Parametrization and Cartesian representation techniques for robust resolution of chemical equilibria

Maxime Jonval, Ibtihel Ben Gharbia, Clément Cancès, Thibault Faney, Quang Huy Tran

Robust resolution of single-phase chemical equilibrium using parametrization and Cartesian representation techniques

Ibtihel Ben Gharbia*, Clément Cancès†, Thibault Faney*,
Maxime Jonval†,*, Quang-Huy Tran*

October 2, 2023

Abstract

Chemical equilibria computations, especially those with vanishing species in aqueous phase, lead to nonlinear systems that are difficult to solve due to blowing up gradients. Instead of the commonly used *ad hoc* treatments, we propose two reformulations of the problem which are in line with the spirit of preconditioning but whose actual aims are to guarantee a better stability of Newton's method. The first reformulation is a parametrization of the graph linking species mole fractions to their chemical potentials. The second is based on an augmented system where this relationship is relaxed for the iterates by means of a Cartesian representation. We theoretically prove the local quadratic convergence of Newton's method for both reformulations. From a numerical point of view, we demonstrate that the two techniques are accurate, allowing to compute equilibria with chemical species having very low concentrations. Moreover, the robustness of the Cartesian representation is superior to that of the literature.

1 Introduction

The simulation of reactive transport is a major issue in various fields: flows in porous media, combustion in engines and gas turbines or the design of chemical reactors for processes. In particular, the computation of reactive transport in porous media plays a central role in CO_2 and H_2 storages or geothermal energy. The performance of current simulators is however limited by the chemical modeling of the problem considered. Most notably, the resolution of nonlinear equations for chemical equilibria is very costly, since it has to be done at each time-step and in each cell of the mesh. In this respect, even the slightest improvement in their resolution may have a direct positive impact on the overall performance.

For chemical modeling, there are mainly two types of reactions: equilibrium reactions and kinetic reactions. The reactions we are interested in are those of equilibrium. Given quantities of chemical elements, a pressure and a temperature, a chemical equilibrium calculation consists in finding the quantities of chemical species minimizing a state function, called Gibbs free energy, and satisfying the conservation of the quantity of matter. This problem involves solving linear equations expressing the conservation of mass as well as non-linear equations related to the chemical reactions involved. The first ones depend on the mole numbers of the species while the second ones are functions of their logarithms. The use of Newton's algorithm for the linearization of these equations encounters a number of difficulties: the iterates can take negative values, which leads to incompatibilities with the logarithm; the values of the solution cover a wide range of values, leading to conditioning problems; the convergence of the algorithm is not ensured if one starts far from the solution. A classical trick consists in using the logarithm of the numbers of moles as unknowns in order to manage the constraint of positivity and to reduce the orders of magnitude between species. However, it is sometimes preferrable to use the number of moles as unknowns. In this article, the parametrization technique developed by Brenner and Cancès [5] is used to automatically switch between the two formulations while ensuring that the partial derivatives of the Jacobian remain bounded, see also [6, 3]. A second approach developed in the article is the well-balanced

*IFP Energies nouvelles, 1 et 4 avenue de Bois Préau, 92852 Rueil-Malmaison Cedex, France. ibtihel.ben-gharbia@ifpen.fr, thibault.faney@ifpen.fr, quang-huy.tran@ifpen.fr

†Inria, Univ. Lille, CNRS, UMR 8524 – Laboratoire Paul Painlevé, 59000 Lille, France. clement.cances@inria.fr, maxime.jonval@inria.fr

Cartesian representation. This method consists in choosing both the numbers of moles and their logarithms as unknowns, a function establishing the relation between these two quantities is then introduced. This function has properties that allow to overcome the problems mentioned above and to control the derivatives of the Jacobian.

1.1 State of the art

Chemical equilibrium was first conceptualized by Berthollet in 1803 [4]. In 1864 Guldberg and Waage [43] defined the law of mass action, which allowed for the calculation of chemical equilibrium. In 1873, Gibbs showed [18] that the minimization of a functional, now known as Gibbs free energy, also enabled this calculation. He demonstrated that the global minimum of this state function is reached for a composition of chemical species at equilibrium. Until the 1940s the calculations of chemical equilibrium involved only a few species and were done analytically [23]. After World War II, Brinkley [7] proposed an algorithm for computer calculation. Storey and Van Zeggeren [42] note that the development of chemical equilibrium resolution methods at this time was mainly motivated by calculation of properties of propellants and rockets motors [37], explosives [11, 41], applications in chemical processing and in the behaviour of multiphase biological cell systems [12]. Smith reviewed the methods of this period [35], and later Smith and Missen [36] proposed a classification of different approaches into two categories: stoichiometric methods and non-stoichiometric methods. Stoichiometric methods are based on the mass action equations while non-stoichiometric methods are based on the minimization of the Gibbs free energy. Our approach belongs to the latter category, according to this classification. As mentioned by Leal et al [26], many computational codes use this kind of method, including ChemSage [17, 16], THERIAK [13, 14], HCh [33, 34], FactSage [2, 1], PERPLEX [10, 9], GEM-Selektor [20, 21, 19, 24, 44] and Reaktoro [25]. Subsequent developments in this field have been reviewed by Leal et al. [26], Tsanas et al. [40, 39], and Coatléven and Michel [8].

1.2 Outline

Section 2 presents the mathematical modeling of the chemical equilibrium problem. In section 3, the mathematical details of the parametrization and Cartesian representation techniques are presented. A link between these two approaches is also established. Section 4 presents different results concerning the invertibility of the Jacobian close to convergence, ensuring the local quadratic convergence of Newton's method. In section 5 are presented the results of numerical experiments validating our methods and comparing their robustness. Section 6 concludes and opens to future works.

2 Mathematical description of the chemical equilibrium problem

This section is devoted to the presentation of the chemical equilibrium problem and the equations derived from it.

2.1 Chemical system

The type of system considered in this article are diluted solutions of aqueous species, they are composed of a strongly majority species called solvent, typically water, as well as diluted aqueous species that are present in very small quantities. For a given fixed temperature T and pressure P, such a chemical system $\mathcal{S} := \mathcal{S}_{\mathrm{P,T}} = \{\mathcal{C}, \mathcal{E}, \mathcal{R}\}$ is a collection of three sets:

- a set of N chemical species $\mathcal{C} = (C_1, \ldots, C_N)$;
- a set of M chemical elements $\mathcal{E} = (E_1, \ldots, E_M)$, $M < N$;
- and a set of $N - M$ chemical reactions $\mathcal{R} = (R_1, \ldots, R_{N-M})$.

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 can be written as

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

where s_{ij} is the stoichiometric coefficients that represents the number of molecules of the species C_i involved in the reaction R_j.

The systems we are studying are closed, so there is 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 elements H and O, and of species $\mathrm{H^+}$, $\mathrm{OH^-}$ and $\mathrm{H_2O}$ verifying the equilibrium reaction

$$\mathrm{H_2O} = \mathrm{H^+} + \mathrm{OH^-}.$$

The corresponding formula matrix is

$$\mathbf{A} = \begin{array}{c} \begin{array}{ccc} \mathrm{H^+} & \mathrm{OH^-} & \mathrm{H_2O} \end{array} \\ \left[\begin{array}{ccc} 1 & 1 & 2 \\ 0 & 1 & 1 \end{array}\right] \begin{array}{c} \mathrm{H} \\ \mathrm{O} \end{array} \end{array}$$

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

$$\mathbf{An} = \mathbf{b}.$$

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

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

It is composed of the stoichiometry coefficients involved in the chemical reactions of $\mathcal{R}$ with $\mathbf{S}_{ij} = s_{ij}$.

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

Lemma 2.1. *One has the following result:*

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

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 $\mathbf{S}$:

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

It follows that $\mathrm{Im}\ \mathbf{S} = \ker \mathbf{A}$, then $(\mathrm{Im}\ \mathbf{S})^\perp = (\ker \mathbf{A})^\perp$. The result is obtained using the property $\ker \mathbf{S}^T = (\mathrm{Im}\ \mathbf{S})^\perp$ and $(\ker \mathbf{A})^\perp = \mathrm{Im}\ \mathbf{A}^T$ 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 [36].

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

Lemma 2.2. *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 $a_{ki} n_i = 0, \forall i, \forall k$. 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 $n_i = 0$, $\forall i$. □

2.2 Gibbs free energy and chemical potentials

The state of a closed system $\mathcal{S}$ 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. This function is extensive with respect to the number of moles, meaning that it is a homogeneous function of degree 1. Its standard expression for the study of chemical equilibrium is as follows:

$$G(\mathbf{n}) = \sum_{i=1}^N n_i \frac{\partial G(\mathbf{n})}{\partial n_i} = \sum_{i=1}^N n_i \mu_i(\mathbf{n}), \tag{2}$$

where $\mu_i(\mathbf{n}) = \partial G(\mathbf{n}) / \partial n_i$ 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 an aqueous species C_i, we consider a chemical potential of the form

$$\mu_i := \mu_i(\mathbf{n}) = \mu_i^\circ(P, T) + RT \ln a_i(\mathbf{n}). \tag{3}$$

In (3), $\mu_i^\circ(P, T)$ is the chemical potential of the species C_i 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 on the concentration of all the species.

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

$$x_i := x_i(\mathbf{n}) = n_i / \textstyle\sum_{j=1}^N n_j = n_i / \langle \mathbf{n}, \mathbf{1} \rangle.$$

There are several, increasingly complex activity models for γ_i in the scientific literature [28, 45], the most simple of which being the ideal activity model $\gamma_i = 1$. It corresponds to a theoretical ideal solution where the mean strength of inter-molecular interactions are the same between all the molecules of the system. The activity in (3) is then reduced to the mole fraction. The resulting ideal Gibbs energy

$$G(\mathbf{n}) = \sum_{i=1}^N n_i [\mu_i^0 + RT \ln x_i(\mathbf{n})]$$

is a convex function on $\mathbb{R}^N_+$ (see [36]).

2.3 Equilibrium equations

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 $\mathcal{S}$ which minimizes, for a fixed temperature T, pressure P and element quantities $\mathbf{b}$, the function G, under constraints of element conservation and nonnegativity. This calculation is often referred to as a speciation and is written as:

$$\min_{\mathbf{An}=\mathbf{b},\, \mathbf{n} \geq 0} G(\mathbf{n}). \tag{4}$$

The existence and uniqueness of a solution to the chemical equilibrium problem (4) for a multi-phase ideal system has been studied by Shapiro and Shapley in [32], in particular they provide a proof for the single-phase ideal problem in their Corollary 12.3. Furthermore, let us prove that the inequality constraint is never saturated at the solution. Let

$$\Omega := \{\mathbf{n} \in \mathbb{R}^N \mid n_i > 0,\ i = 1, \dots, N\}$$

be the set of positive vectors of $\mathbb{R}^N$, one defines the set of vectors verifying the constraints of conservation of elements and positivity by

$$\mathcal{M}_{\mathbf{A},\mathbf{b}} := \{\mathbf{n} \in \Omega \mid \mathbf{A}\mathbf{n} = \mathbf{b}\}.$$

One assumes that $\mathcal{M}_{\mathbf{A},\mathbf{b}} \neq \emptyset$. It is a necessary condition to the existence of a minimizer of G in $\mathcal{M}_{\mathbf{A},\mathbf{b}}$. The set $\mathcal{M}_{\mathbf{A},\mathbf{b}}$ is convex and its closure $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ is bounded.

Lemma 2.3. *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 Ω and $\{\mathbf{n} \in \mathbb{R}^N \mid \mathbf{A}\mathbf{n} = \mathbf{b}\}$. We will know demonstrate that the set $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ is bounded. Indeed one has

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

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

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

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 2.4. *The minimum of G obtains from (4) is in $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}} = \{\mathbf{n} \in \overline{\Omega} \,|\, \mathbf{A}\mathbf{n} = \mathbf{b}\}$.*

Proof. From Lemma 2.3, the set $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ in 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 on $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$. □

Lemma 2.5. *If $\mathbf{n}^\star \in \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ minimises G on $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$, then $\mathbf{n}^\star \in \mathcal{M}_{\mathbf{A},\mathbf{b}}$.*

Proof. We know from Lemma 2.4 that there exists $\mathbf{n}^\star \in \{\arg\min G(\mathbf{n}) \mid \mathbf{n} \in \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}\}$. 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 $j \in \{1, \ldots, N\}$ such that $n_j^\star = 0$. Let $\mathbf{n} \in \mathcal{M}_{\mathbf{A},\mathbf{b}}$ and $\varepsilon \in (0, 1)$, then one defines $\mathbf{n}^0 := \mathbf{n} - \mathbf{n}^\star$ and $\mathbf{n}^\varepsilon := \mathbf{n}^\star + \varepsilon \mathbf{n}^0$. 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 2.3, 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 $\overline{\Omega}$,

$$G(\mathbf{n}^\star) \geq G(\mathbf{n}^\varepsilon) + \langle \boldsymbol{\mu}(\mathbf{n}^\varepsilon), \mathbf{n}^\star - \mathbf{n}^\varepsilon \rangle \Leftrightarrow \frac{G(\mathbf{n}^\star) - G(\mathbf{n}^\varepsilon)}{\varepsilon} \geq -\langle \boldsymbol{\mu}(\mathbf{n}^\varepsilon), \mathbf{n}^0 \rangle, \tag{5}$$

where $\boldsymbol{\mu}(\mathbf{n}^\varepsilon) := (\mu_i^\circ + RT \ln x_i^\varepsilon)_{i=1,\ldots,N}$.

