Parametrization and Cartesian representation techniques for robust resolution of chemical equilibria

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

Parametrization and Cartesian representation techniques for robust resolution of chemical equilibria

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

April 30, 2024

Abstract

Chemical equilibria computations, especially those with vanishing species in the aqueous phase, lead to nonlinear systems that are difficult to solve due to gradient blow up. Instead of the commonly used *ad hoc* treatments, we propose two reformulations of the single-phase chemical equilibrium 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 our methods combined with a globalization strategy is superior to that of the literature.

Keywords Chemical equilibria, Newton's method, parametrization, Cartesian representation

1 Introduction

The simulation of reactive transport poses a significant challenge in various fields, including flows in porous media, combustion in engines and gas turbines, and the design of chemical reactors. In particular, the computation of reactive transport in porous media plays a central role in CO_2 and H_2 storage 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 within each cell of the mesh. Consequently, even a slight enhancement in their resolution could directly and positively impact overall performance.

In chemical modeling, reactions primarily fall into two categories: equilibrium reactions and kinetic reactions. Our focus lies specifically on equilibrium reactions. Given specified quantities of chemical elements, along with pressure and temperature parameters, a chemical equilibrium calculation involves determining the amounts of chemical species that minimize a state function – known as Gibbs free energy – while adhering to the conservation of the quantity of matter. Smith and Missen [25] proposed a classification of different approaches to tackle this problem into two categories: stoichiometric methods and non-stoichiometric methods. Stoichiometric methods use the mass action equations while non-stoichiometric methods use the minimization of the Gibbs free energy. Although our approach is based on Gibbs energy minimization, we reformulate our equations so that we use the law of mass action, making it a stoichiometric method. We can frame this problem as a coupled system of equations, encapsulating the principles of mass conservation and chemical equilibrium. The equations governing mass conservation are linear in the species quantities, while the chemical equilibrium equations are expressed as functions of their logarithms. As reviewed by Leal et al. [15], many geochemical codes use this kind of method, including EQ3/6 [32], PHREEQC [18], WATEQ [27], MINEQL [31], CHESS [30], CHIM-XPT [20] and SOLVED-XPT [21]. For the non-stoichiometric methods, a non-exhaustive list of code using this method includes ChemSage [10], THERIAK [8], HCh [24], FactSage [1], PERPLEX [7, 6], GEM-Selektor [13] and

[*]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. Corresponding authors: clement.cances@inria.fr, maxime.jonval@inria.fr

Reaktoro [14]. Subsequent developments in this field have been reviewed by Tsanas et al. [29, 28], and Coatléven and Michel [5].

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 solution values span several orders of magnitude, leading to conditioning issues; the convergence of the algorithm depends on the distance between the initial solution guess and the true solution. A classical technique relies on using the logarithms of the species quantities as unknowns to manage the positivity constraint and reduce the solution span between species. However, for species present in large quantities, such as solvents, it is preferable to use the quantities of the species as unknowns.

In this article, we introduce and analyze two algorithms aimed at addressing these concerns. First, we use the parametrization technique, as developed by Brenner and Cancès [3], to automatically switch between the two formulations, while ensuring that the partial derivatives of the Jacobian remain bounded [4, 2]. Second, we propose a well-balanced Cartesian representation that includes both the species quantities and their logarithms as unknowns. An additional function is introduced to establish the relationship between these two quantities, possessing properties that enable the resolution of the aforementioned issues and control over the derivatives of the Jacobian.

Section 2 presents the mathematical modeling of the chemical equilibrium problem and the existence and uniqueness of our formulation is established. 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, we present the results of numerical experiments validating our methods and comparing their robustness against three test cases. Section 6 concludes and opens to future works.

2 Mathematical description of the chemical equilibrium problem

This section introduces the chemical system and the notions of thermodynamics required to derive the chemical equilibrium problem and the resulting equations.

2.1 Chemical system

The type of system considered in this article involves diluted solutions of aqueous species. These solutions are composed of a predominant species called the solvent, typically water. Additionally, there are diluted aqueous species present in very small quantities. For a given temperature T and pressure P, such a chemical system $\mathcal{S}_{\mathrm{P,T}} = \{\mathcal{C}, \mathcal{E}, \mathcal{R}\}$ 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 the s_{ij} are the stoichiometric coefficients that represent the number of molecules of the species C_i involved in the reaction R_j.

The 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 H^+, OH^- and H_2O verifying the equilibrium reaction

$$H_2O = H^+ + OH^-.$$

The corresponding formula matrix is

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

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 stoichiometry matrix for the example of dissociation of water is

$$\mathbf{S} = \begin{array}{c} \begin{array}{c} R_1 \end{array} \\ \left[\begin{array}{c} 1 \\ 1 \\ -1 \end{array}\right] \begin{array}{c} H^+ \\ OH^- \\ H_2O \end{array} \end{array}$$

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

Lemma 2.1. *One has the following result:*

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

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

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

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

More details on the stoichiometry matrix and its link with the formula matrix can be found in the book of Smith and Missen [25].

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 $k \in \{1, \ldots, M\}$, $\sum_{i=1}^N a_{ki} n_i = 0$. 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 $\mathcal{S}_{\mathrm{P,T}}$ at constant pressure and temperature can be described by the Gibbs free energy function $G : \mathbb{R}_+^N \to \mathbb{R}$, also known as the Gibbs energy. 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 [16, 32], 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 [25]).

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 $\mathcal{S}_{\mathrm{P,T}}$ which minimizes, for a fixed temperature T, pressure P and element quantities $\mathbf{b}$, the function G, under constraints of element conservation and nonnegativity. To describe this calculation as a constrained minimization problem, let

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

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

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

The single-phase chemical equilibrium problem can be written as

$$\min_{\mathbf{n} \in \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}} G(\mathbf{n}). \tag{4}$$

The existence and uniqueness of a solution to the problem (4) for a multiphase ideal system has been studied by Shapiro and Shapley in [23]. In particular, they proved in Theorem 9.9 and Corollary 12.3 that for a single-phase ideal system, the problem (4) admits a unique solution assuming the compactness of the set $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$.

In this subsection, we prove the convexity of $\mathcal{M}_{\mathbf{A},\mathbf{b}}$ and the compactness of the space $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$ in Lemma 2.3, allowing to establish the existence and uniqueness of a solution in Lemma 2.4 together with a reformulation of the results of Theorem 9.2 from [23] which indicates that the inequality constraint is never saturated for this point. This result allows us to establish in Proposition 2.1 the system of equations to be solved to find the minimum of the problem and, from the strict convexity of G on $\mathcal{M}_{\mathbf{A},\mathbf{b}}$ proved in [25], the equivalence between this system and the minimization problem (4).

Proposition 2.1. *Assuming that $\mathcal{M}_{\mathbf{A},\mathbf{b}}$ is nonempty, there exists a unique solution $\mathbf{n} \in \mathcal{M}_{\mathbf{A},\mathbf{b}}$ of problem* (4) *which coincides with the unique solution of the system*

$$\begin{aligned} \mathbf{An} &= \mathbf{b}, \\ \mathbf{S}^T \boldsymbol{\mu}(\mathbf{n}) &= \mathbf{0}, \end{aligned} \tag{5}$$

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

To prove this proposition, we begin by looking at the properties of the $\mathcal{M}_{\mathbf{A},\mathbf{b}}$ set.

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

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

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

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

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

Lemma 2.4 provides the existence of a minimizer, which is furthermore unique and belongs to $\mathcal{M}_{\mathbf{A},\mathbf{b}}$.

Lemma 2.4. *There exists a unique minimum of problem* (4) *and this minimum is in $\mathcal{M}_{\mathbf{A},\mathbf{b}}$.*

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 $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$. As mentioned above, its uniqueness is demonstrated in [23] and follows from the strict convexity of G on $\mathcal{M}_{\mathbf{A},\mathbf{b}}$ from [25]. Let us prove that it belongs to $\mathcal{M}_{\mathbf{A},\mathbf{b}}$. Let $\mathbf{n}^\star \in \{\arg\min G(\mathbf{n}) \mid \mathbf{n} \in \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}\}$ be this minimum. Let us assume that $\mathbf{n}^\star \in \overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}} \setminus \mathcal{M}_{\mathbf{A},\mathbf{b}}$, meaning that there exists $\mathcal{J} \subsetneq \{1, \ldots, N\}$ such that $n_j^\star = 0$, $\forall j \in \mathcal{J}$. Note that the case $\mathcal{J} = \{1, \ldots, N\}$ is not allowed since $\mathbf{b} > 0$. Let $\mathbf{n} \in \mathcal{M}_{\mathbf{A},\mathbf{b}}$ which is assumed to be nonempty and $\varepsilon \in (0, 1)$, then one defines $\mathbf{n}^0 := \mathbf{n} - \mathbf{n}^\star$ and $\mathbf{n}^\varepsilon := \mathbf{n}^\star + \varepsilon \mathbf{n}^0$. 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 $\overline{\Omega}$,

$$G(\mathbf{n}^\star) \geq G(\mathbf{n}^\varepsilon) + \langle \boldsymbol{\mu}(\mathbf{n}^\varepsilon), \mathbf{n}^\star - \mathbf{n}^\varepsilon \rangle \Leftrightarrow \frac{G(\mathbf{n}^\star) - G(\mathbf{n}^\varepsilon)}{\varepsilon} \geq -\langle \boldsymbol{\mu}(\mathbf{n}^\varepsilon), \mathbf{n}^0 \rangle, \tag{6}$$

