquad (\alpha = 1,\dots,N_{Ph}), \tag{2.41d}$$

$$s_\alpha \geq 0,\ r_\alpha \geq 0, \qquad (\alpha = 1,\dots,N_{Ph}). \tag{2.41e}$$

This concludes the proof of Proposition 2.3.1.

### 2.3.3 Study of the uniqueness of a solution to the algebraic equations

The existence of a solution is ensured by assumption (H1) and Lemma 2.2.1 which states that $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ is a nonempty compact set. We have seen in Theorem 2.3.1 that the minimizer $\mathbf{n}$ for (2.37), defined in (2.39) and (2.40), is unique if and only if the set of vectors $\{\mathbf{A}^\alpha \mathbf{n}^\alpha\}_{\alpha\in\Gamma_\mathbf{n}}$, where $\Gamma_\mathbf{n} = \{\alpha \in \{1,\dots,N_{Ph}\} \mid \mathbf{n}^\alpha \neq \mathbf{0}\}$, is linearly independent. We will demonstrate that a similar condition is essential for the uniqueness of the present phases in (2.41). Additionally, a criterion for constructing the vector $\mathbf{b}$ is required to ensure the uniqueness of $\boldsymbol{\xi}^\alpha$ for all absent phases.

To illustrate cases where uniqueness fails for absent phases, we will analyze an example the following system:

$$\begin{aligned}
\mathcal{C} &= \{\mathrm{H_2O(aq)}, \mathrm{H^+}, \mathrm{OH^-}, \mathrm{H_2O(g)}\},\\
\mathcal{E} &= \{\mathrm{H}, \mathrm{O}\},\\
\mathcal{R} &= \{\quad \mathrm{OH^-} = \mathrm{H_2O(aq)} - \mathrm{H^+},\\
&\qquad \mathrm{H_2O(g)} = \mathrm{H_2O(aq)}\},\\
\mathcal{P} &= \{\text{aqueous, gaseous}\},
\end{aligned} \tag{2.61}$$

with the right-hand side

$$\mathbf{b} = \lambda(2,1)^T,\ \lambda > 0,$$

and with a given pressure and temperature such that $\mu^\circ_{\mathrm{H_2O(g)}} < \mu^\circ_{\mathrm{H_2O(aq)}}$. The associated formula and stoichiometry matrices are

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 & 2 \\ 1 & 0 & 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{S}^T = \begin{pmatrix} 1 & -1 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{pmatrix}.$$

Let

$$\boldsymbol{\xi} = (\xi_{H_2O(aq)}, \xi_{H^+}, \xi_{OH^-}, \xi_{H_2O(g)}), \quad \mathbf{s} = (s_{aq}, s_g) \quad \text{and} \quad \mathbf{r} = (r_{aq}, r_g)$$

be such that

$$\xi_{H_2O(g)} = 1 \quad \text{and} \quad s_{aq} = 0.$$

Let us establish under which conditions the system (2.41) is fulfilled. Equation (2.41a) yields

$$s_{aq}\begin{pmatrix} 2 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}\begin{pmatrix} \xi_{H_2O(aq)} \\ \xi_{H^+} \\ \xi_{OH^-} \end{pmatrix} + s_g\begin{pmatrix} 2 \\ 1 \end{pmatrix}\xi_{H_2O(g)} = \lambda\begin{pmatrix} 2 \\ 1 \end{pmatrix},$$

and is satisfied if $s_g = \lambda$.

It follows from equations (2.41c) and (2.41d) that $(\xi_{H_2O(aq)}, \xi_{H^+}, \xi_{OH^-})$ must satisfy

$$\xi_{H_2O(aq)} + \xi_{H^+} + \xi_{OH^-} \leq 1 \tag{2.62}$$

and that $r_g = 0$, $r_{aq} \geq 0$. Finally, equations (2.41b) yields

$$\frac{\mu^\circ_{H_2O(aq)} - \mu^\circ_{H^+} - \mu^\circ_{OH^-}}{RT} + \ln \xi_{H_2O(aq)} - \ln \xi_{H^+} - \ln \xi_{OH^-} = 0, \tag{2.63a}$$

$$\frac{\mu^\circ_{H_2O(aq)} - \mu^\circ_{H_2O(g)}}{RT} + \ln \xi_{H_2O(aq)} - \underbrace{\ln \xi_{H_2O(g)}}_{=0} = 0. \tag{2.63b}$$

We can rewrite (2.63b) as

$$\xi_{H_2O(aq)} = \underbrace{e^{\frac{\mu^\circ_{H_2O(g)} - \mu^\circ_{H_2O(aq)}}{RT}}}_{:=\kappa_1}. \tag{2.64}$$

Therefore, using (2.64) into (2.63a), we get that

$$\xi_{OH^-} = \frac{1}{\xi_{H^+}} \underbrace{e^{\frac{\mu^\circ_{H_2O(g)} - \mu^\circ_{H^+} - \mu^\circ_{OH^-}}{RT}}}_{:=\kappa_2}. \tag{2.65}$$

It follows from (2.62), (2.64) and (2.65) that:

$$\xi_{H_2O(aq)} + \xi_{H^+} + \xi_{OH^-} \leq 1 \;\Leftrightarrow\; \xi^2_{H^+} + (\kappa_1 - 1)\xi_{H^+} + \kappa_2 \leq 0.$$

The discriminant of this quadratic inequality is given by:

$$\Delta = (\kappa_1 - 1)^2 - 4\kappa_2.$$

Using (2.62), (2.64) and (2.65) we get that:

$$1 - \kappa_1 \geq \xi_{H^+} + \xi_{OH^-} \Leftrightarrow \Delta \geq (\xi_{H^+} + \xi_{OH^-})^2 - 4\xi_{H^+}\xi_{OH^-} \Leftrightarrow \Delta \geq (\xi_{H^+} - \xi_{OH^-})^2 \geq 0.$$

It follows that:

$$\xi_{H^+} \in \left]\frac{1 - \kappa_1 - \sqrt{(1-\kappa_1)^2 - 4\kappa_2}}{2}, \frac{1 - \kappa_1 + \sqrt{(1-\kappa_1)^2 - 4\kappa_2}}{2}\right[.$$

Therefore, the chemical system (2.61) leads to a nonunique solution of equations (2.41). This result indicates that a condition on the vector $\mathbf{b}$ is necessary to ensure the uniqueness of the vector $\boldsymbol{\xi}$. Recalling that $\mathbf{a}_j$ is the formula vector for the species $C_j$ as defined in Section 2.1, let us make the following assumptions about the vector $\mathbf{b}$:

(A1) for all subset $\Gamma \subset \{1, \ldots, N_{Ph}\}$ such that $\{\mathbf{a}_j\}_{j\in\sigma^{-1}(\Gamma)}$ is a non-spanning set of vectors in $\mathbb{R}^M$, one has $\mathbf{b} \notin \operatorname{span}\left(\{\mathbf{a}_j\}_{j\in\sigma^{-1}(\Gamma)}\right)$.

We can then write the following uniqueness proposition.

**Proposition 2.3.2.** *Let $\boldsymbol{\mathcal{X}} = (\xi^\alpha, s_\alpha, r_\alpha)_{\alpha=1,\ldots,N_{Ph}}$ be a solution of system (2.41) where $\mathbf{b}$ satisfies assumption (A1), and let $\Gamma_{\boldsymbol{\mathcal{X}}}$ be the set of present phases. Then $\boldsymbol{\mathcal{X}}$ is unique if and only if the set of vectors $\{\mathbf{A}^\alpha \xi^\alpha\}_{\alpha\in\Gamma_{\boldsymbol{\mathcal{X}}}}$ is linearly independent.*

To prove this proposition, we first need to establish the following two intermediate results.

**Lemma 2.3.3.** *Let $\boldsymbol{\mathcal{X}} = (\xi^\alpha, s_\alpha, r_\alpha)_{\alpha=1,\ldots,N_{Ph}}$ a solution of (2.41). If $\mathbf{b}$ satisfies assumption (A1), then there are at least $M$ species present in $\boldsymbol{\mathcal{X}}$. Furthermore, system (2.41) can be rewritten with these $M$ species as the primary species.*

*Proof.* Let $\Gamma_{\boldsymbol{\mathcal{X}}}$ be the set of present phases of $\boldsymbol{\mathcal{X}}$, in this way $\sigma^{-1}(\Gamma_{\boldsymbol{\mathcal{X}}})$ is the set of all present species. From the assumption (A1), $\{\mathbf{a}_j\}_{j\in\sigma^{-1}(\Gamma_{\boldsymbol{\mathcal{X}}})}$ is a spanning set in $\mathbb{R}^M$. It is then possible to extract a basis of $\mathbb{R}^M$ from the set $\{\mathbf{a}_j\}_{j\in\sigma^{-1}(\Gamma_{\boldsymbol{\mathcal{X}}})}$ and to construct a permutation matrix $\mathbf{P} \in \mathbb{R}^{N\times N}$ such that this basis form a primary block in $\tilde{\mathbf{A}} := \mathbf{AP}$. In this way, $\tilde{\mathbf{A}} = [\tilde{\mathbf{A}}_{Pr}, \tilde{\mathbf{A}}_{Sd}]$ where $\tilde{\mathbf{A}}_{Pr}$ is an invertible matrix. Let $\boldsymbol{\mu} := [\mu_i^\circ + \mathrm{RT}\ln\xi_i]_{i=1,\ldots,N}$, since $\mathbf{S}^T\boldsymbol{\mu} = \mathbf{0}$ and $\ker\mathbf{S}^T = \operatorname{Im}\mathbf{A}^T$ from Lemma 2.1.1, there exists $\mathbf{y} \in \mathbb{R}^M$ such that

$$\boldsymbol{\mu} = \mathbf{A}^T\mathbf{y} \Leftrightarrow \mathbf{P}^T\boldsymbol{\mu} = \tilde{\mathbf{A}}^T\mathbf{y} \Leftrightarrow \tilde{\mathbf{S}}^T\mathbf{P}^T\boldsymbol{\mu} = \mathbf{0},$$

where

$$\tilde{\mathbf{S}} := \begin{bmatrix} \tilde{\mathbf{A}}_{Pr}^{-1}\tilde{\mathbf{A}}_{Sd} \\ -\mathbf{I}_{Sd} \end{bmatrix}. \tag{2.66}$$

In this way, system (2.41) can be rewritten using the matrices $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{S}}^T$. □

**Lemma 2.3.4.** *Under assumption (A1), the vector $\boldsymbol{\xi} = (\xi^\alpha)_{\alpha=1,\dots,N_{Ph}}$, satisfying the system (2.41), is unique.*

To establish this lemma, we require the following result from information theory.

**Lemma 2.3.5** (Gibb's inequality)**.** *Suppose that $\boldsymbol{\xi} = (\xi_i)_{i=1,\dots,n}$ and $\overline{\boldsymbol{\xi}} = (\overline{\xi})_{i=1,\dots,n}$ are two discrete probability distribution then*

$$\sum_{i=1}^{n} \xi_i(\ln \xi_i - \ln \overline{\xi}_i) \geq 0,$$

*with equality if and only if $\xi_i = \overline{\xi}_i$, $i = 1, \dots, n$.*

*Proof of Lemma 2.3.4.* Let $\boldsymbol{\mathcal{X}} = (\boldsymbol{\xi}^\alpha, s_\alpha, r_\alpha)_{\alpha=1,\dots,N_{Ph}}$ and $\overline{\boldsymbol{\mathcal{X}}} = (\overline{\boldsymbol{\xi}}^\alpha, \overline{s_\alpha}, \overline{r_\alpha})_{\alpha=1,\dots,N_{Ph}}$ be two solutions of (2.41), then

$$\mathbf{A}[s_\alpha \boldsymbol{\xi}^\alpha - \overline{s_\alpha}\overline{\boldsymbol{\xi}}^\alpha]_{\alpha=1,\dots,N_{Ph}} = 0, \tag{2.67}$$

$$\mathbf{S}^T[\ln \boldsymbol{\xi}^\alpha - \ln \overline{\boldsymbol{\xi}}^\alpha]_{\alpha=1,\dots,N_{Ph}} = 0. \tag{2.68}$$

Due to Lemma 2.1.1, we have the relation $\ker \mathbf{S}^T = (\ker \mathbf{A})^\perp$, therefore

$$\sum_{\alpha=1}^{N_{Ph}} \langle s_\alpha \boldsymbol{\xi}^\alpha - \overline{s_\alpha}\overline{\boldsymbol{\xi}}^\alpha, \ln \boldsymbol{\xi}^\alpha - \ln \overline{\boldsymbol{\xi}}^\alpha \rangle = 0. \tag{2.69}$$

We will prove that the sum (2.69) is composed of non-negative terms. For each phase $\alpha$, there are three cases:

— if $s_\alpha = \overline{s_\alpha} = 0$, then $\langle s_\alpha \boldsymbol{\xi}^\alpha - \overline{s_\alpha}\overline{\boldsymbol{\xi}}^\alpha, \ln \boldsymbol{\xi}^\alpha - \ln \overline{\boldsymbol{\xi}}^\alpha \rangle = 0$;

— if $s_\alpha = \overline{s_\alpha} > 0$, then $\langle s_\alpha \boldsymbol{\xi}^\alpha - \overline{s_\alpha}\overline{\boldsymbol{\xi}}^\alpha, \ln \boldsymbol{\xi}^\alpha - \ln \overline{\boldsymbol{\xi}}^\alpha \rangle \geq 0$ since the logarithm is and increasing function;

— if $s_\alpha \neq \overline{s_\alpha}$, we can assume, without loss of generality, that $s_\alpha > \overline{s_\alpha}$. This implies that $s_\alpha > 0$ and then $r_\alpha = 0$. We can rewrite the term in the sum (2.69) as

$$\overline{s_\alpha}\langle \boldsymbol{\xi}^\alpha - \overline{\boldsymbol{\xi}}^\alpha, \ln \boldsymbol{\xi}^\alpha - \ln \overline{\boldsymbol{\xi}}^\alpha \rangle + (s_\alpha - \overline{s_\alpha})\langle \boldsymbol{\xi}^\alpha, \ln \boldsymbol{\xi}^\alpha - \ln \overline{\boldsymbol{\xi}}^\alpha \rangle =: \underbrace{\overline{s_\alpha}\mathcal{A}_\alpha}_{\geq 0} + \underbrace{(s_\alpha - \overline{s_\alpha})}_{>0}\mathcal{B}_\alpha. \tag{2.70}$$

Applying the convexity inequality:

$$x(\ln x - \ln y) \geq x - y,$$

to $\mathcal{B}_\alpha$ yields

$$\mathcal{B}_\alpha = \langle \xi^\alpha, \ln \xi^\alpha - \ln \overline{\xi}^\alpha \rangle \geq \sum_{i \in \sigma^{-1}(\alpha)} \xi_i - \overline{\xi}_i = \overline{r_\alpha} - r_\alpha = \overline{r_\alpha} \geq 0 \tag{2.71}$$

since $r_\alpha = 0$. Consequently, (2.70) is non-negative.

Therefore, (2.69) is a vanishing sum of non-negative terms, it follows that for each $\alpha \in \sigma^{-1}(\alpha)$:

$$\langle s_\alpha \xi^\alpha - \overline{s_\alpha}\overline{\xi}^\alpha, \ln \xi^\alpha - \ln \overline{\xi}^\alpha \rangle = \overline{s_\alpha}\mathcal{A}_\alpha + (s_\alpha - \overline{s_\alpha})\mathcal{B}_\alpha = 0. \tag{2.72}$$

In order to prove the uniqueness of $\xi^\alpha$, there are the same three cases to deal with:

— if $s_\alpha = \overline{s_\alpha} > 0$, then (2.72) becomes $\mathcal{A}_\alpha = 0$. We introduce the function

$$\mathbf{g}_\alpha(\xi^\alpha) := \sum_{i \in \sigma^{-1}(\alpha)} \xi_i \ln \xi_i,$$

which is strictly convex on $\Omega_\alpha$. We can then rewrite $\mathcal{A}_\alpha$ as

$$\mathcal{A}_\alpha = \langle \xi^\alpha - \overline{\xi}^\alpha, \nabla \mathbf{g}_\alpha(\xi^\alpha) - \nabla \mathbf{g}_\alpha(\overline{\xi}^\alpha) \rangle.$$

The strict convexity of the function $\mathbf{g}_\alpha$ on the set $\Omega_\alpha$ implies the strict monotonicity of its gradient $\nabla \mathbf{g}_\alpha$, consequently

$$\mathcal{A}_\alpha = 0 \quad \Rightarrow \quad \xi^\alpha = \overline{\xi}^\alpha;$$

— if $s_\alpha > \overline{s_\alpha}$, then $s_\alpha > 0$ we have seen that $r_\alpha = 0$ and

$$\underbrace{\overline{s_\alpha}\mathcal{A}_\alpha}_{\geq 0} + \underbrace{(s_\alpha - \overline{s_\alpha})\mathcal{B}_\alpha}_{\geq 0} = 0. \tag{2.73}$$

As a result, both terms in the sum (2.73) are zero, and therefore, $\mathcal{B}_\alpha = 0$. From (2.71), we deduce that $\overline{r_\alpha} = 0$. It follows that $\langle \xi^\alpha, \mathbf{1} \rangle = 1$ and $\langle \overline{\xi}^\alpha, \mathbf{1} \rangle = 1$. This result allows us to apply Gibb's inequality (Lemma 2.3.5), which states:

$$\mathcal{B}_\alpha = \langle \xi^\alpha, \ln \xi^\alpha - \ln \overline{\xi}^\alpha \rangle = 0 \quad \Leftrightarrow \quad \xi^\alpha = \overline{\xi}^\alpha;$$

— if $s_\alpha = \overline{s_\alpha} = 0$, we cannot conclude about the uniqueness of $\xi^\alpha$ from (2.72) but Lemma 2.3.3 guarantees that there are at least $M$ primary species present in $\mathcal{X}$ and that the formula matrix $\mathbf{A}$ can be reformulated with these $M$ species as the first $M$ columns. Let $\xi_{Pr}$ and $\overline{\xi}_{Pr}$ the corresponding vectors in $\mathcal{X}$ and $\overline{\mathcal{X}}$ respectively. We have established in Proposition 2.3.1 that $\xi^\alpha > 0$ for each phase $\alpha$. This implies that if a species is present in a phase $\alpha$, then $s_\alpha > 0$, indicating that the entire phase is present. It follows that the phases where $\xi_{Pr}$ belongs are all present, and from the previous cases we can conclude that $\xi_{Pr} = \overline{\xi}_{Pr}$. Let $\mathbf{d} := \mathbf{S}^T \boldsymbol{\mu}^\circ/(\mathrm{RT})$, it follows from (2.41) that for each $i \in \sigma^{-1}(\alpha)$,

$$\left(\mathbf{d} + \tilde{\mathbf{A}}_{Sd}^T \tilde{\mathbf{A}}_{Pr}^{-T} [\ln \xi_{Pr}]\right)_i = \ln \xi_i^\alpha,$$
$$\left(\mathbf{d} + \tilde{\mathbf{A}}_{Sd}^T \tilde{\mathbf{A}}_{Pr}^{-T} [\ln \overline{\xi}_{Pr}]\right)_i = \ln \overline{\xi}_i^\alpha,$$

and since $\xi_{Pr} = \overline{\xi}_{Pr}$,

$$\left(\mathbf{d} + \tilde{\mathbf{A}}_{Sd}^T \tilde{\mathbf{A}}_{Pr}^{-T} [\ln \xi_{Pr}]\right)_i = \ln \overline{\xi}_i^\alpha,$$

which implies that $\xi^\alpha = \overline{\xi}^\alpha$.

□

With these two lemmas, we can now demonstrate Proposition 2.3.2.

*Proof of Proposition* 2.3.2. In Lemma 2.3.4, we have proved that under assumption (A1), the vector $\xi$ is unique. The uniqueness of $(r_\alpha)_{\alpha=1,\dots,N_{Ph}}$ follows from (2.41c). To demonstrate the uniqueness of $(s_\alpha)_{\alpha=1,\dots,N_{Ph}}$, let $(\bar{s}_\alpha)_{\alpha=1,\dots,N_{Ph}}$ be another solution. According to equation (2.41a), we can derive the following relation:

$$\sum_{\alpha \in \Gamma_{\mathcal{X}}} (s_\alpha - \bar{s}_\alpha) \mathbf{A}^\alpha \xi^\alpha = \mathbf{0}.$$

Therefore the uniqueness is satisfied since $\{\mathbf{A}^\alpha \xi^\alpha\}_{\alpha \in \Gamma_{\mathcal{X}}}$ is linearly independent. □

Chapter 3

# New algorithms for single-phase chemical equilibrium

**Outline of the current chapter**



This chapter focuses on the numerical resolution of the single-phase chemical equilibrium problem and is divided into three sections. As a reminder of Section 2.2.3, the

system to solve is the following: find $(\mathbf{x}, \omega)$ such that

$$\begin{aligned} \mathbf{A}\mathbf{x} - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T\mathbf{y}(\mathbf{x}) &= \mathbf{d}, \\ \langle \mathbf{x}, \mathbf{1} \rangle &= 1, \end{aligned} \tag{2.33}$$

where

$$(\mathbf{y}(\mathbf{x}))_i = y(x_i) := \ln x_i, \ i = 1, \ldots, N \quad \text{and} \quad \mathbf{d} := -\mathbf{S}^T\boldsymbol{\mu}^\circ/(\mathrm{RT}).$$

In the first part of this chapter, we derive the methods of parametrization and Cartesian representation. These methods aim to address the challenges encountered when solving the system described in (2.33) using Newton's method. The main challenge is to manage the relationship $y = \ln x$. This relationship can be interpreted as the graph that connects species mole fractions to their chemical potentials.

When resolving the system with respect to mole fractions, several issues arise in Newton's algorithm. Notably, the positivity of the iterates is not guaranteed, and very low concentrations of a species can lead to a blow-up in the Jacobian matrix. A common remedy for this problem, referred to as the "log-trick," involves solving the system with respect to chemical potentials. However, this approach introduces exponential terms into the Jacobian, which can also blow up during the iterations of Newton's method. The parametrization approach introduces a fictitious variable along with a parametrization of the graph, allowing us to leverage the advantages of both resolution techniques. The second approach involves a Cartesian representation of this relationship, based on an augmented system where both mole fractions and chemical potentials are treated as unknowns. In this case, the relationship is relaxed into a new function and is recovered only at convergence.

In the second part of this chapter, we provide a theoretical proof of the local quadratic convergence of Newton's method for both approaches.

In the final section, we present numerical experiments that demonstrate the accuracy of the two techniques, enabling the computation of equilibria for chemical species with very low concentrations. Additionally, we compare our results with those obtained using an equivalent of the Arxim solver in our framework, considering various initializations.

The results presented in this chapter are essentially part of the preprint [33].

## 3.1 Towards more robust numerical algorithms

After a brief review of Newton's method, this section presents the parametrization and Cartesian representation techniques and their advantages for solving the chemical equilibrium problem.

### 3.1.1 Newton's method

There are many methods to solve the nonlinear system of equations (2.33) as detailed in [51], however our study will focus on Newton's method which is known for its fast convergence as well as for its lack of stability in many contexts. The resolution of system (2.33) can be viewed as the search for the zeros of a function $\mathcal{U} : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$, associated to a function $\mathcal{F} : \mathbb{R}^{2N+1} \to \mathbb{R}^{N+1}$, which are defined as follows:

$$\mathcal{U}(\mathbf{x}, \omega) := \mathcal{F}(\mathbf{x}, \mathbf{y}(\mathbf{x}), \omega) = \begin{pmatrix} \mathbf{A}\mathbf{x} - \omega\mathbf{b} \\ \mathbf{S}^T\mathbf{y}(\mathbf{x}) - \mathbf{d} \\ \langle \mathbf{x}, \mathbf{1} \rangle - 1 \end{pmatrix}. \tag{3.1}$$

The function $\mathcal{U}$ is called residual.

Let $\boldsymbol{\mathcal{X}} := (\mathbf{x}, \omega)$, we recall that the Newton method is an iterative algorithm that, from an initial value $\boldsymbol{\mathcal{X}}^{(0)}$, builds a sequence $(\boldsymbol{\mathcal{X}}^{(k)})_{k>0}$ defined by solving the linear system

$$\nabla\mathcal{U}\left(\boldsymbol{\mathcal{X}}^{(k)}\right)\delta\boldsymbol{\mathcal{X}}^{(k)} = -\mathcal{U}\left(\boldsymbol{\mathcal{X}}^{(k)}\right), \tag{3.2}$$

where $\nabla\mathcal{U}(\boldsymbol{\mathcal{X}}^{(k)})$ stands for the Jacobian matrix of $\mathcal{U}$ evaluated at $\boldsymbol{\mathcal{X}}^{(k)}$. The solution of (3.2) is called the Newton increment and is used to update the sequence as

$$\boldsymbol{\mathcal{X}}^{(k+1)} = \boldsymbol{\mathcal{X}}^{(k)} + \delta\boldsymbol{\mathcal{X}}^{(k)}. \tag{3.3}$$

The algorithm is considered to have converged if the following condition is satisfied:

$$\|\mathcal{U}(\boldsymbol{\mathcal{X}}^{(k)})\| \leq \varepsilon,$$

where $\varepsilon$ is a predefined, sufficiently small tolerance parameter, and $k$ is the current iteration number not exceeding a predetermined maximum number of iterations.

An important result about Newton's method concerns its local quadratic convergence [35]. It requires the following assumptions:

1. The function (3.1) has a unique zero $\boldsymbol{\mathcal{X}}^{\star}$.

2. The Jacobian $\nabla\mathcal{U} : \mathbb{R}^N \to \mathbb{R}^{N\times N}$ is Lipschitz continuous near $\boldsymbol{\mathcal{X}}^\star$: there exists a neighborhood $\mathcal{V}$ of $\boldsymbol{\mathcal{X}}^\star$ and $L > 0$ such that

$$\|\nabla\mathcal{U}(\boldsymbol{\mathcal{X}}_1) - \nabla\mathcal{U}(\boldsymbol{\mathcal{X}}_2)\|_2 \leq L\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\|_2$$

for all $\boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_2$ in $\mathcal{V}$.

3. The matrix $\nabla\mathcal{U}(\boldsymbol{\mathcal{X}}^\star)$ is nonsingular, *i.e.* invertible.

The local quadratic convergence theorem is as follows [35, Theorem 1.1].

**Theorem 3.1.1.** *Let the previous assumptions hold. If $\boldsymbol{\mathcal{X}}^{(0)}$ is sufficiently close to $\boldsymbol{\mathcal{X}}^\star$, then Newton's sequence (3.2)–(3.3) is well defined for all $k \geq 0$ and converges to $\boldsymbol{\mathcal{X}}^\star$. Moreover, there exist $C > 0$ and $k_C \in \mathbb{N}$ such that*

$$\|\boldsymbol{\mathcal{X}}^{(k+1)} - \boldsymbol{\mathcal{X}}^\star\|_2 \leq C\|\boldsymbol{\mathcal{X}}^{(k)} - \boldsymbol{\mathcal{X}}^\star\|_2^2, \ \forall k \geq k_C. \tag{3.4}$$

The property (3.4) together with $\boldsymbol{\mathcal{X}}^{(k)} \to \boldsymbol{\mathcal{X}}^\star$ is referred to as q-quadratic convergence in the monograph [35].

### 3.1.2 A family of parametrizations

Newton's method applied to the function (3.1) yields the following Jacobian matrix:

$$\nabla\mathcal{U}(\mathbf{x}, \omega) = \begin{bmatrix} \mathbf{A} & -\mathbf{b} \\ \mathbf{S}^T\nabla\mathbf{y}(\mathbf{x}) & \mathbf{0} \\ \mathbf{1}^T & 0 \end{bmatrix}, \tag{3.5}$$

where $\nabla\mathbf{y}(\mathbf{x}) = \operatorname{diag}\{1/x_i\}_{i=1,\dots,N}$. This Jacobian diverges when $x_i$ tends to zero, possibly leading to trouble in Newton's algorithm. Beyond the blow up of the Jacobian when one species vanishes, the iterates can become negative and yield the algorithm failure due to the domain of $y$. Furthermore, the concentrations of chemical species can vary significantly, with some species, typically the solvent, being present in large quantities, while others are highly diluted. This wide range in concentration levels can reduce the precision of computations. A classic cure to these problems, known as the log-trick [79], is to consider $y_i = y(x_i)$ as the unknowns and to define $\mathcal{V} : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ as follows

$$\mathcal{V}(\mathbf{y}, \omega) := \mathcal{F}(\mathbf{x}(\mathbf{y}), \mathbf{y}, \omega) = \mathbf{0} \quad \Leftrightarrow \quad \begin{aligned} \mathbf{A}\mathbf{x}(\mathbf{y}) - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T\mathbf{y} - \mathbf{d} &= \mathbf{0}, \\ \langle\mathbf{x}(\mathbf{y}), \mathbf{1}\rangle - 1 &= 0, \end{aligned} \tag{3.6}$$

with $\mathbf{x}(\mathbf{y}) = (x(y_i))_{i=1,\dots,N}$ where $x(y_i) := y^{-1}(y_i) = \exp y_i$, $\mathcal{F}$ being defined as in (3.1). In this case the Jacobian matrix becomes

$$\nabla\mathcal{V}(\mathbf{y}, \omega) = \begin{bmatrix} \mathbf{A}\nabla\mathbf{x}(\mathbf{y}) & -\mathbf{b} \\ \mathbf{S}^T & \mathbf{0} \\ \mathbf{1}^T\nabla\mathbf{x}(\mathbf{y}) & 0 \end{bmatrix}, \tag{3.7}$$

where $\nabla\mathbf{x}(\mathbf{y}) = \operatorname{diag}\{\exp y_i\}_{i=1,\dots,N}$. The logarithmic transformation ensures that all chemical species are represented on a comparable scale, while simultaneously guaranteeing positive concentration values throughout the iterations. However, the Jacobian (3.7) diverges as $y_i$ approaches infinity, and numerical issues can arise even for moderately positive values of $y_i$.

To address these limitations, we can use a more sophisticated method known as parametrization. This approach aims to combine the strengths of both the $\mathbf{x}$ and $\mathbf{y}$ formulations while ensuring better control over the coefficients in the system's Jacobian. For this purpose, the graph

$$\mathcal{T} = \left\{(x, y) \in \mathbb{R}^2 \;\middle|\; y = \ln(x)\right\} \tag{3.8}$$

is parameterized by two monotonic, Lipschitz continuous functions, $X : \mathbb{R} \to \mathbb{R}$ and $Y : \mathbb{R} \to \mathbb{R}$, such that $x_i = X(\tau_i)$ and $y_i = Y(\tau_i)$. These functions are defined to satisfy the relationship

$$Y(\tau) = \ln(X(\tau)),$$

which implies that $\mathcal{T} = (X, Y)(\mathbb{R})$. Therefore, the unknowns of the system become $\boldsymbol{\tau} = (\tau_i)_{i=1,\dots,N}$ and the parametrization problem is: find $(\boldsymbol{\tau}, \omega) \in \mathbb{R}^{N+1}$ such that

$$\mathcal{P}(\boldsymbol{\tau}, \omega) := \mathcal{F}(\mathbf{X}(\boldsymbol{\tau}), \mathbf{Y}(\boldsymbol{\tau}), \omega) = \mathbf{0} \quad \Leftrightarrow \quad \begin{aligned} \mathbf{A}\mathbf{X}(\boldsymbol{\tau}) - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T\mathbf{Y}(\boldsymbol{\tau}) - \mathbf{d} &= \mathbf{0}, \\ \langle \mathbf{X}(\boldsymbol{\tau}), \mathbf{1}\rangle - 1 &= 0, \end{aligned} \tag{3.9}$$

where $\mathbf{X}(\boldsymbol{\tau}) = (X(\tau_i))_{i=1,\dots,N}$ and $\mathbf{Y}(\boldsymbol{\tau}) = (Y(\tau_i))_{i=1,\dots,N}$. The associated Jacobian matrix is written as follows:

$$\nabla\mathcal{P}(\boldsymbol{\tau}, \omega) = \begin{bmatrix} \mathbf{A}\operatorname{diag}\{\mathbf{X}'(\boldsymbol{\tau})\} & -\mathbf{b} \\ \mathbf{S}^T\operatorname{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\} & \mathbf{0} \\ \mathbf{1}^T\operatorname{diag}\{\mathbf{X}'(\boldsymbol{\tau})\} & 0 \end{bmatrix},$$

where $\mathbf{X}'(\boldsymbol{\tau}) = (X'(\tau_i))_{i=1,\dots,N}$ and $\mathbf{Y}'(\boldsymbol{\tau}) = (Y'(\tau_i))_{i=1,\dots,N}$. It is worth noting that the

parametrization technique used in this context serves as a non-linear right preconditioning method. The relationship between the parametrized problem and the $\mathbf{x}$ and $\mathbf{y}$ formulations can be expressed as follows:

$$\mathcal{P}(\boldsymbol{\tau}, \omega) = \mathbf{0} \quad \Leftrightarrow \quad \begin{matrix} \mathcal{U}(\mathbf{X}(\boldsymbol{\tau}), \omega) = \mathbf{0}, \\ \text{or} \\ \mathcal{V}(\mathbf{Y}(\boldsymbol{\tau}), \omega) = \mathbf{0}, \end{matrix} \quad \Leftrightarrow \quad \begin{matrix} \mathcal{U}(\mathbf{X} \circ \mathbf{X}^{-1}(\mathbf{x}), \omega) = \mathbf{0}, \\ \text{or} \\ \mathcal{V}(\mathbf{Y} \circ \mathbf{Y}^{-1}(\mathbf{y}), \omega) = \mathbf{0}. \end{matrix}$$

We will now introduce the conditions that enable us to control the coefficients of the Jacobian matrix. In the problem under consideration, the Jacobian remains bounded if and only if the derivatives $\mathbf{X}'(\boldsymbol{\tau})$ and $\mathbf{Y}'(\boldsymbol{\tau})$ are bounded. Furthermore, if any components of $X'(\tau)$ and $Y'(\tau)$ simultaneously vanish for a given value of $\tau$, the corresponding column in the Jacobian will be zero, rendering the matrix singular. To ensure proper parametrization and avoid singularities, we must satisfy the following conditions, for each $\tau \in \mathbb{R}$:

(P1) $Y(\tau) = \ln(X(\tau))$ where $X$ and $Y$ are monotonic and Lipschitz continuous functions;

(P2) $X'$ and $Y'$ are bounded and Lipschitz continuous functions;

(P3) there exists $\beta > 0$ such that $|X'(\tau)| + |Y'(\tau)| \geq \beta$ for all $\tau \in \mathbb{R}$.