We will now take the limit when ε tends to 0 in the inequality (5). In the right-hand side one has:

$$\lim_{\varepsilon \to 0} -\langle \boldsymbol{\mu}(\mathbf{n}^\varepsilon), \mathbf{n}^0 \rangle = -n_j^0 \lim_{\varepsilon \to 0} \mu_j(\mathbf{n}^\varepsilon) - \sum_{i=1,\, i \neq j}^{N} n_i^0 \lim_{\varepsilon \to 0} \mu_i(\mathbf{n}^\varepsilon) \tag{6}$$

Noting that $\lim_{\epsilon \to 0} \mathbf{n}^\varepsilon = \mathbf{n}^\star$ and in particular that $\lim_{\varepsilon \to 0} n_j^\varepsilon = n_j^\star = 0$, it follows from the continuity of μ_j that

$$\lim_{\varepsilon \to 0} \mu_j(\mathbf{n}^\varepsilon) = -\infty \quad \text{and} \quad \lim_{\varepsilon \to 0} \mu_{i \neq j}(\mathbf{n}^\varepsilon) = \mu_i(\mathbf{n}^\star) \in \mathbb{R}. \tag{7}$$

By combining (6) and (7) with $n_j^0 = n_j > 0$, one finds that the right-hand side of (5) tends to $+\infty$. However if $\mathbf{n}^\star$ minimises G on $\mathcal{M}_{\mathbf{A},\mathbf{b}}$ then the left-hand side of (5) is non-positive which is a contradiction. Therefore $\mathbf{n}_j^\star > 0$ and $\mathbf{n}^\star \in \mathcal{M}_{\mathbf{A},\mathbf{b}}$. □

Thanks to Lemma 2.5, the problem can be simplified to

$$\min_{\mathbf{A}\mathbf{n} = \mathbf{b}} G(\mathbf{n}). \tag{8}$$

The first order optimality conditions are given by the Euler-Lagrange equations which state that if $\mathbf{n}^\star$ is the unique solution of the problem (8), it must satisfy

$$\mathbf{A}\mathbf{n}^\star - \mathbf{b} = \mathbf{0}, \tag{9}$$

$$\boldsymbol{\nabla} G(\mathbf{n}^\star) + \mathbf{A}^T \boldsymbol{\Lambda} = \mathbf{0}, \tag{10}$$

where $\boldsymbol{\nabla} G$ is the gradient of G and $\boldsymbol{\Lambda} = (\lambda_1, \dots, \lambda_M)^T$ is the Lagrange multiplier vector, also known as dual variables. These equations imply that $\mathbf{n}^\star$ is a critical point of the Lagrangian function

$$\mathcal{L}(n, \boldsymbol{\Lambda}) = G(\mathbf{n}) + \langle \mathbf{A}\mathbf{n} - \mathbf{b}, \boldsymbol{\Lambda} \rangle.$$

associated to the problem (8). Indeed, (9) corresponds to $\boldsymbol{\nabla}_{\boldsymbol{\Lambda}} \mathcal{L} = \mathbf{0}$, while (10) corresponds to $\boldsymbol{\nabla}_{\mathbf{n}} \mathcal{L} = \mathbf{0}$, where the subscripts $\boldsymbol{\Lambda}$ (or $\mathbf{n}$) under $\boldsymbol{\nabla}$ means that only the part of the gradient $\boldsymbol{\nabla}\mathcal{L}$ corresponding to the derivatives according to $\boldsymbol{\Lambda}$ (or $\mathbf{n}$) is taken.

In the case we are considering, we can simplify the Euler-Lagrange equations by eliminating the dual variables. To do so, we multiply (10) by $\mathbf{S}^T$. As shown by Lemma 2.1, the matrix product $\mathbf{S}^T\mathbf{A}^T$ vanishes. Thus, the equations become $\mathbf{S}^T \boldsymbol{\nabla} G(\mathbf{n}^\star) = \mathbf{0}$. Therefore, denoting by $\boldsymbol{\mu} = \boldsymbol{\nabla} G$ the vector of chemical potentials, the system (9)-(10) can be written as

$$\begin{aligned} \mathbf{A}\mathbf{n}^\star &= \mathbf{b}, \\ \mathbf{S}^T \boldsymbol{\mu}(\mathbf{n}^\star) &= \mathbf{0}. \end{aligned} \tag{11}$$

Proposition 2.1. *The system* (11) *admits a unique solution, which coincides with the solution to the problem* (8)*.*

Proof. The existence of a solution to (11) is guaranteed by the existence of a solution to (8). Furthermore, this solution is unique, it remains to show that it is the only one to satisfy the (11). To do so, assume the existence of $\mathbf{n}_1^\star, \mathbf{n}_2^\star \in \Omega$ that satisfy (11). Then, one has $\mathbf{n}_1^\star - \mathbf{n}_2^\star \in \ker \mathbf{A}$ and $\boldsymbol{\mu}(\mathbf{n}_1^\star) - \boldsymbol{\mu}(\mathbf{n}_2^\star) \in \ker \mathbf{S}^T$. By Lemma 2.1, we know that $\ker \mathbf{S}^T = (\ker \mathbf{A})^\perp$, it follows that

$$\langle \mathbf{n}_1^\star - \mathbf{n}_2^\star, \boldsymbol{\mu}(\mathbf{n}_1^\star) - \boldsymbol{\mu}(\mathbf{n}_2^\star) \rangle = 0.$$

Therefore, by the strict monotonicity of the gradient of G inherited from its strict convexity on $\mathcal{M}_{\mathbf{A},\mathbf{b}}$, it follows that $\mathbf{n}_1^\star = \mathbf{n}_2^\star$. □

2.4 Reformulation of the system in terms of mole fractions

To reduce the strong nonlinearities in the expression of chemical potentials in (11), it is interesting to introduce the following new variable:

$$\omega := 1/\sum_{i=1}^{N} n_i. \tag{12}$$

Then, multiplying the element conservation equations by ω 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 ω. Furthermore, since there is only N equations, the addition of one more equation is needed. A fundamental property of the mole fractions is that $\sum_{i=1}^{N} x_i = 1$ which can be the additional equation. Thus the problem to solve becomes : find $(\mathbf{x}, \omega)$ such that

$$\begin{aligned} \mathbf{A}\mathbf{x} - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T \mathbf{y}(\mathbf{x}) &= \mathbf{d}, \\ \langle \mathbf{x}, \mathbf{1} \rangle &= 1, \end{aligned} \tag{13}$$

where

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

Proposition 2.2. *The system* (13) *is equivalent to the system* (11) *and its solution is unique.*

Proof. Let $\mathbf{n}^\star$ be the unique solution of (11) and let $\omega = 1/\sum_{i=1}^{N} n_i^\star > 0$. Then by construction it is clear that $(\omega\mathbf{n}^\star, \omega)$ solves (13). In particular, this ensures the existence of a point $(\mathbf{x}, \omega)$ verifying (13). Moreover, if $(\mathbf{x}, \omega)$ solves (13), then $\omega \neq 0$. Indeed, if this were not the case, then $\mathbf{x}$ would belong to $\ker \mathbf{A}$, implying from Lemma 2.2 that its coefficients are not all of the same sign, which is not compatible with the logarithm. Thus $\mathbf{n} = \mathbf{x}/\omega$ verifies (11). Now suppose there are $(\mathbf{x}^1, \omega^1)$ and $(\mathbf{x}^2, \omega^2)$ satisfying (13). Then by uniqueness $\mathbf{n}^\star = \mathbf{x}^1/\omega^1 = \mathbf{x}^2/\omega^2$. It follows that

$$\frac{1}{\sum_{i=1}^{N} n_i^\star} = \frac{1}{\sum_{i=1}^{N} x_i^1/\omega^1} = \omega^1 \quad \text{and} \quad \frac{1}{\sum_{i=1}^{N} n_i^\star} = \frac{1}{\sum_{i=1}^{N} x_i^2/\omega^2} = \omega^2,$$

meaning that $\omega^1 = \omega^2$. Therefore $\mathbf{x}^1 = \mathbf{x}^2$. □

2.5 Other types of constraint considered

System (13) 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 [45], but the ones we will use are charge conservation and the redox constraint.

2.5.1 Charge conservation constraint

When defining the matrix formula $\mathbf{A}$, 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. The matrix formula for water dissociation becomes

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

It is also necessary to adapt the coefficient of the vector $\mathbf{b}$ corresponding to the charge.

2.5.2 Redox constraint

In the case of oxidation-reduction reactions, the system (13) can be modified by introducing the notion of electron potential [38]. This potential, denoted pE, is written as

$$\mathrm{pE} = -\log_{10}(a_{e^-}), \tag{14}$$

where a_{e^-} is the electron activity. In (14), the pE value is set by the user, so it is necessary to define the notion of electron activity. To do this, we consider the electron chemical potential:

$$\mu_{e^-} = \mu^\circ_{e^-} + RT \ln a_{e^-}, \tag{15}$$

where $\mu^\circ_{e^-}$ is a standard chemical potential for the electron to be computed from a thermodynamic database. Thus, by integrating (14) into (15), it follows that

$$\mu_{e^-} = \mu^\circ_{e^-} - \mathrm{pE} \times RT \ln 10. \tag{16}$$

To take account of this constraint on the electron potential, we need to consider the electron as a fictitious secondary species and introduce a half-reaction involving species present in our system. As an example, let us consider the following chemical system:

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

associated to the half reaction

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

For this kind of system, the number of chemical elements involved is different from the number of primary species. However, it is possible to define the formula and stoichiometric matrices using charge conservation. In addition, the electron is introduced as a virtual secondary species, resulting in the creation of an associated secondary matrix. We thus define the matrices

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

and

$$\mathbf{A} = [A_{Pr}, A_{Sd}], \quad \mathbf{S} = \begin{bmatrix} A_{Pr}^{-1} A_{Sd} \\ -I_{Sd} \end{bmatrix}, \quad \mathbf{S}_{\mathrm{pE}} = \begin{bmatrix} A_{Pr}^{-1} A^{\mathrm{pE}}_{Sd} \\ -I^{\mathrm{pE}}_{Sd} \end{bmatrix}.$$

The system to solve is written as

$$\begin{aligned}\tilde{\mathbf{A}}\mathbf{x} - \omega\mathbf{b} &= \mathbf{0},\\ \mathbf{S}^T\mathbf{y}(\mathbf{x}) &= \mathbf{d},\\ \mathbf{S}_{\mathrm{pE}}^T\begin{bmatrix}\mathbf{y}(\mathbf{x}_{Pr})\\ \mu_{e^-}\end{bmatrix} &= \mathbf{d}_{\mathrm{pE}},\\ \langle\mathbf{x},\mathbf{1}\rangle &= 1,\end{aligned}$$

with

$$\mu_{e^-} := \mu_{e^-}^\circ/(RT) - \mathrm{pE}\times\ln 10 \quad \text{and} \quad \mathbf{d}_{\mathrm{pE}} := -\mathbf{S}_{\mathrm{pE}}^T\begin{bmatrix}\boldsymbol{\mu}_{Pr}^\circ/(RT)\\ 0\end{bmatrix}$$

where $\tilde{\mathbf{A}}$ is obtained by deleting either the Z line or the H line, depending on the quantity we want to conserve.

3 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 Newton's method

There are many methods to solve the nonlinear system of equations (13) as detailed in [29], 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. Let us recall the considered system: find $(\mathbf{x},\omega)\in\mathbb{R}^{N+1}$ such that

$$\begin{aligned}\mathbf{A}\mathbf{x} - \omega\mathbf{b} &= \mathbf{0},\\ \mathbf{S}^T\mathbf{y}(\mathbf{x}) &= \mathbf{d},\\ \langle\mathbf{x},\mathbf{1}\rangle &= 1,\end{aligned} \tag{17}$$

where $(\mathbf{y}(\mathbf{x}))_i = y(x_i) = \ln x_i$. The resolution of the system (17) can be viewed as the search for the zeros of a function $\mathcal{G}:\mathbb{R}^{N+1}\to\mathbb{R}^{N+1}$, associated to a function $\mathcal{F}:\mathbb{R}^{2N+1}\to\mathbb{R}^{N+1}$, which are defined as follows:

$$\mathcal{G}(\mathbf{x},\omega) := \mathcal{F}(\mathbf{x},\mathbf{y}(\mathbf{x}),\omega) = \begin{pmatrix}\mathbf{A}\mathbf{x} - \omega\mathbf{b}\\ \mathbf{S}^T\mathbf{y}(\mathbf{x}) - \mathbf{d}\\ \langle\mathbf{x},\mathbf{1}\rangle - 1\end{pmatrix}. \tag{18}$$

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

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

$$\boldsymbol{\nabla}\mathcal{G}(\mathbf{u}^{(k)})\boldsymbol{\delta}\mathbf{u}^{(k)} = -\mathcal{G}(\mathbf{u}^{(k)}), \tag{19}$$

to compute the Newton increment $\boldsymbol{\delta}\mathbf{u}^{(k)}$ used to update the sequence as

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \boldsymbol{\delta}\mathbf{u}^{(k)}. \tag{20}$$

In (19), $\boldsymbol{\nabla}\mathcal{G}(\mathbf{u}^{(k)})$ stands for the Jacobian matrix of $\mathcal{G}$ evaluated at $\mathbf{u}^{(k)}$. An important result about Newton's method concerns its local quadratic convergence [22]. It requires the following assumptions:

1. The equation (18) has a solution $\mathbf{u}^\star$.
2. $\boldsymbol{\nabla}\mathcal{G}:\mathbb{R}^N\to\mathbb{R}^{N\times N}$ is Lipschitz continuous near $\mathbf{u}^\star$: there exists a neighborhood $\mathcal{V}$ of $\mathbf{u}^\star$ and $L>0$ such that
$$\|\boldsymbol{\nabla}\mathcal{G}(\mathbf{u}_1) - \boldsymbol{\nabla}\mathcal{G}(\mathbf{u}_2)\|_2 \le L\|\mathbf{u}_1 - \mathbf{u}_2\|_2$$
for all $\mathbf{u}_1,\mathbf{u}_2$ in $\mathcal{V}$.
3. $\boldsymbol{\nabla}\mathcal{G}(\mathbf{u}^\star)$ is nonsingular, *i.e.* invertible.

The local quadratic convergence theorem is as follows [22, Theorem 1.1].

Theorem 3.1. *Let the previous assumptions hold. If* $\mathbf{u}^{(0)}$ *is sufficiently close to* $\mathbf{u}^\star$*, then Newton's sequence* (19)–(20) *is well defined for all* $k \geq 0$ *and converges to* $\mathbf{u}^\star$*. Moreover, there exist* $C > 0$ *and* $k_C \in \mathbb{N}$ *such that*

$$\|\mathbf{u}^{(k+1)} - \mathbf{u}^\star\|_2 \leq C\|\mathbf{u}^{(k)} - \mathbf{u}^\star\|_2^2, \ \forall k \geq k_C. \tag{21}$$

The property (21) together with $\mathbf{u}^{(k)} \to \mathbf{u}^\star$ is referred to as q-quadratic convergence in the monograph [22].