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 (6). In the right-hand side one has

$$\lim_{\varepsilon \to 0} -\langle \boldsymbol{\mu}(\mathbf{n}^\varepsilon), \mathbf{n}^0 \rangle = -\sum_{j \in \mathcal{J}} n_j^0 \lim_{\varepsilon \to 0} \mu_j(\mathbf{n}^\varepsilon) - \sum_{i=1,\, i \notin \mathcal{J}}^{N} n_i^0 \lim_{\varepsilon \to 0} \mu_i(\mathbf{n}^\varepsilon). \tag{7}$$

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

$$\lim_{\varepsilon \to 0} \mu_j(\mathbf{n}^\varepsilon) = -\infty, \, \forall j \in \mathcal{J} \quad \text{and} \quad \lim_{\varepsilon \to 0} \mu_i(\mathbf{n}^\varepsilon) = \mu_i(\mathbf{n}^\star) \in \mathbb{R}, \, \forall i \in \{1, \ldots, N\} \setminus \mathcal{J}. \tag{8}$$

By combining (7) and (8) with $n_j^0 = n_j > 0$, $\forall j \in \mathcal{J}$, one finds that the right-hand side of (6) tends to $+\infty$. However if $\mathbf{n}^\star$ minimises G on $\overline{\mathcal{M}_{\mathbf{A},\mathbf{b}}}$, then the left-hand side of (6) is non-positive which is a contradiction. Therefore $\mathcal{J} = \emptyset$ and $\mathbf{n}^\star \in \mathcal{M}_{\mathbf{A},\mathbf{b}}$. □

Thanks to Lemma 2.4, the problem can be simplified into

$$\min_{\mathbf{An}=\mathbf{b}} G(\mathbf{n}). \tag{9}$$

The first order optimality conditions of (9) are given by the Euler-Lagrange equations which state that if $\mathbf{n}^\star$ is the unique solution of the problem (9), it must satisfy

$$\mathbf{An}^\star - \mathbf{b} = \mathbf{0}, \tag{10}$$

$$\boldsymbol{\nabla} G(\mathbf{n}^\star) + \mathbf{A}^T \boldsymbol{\Lambda} = \mathbf{0}, \tag{11}$$

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

In the case we are considering, we can simplify the equations (10)–(11) by eliminating the Lagrange multipliers. To do so, we multiply (11) by $\mathbf{S}^T$ and as shown by Lemma 2.1, the matrix product $\mathbf{S}^T\mathbf{A}^T$ vanishes. Therefore, (11) becomes $\mathbf{S}^T\boldsymbol{\nabla} G(\mathbf{n}^\star) = \mathbf{0}$ and denoting by $\boldsymbol{\mu} = \boldsymbol{\nabla} G$ the vector of chemical potentials, we obtain the system (5).

We can now prove that the solution of this system is the same as that of the constrained minimization problem.

Proof of Proposition 2.1. The existence of a solution to (5) is guaranteed by the existence of a solution to (9). This solution is unique for the problem (9) but it remains to show that it is the only one to satisfy the system (5). To do so, we assume the existence of $\mathbf{n}_1^\star, \mathbf{n}_2^\star \in \Omega$ that satisfy (5). 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 $\mathcal{M}_{\mathbf{A},\mathbf{b}}$ (see [25]), 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 (5), 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 $\mathbf{Ax} = \omega\mathbf{b}$, 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

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

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* (5) *and its solution is unique.*

Proof. Let $\mathbf{n}^\star$ be the unique solution of (5) 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 (5). 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 $\mathbf{x}^1 = \mathbf{x}^2$. □

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 [17], 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

$$\begin{aligned} \mathbf{A}\mathbf{x} - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T\mathbf{y}(\mathbf{x}) &= \mathbf{d}, \\ \langle \mathbf{x}, \mathbf{1} \rangle &= 1, \end{aligned} \tag{14}$$

where $(\mathbf{y}(\mathbf{x}))_i = y(x_i) = \ln x_i$. The resolution of the system (14) 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:

$$\mathcal{G}(\mathbf{x}, \omega) := \mathcal{F}(\mathbf{x}, \mathbf{y}(\mathbf{x}), \omega) = \begin{pmatrix} \mathbf{A}\mathbf{x} - \omega\mathbf{b} \\ \mathbf{S}^T\mathbf{y}(\mathbf{x}) - \mathbf{d} \\ \langle \mathbf{x}, \mathbf{1} \rangle - 1 \end{pmatrix}. \tag{15}$$

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

$$\boldsymbol{\nabla}\mathcal{G}(\mathbf{u}^{(k)})\boldsymbol{\delta}\mathbf{u}^{(k)} = -\mathcal{G}(\mathbf{u}^{(k)}), \tag{16}$$

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

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \boldsymbol{\delta}\mathbf{u}^{(k)}. \tag{17}$$

In (16), $\boldsymbol{\nabla}\mathcal{G}(\mathbf{u}^{(k)})$ 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 [12]. It requires the following assumptions:

1. The equation (15) 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
$$\|\boldsymbol{\nabla}\mathcal{G}(\mathbf{u}_1) - \boldsymbol{\nabla}\mathcal{G}(\mathbf{u}_2)\|_2 \leq L\|\mathbf{u}_1 - \mathbf{u}_2\|_2$$
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 [12, 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* (16)–(17) *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*

$$\|\mathbf{u}^{(k+1)} - \mathbf{u}^\star\|_2 \leq C\|\mathbf{u}^{(k)} - \mathbf{u}^\star\|_2^2, \ \forall k \geq k_C. \tag{18}$$

The property (18) together with $\mathbf{u}^{(k)} \to \mathbf{u}^\star$ is referred to as q-quadratic convergence in the monograph [12].

3.2 A family of parametrizations

The Newton's method applied to the function (15) yields the following Jacobian matrix:

$$\boldsymbol{\nabla}\mathcal{G}(\mathbf{x}, \omega) = \begin{bmatrix} \mathbf{A} & -\mathbf{b} \\ \mathbf{S}^T\boldsymbol{\nabla}\mathbf{y}(\mathbf{x}) & \mathbf{0} \\ \mathbf{1}^T & 0 \end{bmatrix}, \tag{19}$$

where $\boldsymbol{\nabla}\mathbf{y}(\mathbf{x}) = \operatorname{diag}\{1/x_i\}_{i=1,\dots,N}$. The jacobian in (19) 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, known as the log trick [32], 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

$$\mathcal{H}(\mathbf{y},\omega) := \mathcal{F}(\mathbf{x}(\mathbf{y}),\mathbf{y},\omega) = \mathbf{0} \Leftrightarrow \begin{aligned} \mathbf{A}\mathbf{x}(\mathbf{y}) - \mathbf{b} &= \mathbf{0},\\ \mathbf{S}^T\mathbf{y} - \mathbf{d} &= \mathbf{0},\\ \langle \mathbf{x}(\mathbf{y}), \mathbf{1}\rangle - 1 &= 0, \end{aligned} \tag{20}$$

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 (15). In this case the Jacobian matrix becomes

$$\boldsymbol{\nabla}\mathcal{H}(\mathbf{y},\omega) = \begin{bmatrix} \mathbf{A}\boldsymbol{\nabla}\mathbf{x}(\mathbf{y}) & -\mathbf{b} \\ \mathbf{S}^T & \mathbf{0} \\ \mathbf{1}^T\boldsymbol{\nabla}\mathbf{x}(\mathbf{y}) & 0 \end{bmatrix}, \tag{21}$$

where $\boldsymbol{\nabla}\mathbf{x}(\mathbf{y}) = \operatorname{diag}\{\exp y_i\}_{i=1,\dots,N}$. The Jacobian in (21) 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

$$\Gamma = \{(x,y) \in \mathbb{R}^2 \mid y = \ln(x)\} \tag{22}$$

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 nonlinear 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:

$$(|X'(\tau)|^p + |Y'(\tau)|^p)^{1/p} = 1, \quad p \geq 1. \tag{23}$$

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

$$Y'(\tau) = X'(\tau)/X(\tau) \tag{24}$$

from the condition (A1). Indeed, combining (23) and (24), we get:

$$|X'(\tau)|^p + |X'(\tau)/X(\tau)|^p = 1 \Leftrightarrow |X'(\tau)| = \frac{1}{(1 + |1/X(\tau)|^p)^{1/p}} \tag{25}$$

and

$$|Y'(\tau)|^p = 1 - |X'(\tau)|^p \Leftrightarrow |Y'(\tau)| = \frac{1/X(\tau)}{(1 + |1/X(\tau)|^p)^{1/p}}. \tag{26}$$

Furthermore, equation (26) can be expressed in terms of the function Y. To do this, we multiply (24) by $\exp(Y(\tau))$ and using the derivative

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

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

$$X'(\tau) = \exp(Y(\tau))X'(\tau)/X(\tau) \Leftrightarrow \exp'(Y(\tau))/X(\tau) = 1. \tag{27}$$

Thus, the equations (25), (26) and (27) enable us to express the following system of differential equations:

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

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

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 (23) then becomes

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

and the system (28)-(29) is rewritten as

$$X'(\tau) = \pm\frac{1}{\max(1, |1/X(\tau)|)}, \tag{30}$$

$$Y'(\tau) = \pm\frac{1/X(\tau)}{\max(1, |1/X(\tau)|)} = \pm\frac{1}{\max(1, |\exp(Y(\tau))|)}. \tag{31}$$

Since the logarithm function is increasing and strictly concave, a solution of this latter system is given by the following statement, *cf.* [3].