We define a parametrization as admissible if it fulfills conditions (P1)–(P3). To guarantee the satisfaction of conditions (P2) and (P3), we introduce the following normalization condition on the derivatives:

$$(|X'(\tau)|^p + |Y'(\tau)|^p)^{1/p} = 1, \quad p \geq 1. \tag{3.10}$$

This condition will allow us to determine the functions $X$ and $Y$ using the derivative

$$Y'(\tau) = \frac{X'(\tau)}{X(\tau)} \tag{3.11}$$

from condition (P1). By combining (3.10) and (3.11), we obtain:

$$|X'(\tau)|^p + \left|\frac{X'(\tau)}{X(\tau)}\right|^p = 1 \quad \Leftrightarrow \quad |X'(\tau)| = \frac{1}{(1 + |1/X(\tau)|^p)^{1/p}} \tag{3.12}$$

and

$$|Y'(\tau)|^p = 1 - |X'(\tau)|^p \quad \Leftrightarrow \quad |Y'(\tau)| = \frac{1/X(\tau)}{(1 + |1/X(\tau)|^p)^{1/p}}. \tag{3.13}$$

Furthermore, we can express equation (3.13) in terms of the function $Y$. To achieve this, we multiply (3.11) by $\exp(Y(\tau))$. Using the derivative

$$\exp(Y(\tau))Y'(\tau) = X'(\tau) \tag{3.14}$$

from $\exp(Y(\tau)) = X(\tau)$, we obtain

$$X'(\tau) = \exp(Y(\tau))\frac{X'(\tau)}{X(\tau)} \quad \Leftrightarrow \quad \frac{\exp'(Y(\tau))}{X(\tau)} = 1. \tag{3.15}$$

Consequently, equations (3.12), (3.13) and (3.15) allow us to express the following system of differential equations:

$$X'(\tau) = \pm\frac{1}{(1+|1/X(\tau)|^p)^{1/p}}, \tag{3.16}$$

$$Y'(\tau) = \pm\frac{1/X(\tau)}{(1+|1/X(\tau)|^p)^{1/p}} = \pm\frac{1}{(1+|\exp(Y(\tau))|^p)^{1/p}}. \tag{3.17}$$

There is no explicit formula for generic values of $p$ and it is difficult to calculate $X$ and $Y$ for an arbitrary value of $p$. However, although it is possible to find solutions for certain values of $p$, the case with which we obtain the best numerical results is when $p \to \infty$. In that case, the condition (3.10) simplifies to:

$$\max(|X'(\tau)|, |Y'(\tau)|) = 1.$$

Using $X(\tau) > 0$ and the derivatives (3.11) and (3.14), it follows that:

$$|X'(\tau)| = \min(1, X(\tau)), \tag{3.18}$$

$$|Y'(\tau)| = \frac{1}{\max(1, \exp(Y(\tau)))}. \tag{3.19}$$

Since the logarithm function is increasing, we can impose $Y'(\tau) > 0$ and then $X'(\tau) > 0$ from (3.11). By integrating equation (3.18), we get that there exists $c_1, c_2 \in \mathbb{R}$ such that

$$X(\tau) = \begin{cases} \exp(\tau) + c_1, & \text{if} \quad \tau \leq \ln(1 - c_1), \\ \tau + c_2, & \text{if} \quad \tau > 1 - c_2. \end{cases}$$

Let us choose $c_1 = 0$, consequently $c_2 = 1$ and we finally obtain:

$$(X(\tau), Y(\tau)) = \begin{cases} (\exp(\tau), \tau), & \text{if} \quad \tau < 0, \\ (\tau + 1, \ln(\tau + 1)), & \text{if} \quad \tau \geq 0. \end{cases} \tag{3.20}$$

In the following, we will refer to this choice for the parametrization as the *switch* since it can be thought as a mild way to implement the switch of variable procedure [23, 11]. Figure 3.1 illustrates these functions.

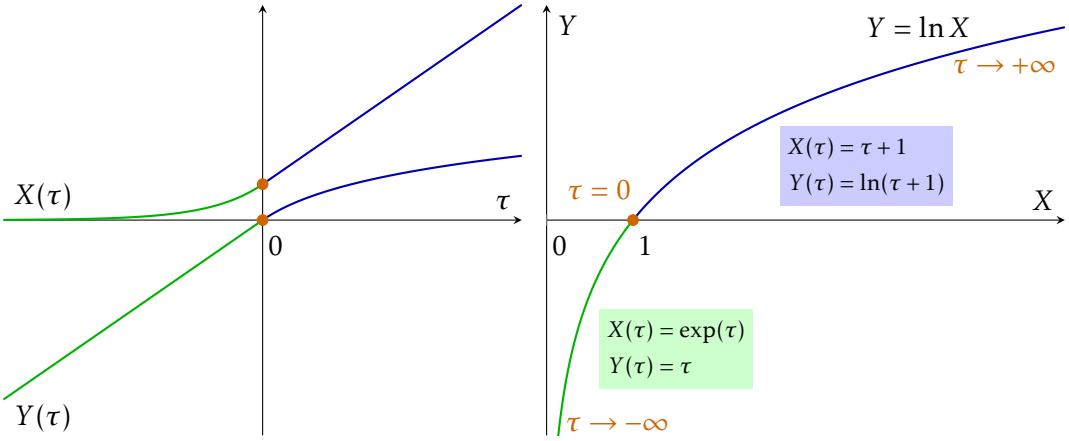


**Figure 3.1 –** The switch function.

### 3.1.3 A family of Cartesian representations

In the parametrization method, the relationship $y = \ln x$ is preserved throughout the iterations. The Cartesian representation technique, however, involves relaxing this relationship, which is equivalent to relaxing the relation $\exp(y) = x$. The core of this technique lies in the utilization of an augmented system, where the set of unknowns is expanded to the variables $(\mathbf{x}, \mathbf{y}, \omega)$. This expansion allows for greater flexibility in problem-solving. The naive approach consists of considering the following systems:

$$\begin{aligned} \mathcal{F}(\mathbf{x}, \mathbf{y}, \omega) &= \mathbf{0}, \\ \mathbf{y} - \ln(\mathbf{x}) &= \mathbf{0}, \end{aligned} \quad \text{or} \quad \begin{aligned} \mathcal{F}(\mathbf{x}, \mathbf{y}, \omega) &= \mathbf{0}, \\ \exp(\mathbf{y}) - \mathbf{x} &= \mathbf{0}. \end{aligned}$$

These two systems lead to the following Jacobian matrices:

$$\begin{bmatrix} \mathbf{A} & \mathbf{0} & -\mathbf{b} \\ \mathbf{0} & \mathbf{S}^T & \mathbf{0} \\ \mathbf{1}^T & \mathbf{0} & 0 \\ \text{diag}\{-1/\mathbf{x}\} & \mathbf{I} & \mathbf{0} \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \mathbf{A} & \mathbf{0} & -\mathbf{b} \\ \mathbf{0} & \mathbf{S}^T & \mathbf{0} \\ \mathbf{1}^T & \mathbf{0} & 0 \\ -\mathbf{I} & \text{diag}\{\exp(\mathbf{y})\} & \mathbf{0} \end{bmatrix}.$$

However, these matrices present problems of the same nature as the matrices $\nabla\mathcal{U}(\mathbf{x}, \omega)$ in (3.5) and $\nabla\mathcal{V}(\mathbf{y}, \omega)$ in (3.7) respectively. To tackle these issues, the idea of Cartesian representation is to introduce two Lipschitz continuous functions $\mathcal{Y} : \mathbb{R} \to \mathbb{R}$ and $\mathcal{X} : \mathbb{R} \to \mathbb{R}$ such that $\mathcal{X} = \mathcal{Y} \circ \ln$. These functions will be defined such that

$$y = \ln(x) \quad \Leftrightarrow \quad \mathcal{Y}(y) = \mathcal{Y}(\ln(x)) \quad \Leftrightarrow \quad \mathcal{Y}(y) = \mathcal{X}(x).$$

We then introduce the function $f : \mathbb{R}^2 \to \mathbb{R}^2$ as:

$$f(x_i, y_i) := \mathcal{Y}(y_i) - \mathcal{X}(x_i).$$

In this way, all the nonlinearities are in this new function and we can define the vector:

$$\boldsymbol{f}(\mathbf{x}, \mathbf{y}) = (f(x_i, y_i))_{i=1,\dots,N}.$$

Therefore, the system to solve is: find $(\mathbf{x}, \mathbf{y}, \omega) \in \mathbb{R}^{2N+1}$ such that

$$\mathcal{C}(\mathbf{x}, \mathbf{y}, \omega) := \begin{bmatrix} \mathcal{F}(\mathbf{x}, \mathbf{y}, \omega) \\ \boldsymbol{f}(\mathbf{x}, \mathbf{y}) \end{bmatrix} = \mathbf{0} \quad \Leftrightarrow \quad \begin{aligned} \mathbf{A}\mathbf{x} - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T\mathbf{y} - \mathbf{d} &= \mathbf{0}, \\ \langle \mathbf{x}, \mathbf{1} \rangle - 1 &= 0, \\ \boldsymbol{\mathcal{Y}}(\mathbf{y}) - \boldsymbol{\mathcal{X}}(\mathbf{x}) &= \mathbf{0}, \end{aligned} \tag{3.21}$$

where $\boldsymbol{\mathcal{Y}}(\mathbf{y}) = (\mathcal{Y}(y_i))_{i=1,\dots,N}$ and $\boldsymbol{\mathcal{X}}(\mathbf{x}) = (\mathcal{X}(x_i))_{i=1,\dots,N}$. The associated Jacobian matrix consists of two main components: a constant block and a variable block. The constant block incorporates the matrices $\mathbf{A}$ and $\mathbf{S}^T$, as well as the vectors $\mathbf{b}$ and $\mathbf{1}^T$. The variable block is composed of the partial derivatives of the function $f$. The complete Jacobian

matrix can be expressed as follows:

$$\nabla\mathcal{C}(\mathbf{x},\mathbf{y},\omega)=\begin{bmatrix}\mathbf{A}&\mathbf{0}&-\mathbf{b}\\ \mathbf{0}&\mathbf{S}^T&\mathbf{0}\\ \mathbf{1}^T&\mathbf{0}&0\\ \nabla_{\mathbf{x}}f&\nabla_{\mathbf{y}}f&\mathbf{0}\end{bmatrix}=\begin{bmatrix}\mathbf{A}&\mathbf{0}&-\mathbf{b}\\ \mathbf{0}&\mathbf{S}^T&\mathbf{0}\\ \mathbf{1}^T&\mathbf{0}&0\\ -\mathrm{diag}\{\mathfrak{X}'(\mathbf{x})\}&\mathrm{diag}\{\mathfrak{Y}'(\mathbf{y})\}&\mathbf{0}\end{bmatrix},$$

where $\mathfrak{X}'(\mathbf{x})=(\mathcal{X}'(x_i))_{i=1,\dots,N}$ and $\mathfrak{Y}'(\mathbf{y})=(\mathcal{Y}'(y_i))_{i=1,\dots,N}$. To prevent the aforementioned issues and to ensure that the coefficients of the Jacobian remain bounded, we aim to satisfy the following conditions for the $\mathcal{Y}$ and $\mathcal{X}$ functions. For all $x, y\in\mathbb{R}$:

(C1) $\mathcal{X}(x)=\mathcal{Y}(y)$ if and only if $y=\ln(x)$ where $\mathcal{X}$ and $\mathcal{Y}$ are monotonic and Lipschitz functions;

(C2) $\partial_x f=-\mathcal{X}'(x)$ and $\partial_y f=\mathcal{Y}'(y)$ are bounded and Lipschitz continuous functions;

(C3) $\partial_x f=-\mathcal{X}'(x)$ and $\partial_y f=\mathcal{Y}'(y)$ do not vanish simultaneously.

We then say that a Cartesian representation is admissible if it satisfies conditions (C1)–(C3). As for the parametrization, we introduce a normalization condition that takes the following form:

$$(|\mathcal{Y}'(y)|^p+|\mathcal{X}'(x)|^p)^{1/p}=1,\quad p\geq 1,\ y=\ln(x). \tag{3.22}$$

Using the same reasoning as we did for the parametrization, we can combine equations (3.22) with the derivative

$$\mathcal{X}'(x)=\frac{\mathcal{Y}'(\ln(x))}{x},$$

from $\mathcal{X}(x)=\mathcal{Y}\circ\ln(x)$, to obtain a system of differential equations:

$$\mathcal{X}'(x)=\pm\frac{1/x}{(1+|1/x|^p)^{1/p}}, \tag{3.23}$$

$$\mathcal{Y}'(v)=\pm\frac{1}{(1+|1/\exp(y)|^p)^{1/p}}=\pm\frac{\exp(y)}{(1+|\exp(y)|^p)^{1/p}}. \tag{3.24}$$

As for the parametrization, it is difficult to find solutions for an arbitrary value of $p$. For the numerical experiments, the case of interest is the asymptotic limit $p\to\infty$. The condition (3.22) then becomes:

$$\max(|\mathcal{Y}'(y)|,|\mathcal{X}'(x)|)=1,\quad y=\ln(x).$$

This condition leads to the following system of differential equations:

$$\mathcal{X}'(x) = \pm\frac{1/x}{\max(1, 1/x)}, \tag{3.25}$$

$$\mathcal{Y}'(y) = \pm\frac{\exp(y)}{\max(1, \exp(y))}. \tag{3.26}$$

Since the logarithm is an increasing function, we can impose that $\mathcal{X}'(x) > 0$. From the equivalence

$$y = \ln(x) \quad \Leftrightarrow \quad \mathcal{Y}(y) = \mathcal{X}(x),$$

we have $\mathcal{Y}'(y) > 0$ and we can impose $\mathcal{X}(1) = 0$, $\mathcal{Y}(0) = 0$. It follows from (3.25) and (3.26) that

$$\begin{aligned} y > 0 \quad &\Rightarrow \exp(y) > 1 \quad \Rightarrow \mathcal{Y}'(y) = 1 \qquad\quad \Rightarrow \mathcal{Y}(y) \underbrace{-\mathcal{Y}(0)}_{=0} = y - 0, \\ y \leq 0 \quad &\Rightarrow \exp(y) \leq 1 \quad \Rightarrow \mathcal{Y}'(y) = \exp(y) \quad \Rightarrow \mathcal{Y}(y) \underbrace{-\mathcal{Y}(0)}_{=0} = \exp(y) - 1. \end{aligned}$$

The function $\mathcal{Y}$ is then defined by:

$$\mathcal{Y}(y) = y\mathbf{1}_{\{y>0\}} + (\exp(y) - 1)\mathbf{1}_{\{y\leq 0\}}. \tag{3.27}$$

Consequently, we obtain the function $\mathcal{X}$:

$$\mathcal{X}(x) = \mathcal{Y}(\ln x) = \ln(x)\mathbf{1}_{\{x>1\}} + (x - 1)\mathbf{1}_{\{x\leq 1\}}. \tag{3.28}$$

Finally, the function $f$ is defined over four distinct regions as follows:

$$f(x, y) = \begin{cases} \mathrm{e}^y - x, & \text{if} \quad x \leq 1, y \leq 0, \\ y - x + 1, & \text{if} \quad x \leq 1, y \geq 0, \\ y - \ln x, & \text{if} \quad x \geq 1, y \geq 0, \\ \mathrm{e}^y - \ln x - 1, & \text{if} \quad x \geq 1, y \leq 0. \end{cases} \tag{3.29}$$

This function belongs to $\mathcal{C}^{1,1}(\mathbb{R}^2)$: it is continuous, differentiable and its gradient is Lipschitz continuous on $\mathbb{R}^2$. The function $f$, referred to as the *discrepancy function* and depicted on Figure 3.2, can readily be shown to be convex.

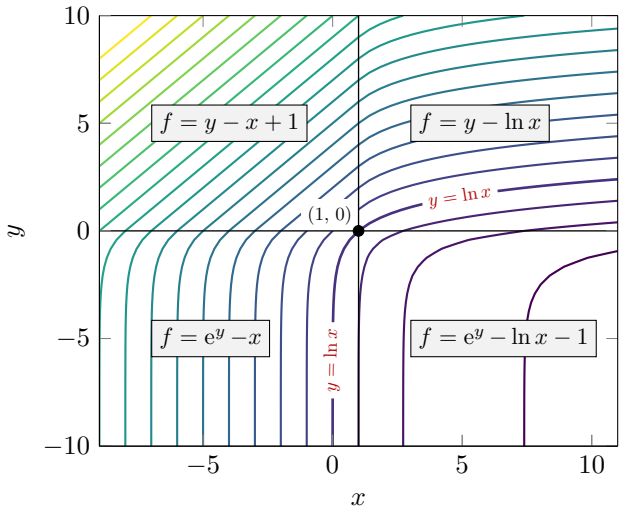


**Figure 3.2 –** Contour lines of the discrepancy function.

### Link between the parametrization and Cartesian representation

To conclude this section, we will demonstrate that the Cartesian representation is naturally associated to the parametrization. This link is given in the two following propositions.

**Proposition 3.1.1.** *Let $X(\tau)$, $Y(\tau)$ be an admissible parametrization in the sense of (P1)–(P3). Then there exists a Cartesian representation $f(x,y) = \mathcal{Y}(y) - \mathcal{X}(x)$ such that, for all $(x,y) \in \mathbb{R}^2$,*

$$\begin{aligned} \mathcal{X}'(x) &= Y'(X^{-1}(x)), \\ \mathcal{Y}'(y) &= X'(Y^{-1}(y)). \end{aligned} \tag{3.30}$$

*This Cartesian representation is admissible in the sense of (C1)–(C3). Moreover, it satisfies the normalization (3.22) if the parametrization satisfies the normalization (3.10).*

*Proof.* If $x = X(\tau)$ and $y = Y(\tau)$, by the invertibility of $X$ and $Y$ one can recover $\tau = X^{-1}(x) = Y^{-1}(y)$. A natural Cartesian representation is then $Y^{-1}(y) - X^{-1}(x) = 0$ or

$$\Psi(Y^{-1}(y)) - \Psi(X^{-1}(x)) = 0$$

for a suitable function $\Psi$. Setting $\mathcal{Y}(y) = \Psi(Y^{-1}(y))$ and $\mathcal{X}(x) = \Psi(X^{-1}(x))$, one has

$$\mathcal{X}'(x) = \frac{\Psi'(X^{-1}(x))}{X'(X^{-1}(x))} \quad \text{and} \quad \mathcal{Y}'(y) = \frac{\Psi'(Y^{-1}(y))}{Y'(Y^{-1}(y))}.$$

The result (3.30) is obtained by taking

$$\Psi(\tau) = \int^{\tau} X'(\theta)Y'(\theta)\,\mathrm{d}\theta.$$

□

**Proposition 3.1.2.** *Let $f(x, y) = \mathcal{Y}(y) - \mathcal{X}(x)$ be an admissible Cartesian representation in the sense of (C1)–(C3). Then, there exists a parametrization $X(\tau)$, $Y(\tau)$ such that, for all $\tau \in \mathbb{R}$,*

$$\begin{aligned} X'(\tau) &= \mathcal{Y}'(Y(\tau)), \\ Y'(\tau) &= \mathcal{X}'(X(\tau)). \end{aligned} \tag{3.31}$$

*This parametrization is admissible in the sense of (P1)–(P3) and satisfies the normalization (3.10) if the Cartesian representation satisfies the normalization (3.22).*

*Proof.* The existence of a solution to the ODE (3.31) is guaranteed by the hypothesis on $\mathcal{Y}'$ and $\mathcal{X}'$ and the Cauchy-Lipschitz theorem. Therefore

$$\frac{\mathrm{d}}{\mathrm{d}\tau} f(X(\tau), Y(\tau)) = \mathcal{Y}'(Y(\tau))Y'(\tau) - \mathcal{X}'(X(\tau))X'(\tau) = 0,$$

and it follows that $f(X(\tau), Y(\tau)) = c \in \mathbb{R}$. Moreover if $f(X(0), Y(0)) = 0$, then $c = 0$.

Since $\mathcal{X}$ and $\mathcal{Y}$ satisfy conditions (C1)–(C3), it follows that $X$ and $Y$ satisfy conditions (P1)–(P3). Furthermore, if $\mathcal{X}'$ and $\mathcal{Y}'$ fulfill the normalization condition given by (3.22), then the normalization conditions for $X'$ and $Y'$ follow from (3.31). □

## 3.2 Elements of theoretical analysis

In this section, we demonstrate the local quadratic convergence of Newton's algorithm applied to parametrization and Cartesian representation techniques for the chemical equilibrium problem.

### 3.2.1 About the parametrization

Let $X(\tau)$, $Y(\tau)$ be an admissible parametrization for the system (2.33) in the sense of (P1)–(P3). To proceed with our analysis, we must first restate the system of equations,

which is presented as follows:

$$\mathcal{P}(\boldsymbol{\tau}, \omega) = \mathbf{0} \quad \Leftrightarrow \quad \begin{aligned} \mathbf{A}\mathbf{X}(\boldsymbol{\tau}) - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T[\boldsymbol{\mu}^\circ/(\mathrm{RT}) + \mathbf{Y}(\boldsymbol{\tau})] &= \mathbf{0}, \\ \langle \mathbf{X}(\boldsymbol{\tau}), \mathbf{1}\rangle - 1 &= 0, \end{aligned} \tag{3.9}$$

together with its Jacobian

$$\nabla\mathcal{P}(\boldsymbol{\tau}, \omega) = \begin{bmatrix} \mathbf{A}\mathrm{diag}\{\mathbf{X}'(\boldsymbol{\tau})\} & -\mathbf{b} \\ \mathbf{S}^T\mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\} & \mathbf{0} \\ \mathbf{X}'(\boldsymbol{\tau})^T & 0 \end{bmatrix}.$$

The local quadratic theorem is as follows.

**Theorem 3.2.1.** *Let $X(\tau)$, $Y(\tau)$ be an admissible parametrization in the sense of assumptions (P1)–(P3). If the Newton sequence (3.2)–(3.3) is applied to the function $\mathcal{P}$ defined as (3.9), then the local quadratic convergence theorem holds.*

First, let us demonstrate that $\nabla\mathcal{P}(\boldsymbol{\tau}, \omega)$ is invertible at the solution point.

**Proposition 3.2.1.** *If $(\boldsymbol{\tau}, \omega)$ is solution of (3.9), then $\nabla\mathcal{P}(\boldsymbol{\tau}, \omega)$ is nonsingular.*

*Proof.* Let $(\delta\boldsymbol{\tau}, \delta\omega)^T \in \ker\nabla\mathcal{P}(\boldsymbol{\tau}, \omega)$, then

$$\mathbf{A}\mathrm{diag}\{\mathbf{X}'(\boldsymbol{\tau})\}\delta\boldsymbol{\tau} - \delta\omega\mathbf{b} = \mathbf{0}, \tag{3.32}$$

$$\mathbf{S}^T\mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\delta\boldsymbol{\tau} = \mathbf{0}, \tag{3.33}$$

$$\langle \mathbf{X}'(\boldsymbol{\tau}), \delta\boldsymbol{\tau}\rangle = 0. \tag{3.34}$$

Since $(\boldsymbol{\tau}, \omega)$ is solution of (3.9), one has

$$\mathbf{b} = \frac{1}{\omega}\mathbf{A}\mathbf{X}(\boldsymbol{\tau})$$

with $\omega > 0$. Equation (3.32) then becomes

$$\mathbf{A}\left[\mathrm{diag}\{\mathbf{X}'(\boldsymbol{\tau})\}\delta\tau - \frac{\delta\omega}{\omega}\mathbf{X}(\boldsymbol{\tau})\right] = \mathbf{0}. \tag{3.35}$$

Moreover, equation (3.33) indicates that

$$\mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\delta\boldsymbol{\tau} \in \ker\mathbf{S}^T \quad \Leftrightarrow \quad \mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\delta\boldsymbol{\tau} \in (\ker\mathbf{A})^\perp. \tag{3.36}$$

The equivalence in (3.36) is due to Lemma 2.1.1. Consequently, using (3.35), the following equality holds:

$$\begin{aligned}\langle \operatorname{diag}\{\mathbf{X}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau}, \operatorname{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau}\rangle &= \frac{\delta\omega}{\omega}\langle \mathbf{X}(\boldsymbol{\tau}), \operatorname{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau}\rangle \\ &= \frac{\delta\omega}{\omega}\langle \operatorname{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\mathbf{X}(\boldsymbol{\tau}), \boldsymbol{\delta\tau}\rangle.\end{aligned} \tag{3.37}$$

Using the derivative of $X(\tau) = \exp(Y(\tau))$, we get that

$$\mathbf{X}'(\boldsymbol{\tau}) = \operatorname{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\mathbf{X}(\boldsymbol{\tau}). \tag{3.38}$$

One can deduce that the right-hand side (3.37) equals zero due to equation (3.34). Therefore

$$\langle \operatorname{diag}\{\mathbf{X}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau}, \operatorname{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau}\rangle = 0,$$

which is only possible if $\boldsymbol{\delta\tau} = \mathbf{0}$. Indeed, using (3.38), we find that

$$\operatorname{diag}\{\mathbf{X}'(\boldsymbol{\tau})\}\operatorname{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\} = \operatorname{diag}\{\mathbf{X}(\boldsymbol{\tau})\}\operatorname{diag}\{\mathbf{Y}'(\boldsymbol{\tau})^2\}$$

is a positive-definite matrix since $X(\tau) > 0$ for all $\tau \in \mathbb{R}$ and $Y'(\tau) \neq 0$ thanks to (P3) and (3.11). Equation (3.32) finally allows to conclude that $\delta\omega = 0$, thereby proving that the Jacobian is nonsingular. □

Based on this result, we can demonstrate the local quadratic convergence of Newton's method.

*Proof of Theorem* 3.2.1. The proof of this result involves verifying that the assumptions of Theorem 3.1.1 are satisfied. The existence of a solution is guaranteed by Proposition 2.2.3 and the assumptions (P1)–(P3) regarding $X(\tau)$ and $Y(\tau)$. The Jacobian $\boldsymbol{\nabla}\mathcal{P}$ is Lipschitz continuous because $X'$ and $Y'$ are Lipschitz continuous, as stated in (P2). Furthermore, according to Proposition 3.2.1, $\boldsymbol{\nabla}\mathcal{P}$ is nonsingular at the solution point. □

### 3.2.2 About the Cartesian representation

Let $f(x, y)$ be an admissible Cartesian representation for the system (2.33) in the sense of (C1)–(C3). We restate that the system to solve is:

$$\mathcal{C}(\mathbf{x}, \mathbf{y}, \omega) = \mathbf{0} \quad \Leftrightarrow \quad \begin{aligned} \mathbf{A}\mathbf{x} - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T[\boldsymbol{\mu}^\circ/(\mathrm{RT}) + \mathbf{y}] &= \mathbf{0}, \\ \langle \mathbf{x}, \mathbf{1} \rangle - 1 &= 0, \\ \boldsymbol{f}(\mathbf{x}, \mathbf{y}) &= \mathbf{0}. \end{aligned} \tag{3.21}$$

The associated Jacobian matrix is written as

$$\nabla\mathcal{C}(\mathbf{x}, \mathbf{y}, \omega) = \begin{bmatrix} \mathbf{A} & \mathbf{0} & -\mathbf{b} \\ \mathbf{0} & \mathbf{S}^T & \mathbf{0} \\ \mathbf{1}^T & \mathbf{0} & 0 \\ \nabla_{\mathbf{x}}\boldsymbol{f} & \nabla_{\mathbf{y}}\boldsymbol{f} & \mathbf{0} \end{bmatrix},$$

where

$$\nabla_{\mathbf{x}}\boldsymbol{f} := \nabla_{\mathbf{x}}\boldsymbol{f}(\mathbf{x}, \mathbf{y}) = \mathrm{diag}\{\partial_{x_i} f(x_i, y_i)\}_{i=1,\ldots,N},$$
$$\nabla_{\mathbf{y}}\boldsymbol{f} := \nabla_{\mathbf{y}}\boldsymbol{f}(\mathbf{x}, \mathbf{y}) = \mathrm{diag}\{\partial_{y_i} f(x_i, y_i)\}_{i=1,\ldots,N}.$$

The local quadratic convergence theorem of Newton's method is as follows.

**Theorem 3.2.2.** *Let $f(x, y) = \mathcal{Y}(y) - \mathcal{X}(x)$ be an admissible Cartesian representation in the sense of assumptions (C1)–(C3). If the Newton sequence (3.2)–(3.3) is applied to the function $\mathcal{C}$ defined as (3.21), then the local quadratic convergence theorem holds.*

We will demonstrate that the Jacobian $\nabla\mathcal{C}$ is invertible for the unique vector $(\mathbf{x}, \mathbf{y}, \omega)^T$ that satisfies (3.21). To facilitate this, we will first establish the following two lemmas.

**Lemma 3.2.1.** *Let $f(x, y) = \mathcal{Y}(y) - \mathcal{X}(x)$ be an admissible Cartesian representation in the sense of (C1)–(C3). If $y = \ln(x)$, then*

$$-(\partial_x f)^{-1}\,\partial_y f = (\mathcal{X}'(x))^{-1}\mathcal{Y}'(y) = x.$$

*Proof.* Using $y = \ln(x)$ in the derivative

$$\mathcal{X}'(x) = \frac{\mathcal{Y}'(\ln(x))}{x},$$

it follows that

$$(\mathcal{X}'(x))^{-1}\mathcal{Y}'(y) = (\mathcal{Y}'(y)/x)^{-1}\mathcal{Y}'(y) = x.$$

□

**Lemma 3.2.2.** *The matrix defined as*

$$\mathbf{J}(\mathbf{x}, \mathbf{y}) := \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^T \\ \nabla_{\mathbf{x}} f & \nabla_{\mathbf{y}} f \end{bmatrix},$$

*which corresponds to the first $N$ rows, the last $N$ rows and the first $2N$ columns of $\nabla\mathcal{C}$, is invertible for all $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{2N}$.*

*Proof.* The proof consists of establishing that $\mathbf{J}^T$ is injective. Let

$$\boldsymbol{\mathcal{X}} = (\delta\mathbf{x}_1, \delta\mathbf{x}_2, \delta\mathbf{y})^T \in \mathbb{R}^M \times \mathbb{R}^{N-M} \times \mathbb{R}^N$$

be such that $\boldsymbol{\mathcal{X}} \in \ker \mathbf{J}^T(\mathbf{x}, \mathbf{y})$, with $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{2N}$. Therefore

$$\mathbf{J}^T(\mathbf{x},\mathbf{y})\boldsymbol{\mathcal{X}} = \mathbf{0} \quad \Leftrightarrow \quad \begin{aligned} \mathbf{A}^T \delta\mathbf{x}_1 &= -\nabla_{\mathbf{x}} f\, \delta\mathbf{y}, \\ \mathbf{S}\delta\mathbf{x}_2 &= -\nabla_{\mathbf{y}} f\, \delta\mathbf{y}. \end{aligned}$$

By performing the scalar product of these two terms, we find that

$$\delta\mathbf{x}_1^T(\mathbf{AS})\delta\mathbf{x}_2 = \delta\mathbf{y}^T(\nabla_{\mathbf{x}} f \nabla_{\mathbf{y}} f)\delta\mathbf{y}.$$

The product $\mathbf{AS}$ vanishes due to Lemma 2.1.1, then

$$\delta\mathbf{y}^T(\nabla_{\mathbf{x}} f \nabla_{\mathbf{y}} f)\delta\mathbf{y} = 0.$$

Given that

$$\begin{aligned} \partial_{x_i} f(x_i, y_i) &= -\mathcal{X}'(x_i), \\ \partial_{y_i} f(x_i, y_i) &= \mathcal{Y}'(y_i), \end{aligned}$$

and noting that $\mathcal{X}'$ and $\mathcal{Y}'$ have the same sign, we can conclude that the matrix $\nabla_{\mathbf{x}} f \nabla_{\mathbf{y}} f$ is negative-definite. Consequently, $\delta\mathbf{y} = \mathbf{0}$. It follows that $\delta\mathbf{x}_1 \in \ker \mathbf{A}^T = \{\mathbf{0}_{\mathbb{R}^M}\}$ and $\delta\mathbf{x}_2 \in \ker \mathbf{S} = \{\mathbf{0}_{\mathbb{R}^{N-M}}\}$ given that $\mathbf{A}^T$ and $\mathbf{S}$ have full rank. Thus, we have demonstrated that $\mathbf{J}^T(\mathbf{x},\mathbf{y})$ is invertible, which implies that $\mathbf{J}(\mathbf{x},\mathbf{y})$ is also invertible. □

We can now demonstrate the invertibility of $\nabla\mathcal{C}$.