3.2 A family of parametrizations

The Newton's method applied to the function (18) yields the following Jacobian matrix:

$$\boldsymbol{\nabla}\mathcal{G}(\mathbf{x},\omega) = \begin{bmatrix} \mathbf{A} & -\mathbf{b} \\ \mathbf{S}^T\boldsymbol{\nabla}\mathbf{y}(\mathbf{x}) & \mathbf{0} \\ \mathbf{1}^T & 0 \end{bmatrix}, \tag{22}$$

where $\boldsymbol{\nabla}\mathbf{y}(\mathbf{x}) = \operatorname{diag}\{1/x_i\}_{i=1,\dots,N}$. The jacobian in (22) 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. A classic cure to these problems is to consider $y_i = y(x_i)$ as the unknowns and to define $\mathcal{H} : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ as follows

$$\mathcal{H}(\mathbf{y},\omega) := \mathcal{F}(\mathbf{x}(\mathbf{y}),\mathbf{y},\omega) = \mathbf{0} \Leftrightarrow \begin{aligned} \mathbf{A}\mathbf{x}(\mathbf{y}) - \mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T\mathbf{y} - \mathbf{d} &= \mathbf{0}, \\ \langle \mathbf{x}(\mathbf{y}), \mathbf{1}\rangle - 1 &= 0, \end{aligned} \tag{23}$$

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 (18). In this case the Jacobian matrix becomes

$$\boldsymbol{\nabla}\mathcal{H}(\mathbf{y},\omega) = \begin{bmatrix} \mathbf{A}\boldsymbol{\nabla}\mathbf{x}(\mathbf{y}) & -\mathbf{b} \\ \mathbf{S}^T & \mathbf{0} \\ \mathbf{1}^T\boldsymbol{\nabla}\mathbf{x}(\mathbf{y}) & 0 \end{bmatrix}, \tag{24}$$

where $\boldsymbol{\nabla}\mathbf{x}(\mathbf{y}) = \operatorname{diag}\{\exp y_i\}_{i=1,\dots,N}$. The Jacobian in (24) diverges when y_i tends to the infinity, and numerical issues can appear already for moderate positive values of y_i. However the positivity constraint on the iterates is not necessary anymore.

The formulation in y is better behaved than the one in x, but it is possible to do even better with the parametrization. The idea of parametrization is to make the best of both formulations while ensuring that the values of the coefficients of the system's Jacobian are controlled. For this purpose, the graph

$$\Gamma = \{(x,y) \in \mathbb{R}^2 \mid y = \ln(x)\} \tag{25}$$

will be 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 $y_i = Y(\tau_i)$ and in such a way that $\Gamma = (X,Y)(\mathbb{R})$. The problem to solve becomes: find $(\boldsymbol{\tau},\omega) \in \mathbb{R}^{N+1}$ such that

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

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:

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

where $\mathbf{X}'(\boldsymbol{\tau}) = (X'(\tau_i))_{i=1,\dots,N}$ and $\mathbf{Y}'(\boldsymbol{\tau}) = (Y'(\tau_i))_{i=1,\dots,N}$. Note that the parametrization is a non-linear right preconditioning, indeed

$$\mathfrak{F}(\boldsymbol{\tau},\omega) = \mathbf{0} \Leftrightarrow \begin{aligned} \mathcal{G}(\mathbf{X}(\boldsymbol{\tau}),\omega) &= \mathbf{0} \\ \text{or} & \\ \mathcal{H}(\mathbf{Y}(\boldsymbol{\tau}),\omega) &= \mathbf{0}. \end{aligned}$$

We will now introduce the conditions that enable us to control the coefficients of the Jacobian. For the problem we are considering, the $\boldsymbol{\nabla}_{\mathbf{X}(\boldsymbol{\tau})}\mathcal{F}$ and $\boldsymbol{\nabla}_{\mathbf{Y}(\boldsymbol{\tau})}\mathcal{F}$ terms do not depend on $\boldsymbol{\tau}$, so the Jacobian is bounded if the $\mathbf{X}'(\boldsymbol{\tau})$ and $\mathbf{Y}'(\boldsymbol{\tau})$ terms are. Moreover, if any of $X'(\tau)$ and $Y'(\tau)$ vanish for the same value of τ, then the corresponding column in the Jacobian will be zero and the Jacobian will become singular. To ensure correct parametrization, we need to satisfy the following conditions, for each $\tau \in \mathbb{R}$:

(A1) $Y(\tau) = \ln(X(\tau))$;

(A2) X' and Y' are strictly monotonic bounded Lipschitz continuous functions;

(A3) $X'(\tau)$ and $Y'(\tau)$ do not vanish for the same value of τ.

We then say that a parametrization is admissible if it satisfies conditions (A1)–(A3). To ensure that conditions (A2) and (A3) are satisfied, we introduce the following normalization condition on the derivatives:

$$(|X'(\tau)|^p + |Y'(\tau)|^p)^{1/p} = 1, \quad p \geq 1. \tag{26}$$

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

$$Y'(\tau) = X'(\tau)/X(\tau) \tag{27}$$

from the condition (A1). Indeed, combining (26) and (27), we get:

$$|X'(\tau)|^p + |X'(\tau)/X(\tau)|^p = 1 \Leftrightarrow |X'(\tau)| = \frac{1}{(1 + |1/X(\tau)|^p)^{1/p}} \tag{28}$$

and

$$|Y'(\tau)|^p = 1 - |X'(\tau)|^p \Leftrightarrow |Y'(\tau)| = \frac{1/X(\tau)}{(1 + |1/X(\tau)|^p)^{1/p}}. \tag{29}$$

Furthermore, equation (29) can be expressed in terms of the function Y. To do this, we multiply (27) by $\exp(Y(\tau))$ and using the derivative

$$\exp(Y(\tau))Y'(\tau) = X'(\tau)$$

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

$$X'(\tau) = \exp(Y(\tau))X'(\tau)/X(\tau) \Leftrightarrow \exp'(Y(\tau))/X(\tau) = 1. \tag{30}$$

Thus, the equations (28), (29) and (30) enable us to express the following system of differential equations:

$$X'(\tau) = \pm\frac{1}{(1 + |1/X(\tau)|^p)^{1/p}}, \tag{31}$$

$$Y'(\tau) = \pm\frac{1/X(\tau)}{(1 + |1/X(\tau)|^p)^{1/p}} = \pm\frac{1}{(1 + |\exp(Y(\tau))|^p)^{1/p}}. \tag{32}$$

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 that of the limit $p \to \infty$, the condition (26) then becomes

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

and the system (31)-(32) is rewritten as

$$X'(\tau) = \pm\frac{1}{\max(1, |1/X(\tau)|}, \tag{33}$$

$$Y'(\tau) = \pm\frac{1/X(\tau)}{\max(1, |1/X(\tau)|)} = \pm\frac{1}{\max(1, |\exp(Y(\tau))|)}. \tag{34}$$

Since the logarithm function is increasing and strictly concave, a solution of this latter system is given by the following statement, *cf.* [5].

Proposition 3.1. *A solution of* (33)-(34) *is given by*

$$(X(\tau), Y(\tau)) = \begin{cases} (\exp(\tau),\ \tau) & \text{if} \quad \tau < 0, \\ (\tau + 1,\ \ln(\tau + 1)) & \text{if} \quad \tau \geq 0. \end{cases} \tag{35}$$

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 [15]. Figure 1 illustrates these functions.

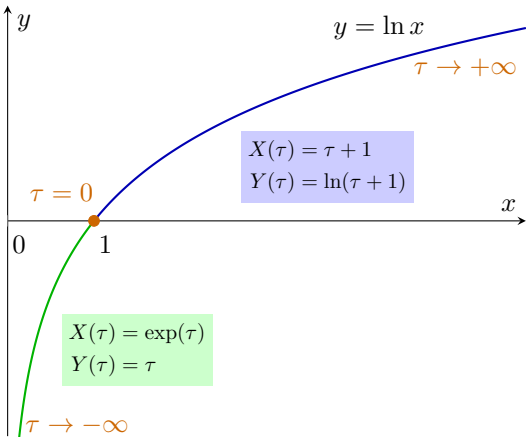

Figure 1: The switch function.

3.3 A family of Cartesian representations

The Cartesian representation technique is based on an augmented system where the relation $y = \ln(x)$, or the equivalent $\exp(y) = x$, is relaxed. The resolution is then on $(\mathbf{x}, \mathbf{y}, \omega)$ and the systems are written as

$$\begin{aligned} \mathcal{F}(\mathbf{x}, \mathbf{y}, \omega) &= \mathbf{0}, \\ \mathbf{y} - \ln(\mathbf{x}) &= \mathbf{0}, \\ \langle \mathbf{x}, \mathbf{1} \rangle - 1 &= \mathbf{0} \end{aligned} \quad \text{or} \quad \begin{aligned} \mathcal{F}(\mathbf{x}, \mathbf{y}, \omega) &= \mathbf{0}, \\ \exp(\mathbf{y}) - \mathbf{x} &= \mathbf{0}, \\ \langle \mathbf{x}, \mathbf{1} \rangle - 1 &= \mathbf{0}. \end{aligned}$$

The corresponding Jacobian matrices are

$$\begin{bmatrix} \mathbf{A} & \mathbf{0} & -\mathbf{b} \\ \mathbf{0} & \mathbf{S}^T & \mathbf{0} \\ -\boldsymbol{\nabla}\boldsymbol{\varphi}(\mathbf{u}) & \mathbf{I} & \mathbf{0} \\ \mathbf{1}^T & \mathbf{0} & 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \mathbf{A} & \mathbf{0} & -\mathbf{b} \\ \mathbf{0} & \mathbf{S}^T & \mathbf{0} \\ -\mathbf{I} & \boldsymbol{\nabla}\boldsymbol{\psi}(\mathbf{v}) & \mathbf{0} \\ \mathbf{1}^T & \mathbf{0} & 0 \end{bmatrix}.$$

These matrices present problems of the same nature as the matrices $\boldsymbol{\nabla}\mathcal{G}(\mathbf{x}, \omega)$ and $\boldsymbol{\nabla}\mathcal{H}(\mathbf{y}, \omega)$ respectively. To tackle these issues, the idea of Cartesian representation is to introduce two Lipschitz continuous functions $H : \mathbb{R} \to \mathbb{R}$ and $G : \mathbb{R} \to \mathbb{R}$ such that $G = H \circ \ln$, and to rewrite the system as

$$\begin{aligned} \mathcal{F}(\mathbf{x}, \mathbf{y}) &= \mathbf{0}, \\ \mathbf{H}(\mathbf{y}) - \mathbf{G}(\mathbf{x}) &= \mathbf{0}, \end{aligned}$$

where $\mathbf{H}(\mathbf{y}) = (H(y_i))_{i=1,\dots,N}$ and $\mathbf{G}(\mathbf{x}) = (G(x_i))_{i=1,\dots,N}$. The aim of this technique is to control the partial derivatives of the function

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

The problem to solve becomes: find $(\mathbf{x}, \mathbf{y}, \omega) \in \mathbb{R}^{2N+1}$ such that

$$\mathfrak{G}(\mathbf{x}, \mathbf{y}, \omega) := \begin{bmatrix} \mathcal{F}(\mathbf{x}, \mathbf{y}, \omega) \\ \boldsymbol{f}(\mathbf{x}, \mathbf{y}) \\ \langle \mathbf{x}, \mathbf{1} \rangle - 1 \end{bmatrix} = \mathbf{0} \Leftrightarrow \begin{aligned} \mathbf{A}\mathbf{x} - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T\mathbf{y} - \mathbf{d} &= \mathbf{0}, \\ \mathbf{H}(\mathbf{y}) - \mathbf{G}(\mathbf{x}) &= \mathbf{0}, \\ \langle \mathbf{x}, \mathbf{1} \rangle - 1 &= 0 \end{aligned}$$

The associated Jacobian matrix is written as follows:

$$\boldsymbol{\nabla}\mathfrak{G}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \mathbf{A} & \mathbf{0} & -\mathbf{b} \\ \mathbf{0} & \mathbf{S}^T & \mathbf{0} \\ \boldsymbol{\nabla}_{\mathbf{x}}\boldsymbol{f} & \boldsymbol{\nabla}_{\mathbf{y}}\boldsymbol{f} & \mathbf{0} \\ \mathbf{1}^T & \mathbf{0} & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{0} & \\ \mathbf{0} & \mathbf{S}^T & \\ -\mathrm{diag}\{\mathbf{G}'(\mathbf{x})\} & \mathrm{diag}\{\mathbf{H}'(\mathbf{y})\} & \\ \mathbf{1}^T & \mathbf{0} & 0 \end{bmatrix},$$

where $\mathbf{G}'(\mathbf{x}) = (G'(x_i))_{i=1,\dots,N}$ and $\mathbf{H}'(\mathbf{y}) = (H'(y_i))_{i=1,\dots,N}$. To avoid the problems mentioned above and to ensure that the coefficients of the Jacobian are bounded, we wish to satisfy the following conditions on the H and G functions, $\forall x, y \in \mathbb{R}$:

(H1) $G(x) = H(y)$ if and only if $y = \ln(x)$;

(H2) $\partial_x f = -G'(x)$ and $\partial_y f = H'(y)$ are strictly monotonic bounded Lipschitz continuous functions;

(H3) $\partial_x f = -G'(x)$ and $\partial_y f = H'(y)$ do not vanish for the same value of τ.

We then say that a Cartesian representation is admissible if it satisfies conditions (H1)–(H3). As for the parametrization, we introduce a normalization condition that takes the following form:

$$(|H'(y)|^p + |G'(x)|^p)^{1/p} = 1, \quad p \geq 1,\ y = \ln(x). \tag{36}$$

Using the same reasoning as we did for the parametrization, we can combine equations (36) with the derivative

$$G'(x) = H'(\ln(x))/x, \tag{37}$$

from $G(x) = H \circ \ln(x)$, to obtain a system of differential equations:

$$G'(x) = \pm \frac{1/x}{(1 + |1/x|^p)^{1/p}}, \tag{38}$$

$$H'(v) = \pm \frac{1}{(1 + |1/\exp(y)|^p)^{1/p}} = \pm \frac{\exp(y)}{(1 + |\exp(y)|^p)^{1/p}}. \tag{39}$$

The case of interest for numerical experiments is that of the limit $p \to \infty$, the condition (36) then becomes

$$\max(|H'(x)|, |G'(y)|) = 1, \quad ,$$

while the differential equations (38)–(39) become

$$G'(x) = \pm \frac{1/x}{\max(1, |1/x|)}, \tag{40}$$

$$H'(y) = \pm \frac{1}{\max(1, |1/\exp(v)|)} = \pm \frac{\exp(v)}{\max(1, |\exp(v)|)}. \tag{41}$$

A first interesting property for studying the Jacobian of this system is the following.

Lemma 3.1. *Let $f(x, y) = H(y) - G(x)$ be an admissible Cartesian representation in the sense of (H1)–(H3). If $y = \ln(x)$, then*

$$-(\partial_x f)^{-1} \partial_y f = (G'(x))^{-1} H'(y) = x$$

Proof. Using $y = \ln(x)$ in (37), it follows that

$$(G'(x))^{-1} H'(y) = (H'(y)/x)^{-1} H'(y) = x.$$

□

The Cartesian representation is naturally associated to the switch parametrization. In particular, the link between parametrizations and Cartesian representations is given in the two following propositions.

Proposition 3.2. *Let $X(\tau), Y(\tau)$ be an admissible parametrization in the sense of (A1)–(A3). Then there exists a Cartesian representation $f(x, y) = H(y) - G(x)$ such that, for all $(x, y) \in \mathbb{R}^2$,*

$$\begin{aligned} G'(x) &= Y'(X^{-1}(x)), \\ H'(y) &= X'(Y^{-1}(y)). \end{aligned} \tag{42}$$

This Cartesian representation is admissible in the sens of (H1)–(H3). Moreover, it satisfies the normalization (36) *if the parametrization satisfies the normalization* (26).