Proposition 3.1. *A solution of* (30)-(31) *is given by*

$$(X(\tau), Y(\tau)) = \begin{cases} (\exp(\tau),\ \tau) & \text{if} \quad \tau < 0, \\ (\tau + 1,\ \ln(\tau + 1)) & \text{if} \quad \tau \geq 0. \end{cases} \tag{32}$$

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 [9]. Figure 1 illustrates these functions.

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

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

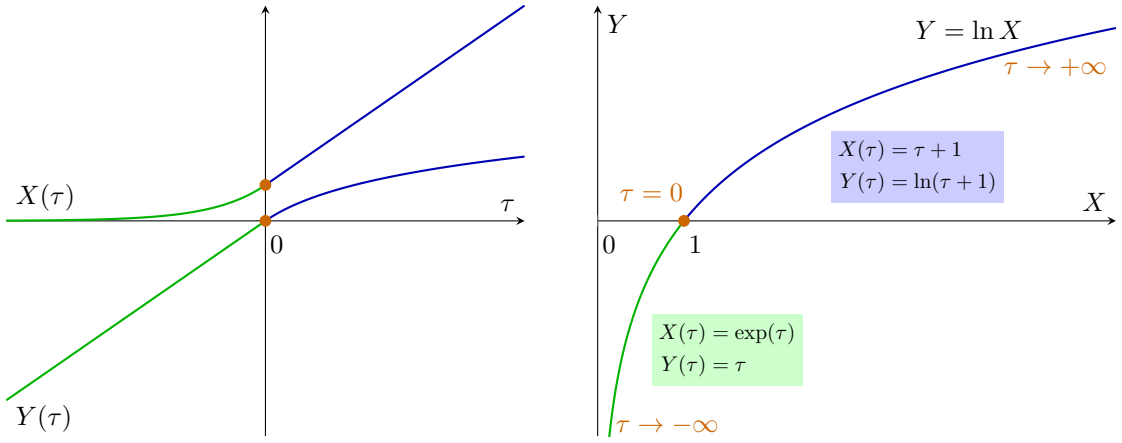

Figure 1: The switch function.

The corresponding Jacobian matrices are

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

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 simultaneously.

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:

$$(|H'(y)|^p + |G'(x)|^p)^{1/p} = 1, \quad p \geq 1,\ y = \ln(x). \tag{33}$$

Using the same reasoning as we did for the parametrization, we can combine equations (33) with the derivative

$$G'(x) = H'(\ln(x))/x, \tag{34}$$

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

$$G'(x) = \pm \frac{1/x}{(1 + |1/x|^p)^{1/p}}, \tag{35}$$

$$H'(v) = \pm \frac{1}{(1 + |1/\exp(y)|^p)^{1/p}} = \pm \frac{\exp(y)}{(1 + |\exp(y)|^p)^{1/p}}. \tag{36}$$

The case of interest for numerical experiments is that of the limit $p \to \infty$, the condition (33) then becomes

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

while the differential equations (35)–(36) become

$$G'(x) = \pm \frac{1/x}{\max(1, |1/x|)}, \tag{37}$$

$$H'(y) = \pm \frac{1}{\max(1, |1/\exp(v)|)} = \pm \frac{\exp(v)}{\max(1, |\exp(v)|)}. \tag{38}$$

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 (34), 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 $(x, y) \in \mathbb{R}^2$,*

$$\begin{aligned} G'(x) &= Y'(X^{-1}(x)), \\ H'(y) &= X'(Y^{-1}(y)). \end{aligned} \tag{39}$$

This Cartesian representation is admissible in the sens of (H1)–(H3). Moreover, it satisfies the normalization (33) *if the parametrization satisfies the normalization* (23)*.*

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 (39) 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,*

$$\begin{aligned} X'(\tau) &= H'(Y(\tau)), \\ Y'(\tau) &= G'(X(\tau)). \end{aligned} \tag{40}$$

This parametrization is admissible in the sense of (A1)–(A3) and satisfies the normalization (23) *if the Cartesian representation satisfies the normalization* (33).

Proof. The existence of a solution to the ODE (40) 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 (37)-(38) 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

$$H(y) = y\mathbf{1}_{\{y>0\}} + (\exp(y) - 1)\mathbf{1}_{\{y\leq 0\}}. \tag{41}$$

It follows that:

$$G(x) = H(\ln x) = \ln x \mathbf{1}_{\{x>1\}} + (x-1)\mathbf{1}_{\{x\leq 1\}}. \tag{42}$$

The function f is then defined in four areas as

$$f(x,y) = \begin{cases} \mathrm{e}^y - x, & \text{if } x \leq 1,\ y \leq 0, \\ y - x + 1, & \text{if } x \leq 1,\ y \geq 0, \\ y - \ln x, & \text{if } x \geq 1,\ y \geq 0, \\ \mathrm{e}^y - \ln x - 1, & \text{if } x \geq 1, y \leq 0. \end{cases} \tag{43}$$

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.

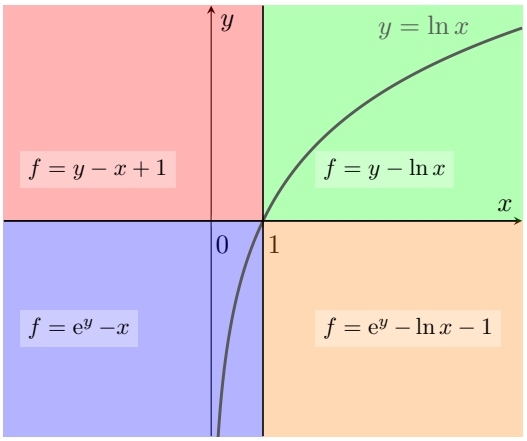

Figure 2: The discrepancy function.

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

$$\begin{aligned} \mathbf{A}\mathbf{X}(\boldsymbol{\tau}) - \omega \mathbf{b} &= \mathbf{0}, \\ \mathfrak{W}(\boldsymbol{\tau}, \omega) = \mathbf{0} \Leftrightarrow \mathbf{S}^T[\boldsymbol{\mu}^\circ/(RT) + \mathbf{Y}(\boldsymbol{\tau})] &= \mathbf{0}, \\ \langle \mathbf{X}(\boldsymbol{\tau}), \mathbf{1} \rangle - 1 &= 0. \end{aligned} \tag{44}$$

The Jacobian matrix $\boldsymbol{\nabla}\mathfrak{W} = \boldsymbol{\nabla}\mathfrak{W}(\boldsymbol{\tau}, \omega)$, associated to (44), 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 (44), 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

$$\mathbf{A}\mathrm{diag}\{\mathbf{X}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau} - \delta\omega\mathbf{b} = \mathbf{0}, \tag{45}$$

$$\mathbf{S}^T\mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau} = \mathbf{0}, \tag{46}$$

$$\langle \mathbf{X}'(\boldsymbol{\tau}), \boldsymbol{\delta\tau} \rangle = 0. \tag{47}$$

Since $(\boldsymbol{\tau}, \omega)$ is solution of (44), one has $\mathbf{b} = \mathbf{A}\mathbf{X}(\boldsymbol{\tau})/\omega$ with $\omega > 0$. The equation (45) then becomes

$$\mathbf{A}\left[\mathrm{diag}\{\mathbf{X}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau} - \frac{\delta\omega}{\omega}\mathbf{X}(\boldsymbol{\tau})\right] = \mathbf{0}. \tag{48}$$

Moreover, equation (46) 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 (48), the following equality holds:

$$\begin{aligned} \langle \mathrm{diag}\{\mathbf{X}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau}, \mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau} \rangle &= \frac{\delta\omega}{\omega} \langle \mathbf{X}(\boldsymbol{\tau}), \mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\boldsymbol{\delta\tau} \rangle \\ &= \frac{\delta\omega}{\omega} \langle \mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\mathbf{X}(\boldsymbol{\tau}), \boldsymbol{\delta\tau} \rangle . \end{aligned} \tag{49}$$

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

$$\mathbf{X}'(\boldsymbol{\tau}) = \mathrm{diag}\{\mathbf{Y}'(\boldsymbol{\tau})\}\mathbf{X}(\boldsymbol{\tau}). \tag{50}$$

One deduces that the right-hand side of (49) vanishes thanks to (47). 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 (50), 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 (45) 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* (16)–(17) *is applied to the function $\mathfrak{W}$ defined as* (44)*, 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

$$\mathfrak{H}(\mathbf{x}, \mathbf{y}, \omega) = \mathbf{0} \Leftrightarrow \begin{aligned} \mathbf{A}\mathbf{x} - \omega\mathbf{b} &= \mathbf{0}, \\ \mathbf{S}^T[\boldsymbol{\mu}^\circ/(RT) + \mathbf{y}] &= \mathbf{0}, \\ \boldsymbol{f}(\mathbf{x}, \mathbf{y}) &= \mathbf{0}, \\ \langle \mathbf{x}, \mathbf{1}\rangle - 1 &= 0. \end{aligned} \tag{51}$$