**Proposition 3.2.2.** *If* $\boldsymbol{\mathcal{X}} = (\mathbf{x}, \mathbf{y}, \omega)^T$ *is solution of* (3.21)*, then* $\boldsymbol{\nabla}\mathcal{C}(\boldsymbol{\mathcal{X}})$ *is nonsingular.*

*Proof.* Let $\alpha \in \mathbb{R}$ be a parameter. For $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{2N}$, one denotes by $\delta\tilde{\boldsymbol{\mathcal{X}}}_\alpha = (\delta\mathbf{x}_\alpha, \delta\mathbf{y}_\alpha)^T$ the unique solution of

$$\mathbf{J}(\mathbf{x}, \mathbf{y})\delta\tilde{\boldsymbol{\mathcal{X}}}_\alpha = \begin{pmatrix} \alpha\mathbf{b} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

The existence of this solution is always guaranteed by the invertibility of $\mathbf{J}$ as established in Lemma 3.2.2. In particular, the solution satisfies

$$\delta\tilde{\boldsymbol{\mathcal{X}}}_\alpha = \alpha\delta\tilde{\boldsymbol{\mathcal{X}}}_1.$$

We then define the vector

$$\delta\boldsymbol{\mathcal{X}} := (\delta\tilde{\boldsymbol{\mathcal{X}}}_{\delta\omega}, \delta\omega)^T = \delta\omega(\delta\tilde{\boldsymbol{\mathcal{X}}}_1, 1)^T.$$

It follows that

$$\boldsymbol{\nabla}\mathcal{C}(\mathbf{x}, \mathbf{y}, \omega)\delta\boldsymbol{\mathcal{X}} = \mathbf{0} \quad \Leftrightarrow \quad \begin{aligned} \delta\omega\left[\mathbf{J}(\mathbf{x}, \mathbf{y})\delta\tilde{\boldsymbol{\mathcal{X}}}_1 - (\mathbf{b}, \mathbf{0}, \mathbf{0})^T\right] &= \mathbf{0}, \\ \delta\omega\langle\delta\mathbf{x}_1, \mathbf{1}\rangle &= \mathbf{0}. \end{aligned}$$

By the definition of $\tilde{\boldsymbol{\mathcal{X}}}_1$ one has

$$\mathbf{J}(\mathbf{x}, \mathbf{y})\delta\tilde{\boldsymbol{\mathcal{X}}}_1 - (\mathbf{b}, \mathbf{0}, \mathbf{0})^T = \mathbf{0}.$$

Therefore, the invertibility of $\boldsymbol{\nabla}\mathcal{C}$ is determined by $\delta\omega\langle\delta\mathbf{x}_1, \mathbf{1}\rangle = \mathbf{0}$:

- if $\langle\delta\mathbf{x}_1, \mathbf{1}\rangle \neq 0$, then $\delta\omega = 0$ and it follows that the matrix $\mathbf{J}(\mathbf{x}, \mathbf{y}, \omega)$ is invertible;
- otherwise if $\langle\delta\mathbf{x}_1, \mathbf{1}\rangle = 0,\ \delta\mathbf{x}_1 \neq \mathbf{0}$, then $\ker\boldsymbol{\nabla}\mathcal{C}(\mathbf{x}, \mathbf{y}, \omega) = \mathrm{Vect}\{(\delta\tilde{\boldsymbol{\mathcal{X}}}_1, 1)^T\}$.

To demonstrate that $\boldsymbol{\nabla}\mathcal{C}(\mathbf{x}, \mathbf{y}, \omega)$ is invertible for the solution $(\mathbf{x}, \mathbf{y}, \omega)$ of (3.21), it is sufficient to establish that $\langle\delta\mathbf{x}_1, \mathbf{1}\rangle \neq 0$, where $(\delta\mathbf{x}_1, \delta\mathbf{y}_1)$ the unique solution of:

$$\mathbf{A}\delta\mathbf{x}_1 = \mathbf{b}, \tag{3.39}$$

$$\mathbf{S}^T\delta\mathbf{y}_1 = \mathbf{0}, \tag{3.40}$$

$$\boldsymbol{\nabla}_{\mathbf{x}}f\,\delta\mathbf{x}_1 + \boldsymbol{\nabla}_{\mathbf{y}}f\,\delta\mathbf{y}_1 = \mathbf{0}. \tag{3.41}$$

By defining $\mathbf{D} := -(\boldsymbol{\nabla}_{\mathbf{x}}f)^{-1}\boldsymbol{\nabla}_{\mathbf{y}}f$, we obtain

$$\delta\mathbf{x}_1 = \mathbf{D}\delta\mathbf{y}_1 \tag{3.42}$$

from (3.41). Additionally, from (3.40) and Lemma 2.1.1, we know that $\delta \mathbf{y}_1 \in \ker \mathbf{S}^T = \operatorname{Im} \mathbf{A}^T$. This implies the existence of $\delta \mathbf{h}_1$ such that

$$\delta \mathbf{y}_1 = \mathbf{A}^T \delta \mathbf{h}_1. \tag{3.43}$$

Consequently, equation (3.39) can be rewritten as:

$$\mathbf{A}\mathbf{D}\mathbf{A}^T \delta \mathbf{h}_1 = \mathbf{b}. \tag{3.44}$$

The matrix $\mathbf{A}\mathbf{D}\mathbf{A}^T = \mathbf{A}\mathbf{D}^{1/2}\left(\mathbf{A}\mathbf{D}^{1/2}\right)^T$ is invertible because the rank of $\mathbf{A}\mathbf{D}^{1/2}$ is maximal. Furthermore, since $(\mathbf{x}, \mathbf{y}, \omega)$ is a solution to (3.21), we have $\mathbf{b} = \frac{1}{\omega}\mathbf{A}\mathbf{x}$. From (3.44), it follows that

$$\delta \mathbf{h}_1 = \frac{1}{\omega}(\mathbf{A}\mathbf{D}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{x}.$$

By multiplying this equation on the left by $\mathbf{D}\mathbf{A}^T$, we obtain

$$\mathbf{D}\mathbf{A}^T \delta \mathbf{h}_1 = \frac{1}{\omega}\mathbf{D}\mathbf{A}^T(\mathbf{A}\mathbf{D}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{x}.$$

Using (3.42) and (3.43), we derive:

$$\begin{aligned} \delta \mathbf{x}_1 &= \frac{1}{\omega}\mathbf{D}\mathbf{A}^T(\mathbf{A}\mathbf{D}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{x} \\ &= \frac{1}{\omega}\mathbf{D}^{1/2}\left[(\mathbf{A}\mathbf{D}^{1/2})^T(\mathbf{A}\mathbf{D}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{D}^{1/2}\right]\mathbf{D}^{-1/2}\mathbf{x}. \end{aligned} \tag{3.45}$$

Let $\mathbf{B} := \mathbf{A}\mathbf{D}^{1/2}$. Then, the matrix $\boldsymbol{\Pi} := \mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}$ represents the orthogonal projection onto $(\ker \mathbf{B})^\perp$. Consequently, equation (3.45) can be rewritten as:

$$\delta \mathbf{x}_1 = \frac{1}{\omega}\mathbf{D}^{1/2}\boldsymbol{\Pi}\mathbf{D}^{-1/2}\mathbf{x},$$

Given that an orthogonal projection satisfies the property $\boldsymbol{\Pi}^2 = \boldsymbol{\Pi}$, we further simplify this equation to obtain:

$$\delta \mathbf{x}_1 = \frac{1}{\omega}\mathbf{D}^{1/2}\boldsymbol{\Pi}^2\mathbf{D}^{-1/2}\mathbf{x}. \tag{3.46}$$

Since $(\mathbf{x}, \mathbf{y}, \omega)$ is a solution to the system (3.21), we can apply Lemma 3.2.1 to conclude that $\mathbf{D} = \operatorname{diag}\{x_i\}_{i=1,\dots,N}$. Consequently, we have:

$$\mathbf{1}^T\mathbf{D}^{1/2} = (\mathbf{D}^{-1/2}\mathbf{x})^T = (\mathbf{x}^{1/2})^T.$$

Taking the scalar product of the vector $\mathbf{1}$ with equation (3.46), we obtain:

$$\langle \delta \mathbf{x}_1, \mathbf{1} \rangle = \frac{1}{\omega} \left\| \boldsymbol{\Pi} \mathbf{x}^{1/2} \right\|^2.$$

This leads us to the equivalence:

$$\begin{aligned} \langle \delta \mathbf{x}_1, \mathbf{1} \rangle = 0 &\Leftrightarrow \mathbf{x}^{1/2} \in \ker \boldsymbol{\Pi} \\ &\Leftrightarrow \mathbf{x}^{1/2} \in \ker \mathbf{B} \\ &\Leftrightarrow \mathbf{A}\mathbf{D}^{1/2}\mathbf{x}^{1/2} = \mathbf{A}\mathbf{x} = 0. \end{aligned}$$

However, this scenario is impossible because $\mathbf{A}\mathbf{x} = \omega \mathbf{b} \neq 0$. Therefore, we conclude that $\langle \delta \mathbf{x}_1, \mathbf{1} \rangle \neq 0$, which implies that the Jacobian is invertible. □

Finally, the proof of the local quadratic convergence is as follows.

*Proof of Theorem 3.2.2.* The proof consists of verifying that the assumptions of Theorem 3.1.1 are satisfied. The existence of a solution is guaranteed by Proposition 2.2.3 and the assumptions (C1)–(C3) regarding $\mathcal{Y}(y)$ and $\mathcal{X}(x)$. The Jacobian $\nabla \mathcal{C}$ is Lipschitz continuous since $\mathcal{Y}'$ and $\mathcal{X}'$ are Lipschitz continuous according to (C2). Moreover, from Proposition 3.2.2, $\nabla \mathcal{C}$ is nonsingular at the solution point. □

To conclude this section, we will demonstrate an interesting property of the Cartesian representation associated with the function (3.29): the iterates of Newton's method consistently lie above the graph of the logarithm.

**Proposition 3.2.3.** *Let $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}, \omega^{(k)})$ be a Newton iterate for the Cartesian representation formulation described in Section 3.1.3 with discrepancy function $f$ defined by (3.29). Then, for $k \geq 1$, the linear equations*

$$\begin{aligned} \mathbf{A}\mathbf{x}^{(k)} &= \omega^{(k)} \mathbf{b}, \\ \mathbf{S}^T \mathbf{y}^{(k)} &= \mathbf{d}, \\ \langle \mathbf{x}^{(k)}, \mathbf{1} \rangle &= 1, \end{aligned}$$

*are satisfied, whereas $f(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) \geq \mathbf{0}$ componentwise.*

*Proof.* The fact that the linear equations are solved exactly by Newton's method is a

well-known fact. As the discrepancy function $f$ is convex, one has

$$f\left(x_i^{(k)}, y_i^{(k)}\right) \geq f\left(x_i^{(k-1)}, y_i^{(k-1)}\right) + \partial_x f\left(x_i^{(k-1)}, y_i^{(k-1)}\right)\delta x_i^{(k-1)} + \partial_y f\left(x_i^{(k-1)}, y_i^{(k-1)}\right)\delta y_i^{(k-1)}.$$

It follows that the right hand side this inequality is zero from the definition of the increment $\delta\mathbf{x}^{(k-1)}$ by Newton's method. □

## 3.3 Numerical experiments

In this section, we present a comparative study of the convergence of our proposed methods against the traditional log-trick approach found in existing literature. This comparison is based on three test cases, each increasing in complexity.

In the first two scenarios, our methods demonstrate strong accuracy and effectively stabilize the initial iterations of the Newton method. However, the complexity of the final test case requires the integration of a linear search technique to ensure global convergence of the Newton method. The log-trick, combined with this line search, is the method implemented in the Arxim solver [47]. This additional technique enhances the convergence of our methods, which the traditional approach struggled to achieve. This strategy is detailed in Chapter 9.7.1 of the book Numerical Recipes [54] and in Appendix B.2.

Finally, we conduct a comparison across various initializations, highlighting the superior performance of the Cartesian representation method.

### 3.3.1 Numerical parameters

The functions $X$ and $Y$ for the parametrization are those of the switch defined in (3.20). For the Cartesian representation technique, the function $f$ is the discrepancy function defined in (3.29). In all numerical experiments, pressure and temperature values are set at P= 1 Bar and T = 298.15 K. Moreover, if $\mathcal{F}(\boldsymbol{\mathcal{X}})$ represents the function for which we are seeking the root, the convergence criterion for Newton's algorithm is

$$\|\mathcal{F}(\boldsymbol{\mathcal{X}}^{(k+1)})\|_\infty \leq 10^{-10} \quad \text{and} \quad \|\boldsymbol{\mathcal{X}}^{(k+1)} - \boldsymbol{\mathcal{X}}^{(k)}\|_\infty \leq 10^{-10}$$

where $k+1$ is the current Newton iteration.

It is important to mention the different ways of initializing the Newton algorithm, depending on the method. Starting from an initial guess $\mathbf{n} = (n_1, \ldots, n_N)$, the variable

$\omega$ is defined as $1/\langle \mathbf{n}, \mathbf{1} \rangle$ and the mole fractions $\mathbf{x}$ as $x_i = \omega n_i$ (see Section 2.2.3). The initialization for each method is described below.

— For the log-trick, the variables are the logarithms of the mole fractions, so the initial guess is

$$\boldsymbol{\mathcal{X}}^{(0)} = [\omega, \ln(x_1), \ldots, \ln(x_N)].$$

— For the parametrization, the function $X(\tau)$ is inverted to initialize $\boldsymbol{\tau}$:

$$\boldsymbol{\mathcal{X}}^{(0)} = [\omega, X^{-1}(x_1), \ldots, X^{-1}(x_N)].$$

— For the Cartesian representation, the variables of $\mathbf{y}$ are initialized as the logarithms of the variables of $\mathbf{x}$:

$$\boldsymbol{\mathcal{X}}^{(0)} = [\omega, x_1, \ldots, x_N, \ln(x_1), \ldots, \ln(x_N)].$$

### 3.3.2 Test cases presentation

We will study the following 3 test cases.

— The $H_2O$ test case: 3 species, 2 elements and 1 reaction.

— The *Seawater* test case: 37 species, 10 elements and 27 reactions.

— The *Water-Concrete* test cases: 88 species, 12 elements and 75 reactions.

All the chemical systems involved in these tests cases are detailed in Appendix A.2 together with the solution of the chemical equilibria. The first two, $H_2O$ and *Seawater*, use the charge constraint defined in section 2.2.4 while *Water-Concrete* use the charge constraint and the pE constraint defined in section 2.2.4. Let $n_1 = n_{H_2O}$, the initial vector $\mathbf{n}^{(0)}$ for Newton's method is as follows:

$$\mathbf{n}^{(0)} = (n_1, n_{j\neq 1}) = (55, 1).$$

### 3.3.3 Numerical results

#### $H_2O$ test case

For this test case, we will investigate the evolution of residuals. We will then explain the differences in convergence speed between our parametrization and Cartesian representation methods and the classical log-trick approach.

##### Evolution of residuals

Figure 3.3 shows the evolution of the residuals of our methods compared with the classical log-trick approach. The graph on the left shows the evolution of the norm of the function at iterate $k$, while that on the right shows the evolution of the norm between iterates $k$ and $k-1$. There is a significant decrease in the norm of the function between the second and third iterations for the parametrization and Cartesian representation method, in contrast to the log-trick, which converges more slowly to the solution. The plateau observed on the left-hand graph indicates that the convergence of the iterates $\|\boldsymbol{\mathcal{X}}^{(k)} - \boldsymbol{\mathcal{X}}^{(k-1)}\|_\infty$ is slower than that of the residuals $\|\mathcal{F}(\boldsymbol{\mathcal{X}}^{(k)})\|_\infty$, justifying the use of both convergence criteria to guarantee the accuracy of the results.

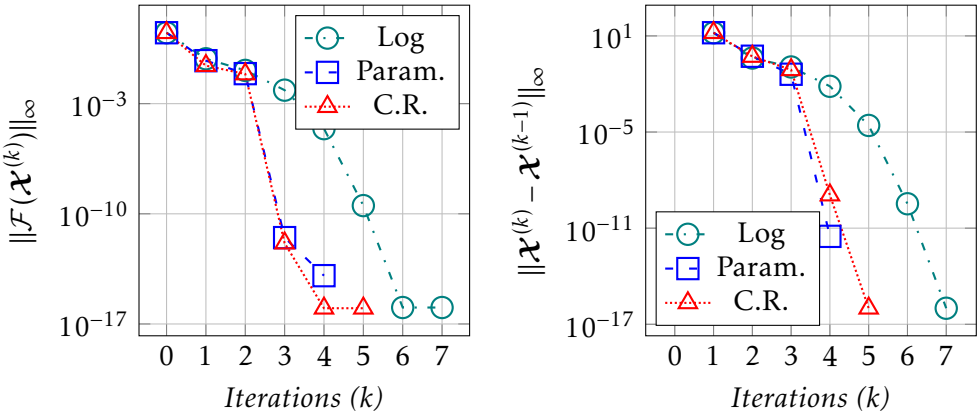


**Figure 3.3 –** Evolution of residuals for the $H_2O$ test case.

**Comparison between the log-trick and the parametrization**

To explain the difference in convergence speed between the parametrization and the log-trick, Figure 3.4 shows the evolution of the norm of the function at the $k$-th iterate restricted to the equations of conservation of elements for the graph on the left, and to the equilibrium equations for the graph on the right. Figure 3.5 describes the evolution of iterates for the species $H_2O$ and its mole fraction. Table 3.1 shows which parametrization function is used at each iteration. Finally, Figure 3.6 shows the evolution of the iterates of species $H^+$ and $OH^-$.

Figure 3.4 shows that the significant decrease in residual is due to the chemical equilibrium equation being completely solved from iteration 3 onward. Indeed, from iteration 2 onward, the equilibrium equation is linear, and Newton's method is known to solve linear equations exactly. Figure 3.6 shows that the iterates $\tau_{H^+}$ and $\tau_{OH^-}$ are always negative, so that $Y(\tau_{H^+})$ and $Y(\tau_{OH^-})$ are always linear. The linearity or non-linearity of the equilibrium equation therefore depends only on the value of $\tau_{H_2O}$, and Figure 3.5

and Table 3.1 clearly show that this value is negative from the second iteration onward, making the equilibrium equation linear.

From Figure 3.5, it is important to note that on the first iteration, the $\tau_{H_2O}$ of the parameterization and the $\ln(x_{H_2O})$ of the log-trick have identical values, but that the corresponding mole fractions $X(\tau_{H_2O})$ and $x_{H_2O}$ are not the same, due to the fact that $X(\tau)$ is linear for a positive value of $\tau$. This prevents the mole fraction value from being too high, and leads to a value of $\tau_{H_2O}$ much closer to the solution at iteration 2, in contrast to log-trick. The parametrization mechanism therefore accelerated convergence towards the solution in this test case.

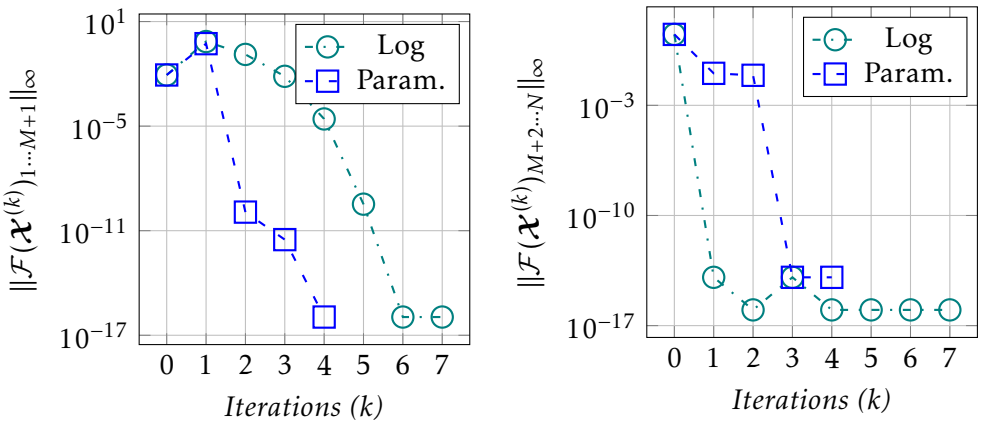


**Figure 3.4 –** Evolution of residuals of the log-trick and the parametrization for the $H_2O$ test case. The left graph represents the residuals of the conservation equations while the right one represents the residuals of the equilibrium equation.

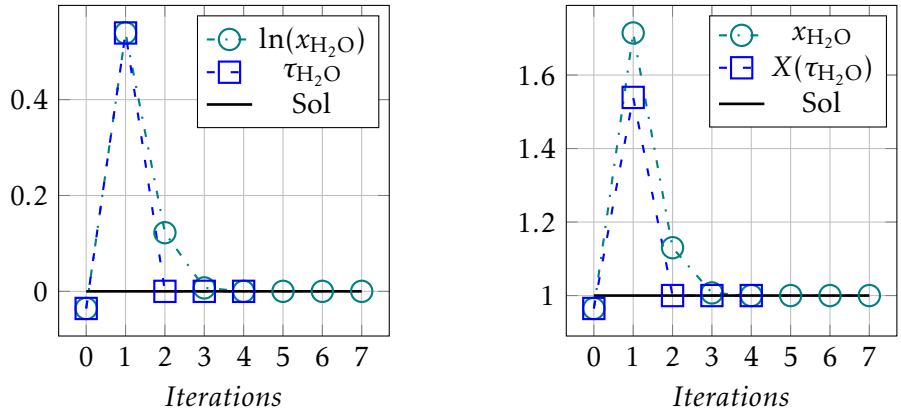


**Figure 3.5 –** Evolution of the species $H_2O$ with the log-trick and parametrization methods.

| Iteration | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $\text{sign}(\tau_{H_2O})$ | $-$ | $+$ | $-$ | $-$ | $-$ |
| $X(\tau_{H_2O})$ | $\exp \tau_{H_2O}$ | $\tau_{H_2O} + 1$ | $\exp \tau_{H_2O}$ | $\exp \tau_{H_2O}$ | $\exp \tau_{H_2O}$ |
| $Y(\tau_{H_2O})$ | $\tau_{H_2O}$ | $\ln \tau_{H_2O}$ | $\tau_{H_2O}$ | $\tau_{H_2O}$ | $\tau_{H_2O}$ |

**Table 3.1 –** Evolution of $X(\tau_{H_2O})$ and $Y(\tau_{H_2O})$.

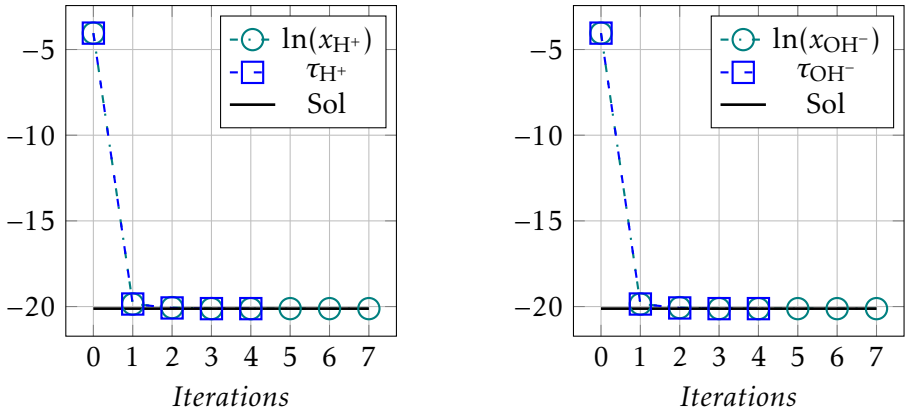


**Figure 3.6 –** Evolution of the species $H^+$ and $OH^-$ for the log-trick and parametrization method.

**Convergence of the Cartesian representation compared to the log-trick**

The difference in convergence speed between the Cartesian representation and the log-trick observed at the third iteration corresponds to the moment when the $x$ and $y$ variables of the species $H_2O$ are again in the lower left-hand zone in the Figure 3.7, left-hand graph. The graph on the right shows the decay of the residuals for equations linked to the $f$ function. As expected, the linear equations are solved starting from the first iteration. We observe in Figure 3.7 and Figure 3.8 that the link $y = \ln x$ is strongly broken after the first iteration, expressing the fact that the $x$ and $y$ variables evolve separately.

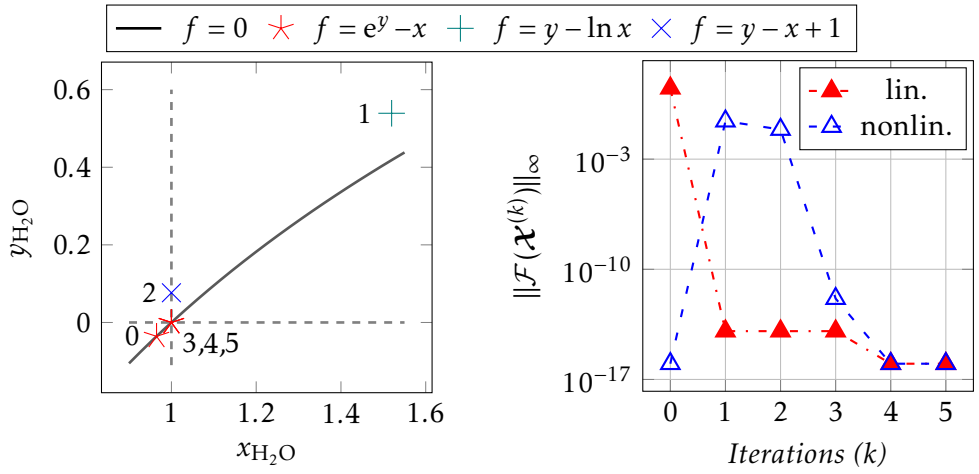


**Figure 3.7 –** Evolution of iterates for the species $H_2O$ (left) and evolution of the residuals for linear and nonlinear equations of the system (right).

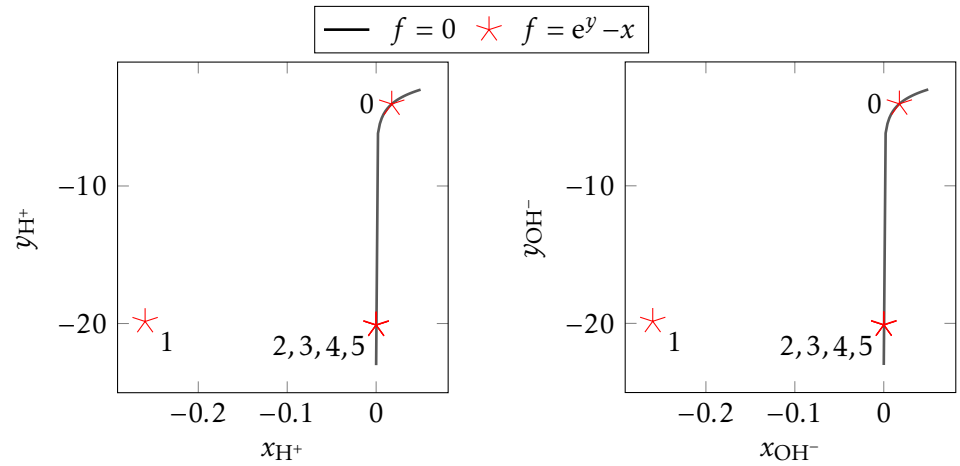


**Figure 3.8 –** Evolution of iterates for the species $H^+$ and $OH^-$.

### *Seawater* test case

This test case, presented in A.2.2, contains many more species but remains fairly simple, allowing us to validate the convergence of our methods. For the initialization considered, the number of iterations is the same for all methods. However, we observe an increase in the residual during the first iteration for the log-trick, as shown in Figure 3.9. In the case of parametrization, Table 3.2 indicates that there are only three species for which the $\tau$ parameter changes sign at the first iteration. From the second iteration onward, the chemical equilibrium equations are linear and thus resolved from the

| Iteration | 0 | 1 | 2 | ⋯ | 21 |
|---|---|---|---|---|---|
| $\text{sign}(\tau_i)$ | $-$ | $+$ | $-$ | ⋯ | $-$ |
| $X(\tau_i)$ | $\exp\tau_i$ | $\tau_i + 1$ | $\exp\tau_i$ | ⋯ | $\exp\tau_i$ |
| $Y(\tau_i)$ | $\tau_i$ | $\ln\tau_i$ | $\tau_i$ | ⋯ | $\tau_i$ |

**Table 3.2 –** Evolution of $\tau_i$ for $i \in \{H_2O, K^+, KSO_4^-\}$

third iteration onward, as illustrated in Figure 3.10. Additionally, Figure 3.11 and 3.12 show that there are only three species whose variables $(x, y)$ change zones, and from the third iteration onward, there is no longer any change in the function for the Cartesian representation.

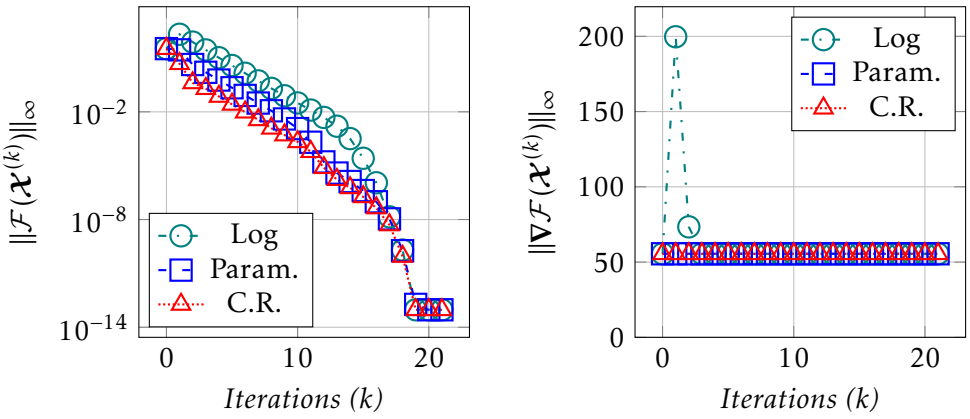


**Figure 3.9 –** Evolution of residuals and norm of the Jacobian matrix.

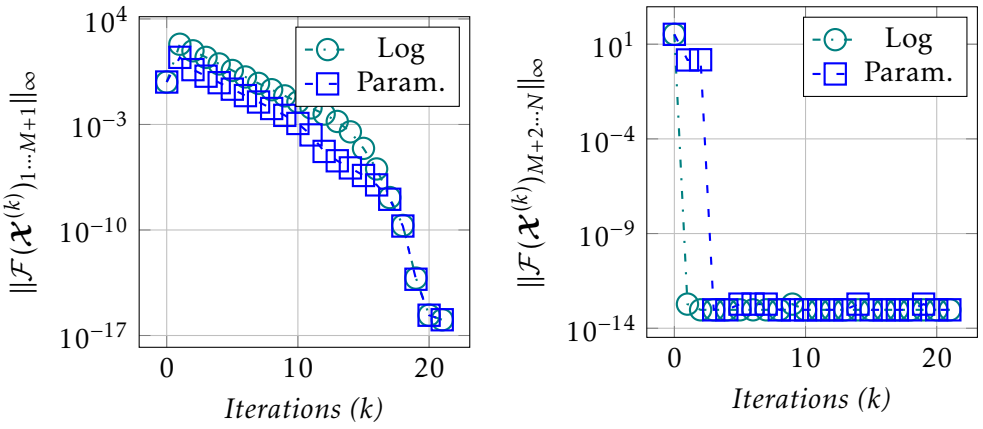


**Figure 3.10 –** Evolution of residuals of the log-trick and the parametrization for the *Seawater* test case. The left graph represents the residuals of the conservation equations while the right one represents the residuals of the equilibrium equations.

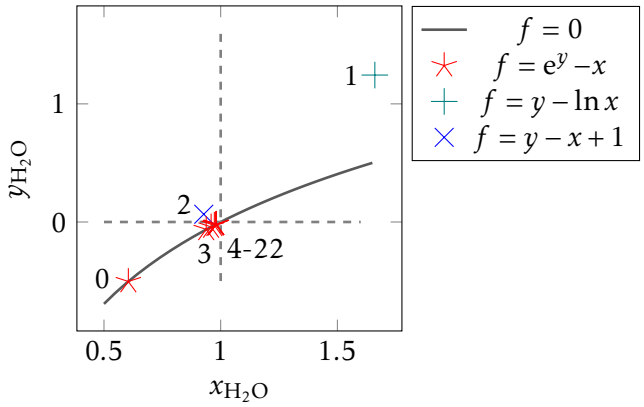


**Figure 3.11 –** Evolution of iterates for the species $H_2O$ for the Cartesian representation

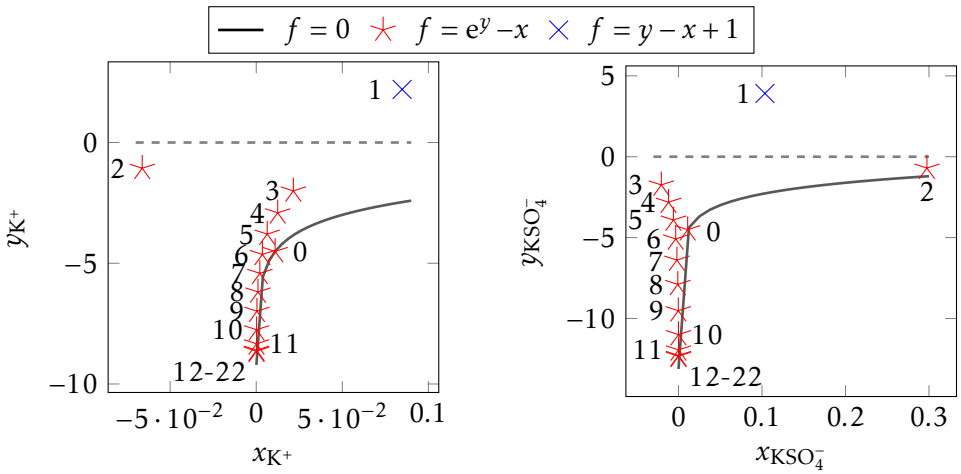


**Figure 3.12 –** Evolution of iterates for the species $K^+$ and $KSO_4^-$ for the Cartesian representation.

***Water-Concrete* test case**

The *Water-Concrete* test case differs from the previous ones in that it includes a redox constraint. This constraint leads to very low concentrations of certain chemical species. Figure 3.13 illustrates the evolution of the residuals. Unlike the other test cases, this one exhibits no convergence: the log-trick diverges very quickly, the parametrization diverges after 33 iterations, while a cycle forms for the Cartesian representation.