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 Ψ. Setting $H(y) = \Psi(Y^{-1}(y))$ and $G(x) = \Psi(X^{-1}(x))$, one has

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

The result (42) is obtained by taking

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

□

Proposition 3.3. *Let $f(x, y) = H(y) - G(x)$ be an admissible Cartesian representation in the sense of (H1)–(H3). Then, there exists a parametrization $X(\tau), Y(\tau)$ such that, for all τ,*

$$\begin{aligned} X'(\tau) &= H'(Y(\tau)), \\ Y'(\tau) &= G'(X(\tau)). \end{aligned} \tag{43}$$

This parametrization is admissible in the sense of (A1)–(A3) and satisfies the normalization (26) *if the Cartesian representation satisfies the normalization* (36).

Proof. The existence of a solution to the ODE (43) is guaranteed by the hypothesis on H' and G' and the Cauchy-Lipschitz theorem. Therefore

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

and it follows that $f(X(\tau), Y(\tau)) = cst$. Moreover if $f(X(0), Y(0)) = 0$, then $cst = 0$. □

Let us go back to the case we are interested in, if $y = \ln(x)$ then $(\ln(x))' = 1/x > 0$ and $x = \exp(y) > 0$, then we impose that $G'(x) > 0$, $H'(y) > 0$ and $H(0) = 0$, $G(1) = 0$. We can then remove the absolute values in (40)-(41) and the system is rewritten as

$$G'(x) = \frac{1/x}{\max(1, 1/x)}, \quad H'(y) = \frac{\exp(y)}{\max(1, \exp(y))}.$$

Therefore

$$\begin{aligned} &y > 0 \Rightarrow \exp(y) > 1 \Rightarrow H'(y) = 1 \Rightarrow H(y) \underbrace{-H(0)}_{=0} = y - 0, \\ &y \leq 0 \Rightarrow \exp(y) \leq 1 \Rightarrow H'(y) = \exp(y) \Rightarrow H(y) \underbrace{-H(0)}_{=0} = \exp(y) - 1, \end{aligned}$$

leading to

$$H(y) = y\mathbf{1}_{\{y>0\}} + (\exp(y) - 1)\mathbf{1}_{\{y\leq 0\}}. \tag{44}$$

It follows that:

$$G(x) = H(\ln x) = \ln x \mathbf{1}_{\{x>1\}} + (x - 1)\mathbf{1}_{\{x\leq 1\}}. \tag{45}$$

The function f is then defined in four areas as

$$f(x, y) = \begin{cases} \mathrm{e}^y - x, & \text{if } x \leq 1,\ y \leq 0, \\ y - x + 1, & \text{if } x \leq 1,\ y \geq 0, \\ y - \ln x, & \text{if } x \geq 1,\ y \geq 0, \\ \mathrm{e}^y - \ln x - 1, & \text{if } x \geq 1, y \leq 0. \end{cases} \tag{46}$$

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 2, can readily be shown to be convex.

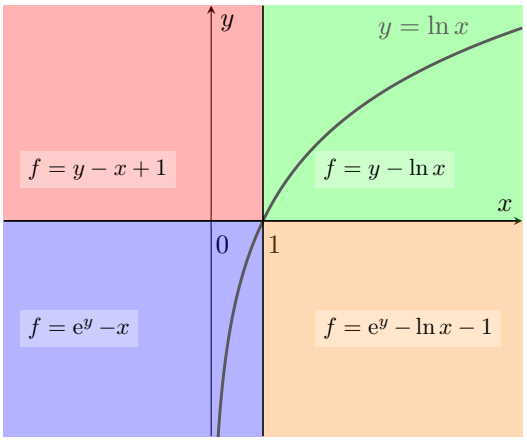

Figure 2: The discrepancy function.

4 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.

4.1 About the parametrization

Let $X(\tau), Y(\tau)$ be an admissible parametrization for the formulation (13) in the sense of (A1)–(A3). One defines the function $\mathfrak{W} : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ such that

$$\begin{aligned} \mathbf{A}\mathbf{X}(\boldsymbol{\tau}) - \omega \mathbf{b} &= \mathbf{0}, \\ \mathfrak{W}(\boldsymbol{\tau}, \omega) = \mathbf{0} \Leftrightarrow \mathbf{S}^T[\boldsymbol{\mu}^\circ/(RT) + \mathbf{Y}(\boldsymbol{\tau})] &= \mathbf{0}, \\ \langle \mathbf{X}(\boldsymbol{\tau}), \mathbf{1} \rangle - 1 &= 0. \end{aligned} \tag{47}$$

The Jacobian matrix $\boldsymbol{\nabla}\mathfrak{W} = \boldsymbol{\nabla}\mathfrak{W}(\boldsymbol{\tau}, \omega)$, associated to (47), is written as

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

Let us demonstrate that this Jacobian is invertible at the solution point.

Proposition 4.1. *If $(\boldsymbol{\tau}, \omega)$ is solution of (47), then $\boldsymbol{\nabla}\mathfrak{W}(\boldsymbol{\tau}, \omega)$ is nonsingular.*

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

$$\mathbf{A}\mathrm{diag}\{\mathbf{X}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau} - \delta\omega\mathbf{b} = \mathbf{0}, \tag{48}$$

$$\mathbf{S}^T\mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau} = \mathbf{0}, \tag{49}$$

$$\langle \mathbf{X}'(\boldsymbol{\tau}), \boldsymbol{\delta\tau} \rangle = 0. \tag{50}$$

Since $(\boldsymbol{\tau}, \omega)$ is solution of (47), one has $\mathbf{b} = \mathbf{A}\mathbf{X}(\boldsymbol{\tau})/\omega$ with $\omega > 0$. The equation (48) then becomes

$$\mathbf{A}\left[\mathrm{diag}\{\mathbf{X}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau} - \frac{\delta\omega}{\omega}\mathbf{X}(\boldsymbol{\tau})\right] = \mathbf{0}. \tag{51}$$

Moreover, equation (49) means that $\mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau} \in \ker \mathbf{S}^T$ which can also be expressed as $\mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau} \in (\ker \mathbf{A})^\perp$, as indicated by Lemma 2.1. Consequently, using (51), the following equality holds:

$$\begin{aligned} \langle \mathrm{diag}\{\mathbf{X}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau}, \mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau} \rangle &= \frac{\delta\omega}{\omega} \langle \mathbf{X}(\boldsymbol{\tau}), \mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau} \rangle \\ &= \frac{\delta\omega}{\omega} \langle \mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\mathbf{X}(\boldsymbol{\tau}), \boldsymbol{\delta\tau} \rangle. \end{aligned} \tag{52}$$

By deriving the relationship $X(\tau) = \exp(Y(\tau))$, we get that

$$\mathbf{X}'(\boldsymbol{\tau}) = \mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\mathbf{X}(\boldsymbol{\tau}). \tag{53}$$

One deduces that the right-hand side of (52) vanishes thanks to (50). Therefore

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

which is only possible if $\boldsymbol{\delta\tau} = \mathbf{0}$. Indeed, using (53), it turns out that

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

is a positive-definite matrix since $X(\tau) > 0$. Equation (48) finally allows to conclude that $\delta\omega = 0$, meaning that the Jacobian is nonsingular. □

Theorem 4.1. *Let $X(\tau), Y(\tau)$ be an admissible parametrization in the sense of assumptions (A1)–(A3). If the Newton sequence* (19)–(20) *is applied to the function $\mathfrak{W}$ defined as* (47)*, then the local quadratic convergence theorem holds.*

Proof. The proof consists of verifying that the assumptions of Theorem 3.1 are satisfied. The existence of a solution come from Proposition 2.2 and the assumptions (A1)–(A3) on $X(\tau)$ and $Y(\tau)$. The Jacobian $\boldsymbol{\nabla}\mathfrak{W}$ is Lipschitz continuous since X' and Y' are Lipschitz continuous according to (A2). Moreover, from Proposition 4.1, $\boldsymbol{\nabla}\mathfrak{W}$ is nonsingular at the solution point. □

4.2 About the Cartesian representation

Let $f(x, y)$ be an admissible Cartesian representation for the formulation (13) in the sense of (H1)–(H3). In order to apply Newton's method, one defines the function $\mathfrak{H} : \mathbb{R}^{2N+1} \to \mathbb{R}^{2N+1}$ such that

$$\mathfrak{H}(\mathbf{x}, \mathbf{y}, \omega) = \mathbf{0} \Leftrightarrow \begin{aligned} \mathbf{A}\mathbf{x} - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T[\boldsymbol{\mu}^\circ/(RT) + \mathbf{y}] &= \mathbf{0}, \\ \boldsymbol{f}(\mathbf{x}, \mathbf{y}) &= \mathbf{0}, \\ \langle \mathbf{x}, \mathbf{1}\rangle - 1 &= 0. \end{aligned} \tag{54}$$

The associated Jacobian matrix $\boldsymbol{\nabla}\mathfrak{H} := \boldsymbol{\nabla}\mathfrak{H}(\mathbf{x}, \mathbf{y}, \omega)$ of this formulation is written as

$$\boldsymbol{\nabla}\mathfrak{H} = \begin{bmatrix} \mathbf{A} & \mathbf{0} & -\mathbf{b} \\ \mathbf{0} & \mathbf{S}^T & \mathbf{0} \\ \boldsymbol{\nabla}_{\mathbf{x}}\boldsymbol{f} & \boldsymbol{\nabla}_{\mathbf{y}}\boldsymbol{f} & \mathbf{0} \\ \mathbf{1}^T & \mathbf{0} & 0 \end{bmatrix}, \quad \text{with} \quad \begin{aligned} \boldsymbol{\nabla}_{\mathbf{x}}\boldsymbol{f} &:= \boldsymbol{\nabla}_{\mathbf{x}}\boldsymbol{f}(\mathbf{x}, \mathbf{y}) = \mathrm{diag}\{\partial_{x_i} f(x_i, y_i)\}_{i=1,\dots,N}, \\ \boldsymbol{\nabla}_{\mathbf{y}}\boldsymbol{f} &:= \boldsymbol{\nabla}_{\mathbf{y}}\boldsymbol{f}(\mathbf{x}, \mathbf{y}) = \mathrm{diag}\{\partial_{y_i} f(x_i, y_i)\}_{i=1,\dots,N}. \end{aligned}$$

We will show that for the unique vector $(\mathbf{x}, \mathbf{y}, \omega)^T$ satisfying (54), the Jacobian $\boldsymbol{\nabla}\mathfrak{H}(\mathbf{x}, \mathbf{y}, \omega)$ is invertible.

Lemma 4.1. *The matrix defined as*

$$\mathbf{J}(\mathbf{x}, \mathbf{y}) := [\boldsymbol{\nabla}\mathfrak{H}(\mathbf{x}, \mathbf{y}, \omega)]_{1-2N,1-2N} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^T \\ \boldsymbol{\nabla}_{\mathbf{x}}\boldsymbol{f} & \boldsymbol{\nabla}_{\mathbf{y}}\boldsymbol{f} \end{bmatrix},$$

corresponding to the first $2N$ rows and columns of $\boldsymbol{\nabla}\mathfrak{H}$, is invertible for all $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{2N}$.

Proof. The proof consists in showing that $\mathbf{J}^T$ is injective. Let $\boldsymbol{\delta}\mathbf{U} \in \ker \mathbf{J}^T(\mathbf{x}, \mathbf{y})$ be such that $\boldsymbol{\delta}\mathbf{U} = (\boldsymbol{\delta}\mathbf{x}_1, \boldsymbol{\delta}\mathbf{x}_2, \boldsymbol{\delta}\mathbf{y})^T \in \mathbb{R}^M \times \mathbb{R}^{N-M} \times \mathbb{R}^N$, and let $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{2N}$, then

$$\mathbf{J}^T(\mathbf{x}, \mathbf{y})\boldsymbol{\delta}\mathbf{U} = \mathbf{0} \Leftrightarrow \begin{aligned} \mathbf{A}^T\boldsymbol{\delta}\mathbf{x}_1 &= -\boldsymbol{\nabla}_{\mathbf{x}}\boldsymbol{f}\boldsymbol{\delta}\mathbf{y} \\ \mathbf{S}\boldsymbol{\delta}\mathbf{x}_2 &= -\boldsymbol{\nabla}_{\mathbf{y}}\boldsymbol{f}\boldsymbol{\delta}\mathbf{y} \end{aligned} \Rightarrow \left[\boldsymbol{\delta}\mathbf{x}_1^T \underbrace{(\mathbf{A}\mathbf{S})}_{=\mathbf{0}} \boldsymbol{\delta}\mathbf{x}_2 = \boldsymbol{\delta}\mathbf{y}^T(\boldsymbol{\nabla}_{\mathbf{x}}\boldsymbol{f}\boldsymbol{\nabla}_{\mathbf{y}}\boldsymbol{f})\boldsymbol{\delta}\mathbf{y} \right] \Rightarrow \boldsymbol{\delta}\mathbf{y} = \mathbf{0}$$

since $\boldsymbol{\nabla}_{\mathbf{x}}\boldsymbol{f}\boldsymbol{\nabla}_{\mathbf{y}}\boldsymbol{f}$ is negative-definite. Hence $\boldsymbol{\delta}\mathbf{x}_1 \in \ker \mathbf{A}^T = \{\mathbf{0}_{\mathbb{R}^M}\}$ and $\boldsymbol{\delta}\mathbf{x}_2 \in \ker \mathbf{S} = \{\mathbf{0}_{\mathbb{R}^{N-M}}\}$ given that $\mathbf{A}^T$ and $\mathbf{S}$ have full rank. Therefore $\mathbf{J}^T(\mathbf{x}, \mathbf{y})$ is invertible and it follows that $\mathbf{J}(\mathbf{x}, \mathbf{y})$ is also invertible. □

Proposition 4.2. *If $\mathbf{U} = (\mathbf{x}, \mathbf{y}, \omega)^T$ is solution of* (54)*, then $\boldsymbol{\nabla}\mathfrak{H}(\mathbf{U})$ is nonsingular.*

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

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

which always exists thanks to the invertibility of $\mathbf{J}$ from Lemma 4.1. Noting that the solution satisfies $\boldsymbol{\delta}\tilde{\mathbf{U}}_\alpha = \alpha\boldsymbol{\delta}\tilde{\mathbf{U}}_1$, we define the vector $\boldsymbol{\delta}\mathbf{U} := (\boldsymbol{\delta}\tilde{\mathbf{U}}_{\delta\omega}, \delta\omega)^T = \delta\omega(\boldsymbol{\delta}\tilde{\mathbf{U}}_1, 1)^T$. It follows that

$$\boldsymbol{\nabla}\mathfrak{H}(\mathbf{x},\mathbf{y},\omega)\boldsymbol{\delta}\mathbf{U} = \mathbf{0} \Leftrightarrow \begin{array}{r} \delta\omega\left[\mathbf{J}(\mathbf{x},\mathbf{y})\boldsymbol{\delta}\tilde{\mathbf{U}}_1 - (\mathbf{b},\mathbf{0},\mathbf{0})^T\right] = \mathbf{0} \\ \delta\omega\langle\boldsymbol{\delta}\mathbf{x}_1, \mathbf{1}\rangle = \mathbf{0}. \end{array}$$