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 (51), 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* (51)*, 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 (51), 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

$$\mathbf{A}\boldsymbol{\delta}\mathbf{x}_1 = \mathbf{b}, \tag{52}$$

$$\mathbf{S}^T\boldsymbol{\delta}\mathbf{y}_1 = \mathbf{0}, \tag{53}$$

$$\boldsymbol{\nabla}_{\mathbf{x}}\boldsymbol{f}\boldsymbol{\delta}\mathbf{x}_1 + \boldsymbol{\nabla}_{\mathbf{y}}\boldsymbol{f}\boldsymbol{\delta}\mathbf{y}_1 = \mathbf{0}. \tag{54}$$

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 (54). Furthermore from (53) 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 (52) can be rewritten as

$$\mathbf{A}\mathbf{D}\mathbf{A}^T\boldsymbol{\delta}\mathbf{h}_1 = \mathbf{b}. \tag{55}$$

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 (51), then from (55) 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

$$\begin{aligned} \mathbf{D}\mathbf{A}^T\boldsymbol{\delta}\mathbf{h}_1 = \mathbf{D}\boldsymbol{\delta}\mathbf{y}_1 = \boldsymbol{\delta}\mathbf{x}_1 &= \frac{1}{\omega}\mathbf{D}\mathbf{A}^T(\mathbf{A}\mathbf{D}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{x} \\ &= \frac{1}{\omega}\mathbf{D}^{1/2}\left[(\mathbf{A}\mathbf{D}^{1/2})^T(\mathbf{A}\mathbf{D}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{D}^{1/2}\right]\mathbf{D}^{-1/2}\mathbf{x}. \end{aligned} \tag{56}$$

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 (56) becomes

$$\boldsymbol{\delta}\mathbf{x}_1 = \frac{1}{\omega}\mathbf{D}^{1/2}\boldsymbol{\Pi}\mathbf{D}^{-1/2}\mathbf{x} = \frac{1}{\omega}\mathbf{D}^{1/2}\boldsymbol{\Pi}^2\mathbf{D}^{-1/2}\mathbf{x}, \tag{57}$$

since an orthogonal projection always satisfies $\boldsymbol{\Pi}^2 = \boldsymbol{\Pi}$. Therefore since $(\mathbf{x},\mathbf{y},\omega)$ solves (51), 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 (57) 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* (16)–(17) *is applied to the function $\mathfrak{H}$ defined as* (51)*, 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 (43) 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* (43)*. 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 experiments

In this section we will present various test cases to validate our methods and compare them with the log trick approach presented in (20). For the most challenging text case, we add 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 [19]. Our code has been developed with the Julia Programming Language and uses the automatic differentiation package ForwardDiff [22].

5.1 Numerical parameters

The function X and Y for the parametrization are those of the switch defined in (32). For the Cartesian representation technique, the function f is the discrepancy function defined in (43). 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 for which we are seeking the root, the convergence criterion for Newton's algorithm is

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

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

It is important to mention the different ways of initializing the Newton algorithm, depending on the method. Starting from an initial guess $\mathbf{n} = (n_1, \ldots, n_N)$, the variable ω is defined as $1/\langle \mathbf{1}, \mathbf{n}\rangle$ and the mole fractions $\mathbf{x}$ as $x_i = \omega n_i$ (see Section 2.4). The initialization for each method is described below.

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

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

- For the parametrization, the function $X(\tau)$, defined in (32), is inverted to initialize $\boldsymbol{\tau} = (\tau_i)_i$:

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

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

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

5.2 Test cases presentation

We will study the following 3 test cases

- The H_2O test case: 3 species, 2 elements and 1 reaction
- The *Seawater* test case: 37 species, 10 elements and 27 reactions.
- The *Water-Concrete* test cases: 88 species, 12 elements and 75 reactions.

All the chemical systems involved in these tests cases are detailed in Appendix C together with the solution of the chemical equilibria. The first two, H_2O and *Seawater*, use the charge constraint defined in section Appendix A.1 while *Water-Concrete* use the charge constraint and the pE constraint defined in section Appendix A.2. For all these test cases, the initial vector $\mathbf{n}$ is as follows:

$$(n_{\mathrm{H_2O}}, n_j) = (55,\ 1).$$

5.3 Numerical results

5.3.1 H2O test case

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

Evolution of residuals

Figure 3 shows the evolution of the residuals of our methods compared with the classical log-trick approach. The graph on the left shows the evolution of the norm of the function at iterate k, while that on the right shows the evolution of the norm between iterates k and $k-1$. There is a significant decrease in the norm of the function between the second and third iterations for the parametrization and Cartesian representation method, in contrast to the log-trick, which converges more slowly to the solution. The plateau observed on the left-hand graph indicates that the convergence of the iterates $\|\mathcal{X}^{(k)} - \mathcal{X}^{(k-1)}\|_\infty$ is slower than that of the residuals $\|\mathfrak{N}(\mathcal{X}^{(k)})\|_\infty$, justifying the use of both convergence criteria to guarantee the accuracy of the results.

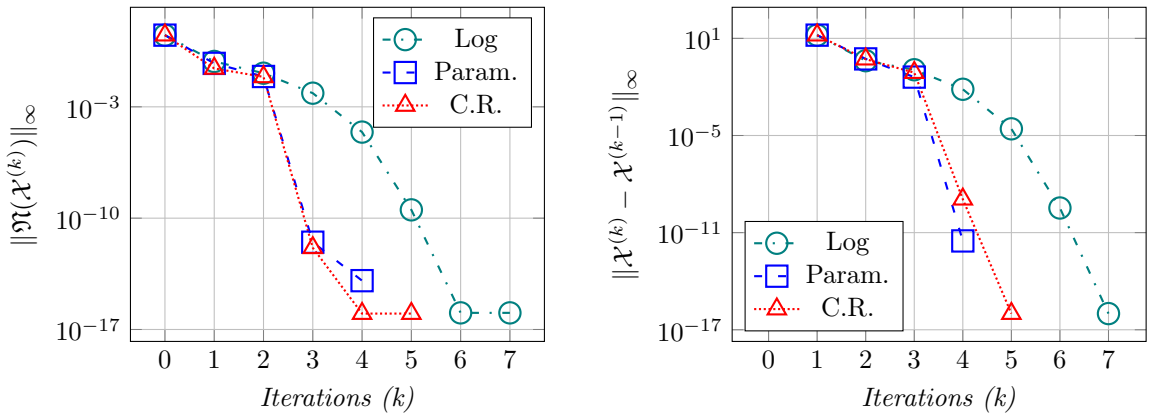

Figure 3: Evolution of residuals for the H_2O test case.

Comparison between the log-trick and the parametrization

To explain the difference in convergence speed between the parametrization and the Cartesian representation, Figure 4 shows the evolution of the norm of the function at the k-th iterate restricted to the equations of conservation of elements for the graph on the left, and to the equilibrium equations for the graph on the right. Figure 5 describes the evolution of iterates for the species H_2O and its mole fraction. Table 1 shows which parametrization function is used at each iteration. Finally, Figure 6 shows the evolution of the iterates of species H^+ and OH^- as well as ω.

Figure 4 shows that the significant decrease in residual is due to the chemical equilibrium equation being completely solved from iteration 3 onwards. Indeed, from iteration 2 onwards, the equilibrium

Iteration	0	1	2	3	4
$\text{sign}(\tau_{H_2O})$	$-$	$+$	$-$	$-$	$-$
$X(\tau_{H_2O})$	$\exp \tau_{H_2O}$	$\tau_{H_2O} + 1$	$\exp \tau_{H_2O}$	$\exp \tau_{H_2O}$	$\exp \tau_{H_2O}$
$Y(\tau_{H_2O})$	τ_{H_2O}	$\ln \tau_{H_2O}$	τ_{H_2O}	τ_{H_2O}	τ_{H_2O}

Table 1: Evolution of τ_{H_2O}.

equation is linear, and Newton's method is known to solve linear equations exactly. Figure 6 shows that the iterates τ_{H^+} and τ_{OH^-} are always negative, so that $Y(\tau_{H^+})$ and $Y(\tau_{OH^-})$ are always linear. The linearity or non-linearity of the equilibrium equation therefore depends only on the value of τ_{H_2O}, and figure 5 and table 1 clearly show that this value is negative from the second iteration onwards, making the equilibrium equation linear.

From figure 5, it is important to note that on the first iteration, the τ_{H_2O} of the parameterization and the $\ln(x_{H_2O})$ of the log trick have identical values, but that the corresponding mole fractions $X(\tau_{H_2O})$ and x_{H_2O} are not the same, thanks to the fact that $X(\tau)$ is linear for a positive value of τ. This prevents the mole fraction value from being too high, and leads to a value of τ_{H_2O} much closer to the solution at iteration 2, in contrast to log trick. The parametrization mechanism therefore accelerated convergence towards the solution in this test case.