To achieve convergence for this test case and initialization, we incorporated a line search into Newton's method. Figure 3.14 presents the results of the three methods with the line search: both our parametrization and Cartesian representation methods converge through a plateau at around $10^{-3}$, whereas the log-trick diverges after the sixth iteration. It should be emphasized that the Arxim solver developed by IFPEN implements a method that combines the log-trick with line-search techniques.

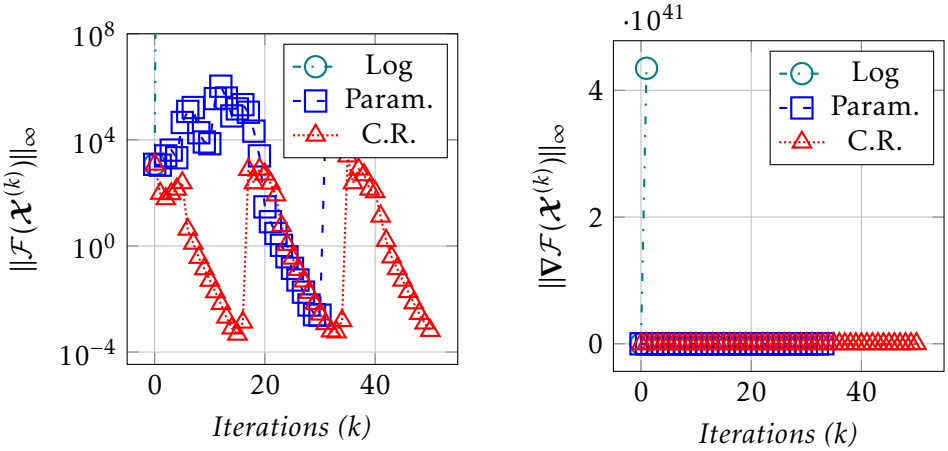


**Figure 3.13 –** Evolution of residuals and norm of the Jacobian matrix without the line search.

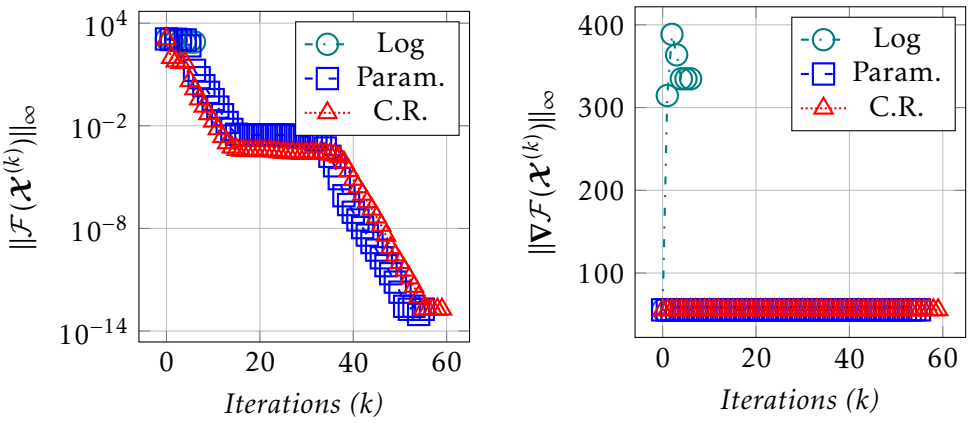


**Figure 3.14 –** Evolution of residuals and norm of the Jacobian matrix with the line search.

### 3.3.4 Study of sensitivity to initialization

In the final test case, we observed that the convergence of Newton's algorithm is not always guaranteed with the provided initialization. Although the line search technique ensures convergence, it is worthwhile to investigate the sensitivity to initializations to determine if better performance can be achieved. In this section, we will evaluate the methods on the Water-Concrete test case using the following four different initializations:

$$\begin{aligned}
\mathbf{n}_1^{(0)} &= (n_1, n_{j\neq 1}) = (55, 10^{-2});\\
\mathbf{n}_2^{(0)} &= (n_1, n_{j\neq 1}) = (55, 10^{-4});\\
\mathbf{n}_3^{(0)} &= (n_1, n_{j\neq 1}) = (55, \epsilon_{32});\\
\mathbf{n}_4^{(0)} &= (n_1, n_{j\neq 1}) = (55, \epsilon_{64});
\end{aligned}$$

where $\epsilon_{32} = 1.1920929 \times 10^{-7}$ and $\epsilon_{64} = 2.220446049250313 \times 10^{-16}$ are the machine epsilons in single and double precision, respectively.

Additionally, the Cartesian representation requires the initialization of the vector $\mathbf{y}$. The graphs in $(x, y)$ space of Figures 3.11 and 3.12 indicate that the relationship $y = \ln x$ is only satisfied in the final iterations, with the trend of $y$ appearing to decrease. Consequently, we propose a new initialization for $y$:

$$y^{(0)} = x^{(0)} - 1. \tag{3.47}$$

This approach aims to avoid starting with a $y$ value that is too low while maintaining a connection with $x$.

The results presented in Table 3.3 demonstrate that the parametrization and log-trick approaches are not robust against initialization for this challenging test case. In contrast, the Cartesian representation is very robust when the initialization (3.47) is used instead of $y = \ln x$.

| **Method** \ **Initial guess** | $\mathbf{n}_1^{(0)}$ | $\mathbf{n}_2^{(0)}$ | $\mathbf{n}_3^{(0)}$ | $\mathbf{n}_4^{(0)}$ |
|---|---|---|---|---|
| Parametrization | × | 60 | × | × |
| CR + y=ln x | × | × | × | × |
| CR + y=x-1 | 50 | 69 | 69 | 69 |
| log-trick | × | × | × | × |

**Table 3.3 –** Number of iterations of Newton's method for *Water-Concrete* with various initializations.

To further compare our methods, Table 3.4 presents the results obtained with the addition of the line search. As mentioned earlier, the log-trick approach combined with the line search is comparable to the ArXim solver. Table 3.4 demonstrates the strong robustness of the Cartesian representation method compared to the other approaches, including the ArXim solver.

| **Method** \ **Initial guess** | $\mathbf{n}_1^{(0)}$ | $\mathbf{n}_2^{(0)}$ | $\mathbf{n}_3^{(0)}$ | $\mathbf{n}_4^{(0)}$ |
|---|---|---|---|---|
| Parametrization + ls | 57 | 58 | × | × |
| CR + y=ln x + ls | 57 | 59 | 37 | × |
| CR + y=x-1 + ls | 58 | 58 | 58 | 58 |
| log-trick + ls | 57 | 53 | × | 37 |

**Table 3.4 –** Number of iterations of Newton's method for *Water-Concrete* with the line search.

Chapter 4

# Innovative numerical methods for multiphase chemical equilibrium

**Outline of the current chapter**



This chapter focuses on the development of numerical algorithms designed to address the complexities of multiphase chemical equilibrium problems. We begin by exploring various approaches to solving these problems, followed by a series of numerical experiments across a range of test cases. The core objective is to determine the variables $(\xi^\alpha, s_\alpha, r_\alpha)_{\alpha=1,\dots,N_{Ph}}$ that satisfy the system of equations obtained in Sec-

tion 2.3.2:

$$\sum_{\alpha=1}^{N_{Ph}} s_\alpha \mathbf{A}^\alpha \boldsymbol{\xi}^\alpha - b = 0, \tag{2.41a}$$

$$\mathbf{S}^T [\mu_i^\circ/(\mathrm{RT}) + \ln \xi_i]_{i=1,\dots,N} = \mathbf{0}, \tag{2.41b}$$

$$\langle \boldsymbol{\xi}^\alpha, \mathbf{1} \rangle + r_\alpha - 1 = 0, \qquad (\alpha = 1, \dots, N_{Ph}), \tag{2.41c}$$

$$s_\alpha r_\alpha = 0, \qquad (\alpha = 1, \dots, N_{Ph}), \tag{2.41d}$$

$$s_\alpha \geq 0, r_\alpha \geq 0, \qquad (\alpha = 1, \dots, N_{Ph}), \tag{2.41e}$$

where the term $\mu_i^\circ$ is normalized by RT to facilitate numerical stability.

The system presents two primary challenges:

— Nonlinearities of the logarithm: the presence of logarithmic terms introduces stiff nonlinearity, complicating the convergence of iterative methods.

— Complementarity problem: the constraints inherent in the system create a complementarity problem (2.41d)–(2.41e), necessitating specialized techniques for resolution.

In the first section, we recall the parametrization and Cartesian representation techniques introduced in Chapter 3, which were applied to single-phase chemical equilibrium in order to address the issues caused by the logarithmic terms.

The subsequent section delves into classical methods for addressing the complementarity problem. We examine the minimum and Fischer-Burmeister complementarity functions, alongside the interior point method. Additionally, we propose an innovative approach, referred to as complementarity parametrization, that integrates the complementarity equations directly into the system through parametrization techniques. This approach offers a new framework for solving these problems while maintaining essential conditions during iterations.

Finally, we validate our proposed approaches using a simple test case, followed by a challenging test case aimed at assessing the robustness of methods. This challenging test case involves 22 phases, providing a rigorous benchmark for our algorithms's performance. Through these experiments, we aim to demonstrate the efficacy and reliability of our numerical algorithms in solving complex multiphase chemical equilibrium problems. We provide, in particular, evidence of the robustness and efficiency of the complementarity parametrization approach in solving the complementarity conditions for phase presence or disappearance.

## 4.1 Addressing log nonlinearity: notable reminders

In this section, we provide a concise overview of the system utilizing the classical log-trick approach, along with the parametrization and Cartesian representation techniques introduced in the previous chapter. The relevant equations are given by (2.41a)–(2.41c).

### 4.1.1 The log-trick

As for the single-phase case, it could be interesting to test the robustness of the log-trick approach as a reference for comparing the methods. The idea of this approach is to make the following change of variable:

$$y = \ln \xi.$$

This ensures that the mole fractions are always positive and prevents issues with the Jacobian when $\xi$ is small. Applying this change of variable to (2.41a)–(2.41c) yields:

$$\sum_{\alpha=1}^{N_{Ph}} s_\alpha \mathbf{A}^\alpha \mathbf{exp}(\mathbf{y}^\alpha) - b = 0, \tag{4.2a}$$

$$\mathbf{S}^T [\mu_i^\circ/(\mathrm{RT}) + y_i]_{i=1,\dots,N} = \mathbf{0}, \tag{4.2b}$$

$$\langle \mathbf{exp}(\mathbf{y}^\alpha), \mathbf{1} \rangle + r_\alpha - 1 = 0, \qquad (\alpha = 1, \dots, N_{Ph}), \tag{4.2c}$$

where $\mathbf{exp}(\mathbf{y}) = (\exp(y_i))_{i=1,\dots,N}$. The associated block in the Jacobian is written as:

$$\begin{bmatrix} s_1 \mathbf{A}^1 \mathrm{diag}\{\mathbf{exp}(\mathbf{y}^1)\} & \cdots & s_{N_{Ph}} \mathbf{A}^{N_{Ph}} \mathrm{diag}\{\mathbf{exp}(\mathbf{y}^{N_{Ph}})\} & \mathbf{A}\mathbf{exp}(\mathbf{y}) & \mathbf{0} \\ & & \mathbf{S}^T & \mathbf{0} & \mathbf{0} \\ \mathbf{exp}(\mathbf{y}^1)^T & & & & \\ & \ddots & & \mathbf{0} & \mathbf{I}_{N_{Ph}} \\ & & \mathbf{exp}(\mathbf{y}^{N_{Ph}})^T & & \end{bmatrix}.$$

### 4.1.2 Parametrization

The parametrization technique introduces a fictitious variable $\tau_i$ for each species and two functions $X : \mathbb{R} \to \mathbb{R}$ and $Y : \mathbb{R} \to \mathbb{R}$ such that

$$Y(\tau_i) = \ln(X(\tau_i)).$$

These functions satisfy properties (P1)–(P3) and allow us to parameterize the graph

$$\left\{(\xi, y) \in \mathbb{R}^2 \;\middle|\; y = \ln \xi\right\}.$$

There are various possibilities to define these functions, but the choice made here is to use the *switch* function:

$$(X(\tau), Y(\tau)) = \begin{cases} (\exp(\tau), \tau), & \text{if} \quad \tau < 0, \\ (\tau + 1, \ln(\tau + 1)), & \text{if} \quad \tau \geq 0. \end{cases} \tag{3.20}$$

Applying this parametrization to (2.41a)–(2.41c) gives:

$$\sum_{\alpha=1}^{N_{Ph}} s_\alpha \mathbf{A}^\alpha \mathbf{X}(\boldsymbol{\tau}^\alpha) - b = 0, \tag{4.3a}$$

$$\mathbf{S}^T [\mu_i^\circ/(\mathrm{RT}) + Y(\tau_i)]_{i=1,\ldots,N} = \mathbf{0}, \tag{4.3b}$$

$$\langle \mathbf{X}(\boldsymbol{\tau}^\alpha), \mathbf{1} \rangle + r_\alpha - 1 = 0, \qquad (\alpha = 1, \ldots, N_{Ph}). \tag{4.3c}$$

The associated block in the Jacobian is written as:

$$\begin{bmatrix} s_1 \mathbf{A}^1 \mathrm{diag}\{\mathbf{X}'(\boldsymbol{\tau}^1)\} & \cdots & s_{N_{Ph}} \mathbf{A}^{N_{Ph}} \mathrm{diag}\{\mathbf{X}'(\boldsymbol{\tau}^{N_{Ph}})\} & \mathbf{A}\mathbf{X}(\boldsymbol{\tau}) & \mathbf{0} \\ & & \mathbf{S}^T \mathrm{diag}\{Y'(\boldsymbol{\tau})\} & \mathbf{0} & \mathbf{0} \\ X'(\boldsymbol{\tau}^1)^T & & & & \\ & \ddots & & \mathbf{0} & \mathbf{I}_{N_{Ph}} \\ & & X'(\boldsymbol{\tau}^{N_{Ph}})^T & & \end{bmatrix}.$$

### 4.1.3 Cartesian representation

The Cartesian representation technique relaxes the relation $y = \ln(\xi)$ by introducing a new set of unknowns $\mathbf{y} = (\mathbf{y}^\alpha)_{\alpha=1,\ldots,N_{Ph}}$ together with a nonlinear function $f$ such that

$$f(\xi, y) = 0 \Leftrightarrow y = \ln(\xi).$$

This function must satisfy conditions ((C1))–((C3)), and is defined as

$$f(x, y) = \begin{cases} \mathrm{e}^y - x, & \text{if} \quad x \leq 1, y \leq 0, \\ y - x + 1, & \text{if} \quad x \leq 1, y \geq 0, \\ y - \ln x, & \text{if} \quad x \geq 1, y \geq 0, \\ \mathrm{e}^y - \ln x - 1, & \text{if} \quad x \geq 1, y \leq 0. \end{cases} \tag{3.29}$$

Using this Cartesian representation, (2.41a)–(2.41c) becomes:

$$\sum_{\alpha=1}^{N_{Ph}} s_\alpha \mathbf{A}^\alpha \xi^\alpha - b = 0, \tag{4.4a}$$

$$\mathbf{S}^T [\mu_i^\circ/(\mathrm{RT}) + y_i]_{i=1,\dots,N} = \mathbf{0}, \tag{4.4b}$$

$$\langle \xi^\alpha, \mathbf{1} \rangle + r_\alpha - 1 = 0, \qquad (\alpha = 1, \dots, N_{Ph}), \tag{4.4c}$$

$$f(\xi, \mathbf{y}) = \mathbf{0}, \tag{4.4d}$$

where $f(\xi, \mathbf{y}) = (f(\xi_i, y_i))_{i=1,\dots,N}$. The associated block in the Jacobian corresponding to these equations is written as follows:

$$\begin{bmatrix} s_1\mathbf{A}^1 & \cdots & s_{N_{Ph}}\mathbf{A}^{N_{Ph}} & \mathbf{0} & \mathbf{A}\xi & \mathbf{0} \\ & \mathbf{0} & & \mathbf{S}^T & \mathbf{0} & \mathbf{0} \\ \mathbf{1}^T_{\#\sigma^{-1}(1)} & & & & & \\ & \ddots & & \mathbf{0} & \mathbf{0} & \mathbf{I}_{N_{Ph}} \\ & & \mathbf{1}^T_{\#\sigma^{-1}(N_{Ph})} & & & \\ & \nabla_{\mathbf{x}} f & & \nabla_{\mathbf{y}} f & \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where

$$\nabla_{\mathbf{x}} f := \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) = \mathrm{diag}\{\partial_{x_i} f(x_i, y_i)\}_{i=1,\dots,N},$$
$$\nabla_{\mathbf{y}} f := \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = \mathrm{diag}\{\partial_{y_i} f(x_i, y_i)\}_{i=1,\dots,N}.$$

## 4.2 Solving the complementarity problem

The complementarity problem associated with system (2.41) is defined by the following conditions:

$$s_\alpha r_\alpha = 0 \quad \text{and} \quad s_\alpha, r_\alpha \geq 0.$$

This formulation indicates that for each pair of non-negative variables $s_\alpha$ and $r_\alpha$, at least one of them must be zero, which is a characteristic of complementarity conditions. Complementarity problems encompass a broad class of mathematical optimization challenges and have generated extensive literature, as noted in works such as those by Acary and Brogliato [1] and Facchinei and Pang [25, 26]. Graphically, the complementarity problem can be visualized as the intersection of the non-negative axes in a Cartesian coordinate system, specifically represented by the two demi-axes $\{s_\alpha \in \mathbb{R} \mid s_\alpha \geq 0\}$ and $\{r_\alpha \in \mathbb{R} \mid r_\alpha \geq 0\}$. Figure 4.1 highlights the feasible region defined by the complementarity

condition.

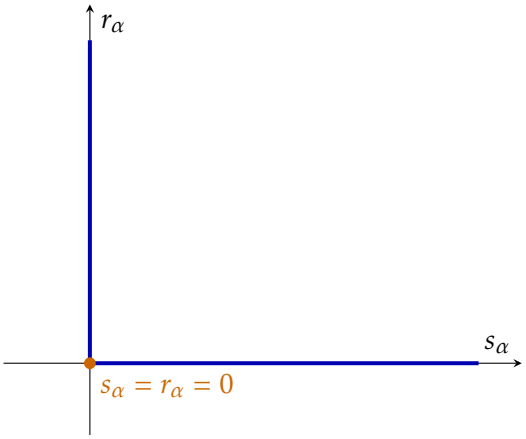


**Figure 4.1 –** The complementarity problem.

The resolution of the complementarity problem using Newton's method presents two significant challenges:

— **Non-differentiability at the origin**: the case where $s_\alpha = r_\alpha$ is not differentiable, which complicates the definition of the Jacobian for the system.

— **Nonnegativity constraint violation**: the nonnegativity constraint is not inherently preserved by the Newton iterates, which can lead to solutions that fall outside the feasible region.

To tackle these issues, classical approaches are typically categorized into two main classes of methods.

The first class consists of semismooth methods, where the complementarity problem is approached using a complementarity function. This function is Lipschitz continuous but may not be differentiable at specific points, particularly when $s_\alpha = r_\alpha$. However, by using the concept of subdifferentials, we can define a subgradient that allows the application of Newton's algorithm effectively. This approach facilitates the handling of non-differentiability while still leveraging the advantages of Newton's method. In Section 4.2.1, we will present the minimum and Fischer-Burmeister complementarity functions, which are commonly used in semismooth approaches to effectively handle the complementarity conditions.

The second class focuses on smoothing techniques for the complementarity equations. In this approach, a regularization parameter is introduced to create a smooth approximation of the complementarity problem. This enables the use of classical Newton's method on the modified problem. As the iterations progress, the regularization

parameter is gradually reduced to zero, guiding the solution back to the original complementarity problem as the iterates converge towards the solution. In Section 4.2.2, we will explore the interior point method (IPM), a well-established technique for addressing the smoothing of complementarity equations.

In addition to these established methods, we propose in Section 4.2.3 a novel approach to the complementarity problem based on parametrization techniques. In this method, referred to as complementarity parametrization method, the complementarity equations are integrated into the system, and the relationships between the complementarity variables are maintained through a parameter and two parametric functions. This innovative approach provides a new framework for solving complementarity problems while ensuring that the essential conditions are met throughout the iterative process.

### 4.2.1 Semismooth methods

The treatment of the complementarity problem using semismooth methods involves the use of a complementarity function $\Psi : \mathbb{R}^2 \to \mathbb{R}$ that satisfies specific conditions. Different functions lead to distinct smoothness and regularity characteristics, which are crucial for the effectiveness of the numerical techniques applied to solve nonlinear complementarity problems. To be a complementarity function, $\Psi$ must satisfy the following condition:

$$\Psi(s_\alpha, r_\alpha) = 0 \quad \Leftrightarrow \quad s_\alpha r_\alpha = 0 \text{ and } s_\alpha \geq 0, r_\alpha \geq 0.$$

There are several options available for the function $\Psi$, but our study will concentrate on the following two functions:

— **The min function** [52]:

$$\Psi_{\min}(s_\alpha, r_\alpha) = \min(s_\alpha, r_\alpha),$$

which is Lipschitz continuous but lacks differentibility at the points where $s_\alpha = r_\alpha$. This function is depicted in Figure 4.2.

— **The Fischer-Burmeister function** [27]:

$$\Psi_{\text{FB}}(s_\alpha, r_\alpha) = s_\alpha + r_\alpha - \sqrt{s_\alpha^2 + r_\alpha^2},$$

which is differentiable everywhere except at the point where $s_\alpha = r_\alpha = 0$. This function is illustrated in Figure 4.3.

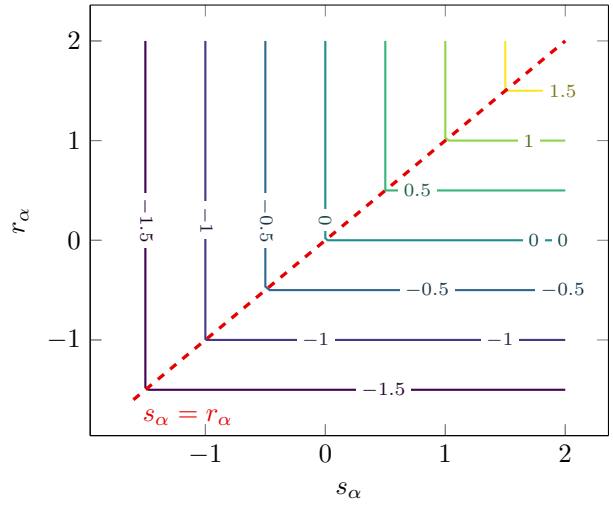


**Figure 4.2 –** Contour plot of the function $\Psi_{\text{min}}$.

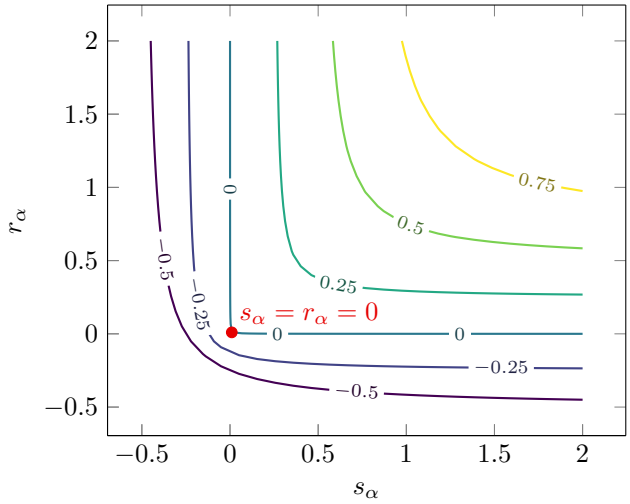


**Figure 4.3 –** Contour plot of the function $\Psi_{\text{FB}}$.

The non-differentiality of the min and Fischer-Burmeister (FB) functions leads to the use of semismooth Newton's method to solve the complementarity problem. In this approach, the Jacobian is replaced by an element of the Clarke subdifferential of the function for which we seek the root.

Let $\mathcal{F} : \boldsymbol{\mathcal{X}} \in \mathcal{D}_{\mathcal{F}} \to \mathbb{R}$ be a locally Lipschitz continuous function defined in a set $\mathcal{D}_{\mathcal{F}}$. By Rademacher's theorem [16, Section 3.4.1], such a function is continuously differentiable almost everywhere. Let $\Omega_{\text{F}}$ be the set where $\mathcal{F}$ is Fréchet differentiable, *i.e.*, where $\boldsymbol{\nabla}\mathcal{F}$ exists. The Bouligand subdifferential $\partial_{\text{B}}\mathcal{F}(\boldsymbol{\mathcal{X}})$ of $\mathcal{F}$ at $\boldsymbol{\mathcal{X}} \in \Omega_{\text{F}}$ is defined

as:

$$\partial_{\mathrm{B}}\mathcal{F}(\boldsymbol{\mathcal{X}}) := \left\{ \mathcal{J} \;\middle|\; \exists (\boldsymbol{\mathcal{X}}^{(k)})_{k\in\mathbb{N}} \subset \Omega_{\mathrm{F}},\ \boldsymbol{\mathcal{X}}^{(k)} \to \boldsymbol{\mathcal{X}},\ \boldsymbol{\nabla}\mathcal{F}(\boldsymbol{\mathcal{X}}^{(k)}) \to \mathcal{J} \right\}.$$

The Clarke subdifferential [14] is then given by:

$$\partial_{\mathrm{C}}\mathcal{F}(\boldsymbol{\mathcal{X}}) = \mathrm{conv}\left(\partial_{\mathrm{B}}\mathcal{F}(\boldsymbol{\mathcal{X}})\right),$$

where conv denotes the convex hull.

As a simple example, consider the absolute value function $f(x) = |x|$ at $x = 0$. In this case, $\partial_{\mathrm{B}} f(0) = \{-1, 1\}$ and $\partial_{\mathrm{C}} f(0) = [-1, 1]$.

The semismooth Newton's algorithm then start from an initial value $\boldsymbol{\mathcal{X}}^{(0)}$, and builds a sequence $(\boldsymbol{\mathcal{X}}^{(k)})_{k>0}$ by solving the linear system [55]:

$$\mathcal{J}\delta\boldsymbol{\mathcal{X}}^{(k)} = -\mathcal{F}\left(\boldsymbol{\mathcal{X}}^{(k)}\right), \quad \mathcal{J} \in \partial_{\mathrm{C}}\mathcal{F}\left(\boldsymbol{\mathcal{X}}^{(k)}\right),$$

and updating the sequence as:

$$\boldsymbol{\mathcal{X}}^{(k+1)} = \boldsymbol{\mathcal{X}}^{(k)} + \delta\boldsymbol{\mathcal{X}}^{(k)}.$$

The selection of $\mathcal{J} \in \partial_{\mathrm{C}}\mathcal{F}\left(\boldsymbol{\mathcal{X}}^{(k)}\right)$ when the latter is multivalued can have an impact on the performance of the method. Some authors [31] suggest choosing an element $\mathcal{J}$ from the Bouligand subdifferential $\partial_{\mathrm{B}}\mathcal{F}\left(\boldsymbol{\mathcal{X}}^{(k)}\right)$ instead. In our case, the corresponding Bouligand subdifferentials for the min and FB functions are:

$$\begin{aligned} \partial_{\mathrm{B}}\Psi_{\min}(s_\alpha, r_\alpha)|_{s_\alpha = r_\alpha} &= \{(1,\ 0)^T, (1,\ 0)^T\} \\ \partial_{\mathrm{B}}\Psi_{\mathrm{FB}}(0, 0) &= \{(1, 1)^T - \mathbf{u} \mid \|\mathbf{u}\|_2 = 1\}. \end{aligned}$$

In our algorithms, the chosen subdifferentials are:

$$\begin{aligned} \partial_{\mathrm{B}}\Psi_{\min}(s_\alpha, r_\alpha)|_{s_\alpha = r_\alpha} &= (1, 0)^T \\ \partial_{\mathrm{B}}\Psi_{\mathrm{FB}}(0, 0) &= (1 - \sqrt{2}/2,\ 1 - \sqrt{2}/2)^T. \end{aligned}$$

However, these specific cases almost never occur in our numerical experiments.

### 4.2.2 Smoothing methods

The central concept of smoothing techniques involves the introduction of a new parameter, denoted as $\nu \in \mathbb{R}$, known as the regularization parameter, into the complementarity equation. This parameter enables the "smoothing" of the equation, rendering

it differentiable and thus suitable for classical optimization methods, such as the Newton method. In this context, we will focus on a specific class of smoothing methods that are part of interior point techniques [30]. Here, the complementarity equation (2.41d) is regularized as follows:

$$s_\alpha r_\alpha = \nu.$$

The system is then transformed into a new, smooth equation given by:

$$\mathcal{F}(\boldsymbol{\mathcal{X}}; \nu) = \mathbf{0}. \tag{4.5}$$

The smooth equation can be effectively solved using the Newton method. During the Newton iterations, the regularization parameter is gradually decreased, allowing the smooth equation to converge toward the original complementarity equation. As the iterations progress and $\nu$ approaches zero, the solution of the smooth equation converges to that of the original complementarity equation. The Newton step is defined by solving:

$$\boldsymbol{\nabla}\mathcal{F}(\boldsymbol{\mathcal{X}}^{(k)};\, \nu^{(k)})\boldsymbol{\delta\mathcal{X}}^{(k)} = -\mathcal{F}(\boldsymbol{\mathcal{X}}^{(k)}; \nu^{(k)}).$$

Subsequently, $\nu$ is updated according to a predefined sequence:

$$\nu^{(k+1)} = \Theta(\nu^{(k)}),$$

which converges toward zero. In our study, we consider the following strategy for reducing $\nu^{(k)}$ [76]:

$$\nu^{(k+1)} = \max\left(10^{-32}, \min\left(\kappa_\nu \nu^{(k)}, \left(\nu^{(k)}\right)^{\theta_\nu}\right)\right),$$

with $\kappa_\nu \in (0, 1)$ and $\theta_\nu \in (1, 2)$. At this stage, the complementarity equation is not fully satisfied, as we must ensure $s_\alpha, r_\alpha \geq 0$. While several strategies exist to maintain the positivity of these iterates, we present a method that treats $s_\alpha$ and $r_\alpha$ individually. The update for $\mathbf{s} = (s_\alpha)_{\alpha=1,\dots,N_{Ph}}$ and $\mathbf{r} = (r_\alpha)_{\alpha=1,\dots,N_{Ph}}$ are as follows:

$$\mathbf{s}^{(k+1)} = \mathbf{s}^{(k)} + \beta_{\mathbf{s}}^{(k)} \boldsymbol{\delta}\mathbf{s}^{(k)},$$
$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} + \beta_{\mathbf{r}}^{(k)} \boldsymbol{\delta}\mathbf{r}^{(k)}.$$

Following the approach outlined in [76], $\beta_{\mathbf{u}}$ for $\mathbf{u} \in \{\mathbf{s}, \mathbf{r}\}$ is defined as:

$$\beta_{\mathbf{v}} = \max\{\beta \in (0, 1] \,|\, \mathbf{u} + \beta\boldsymbol{\delta}\mathbf{u} \geq (1 - \gamma^{(k)})\mathbf{u}\},$$

where $\gamma^{(k)}$ is referred to the fraction-to-the-boundary parameter in [76]:

$$\gamma^{(k)} = \max(\gamma_{\min}, 1 - \nu^{(k)}),$$

with $\gamma_{\min} \in (0, 1)$. We will refer to this update as the IPOPT line search in the following sections.

Let us also mention alternative method where $\nu$ is considered as an unknowns. This approach is named nonparametric IPM (NPIPM) and has been developed by Vu *et al.* [73, 74].

### 4.2.3 The complementarity parametrization method

Similarly to our approach of parametrization for the relationship $y = \ln \xi$, the graph

$$\mathcal{T} = \{(s, r) \in \mathbb{R}^2 \mid \min(s, r) = 0\}$$

is parameterized by two monotonic piecewise continuously differentiable functions, $S : \mathbb{R} \to \mathbb{R}$ and $R : \mathbb{R} \to \mathbb{R}$, such that

$$s_\alpha = S(\eta_\alpha) \quad \text{and} \quad r_\alpha = R(\eta_\alpha).$$

These functions are defined to satisfy:

$$\min(S(\eta), R(\eta)) = 0,$$

which implies that $\mathcal{T} = (S, R)(\mathbb{R})$. To ensure proper parametrization and avoid singularities, the functions $S$ and $R$ must satisfy the following conditions, there exists $M > 0$, $\varepsilon > 0$ such that for each $\eta \in \mathbb{R}$:

(PC1) $S(\eta)R(\eta) = 0$ and $S(\eta), R(\eta) \geq 0$ with $S$ and $R$ monotonic and $M$-Lipschitz continuous;

(PC2) the derivatives $S'$ and $R'$ are bounded and piecewise continuous with $0 \leq S'(\eta) \leq M$ and $-M \leq R'(\eta) \leq 0$;

(PC3) $S'(\eta) - R'(\eta) > \varepsilon$.