By the definition of $\tilde{\mathbf{U}}_1$ one has $\mathbf{J}(\mathbf{x},\mathbf{y})\boldsymbol{\delta}\tilde{\mathbf{U}}_1 - (\mathbf{b},\mathbf{0},\mathbf{0})^T = \mathbf{0}$, hence the invertibility of $\boldsymbol{\nabla}\mathfrak{H}$ is determined by $\delta\omega\langle\boldsymbol{\delta}\mathbf{x}_1, \mathbf{1}\rangle = \mathbf{0}$:

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

Therefore to prove that $\boldsymbol{\nabla}\mathfrak{H}(\mathbf{x},\mathbf{y},\omega)$ is invertible for $(\mathbf{x},\mathbf{y},\omega)$ solution of (54), it is sufficient to show that $\langle\boldsymbol{\delta}\mathbf{x}_1, \mathbf{1}\rangle \neq 0$ for $(\boldsymbol{\delta}\mathbf{x}_1, \boldsymbol{\delta}\mathbf{y}_1)$ the unique solution of

$$\mathbf{A}\boldsymbol{\delta}\mathbf{x}_1 = \mathbf{b}, \tag{55}$$

$$\mathbf{S}^T\boldsymbol{\delta}\mathbf{y}_1 = \mathbf{0}, \tag{56}$$

$$\boldsymbol{\nabla}_{\mathbf{x}}\boldsymbol{f}\boldsymbol{\delta}\mathbf{x}_1 + \boldsymbol{\nabla}_{\mathbf{y}}\boldsymbol{f}\boldsymbol{\delta}\mathbf{y}_1 = \mathbf{0}. \tag{57}$$

By denoting $\mathbf{D} := -(\boldsymbol{\nabla}_{\mathbf{x}}\boldsymbol{f})^{-1}\boldsymbol{\nabla}_{\mathbf{y}}\boldsymbol{f}$, one has $\boldsymbol{\delta}\mathbf{x}_1 = \mathbf{D}\boldsymbol{\delta}\mathbf{y}_1$ from (57). Furthermore from (56) and Lemma 2.1 one has that $\boldsymbol{\delta}\mathbf{y}_1 \in \ker\mathbf{S}^T = \mathrm{Im}\,\mathbf{A}^T$, so there exists $\boldsymbol{\delta}\mathbf{h}_1$ such that $\boldsymbol{\delta}\mathbf{y}_1 = \mathbf{A}^T\boldsymbol{\delta}\mathbf{h}_1$. Therefore (55) can be rewritten as

$$\mathbf{A}\mathbf{D}\mathbf{A}^T\boldsymbol{\delta}\mathbf{h}_1 = \mathbf{b}. \tag{58}$$

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 since the rank of $\mathbf{A}\mathbf{D}^{1/2}$ is maximal. Moreover one has $\mathbf{b} = \frac{1}{\omega}\mathbf{A}\mathbf{x}$ since $(\mathbf{x},\mathbf{y},\omega)$ solves (54), then from (58) one finds that

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

By multiplying both sides of this equation by $\mathbf{D}\mathbf{A}^T$ one obtains

$$\begin{aligned} \mathbf{D}\mathbf{A}^T\boldsymbol{\delta}\mathbf{h}_1 = \mathbf{D}\boldsymbol{\delta}\mathbf{y}_1 = \boldsymbol{\delta}\mathbf{x}_1 &= \frac{1}{\omega}\mathbf{D}\mathbf{A}^T(\mathbf{A}\mathbf{D}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{x} \\ &= \frac{1}{\omega}\mathbf{D}^{1/2}\left[(\mathbf{A}\mathbf{D}^{1/2})^T(\mathbf{A}\mathbf{D}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{D}^{1/2}\right]\mathbf{D}^{-1/2}\mathbf{x}. \end{aligned} \tag{59}$$

Let $\mathbf{B} := \mathbf{A}\mathbf{D}^{1/2}$, then $\boldsymbol{\Pi} := \mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}$ is the orthogonal projection on $(\ker\mathbf{B})^\perp$. Thus (59) becomes

$$\boldsymbol{\delta}\mathbf{x}_1 = \frac{1}{\omega}\mathbf{D}^{1/2}\boldsymbol{\Pi}\mathbf{D}^{-1/2}\mathbf{x} = \frac{1}{\omega}\mathbf{D}^{1/2}\boldsymbol{\Pi}^2\mathbf{D}^{-1/2}\mathbf{x}, \tag{60}$$

since an orthogonal projection always satisfies $\boldsymbol{\Pi}^2 = \boldsymbol{\Pi}$. Therefore since $(\mathbf{x},\mathbf{y},\omega)$ solves (54), Lemma 3.1 yields $\mathbf{D} = \mathrm{diag}\{x_i\}_{i=1,\dots,N}$, then $\mathbf{1}^T\mathbf{D}^{1/2} = (\mathbf{D}^{-1/2}\mathbf{x})^T = (\mathbf{x}^{1/2})^T$. Hence the scalar product between the vector $\mathbf{1}$ and (60) gives

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

Therefore

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

which is not possible since $\mathbf{A}\mathbf{x} = \omega\mathbf{b} \neq 0$. Thus $\langle\mathbf{1}, \boldsymbol{\delta}\mathbf{x}_1\rangle \neq 0$ and the Jacobian is invertible. □

Theorem 4.2. *Let $f(x,y) = H(y) - G(x)$ be an admissible Cartesian representation in the sense of assumptions (H1)–(H3). If the Newton sequence* (19)–(20) *is applied to the function $\mathfrak{H}$ defined as* (54)*, then the local quadratic convergence theorem holds.*

Proof. The proof consists of verifying that the assumptions of Theorem 3.1 are satisfied. The existence of a solution come from Proposition 2.2 and the assumptions (H1)–(H3) on $H(y)$ and $G(x)$. The Jacobian $\boldsymbol{\nabla}\mathfrak{H}$ is Lipschitz continuous since H' and G' are Lipschitz continuous according to (H2). Moreover, from Proposition 4.2, $\boldsymbol{\nabla}\mathfrak{H}$ is nonsingular at the solution point. □

An interesting property of the Cartesian representation associated with the function (46) is that the iterates of Newton's method always lie above the logarithm graph.

Proposition 4.3. *Let $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}, \omega^{(k)})$ be a Newton iterate for the Cartesian representation formulation described in Section 3.3 with discrepancy function f defined by* (46)*. Then, for $k \geq 1$, the linear equations $\mathbf{A}\mathbf{x}^{(k)} = \omega^{(k)}\mathbf{b}$, $\mathbf{S}^T\mathbf{y}^{(k)} = \mathbf{d}$ and $\langle \mathbf{x}^{(k)}, \mathbf{1}\rangle = 1$ 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

$$f(x_i^{(k)}, y_i^{(k)}) \geq f(x_i^{(k-1)}, y_i^{(k-1)}) + \partial_x f(x_i^{(k-1)}, y_i^{(k-1)})\delta x_i^{(k-1)} + \partial_y f(x_i^{(k-1)}, y_i^{(k-1)})\delta y_i^{(k-1)} = 0,$$

the last equality stemming from the definition of the increment $\boldsymbol{\delta}\mathbf{x}^{(k-1)} = \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}$ by Newton's method. □

5 Numerical results

In this section we will present various test cases to validate our methods and compare them with the Arxim geochemical modeling library [27]. Arxim is written in C++ and it is based on the log formulation as described in (23). Moreover, Arxim uses a line search to globalize the Newton method, this strategy is the one described in the chapter 9.7.1 of the book Numerical recipes [30]. Our code has been developed with the Julia Programming Language and uses the automatic differentiation package ForwardDiff [31]. No globalization strategy for Newton's method has been implemented in our code. The function X and Y for the parametrization are those of the switch defined in (35). For the Cartesian representation technique, the function f is the discrepancy function defined in (46). In all numerical experiments, pressure and temperature values are set at P= 1 Bar and T = 298.15 K. Moreover, if $\mathfrak{N}(\mathcal{X})$ represents the function whose root we seek, the convergence criterion for Newton's algorithm is

$$\|\mathfrak{N}(\mathcal{X}^{(k+1)})\|_\infty \leq 1e^{-7} \quad \text{and} \quad \|\mathcal{X}^{(k+1)} - \mathcal{X}^{(k)}\|_\infty \leq 1e^{-7}$$

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

All the systems we will be studying are dilute solutions with water as the solvent. In Arxim, the ideal activity model corresponds to molalities, i.e. the quantity of solute present in 1 kg of solvent:

$$a_i(\mathbf{n}) = \begin{cases} 1 & \text{if} \quad i = 1 \text{ (solvent)}, \\ \dfrac{n_i}{n_1 M_{\mathrm{H_2O}}} & \text{if} \quad i > 1 \text{ (solute)}, \end{cases} \tag{61}$$

where $n_1 = n_{H_2O}$ and with $M_{\mathrm{H_2O}} = 0.0180152$ kg/mol the molar mass of water. The activity (61) is an approximation based on the fact that in a dilute solution the quantities of solute species are negligible compared to the quantity of solvent. Therefore, the chemical potential of the solvent is reduced to its standard chemical potential and for a dilute species, the transition from a chemical potential in mole fraction to a chemical potential in molality is as follows:

$$\mu_i^\circ + RT \ln \frac{n_i}{\sum_{j=1}^N n_j} \approx \mu_i^\circ + RT \ln \frac{n_i}{n_1} = \tilde{\mu}_i^\circ + RT \ln \frac{n_i}{n_1 M_{\mathrm{H_2O}}},$$

where $\tilde{\mu}_i^\circ := \mu_i^\circ + RT \ln M_{\mathrm{H_2O}}$.

In order to compare our results with those of Arxim, it is necessary to adapt the system (13) to the molality activity model. We thus obtain

$$\begin{aligned} \mathbf{A}\begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T\begin{bmatrix} 0 \\ \mathbf{y}(\mathbf{x}) - \ln(M_{\mathrm{H_2O}}) \end{bmatrix} - \mathbf{d} &= \mathbf{0}. \end{aligned} \tag{62}$$

Compared to (13), the quantity of water is no longer a variable and $x_i = \omega n_i$ where the associated ω is defined as

$$\omega = \frac{1}{n_1}. \tag{63}$$

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 initializations are as follows.

- In Arxim, the variables are the logarithm of the quantities, namely:

```
init_guess = [ln(n[1]), ..., ln(n[N])].
```

- For the parametrization, the variable ω is defined as in (12) or (63), depending of the ideal activity model used, whereas the function $X(\tau)$, defined in (35), is inverted to define the others variables from $x_i = \omega n_i$:

```
init_guess_molality = [1/n[1], Xinv(x[2]), ..., Xinv(x[N])],
init_guess_mole_frac = [1/sum(n), Xinv(x[1]), ..., Xinv(x[N])].
```

- For the Cartesian representation, the main question concerns the initialization of the vector $\mathbf{y}$. The first choice is to initialize with $y_i = \ln(x_i)$. However, we will see that this choice is too restrictive, hence we propose the second initialisation $y_i = x_i - 1$. In this way, the point (x_i, y_i) always lies above the curve $y_i = \ln x_i$, which will always be the case during the iterations according to Proposition 4.3. A comparison of these initializations will be made on the Seawater test case. The resulting initializations are:

```
init1_guess_molality = [1/n[1], x[2], ..., x[N], ln(x[2]), ..., ln(x[N])],
init1_guess_mole_frac = [1/sum(n), x[1], ..., x[N], ln(x[1]), ..., ln(x[N])],
init2_guess_molality = [1/n[1], x[2], ..., x[N], x[2]-1, ..., x[N]-1],
init2_guess_mole_frac = [1/sum(n), x[1], ..., x[N], x[1]-1, ..., x[N]-1].
```

We present five different test cases. The first two, *NaCl* and *Seawater*, use the charge constraint defined in section 2.5.1. For *NaCl*, a comparison of the solution will be made with Arxim to ensure the accuracy of our methods. For *Seawater*, a comparison of the number of iterations over four different initializations will test the robustness of our methods against Arxim. For each of these initializations, the n_1 value for water will be dictated by the amount of oxygen O, which is dominated by the presence of H_2O. The other quantities $(n_2, \ldots, n_N)$ will all be initialized to the same value $\chi \in \{1e^{-2}, 1e^{-4}, \epsilon_{32}, \epsilon_{64}\}$ where $\epsilon_{32} = 1.1920929e^{-7}$ and $\epsilon_{64} = 2.220446049250313e^{-16}$ are the machine epsilons in single and double precision respectively. The three other test cases, *H_ 2 pE*, *Water-Clay* and *Water-Concrete*, use the charge constraint and the pE constraint defined in section 2.5.2. The first is used to test the precision of our methods, while the other two are intended to test their robustness.

We will additionally assess the precision of certain results represented by $\mathbf{u}$ and $\mathbf{v}$ by employing the following estimation method for comparison:

$$\varepsilon_{prec} = \max_i \frac{\|u_i - v_i\|}{\max(\|u_i\|, \|v_i\|)}.$$

5.1 The *NaCl* test case

This first test case, *NaCl*, will serve to validate our methods in terms of solution accuracy. The chemical system is composed of 8 species, 4 elements and 4 reactions. This system corresponds to the following three sets:

$$\begin{aligned}
\mathcal{C} &= (\mathrm{H_2O}, \mathrm{Cl^-}, \mathrm{Na^+}, \mathrm{H^+}, \mathrm{NaCl}, \mathrm{NaOH}, \mathrm{OH^-}, \mathrm{HCl}),\\
\mathcal{E} &= (\mathrm{H}, \mathrm{O}, \mathrm{Na}, \mathrm{Cl}),\\
\mathcal{R} &= (\mathrm{NaCl(aq)} = \mathrm{Cl^-} + \mathrm{Na^+},\\
&\quad\ \mathrm{NaOH(aq)} = \mathrm{H_2O} + \mathrm{Na^+} - \mathrm{H^+},\\
&\quad\ \mathrm{OH^-} = \mathrm{H_2O} - \mathrm{H^+},\\
&\quad\ \mathrm{HCl(aq)} = \mathrm{Cl^-} + \mathrm{H^+}).
\end{aligned}$$

Table 1: Number of moles of the solution for *NaCl* with 8 significant figures, computed with the discrepancy function.