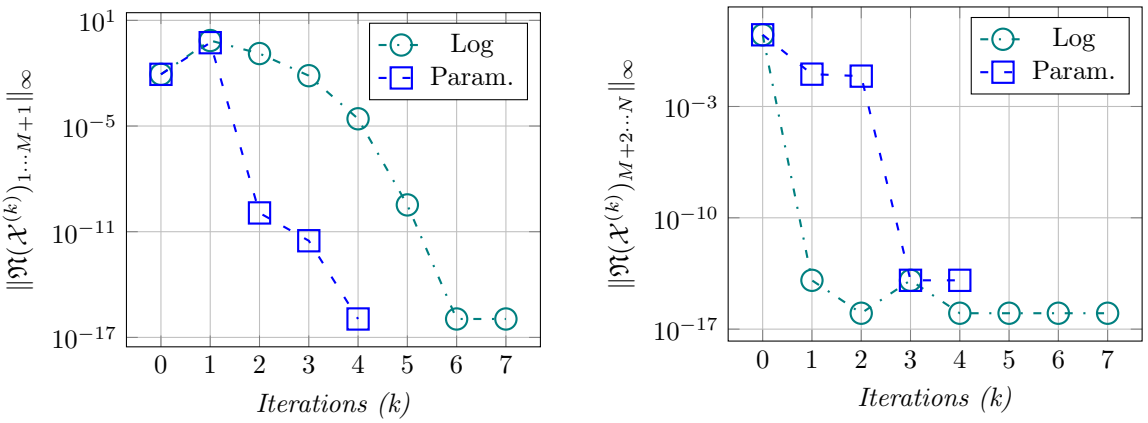

Figure 4: Evolution of residuals of the log trick and the parametrization for the H_2O test case. The left graph represents the residuals of the conservation equations while the right one represents the residuals of the equilibrium equations.

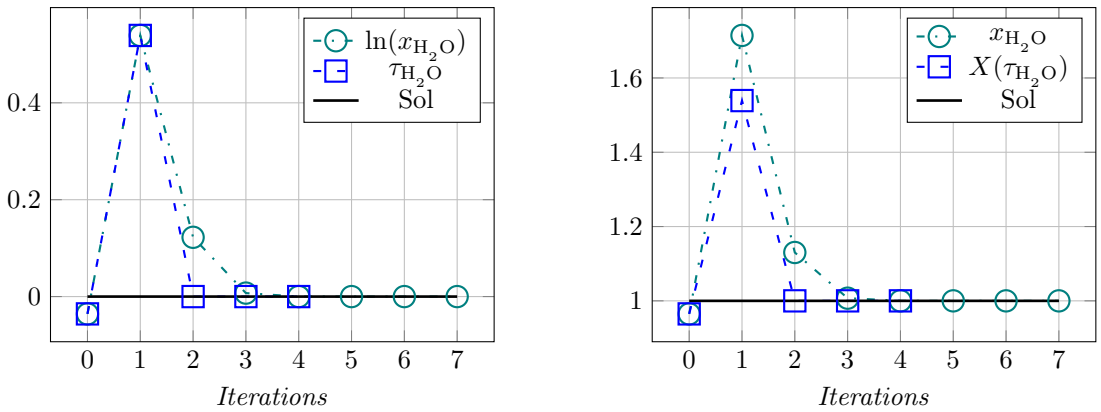

Figure 5: Evolution of the species H_2O with the log trick and parametrization methods.

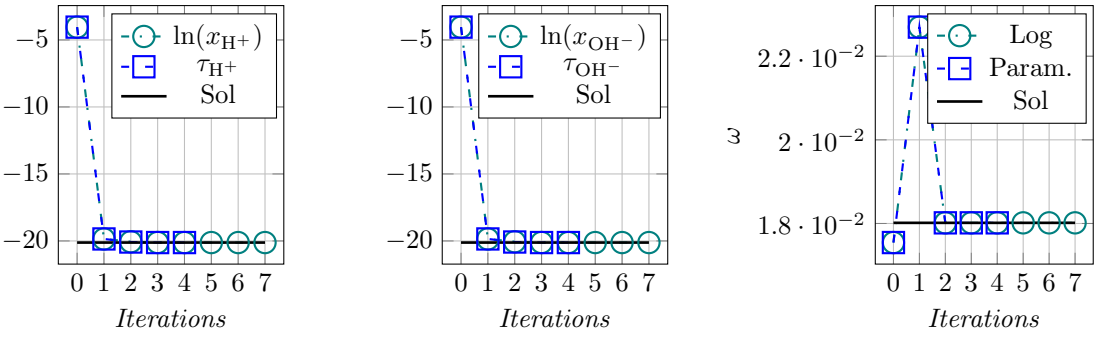

Figure 6: Evolution of the species H^+ and OH^- and the unknowns ω for the log trick and parametrization method.

Convergence of the Cartesian representation

The difference in convergence speed between the Cartesian representation and the log-trick observed at the third iteration corresponds to the moment when the x and y variables of the species H_2O are again in the lower left-hand zone in the Figure 7, left-hand graph. The graph on the right shows the decay of the residuals for equations linked to the f function. As expected, the linear equations are solved starting from the first iteration. We observe that the link $y = \ln x$ is strongly broken after the first iteration, expressing the fact that the x and y variables evolve separately.

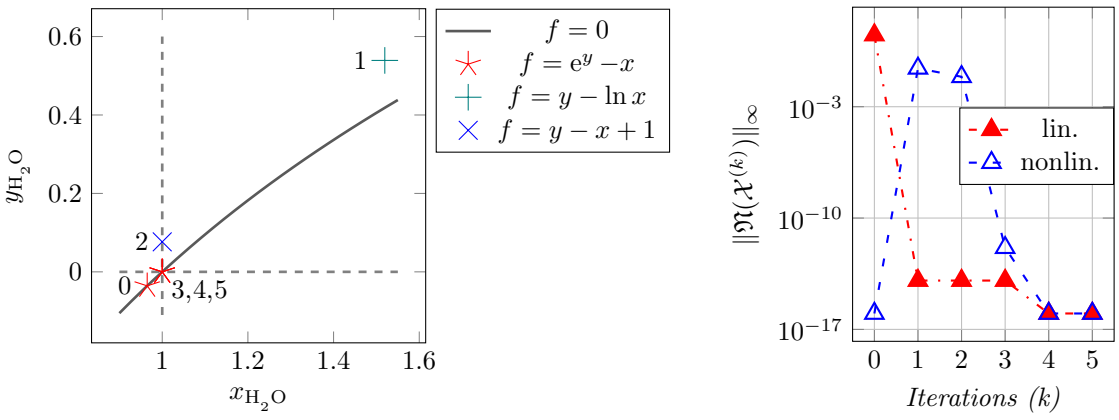

Figure 7: Evolution of iterates for the species H_2O (left) and evolution of the residuals for linear and nonlinear equations of the system (right).

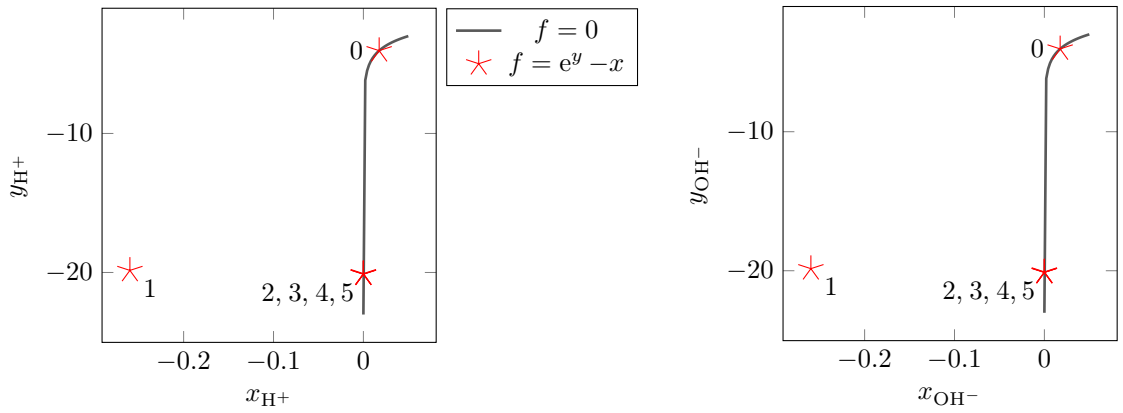

Figure 8: Evolution of iterates for the species H^+ and OH^-.

5.3.2 *Seawater* test case

This test case presented in Appendix C.2 contains many more species and remains fairly simple, allowing us to validate the convergence of our methods. For the initialization considered, the number of iterations is the same for all methods, but we observe an increase in the residual at the first iteration for the log-trick in Figure 9. In the case of parametrization, the Table 2 shows that there are only three species for which the τ parameter changes sign at the first iteration. So from the second iteration onwards, the chemical equilibrium equations are linear and therefore resolved from the third iteration onwards, as shown in Figure 10. We also observe in Figure 11 and Figure 12 there are only three species whose variables (x, y) change zone and from the third iteration onwards, there is no longer any change of function for the Cartesian representation.