We define a parametrization as admissible if it fulfills conditions (PC1)–(PC3). To define these specific functions, we require that $S$ and $R$ are both monotonic and $M$-Lipschitz continuous, which implies that they are differentiable almost everywhere, as established by Rademacher's theorem. For the points at which $S$ and $R$ are not differentiable, we

assign specific values to $S'$ and $R'$. In practice, $S'$ and $R'$ exhibit piecewise continuity. Furthermore, we ensure that $|S'|$ and $|R'|$ are upper semi-continuous by choosing the largest possible value (in magnitude) for both $S'$ and $R'$. We also impose a normalization condition:

$$\max(|S'(\eta)|, |R'(\eta)|) = 1, \quad \forall \eta \in \mathbb{R}.$$

Condition (PC1)–(PC3) implies the existence of a point $\eta^0 \in \mathbb{R}$ such that $S(\eta^0) = R(\eta^0) = 0$. By selecting $\eta^0 = 0$ and enforcing $S' \geq 0$, $R' \leq 0$, we derive the following implications:

$$\begin{aligned} \eta > 0 \quad &\Rightarrow \quad S(\eta) > 0 \text{ and } R(\eta) = 0 \quad \Rightarrow \quad R'(\eta) = 0, \\ \eta < 0 \quad &\Rightarrow \quad S(\eta) = 0 \text{ and } R(\eta) > 0 \quad \Rightarrow \quad S'(\eta) = 0. \end{aligned}$$

From the normalization condition, we have:

$$\max(|S'(\eta)|, |R'(\eta)|) = 1 \quad \Leftrightarrow \quad \max(S'(\eta), -R'(\eta)) = 1$$

This leads to the following conclusions:

$$\begin{aligned} \eta > 0 \quad &\Rightarrow \quad S'(\eta) = 1 \quad \Rightarrow \quad S(\eta) = \eta, \\ \eta < 0 \quad &\Rightarrow \quad R'(\eta) = -1 \quad \Rightarrow \quad R(\eta) = -\eta. \end{aligned}$$

Thus, we define the complementarity parametrization method as follows:

$$S(\eta) = \begin{cases} 0 & \text{if } \eta < 0, \\ \eta & \text{if } \eta \geq 0, \end{cases} \quad \text{and} \quad R(\eta) = \begin{cases} -\eta & \text{if } \eta \leq 0, \\ 0 & \text{if } \eta > 0. \end{cases}$$

These functions are illustrated in Figure 4.4. Furthermore, the derivatives of these functions are defined as:

$$S'(\eta) = \begin{cases} 0 & \text{if } \eta < 0, \\ 1 & \text{if } \eta \geq 0, \end{cases} \quad \text{and} \quad R'(\eta) = \begin{cases} -1 & \text{if } \eta \leq 0, \\ 0 & \text{if } \eta > 0. \end{cases}$$

At the point $\eta = 0$, we have chosen to set both $|S'(0)|$ and $|R'(0)|$ equal to 1 to ensure the invertibility of the Jacobian matrix.

These functions lead to the a reformulation of system (2.41) which becomes: find $(\boldsymbol{\xi}^\alpha, \eta_\alpha)_{\alpha=1,\dots,N_{Ph}}$ such that:

$$\sum_{\alpha=1}^{N_{Ph}} S(\eta_\alpha) \mathbf{A}^\alpha \boldsymbol{\xi}^\alpha - b = 0,$$

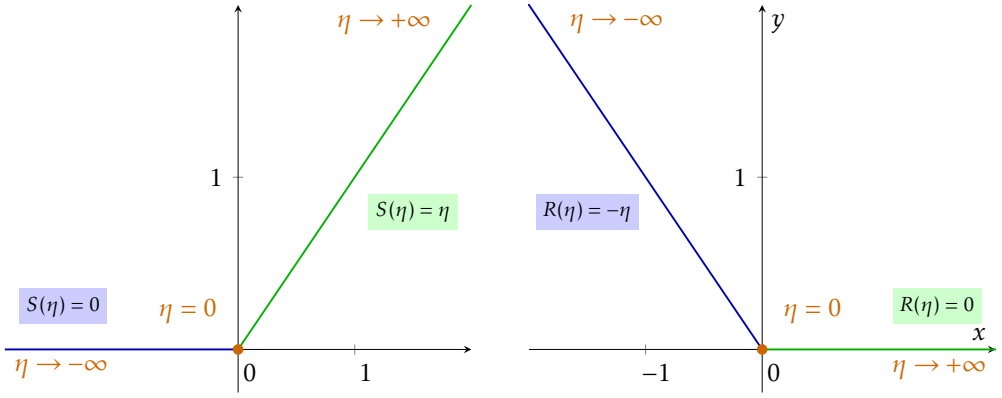


**Figure 4.4 –** Functions $S$ and $R$ for the complementarity parametrization method.

$$\begin{aligned} \mathbf{S}^T[\mu_i^\circ/(\mathrm{RT}) + \ln \xi_i]_{i=1,\dots,N} &= \mathbf{0}, \\ \langle \boldsymbol{\xi}^\alpha, \mathbf{1} \rangle + R(\eta_\alpha) - 1 &= 0, \quad (\alpha = 1, \dots, N_{Ph}). \end{aligned}$$

For each of the method to treat the log nonlinearity, the Jacobian are detailed in the following. The Jacobian of the log-trick is:

$$\begin{bmatrix} S(\eta_1)\mathbf{A}^1 \mathrm{diag}\{\mathbf{exp}(\mathbf{y}^1)\} & \cdots & S(\eta_{N_{Ph}})\mathbf{A}^{N_{Ph}} \mathrm{diag}\{\mathbf{exp}(\mathbf{y}^{N_{Ph}})\} & S'(\eta_1)\mathbf{A}^1\mathbf{exp}(\mathbf{y}^1) & \cdots & S'(\eta_{N_{Ph}})\mathbf{A}^{N_{Ph}}\mathbf{exp}(\mathbf{y}^{N_{Ph}}) \\ & & \mathbf{S}^T & & & \mathbf{0} \\ \mathbf{exp}(\mathbf{y}^1)^T & & & R'(\eta_1) & & \\ & \ddots & & & \ddots & \\ & & \mathbf{exp}(\mathbf{y}^{N_{Ph}})^T & & & R'(\eta_{N_{Ph}}) \end{bmatrix}.$$

The Jacobian of the parametrization is:

$$\begin{bmatrix} S(\eta_1)\mathbf{A}^1 \mathrm{diag}\{\mathbf{X}'(\boldsymbol{\tau}^1)\} & \cdots & S(\eta_{N_{Ph}})\mathbf{A}^{N_{Ph}} \mathrm{diag}\{\mathbf{X}'(\boldsymbol{\tau}^{N_{Ph}})\} & S'(\eta_1)\mathbf{A}^1\mathbf{X}(\boldsymbol{\tau}^1) & \cdots & S'(\eta_{N_{Ph}})\mathbf{A}^{N_{Ph}}\mathbf{X}(\boldsymbol{\tau}^{N_{Ph}}) \\ & & \mathbf{S}^T \mathrm{diag}\{Y'(\boldsymbol{\tau})\} & & & \mathbf{0} \\ X'(\boldsymbol{\tau}^1)^T & & & R'(\eta_1) & & \\ & \ddots & & & \ddots & \\ & & X'(\boldsymbol{\tau}^{N_{Ph}})^T & & & R'(\eta_{N_{Ph}}) \end{bmatrix}.$$

The Jacobian of the Cartesian representation is:

$$\begin{bmatrix} S(\eta_1)\mathbf{A}^1 & \cdots & S(\eta_{N_{Ph}})\mathbf{A}^{N_{Ph}} & \mathbf{0} & S'(\eta_1)\mathbf{A}^1\xi^1 & \cdots & S'(\eta_{N_{Ph}})\mathbf{A}^{N_{Ph}}\xi^{N_{Ph}} \\ & \mathbf{0} & & \mathbf{S}^T & & \mathbf{0} & \\ \mathbf{1}^T_{\#\sigma^{-1}(1)} & & & & R'(\eta_1) & & \\ & \ddots & & \mathbf{0} & & \ddots & \\ & & \mathbf{1}^T_{\#\sigma^{-1}(N_{Ph})} & & & & R'(\eta_{N_{Ph}}) \\ & \nabla_{\mathbf{x}} f & & \nabla_{\mathbf{y}} f & & \mathbf{0} & \end{bmatrix}.$$

## 4.3 Numerical experiments

In this section, we will compare the three formulations for addressing the log problem with each of the four methods for tackling the complementarity problem. This analysis will initially focus on a straightforward test case involving an aqueous phase and a mineral phase under various presence/absence configurations. Subsequently, we will examine the limitations of these approaches using a more complex test case that includes 22 phases.

### 4.3.1 Numerical parameters

The functions $X$ and $Y$ for the parametrization are those of the switch defined in (3.20). For the Cartesian representation technique, the function $f$ is the discrepancy function defined in (3.29). The parameters used for IPM are $\kappa_\nu = 0.5, \theta_\nu = 2$ and $\gamma_{\min} = 0.1$. If $\mathcal{F}(\boldsymbol{\mathcal{X}})$ represents the function for which we are seeking the root, the convergence criterion for Newton's algorithm is

$$\|\mathcal{F}(\boldsymbol{\mathcal{X}}^{(k+1)})\|_\infty \leq 10^{-10} \quad \text{and} \quad \|\boldsymbol{\mathcal{X}}^{(k+1)} - \boldsymbol{\mathcal{X}}^{(k)}\|_\infty \leq 10^{-10}$$

where $k+1$ is the current Newton iteration.

However, in the context of IPM, the variables $s_\alpha$ and $r_\alpha$ are never exactly zero and are always positive. This presents challenges when both variables must equal zero at equilibrium. Consequently, we must impose stricter precision requirements on the residuals associated with complementarity. The tolerance on $\mathcal{F}$ then becomes:

$$\|\mathcal{F}(\boldsymbol{\mathcal{X}}^{(k)})\|_\infty \leq 10^{-10} \quad \text{and} \quad \|\mathcal{F}_{\text{compl}}(\boldsymbol{\mathcal{X}}^{(k)})\|_\infty \leq \epsilon_{64}^2,$$

where $\epsilon_{64} \approx 2.22 \times 10^{-16}$ is the epsilon machine in double precision.

The way to initialize Newton's method for the different methods is as follows. We first start with vectors of quantities defined as:

$$n_{aq} = (\mathbf{b}_{\mathrm{O}}, 1, \ldots, 1), \quad n_{min} = 1, \text{ for all mineral}, \quad n_{gas} = (1, \ldots, 1),$$

where $\mathbf{b}_{\mathrm{O}}$ is the fixed quantities of oxygen. We then defined the total quantities of each phase as:

$$s_{aq} = \langle n_{aq}, \mathbf{1} \rangle, \quad s_{min} = 1, \text{ for all mineral}, \quad s_{gas} = \langle n_{gas}, \mathbf{1} \rangle.$$

This leads to the definition of the vector $\xi$ as:

$$\xi_{aq} = n_{aq}/s_{aq}, \quad \xi_{min} = 1, \text{ for all mineral}, \quad \xi_{gas} = n_{gas}/s_{gas}.$$

The vector $\mathbf{r}$ is initialized as:

$$r_{aq} = 1, \quad r_{min} = 1, \text{ for all mineral}, \quad r_{gas} = 1.$$

For the complementarity parametrization method, the vector $\eta$ is initialized as:

$$\eta_{\mathrm{aq}} = s_{\mathrm{aq}}, \qquad \eta_{min} = 1, \text{ for all mineral}, \quad \eta_{gas} = 1.$$

The variable $\mathbf{y}$ for the Cartesian representation is initialized as in the single phase case:

$$\mathbf{y} = \xi - 1.$$

Finally, the parameter $\nu$ for IPM is defined as:

$$\nu = \max\left\{ \min_{\alpha \in \{1,\ldots,N_{ph}\}} (s_\alpha r_\alpha) - 0.1, 0.5 \right\}$$

### 4.3.2 Test cases presentation

Our first test case, referred to as the $SiO_2$ test case, is composed of 4 species in an aqueous phase:

$$H_2O, H^+, SiO_2(aq), OH^-,$$

and 1 mineral: the Quartz ($SiO_2(s)$).

The more challenging test case, referred to as *multiphase Seawater*, is composed of 50 species in a aqueous phase, 20 minerals in 20 pure phases and 2 species in an gaseous phase.

The details of these two test cases are in Appendix A.3. Both use the charge constraint defined in section 2.2.4.

### 4.3.3 Numerical results

#### SiO2 test case

For this test case, we have run our algorithms with three different vectors **b** in order to obtain the following subcases:

— **A**: the mineral is present in a large amount;

— **B**: the mineral is absent with condition that are close to precipitation, indicating that both variables of the complementarity conditions vanish at equilibrium;

— **C**: the mineral is absent with conditions that are far from precipitation.

The three vectors **b** are detailed in Appendix A.3. The results obtained from the various combinations of methods are presented in Table 4.1. The evolution of residuals for setting **A** is illustrated in Figure 4.6 and the evolution of IPM in setting **B** in Figure 4.5. Notably, for the methods that converge, local quadratic convergence is observed. These results enable us to draw the following conclusions:

1. The complementarity parametrization method is the faster method for the log-trick and parametrization formulations.
2. The Cartesian representation techniques diverges except for the IPM.
3. In **B**, the IPM is very slow, due to the needed precision on the complementarity equations and to a linear convergence.
4. In **C**, the min function diverge with the parametrization formulation and the FB function is slow with the log-trick and parametrization formulations.
5. The complementarity parametrization method and the IPM approaches are robust except for the Cartesian representation with the parametrization. However, the convergence of the IPM is slower.

In the remainder of this section, we will conduct a comprehensive analysis of the diverse behaviors exhibited by algorithms.

| Log formulation | Complementarity | Setting | | |
|---|---|---|---|---|
| | | A | B | C |
| Log-trick | param | 7 | 7 | 10 |
| | min | 9 | 7 | 10 |
| | FB | 8 | 7 | 15 |
| | IPM | 9 | 52 | 12 |
| Param | param | 5 | 5 | 10 |
| | min | 6 | 6 | × |
| | FB | 7 | 6 | 14 |
| | IPM | 9 | 52 | 12 |
| Cart. repr. | param | × | × | × |
| | min | × | × | × |
| | FB | × | × | × |
| | IPM | 9 | 49 | 13 |

**Table 4.1 –** Number of iterations for Newton's method.

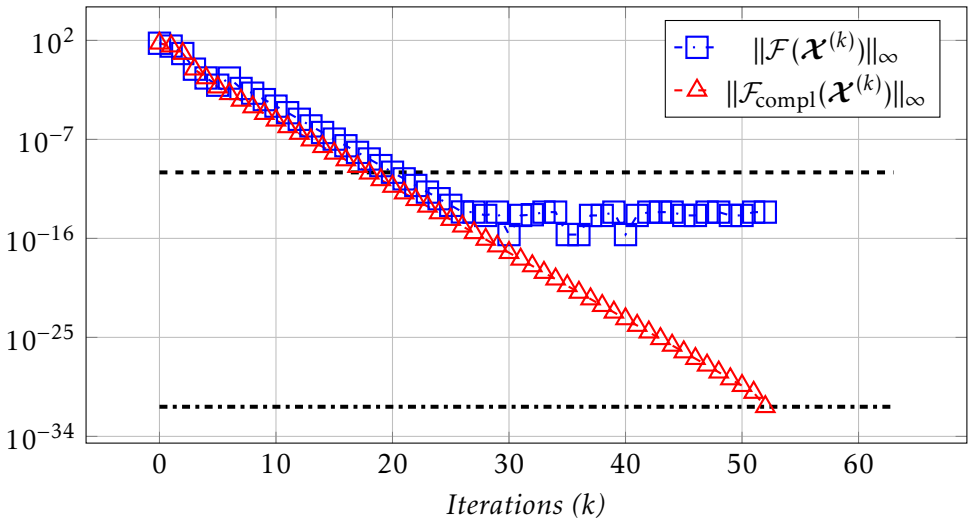


**Figure 4.5 –** Evolution of residuals of IPM for setting **B** with the parametrization formulation.

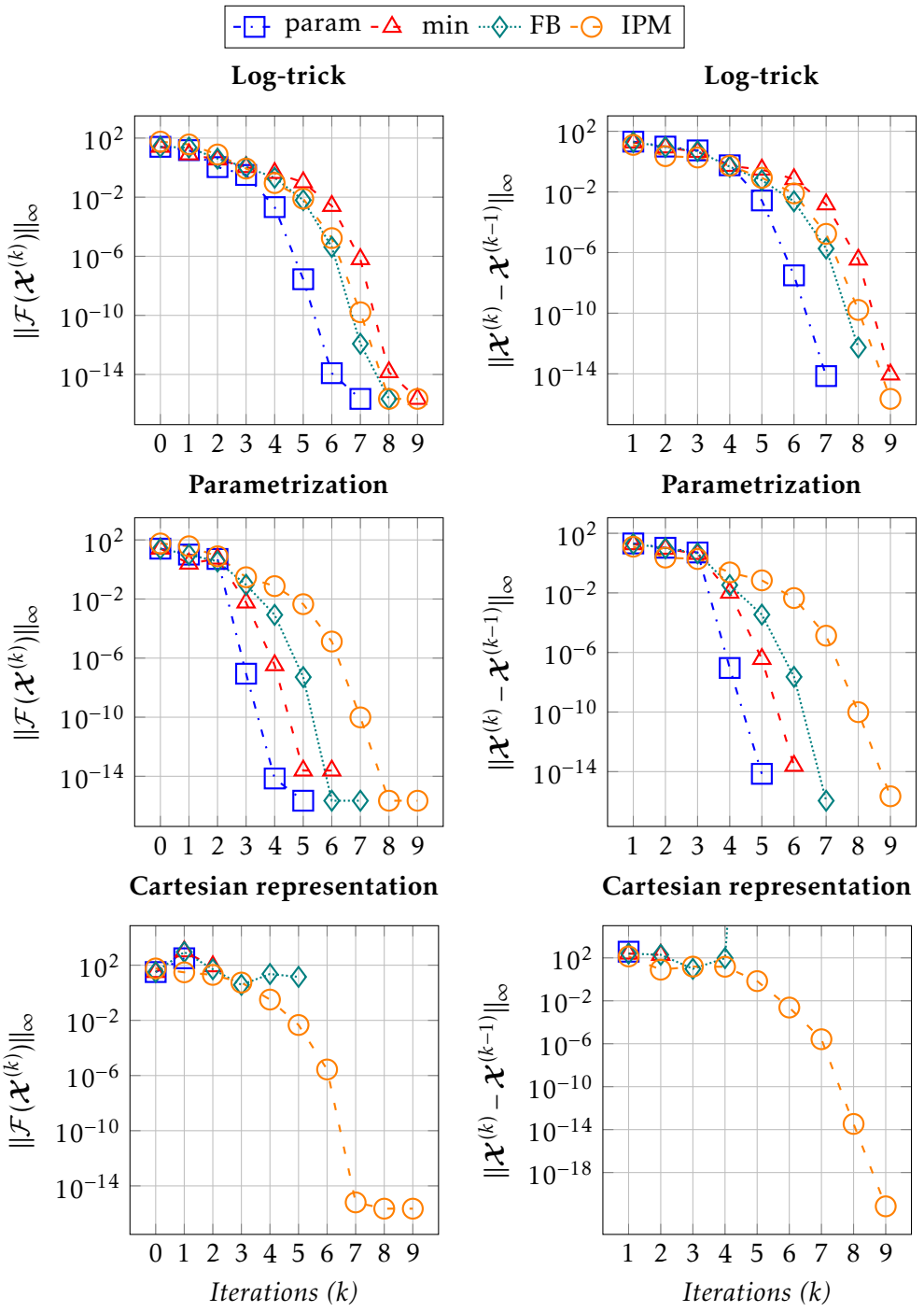


**Figure 4.6 –** Evolution of residuals for the $SiO_2$ text case in configuration **A**.

**Comparison between the log-trick and the parametrization with complementarity parametrization**

The difference in convergence speed between the parametrization method and the log-trick arises from the same mechanisms discussed in Section 3.3 for the single-phase case. Figure 4.7 illustrates the evolution of the norm of the function $\mathcal{F}$, which is restricted to the conservation equations and sum of mole fractions equations on the left graph, and to the equilibrium equations on the right graph. Meanwhile, Figure 4.8 depicts the evolution of iterates for the species $H_2O$ and its mole fraction. Table 4.2 indicates the parametrization function used at each iteration.

As shown in Figure 4.7, the decrease in residual can be attributed to the complete resolution of the chemical equilibrium equation from the third iteration onward. Specifically, from the second iteration onward, the equilibrium equation becomes linear, and it is well-known that Newton's method can solve linear equations exactly. Similar to the single-phase case, the $\tau_i$ variables, excluding $H_2O$, are negative, which ensures that the $Y(\tau_i)$ values remain linear. Consequently, the linearity or non-linearity of the equilibrium equations is determined solely by the value of $\tau_{H_2O}$. Figures 3.5 and Table 4.2 show that this value is negative starting from the second iteration, resulting in a linear equilibrium equation.

It is noteworthy from Figure 4.8 that during the first iteration, the $\tau_{H_2O}$ from the parametrization and the $\ln(x_{H_2O})$ from the log-trick are equal. However, the corresponding mole fractions $X(\tau_{H_2O})$ and $x_{H_2O}$ differ because $X(\tau)$ is linear for positive $\tau$ values. This characteristic prevents the mole fraction from being excessively high, leading to a $\tau_{H_2O}$ value that is much closer to the solution by the second iteration, unlike the log-trick. Thus, the parametrization method accelerated convergence towards the solution in this test case.

| Iteration | 0 | 1 | 2 | $\cdots$ | 5 |
|---|---|---|---|---|---|
| $\text{sign}(\tau_{H_2O})$ | $-$ | $+$ | $-$ | $\cdots$ | $-$ |
| $X(\tau_{H_2O})$ | $\exp \tau_{H_2O}$ | $\tau_{H_2O} + 1$ | $\exp \tau_{H_2O}$ | $\cdots$ | $\exp \tau_{H_2O}$ |
| $Y(\tau_{H_2O})$ | $\tau_{H_2O}$ | $\ln \tau_{H_2O}$ | $\tau_{H_2O}$ | $\cdots$ | $\tau_{H_2O}$ |

**Table 4.2 –** Evolution of $X(\tau_{H_2O})$ and $Y(\tau_{H_2O})$.

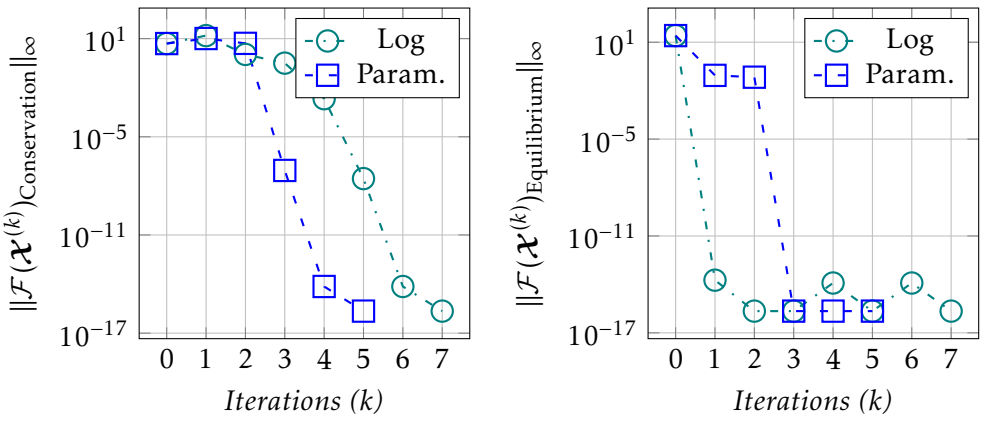


**Figure 4.7 –** Evolution of residuals of the log-trick and the parametrization for the $SiO_2$ test case. The left graph represents the residuals of the conservation equations while the right one represents the residuals of the equilibrium equations.

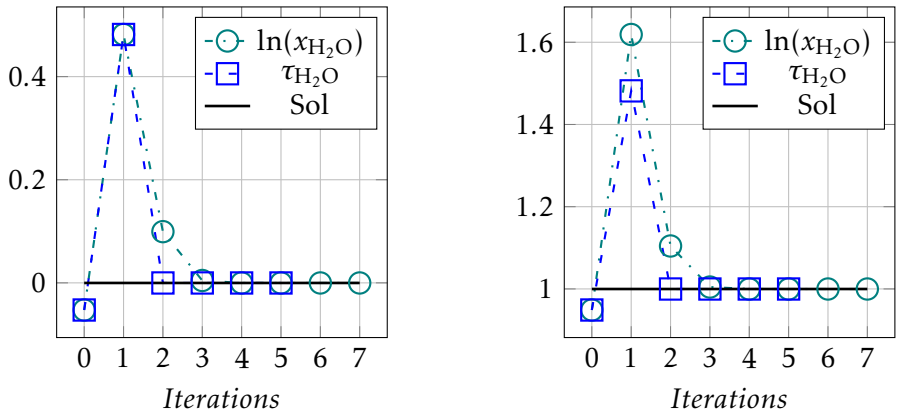


**Figure 4.8 –** Evolution of the species $H_2O$ with the log-trick and parametrization methods.

**Superiority of the parametrization formulation with complementarity parametrization method**

The Superiority of the parametrization log formulation combined with complementarity parametrization method in Table 4.1 is due to a more profitable initialization. Indeed, the nature of complementarity parametrization prevent to have the same sign for $S(\eta)$ and $R(\eta)$. In settings **A** and **B**, there is no switch in the functions $S$ and $R$ and in particular, $R(\eta_{\text{Quart}})$ is always zero. This accelerates the computation compared to the others approaches which start from a positive $r_{\text{Quart}}$. We can observe these evolution in Figure 4.9 and Figure 4.10. The switch of the complementarity parametrization methods

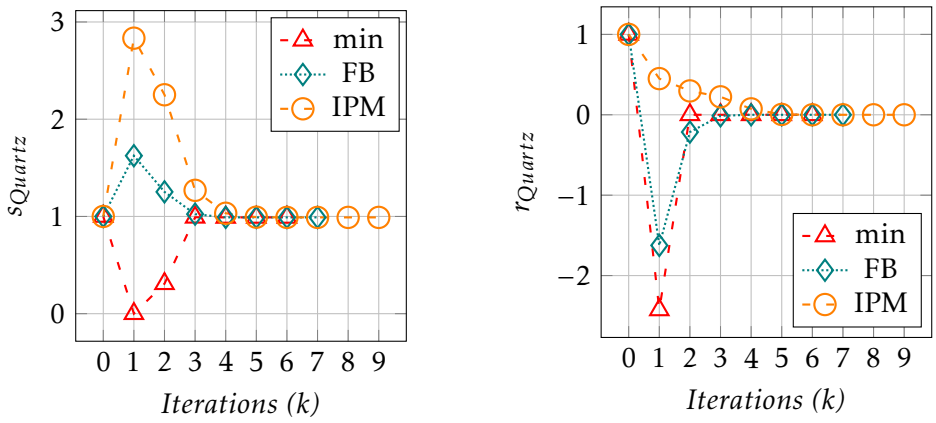


**Figure 4.9 –** Evolution of $s_{Quartz}$ and $r_{Quartz}$ for the parametrization formulation combined with various complementarity approaches in setting **A**.

for setting **C** are depicted in Figure 4.11 and Table 4.3. We observe that from iteration 2 onward, there is no change is the sign of $\eta_{\text{Quartz}}$.

The superiority of the parametrization formulation combined with the complementarity parametrization method, as shown in Table 4.1, can be attributed to a more effective initialization. The inherent nature of the complementarity parametrization prevents both $S(\eta)$ and $R(\eta)$ from having the same sign. In settings **A** and **B**, there is no switch inside the functions $S$ and $R$; notably, $R(\eta_{\text{Quartz}})$ remains consistently zero. This characteristic accelerates computations compared to other approaches that initiate with a positive $r_{\text{Quart}}$.

We can observe these evolutions in Figures 4.9 and 4.10. The switching behavior of the complementarity parametrization methods for setting **C** is illustrated in Figure 4.11 and detailed in Table 4.3. Notably, from iteration 2 onward, there is no change in the sign of $\eta_{\text{Quartz}}$.

| Iteration | 0 | 1 | 2 | 3 | ⋯ | 10 |
|---|---|---|---|---|---|---|
| sign($\eta_{\text{Quartz}}$) | + | + | − | − | ⋯ | − |

**Table 4.3 –** Evolution of the sign of $\eta_{\text{Quartz}}$ in setting **C**.

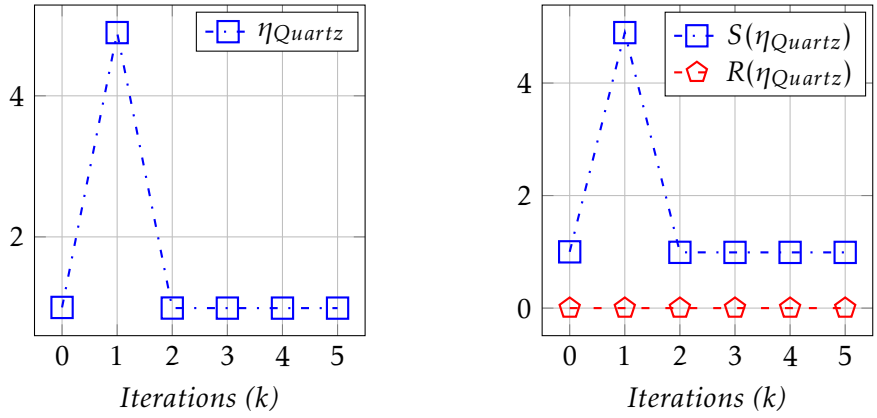


**Figure 4.10 –** Evolution of $\eta_{Quartz}$ and the associated functions $S(\eta_{Quartz})$ and $R(\eta_{Quartz})$ for the parametrization formulation combined with the complementarity parametrization method in setting **A**.

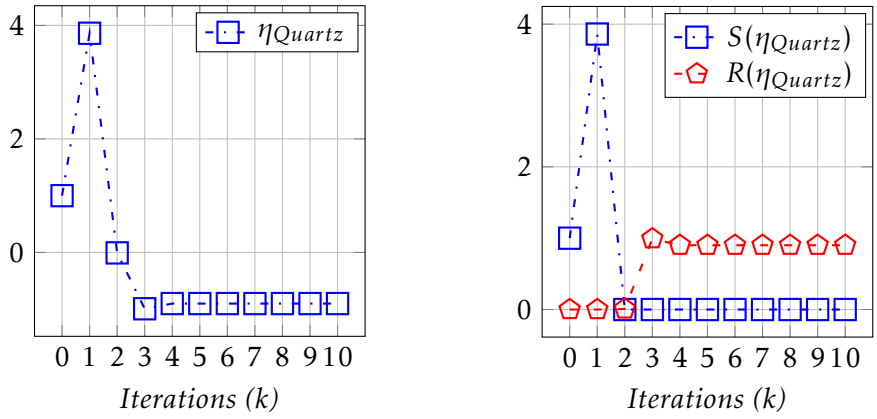


**Figure 4.11 –** Evolution of $\eta_{Quartz}$ and the associated functions $S(\eta_{Quartz})$ and $R(\eta_{Quartz})$ for the parametrization formulation combined with the complementarity parametrization method in setting **C**.

### Failure of the Cartesian representation

The equation for the conservation of the charge Z in the Cartesian representation formulation is given by

$$s_{aq}\xi_{H^+} - s_{aq}\xi_{OH^-} - \mathbf{b}_Z = 0. \tag{4.6}$$

This equations leads to the following row in the Jacobian matrix:

$$\begin{pmatrix} 0 & s_{aq} & 0 & s_{aq} & 0 & \mathbf{0}_5 & \xi_{H^+} - \xi_{OH^-} & 0 & 0 & 0 \end{pmatrix}, \tag{4.7}$$

where $\mathbf{0}_5$ is the zero of $\mathbb{R}^5$. We can observe in Figure 4.12 that at iteration 2, $s_{\text{aq}} = 0$ and $\xi_{\text{H}^+} = \xi_{\text{OH}^-}$, leading to a row of zero in (4.7). Other cases of failure are of the same nature.

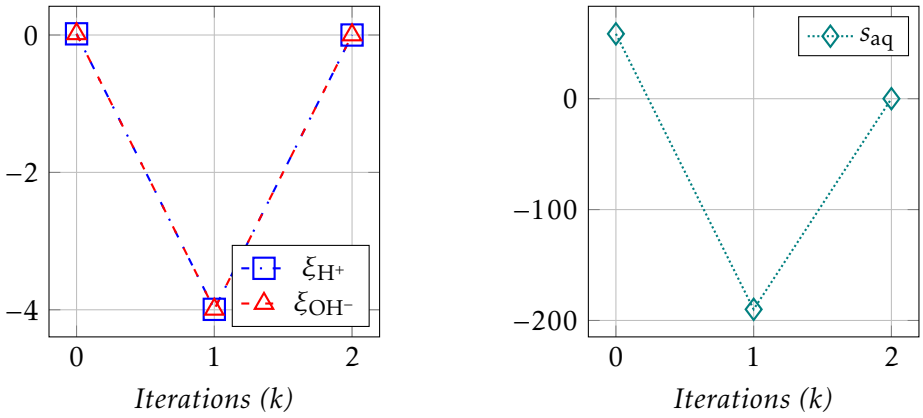


**Figure 4.12 –** Evolution of $s_{\text{aq}}$, $\xi_{\text{H}^+}$ and $\xi_{\text{OH}^-}$ for the Cartesian representation with min function in setting **A**.

### IPM slowdown

Test case **B** has the unique characteristic of possessing a doubly active complementarity constraint for the mineral. In Figure 4.5, we observe that the tolerance on $\|\mathcal{F}(\boldsymbol{\mathcal{X}}^{(k)})\|_\infty$ is reached at iteration 21. A plateau is then observed around the machine epsilon until the convergence criteria is satisfied for $\|\mathcal{F}\text{compl}(\boldsymbol{\mathcal{X}}^{(k)})\|_\infty$. Furthermore, we note that the convergence is no longer quadratic but instead exhibits a linear behavior.

### Trouble for the complementarity functions

The observed slowdown in the FB function, as illustrated in Figure 4.13, can be attributed to an abrupt decay in the concentrations of $SiO_2$ and Quartz, as depicted in Figure 4.15. This variation does not effectively reduce the residual, resulting in the iterates falling below the solution value, which is subsequently followed by a significant compensatory adjustment that elevates the iterates above the solution value. A similar phenomenon is evident in the case of the parametrization using the min function, as shown in Figure 4.14, although this time the adjustment is unsuccessful. Specifically, Newton's algorithm for the parametrization method utilizing the min function encounters termination due to a singular Jacobian. At iteration 23, the variable $n_{\text{aq}}$ is assigned a value of 0, leading to the loss of a substantial portion of the Jacobian matrix (4.8).