Feeds (mol)		
O	55.5087	
Na	0.1	
Cl	0.1	
Z (charge)	0	
Results		
Amount (mol)	Molality model	Mole fraction model
H_2O	5.55086999e+01	5.55086999e+01
Cl^-	9.83824576e-02	9.83880354e-02
Na^+	9.83824536e-02	9.83880313e-02
H^+	1.61754040e-03	1.61196265e-03
NaCl(aq)	1.03490312e-07	1.03668102e-07
NaOH(aq)	9.94371673e-08	9.96214123e-08
OH^-	6.03710854e-09	6.02709662e-09
HCl(aq)	1.98396398e-09	1.98040700e-09

Our two methods give the same results with a precision of $\varepsilon_{prec} \approx 7.47e^{-14}$ for the molality models and $\varepsilon_{prec} \approx 3.089e^{-13}$ for the mole fraction models. A comparison between the results obtained from the discrepancy function and Arxim gives a precision of $\varepsilon_{prec} \approx 5.73e^{-11}$, meaning that our methods are accurate. Moreover, a comparison of the results in Table 1 gives $\varepsilon_{prec} \approx 0.0035$, which is the loss of precision between the two activity models.

5.2 The *Seawater* test case

The *Seawater* test case aims to compare the required number of iterations for the Newton's method to converge. The associated system comprises 37 species, 10 elements and 27 reactions. It is composed of the three sets described in appendix Appendix B.1.

The results of Table 2 demonstrate the robustness of our methods to the input data. In particular, the Cartesian representation always converges in 23 iterations with the second initialization. The robustness of this second initialization is better than the first. In the following, we will only use this second initialization. It is worth noting that Arxim's line search method reduces the number of iterations on the first two initializations, but there is a notable degradation in convergence when initializing with the epsilon machine ϵ_{64}.

A comparison between the results obtained from the discrepancy function and Arxim gives a precision of $\varepsilon_{prec} \approx 3.0024e^{-12}$. A comparison of between the two activity models gives $\varepsilon_{prec} \approx 0.029$.

5.3 The *H_ 2 pE* test case

The *H_ 2 pE* test case is a system for testing the accuracy of our method when adding a pE constraint. The system is composed of 16 species, 4 elements and 11 reactions. The corresponding sets are described in appendix Appendix B.2.

Our two methods give the same results with a precision of $\varepsilon_{prec} \approx 3.98e^{-13}$ for both activity models. A comparison between the results obtained from the discrepancy function and Arxim gives a precision of $\varepsilon_{prec} \approx 3.17e^{-12}$, meaning that our methods are also accurate when using a pE constraint. A comparison of between the two activity models in Table 3 gives $\varepsilon_{prec} \approx 5.67e^{-5}$.

5.4 The *Water-Clay* and *Water-Concrete* test cases

The purpose of the *Water-Clay* and *Water-Concrete* test cases is to test the robustness of our method when adding a pE constraint. They are both based on the same chemical system, but differ in their quantities of chemical elements. This system is composed of 88 species, 12 elements and 75 reactions distributed in the sets detailed in appendix Appendix B.3.

The results in Table 4 show that the Cartesian representation method is robust to the input data. Furthermore, the number of iterations is almost always the same. The parametrization method converges

Table 2: Number of iterations of Newton's method before convergence for *Seawater*, computed with the discrepancy function.

Feeds (mol)				
O	55.5087			
Na	0.469			
Mg	0.0528			
S	0.0282			
Ca	0.0103			
K	0.0102			
C	0.00206			
Sr	1e-5			
Cl	0.546			
Z (charge)	0			
Initial guess (mol)				
H_2O	55.5087	55.5087	55.5087	55.5087
Other species	1e-2	1e-4	ϵ_{32}	ϵ_{64}
Results (Mole fraction)				
Switch	18	29	31	29
Discrepancy init1	18	24	24	$\times$
Discrepancy init2	23	23	23	23
Results (Molality)				
Arxim	16	13	22	33
Switch	18	25	27	28
Discrepancy init1	18	24	24	24
Discrepancy init2	23	23	23	23

Table 3: Number of moles of the solution for *H_ 2 pE* with 8 significant figures.

Feeds (mol)		
O	55.5087	
C	0.001	
Ca	0.001	
Z (charge)	0	
pE	10	
Results		
Amount (mol)	Molality model	Mole fraction model
H_2O	5.55053359e+01	5.55053359e+01
Ca^{2+}	5.89499320e-04	5.89504780e-04
$CaCO_3$(aq)	4.06465710e-04	4.06460327e-04
CO_3^{2-}	3.24657039e-04	3.24660881e-04
HCO_3^-	2.67101320e-04	2.67102894e-04
OH^-	2.66618250e-04	2.66619835e-04
O_2(aq)	4.75961472e-05	4.75956445e-05
$CaOH^+$	2.28144885e-06	2.28140520e-06
$CaHCO_3^+$	1.75352116e-06	1.75348751e-06
CO_2(aq)	2.24095166e-08	2.24102849e-08
H^+	3.85927676e-11	3.85938635e-11
H_2O_2(aq)	1.66424972e-19	1.66421236e-19
HO_2^-	8.92582992e-21	8.92568259e-21
H_2(aq)	1.15434611e-44	1.15437203e-44
CO(aq)	2.93159368e-54	2.93176001e-54
CH_4(aq)	3.35322590e-150	3.35341170e-150

in only 7 out of 16 cases, and is therefore not robust to this type of system. It should be noted that Arxim does not converge for initializations using the epsilon machines ϵ_{64}.

Comparisons of the number of moles of the solution for *Water-Clay* and *Water-Concrete* with Arxim gives a precision of $\varepsilon_{prec} \approx 1.89e^{-10}$ and $\varepsilon_{prec} \approx 1.71e^{-11}$ respectively. Comparisons of between the two activity models gives $\varepsilon_{prec} \approx 0.043$ and $\varepsilon_{prec} \approx 0.104$ respectively.

Table 4: Number of iterations of Newton's method before convergence for *Water-Clay* and *Water-Concrete*.

Feeds (mol)								
	Water-Clay				*Water-Concrete*			
O	55.5078				55.5078			
Na	0.0401				0.0601			
Mg	0.0057449				1.5079e-09			
K	0.000523301				0.1402			
Ca	0.00846445				0.00196384			
Fe	7.5646e-05				4.58364e-07			
C	0.003783				5.29145e-05			
Al	8.32493e-08				3.80016e-05			
S	0.0126881				0.000974141			
Cl	0.04096				1.42825e-10			
Sr	0.000231597				1e-10			
Z (charge)	0				0			
pE	-2.807				-2.98873			
Initial guess (mol)								
H_2O	55.5078	55.5078	55.5078	55.5078	55.5078	55.5078	55.5078	55.5078
Others	1e-2	1e-4	ϵ_{32}	ϵ_{64}	1e-2	1e-4	ϵ_{32}	ϵ_{64}
Results (Mole fraction)								
Switch	31	31	×	35	63	×	×	×
Discrepancy	34	34	34	34	50	68	68	68
Results (molality)								
Arxim	28	25	28	×	65	47	34	×
Switch	31	29	×	×	×	×	59	×
Discrepancy	32	32	32	32	52	52	52	52

6 Conclusion and future works

We have presented parametrization and Cartesian representation techniques for stabilizing Newton's algorithm in the context of calculating chemical equilibria in an aqueous phase. For each of them, we have proved the local quadratic convergence of Newton's procedure. The numerical results demonstrate excellent accuracy on the solution of the chemical systems considered. The Cartesian representation technique is particularly robust with respect to the system data and the initial Newton point. Moreover, contrary to existing approaches, no globalization strategy is required for the Cartesian representation to converge, even on challenging test cases. Although the robustness results of the parametrization are not satisfactory for difficult test cases, they are still notable for simpler chemical systems. Future work will focus on extending these methods to multiphase chemical equilibrium problems.

Appendix A Standard chemical potentials

The standard chemical potentials $\tilde{\mu}_i^\circ$(P, T) of a species C_i for a constant pressure P and temperature T is calculated using the Helgeson-Kirkham-Flowers (HKF) model.

Table 5: Standard chemical potentials based on molality at P= 1 Bar and T = 298.15 K

Formula	$\tilde{\mu}_i^\circ$(P,T)	Formula	$\tilde{\mu}_i^\circ$(P,T)
H_2O	-237138.97589284607	H^+	0.0
O_2(aq)	16543.49301855645	Na^+	-261880.68093357404
Mg^{2+}	-453984.787558262	K^+	-282461.78508303704
Ca^{2+}	-552789.9276571237	Fe^{2+}	-91504.0330259397
HCO_3^-	-586939.7841306784	Al^{3+}	-483707.8898133192
SO_4^{2-}	-744458.9698933255	Cl^-	-131289.73255268394
Sr^{2+}	-563835.6855778407	AlO^+	-661858.4374760946
$AlOH^{2+}$	-692347.2367805466	$HAlO_2$(aq)	-869016.6094354176
AlO_2^-	-831331.3317788806	$CaOH^+$	-716719.0378653684
CO(aq)	-120005.50268817003	CO_2(aq)	-385973.95119834674

CO_3^{2-}	-527983.0213240985	$CaHCO_3^+$	-1.1457044648801715e6
$CaCl^+$	-682410.2459630198	$CaCl_2(aq)$	-811695.8491350176
$CaSO_4(aq)$	-1.3092988132949607e6	$HClO(aq)$	-79914.41801087881
ClO^-	-36819.196554506314	ClO_2^-	17154.36282858776
ClO_3^-	-7949.650116475296	ClO_4^-	-8535.41943368159
Fe^{3+}	-17238.03459232613	$FeCl^+$	-221877.47047403525
$FeCl_2(aq)$	-307440.2827514088	$FeOH^{2+}$	-241835.11351838848
$FeOH^+$	-275516.3143866307	FeO^+	-222170.29366991075
FeO	-212212.40732728643	$HFeO_2^-$	-399153.4930723803
$HFeO_2(aq)$	-423002.3083502744	FeO_2^-	-368191.91758807504
$H_2(aq)$	17723.386991291925	$H_2S(aq)$	-27919.871309964183
HO_2^-	-67320.55211902864	HS^-	11966.205458423023
HSO_3^-	-527727.8413605248	HSO_4^-	-755755.7895325182
HSO_5^-	-637515.9968599698	$KCl(aq)$	-399279.0752045903
$KHSO_4(aq)$	-1.0183854278434662e6	$KOH(aq)$	-437227.9151528983
KSO_4^-	-1.031941553491834e6	$CH_4(aq)$	-34451.09621924444
$Mg(CO_3)(aq)$	-998971.5788823196	$Mg(HCO_3)^+$	-1.0468365642174666e6
$MgCl^+$	-584504.6645023803	$MgOH^+$	-624482.7757953044
$NaCl(aq)$	-388735.37877663504	$NaOH(aq)$	-417981.50464769645
OH^-	-157297.44203625448	S_2^{2-}	79495.94491610056
$S_2O_3^{2-}$	-522581.52176511387	$HS_2O_3^-$	-532204.7153883826
$H_2S_2O_3(aq)$	-535551.9268921935	$S_2O_4^{2-}$	-600403.9032858713
$HS_2O_4^-$	-614629.5094590015	$H_2S_2O_4(aq)$	-616721.5194100296
$S_2O_5^{2-}$	-790775.8644150125	$S_2O_6^{2-}$	-966503.8308009355
$S_2O_8^{2-}$	-1.1150358451560326e6	S_3^{2-}	73638.33479142259
$S_3O_6^{2-}$	-958135.8346519175	S_4^{2-}	69035.92450587676
$S_4O_6^{2-}$	-1.0405606330002927e6	S_5^{2-}	65688.7139146296
$S_5O_6^{2-}$	-958135.8391056098	$SO_2(aq)$	-301164.29824334016
SO_3^{2-}	-486599.08954829833	$Sr(HCO_3)^+$	-1.1577962219631393e6
$SrCl^+$	-693707.0438437777	$SrOH^+$	-725087.0372310993
$H_2O_2(aq)$	-134013.53088829323	$HClO_2(aq)$	5857.54132274524
$NaSO_4^-$	-1.0103353876813294e6	$MgSO_4(aq)$	-1.211171480195382e6
$HCl(aq)$	-127235.40484114335	$CaCO_3(aq)$	-1.0997641162865811e6
$SrCO_3(aq)$	-1.10817395483314e6	$FeCl^{2+}$	-156975.24998697112
e^-	-16.315331966024218		

Appendix B Chemical reaction equations

Appendix B.1 The *Seawater* test case

$$\begin{aligned}\mathcal{C} = (&H_2O, Na^+, Mg^{2+}, SO_4^{2-}, Ca^{2+}, K^+, HCO_3^-, Sr^{2+}, Cl^-, H+,\\ &CaOH^+, CO_2(aq), CO_3^{2-}, CaHCO_3^+, CaCl^+, CaCl_2(aq), CaSO_4(aq), HSO_4^-, KCl(aq),\\ &KHSO_4(aq), KOH(aq), KSO_4^-, Mg(CO_3)(aq), Mg(HCO_3)^+, MgCl^+, MgOH^+, NaCl(aq),\\ &NaOH(aq), OH^-, Sr(CO_3)(aq), Sr(HCO_3)^+, SrCl^+, SrOH^+, NaSO_4^-, MgSO_4(aq),\\ &HCl(aq), CaCO_3(aq), SrCO_3(aq)),\end{aligned}$$

$$\mathcal{E} = (H, O, Na, Mg, S, Ca, K, C, Sr, Cl),$$

and the set $\mathcal{R}$ composed of the reactions:

$$CaOH^+ = H_2O + Ca^{2+} - H^+; \qquad CO_2(aq) = -H_2O + HCO_3^- + H^+;$$

$$CO_3^{2-} = HCO_3^- - H^+; \qquad CaHCO_3^+ = Ca^{2+} + HCO_3^-;$$

$$CaCl^+ = Ca^{2+} + Cl^-; \qquad CaCl_2(aq) = Ca^{2+} + 2Cl^-;$$

$$CaSO_4(aq) = SO_4^{2-} + Ca^{2+}; \qquad HSO_4^- = SO_4^{2-} + H^+;$$

$$KCl(aq) = K^+ + Cl^-; \qquad KHSO_4(aq) = SO_4^{2-} + K^+ + H^+;$$

$$KOH(aq) = H_2O + K^+ - H^+; \qquad KSO_4^- = SO_4^{2-} + K^+;$$

$$Mg(CO_3)(aq) = Mg^{2+} + HCO_3^- - H^+; \qquad Mg(HCO_3)^+ = Mg^{2+} + HCO_3^-;$$

$$\begin{aligned}
\mathrm{MgCl^+} &= \mathrm{Mg^{2+}} + \mathrm{Cl^-};\\
\mathrm{NaCl(aq)} &= \mathrm{Na^+} + \mathrm{Cl^-};\\
\mathrm{OH^-} &= \mathrm{H_2O} - \mathrm{H^+};\\
\mathrm{SrCl^+} &= \mathrm{Sr^{2+}} + \mathrm{Cl^-};\\
\mathrm{NaSO_4^-} &= \mathrm{Na^+} + \mathrm{SO_4^{2-}};\\
\mathrm{HCl(aq)} &= \mathrm{Cl^-} + \mathrm{H^+};\\
\mathrm{SrCO_3(aq)} &= \mathrm{HCO_3^-} + \mathrm{Sr^{2+}} - \mathrm{H^+}.
\end{aligned}$$

$$\begin{aligned}
\mathrm{MgOH^+} &= \mathrm{H_2O} + \mathrm{Mg^{2+}} - \mathrm{H^+};\\
\mathrm{NaOH(aq)} &= \mathrm{H_2O} + \mathrm{Na^+} - \mathrm{H^+};\\
\mathrm{Sr(HCO_3)^+} &= \mathrm{HCO_3^-} + \mathrm{Sr^{2+}};\\
\mathrm{SrOH^+} &= \mathrm{H_2O} + \mathrm{Sr^{2+}} - \mathrm{H^+};\\
\mathrm{MgSO_4(aq)} &= \mathrm{Mg^{2+}} + \mathrm{SO_4^{2-}};\\
\mathrm{CaCO_3(aq)} &= \mathrm{Ca^{2+}} + \mathrm{HCO_3^-} - \mathrm{H^+};
\end{aligned}$$