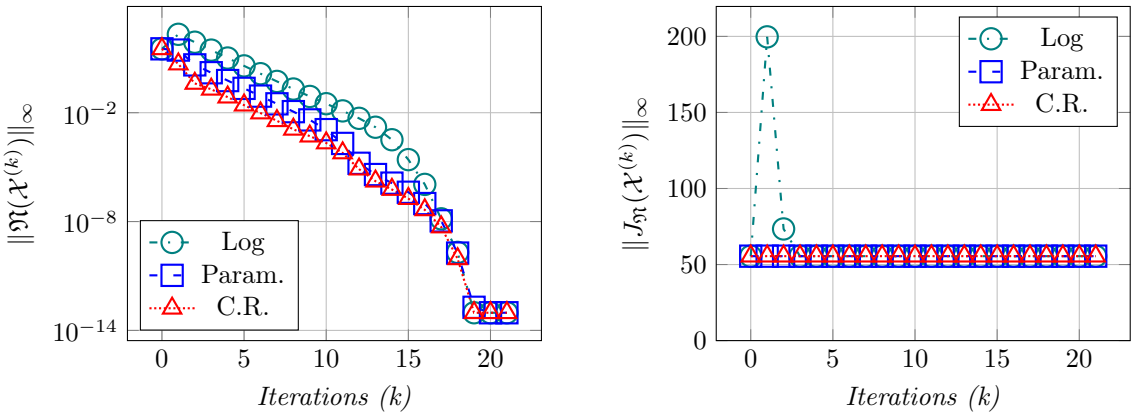

Figure 9: Evolution of residuals and norm of the Jacobian matrix.

Iteration	0	1	2	$\cdots$	21
$\mathrm{sign}(\tau_i)$	$-$	$+$	$-$	$\cdots$	$-$
$X(\tau_i)$	$\exp \tau_i$	$\tau_i + 1$	$\exp \tau_i$	$\cdots$	$\exp \tau_i$
$Y(\tau_i)$	τ_i	$\ln \tau_i$	τ_i	$\cdots$	τ_i

Table 2: Evolution of τ_i for $i \in \{H_2O, K^+, KSO_4^-\}$

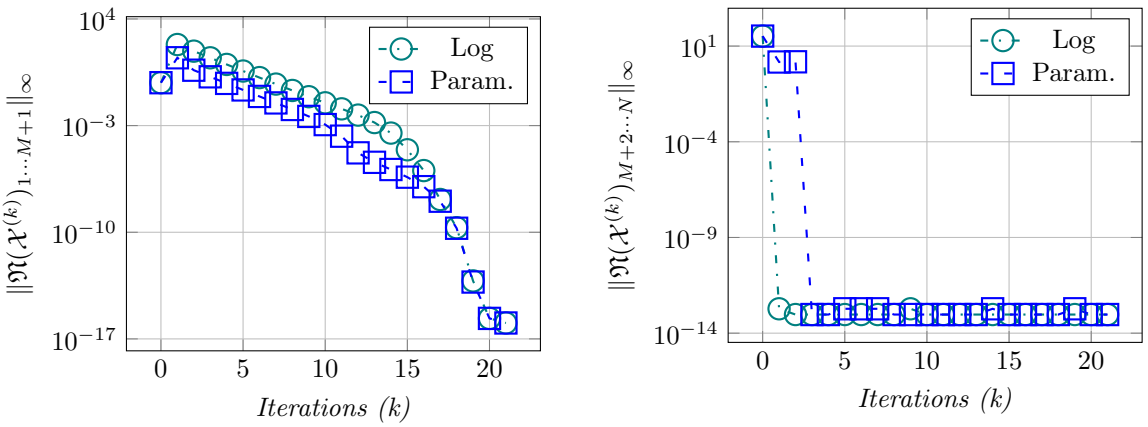

Figure 10: Evolution of residuals of the log trick and the parametrization for the *Seawater* test case. The left graph represents the residuals of the conservation equations while the right one represents the residuals of the equilibrium equations.

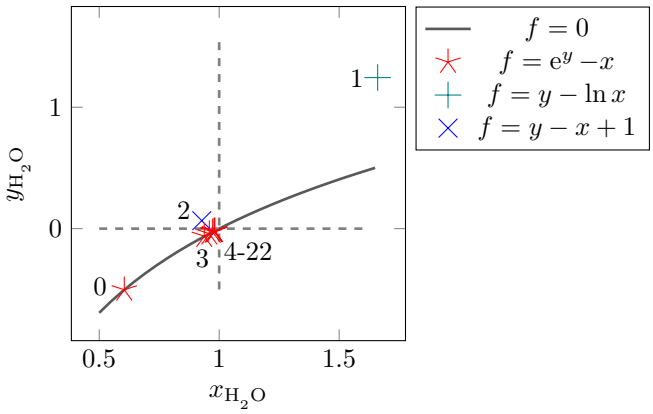

Figure 11: Evolution of iterates for the species H_2O for the Cartesian representation

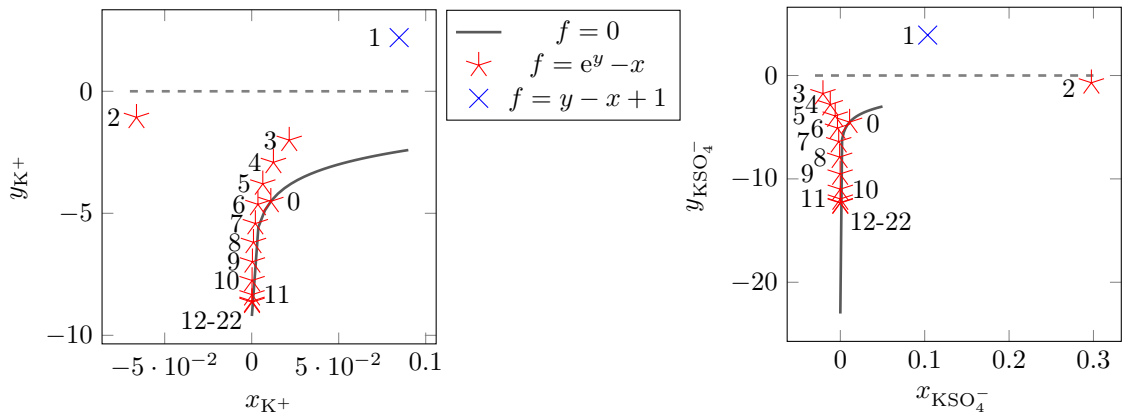

Figure 12: Evolution of iterates for the species K^+ and KSO_4^- for the Cartesian representation.

5.3.3 *Water-Concrete* test case

The *Water-Concrete* test case differs from the previous ones in that it contains a redox constraint. This constraint results in very low concentrations of certain chemical species. Figure 13 shows the evolution of the residuals and, unlike the other test cases, there is no convergence: the log-trick diverges very quickly, the parametrization diverges after 33 iterations, while a cycle is formed for the Cartesian representation. To achieve convergence for this test case and initialization, we added a line search to Newton's method. Figure 14 shows the results of the three methods with line search: our parametrization and Cartesian representation methods converge through a plateau at around 10^{-3}, while the log-trick diverges after the sixth iteration.

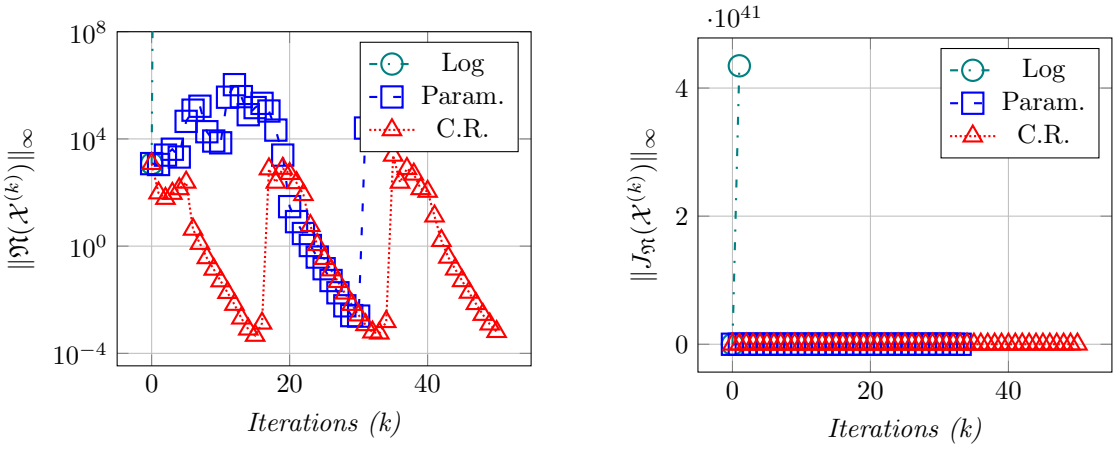

Figure 13: Evolution of residuals and norm of the Jacobian matrix.