Consequently, the matrix lacks sufficient information to maintain linear independence between rows 3 and 8. Moreover, the values in row 3 are approaching zero.

$$\mathcal{J}(\boldsymbol{\mathcal{X}}^{(23)}) \approx \begin{pmatrix} 0.0 & 0.0 & 0.0 & 0.0 & -0.037 & 0.47 & 2.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & -0.018 & 0.00018 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 3\times10^{-22} & 0.0 & 0.0 & 0.0 \\ 1.0 & -1.0 & 0.0 & -1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & -1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.47 & 4\times10^{-8} & 0.00018 & 4\times10^{-8} & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \end{pmatrix} \tag{4.8}$$

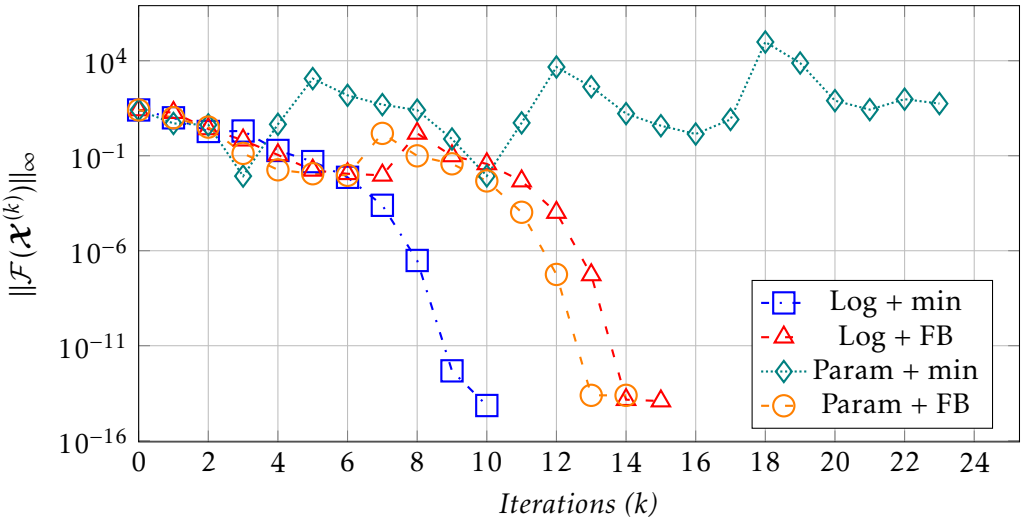


**Figure 4.13 –** Evolution of residuals in setting **C**.

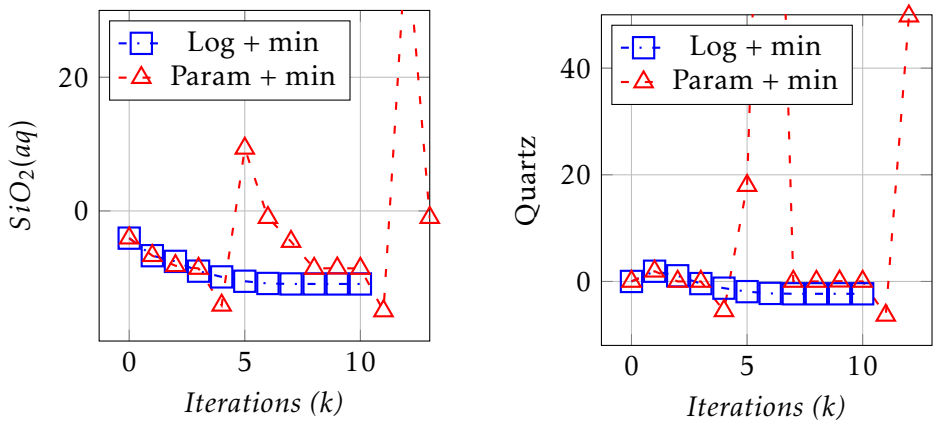


**Figure 4.14 –** Evolution of the concentrations of $SiO_2$(aq) and Quartz in setting **C** with the min function.

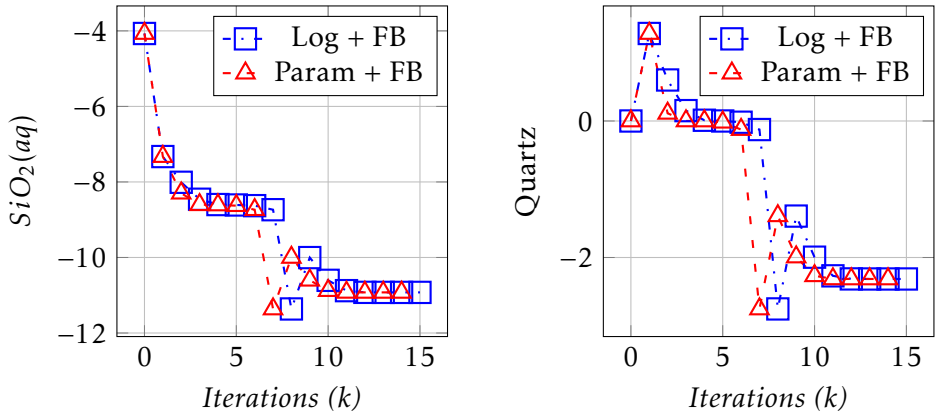


**Figure 4.15 –** Evolution of the concentrations of $SiO_2$(aq) and Quartz in setting **C** with the FB function.

### Additional comments on the $SiO_2$ test case

The Cartesian representation demonstrates convergence exclusively in the context of IPM, which incorporates a line search on the complementarity variables. Consequently, it is worthwhile to explore the application of this line search for both the min and FB cases. Table 4.4 presents the results obtained using the IPOPT line search, presented in Section 4.2.2, with the min and FB methods across all formulations. Our findings indicate that the Cartesian representation achieves convergence with the FB approach, although we also observe a slight deceleration in convergence in certain instances. Furthermore, the Cartesian representation methodology could potentially benefit from the integration of alternative line search strategies to enhance its performance.

| Log formulation | Complementarity | Setting | | |
|---|---|---|---|---|
| | | A | B | C |
| Log-trick | min | 9 | 7 | 10 |
| | FB | 8 | 8 | 18 |
| Param | min | 6 | 8 | 10 |
| | FB | 6 | 7 | 14 |
| Cart. repr. | min | × | × | × |
| | FB | 8 | 8 | 23 |

**Table 4.4 –** Number of iterations for Newton's method with IPOPT line search.

***Multiphase Seawater* test case**

This second test case is composed of 72 species divided into 22 phases:

— 50 aqueous species in 1 phase;

— 20 minerals in 20 pure phases;

— 2 gaseous species in 1 phase.

It is a challenging test case that allows us to evaluate the limits of each method. There are no methods that can converge without a line search strategy. For the min and FB approaches, which use the same variables as IPM, we will uses the line search from IPOPT. The case of complementarity parametrization is slightly different in that there is only one variable per phase. We note that the issue of invertibility, when a significant portion of the Jacobian matrix cancels out, also arises in the case of the complementarity parametrization. We then propose a strategy that "pauses" at zero when the sign of one of the $\eta_\alpha$ variables changes. This strategy gives the iterate a chance to remain in its zone at the next iteration, due to the specific choice made for $S'(0)$ and $R'(0)$ in Section 4.2.3. Although this strategy can be applied on a case-by-case basis, we have decided to retain Newton's direction on the $\boldsymbol{\eta}$ vector. The line search is then written as follows:

$$\boldsymbol{\eta}^{(k+1)} = \boldsymbol{\eta}^{(k)} + \max_{j\in\{1,\dots,N_{Ph}\}} \left\{\beta_j^{(k)}\right\} \delta\boldsymbol{\eta}^{(k)},$$

$$\beta_j^{(k)} = \begin{cases} -\eta_j^{(k)}/\delta\eta_j^{(k)} & \text{if } \operatorname{sign}\left(\eta_j^{(k)} + \delta\eta_j^{(k)}\right) \neq \operatorname{sign}\left(\eta_j^{(k)}\right), \\ 1 & \text{otherwise.} \end{cases}$$

However, the chosen initialization significantly constrains the variation of $\eta$ during the first iteration, resulting in an uncontrolled augmentation of subsequent iterates.

To address this issue, we recommend initializing with a present aqueous phase, as it contains all the chemical elements, while considering mineral and gaseous phases as absent. Consequently, we propose the following initialization for $\eta$:

$$\eta_{\text{aq}} = s_{\text{aq}}, \quad \eta_{\text{min}} = -s_{\text{aq}}, \quad \eta_{\text{gas}} = -s_{\text{aq}}$$

The $\eta$ variables of the absent phases are initialized with significantly negative values to enhance the potential for evolution.

To test the robustness of our methods, we will decrease the quantities of oxygen $\mathbf{b}_{\text{O}}$ present in the chemical system under investigation. This variation in the oxygen content will allow the appearance of certain minerals and the gas phase. The specific variations in oxygen concentration and the number of phases present in each case are in Table 4.5.

| $\mathbf{b}_{\text{O}}$ | **Nb of phases** | | |
|---|---|---|---|
| | **aqueous** | **mineral** | **gaseous** |
| **55** | 1 | 1 | 0 |
| **15** | 1 | 2 | 0 |
| **5** | 1 | 3 | 0 |
| **1** | 1 | 5 | 0 |
| **0.5** | 1 | 5 | 1 |

**Table 4.5 –** Number of present phases for each quantities of oxygen considered.

The results obtained are summarized in Table 4.6 and the associated solutions are in Appendix A.3. We can conclude that the complementarity parametrization method is highly robust when combined with the log-trick and parametrization formulations, although the parametrization formulation exhibits a slight advantage. Indeed, we can observe in Figure 4.16 that the log-trick can suffer from abrupt variations in residual values, leading it to diverge in the case of $\mathbf{b}_{\text{O}} = 0.5$. The parametrization formulation using the FB function is also very robust, but demonstrates a slower convergence rate compared to the other methods.

| Log formulation | Complementarity | Quantities of oxygen | | | | |
|---|---|---|---|---|---|---|
| | | 55 | 15 | 5 | 1 | 0.5 |
| Log-trick | param | 21 | 26 | 46 | 27 | × |
| | min | × | × | × | × | × |
| | FB | 91 | × | × | × | × |
| | IPM | × | × | × | × | × |
| Param | param | 27 | 31 | 36 | 29 | 31 |
| | min | × | × | × | × | × |
| | FB | 70 | 67 | 65 | 52 | 63 |
| | IPM | × | × | × | × | 55 |
| Cart. repr. | param | × | × | × | × | × |
| | min | × | × | × | × | × |
| | FB | × | 67 | × | 78 | × |
| | IPM | × | × | × | × | × |

**Table 4.6 –** Number of iterations for Newton's method.

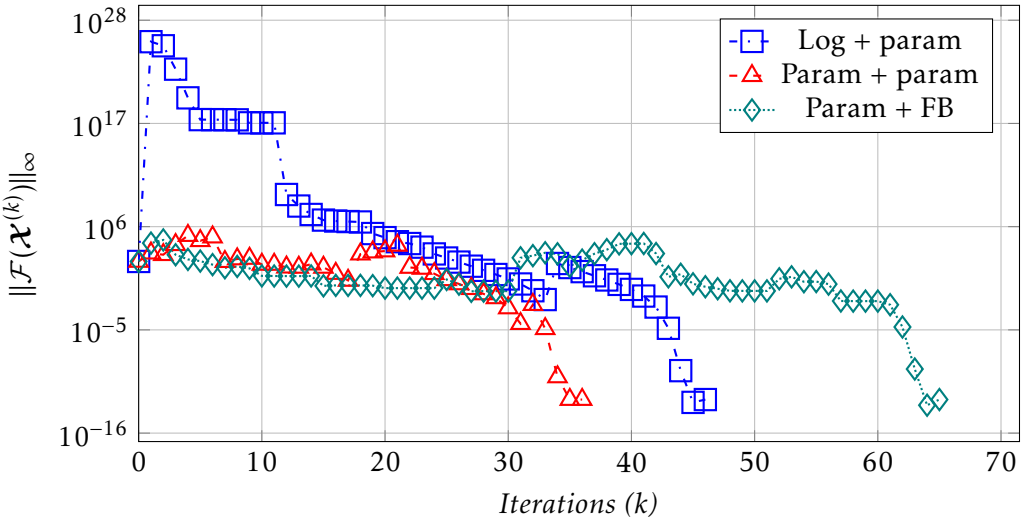


**Figure 4.16 –** Evolution of residuals with $\mathbf{b}_O = 5$.

# Conclusions and perspectives

**Outline of the current chapter**



## Summary of contributions

This summary highlights the key contributions of the thesis, organized by chapters that focus on significant advancements in the field of chemical equilibrium.

### Original theoretical contributions in chemical equilibrium

In Chapter 2, we conducted a thorough analysis of ideal single-phase equilibrium, confirming the existence of a unique solution for ideal systems. We extended this understanding to mixtures with convex Gibbs energy that may be non-ideal. A notable contribution is the introduction of a new system of algebraic equations based on mole fractions, which enhances the theoretical framework for single-phase equilibria.

Furthermore, we addressed the complexities of multiphase equilibrium by introducing new conditions that guarantee uniqueness even when certain phases may vanish.

This contribution provides a deeper analytical framework for multiphase systems. Additionally, we established an equivalence between the minimization problem and a new set of algebraic equations based on extended mole fractions, leveraging concepts from the subdifferential of Gibbs energy.

### More robust numerical algorithms for single-phase systems

In Chapter 3, we focused on developing numerical methods to resolve single-phase chemical equilibrium problems. We derived parametrization and Cartesian representation methods to effectively tackle challenges encountered when applying Newton's method. A significant contribution includes providing a theoretical proof for the local quadratic convergence of Newton's method for both approaches, ensuring reliability in our numerical techniques.

We conducted extensive numerical experiments demonstrating the accuracy of our techniques, enabling computations of equilibria for chemical species with very low concentrations. Our comparative analysis with results obtained using an equivalent of the Arxim solver revealed superior robustness in our Cartesian representation method. These findings are documented in the preprint [33].

### Innovative numerical methods for multiphase systems

In Chapter 4, we introduced the complementarity parametrization method, which integrates complementarity equations directly into the system through parameterization techniques. This innovative approach significantly enhances the modeling of phase behaviors in multiphase systems.

We provided compelling evidence showcasing the robustness and efficiency of the complementarity parametrization approach in addressing complementarity conditions related to phase presence or disappearance. This contribution is pivotal for advancing computational methods in multiphase chemical equilibrium analysis.

## Research perspectives

As pointed out previously, significant progresses on the numerical resolution of chemical equilibrium problems are collected in this manuscript. However, there are several promising research directions that can further build upon these achievements. By pursuing these research perspectives, the methods developed in this thesis can

be further extended and applied to a wider range of problems, contributing to the advancement of reactive transport modeling and its applications in various fields.

### Coupling with reactive transport codes

To fully leverage the potential of the developed methods, it would be beneficial to couple them with existing reactive transport codes. By incorporating these techniques into reactive transport codes, the computational efficiency and robustness of solving the chemical part can be significantly improved. This would enable more accurate and computationally feasible simulations of complex geochemical processes in porous media, significantly aiding the evaluation of CO2 sequestration programs.

### Extension to kinetic equilibria

The current work has focused on solving chemical equilibrium problems. However, many real-world systems involve both equilibrium and kinetic reactions. An important future research direction would be to extend the developed methods to handle kinetic equilibrium problems. This would involve coupling the equilibrium solvers with kinetic rate expressions and solving the resulting system of differential-algebraic equations. Such an approach could provide a more comprehensive framework for modeling complex geochemical systems.

### Extension of the multiphase formulation to non-ideal Gibbs energies

The formulation we have proposed is limited to chemical equilibrium in ideal multiphase systems. However, real-world applications often involve non-ideal mixtures where the Gibbs energy landscape can be complex and non-convex. Extending our formulation to accommodate non-ideal Gibbs energy is crucial for accurately modeling these systems.

### A novel approach for addressing complementarity constraints

The complementarity parametrization method for addressing the complementarity constraint, as detailed in Chapter 4, appears to be novel to our knowledge. The promising results we obtained encourage further exploration of this approach and its application to other problems. If the complementarity parametrization continues to produce compelling results in different contexts, it will be essential to develop a mathematical theory that provides additional theoretical guarantees.

# Bibliography


[1] V. Acary and B. Brogliato. Numerical Methods for Nonsmooth Dynamical Systems. Lecture Notes in Applied and Computational Mechanics. Springer Berlin, 2008. DOI: https://doi.org/10.1007/978-3-540-75392-6.

[2] C. Bale, P. Chartrand, S. Degterov, G. Eriksson, K. Hack, R. Ben Mahfoud, J. Melançon, A. Pelton, and S. Petersen. "FactSage Thermochemical Software and Databases". In: Calphad 26.2 (2002), pp. 189–228. DOI: 10.1016/S0364-5916(02)00035-4.

[3] S. Bassetto. "Towards more robust and accurate computations of capillary effects in the simulation of multiphase flows in porous media". PhD thesis. Université de Lille, 2021. URL: https://theses.hal.science/tel-03512051.

[4] S. Bassetto, C. Cancès, G. Enchéry, and Q. H. Tran. "Robust Newton solver based on variable switch for a finite volume discretization of Richards equation". In: Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples. Ed. by R. Klöfkorn, E. Keilegavlen, F. A. Radu, and J. Fuhrmann. Vol. 323. Springer Proceedings in Mathematics & Statistics. Cham: Springer, 2020. DOI: 10.1007/978-3-030-43651-3_35.

[5] Study of Compositional Multi-Phase Flow Formulations with Cubic EOS. SPE Reservoir Simulation Conference. 2015. URL: https://doi.org/SPE-173249-MS.

[6] I. Ben Gharbia. "Résolution de problèmes de complémentarité. : Application à un écoulement diphasique dans un milieu poreux". PhD thesis. Université Paris Dauphine - Paris IX, 2012. URL: https://theses.hal.science/tel-00776617.

[7] I. Ben Gharbia and E. Flauraud. "Study of compositional multiphase flow formulation using complementarity conditions". In: Oil Gas Sci. Technol. – Rev. IFP Energies nouvelles 74.43 (2019). DOI: https://doi.org/10.2516/ogst/2019012.

[8] I. Ben Gharbia and J. Jaffré. "Gas phase appearance and disappearance as a problem with complementarity constraints". In: Mathematics and Computers in Simulation 99 (2014), pp. 28–36. ISSN: 0378-4754. DOI: https://doi.org/10.1016/j.matcom.2013.04.021.

[9] C.-L. Berthollet. Essai de statique chimique. Paris: Firmin Didot, 1803.

[10] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. "Julia: A fresh approach to numerical computing". In: SIAM review 59.1 (2017), pp. 65–98. URL: https://doi.org/10.1137/141000671.

[11] K. Brenner and C. Cancès. "Improving Newton's Method Performance by Parametrization: The Case of the Richards Equation". In: SIAM Journal on Numerical Analysis 55.4 (2017), pp. 1760–1785. DOI: 10.1137/16M1083414.

[12] S. R. Brinkley. "Calculation of the Equilibrium Composition of Systems of Many Constituents". In: The Journal of Chemical Physics 15.2 (1947), pp. 107–110. DOI: 10.1063/1.1746420.

[13] C. de Capitani and K. Petrakakis. "The computation of equilibrium assemblage diagrams with Theriak/Domino software". In: American Mineralogist 95.7 (2010), pp. 1006–1016. DOI: 10.2138/am.2010.3354.

[14] F. H. Clarke. "Generalized Gradients and Applications". In: Transactions of the American Mathematical Society 205 (1975), pp. 247–262. DOI: 10.1090/S0002-9947-1975-0367131-6.

[15] J. Coatléven and A. Michel. "A successive substitution approach with embedded phase stability for simultaneous chemical and phase equilibrium calculations". In: Computers & Chemical Engineering 168 (2022), p. 108041. DOI: 10.1016/j.compchemeng.2022.108041.

[16] Ş. Cobzaş, R. Miculescu, and A. Nicolae. Lipschitz Functions. Vol. 2241. Lecture Notes in Mathematics. Springer Cham, 2019. ISBN: 978-3-030-16489-8. URL: https://link.springer.com/book/10.1007/978-3-030-16489-8.

[17] J. A. D. Connolly and K. Petrini. "An automated strategy for calculation of phase diagram sections and retrieval of rock properties as a function of physical conditions". In: Journal of Metamorphic Geology 20.7 (2002), pp. 697–708. DOI: 10.1046/j.1525-1314.2002.00398.x.

[18] J. Connolly. "Computation of phase equilibria by linear programming: A tool for geodynamic modeling and its application to subduction zone decarbonation". In: Earth and Planetary Science Letters 236.1-2 (2005), pp. 524–541. DOI: 10.1016/j.epsl.2005.04.033.

[19] M. A. Cook. The science of high explosives. Vol. 139. American Chemical Society Monoggraph. New York: Reinhold Publishing Corporation, 1958.

[20] G. B. Dantzig and J. C. DeHaven. "On the Reduction of Certain Multiplicative Chemical Equilibrium Systems to Mathematically Equivalent Additive Systems". In: The Journal of Chemical Physics 36.10 (1962), pp. 2620–2627. DOI: 10.1063/1.1732342.

[21] J. E. Dennis and R. B. Schnabel. Numerical methods for unconstrained optimization and nonlinear equations. Vol. 16. 1996. URL: https://epubs.siam.org/doi/book/10.1137/1.9781611971200.

[22] P. Deuflhard. Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms. Springer Series in Computational Mathematics. Springer Berlin, 2011. DOI: https://doi.org/10.1007/978-3-642-23899-4.

[23] H.-J. G. Diersch and P. Perrochet. "On the primary variable switching technique for simulating unsaturated–saturated flows". In: Advances in Water Resources 23.3 (1999), pp. 271–301. DOI: 10.1016/S0309-1708(98)00057-8.

[24] G. Eriksson and K. Hack. "ChemSage—A computer program for the calculation of complex chemical equilibria". In: Metallurgical Transactions B 21.6 (1990), pp. 1013–1023. DOI: 10.1007/BF02670272.

[25] F. Facchinei and J.-S. Pang. Finite-Dimensional Variational Inequalities and Complementarity Problems. Vol. Volume I. Springer Series in Operations Research and Financial Engineering. Springer New York, 2003. DOI: https://doi.org/10.1007/b97543.

[26] F. Facchinei and J.-S. Pang. Finite-Dimensional Variational Inequalities and Complementarity Problems. Vol. Volume II. Springer Series in Operations Research and Financial Engineering. Springer New York, 2003. DOI: https://doi.org/10.1007/b97544.

[27] A. Fischer. "A special Newton-type optimization method". In: Optimization 24.3-4 (1992), pp. 269–284. URL: https://doi.org/10.1080/02331939208843795.

[28] I. B. Gharbia, M. Haddou, Q. H. Tran, and D. T. S. Vu. "An analysis of the unified formulation for the equilibrium problem of compositional multiphase mixtures". In: ESAIM: M2AN 55.6 (2021), p. 38. DOI: https://doi.org/10.1051/m2an/2021075.

[29] J. W. Gibbs. "A method of geometrical representation of the thermodynamic properties of substances by means of surfaces". In: Transactions of the Connecticut Academy of Arts and Sciences 2 (1873), pp. 382–404. URL: https://www3.nd.edu/~powers/ame.20231/gibbs1873b.pdf.

[30] J. Gondzio. "Interior point methods 25 years later". In: European Journal of Operational Research 218.3 (2012), pp. 587–601. ISSN: 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2011.09.017.

[31] 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, 2014. DOI: https://doi.org/10.1007/978-3-319-04247-3.

[32] J. W. Johnson, E. H. Oelkers, and H. C. Helgeson. "SUPCRT92: A software package for calculating the standard molal thermodynamic properties of minerals, gases, aqueous species, and reactions from 1 to 5000 bar and 0 to 1000°C". In: Computers & Geosciences 18.7 (1992), pp. 899–947. DOI: 10.1016/0098-3004(92)90029-Q.

[33] M. Jonval, I. Ben Gharbia, C. Cancès, T. Faney, and Q.-H. Tran. "Parametrization and Cartesian representation techniques for robust resolution of chemical equilibria". In: (2024). URL: https://hal.science/hal-04225504v2.

[34] C. T. Kelley. Iterative methods for linear and nonlinear equations. 1995. URL: https://epubs.siam.org/doi/pdf/10.1137/1.9781611970944.fm.

[35] C. T. Kelley. Solving nonlinear equations with Newton's method. Fundamentals of algorithms. Philadelphia: Society for Industrial and Applied Mathematics, 2003. ISBN: 978-0-89871-546-0. DOI: 10.1137/1.9780898718898.

[36] K. A. Kobe and T. W. Leland. "The Calculation of Chemical Equilibrium in a Complex System". In: Special Publication No. 26 of the Bureau of Engineering Research, The University of Texas, Austin, TX. (1954).

[37] T. Koch, D. Gläser, K. Weishaupt, et al. "DuMux 3 – an open-source simulator for solving flow and transport problems in porous media with a focus on model coupling". In: Computers & Mathematics with Applications 81 (2021), pp. 423–443. DOI: 10.1016/j.camwa.2020.02.012.

[38] D. A. Kulik, T. Wagner, S. V. Dmytrieva, G. Kosakowski, F. F. Hingerl, K. V. Chudnenko, and U. R. Berner. "GEM-Selektor geochemical modeling package: revised algorithm and GEMS3K numerical kernel for coupled simulation codes". In: Computational Geosciences 17 (2012), pp. 1–24. DOI: 10.1007/s10596-012-9310-6.

[39] A. Lauser, C. Hager, R. Helmig, and B. Wohlmuth. "A new approach for phase transitions in miscible multi-phase flow in porous media". In: Advances in Water Resources 34.8 (2011), pp. 957–966. DOI: 10.1016/j.advwatres.2011.04.021.

[40] A. M. M. Leal. Reaktoro: A unified framework for modeling chemically reactive systems. 2015. URL: https://reaktoro.org.

[41] A. M. M. Leal, D. A. Kulik, W. R. Smith, and M. O. Saar. "An overview of computational methods for chemical equilibrium and kinetic calculations for geochemical and reactive transport modeling". In: Pure and Applied Chemistry 89.5 (2017), pp. 597–643. DOI: 10.1515/pac-2016-1107.

[42] J. van der Lee, L. De Windt, V. Lagneau, and P. Goblet. "Module-oriented modeling of reactive transport with HYTEC". In: 29.3 (), pp. 265–275. URL: https://dl.acm.org/doi/10.1016/S0098-3004%2803%2900004-9.

[43] J. van der Lee and L. De Windt. CHESS Tutorial and Cookbook. Tech. rep. LHM/RD/02/13. Ecole Nationale Supérieure des Mines de Paris, Centre d'Informatique Géologique, 2002. URL: https://radiochemistry.faculty.unlv.edu/readings/chess-tutorial3-0.pdf.

[44] R. Masson, L. Trenty, and Y. Zhang. "Formulations of two phase liquid gas compositional Darcy flows with phase transitions". In: International Journal of Finite Volume (2014).

[45] B. Metz, O. Davidson, H. de Coninck, M. Loos, and L. Meyer, eds. IPCC Special Report on Carbon Dioxide Capture and Storage. Cambridge, UK: Cambridge University Press, 2005, p. 431.

[46] R. Meyer. La Capture et Séquestration de Carbone pour réduire nos émissions de CO2. URL: https://lereveilleur.com/csc-reduction-des-emissions/.

[47] J. Moutte, A. Michel, G. Battaia, T. Parra, D. Garcia, and S. Wolf. "Arxim, a library for thermodynamic modeling of reactive heterogeneous systems, with applications to the simulation of fluid-rock systems". In: Proceedings of the 21st IUPAC International Conference on Chemical Thermodynamics. Tsukuba, Japan, 2010.

[48] Multicomponent reactive transport modeling in variably saturated porous media. URL: https://www.min3p.com/.

[49] D. Nordstrom and K. Campbell. "Modeling Low-Temperature Geochemical Processes". In: Treatise on Geochemistry. Elsevier, 2014, pp. 27–68. ISBN: 978-0-08-098300-4. DOI: 10.1016/B978-0-08-095975-7.00502-7.

[50] Northern Lights (projet industriel). URL: https://fr.wikipedia.org/wiki/Northern_Lights_(projet_industriel).

[51] J. M. Ortega and W. C. Rheinboldt. Iterative solution of nonlinear equations in several variables. Vol. 30. Classics in Applied Mathematics. Philadelphia: Society for Industrial and Applied Mathematics, 2000. ISBN: 978-0-89871-461-6. DOI: 10.1137/1.9780898719468.

[52] J.-S. Pang. "Newton's Method for B-Differentiable Equations". In: Mathematics of Operations Research 15.2 (1990), pp. 311–341. URL: http://www.jstor.org/stable/3689785.

[53] D. L. Parkhurst and C. A. J. Appelo. "Description of input and examples for PHREEQC version 3–A computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations". In: Modeling techniques. chap. A43 6. U.S. Geological Survey Techniques and Methods, 2013, p. 497. URL: https://pubs.usgs.gov/tm/06/a43/.

[54] W. H. Press and S. A. Teukolsky. Numerical Recipes 3rd Edition: The Art of Scientific Computing. Cambridge University Press, 2007. ISBN: 978-0-521-88068-8. URL: http://nrbook.com.

[55] L. Qi and J. Sun. "A nonsmooth version of Newton's method". In: Mathematical Programming 58 (1993), pp. 353–367. URL: https://doi.org/10.1007/BF01581275.

[56] M. H. Reed, N. F. Spycher, and J. Palandri. "SOLVEQ-XPT : A Computer Program for Computing Aqueous-Mineral-Gas Equilibria". In: Department of Geological Sciences, University of Oregon (2018). URL: https://pages.uoregon.edu/palandri/data/solveq-xpt%20guide_v.2.25.pdf.

[57] M. H. Reed, N. F. Spycher, and J. Palandri. "Users Guide for CHIM-XPT: A Program for Computing Reaction Processes in Aqueous-Mineral-Gas Systems and MINTAB Guide". In: Department of Geological Sciences, University of Oregon (2016). URL: https://pages.uoregon.edu/palandri/data/chim-xpt%20guide%20V.2.50.pdf.

[58] J. Revels, M. Lubin, and T. Papamarkou. Forward-Mode Automatic Differentiation in Julia. 2016. arXiv: 1607.07892 [cs.MS].

[59] R. T. Rockafellar. Convex Analysis. Princeton: Princeton University Press, 1970. ISBN: 9781400873173. DOI: doi:10.1515/9781400873173.

[60] N. Z. Shapiro and L. S. Shapley. "Mass Action Laws and the Gibbs Free Energy Function". In: Journal of the Society for Industrial and Applied Mathematics 13.2 (1965), pp. 353–375. DOI: 10.1137/0113020.

[61] Y. V. Shvarov. "HCh: New potentialities for the thermodynamic simulation of geochemical systems offered by windows". In: Geochemistry International 46.8 (2008), pp. 834–839. DOI: 10.1134/S0016702908080089.

[62] J. V. Smith, R. W. Missen, and W. R. Smith. "General optimality criteria for multiphase multireaction chemical equilibrium". In: AIChE Journal 39.4 (1993), pp. 707–710. DOI: 10.1002/aic.690390421.

[63] W. R. Smith. "The Computation of Chemical Equilibria in Complex Systems". In: Industrial & Engineering Chemistry Fundamentals 19.1 (1980), pp. 1–10. DOI: 10.1021/i160073a001.

[64] W. R. Smith and R. W. Missen. Chemical reaction equilibrium analysis: Theory and algorithms. New York: John Wiley & Sons, 1982.

[65] Stockage du CO2 : la Norvège espère capter le marché. URL: https://www.radiofrance.fr/franceculture/podcasts/le-reportage-de-la-redaction/stockage-du-co2-la-norvege-espere-capter-le-marche-9798992.

[66] G. P. Sutton. Rocket propulsion elements: an introduction to engineering of rockets. eng. 3 ed. New York: John Wiley & Sons, 1963. ISBN: 978-0-471-83835-7.

[67] The Intergovernmental Panel on Climate Change. URL: https://www.ipcc.ch/.

[68] D. C. Thorstenson. The concept of electron activity and its relation to redox potentials in aqueous geochemical systems. Tech. rep. 84-072. US Geological survey, 1984.

[69] A. H. Truesdell and B. F. Jones. "WATEQ, a computer program for calculating chemical equilibria of natural waters". In: Journal of Research of the U.S. Geological Survey 2.2 (1974), pp. 233–248. URL: https://pubs.usgs.gov/journal/1974/vol2issue2/report.pdf#page=105.

[70] C. Tsanas, E. H. Stenby, and W. Yan. "Calculation of Multiphase Chemical Equilibrium by the Modified RAND Method". In: Industrial & Engineering Chemistry Research 56.41 (2017), pp. 11983–11995. DOI: 10.1021/acs.iecr.7b02714.

[71] C. Tsanas, E. H. Stenby, and W. Yan. "Calculation of simultaneous chemical and phase equilibrium by the method of Lagrange multipliers". In: Chemical Engineering Science 174 (2017), pp. 112–126. DOI: 10.1016/j.ces.2017.08.033.

[72] D. Voskov, I. Saifullin, M. Wapperom, X. Tian, A. Palha, L. Orozco, and A. Novikov. open Delft Advanced Research Terra Simulator (open-Darts). 2023. URL: https://zenodo.org/records/8116928.

[73] D. T. S. Vu. "Numerical resolution of algebraic systems with complementarity conditions: Application to the thermodynamics of compositional multiphase mixtures". In: (2020). URL: https://theses.fr/2020UPASG006.

[74] D. T. S. Vu, I. Ben Gharbia, M. Haddou, and Q. H. Tran. "A new approach for solving nonlinear algebraic systems with complementarity conditions. Application to compositional multiphase equilibrium problems". In: Mathematics and Computers in Simulation 190 (2021), pp. 1243–1274. ISSN: 0378-4754. DOI: https://doi.org/10.1016/j.matcom.2021.07.015.