Appendix B.2 The *H_ 2 pE* test case

$$\begin{aligned}
\mathcal{C} = (&\mathrm{H_2O}, \mathrm{H^+}, \mathrm{CO_2(aq)}, \mathrm{Ca^{2+}}, \mathrm{H_2(aq)}, \mathrm{CaOH^+}, \mathrm{CO(aq)}, \mathrm{CO_3^{2-}},\\
&\mathrm{CaHCO_3^+}, \mathrm{HCO_3^-}, \mathrm{HO_2^-}, \mathrm{CH_4(aq)}, \mathrm{O_2(aq)}, \mathrm{OH^-}, \mathrm{H_2O_2(aq)}, \mathrm{CaCO_3(aq)}),\\
\mathcal{E} = (&\mathrm{H}, \mathrm{O}, \mathrm{C}, \mathrm{Ca}),
\end{aligned}$$

and the set $\mathcal{R}$ composed of the reactions:

$$\begin{aligned}
\mathrm{CaOH^+} &= \mathrm{H_2O} - \mathrm{H^+} + \mathrm{Ca^{2+}};\\
\mathrm{CO_3^{2-}} &= \mathrm{H_2O} - 2\mathrm{H^+} + \mathrm{CO_2(aq)};\\
\mathrm{HCO_3^-} &= \mathrm{H_2O} - \mathrm{H^+} + \mathrm{CO_2(aq)};\\
\mathrm{CH_4(aq)} &= -2\mathrm{H_2O} + \mathrm{CO_2(aq)} + 4\mathrm{H_2(aq)};\\
\mathrm{OH^-} &= \mathrm{H_2O} - \mathrm{H^+};\\
\mathrm{CaCO_3(aq)} &= \mathrm{H_2O} - 2\mathrm{H^+} + \mathrm{CO_2(aq)} + \mathrm{Ca^{2+}}.
\end{aligned}$$

$$\begin{aligned}
\mathrm{CO(aq)} &= -\mathrm{H_2O} + \mathrm{CO_2(aq)} + \mathrm{H_2(aq)};\\
\mathrm{CaHCO_3^+} &= \mathrm{H_2O} - \mathrm{H^+} + \mathrm{CO_2(aq)} + \mathrm{Ca^{2+}};\\
\mathrm{HO_2^-} &= 2\mathrm{H_2O} - \mathrm{H^+} - \mathrm{H_2(aq)};\\
\mathrm{O_2(aq)} &= 2\mathrm{H_2O} - 2\mathrm{H_2(aq)};\\
\mathrm{H_2O_2(aq)} &= 2\mathrm{H_2O} - \mathrm{H_2(aq)};
\end{aligned}$$

Appendix B.3 The *Water-Clay* and *Water-Concrete* test cases

$$\begin{aligned}
\mathcal{C} = (&\mathrm{H_2O}, \mathrm{H^+}, \mathrm{O_2(aq)}, \mathrm{Na^+}, \mathrm{Mg^{2+}}, \mathrm{K^+}, \mathrm{Ca^{2+}}, \mathrm{Fe^{2+}}, \mathrm{HCO_3^-}, \mathrm{Al^{3+}},\\
&\mathrm{SO_4^{2-}}, \mathrm{Cl^-}, \mathrm{Sr^{2+}}, \mathrm{AlO^+}, \mathrm{AlOH^{2+}}, \mathrm{HAlO_2(aq)}, \mathrm{AlO_2^-}, \mathrm{CaOH^+}, \mathrm{CO(aq)},\\
&\mathrm{CO_2(aq)}, \mathrm{CO_3^{2-}}, \mathrm{CaHCO_3^+}, \mathrm{CaCl^+}, \mathrm{CaCl_2(aq)}, \mathrm{CaSO_4(aq)}, \mathrm{HClO(aq)}, \mathrm{ClO^-},\\
&\mathrm{ClO_2^-}, \mathrm{ClO_3^-}, \mathrm{ClO_4^-}, \mathrm{Fe^{3+}}, \mathrm{FeCl^+}, \mathrm{FeCl_2(aq)}, \mathrm{FeOH^{2+}}, \mathrm{FeOH^+}, \mathrm{FeO^+},\\
&\mathrm{FeO}, \mathrm{HFeO_2^-}, \mathrm{HFeO_2(aq)}, \mathrm{FeO_2^-}, \mathrm{H_2(aq)}, \mathrm{H_2S(aq)}, \mathrm{HO_2^-}, \mathrm{HS^-}, \mathrm{HSO_3^-}, \mathrm{HSO_4^-},\\
&\mathrm{HSO_5^-}, \mathrm{KCl(aq)}, \mathrm{KHSO_4(aq)}, \mathrm{KOH(aq)}, \mathrm{KSO_4^-}, \mathrm{CH_4(aq)}, \mathrm{Mg(CO_3)(aq)}, \mathrm{Mg(HCO_3)^+}, \mathrm{MgCl^+},\\
&\mathrm{MgOH^+}, \mathrm{NaCl(aq)}, \mathrm{NaOH(aq)}, \mathrm{OH^-}, \mathrm{S_2^{2-}}, \mathrm{S_2O_3^{2-}}, \mathrm{HS_2O_3^-}, \mathrm{H_2S_2O_3(aq)}, \mathrm{S_2O_4^{2-}},\\
&\mathrm{HS_2O_4^-}, \mathrm{H_2S_2O_4(aq)}, \mathrm{S_2O_5^{2-}}, \mathrm{S_2O_6^{2-}}, \mathrm{S_2O_8^{2-}}, \mathrm{S_3^{2-}}, \mathrm{S_3O_6^{2-}}, \mathrm{S_4^{2-}},\\
&\mathrm{S_4O_6^{2-}}, \mathrm{S_5^{2-}}, \mathrm{S_5O_6^{2-}}, \mathrm{SO_2(aq)}, \mathrm{SO_3^{2-}}, \mathrm{Sr(CO_3)(aq)}, \mathrm{Sr(HCO_3)^+}, \mathrm{SrCl^+}, \mathrm{SrOH^+},\\
&\mathrm{H_2O_2(aq)}, \mathrm{HClO_2(aq)}, \mathrm{NaSO_4^-}, \mathrm{MgSO_4(aq)}, \mathrm{HCl(aq)}, \mathrm{CaCO_3(aq)}, \mathrm{SrCO_3(aq)}, \mathrm{FeCl^{2+}})\\
\mathcal{E} = (&\mathrm{H}, \mathrm{O}, \mathrm{Na}, \mathrm{Mg}, \mathrm{S}, \mathrm{Ca}, \mathrm{K}, \mathrm{C}, \mathrm{Sr}, \mathrm{Cl}),
\end{aligned}$$

and the set $\mathcal{R}$ composed of the reactions:

$$\begin{aligned}
\mathrm{AlO^+} &= \mathrm{H_2O} - 2\mathrm{H^+} + \mathrm{Al^{3+}};\\
\mathrm{HAlO_2(aq)} &= 2\mathrm{H_2O} - 3\mathrm{H^+} + \mathrm{Al^{3+}};\\
\mathrm{CaOH^+} &= \mathrm{H_2O} - \mathrm{H^+} + \mathrm{Ca^{2+}};\\
\mathrm{CO_2(aq)} &= -\mathrm{H_2O} + \mathrm{H^+} + \mathrm{HCO_3^-};\\
\mathrm{CaHCO_3^+} &= \mathrm{Ca^{2+}} + \mathrm{HCO_3^-};\\
\mathrm{CaCl_2(aq)} &= \mathrm{Ca^{2+}} + 2\mathrm{Cl^-};\\
\mathrm{HClO(aq)} &= \mathrm{H^+} + 0.5\mathrm{O_2(aq)} + \mathrm{Cl^-};\\
\mathrm{ClO_2^-} &= \mathrm{O_2(aq)} + \mathrm{Cl^-};\\
\mathrm{ClO_4^-} &= 2\mathrm{O_2(aq)} + \mathrm{Cl^-};\\
\mathrm{FeCl^+} &= \mathrm{Fe^{2+}} + \mathrm{Cl^-};\\
\mathrm{FeOH^{2+}} &= 0.5\mathrm{H_2O} + 0.25\mathrm{O_2(aq)} + \mathrm{Fe^{2+}};\\
\mathrm{FeO^+} &= 0.5\mathrm{H_2O} - \mathrm{H^+} + 0.25\mathrm{O_2(aq)} + \mathrm{Fe^{2+}};\\
\mathrm{HFeO_2^-} &= 2\mathrm{H_2O} - 3\mathrm{H^+} + \mathrm{Fe^{2+}};
\end{aligned}$$

$$\begin{aligned}
\mathrm{AlOH^{2+}} &= \mathrm{H_2O} - \mathrm{H^+} + \mathrm{Al^{3+}};\\
\mathrm{AlO_2^-} &= 2\mathrm{H_2O} - 4\mathrm{H^+} + \mathrm{Al^{3+}};\\
\mathrm{CO(aq)} &= -\mathrm{H_2O} + \mathrm{H^+} - 0.5\mathrm{O_2(aq)} + \mathrm{HCO_3^-};\\
\mathrm{CO_3^{2-}} &= -\mathrm{H^+} + \mathrm{HCO_3^-};\\
\mathrm{CaCl^+} &= \mathrm{Ca^{2+}} + \mathrm{Cl^-};\\
\mathrm{CaSO_4(aq)} &= \mathrm{Ca^{2+}} + \mathrm{SO_4^{2-}};\\
\mathrm{ClO^-} &= 0.5\mathrm{O_2(aq)} + \mathrm{Cl^-};\\
\mathrm{ClO_3^-} &= 1.5\mathrm{O_2(aq)} + \mathrm{Cl^-};\\
\mathrm{Fe^{3+}} &= -0.5\mathrm{H_2O} + \mathrm{H^+} + 0.25\mathrm{O_2(aq)} + \mathrm{Fe^{2+}};\\
\mathrm{FeCl_2(aq)} &= \mathrm{Fe^{2+}} + 2\mathrm{Cl^-};\\
\mathrm{FeOH^+} &= \mathrm{H_2O} - \mathrm{H^+} + \mathrm{Fe^{2+}};\\
\mathrm{FeO} &= \mathrm{H_2O} - 2\mathrm{H^+} + \mathrm{Fe^{2+}};\\
\mathrm{HFeO_2(aq)} &= 1.5\mathrm{H_2O} - 2\mathrm{H^+} + 0.25\mathrm{O_2(aq)} + \mathrm{Fe^{2+}};
\end{aligned}$$