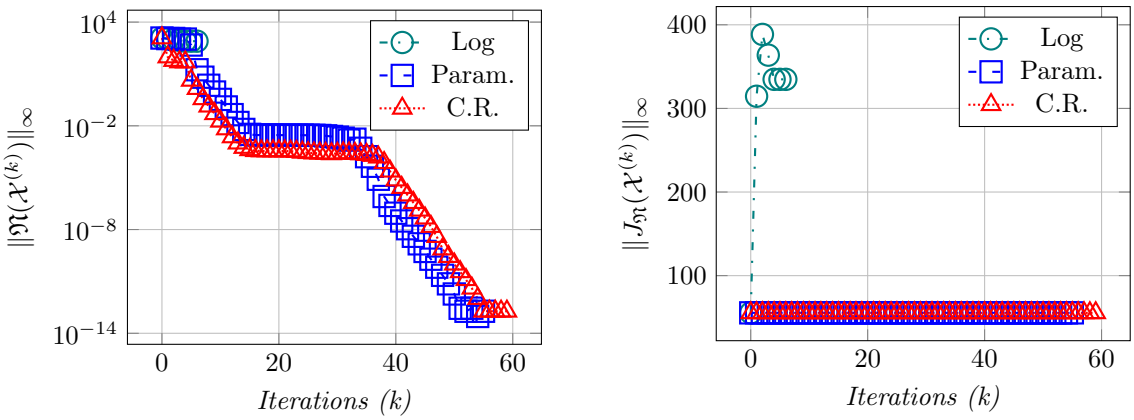

Figure 14: Evolution of residuals and norm of the Jacobian matrix for all methods with a line search.

6 Conclusion and future works

We 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 proved the local quadratic convergence of Newton's procedure. We conducted a comparative study of the convergence of our proposed methods against the traditional log trick approach, from the existing literature. This comparison was based on three test cases, each escalating in complexity. In the initial two scenarios, our methods exhibited good accuracy and effectively stabilized the initial iterations of the Newton method. However, the complexity of the final test case necessitated the incorporation of a linear search technique to globalize the convergence of the Newton method. This additional technique facilitated the convergence of our methods, which the traditional approach failed to do.

Future work will focus on expanding the parametrization and Cartesian representation techniques to the more complex case of multiphase chemical equilibrium calculations. The results obtained in this

article are promising as regards the possible improvement of the robustness of Newton's method in this field.

CRediT authorship contribution statement

Maxime Jonval: Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Ibtihel Ben Gharbia:** Conceptualization, Methodology, Validation, Writing - Review & Editing, Supervision. **Clément Cancès:** Conceptualization, Methodology, Formal analysis, Writing - Review & Editing, Supervision. **Thibault Faney**: Conceptualization, Methodology, Validation, Writing - Review & Editing, Supervision. **Quang-Huy Tran:** Conceptualization, Methodology, Writing - Review & Editing, Supervision.

Acknowledgement

This work was jointly supported by IFPEN and Inria. Maxime Jonval and Clément Cancès acknowledge support from the Labex CEMPI (ANR-11-LABX-0007-01).

Appendix A 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 [32], but the ones we will use are charge conservation and the redox constraint.

Appendix A.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

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

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

Appendix A.2 Redox constraint

In the case of oxidation-reduction reactions, the system (13) can be modified by introducing the notion of electron potential [26]. This potential, denoted pE, is written as

$$\mathrm{pE} = -\log_{10}(a_{e^-}), \tag{58}$$

where a_{e^-} is the electron activity. In (58), 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:

$$\mu_{e^-} = \mu^\circ_{e^-} + RT \ln a_{e^-}, \tag{59}$$

where $\mu^\circ_{e^-}$ is a standard chemical potential for the electron to be computed from a thermodynamic database. Thus, by integrating (58) into (59), it follows that

$$\mu_{e^-} = \mu^\circ_{e^-} - \mathrm{pE} \times RT \ln 10. \tag{60}$$

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:

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

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

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

and

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

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.

Appendix B Standard chemical potentials

The standard chemical potential μ_i°(P, T) of a species C_i for a constant pressure P and temperature T is calculated from the SUPCRT92 database [11].

Table 3: Standard chemical potentials at P = 1 Bar and T = 298.15 K.

Formula	μ_i°(P,T)	Formula	μ_i°(P,T)
$\mathrm{H_2O}$	-237138.97589284607	$\mathrm{H^+}$	9956.885403312557
$\mathrm{O_2(aq)}$	26500.37842186901	$\mathrm{Na^+}$	-251923.79553026147
$\mathrm{Mg^{2+}}$	-444027.90215494944	$\mathrm{K^+}$	-272504.8996797245
$\mathrm{Ca^{2+}}$	-542833.0422538111	$\mathrm{Fe^{2+}}$	-81547.14762262715
$\mathrm{HCO_3^-}$	-576982.8987273659	$\mathrm{Al^{3+}}$	-473751.00441000663
$\mathrm{SO_4^{2-}}$	-734502.084490013	$\mathrm{Cl^-}$	-121332.84714937139
$\mathrm{Sr^{2+}}$	-553878.8001745282	$\mathrm{AlO^+}$	-651901.5520727821
$\mathrm{AlOH^{2+}}$	-682390.3513772341	$\mathrm{HAlO_2(aq)}$	-859059.7240321051
$\mathrm{AlO_2^-}$	-821374.446375568	$\mathrm{CaOH^+}$	-706762.1524620559
$\mathrm{CO(aq)}$	-110048.61728485748	$\mathrm{CO_2(aq)}$	-376017.06579503417

CO_3^{2-}	-518026.1359207859	$CaHCO_3^+$	-1.135747579476859e6
$CaCl^+$	-672453.3605597073	$CaCl_2(aq)$	-801738.9637317051
$CaSO_4(aq)$	-1.2993419278916481e6	HClO(aq)	-69957.53260756626
ClO^-	-26862.311151193757	ClO_2^-	27111.248231900317
ClO_3^-	2007.235286837261	ClO_4^-	1421.4659696309664
Fe^{3+}	-7281.149189013573	$FeCl^+$	-211920.58507072268
$FeCl_2(aq)$	-297483.39734809624	$FeOH^{2+}$	-231878.2281150759
$FeOH^+$	-265559.42898331815	FeO^+	-212213.40826659818
FeO	-202255.52192397387	$HFeO_2^-$	-389196.60766906774
$HFeO_2(aq)$	-413045.42294696183	FeO_2^-	-358235.0321847625
$H_2(aq)$	27680.272394604483	$H_2S(aq)$	-17962.985906651626
HO_2^-	-57363.666715716085	HS^-	21923.09086173558
HSO_3^-	-517770.9559572122	HSO_4^-	-745798.9041292057
HSO_5^-	-627559.1114566573	KCl(aq)	-389322.1898012777
$KHSO_4(aq)$	-1.0084285424401537e6	KOH(aq)	-427271.0297495857
KSO_4^-	-1.0219846680885215e6	$CH_4(aq)$	-24494.210815931885
$Mg(CO_3)(aq)$	-989014.6934790071	$Mg(HCO_3)^+$	-1.0368796788141541e6
$MgCl^+$	-574547.7790990678	$MgOH^+$	-614525.8903919919
NaCl(aq)	-378778.4933733225	NaOH(aq)	-408024.6192443839
OH^-	-147340.5566329419	S_2^{2-}	89452.83031941311
$S_2O_3^{2-}$	-512624.6363618013	$HS_2O_3^-$	-522247.82998507
$H_2S_2O_3(aq)$	-525595.041488881	$S_2O_4^{2-}$	-590447.0178825587
$HS_2O_4^-$	-604672.624055689	$H_2S_2O_4(aq)$	-606764.6340067171
$S_2O_5^{2-}$	-780818.9790117	$S_2O_6^{2-}$	-956546.945397623
$S_2O_8^{2-}$	-1.10507895975272e6	S_3^{2-}	83595.22019473514
$S_3O_6^{2-}$	-948178.949248605	S_4^{2-}	78992.80990918931
$S_4O_6^{2-}$	-1.0306037475969802e6	S_5^{2-}	75645.59931794215
$S_5O_6^{2-}$	-948178.9537022973	$SO_2(aq)$	-291207.4128400276
SO_3^{2-}	-476642.20414498576	$Sr(HCO_3)^+$	-1.1478393365598267e6
$SrCl^+$	-683750.1584404652	$SrOH^+$	-715130.1518277868
$H_2O_2(aq)$	-124056.64548498068	$HClO_2(aq)$	15814.426726057798
$NaSO_4^-$	-1.0003785022780169e6	$MgSO_4(aq)$	-1.2012145947920694e6
HCl(aq)	-117278.5194378308	$CaCO_3(aq)$	-1.0898072308832686e6
$SrCO_3(aq)$	-1.0982170694298274e6	$FeCl^{2+}$	-147018.36458365855
e^-	-16.315331966024218		

Appendix C Test case chemical systems

Appendix C.1 The *dissociation of water* test case

The H_2O test case is composed of

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

The vector of constraints $\mathbf{b}$ for this test case is composed of $n_O = 55.5087$ and a charge of 0. The solution obtained is $\mathbf{n} = (55.5086998985565, 1.0144349897420512 \times 10^{-7}, 1.0144349897420512 \times 10^{-7})$.

Appendix C.2 The *Seawater* test case

The *Seawater* test case is composed of

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

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

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

The vector **b** for this test case is given in Table 4.