[75] P. Waage and C. M. Gulberg. "Studies Concerning Affinity". In: Journal of Chemical Education 63.12 (1986), pp. 1044–1047. DOI: 10.1021/ed063p1044.

[76] A. Wächter and L. T. Biegler. "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming". In: Mathematical Programming 106.1 (2006), pp. 25–57. DOI: 10.1007/s10107-004-0559-y.

[77] J. C. Westall, J. L. Zachary, and F. M. M. Morel. MINEQL: A computer program for the calculation of chemical equilibrium composition of aqueous systems. Tech. rep. Cambridge, Mass. : Water Quality Laboratory, Ralph M. Parsons Laboratory for Water Resources and Environmental Engineering sic, Dept. of Civil Engineering, Massachusetts Institute of Technology, 1976. URL: https://dspace.mit.edu/handle/1721.1/142980.

[78] T. Wolery. EQ3/6, a software package for geochemical modeling of aqueous systems: Package overview and installation guide (Version 7.0). Tech. rep. UCRL-MA–110662-Pt.1, 138894. Lawrence Livermore National Laboratory, 1992. URL: http://www.osti.gov/servlets/purl/138894/.

[79] T. Wolery. EQ3NR, a computer program for geochemical aqueous speciation-solubility calculations: Theoretical manual, user's guide, and related documentation (Version 7.0); Part 3. Tech. rep. UCRL-MA–110662-Pt.3, 138643. Lawrence Livermore National Laboratory, 1992, UCRL–MA–110662–Pt.3, 138643. DOI: 10.2172/138643.

[80] F. van Zeggeren and S. H. Storey. The computation of chemical equilibria. London: Cambridge University Press, 1970. ISBN: 978-0-521-07630-2.

[81] F. van Zeggeren and S. H. Storey. "The effect of changes in initial reactant composition on solid-gas equilibria resulting from constant-volume, adiabatic processes". In: The Canadian Journal of Chemical Engineering 47.1 (1969), pp. 81–84. DOI: 10.1002/cjce.5450470115.

Appendix A

# Chemical systems for numerical experiments

**Outline of the current chapter**



## A.1 Standard chemical potentials

The standard chemical potential $\mu_i^\circ(P, T)$ of a species $C_i$ for a constant pressure P and temperature T is calculated from the SUPCRT92 database [32].

| Formula | $\mu_i^\circ$(P,T) | Formula | $\mu_i^\circ$(P,T) |
|---|---|---|---|
| $H_2O$ | -237138.97589284607 | $H^+$ | 9956.885403312557 |
| $O_2(aq)$ | 26500.37842186901 | $Na^+$ | -251923.79553026147 |
| $Mg^{2+}$ | -444027.90215494944 | $K^+$ | -272504.8996797245 |
| $Ca^{2+}$ | -542833.0422538111 | $Fe^{2+}$ | -81547.14762262715 |
| $HCO_3^-$ | -576982.8987273659 | $Al^{3+}$ | -473751.00441000663 |
| $SO_4^{2-}$ | -734502.084490013 | $Cl^-$ | -121332.84714937139 |
| $Sr^{2+}$ | -553878.8001745282 | $AlO^+$ | -651901.5520727821 |
| $AlOH^{2+}$ | -682390.3513772341 | $HAlO_2(aq)$ | -859059.7240321051 |
| $AlO_2^-$ | -821374.446375568 | $CaOH^+$ | -706762.1524620559 |

| | | | |
|---|---|---|---|
| $CO(aq)$ | -110048.61728485748 | $CO_2(aq)$ | -376017.06579503417 |
| $CO_3^{2-}$ | -518026.1359207859 | $CaHCO_3^+$ | -1.135747579476859e6 |
| $CaCl^+$ | -672453.3605597073 | $CaCl_2(aq)$ | -801738.9637317051 |
| $CaSO_4(aq)$ | -1.2993419278916481e6 | $HClO(aq)$ | -69957.53260756626 |
| $ClO^-$ | -26862.311151193757 | $ClO_2^-$ | 27111.248231900317 |
| $ClO_3^-$ | 2007.235286837261 | $ClO_4^-$ | 1421.4659696309664 |
| $Fe^{3+}$ | -7281.149189013573 | $FeCl^+$ | -211920.58507072268 |
| $FeCl_2(aq)$ | -297483.39734809624 | $FeOH^{2+}$ | -231878.2281150759 |
| $FeOH^+$ | -265559.42898331815 | $FeO^+$ | -212213.40826659818 |
| $FeO$ | -202255.52192397387 | $HFeO_2^-$ | -389196.60766906774 |
| $HFeO_2(aq)$ | -413045.42294696183 | $FeO_2^-$ | -358235.0321847625 |
| $H_2(aq)$ | 27680.272394604483 | $H_2S(aq)$ | -17962.985906651626 |
| $HO_2^-$ | -57363.666715716085 | $HS^-$ | 21923.09086173558 |
| $HSO_3^-$ | -517770.9559572122 | $HSO_4^-$ | -745798.9041292057 |
| $HSO_5^-$ | -627559.1114566573 | $KCl(aq)$ | -389322.1898012777 |
| $KHSO_4(aq)$ | -1.0084285424401537e6 | $KOH(aq)$ | -427271.0297495857 |
| $KSO_4^-$ | -1.0219846680885215e6 | $CH_4(aq)$ | -24494.210815931885 |
| $Mg(CO_3)(aq)$ | -989014.6934790071 | $Mg(HCO_3)^+$ | -1.0368796788141541e6 |
| $MgCl^+$ | -574547.7790990678 | $MgOH^+$ | -614525.8903919919 |
| $NaCl(aq)$ | -378778.4933733225 | $NaOH(aq)$ | -408024.6192443839 |
| $OH^-$ | -147340.5566329419 | $S_2^{2-}$ | 89452.83031941311 |
| $S_2O_3^{2-}$ | -512624.6363618013 | $HS_2O_3^-$ | -522247.82998507 |
| $H_2S_2O_3(aq)$ | -525595.041488881 | $S_2O_4^{2-}$ | -590447.0178825587 |
| $HS_2O_4^-$ | -604672.624055689 | $H_2S_2O_4(aq)$ | -606764.6340067171 |
| $S_2O_5^{2-}$ | -780818.9790117 | $S_2O_6^{2-}$ | -956546.945397623 |
| $S_2O_8^{2-}$ | -1.10507895975272e6 | $S_3^{2-}$ | 83595.22019473514 |
| $S_3O_6^{2-}$ | -948178.949248605 | $S_4^{2-}$ | 78992.80990918931 |
| $S_4O_6^{2-}$ | -1.0306037475969802e6 | $S_5^{2-}$ | 75645.59931794215 |
| $S_5O_6^{2-}$ | -948178.9537022973 | $SO_2(aq)$ | -291207.4128400276 |
| $SO_3^{2-}$ | -476642.20414498576 | $Sr(HCO_3)^+$ | -1.1478393365598267e6 |
| $SrCl^+$ | -683750.1584404652 | $SrOH^+$ | -715130.1518277868 |
| $H_2O_2(aq)$ | -124056.64548498068 | $HClO_2(aq)$ | 15814.426726057798 |
| $NaSO_4^-$ | -1.0003785022780169e6 | $MgSO_4(aq)$ | -1.2012145947920694e6 |
| $HCl(aq)$ | -117278.5194378308 | $CaCO_3(aq)$ | -1.0898072308832686e6 |
| $SrCO_3(aq)$ | -1.0982170694298274e6 | $FeCl^{2+}$ | -147018.36458365855 |
| $Br^-$ | -94099.20681114095 | $B(OH)_3(aq)$ | -958806.2944243755 |
| $F^-$ | -271793.6108649851 | $BF_4^-$ | -1.477036419104179e6 |
| $BO_2^-$ | -668855.0735514564 | $CaF^+$ | -828474.6584440577 |
| $HF(aq)$ | -289876.86749552627 | $HF_2^-$ | -568104.4495969338 |
| $KBr(aq)$ | -366644.9255226958 | $MgF^+$ | -733497.8766684392 |
| $NaBr(aq)$ | -348235.31123035855 | $NaF(aq)$ | -527979.8734791129 |
| $SrF^+$ | -836424.2572999222 | $CO_2(g)$ | -394391.27240993205 |
| $H_2O(g)$ | -228164.33933913204 | ANHYDRITE | -1.3218299449455e6 |
| ARAGONITE | -1.12835344756575e6 | ARTINITE | -2.5686198744775e6 |
| BRUCITE | -835318.69698875 | CALCITE | -1.1291776990575502e6 |
| CELESTITE | -1.34069978084625e6 | DOLOMITE-DIS | -2.1574917216637502e6 |
| DOLOMITE-ORD | -2.1663074074905002e6 | FLUORITE | -1.17358246402515e6 |
| HALITE | -384120.431987875 | HUNTITE | -4.2037057981325e6 |
| HYDROMAGNESITE | -5.864657282973e6 | LIME | -604027.2218463 |
| MAGNESITE | -1.02783286346585e6 | PERICLASE | -569383.7028176 |

| | | | |
|---|---|---|---|
| POTASSIUM-OXIDE | -322402.2804475 | SODIUM-OXIDE | -376070.415242 |
| STRONTIANITE | -1.15256625621825e6 | SYLVITE | -408923.19198430004 |
| NESQUEHONITE | -1.7239541270617498e6 | $e^-$ | -16.315331966024218 |

**Table A.1 –** Standard chemical potentials at P = 1 Bar and T = 298.15 K.

## A.2 Test case: single-phase chemical systems

### A.2.1 The *dissociation of water* test case

The $H_2O$ test case is composed of

$$\mathcal{C} = (H_2O, H^+, OH^-),\ \mathcal{E} = (H, O),\ \mathcal{R} = (OH^- = H_2O - H^+).$$

The vector of constraints **b** for this test case is composed of $n_O = 55.5087$ and a charge of 0. The solution obtained is $\mathbf{n} = (55.5086998985565, 1.0144349897420512 \times 10^{-7}, 1.0144349897420512 \times 10^{-7})$.

### A.2.2 The *Seawater* test case

The *Seawater* test case is composed of:

$$\begin{aligned}\mathcal{C} = (&H_2O, Na^+, Mg^{2+}, SO_4^{2-}, Ca^{2+}, K^+, HCO_3^-, Sr^{2+}, Cl^-, H+,\\ &CaOH^+, CO_2(aq), CO_3^{2-}, CaHCO_3^+, CaCl^+, CaCl_2(aq), CaSO_4(aq), HSO_4^-, KCl(aq),\\ &KHSO_4(aq), KOH(aq), KSO_4^-, Mg(CO_3)(aq), Mg(HCO_3)^+, MgCl^+, MgOH^+, NaCl(aq),\\ &NaOH(aq), OH^-, Sr(HCO_3)^+, SrCl^+, SrOH^+, NaSO_4^-, MgSO_4(aq),\\ &HCl(aq), CaCO_3(aq), SrCO_3(aq)),\\ \mathcal{E} = (&H, O, Na, Mg, S, Ca, K, C, Sr, Cl),\end{aligned}$$

and the set $\mathcal{R}$ composed of the reactions:

$$\begin{aligned}
CaOH^+ &= H_2O + Ca^{2+} - H^+; & CO_2(aq) &= -H_2O + HCO_3^- + H^+;\\
CO_3^{2-} &= HCO_3^- - H^+; & CaHCO_3^+ &= Ca^{2+} + HCO_3^-;\\
CaCl^+ &= Ca^{2+} + Cl^-; & CaCl_2(aq) &= Ca^{2+} + 2Cl^-;\\
CaSO_4(aq) &= SO_4^{2-} + Ca^{2+}; & HSO_4^- &= SO_4^{2-} + H^+;\\
KCl(aq) &= K^+ + Cl^-; & KHSO_4(aq) &= SO_4^{2-} + K^+ + H^+;\\
KOH(aq) &= H_2O + K^+ - H^+; & KSO_4^- &= SO_4^{2-} + K^+;\\
Mg(CO_3)(aq) &= Mg^{2+} + HCO_3^- - H^+; & Mg(HCO_3)^+ &= Mg^{2+} + HCO_3^-;\\
MgCl^+ &= Mg^{2+} + Cl^-; & MgOH^+ &= H_2O + Mg^{2+} - H^+;\\
NaCl(aq) &= Na^+ + Cl^-; & NaOH(aq) &= H_2O + Na^+ - H^+;\\
OH^- &= H_2O - H^+; & Sr(HCO_3)^+ &= HCO_3^- + Sr^{2+};\\
SrCl^+ &= Sr^{2+} + Cl^-; & SrOH^+ &= H_2O + Sr^{2+} - H^+;\\
NaSO_4^- &= Na^+ + SO_4^{2-}; & MgSO_4(aq) &= Mg^{2+} + SO_4^{2-};
\end{aligned}$$

$$\mathrm{HCl(aq)} = \mathrm{Cl^-} + \mathrm{H^+}; \qquad \mathrm{CaCO_3(aq)} = \mathrm{Ca^{2+}} + \mathrm{HCO_3^-} - \mathrm{H^+};$$
$$\mathrm{SrCO_3(aq)} = \mathrm{HCO_3^-} + \mathrm{Sr^{2+}} - \mathrm{H^+}.$$

The vector **b** for this test case is given in Table A.2.

| Feeds (mol) | | | | |
|---|---|---|---|---|
| O | Na | Mg | S | Ca |
| 55.5087 | 0.469 | 0.0528 | 0.0282 | 0.0103 |
| K | C | Sr | Cl | Z (charge) |
| 0.0102 | 0.00206 | $1 \times 10^{-5}$ | 0.546 | 0 |

**Table A.2 –** Vector **b** of the *Seawater* test case for the elements conservation.

The solution obtained for the *Seawater* test case is

$$\begin{aligned}\mathbf{n} = (&55.38968051132381, 0.42692804180975974, 0.02705950281438023,\\ &0.0033461882935940023, 0.005755697020771243, 0.009937871289383643,\\ &0.0008404617918854967, 7.642853227990452 \times 10^{-6}, 0.4991461549132867,\\ &1.6714211435836995 \times 10^{-9}, 5.132624707658365 \times 10^{-7}, 3.0602737036270006 \times 10^{-6},\\ &2.3982990446766207 \times 10^{-5}, 5.298476689811943 \times 10^{-5}, 0.00144101412325631,\\ &0.0003152095229661645 7, 0.0024462442193710124, 5.242661308469255 \times 10^{-10},\\ &1.4219455529061577 \times 10^{-5}, 1.7188295621737205 \times 10^{-15}, 2.1991731951952572 \times 10^{-8},\\ &0.00024788726335362165, 0.0006080613936706859, 0.0002428721982493850 8,\\ &0.009738343343159612, 3.4148566766825304 \times 10^{-5}, 0.0350277315444541,\\ &1.6186301459231793 \times 10^{-6}, 6.246267369519702 \times 10^{-6}, 1.0728986616184622 \times 10^{-7},\\ &2.1174144853264242 \times 10^{-6}, 2.3140715139129212 \times 10^{-10}, 0.007042608015640257,\\ &0.015117071683773249, 1.598968187418331 \times 10^{-10}, 0.0002883370842663884,\\ &1.3221101336988816 \times 10^{-7})\end{aligned}$$

### A.2.3 The *Water-Concrete* test cases

The *Water-Concrete* test case is composed of:

$$\mathcal{C} = (\mathrm{H_2O}, \mathrm{H^+}, \mathrm{O_2(aq)}, \mathrm{Na^+}, \mathrm{Mg^{2+}}, \mathrm{K^+}, \mathrm{Ca^{2+}}, \mathrm{Fe^{2+}},$$
$$\mathrm{HCO_3^-}, \mathrm{Al^{3+}}, \mathrm{SO_4^{2-}}, \mathrm{Cl^-}, \mathrm{Sr^{2+}}, \mathrm{AlO^+}, \mathrm{AlOH^{2+}}, \mathrm{HAlO_2(aq)},$$
$$\mathrm{AlO_2^-}, \mathrm{CaOH^+}, \mathrm{CO(aq)}, \mathrm{CO_2(aq)}, \mathrm{CO_3^{2-}}, \mathrm{CaHCO_3^+}, \mathrm{CaCl^+}, \mathrm{CaCl_2(aq)},$$
$$\mathrm{CaSO_4(aq)}, \mathrm{HClO(aq)}, \mathrm{ClO^-}, \mathrm{ClO_2^-}, \mathrm{ClO_3^-}, \mathrm{ClO_4^-}, \mathrm{Fe^{3+}}, \mathrm{FeCl^+},$$
$$\mathrm{FeCl_2(aq)}, \mathrm{FeOH^{2+}}, \mathrm{FeOH^+}, \mathrm{FeO^+}, \mathrm{FeO}, \mathrm{HFeO_2^-}, \mathrm{HFeO_2(aq)}, \mathrm{FeO_2^-}, \mathrm{H_2(aq)},$$
$$\mathrm{H_2S(aq)}, \mathrm{HO_2^-}, \mathrm{HS^-}, \mathrm{HSO_3^-}, \mathrm{HSO_4^-}, \mathrm{HSO_5^-}, \mathrm{KCl(aq)}, \mathrm{KHSO_4(aq)},$$
$$\mathrm{KOH(aq)}, \mathrm{KSO_4^-}, \mathrm{CH_4(aq)}, \mathrm{Mg(CO_3)(aq)}, \mathrm{Mg(HCO_3)^+}, \mathrm{MgCl^+}, \mathrm{MgOH^+}, \mathrm{NaCl(aq)},$$
$$\mathrm{NaOH(aq)}, \mathrm{OH^-}, \mathrm{S_2^{2-}}, \mathrm{S_2O_3^{2-}}, \mathrm{HS_2O_3^-}, \mathrm{H_2S_2O_3(aq)}, \mathrm{S_2O_4^{2-}}, \mathrm{HS_2O_4^-},$$
$$\mathrm{H_2S_2O_4(aq)}, \mathrm{S_2O_5^{2-}}, \mathrm{S_2O_6^{2-}}, \mathrm{S_2O_8^{2-}}, \mathrm{S_3^{2-}}, \mathrm{S_3O_6^{2-}}, \mathrm{S_4^{2-}}, \mathrm{S_4O_6^{2-}},$$
$$\mathrm{S_5^{2-}}, \mathrm{S_5O_6^{2-}}, \mathrm{SO_2(aq)}, \mathrm{SO_3^{2-}}, \mathrm{Sr(HCO_3)^+}, \mathrm{SrCl^+}, \mathrm{SrOH^+}, \mathrm{H_2O_2(aq)},$$
$$\mathrm{HClO_2(aq)}, \mathrm{NaSO_4^-}, \mathrm{MgSO_4(aq)}, \mathrm{HCl(aq)}, \mathrm{CaCO_3(aq)}, \mathrm{SrCO_3(aq)}, \mathrm{FeCl^{2+}})$$
$$\mathcal{E} = (\mathrm{H}, \mathrm{O}, \mathrm{Na}, \mathrm{Mg}, \mathrm{K}, \mathrm{Ca}, \mathrm{Fe}, \mathrm{C}, \mathrm{Al}, \mathrm{S}, \mathrm{Cl}, \mathrm{Sr}),$$

and the set $\mathcal{R}$ composed of the reactions:

$$\mathrm{AlO^+ = H_2O - 2H^+ + Al^{3+}};$$
$$\mathrm{AlOH^{2+} = H_2O - H^+ + Al^{3+}};$$
$$\mathrm{HAlO_2(aq) = 2H_2O - 3H^+ + Al^{3+}};$$
$$\mathrm{AlO_2^- = 2H_2O - 4H^+ + Al^{3+}};$$
$$\mathrm{CaOH^+ = H_2O - H^+ + Ca^{2+}};$$
$$\mathrm{CO(aq) = -H_2O + H^+ - 0.5O_2(aq) + HCO_3^-};$$
$$\mathrm{CO_2(aq) = -H_2O + H^+ + HCO_3^-};$$
$$\mathrm{CO_3^{2-} = -H^+ + HCO_3^-};$$
$$\mathrm{CaHCO_3^+ = Ca^{2+} + HCO_3^-};$$
$$\mathrm{CaCl^+ = Ca^{2+} + Cl^-};$$
$$\mathrm{CaCl_2(aq) = Ca^{2+} + 2Cl^-};$$
$$\mathrm{CaSO_4(aq) = Ca^{2+} + SO_4^{2-}};$$
$$\mathrm{HClO(aq) = H^+ + 0.5O_2(aq) + Cl^-};$$
$$\mathrm{ClO^- = 0.5O_2(aq) + Cl^-};$$
$$\mathrm{ClO_2^- = O_2(aq) + Cl^-};$$
$$\mathrm{ClO_3^- = 1.5O_2(aq) + Cl^-};$$
$$\mathrm{ClO_4^- = 2O_2(aq) + Cl^-};$$
$$\mathrm{Fe^{3+} = -0.5H_2O + H^+ + 0.25O_2(aq) + Fe^{2+}};$$
$$\mathrm{FeCl^+ = Fe^{2+} + Cl^-};$$
$$\mathrm{FeCl_2(aq) = Fe^{2+} + 2Cl^-};$$
$$\mathrm{FeOH^{2+} = 0.5H_2O + 0.25O_2(aq) + Fe^{2+}};$$
$$\mathrm{FeOH^+ = H_2O - H^+ + Fe^{2+}};$$
$$\mathrm{FeO^+ = 0.5H_2O - H^+ + 0.25O_2(aq) + Fe^{2+}};$$
$$\mathrm{FeO = H_2O - 2H^+ + Fe^{2+}};$$
$$\mathrm{HFeO_2^- = 2H_2O - 3H^+ + Fe^{2+}};$$
$$\mathrm{HFeO_2(aq) = 1.5H_2O - 2H^+ + 0.25O_2(aq) + Fe^{2+}};$$
$$\mathrm{FeO_2^- = 1.5H_2O - 3H^+ + 0.25O_2(aq) + Fe^{2+}};$$
$$\mathrm{H_2(aq) = H_2O - 0.5O_2(aq)};$$
$$\mathrm{H_2S(aq) = 2H^+ - 2O_2(aq) + SO_4^{2-}};$$
$$\mathrm{HO_2^- = H_2O - H^+ + 0.5O_2(aq)};$$
$$\mathrm{HS^- = H^+ - 2O_2(aq) + SO_4^{2-}};$$
$$\mathrm{HSO_3^- = H^+ - 0.5O_2(aq) + SO_4^{2-}};$$
$$\mathrm{HSO_4^- = H^+ + SO_4^{2-}};$$
$$\mathrm{HSO_5^- = H^+ + 0.5O_2(aq) + SO_4^{2-}};$$
$$\mathrm{KCl(aq) = K^+ + Cl^-};$$
$$\mathrm{KHSO_4(aq) = H^+ + K^+ + SO_4^{2-}};$$
$$\mathrm{KOH(aq) = H_2O - H^+ + K^+};$$
$$\mathrm{KSO_4^- = K^+ + SO_4^{2-}};$$
$$\mathrm{CH_4(aq) = H_2O + H^+ - 2O_2(aq) + HCO_3^-};$$
$$\mathrm{Mg(CO_3)(aq) = -H^+ + Mg^{2+} + HCO_3^-};$$
$$\mathrm{Mg(HCO_3)^+ = Mg^{2+} + HCO_3^-};$$
$$\mathrm{MgCl^+ = Mg^{2+} + Cl^-};$$
$$\mathrm{MgOH^+ = H_2O - H^+ + Mg^{2+}};$$
$$\mathrm{NaCl(aq) = Na^+ + Cl^-};$$
$$\mathrm{NaOH(aq) = H_2O - H^+ + Na^+};$$
$$\mathrm{OH^- = H_2O - H^+};$$

$$S_2^{2-} = -H_2O + 2H^+ - 3.5O_2(aq) + 2SO_4^{2-};$$
$$S_2O_3^{2-} = -H_2O + 2H^+ - 2O_2(aq) + 2SO_4^{2-};$$
$$HS_2O_3^- = -H_2O + 3H^+ - 2O_2(aq) + 2SO_4^{2-};$$
$$H_2S_2O_3(aq) = -H_2O + 4H^+ - 2O_2(aq) + 2SO_4^{2-};$$
$$S_2O_4^{2-} = -H_2O + 2H^+ - 1.5O_2(aq) + 2SO_4^{2-};$$
$$HS_2O_4^- = -H_2O + 3H^+ - 1.5O_2(aq) + 2SO_4^{2-};$$
$$H_2S_2O_4(aq) = -H_2O + 4H^+ - 1.5O_2(aq) + 2SO_4^{2-};$$
$$S_2O_5^{2-} = -H_2O + 2H^+ - O_2(aq) + 2SO_4^{2-};$$
$$S_2O_6^{2-} = -H_2O + 2H^+ - 0.5O_2(aq) + 2SO_4^{2-};$$
$$S_2O_8^{2-} = -H_2O + 2H^+ + 0.5O_2(aq) + 2SO_4^{2-};$$
$$S_3^{2-} = -2H_2O + 4H^+ - 5O_2(aq) + 3SO_4^{2-};$$
$$S_3O_6^{2-} = -2H_2O + 4H^+ - 2O_2(aq) + 3SO_4^{2-};$$
$$S_4^{2-} = -3H_2O + 6H^+ - 6.5O_2(aq) + 4SO_4^{2-};$$
$$S_4O_6^{2-} = -3H_2O + 6H^+ - 3.5O_2(aq) + 4SO_4^{2-};$$
$$S_5^{2-} = -4H_2O + 8H^+ - 8O_2(aq) + 5SO_4^{2-};$$
$$S_5O_6^{2-} = -4H_2O + 8H^+ - 5O_2(aq) + 5SO_4^{2-};$$
$$SO_2(aq) = -H_2O + 2H^+ - 0.5O_2(aq) + SO_4^{2-};$$
$$SO_3^{2-} = -0.5O_2(aq) + SO_4^{2-};$$
$$Sr(HCO_3)^+ = HCO_3^- + Sr^{2+};$$
$$SrCl^+ = Cl^- + Sr^{2+};$$
$$SrOH^+ = H_2O - H^+ + Sr^{2+};$$
$$H_2O_2(aq) = H_2O + 0.5O_2(aq);$$
$$HClO_2(aq) = H^+ + O_2(aq) + Cl^-;$$
$$NaSO_4^- = Na^+ + SO_4^{2-};$$
$$MgSO_4(aq) = Mg^{2+} + SO_4^{2-};$$
$$HCl(aq) = H^+ + Cl^-;$$
$$CaCO_3(aq) = -H^+ + Ca^{2+} + HCO_3^-;$$
$$SrCO_3(aq) = -H^+ + HCO_3^- + Sr^{2+};$$
$$FeCl^{2+} = -0.5H_2O + H^+ + 0.25O_2(aq) + Fe^{2+} + Cl^-.$$

The vector **b** for this test case is given in Table A.3.

| Feeds (mol) | | | | |
|---|---|---|---|---|
| O | Na | Mg | K | Ca |
| 55.5078 | 0.0601 | $1.5079 \times 10^{-9}$ | 0.1402 | 0.00196384 |
| Fe | C | Al | S | Cl |
| $4.58364 \times 10^{-7}$ | $5.29145 \times 10^{-5}$ | $3.80016 \times 10^{-5}$ | 0.000974141 | $1.42825 \times 10^{-10}$ |
| Sr | Z (charge) | pE | | |
| $1 \times 10^{-10}$ | 0 | -2.98873 | | |

**Table A.3 –** Vector **b** of the *Water-Concrete* test case for the elements conservation.

The solution obtained for the *Water-Concrete* test case is

$$\mathbf{n} = (55.30153263049913, 1.2515957847430055 \times 10^{-45}, 0.05383341189709943,$$
$$3.85205612592884 \times 10^{-11}, 0.13104185508007585, 0.0005173318114609391,$$
$$1.7211732119737842 \times 10^{-22}, 2.968381612134554 \times 10^{-8}, 5.512454738276492 \times 10^{-36},$$
$$0.0004189076148432289, 1.414654075260127 \times 10^{-10}, 5.177690711306953 \times 10^{-11},$$
$$5.534605441939609 \times 10^{-14}, 8.331028635348468 \times 10^{-20}, 1.0086583996023467 \times 10^{-27},$$
$$8.387036717264553 \times 10^{-12}, 3.8001591612963186 \times 10^{-5}, 0.0013909717301057072,$$
$$9.16012434284405 \times 10^{-41}, 3.584713859640573 \times 10^{-15}, 2.5242434859636466 \times 10^{-5},$$
$$1.7045004485556293 \times 10^{-10}, 3.7199328160101444 \times 10^{-14}, 2.3370165495489375 \times 10^{-24},$$
$$2.7894078954109622 \times 10^{-5}, 7.744041296711739 \times 10^{-54}, 3.9571740772053503 \times 10^{-48},$$
$$1.3767116074490875 \times 10^{-78}, 3.4198978487645584 \times 10^{-95}, 4.303257445391223 \times 10^{-116},$$
$$1.734550767970587 \times 10^{-38}, 1.6769901176709222 \times 10^{-32}, 2.305681346281873 \times 10^{-50},$$
$$1.9826398981963012 \times 10^{-27}, 1.526767859061291 \times 10^{-18}, 1.2896347040880048 \times 10^{-17},$$

| Formula | $\mu_i^\circ$(P,T) | Formula | $\mu_i^\circ$(P,T) |
|---|---|---|---|
| $H_2O$ | -263568.5484586815 | $H^+$ | 19140.66365557133 |
| $SiO_2$(aq) | -831259.6930622912 | $OH^-$ | -121912.89750247078 |
| Quartz | -872288.8744824057 | | |

**Table A.4 –** Standard chemical potentials at P = 200 Bar and T = 573,15 K.

$2.2480301791284984\times10^{-16}$, $6.4990813411481995\times10^{-12}$, $1.0117847734211808\times10^{-10}$, $4.582563222020898\times10^{-7}$, $2.2465395189279816\times10^{-24}$, $3.8486081967306244\times10^{-72}$, $3.185210777712516\times10^{-38}$, $7.176882182296528\times10^{-66}$, $2.495689820120032\times10^{-34}$, $2.2023864652584633\times10^{-15}$, $4.2214199103551675\times10^{-57}$, $5.385118088684071\times10^{-14}$, $9.648599524565427\times10^{-20}$, 0.008743466121629882, 0.0004146787982403328, $7.700254829779092\times10^{-76}$, $9.23254021400711\times10^{-13}$, $1.237443643602187\times10^{-17}$, $3.981561928493589\times10^{-21}$, $1.4657260543330728\times10^{-9}$, $1.2685419568327648\times10^{-12}$, 0.006153927596402049, 0.18584677511174025, $3.3198439612835856\times10^{-120}$, $9.804207509062558\times10^{-78}$, $2.6241149020481446\times10^{-89}$, $5.585408849329376\times10^{-102}$, $4.191684712607641\times10^{-85}$, $7.1822771665537554\times10^{-96}$, $9.213692990260926\times10^{-109}$, $9.355900949612459\times10^{-73}$, $5.679111032409044\times10^{-63}$, $5.890335493096298\times10^{-79}$, $6.092605128626266\times10^{-175}$, $3.3555331102837764\times10^{-120}$, $6.738886584218897\times10^{-230}$, $1.5972079714124394\times10^{-161}$, $4.4923443994746124\times10^{-285}$, $1.0016752314700165\times10^{-231}$, $9.877024332443298\times10^{-46}$, $2.81914046857951\times10^{-28}$, $2.6014418358334538\times10^{-17}$, $4.119856257314945\times10^{-21}$, $4.7267746136716166\times10^{-11}$, $8.488781208032296\times10^{-40}$, $7.237999946744915\times10^{-90}$, 0.000112660505230008, $2.7301180078217594\times10^{-12}$, $1.5206742654457507\times10^{-24}$, $2.764220899199947\times10^{-5}$, $9.553207316759742\times10^{-13}$, $7.384932228127377\times10^{-47}$)

## A.3 Test case: multiphase chemical systems

### A.3.1 The $SiO_2$est case

The $SiO_2$ test case contains one aqueous phase and one mineral and it is composed of:

$$\begin{aligned}
\mathcal{C} &= \{H_2O,\ H^+,\ SiO_2(aq),\ OH^-,\ Quartz(SiO_2(s))\},\\
\mathcal{E} &= \{H,\ O,\ Si,\\
\mathcal{R} &= \{\quad OH^- = H_2O - H^+\\
&\qquad SiO_2(s) = SiO_2(aq)\}.
\end{aligned}$$

For this test case, P=200 Bar and T=573,15 K. The associated standard chemical potentials are given in Table A.4 and vectors **b** for the different settings are given in Table A.5.

| | Feeds (mol) | | |
|---|---|---|---|
| | O | Si | Z (charge) |
| **A** | 55.5087 | 1 | 0 |
| **B** | 53.050787670647416 | 0.025393835323706666 | 0 |
| **C** | 55.5087 | 0.001 | 0 |

**Table A.5 –** Vector **b** of the $SiO_2$ test case for the elements conservation.

| Setting | **A** | **B** | **C** |
|---|---|---|---|
| $H_2O$ | 53.508697479559125 | 52.99999750315028 | 55.506697385661376 |
| $H^+$ | 2.520440870534497e-6 | 2.496849719735878e-6 | 2.6143386168330353e-6 |
| $SiO_2$(aq) | 0.009757969875539042 | 0.02539383532370667 | 0.0010000000000000002 |
| $OH^-$ | 2.520440870534497e-6 | 2.496849719735878e-6 | 2.6143386168330353e-6 |
| Quartz | 0.990242030124461 | 0.0 | 0.0 |