$$\mathrm{FeO_2^-} = 1.5\mathrm{H_2O} - 3\mathrm{H^+} + 0.25\mathrm{O_2(aq)} + \mathrm{Fe^{2+}};$$
$$\mathrm{H_2S(aq)} = 2\mathrm{H^+} - 2\mathrm{O_2(aq)} + \mathrm{SO_4^{2-}};$$
$$\mathrm{HS^-} = \mathrm{H^+} - 2\mathrm{O_2(aq)} + \mathrm{SO_4^{2-}};$$
$$\mathrm{HSO_4^-} = \mathrm{H^+} + \mathrm{SO_4^{2-}};$$
$$\mathrm{KCl(aq)} = \mathrm{K^+} + \mathrm{Cl^-};$$
$$\mathrm{KOH(aq)} = \mathrm{H_2O} - \mathrm{H^+} + \mathrm{K^+};$$
$$\mathrm{CH_4(aq)} = \mathrm{H_2O} + \mathrm{H^+} - 2\mathrm{O_2(aq)} + \mathrm{HCO_3^-};$$
$$\mathrm{Mg(HCO_3)^+} = \mathrm{Mg^{2+}} + \mathrm{HCO_3^-};$$
$$\mathrm{MgOH^+} = \mathrm{H_2O} - \mathrm{H^+} + \mathrm{Mg^{2+}};$$
$$\mathrm{NaOH(aq)} = \mathrm{H_2O} - \mathrm{H^+} + \mathrm{Na^+};$$
$$\mathrm{S_2^{2-}} = -\mathrm{H_2O} + 2\mathrm{H^+} - 3.5\mathrm{O_2(aq)} + 2\mathrm{SO_4^{2-}};$$
$$\mathrm{HS_2O_3^-} = -\mathrm{H_2O} + 3\mathrm{H^+} - 2\mathrm{O_2(aq)} + 2\mathrm{SO_4^{2-}};$$
$$\mathrm{S_2O_4^{2-}} = -\mathrm{H_2O} + 2\mathrm{H^+} - 1.5\mathrm{O_2(aq)} + 2\mathrm{SO_4^{2-}};$$
$$\mathrm{H_2S_2O_4(aq)} = -\mathrm{H_2O} + 4\mathrm{H^+} - 1.5\mathrm{O_2(aq)} + 2\mathrm{SO_4^{2-}};$$
$$\mathrm{S_2O_6^{2-}} = -\mathrm{H_2O} + 2\mathrm{H^+} - 0.5\mathrm{O_2(aq)} + 2\mathrm{SO_4^{2-}};$$
$$\mathrm{S_3^{2-}} = -2\mathrm{H_2O} + 4\mathrm{H^+} - 5\mathrm{O_2(aq)} + 3\mathrm{SO_4^{2-}};$$
$$\mathrm{S_4^{2-}} = -3\mathrm{H_2O} + 6\mathrm{H^+} - 6.5\mathrm{O_2(aq)} + 4\mathrm{SO_4^{2-}};$$
$$\mathrm{S_5^{2-}} = -4\mathrm{H_2O} + 8\mathrm{H^+} - 8\mathrm{O_2(aq)} + 5\mathrm{SO_4^{2-}};$$
$$\mathrm{SO_2(aq)} = -\mathrm{H_2O} + 2\mathrm{H^+} - 0.5\mathrm{O_2(aq)} + \mathrm{SO_4^{2-}};$$
$$\mathrm{Sr(HCO_3)^+} = \mathrm{HCO_3^-} + \mathrm{Sr^{2+}};$$
$$\mathrm{SrOH^+} = \mathrm{H_2O} - \mathrm{H^+} + \mathrm{Sr^{2+}};$$
$$\mathrm{HClO_2(aq)} = \mathrm{H^+} + \mathrm{O_2(aq)} + \mathrm{Cl^-};$$
$$\mathrm{MgSO_4(aq)} = \mathrm{Mg^{2+}} + \mathrm{SO_4^{2-}};$$
$$\mathrm{CaCO_3(aq)} = -\mathrm{H^+} + \mathrm{Ca^{2+}} + \mathrm{HCO_3^-};$$
$$\mathrm{FeCl^{2+}} = -0.5\mathrm{H_2O} + \mathrm{H^+} + 0.25\mathrm{O_2(aq)} + \mathrm{Fe^{2+}} + \mathrm{Cl^-}.$$
$$\mathrm{H_2(aq)} = \mathrm{H_2O} - 0.5\mathrm{O_2(aq)};$$
$$\mathrm{HO_2^-} = \mathrm{H_2O} - \mathrm{H^+} + 0.5\mathrm{O_2(aq)};$$
$$\mathrm{HSO_3^-} = \mathrm{H^+} - 0.5\mathrm{O_2(aq)} + \mathrm{SO_4^{2-}};$$
$$\mathrm{HSO_5^-} = \mathrm{H^+} + 0.5\mathrm{O_2(aq)} + \mathrm{SO_4^{2-}};$$
$$\mathrm{KHSO_4(aq)} = \mathrm{H^+} + \mathrm{K^+} + \mathrm{SO_4^{2-}};$$
$$\mathrm{KSO_4^-} = \mathrm{K^+} + \mathrm{SO_4^{2-}};$$
$$\mathrm{Mg(CO_3)(aq)} = -\mathrm{H^+} + \mathrm{Mg^{2+}} + \mathrm{HCO_3^-};$$
$$\mathrm{MgCl^+} = \mathrm{Mg^{2+}} + \mathrm{Cl^-};$$
$$\mathrm{NaCl(aq)} = \mathrm{Na^+} + \mathrm{Cl^-};$$
$$\mathrm{OH^-} = \mathrm{H_2O} - \mathrm{H^+};$$
$$\mathrm{S_2O_3^{2-}} = -\mathrm{H_2O} + 2\mathrm{H^+} - 2\mathrm{O_2(aq)} + 2\mathrm{SO_4^{2-}};$$
$$\mathrm{H_2S_2O_3(aq)} = -\mathrm{H_2O} + 4\mathrm{H^+} - 2\mathrm{O_2(aq)} + 2\mathrm{SO_4^{2-}};$$
$$\mathrm{HS_2O_4^-} = -\mathrm{H_2O} + 3\mathrm{H^+} - 1.5\mathrm{O_2(aq)} + 2\mathrm{SO_4^{2-}};$$
$$\mathrm{S_2O_5^{2-}} = -\mathrm{H_2O} + 2\mathrm{H^+} - \mathrm{O_2(aq)} + 2\mathrm{SO_4^{2-}};$$
$$\mathrm{S_2O_8^{2-}} = -\mathrm{H_2O} + 2\mathrm{H^+} + 0.5\mathrm{O_2(aq)} + 2\mathrm{SO_4^{2-}};$$
$$\mathrm{S_3O_6^{2-}} = -2\mathrm{H_2O} + 4\mathrm{H^+} - 2\mathrm{O_2(aq)} + 3\mathrm{SO_4^{2-}};$$
$$\mathrm{S_4O_6^{2-}} = -3\mathrm{H_2O} + 6\mathrm{H^+} - 3.5\mathrm{O_2(aq)} + 4\mathrm{SO_4^{2-}};$$
$$\mathrm{S_5O_6^{2-}} = -4\mathrm{H_2O} + 8\mathrm{H^+} - 5\mathrm{O_2(aq)} + 5\mathrm{SO_4^{2-}};$$
$$\mathrm{SO_3^{2-}} = -0.5\mathrm{O_2(aq)} + \mathrm{SO_4^{2-}};$$
$$\mathrm{SrCl^+} = \mathrm{Cl^-} + \mathrm{Sr^{2+}};$$
$$\mathrm{H_2O_2(aq)} = \mathrm{H_2O} + 0.5\mathrm{O_2(aq)};$$
$$\mathrm{NaSO_4^-} = \mathrm{Na^+} + \mathrm{SO_4^{2-}};$$
$$\mathrm{HCl(aq)} = \mathrm{H^+} + \mathrm{Cl^-};$$
$$\mathrm{SrCO_3(aq)} = -\mathrm{H^+} + \mathrm{HCO_3^-} + \mathrm{Sr^{2+}};$$

References

[1] C. Bale, E. Bélisle, P. Chartrand, S. Decterov, G. Eriksson, K. Hack, I.-H. Jung, Y.-B. Kang, J. Melançon, A. Pelton, C. Robelin, and S. Petersen, *FactSage thermochemical software and databases — recent developments*, Calphad, 33 (2009), pp. 295–311, https://doi.org/10.1016/j.calphad.2008.09.009.

[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*, Calphad, 26 (2002), pp. 189–228, https://doi.org/10.1016/S0364-5916(02)00035-4.

[3] 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, R. Klöfkorn, E. Keilegavlen, F. A. Radu, and J. Fuhrmann, eds., vol. 323 of Springer Proceedings in Mathematics & Statistics, Cham, June 2020, Springer, https://doi.org/10.1007/978-3-030-43651-3_35.

[4] C.-L. Berthollet, *Essai de statique chimique*, Firmin Didot, Paris, 1803.

[5] K. Brenner and C. Cancès, *Improving Newton's method performance by parametrization: The case of the Richards equation*, SIAM Journal on Numerical Analysis, 55 (2017), pp. 1760–1785, https://doi.org/10.1137/16M1083414.

[6] K. Brenner, M. Groza, L. Jeannin, R. Masson, and J. Pellerin, *Immiscible two-phase Darcy flow model accounting for vanishing and discontinuous capillary pressures: application to the flow in fractured porous media*, Computational Geosciences, 21 (2017), pp. 1075–1094, https://doi.org/10.1007/s10596-017-9675-7.

[7] S. R. Brinkley, *Calculation of the equilibrium composition of systems of many constituents*, The Journal of Chemical Physics, 15 (1947), pp. 107–110, https://doi.org/10.1063/1.1746420.

[8] J. Coatléven and A. Michel, *A successive substitution approach with embedded phase stability for simultaneous chemical and phase equilibrium calculations*, Computers & Chemical Engineering, 168 (2022), p. 108041, https://doi.org/10.1016/j.compchemeng.2022.108041.

[9] J. Connolly, *Computation of phase equilibria by linear programming: A tool for geodynamic modeling and its application to subduction zone decarbonation*, Earth and Planetary Science Letters, 236 (2005), pp. 524–541, https://doi.org/10.1016/j.epsl.2005.04.033.

[10] 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*, Journal of Metamorphic Geology, 20 (2002), pp. 697–708, https://doi.org/10.1046/j.1525-1314.2002.00398.x.

[11] M. A. Cook, *The science of high explosives*, vol. 139 of American Chemical Society Monoggraph, Reinhold Publishing Corporation, New York, 1958.

[12] G. B. Dantzig and J. C. DeHaven, *On the reduction of certain multiplicative chemical equilibrium systems to mathematically equivalent additive systems*, The Journal of Chemical Physics, 36 (1962), pp. 2620–2627, https://doi.org/10.1063/1.1732342.

[13] C. de Capitani and T. H. Brown, *The computation of chemical equilibrium in complex systems containing non-ideal solutions*, Geochimica et Cosmochimica Acta, 51 (1987), pp. 2639–2652, https://doi.org/10.1016/0016-7037(87)90145-1.

[14] C. de Capitani and K. Petrakakis, *The computation of equilibrium assemblage diagrams with Theriak/Domino software*, American Mineralogist, 95 (2010), pp. 1006–1016, https://doi.org/10.2138/am.2010.3354.

[15] H. J. G. Diersch and P. Perrochet, *On the primary variable switching technique for simulating unsaturated–saturated flows*, Advances in Water Resources, 23 (1999), pp. 271–301, https://doi.org/10.1016/S0309-1708(98)00057-8, https://www.sciencedirect.com/science/article/pii/S0309170898000578 (accessed 2023-09-12).

[16] G. Eriksson and K. Hack, *ChemSage—A computer program for the calculation of complex chemical equilibria*, Metallurgical Transactions B, 21 (1990), pp. 1013–1023, https://doi.org/10.1007/BF02670272.

[17] G. Eriksson and W. T. Thompson, *A procedure to estimate equilibrium concentrations in multicomponent systems amd related applications*, Calphad, 13 (1989), pp. 389–400, https://doi.org/https://doi.org/10.1016/0364-5916(89)90027-8.

[18] J. W. Gibbs, *A method of geometrical representation of the thermodynamic properties of substances by means of surfaces*, Transactions of the Connecticut Academy of Arts and Sciences, 2 (1873), pp. 382–404, https://www3.nd.edu/~powers/ame.20231/gibbs1873b.pdf.

[19] I. K. Karpov, *The convex programming minimization of five thermodynamic potentials other than Gibbs energy in geochemical modeling*, American Journal of Science, 302 (2002), pp. 281–311, https://doi.org/10.2475/ajs.302.4.281.

[20] I. K. Karpov, K. V. Chudnenko, and D. A. Kulik, *Modeling chemical mass transfer in geochemical processes; thermodynamic relations, conditions of equilibria and numerical algorithms*, American Journal of Science, 297 (1997), pp. 767–806, https://doi.org/10.2475/ajs.297.8.767.

[21] I. K. Karpov, K. V. Chudnenko, D. A. Kulik, and O. Avchenko, *Minimization of Gibbs free energy in geochemical systems by convex programming*, Geochemistry International, 39 (2001), pp. 1108–1119, http://repository.geologyscience.ru/handle/123456789/24918.

[22] C. T. Kelley, *Solving nonlinear equations with Newton's method*, Fundamentals of algorithms, Society for Industrial and Applied Mathematics, Philadelphia, 2003, https://doi.org/10.1137/1.9780898718898.

[23] K. A. Kobe and T. W. Leland, *The calculation of chemical equilibrium in a complex system*, Special Publication No. 26 of the Bureau of Engineering Research, The University of Texas, Austin, TX., (1954).

[24] 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*, Computational Geosciences, 17 (2012), pp. 1–24, https://doi.org/10.1007/s10596-012-9310-6.

[25] A. M. M. Leal, *Reaktoro: A unified framework for modeling chemically reactive systems*, 2015.

[26] 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*, Pure and Applied Chemistry, 89 (2017), pp. 597–643, https://doi.org/10.1515/pac-2016-1107.

[27] 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.

[28] D. Nordstrom and K. Campbell, *Modeling low-temperature geochemical processes*, in Treatise on Geochemistry, Elsevier, 2014, pp. 27–68, https://doi.org/10.1016/B978-0-08-095975-7.00502-7.

[29] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, vol. 30 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, 2000, https://doi.org/10.1137/1.9780898719468.

[30] W. H. Press and S. A. Teukolsky, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, Cambridge University Press, 2007, http://nrbook.com.

[31] J. Revels, M. Lubin, and T. Papamarkou, *Forward-mode automatic differentiation in Julia*, 2016, https://arxiv.org/abs/1607.07892.

[32] N. Z. Shapiro and L. S. Shapley, *Mass action laws and the Gibbs free energy function*, Journal of the Society for Industrial and Applied Mathematics, 13 (1965), pp. 353–375, https://doi.org/10.1137/0113020.

[33] Y. V. Shvarov, *Algorithmization of the numeric equilibrium modeling of dynamic geochemical processes*, Geochemistry International, 37 (1999), pp. 571–576.

[34] Y. V. Shvarov, *HCh: New potentialities for the thermodynamic simulation of geochemical systems offered by windows*, Geochemistry International, 46 (2008), pp. 834–839, https://doi.org/10.1134/S0016702908080089.

[35] W. R. Smith, *The computation of chemical equilibria in complex systems*, Industrial & Engineering Chemistry Fundamentals, 19 (1980), pp. 1–10, https://doi.org/10.1021/i160073a001.

[36] W. R. Smith and R. W. Missen, *Chemical reaction equilibrium analysis: Theory and algorithms*, John Wiley & Sons, New York, 1982.

[37] G. P. Sutton, *Rocket propulsion elements: an introduction to engineering of rockets*, John Wiley & Sons, New York, 3 ed ed., 1963.

[38] D. C. Thorstenson, *The concept of electron activity and its relation to redox potentials in aqueous geochemical systems*, Tech. Report 84-072, US Geological survey, 1984.

[39] C. Tsanas, E. H. Stenby, and W. Yan, *Calculation of multiphase chemical equilibrium by the modified RAND method*, Industrial & Engineering Chemistry Research, 56 (2017), pp. 11983–11995, https://doi.org/10.1021/acs.iecr.7b02714.

[40] C. Tsanas, E. H. Stenby, and W. Yan, *Calculation of simultaneous chemical and phase equilibrium by the method of Lagrange multipliers*, Chemical Engineering Science, 174 (2017), pp. 112–126, https://doi.org/10.1016/j.ces.2017.08.033.

[41] 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*, The Canadian Journal of Chemical Engineering, 47 (1969), pp. 81–84, https://doi.org/10.1002/cjce.5450470115.

[42] F. van Zeggeren and S. H. Storey, *The computation of chemical equilibria*, Cambridge University Press, London, 1970.

[43] P. Waage and C. M. Gulberg, *Studies concerning affinity*, Journal of Chemical Education, 63 (1986), pp. 1044–1047, https://doi.org/10.1021/ed063p1044.

[44] T. Wagner, D. A. Kulik, F. F. Hingerl, and S. V. Dmytrieva, *GEM-Selektor geochemical modeling package: TSolMod library and data interface for multicomponent phase models*, The Canadian Mineralogist, 50 (2012), pp. 1173–1195, https://doi.org/10.3749/canmin.50.5.1173.

[45] 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. Report UCRL-MA–110662-Pt.3, 138643, Sept. 1992, https://doi.org/10.2172/138643.