Feeds (mol)									
O	Na	Mg	S	Ca	K	C	Sr	Cl	Z (charge)
55.5087	0.469	0.0528	0.0282	0.0103	0.0102	0.00206	1×10^{-5}	0.546	0

Table 4: Vector **b** of the *Seawater* test case for the elements conservation.

The solution obtained for the *Seawater* test case is

$\mathbf{n} = (55.38968051132381, 0.42692804180975974, 0.02705950281438023, 0.0033461882935940023$
$0.005755697020771243, 0.009937871289383643, 0.0008404617918854967, 7.642853227990452 \times 10^{-6},$
$0.4991461549132867, 1.6714211435836995 \times 10^{-9}, 5.132624707658365 \times 10^{-7}, 3.0602737036270006 \times 10^{-6},$
$2.3982990446766207 \times 10^{-5}, 5.298476689811943 \times 10^{-5}, 0.00144101412325631, 0.00031520952296616457,$
$0.0024462442193710124, 5.242661308469255 \times 10^{-10}, 1.4219455529061577 \times 10^{-5}, 1.7188295621737205 \times 10^{-15},$
$2.1991731951952572 \times 10^{-8}, 0.00024788726335362165, 0.0006080613936706859, 0.00024287219824938508,$
$0.009738343343159612, 3.4148566766825304 \times 10^{-5}, 0.0350277315444541, 1.6186301459231793 \times 10^{-6},$
$6.246267369519702 \times 10^{-6}, 1.0728986616184622 \times 10^{-7}, 2.1174144853264242 \times 10^{-6}, 2.3140715139129212 \times 10^{-10},$
$0.007042608015640257, 0.015117071683773249, 1.598968187418331 \times 10^{-10}, 0.0002883370842663884,$
$1.3221101336988816 \times 10^{-7})$

Appendix C.3 The *Water-Concrete* test cases

The *Water-Concrete* test case is composed of

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

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+}};\\
\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^-}.
\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+}};\\
\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+}};
\end{aligned}$$

The vector **b** for this test case is given in Table 5.

Feeds (mol)						
O	Na	Mg	K	Ca	Fe	C
55.5078	0.0601	1.5079×10^{-9}	0.1402	0.00196384	4.58364×10^{-7}	5.29145×10^{-5}
Al	S	Cl	Sr	Z (charge)	pE	
3.80016×10^{-5}	0.000974141	1.42825×10^{-10}	1×10^{-10}	0	-2.98873	

Table 5: Vector **b** of the *Water-Concrete* test case for the elements conservation.

The solution obtained for the *Water-Concrete* test case is

$$\mathbf{n} = (55.30153263049913, 1.2515957847430055 \times 10^{-45}, 0.05383341189709943, 3.85205612592884 \times 10^{-11}$$

$0.13104185508007585, 0.0005173318114609391, 1.7211732119737842 \times 10^{-22}, 2.968381612134554 \times 10^{-8}$
$5.512454738276492 \times 10^{-36}, 0.0004189076148432289, 1.414654075260127 \times 10^{-10}, 5.177690711306953 \times 10^{-11},$
$5.534605441939609 \times 10^{-14}, 8.331028635348468 \times 10^{-20}, 1.0086583996023467 \times 10^{-27}, 8.387036717264553 \times 10^{-12},$
$3.8001591612963186 \times 10^{-5}, 0.0013909717301057072, 9.16012434284405 \times 10^{-41}, 3.584713859640573 \times 10^{-15},$
$2.5242434859636466 \times 10^{-5}, 1.7045004485556293 \times 10^{-10}, 3.7199328160101444 \times 10^{-14}, 2.3370165495489375 \times 10^{-24},$
$2.7894078954109622 \times 10^{-5}, 7.744041296711739 \times 10^{-54}, 3.9571740772053503 \times 10^{-48}, 1.3767116074490875 \times 10^{-78},$
$3.4198978487645584 \times 10^{-95}, 4.303257445391223 \times 10^{-116}, 1.734550767970587 \times 10^{-38}, 1.6769901176709222 \times 10^{-32},$
$2.305681346281873 \times 10^{-50}, 1.9826398981963012 \times 10^{-27}, 1.526767859061291 \times 10^{-18}, 1.2896347040880048 \times 10^{-17},$
$2.2480301791284984 \times 10^{-16}, 6.4990813411481995 \times 10^{-12}, 1.0117847734211808 \times 10^{-10}, 4.582563222020898 \times 10^{-7},$
$2.2465395189279816 \times 10^{-24}, 3.8486081967306244 \times 10^{-72}, 3.185210777712516 \times 10^{-38}, 7.176882182296528 \times 10^{-66},$
$2.495689820120032 \times 10^{-34}, 2.2023864652584633 \times 10^{-15}, 4.2214199103551675 \times 10^{-57}, 5.385118088684071 \times 10^{-14},$
$9.648599524565427 \times 10^{-20}, 0.008743466121629882, 0.0004146787982403328, 7.700254829779092 \times 10^{-76},$
$9.23254021400711 \times 10^{-13}, 1.237443643602187 \times 10^{-17}, 3.981561928493589 \times 10^{-21}, 1.4657260543330728 \times 10^{-9},$
$1.2685419568327648 \times 10^{-12}, 0.006153927596402049, 0.18584677511174025, 3.3198439612835856 \times 10^{-120},$
$9.804207509062558 \times 10^{-78}, 2.6241149020481446 \times 10^{-89}, 5.585408849329376 \times 10^{-102}, 4.191684712607641 \times 10^{-85},$
$7.1822771665537554 \times 10^{-96}, 9.213692990260926 \times 10^{-109}, 9.355900949612459 \times 10^{-73}, 5.679111032409044 \times 10^{-63},$
$5.890335493096298 \times 10^{-79}, 6.092605128626266 \times 10^{-175}, 3.3555331102837764 \times 10^{-120}, 6.738886584218897 \times 10^{-230},$
$1.5972079714124394 \times 10^{-161}, 4.4923443994746124 \times 10^{-285}, 1.0016752314700165 \times 10^{-231}, 9.877024332443298 \times 10^{-46},$
$2.81914046857951 \times 10^{-28}, 2.6014418358334538 \times 10^{-17}, 4.119856257314945 \times 10^{-21}, 4.7267746136716166 \times 10^{-11},$
$8.488781208032296 \times 10^{-40}, 7.237999946744915 \times 10^{-90}, 0.000112660505230008, 2.7301180078217594 \times 10^{-12},$
$1.5206742654457507 \times 10^{-24}, 2.764220899199947 \times 10^{-5}, 9.553207316759742 \times 10^{-13}, 7.384932228127377 \times 10^{-47})$

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] J. W. Johnson, E. H. Oelkers, and H. C. Helgeson, *SUPCRT92: A software package for calculating the standard molal thermodynamic properties of minerals, gases, aqueous species, and reactions from 1 to 5000 bar and 0 to 1000°C*, Computers & Geosciences, 18 (1992), pp. 899–947, https://doi.org/10.1016/0098-3004(92)90029-Q.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] D. L. Parkhurst and C. A. J. Appelo, *Description of input and examples for PHREEQC version 3–A computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations*, in Modeling techniques, no. 6 in chap. A43, U.S. Geological Survey Techniques and Methods, 2013, p. 497, https://pubs.usgs.gov/tm/06/a43/.

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

[20] M. H. Reed, N. F. Spycher, and J. Palandri, *Users Guide for CHIM-XPT: A Program for Computing Reaction Processes in Aqueous-Mineral-Gas Systems and MINTAB Guide*, Department of Geological Sciences, University of Oregon, (2016), https://pages.uoregon.edu/palandri/data/chim-xpt%20guide%20V.2.50.pdf.

[21] M. H. Reed, N. F. Spycher, and J. Palandri, *SOLVEQ-XPT : A Computer Program for Computing Aqueous-Mineral-Gas Equilibria*, Department of Geological Sciences, University of Oregon, (2018), https://pages.uoregon.edu/palandri/data/solveq-xpt%20guide_v.2.25.pdf.

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

[23] 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.

[24] 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.

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

[26] 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.

[27] A. H. Truesdell and B. F. Jones, *WATEQ, a computer program for calculating chemical equilibria of natural waters*, Journal of Research of the U.S. Geological Survey, 2 (1974), pp. 233–248, https://pubs.usgs.gov/journal/1974/vol2issue2/report.pdf#page=105.

[28] 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.

[29] 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.

[30] J. van der Lee and L. De Windt, *CHESS Tutorial and Cookbook*, Tech. Report LHM/RD/02/13, Ecole Nationale Supérieure des Mines de Paris, Centre d'Informatique Géologique, 2002, https://radiochemistry.faculty.unlv.edu/readings/chess-tutorial3-0.pdf.

[31] J. C. Westall, J. L. Zachary, and F. M. M. Morel, *MINEQL: A computer program for the calculation of chemical equilibrium composition of aqueous systems*, tech. report, Cambridge, Mass. : Water Quality Laboratory, Ralph M. Parsons Laboratory for Water Resources and Environmental Engineering sic, Dept. of Civil Engineering, Massachusetts Institute of Technology, 1976, https://dspace.mit.edu/handle/1721.1/142980.

[32] 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, Lawrence Livermore National Laboratory, Sept. 1992, https://doi.org/10.2172/138643.