**Table A.6 –** Solution vectors of the $SiO_2$ test case.

### A.3.2 The *Multiphase Seawater* test case

The *Multiphase Seawater* test case contains one aqueous phase, 20 minerals in pure phase and one gaseous phase. It is composed of:

$$
\begin{aligned}
\mathcal{C} = \{&H_2O, H^+, Cl^-, Na^+, Mg^{+2}, SO_4^{-2}, Ca^{+2}, K^+,\\
&HCO_3^-, Br^-, B(OH)_3(aq), Sr^{+2}, F^-, BF_4^-, BO_2^-, CaOH^+,\\
&CO_2(aq), CO_3^{-2}, CaHCO_3^+, CaCl^+, CaCl_2(aq), CaF^+, CaSO_4(aq), HF(aq),\\
&HF_2^-, HSO_4^-, KBr(aq), KCl(aq), KHSO_4(aq), KOH(aq), KSO_4^-, Mg(CO_3)(aq),\\
&Mg(HCO_3)^+, MgCl^+, MgF^+, MgOH^+, NaBr(aq), NaCl(aq), NaF(aq), NaOH(aq),\\
&OH^-, Sr(HCO_3)^+, SrCl^+, SrF^+, SrOH^+, NaSO_4^-, MgSO_4(aq), HCl(aq),\\
&CaCO_3(aq), SrCO_3(aq), Anhydrite, Aragonite, Artinite, Brucite,\\
&Calcite, Celestite, Dolomite – DIS, Dolomite – ORD, Fluorite, Halite, Huntite,\\
&Hydromagnesite, Lime, Magnesite, Periclase, Potassium – Oxide,\\
&Sodium – Oxide, Strontianite, Sylvite, Nesquehonite, CO_2(g), H_2O(g)\},\\
\mathcal{E} = \{&H, O, Cl, Na, Mg, S, Ca, K, C, Br, B, Sr, F\},
\end{aligned}
$$

and the set $\mathcal{R}$ composed of the reactions:

$$
\begin{aligned}
BF_4^- &= -3H_2O3H^+ + B(OH)_3(AQ)4F^-; & BO_2^- &= -H_2O - H^+ + B(OH)_3(AQ)\\
CaOH^+ &= H_2O - H^+ + Ca^{+2}; & CO_2(AQ) &= -H_2O + H^+ + HCO_3^-\\
CO_3^{-2} &= -H^+ + HCO_3^-; & CaHCO_3^+ &= Ca^{+2} + HCO_3^-\\
CaCl^+ &= Cl^- + Ca^{+2}; & CaCl_2(AQ) &= 2Cl^- + Ca^{+2}\\
CaF^+ &= Ca^{+2} + F^-; & CaSO_4(AQ) &= SO_4^{-2} + Ca^{+2}\\
HF(AQ) &= H^+ + F^-; & HF_2^- &= H^+ 2F^-\\
HSO_4^- &= H^+ + SO_4^{-2}; & KBr(AQ) &= K^+ + Br^-
\end{aligned}
$$

$$KCl(AQ) = Cl^- + K^+;$$
$$KOH(AQ) = H_2O - H^+ + K^+;$$
$$Mg(CO_3)(AQ) = -H^+ + Mg^{+2} + HCO_3^-;$$
$$MgCl^+ = Cl^- + Mg^{+2};$$
$$MgOH^+ = H_2O - H^+ + Mg^{+2};$$
$$NaCl(AQ) = Cl^- + Na^+;$$
$$NaOH(AQ) = H_2O - H^+ + Na^+;$$
$$Sr(HCO_3)^+ = HCO_3^- + Sr^{+2};$$
$$SrF^+ = Sr^{+2} + F^-;$$
$$NaSO_4^- = Na^+ + SO_4^{-2};$$
$$HCl(AQ) = H^+ + Cl^-;$$
$$SrCO_3(AQ) = -H^+ + HCO_3^- + Sr^{+2};$$
$$Aragonite = -H^+ + Ca^{+2} + HCO_3^-;$$
$$Brucite = 2H_2O - 2H^+ + Mg^{+2};$$
$$Celestite = SO_4^{-2} + Sr^{+2};$$
$$Dolomite - ORD = -2H^+ + Mg^{+2} + Ca^{+2}2HCO_3^-;$$
$$Halite = Cl^- + Na^+;$$
$$Hydromagnesite = 6H_2O - 6H^+5Mg^{+2}4HCO_3^-;$$
$$Magnesite = -H^+ + Mg^{+2} + HCO_3^-;$$
$$Potassium - Oxide = H_2O - 2H^+2K^+;$$
$$Strontianite = -H^+ + HCO_3^- + Sr^{+2};$$
$$Nesquehonite = 3H_2O - H^+ + Mg^{+2} + HCO_3^-;$$
$$H_2O(G) = H_2O;$$
$$KHSO_4(AQ) = H^+ + SO_4^{-2} + K^+$$
$$KSO_4^- = SO_4^{-2} + K^+$$
$$Mg(HCO_3)^+ = Mg^{+2} + HCO_3^-$$
$$MgF^+ = Mg^{+2} + F^-$$
$$NaBr(AQ) = Na^+ + Br^-$$
$$NaF(AQ) = Na^+ + F^-$$
$$OH^- = H_2O - H^+$$
$$SrCl^+ = Cl^- + Sr^{+2}$$
$$SrOH^+ = H_2O - H^+ + Sr^{+2}$$
$$MgSO_4(AQ) = Mg^{+2} + SO_4^{-2}$$
$$CaCO_3(AQ) = -H^+ + Ca^{+2} + HCO_3^-$$
$$Anhydrite = SO_4^{-2} + Ca^{+2}$$
$$Artinite = 5H_2O - 3H^+2Mg^{+2} + HCO_3^-$$
$$Calcite = -H^+ + Ca^{+2} + HCO_3^-$$
$$Dolomite - DIS = -2H^+ + Mg^{+2} + Ca^{+2}2HCO_3^-$$
$$Fluorite = Ca^{+2}2F^-$$
$$Huntite = -4H^+3Mg^{+2} + Ca^{+2}4HCO_3^-$$
$$Lime = H_2O - 2H^+ + Ca^{+2}$$
$$Periclase = H_2O - 2H^+ + Mg^{+2}$$
$$Sodium - Oxide = H_2O - 2H^+2Na^+$$
$$Sylvite = Cl^- + K^+$$
$$CO_2(G) = -H_2O + H^+ + HCO_3^-$$

The vector **b** for this test case is given in Table A.7.

| Feeds (mol) | | | | |
|---|---|---|---|---|
| Cl | Na | Mg | S | Ca |
| 0.546 | 0.469 | 0.0528 | 0.0282 | 0.0103 |
| K | C | Br | B | Sr |
| 0.0102 | 0.00206 | 0.000844 | 0.000416 | 0.000091 |
| F | Z (charge) | | | |
| 0.000068 | 0 | | | |

**Table A.7 –** Vector **b** of the *multiphase Seawater* test case for the elements conservation.

| $\mathbf{b}_O$ | 55 | 15 | 5 | 1 | 0.5 |
|---|---|---|---|---|---|
| $H_2O$ | 54.88029693727669 | 14.880474024263984 | 4.880538708333894 | 0.8805912926337224 | 0.3806208167151429 |
| $H^+$ | 4.041763217411945e-7 | 3.4907669775495055e-7 | 2.2524571122985246e-7 | 1.0557335944080568e-7 | 5.333769974506123e-8 |
| $Cl^-$ | 0.498728465043812 7 | 0.42867723798721075 | 0.34729774242758654 | 0.1876224611569889 | 0.10407557139746387 |
| $Na^+$ | 0.4266550490209221 | 0.36785153418144856 | 0.29488766993527915 | 0.14028726937861835 | 0.05735437043498458 |
| $Mg^{+2}$ | 0.02718930364822258 | 0.017898739048400652 | 0.011539378967232963 | 0.006446946119913651 | 0.00536796823782442 |
| $SO_4^{-2}$ | 0.0033138186182350626 | 0.0013672834041931396 | 0.0006525471793323628 | 0.0002808383683534976 | 0.00018232837158243588 |
| $Ca^{+2}$ | 0.005710282498179524 | 0.0029461084415922335 | 0.0008169535313956894 | 0.0001202205957010383 1 | 4.1994585009552795e-5 |
| $K^+$ | 0.009937792119902754 | 0.009801006465504414 | 0.009645618299636883 | 0.009249241227901526 | 0.008976033198470278 |
| $HCO_3^-$ | 0.0005902333900327811 | 0.00024794653157847457 | 0.00013765449462787492 | 5.662714514372406e-5 | 2.4321895660416842e-5 |
| $Br^-$ | 0.0008284230563099871 | 0.0007982817207710564 | 0.0007492116591658581 | 0.0006793137951859043 | 0.0006942115035790783 |

| | | | | | |
|---|---|---|---|---|---|
| $B(OH)_3$(AQ) | 0.00041540964972943133 | 0.00041579725575835544 | 0.00041587319301714435 | 0.0004159050274667182 | 0.0004159013701566526 |
| $Sr^{+2}$ | 7.05762115603475e-5 | 4.887434447413311e-5 | 3.128231490764926e-5 | 5.1199713260411705e-6 | 1.7884711836947215e-6 |
| $F^-$ | 4.05250286846291e-5 | 2.623558042421954e-5 | 1.1228099328232325e-5 | 3.6951750725534738e-6 | 2.054645887124914e-6 |
| $BF_4^-$ | 7.135549162850914e-23 | 6.280334863937405e-20 | 9.161665125414996e-19 | 4.697255859986805e-17 | 1.3941489117878202e-16 |
| $BO_2^-$ | 5.903502705689351e-7 | 2.0274424164423376e-7 | 1.268069828545495e-7 | 9.49725332347351e-8 | 9.862984320809484e-8 |
| $CaOH^+$ | 2.0864202856521143e-9 | 3.3794284261512066e-10 | 4.7632919679061716e-11 | 2.6983453854978144e-12 | 8.063993436869207e-13 |
| $CO_2$(AQ) | 0.0005245216915081791 | 0.0007018549939493414 | 0.0007665922589228328 | 0.0008191998911631415 | 0.0004112675373116612 |
| $CO_3^{-2}$ | 6.902234991833261e-8 | 9.516254037409722e-9 | 2.978610487682658e-9 | 6.579050154260229e-10 | 2.6635577987087164e-10 |
| $CaHCO_3^+$ | 3.7252113225008486e-5 | 2.8482906054914238e-5 | 1.2053530803286375e-5 | 2.8994595301533556e-6 | 9.134780335556056e-7 |
| $CaCl^+$ | 0.0014414472387463885 | 0.002255091343868703 | 0.0013926221553935021 | 0.0004399311022948949 5 | 0.0001790016133719773 7 |
| $CaCl_2$(AQ) | 0.0003179074362411906 | 0.0015081353756637252 | 0.002074101595641337 | 0.0014065372486406966 | 0.0006666246621023249 |
| $CaF^+$ | 1.1036104724481043e-6 | 1.3004159234463986e-6 | 4.2422423355253894e-7 | 8.163807312959679e-8 | 3.329683948024087e-8 |
| $CaSO_4$(AQ) | 0.002425337921770255 | 0.0018213826112624767 | 0.000662601151307285 | 0.0001667497567120048 2 | 7.94092920359596e-5 |
| HF(AQ) | 2.393699872603126e-8 | 4.7216633577133175e-8 | 3.5842249651371845e-8 | 2.196892580519187e-8 | 1.295939154222947e-8 |
| $HF_2^-$ | 2.3246241272917314e-13 | 1.0472572981234125e-12 | 9.3522951712584e-13 | 7.496309651693842e-13 | 5.163202465616108e-13 |
| $HSO_4^-$ | 1.2669210523237003e-7 | 1.592712318039473e-7 | 1.3482653852001373e-7 | 1.0806988696400522e-7 | 7.443485298200283e-8 |
| KBr(AQ) | 1.4964988011652368e-7 | 5.017289190674652e-7 | 1.2738722363947157e-6 | 4.4010346825307095e-6 | 9.165336016071712e-6 |
| KCl(AQ) | 1.4336738085706919e-5 | 4.287512232647275e-5 | 9.396915132414181e-5 | 0.00019343327488544274 | 0.0002186590156230979 3 |
| $KHSO_4$(AQ) | 4.191423296940981e-13 | 1.8333190973221227e-12 | 4.1984066797712935e-12 | 1.2822621665494428e-11 | 1.799785277244005e-11 |
| KOH(AQ) | 9.010701242235141e-11 | 2.789906702261677e-11 | 1.395611052557833e-11 | 5.1516843935264735e-12 | 4.277261494430084e-12 |
| $KSO_4^-$ | 0.00024772140160527 4 | 0.0003556166535176565 | 0.0004591386586480 6387 | 0.0007529244445561985 | 0.00099614242761544 |
| $Mg(CO_3)$(AQ) | 1.7743795631232039e-6 | 5.681397158348838e-7 | 3.151460272281 88e-7 | 1.5453252512143826e-7 | 1.0938773024969942e-7 |
| $Mg(HCO_3)^+$ | 0.0001729403337050039 9 | 0.0001687185727458804 7 | 0.0001659985336031031 | 0.0001515992678858817 3 | 0.0001138464884274313 |
| $MgCl^+$ | 0.009865842960419294 | 0.019693948721101855 | 0.028275698294683155 | 0.033912108171427514 | 0.03289035385120167 |
| $MgF^+$ | 2.4617947555855274e-5 | 3.701276459176375e-5 | 2.807217703432187e-5 | 2.0509895572693163e-5 | 1.9939542725652427e-5 |
| $MgOH^+$ | 1.4058965652880176e-7 | 2.9055479251778692e-8 | 9.521456049183482e-9 | 2.0477823041014494e-9 | 1.4587399994658705e-9 |
| NaBr(AQ) | 1.5427293809896016e-5 | 4.5216550309876995e-5 | 9.351446859774685e-5 | 0.00016028517013156465 | 0.0001406231604048499 5 |
| NaCl(AQ) | 0.035294339944005076 | 0.0922730762574767 | 0.16473260952317667 | 0.1682329543856544 | 0.08011561796732034 |
| NaF(AQ) | 1.725563848585748e-6 | 3.3978331348743527e-6 | 3.2044290367434693e-6 | 1.993554244634957e-6 | 9.516388439683796e-7 |
| NaOH(AQ) | 6.627846619291228e-9 | 1.7939835382601514e-9 | 7.310007451281477e-10 | 1.338715470199396e-10 | 4.68246722586911e-11 |
| $OH^-$ | 2.5362307845996888e-8 | 2.256997645358346e-9 | 4.17347128682095e-10 | 4.043144537512865e-11 | 1.6472654571855 67e-11 |
| $Sr(HCO_3)^+$ | 7.021044781510132e-7 | 7.205541022256185e-7 | 7.038269493913138e-7 | 1.8830236955943448e-7 | 5.932487640202284e-8 |
| $SrCl^+$ | 1.9714217618556074e-5 | 4.139771806453778e-5 | 5.900854102045175e-5 | 2.073258084649816e-5 | 8.435787789335735e-6 |
| $SrF^+$ | 3.911974830921306e-9 | 6.187197603969232e-9 | 4.658826034109734e-9 | 9.971509093920356e-10 | 4.066971756535436e-10 |
| $SrOH^+$ | 8.755525472003309e-12 | 1.9035099044780 16e-12 | 6.192819227564647e-13 | 3.901807708525891e-14 | 1.1660535349768524e-14 |
| $NaSO_4^-$ | 0.007033451549567748 | 0.0088267733836465 | 0.009283000912908844 | 0.007552340166365979 | 0.004209412580943009 |
| $MgSO_4$(AQ) | 0.01517954381629728 | 0.014545239140374755 | 0.012302212943341002 | 0.011754017852496497 | 0.013342399483850735 |
| HCl(AQ) | 3.898482979895622e-8 | 1.020986235648141e-7 | 1.467155328199435e-7 | 1.4761950732307423e-7 | 8.687234654374273e-8 |
| $CaCO_3$(AQ) | 8.30770364555633e-7 | 2.08476161327439e-7 | 4.9739534311461953e-8 | 6.424209512005336e-9 | 1.907774007284436e-9 |
| $SrCO_3$(AQ) | 3.545612588971577e-9 | 1.1942579896890575e-9 | 6.576771916325291e-10 | 9.447521434735164e-11 | 2.8055958935908545e-11 |
| Anhydrite | 0.0 | 0.0012835455339403473 | 0.004840364323725513 | 0.007628063275013475 | 0.009309517409735813 |
| Aragonite | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Artinite | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Brucite | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Calcite | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Celestite | 0.0 | 0.0 | 0.0 | 6.495805379275943e-5 | 8.07159813857724e-5 |
| Dolomite – DIS | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Dolomite – ORD | 0.0003658363245803452 | 0.0004557445575899873 | 0.0004883144166220993 | 0.0005146621123963387 | 0.0 |
| Fluorite | 0.0 | 0.0 | 1.2515283710500822e-5 | 2.0848384730412202e-5 | 2.2503754290929004e-5 |
| Halite | 0.0 | 0.0 | 0.0 | 0.1527651572111136 | 0.32717902417067857 |
| Huntite | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Hydromagnesite | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Lime | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Magnesite | 0.0 | 0.0 | 0.0 | 0.0 | 0.00106538154949984 |
| Periclase | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Potassium – Oxide | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Sodium – Oxide | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Strontianite | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Sylvite | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Nesquehonite | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| $CO_2$(G) | 0.0 | 0.0 | 0.0 | 0.0 | 0.00044409813627469794 |
| $H_2O$(G) | 0.0 | 0.0 | 0.0 | 0.0 | 6.646061154477252e-6 |

**Table A.8 –** Solution vectors for the *multiphase Seawater* test case.

Appendix B

# Globalization methods for Newton's algorithm

**Outline of the current chapter**



## B.1 Limiter for Newton's method in the single-phase case

Depending on the methods and formulation, the limiter is defined as:

1. Activity formulation:
   — *Log-trick*:
   $$\delta\mathbf{y} \leftarrow \text{sign}(\delta\mathbf{y}) \min\left(|\delta\mathbf{y}|, \max\left(|\mathbf{y}^{(k)}|, \ln\left(\|\mathbf{b}\|_1\right)\right)\right)$$
   — *Parametrization*:
   $$\delta\boldsymbol{\tau} \leftarrow \text{sign}(\delta\boldsymbol{\tau}) \min\left(|\delta\boldsymbol{\tau}|, \max\left(|\boldsymbol{\tau}^{(k)}|, X^{-1}\left(\|\mathbf{b}\|_1\right)\right)\right)$$
   — *Cartesian representation*:
   $$\delta\mathbf{n} \leftarrow \text{sign}(\delta\mathbf{n}) \min\left(|\delta\mathbf{n}|, \max\left(|\mathbf{n}^{(k)}|, \|\mathbf{b}\|_1\right)\right)$$
   $$\delta\mathbf{y} \leftarrow \text{sign}(\delta\mathbf{y}) \min\left(|\delta\mathbf{y}|, \max\left(|\mathbf{y}^{(k)}|, \log(\|\mathbf{b}\|_1\right)\right)$$
2. Quantity formulation:
   — *Log-trick*:
   $$\delta\mathbf{y} \leftarrow \text{sign}(\delta\mathbf{y}) \min\left(|\delta\mathbf{y}|, \max\left(|\mathbf{y}^{(k)}|, \ln\left(\|\mathbf{b}\|_1\right)\right)\right)$$

$$\delta s \leftarrow \text{sign}(\delta s)\min\left(|\delta s|, \max\left(|s^{(k)}|, \ln\left(\|\mathbf{b}\|_1\right)\right)\right)$$
$$\delta l \leftarrow \text{sign}(\delta l)\min\left(|\delta l|, \max\left(|l^{(k)}|, \log(\|\mathbf{b}\|_1\right)\right)$$

— *Parametrization*:

$$\delta \boldsymbol{\tau} \leftarrow \text{sign}(\delta \boldsymbol{\tau})\min\left(|\delta \boldsymbol{\tau}|, \max\left(|\boldsymbol{\tau}^{(k)}|, X^{-1}\left(\|\mathbf{b}\|_1\right)\right)\right)$$
$$\delta s \leftarrow \text{sign}(\delta s)\min\left(|\delta s|, \max\left(|s^{(k)}|, X^{-1}\left(\|\mathbf{b}\|_1\right)\right)\right)$$
$$\delta l \leftarrow \text{sign}(\delta l)\min\left(|\delta l|, \max\left(|l^{(k)}|, \log(\|\mathbf{b}\|_1\right)\right)$$

— *Cartesian representation*:

$$\delta \mathbf{n} \leftarrow \text{sign}(\delta \mathbf{n})\min\left(|\delta \mathbf{n}|, \max\left(|\mathbf{n}^{(k)}|, \|\mathbf{b}\|_1\right)\right)$$
$$\delta \mathbf{y} \leftarrow \text{sign}(\delta \mathbf{y})\min\left(|\delta \mathbf{y}|, \max\left(|\mathbf{y}^{(k)}|, \log(\|\mathbf{b}\|_1\right)\right)$$
$$\delta s \leftarrow \text{sign}(\delta s)\min\left(|\delta s|, \max\left(|s^{(k)}|, \|\mathbf{b}\|_1\right)\right)$$
$$\delta l \leftarrow \text{sign}(\delta l)\min\left(|\delta l|, \max\left(|l^{(k)}|, \log(\|\mathbf{b}\|_1\right)\right)$$

3. Mole fraction formulation:

— *Log-trick*:

$$\delta \mathbf{y} \leftarrow \text{sign}(\delta \mathbf{y})\min\left(|\delta \mathbf{y}|, \max\left(|\mathbf{y}^{(k)}|, \ln\left(\|\mathbf{b}\|_1\right)\right)\right)$$
$$\delta \omega \leftarrow \text{sign}(\delta \omega)\min\left(|\delta \omega|, \max\left(|\omega^{(k)}|, 1/\|\mathbf{b}\|_1\right)\right)$$

— *Parametrization*:

$$\delta \boldsymbol{\tau} \leftarrow \text{sign}(\delta \boldsymbol{\tau})\min\left(|\delta \boldsymbol{\tau}|, \max\left(|\boldsymbol{\tau}^{(k)}|, X^{-1}\left(\|\mathbf{b}\|_1\right)\right)\right)$$
$$\delta \omega \leftarrow \text{sign}(\delta \omega)\min\left(|\delta \omega|, \max\left(|\omega^{(k)}|, 1/\|\mathbf{b}\|_1\right)\right)$$

— *Cartesian representation*:

$$\delta \mathbf{x} \leftarrow \text{sign}(\delta \mathbf{x})\min\left(|\delta \mathbf{x}|, \max\left(|\mathbf{x}^{(k)}|, \|\mathbf{b}\|_1\right)\right)$$
$$\delta \mathbf{y} \leftarrow \text{sign}(\delta \mathbf{y})\min\left(|\delta \mathbf{y}|, \max\left(|\mathbf{y}^{(k)}|, \log(\|\mathbf{b}\|_1\right)\right)$$
$$\delta \omega \leftarrow \text{sign}(\delta \omega)\min\left(|\delta \omega|, \max\left(|\omega^{(k)}|, 1/\|\mathbf{b}\|_1\right)\right)$$

## B.2 Line search strategy from Numerical Recipes

This Line search strategy from Numerical Recipes [54] is as follows.

**Algorithm 1** Newton Line Search Algorithm 1/2

$\tau_{guess}$
step_max_factor $\leftarrow 100$
step_max $\leftarrow$ step_max_factor $\cdot \max(\|\tau_{guess}\|, \text{length}(\tau_{guess}))$
$\text{norm}_{\delta\tau} \leftarrow \|\delta\tau\|$
**if** $\text{norm}_{\delta\tau} >$ step_max **then**
    $\delta\tau \leftarrow \delta\tau \cdot \text{step\_max}/\text{norm}_{\delta\tau}$
**end if**
$(F, \text{Jac}) \leftarrow \text{evalFJ}(\text{system}, \tau)$
$\text{Grad} \leftarrow F^T \cdot \text{Jac}$
slope $\leftarrow \text{dot}(\text{Grad}^T, \delta\tau)$
**if** slope $\geq 0$ **then**
    **return** $\delta\tau$
**end if**
$\lambda \leftarrow 1.0$, $\lambda_{old} \leftarrow 1.0$
$\sigma_{max} \leftarrow 0.5$, $\sigma_{min} \leftarrow 0.1$
max_armijo_iter $\leftarrow 100$
$\tau_{old} \leftarrow \tau$
$\tau \leftarrow \tau_{old} + \lambda \cdot \delta\tau$
$\text{norm}_{F0} \leftarrow \|F\|$
$\text{norm}_{fmin0} \leftarrow 0.5 \cdot \text{norm}_{F0}^2$
$F \leftarrow \text{evalF}(\text{system}, \tau)$
$\text{norm}_F \leftarrow \|F\|$
$n_{armijo\_iter} \leftarrow 0$
max_exponent $\leftarrow 300$
max_value $\leftarrow 10^{0.5 \cdot \text{max\_exponent}}$
**if** $\text{norm}_F >$ max_value **or** isnan($\text{norm}_F$) **then**
    $\text{norm}_F \leftarrow$ max_value
**end if**
$\text{norm}_{fmin} \leftarrow 0.5 \cdot \text{norm}_F^2$
$\text{norm}_{fmin\_old} \leftarrow \text{norm}_{fmin}$
$\lambda_{min} \leftarrow \epsilon_{64}/\max(\text{abs}(\delta\tau)/\max(\text{abs}(\tau_{old}), \text{ones}(\text{length}(\tau_{old}))))$
residual $\leftarrow \text{norm}_F$

**Algorithm 2** Newton Line Search Algorithm 2/2

**while** $\text{norm}_{fmin} \geq \text{norm}_{fmin0} + 10^{-4} \cdot \lambda \cdot \text{slope}$ **and** residual $> 1e-10$ **do**
  **if** $\lambda < \lambda_{min}$ **or** $n_{armijo\_iter} >$ max_armijo_iter **then**
    **break**
  **end if**
  $\lambda_{tmp} \leftarrow 0.0$
  **if** $n_{armijo\_iter} = 0$ **then**
    $\lambda_{tmp} \leftarrow -\text{slope}/(2 \cdot (\text{norm}_{fmin} - \text{norm}_{fmin0} - \text{slope}))$
  **else**
    $\text{RHS}_1 \leftarrow \text{norm}_{fmin} - \text{norm}_{fmin0} - \lambda \cdot \text{slope}$
    $\text{RHS}_2 \leftarrow \text{norm}_{fmin\_old} - \text{norm}_{fmin0} - \lambda_{old} \cdot \text{slope}$
    $A \leftarrow (\text{RHS}_1/\lambda^2 - \text{RHS}_2/\lambda_{old}^2)/(\lambda - \lambda_{old})$
    $B \leftarrow (-\lambda_{old} \cdot \text{RHS}_1/\lambda^2 + \lambda \cdot \text{RHS}_2/\lambda_{old}^2)/(\lambda - \lambda_{old})$
    **if** $|A| < 100 \cdot \epsilon_{64}$ **then**
      $\lambda_{tmp} \leftarrow -\text{slope}/(2 \cdot B)$
    **else**
      $\text{DISC} \leftarrow B^2 - 3 \cdot A \cdot \text{slope}$
      **if** $\text{DISC} < 0$ **then**
        $\lambda_{tmp} \leftarrow \sigma_{max} \cdot \lambda$
      **else if** $B \leq 0$ **then**
        $\lambda_{tmp} \leftarrow (-B + \sqrt{\text{DISC}})/(3 \cdot A)$
      **else**
        $\lambda_{tmp} \leftarrow -\text{slope}/(B + \sqrt{\text{DISC}})$
      **end if**
    **end if**
    $\lambda_{tmp} \leftarrow \min(\lambda_{tmp}, \sigma_{max} \cdot \lambda)$
  **end if**
  $\lambda_{old} \leftarrow \lambda$
  $\text{norm}_{fmin\_old} \leftarrow \text{norm}_{fmin}$
  $\lambda \leftarrow \max(\lambda_{tmp}, \sigma_{min} \cdot \lambda)$
  $\tau \leftarrow \tau_{old} + \lambda \cdot \delta\tau$
  $F \leftarrow \text{evalF}(\text{system}, \tau)$
  $\text{norm}_F \leftarrow \|F\|$
  **if** $\text{norm}_F >$ max_value **or** isnan($\text{norm}_F$) **then**
    $\text{norm}_F \leftarrow$ max_value
  **end if**
  $\text{norm}_{fmin} \leftarrow 0.5 \cdot \text{norm}_F^2$
  $n_{armijo\_iter} \leftarrow n_{armijo\_iter} + 1$
**end while**
$\delta\tau \leftarrow \delta\tau \cdot \lambda$

**Méthodes numériques avancées pour les problèmes à forte raideur en transport réactif**

**Résumé**

La simulation du transport réactif en milieu poreux est un enjeu majeur pour la transition énergétique, avec des applications pour la séquestration du CO2, la géothermie et le stockage d'hydrogène. La performance des codes de transport réactif est aujourd'hui fortement limitée par les difficultés numériques liées à la modélisation chimique. Ces difficultés se rattachent à un problème de raideur des équations à résoudre. Dans cette thèse, on s'intéresse aux réactions d'équilibre dans le cas d'une seule phase aqueuse et dans celui d'un mélange multiphasique pouvant contenir des phases aqueuses, gazeuses et minérales. Ces deux problèmes mènent à la résolution d'équations algébriques non linéaires par la méthode de Newton.

Les réactions d'équilibre monophasiques considérées recouvrent une large gamme de valeurs du domaine de définition des inconnues, avec des plages de fonctionnement et des ordres de grandeurs tout à fait différents. Cela pose des problèmes lors de la résolution par la méthode de Newton. Tantôt il est préférable de choisir comme inconnues les nombres de moles des espèces, tantôt il est souhaitable de prendre leurs logarithmes (ou potentiels chimiques), sous peine d'accroître le nombre d'itérations nécessaires voire de faire diverger la méthode de Newton. Les réactions d'équilibre multiphasiques, en plus de contenir les difficultés précédentes, peuvent contenir un nombre important de phase potentiellement présentes. La présence ou l'absence d'une phase est modélisée par un problème de complémentarité. La non différentiabilité des conditions de complémentarité met en défaut la méthode de Newton dans des cas difficiles mais réalistes.

Dans cette thèse, nous avons amélioré la robustesse de la méthode de Newton pour le cas monophasique grâce à une reformulation du système basé sur les fractions molaires et à l'utilisation des méthodes de paramétrage et de représentation Cartésienne. La méthode de paramétrage permet, grâce à une variable fictive, de rendre automatique le choix d'une résolution en fraction molaires ou en potentiels chimiques. La représentation Cartésienne, quant à elle, considère un système élargi où fractions molaires et potentiels chimiques sont des inconnues et où leur relation est relâchée et intégrée aux équations sous la forme d'une fonction non linéaire vérifiant ce lien à convergence. Nous avons démontré que la méthode de Newton appliquée à ces formulations vérifie la propriété de convergence quadratique locale. Les résultats numériques obtenus démontrent une robustesse accrue de nos méthodes comparées à la littérature.

Pour le cas multiphasique, nous avons établi une nouvelle modélisation du problème d'équilibres chimiques. Le système obtenu permet de considérer l'absence ou la présence des phases dans un cadre rigoureux et unifié. Cette unification fait référence à la notion de fractions molaires étendues, il s'agit d'une extension de la notion de fractions molaires aux phases absentes et permet de traiter indifféremment les phases grâce à une condition de complémentarité. Nous avons appliqué les méthodes de paramétrage et de représentation Cartésienne à ce problème ainsi qu'une nouvelle méthode de paramétrage de la complémentarité. Cette dernière a été comparée à différentes approches de la littérature pour le traitement de la complémentarité. Les expériences numériques obtenues ont montré une nette amélioration en termes de robustesse et de rapidité par rapport à l'état de l'art.

**Mots clés :** équilibres chimiques, méthodes de Newton, paramétrage, représentation cartésienne

---

**Centre Inria de l'Université de Lille**
Parc scientifique de la Haute-Borne – 40, avenue Halley – 59650 Villeneuve d'Ascq – France

**Advanced numerical methods for stiff problems in the context of reactive transport**


## Abstract

The simulation of reactive transport in porous media is a major challenge for the energy transition, with applications in CO2 sequestration, geothermal energy and hydrogen storage. Today, the performance of reactive transport codes is severely limited by the numerical difficulties associated with chemical modeling. These difficulties are related to the stiffness of the equations to be solved. In this thesis, we focus on equilibrium reactions in the case of a single aqueous phase and in the case of a multiphase mixture that may contain aqueous, gaseous and mineral phases. Both problems lead to the solution of nonlinear algebraic equations using Newton's method.

The single-phase equilibrium reactions considered cover a wide range of values in the domain of definition of the unknowns, with different operating ranges and orders of magnitude. This poses problems when solving by Newton's method. Sometimes it is preferable to choose the mole numbers of the species as unknowns, sometimes it is desirable to take their logarithms (or chemical potentials), which may increase the number of iterations or cause Newton's method to diverge. Multiphase equilibrium reactions, in addition to containing the above difficulties, may also include a significant number of potentially present phases. The presence or absence of a phase is modeled by a complementarity problem. The non-differentiability of complementarity conditions causes Newton's method to fail in difficult but realistic cases.

In this thesis, we have improved the robustness of Newton's method for the single-phase case by reformulating the system based on mole fractions and using the parameterization and Cartesian representation methods. The parametrization method uses a fictitious variable to automatically select a resolution in mole fractions or chemical potentials. The Cartesian representation, on the other hand, considers an extended system in which mole fractions and chemical potentials are unknowns and their relationship is relaxed and integrated into the equations in the form of a non-linear function verifying this link only at convergence. We have demonstrated that Newton's method applied to these formulations verifies the property of local quadratic convergence. The numerical results obtained demonstrate the increased robustness of our methods compared with the literature.

For the multiphase case, we have established a new model of the chemical equilibrium problem. The resulting system allows us to consider the absence or presence of phases within a rigorous, unified framework. This unification refers to the notion of extended mole fractions, which is an extension of the notion of mole fractions to absent phases, and enables phases to be treated indifferently thanks to a complementarity condition. We have applied Cartesian parametrization and representation methods to this problem, as well as a new approach called the complementarity parametrization method. This method has been compared with various approaches to complementarity treatment from the literature. The numerical experiments obtained has shown a clear improvement in terms of robustness and speed compared with the state of the art.



**Keywords:** chemical equilibria, Newton's method, parametrization, Cartesian representation