SOLUTION OF REACTIVE TRANSPORT PROBLEMS INCLUDING MINERAL PRECIPITATION-DISSOLUTION REACTIONS BY A SEMISMOOTH NEWTON METHOD[1]

Hannes Buchholzer[1], Christian Kanzow[1], Peter Knabner[2], and Serge Kräutle[2]

Preprint 288 April 2009

[1] University of Würzburg
Institute of Mathematics
Am Hubland, 97074 Würzburg, Germany
e-mail: buchholzer@mathematik.uni-wuerzburg.de
kanzow@mathematik.uni-wuerzburg.de

[2] University of Erlangen-Nürnberg
Department of Mathematics
Martensstraße 3, 91058 Erlangen, Germany
e-mail: kraeutle@am.uni-erlangen.de
knabner@am.uni-erlangen.de

April 14, 2009

[1]This research was partially supported by the DFG (Deutsche Forschungsgemeinschaft) under the grants Ka 1296/16-1 and KN 229/12-1.

Abstract. The modeling of reactive transport in the subsurface including mineral precipitation-dissolution reactions involves a coupling of PDEs, ODEs, and algebraic equations to inequalities. In the geoscientists' community, the most frequently used algorithms to solve these kinds of systems apply some kind of trial-and-error strategies. The aim of this article is to apply a modern and efficient solution strategy, the semismooth Newton method, to this geoscientific problem, and to investigate its applicability and efficiency both from a theoretical and a numerical point of view. In particular, it turns out that the method is typically quadratically convergent.

Key Words: reactive transport, mineral precipitation-dissolution, complementarity problems, semismooth Newton method, quadratic convergence

1. Introduction

The modeling of reactive transport problems in porous media leads to systems containing partial differential equations (PDEs) and ordinary differential equations (ODEs); the PDEs for the concentration of species which are dissolved in the water (mobile), and the ODEs for the concentration of species which are attached to the soil matrix (immobile). The immobile species can be sorbed species or minerals. If the reactions are sufficiently fast, then the assumption of local equilibrium is reasonable. This equilibrium is usually described by a set of (nonlinear) algebraic equations (AEs) coupling the PDEs and the ODEs. However, the equilibrium description only by AEs is no longer valid when reactions with minerals are involved. In this situation, the equilibrium description of the mineral precipitation-dissolution reactions has to take into account two possibilities for equilibrium: the case of a saturated fluid, and the case of a complete dissolution of the mineral (see Sec. 2). Such an equilibrium condition can be expressed by using a combination of equations and inequalities, having the shape of a (nonlinear) complementarity problem. The resulting system consists of PDEs, ODEs, AEs, and complementarity conditions (CCs).

For the numerical solution, many publications on reactive transport in porous media suggest to enforce a decoupling between transport and reaction by applying an operator splitting technique. By this, the reaction subproblem is fully local, i.e., it consists only of AEs and CCs, while only the transport subproblem contains the PDEs and ODEs. However, operator splitting either introduces splitting errors or requires a fixed-point type iteration between transport and reaction within each time step. In the first case, accuracy considerations, and in the second case, convergence issues often lead to severe time step restrictions for splitting methods.

A very popular way to handle the PDE-ODE-AE-CC system in computational geosciences is the following [2, 3]: For the current time step, for each mineral and each discretization point, an assumption is made (usually based on the previous time step) whether saturation or complete dissolution will hold. Under this assumption, a Newton iteration is performed. If the result has no physical meaning (negative mineral concentration, or supersaturated fluid), then the assumptions are modified in some way and the Newton iteration is repeated, until (hopefully) a physically meaningful solution is obtained. Besides its heuristic motivation, another drawback of this procedure is that the CPU time required is significantly higher than for reactive transport problems without minerals, since several Newton iterations are required per time step. The lack of efficiency becomes even more troublesome if fully implicit methods (avoiding the splitting of transport and reactions) are considered, since systems containing PDEs have to be solved again and again.

Other authors from the geosciences community propose to use a formulation as a free boundary problem for front tracking approaches [13]. However, this approach lacks simplicity as soon as more than one space dimension is involved and topology changes of the precipitation-dissolution fronts appear. Another approach is to approximate the equilibrium, i.e., very fast reactions, by a kinetic description with large rate coefficients. Besides the approximative nature of this approach, large rate coefficients may increase the stiffness of the problem to solve.

Modern techniques from the optimization theory for the reactive transport problem are considered in [17, 18] and in [10]. In [17, 18], an operator splitting is performed, and the now fully

local reaction problem is replaced by an equivalent constrained minimization problem for the so-called Gibbs free energy. Its KKT conditions are solved with an interior-point algorithm. Numerical test runs are performed without any deeper theoretical investigation. Note that this procedure leads to additional unknowns, the Lagrange multipliers for the equality and inequality constraints.

In [10, Sec. 4], to our knowledge for the first time, the application of a semismooth Newton method to the reactive transport mineral precipitation-dissolution problem is carried out. There the reactive transport problem is tackled fully implicit, avoiding any operator splitting. The author considers a rather general situation of reactive problems including equilibrium and kinetic reactions, where the equilibrium reactions may be of the aqueous, the sorption, or the mineral precipitation-dissolution type. The implementation of the solution strategy is described and some results on the nonsingularity of the Jacobian of the system are given.

The following article propagates and investigates similar solution strategies as in [10], but it focusses on those reactive systems without kinetic reactions, and where all the (equilibrium) reactions are of aquatic and of mineral type, i.e., no sorption is involved. This restriction allows to prove stronger theoretical results. The structure of the article is the following: In Sec. 2 the problem is formulated and its mathematical model is given. Sec. 3 contains an equivalence transformation (going back to [11, 12, 10]) being applied to the PDE-ODE-AE-CC system. The motivation for this reformulation is a decoupling of some (linear) PDEs, leading to a smaller nonlinear system. The resulting discretized system is a mixed complementarity problem that can be reformulated as a nonlinear (but nonsmooth) system of equations. The theoretical properties of this nonsmooth system of equations will be investigated in the subsequent sections, cf. Sec. 4–7. In particular, it is shown that a nonsmooth (semismooth) Newton-type method applied to this system is (usually) locally quadratically convergent since the resulting (generalized) Jacobian has no inherent singularity properties. Sec. 8 gives some numerical results for the semismooth Newton method applied to a special instance of our problem, and we close with some final remarks in Sec. 9.

2. Problem Formulation

This section gives a precise formulation of the mathematical model for the application that was outlined in the introduction. This formulation will be the basis for our subsequent theoretical and numerical investigations.

To this end, let us consider the concentrations of I mobile species $c = (c_1, c_2, \ldots, c_I)^T$. These species are dissolved in the groundwater. Their concentrations are time- and space-dependent. They are convected by a given Darcy flow field q and are subject to dispersion. The convection-diffusion operator for these species is given by

$$L_i c_i = -\nabla \cdot (D \nabla c_i - q c_i), \quad i = 1, \ldots, I,$$

with dispersion tensor $D = D(q)$ which depends on the flow field q. Clearly, this operator $L = (L_1, \ldots, L_I)^T$ is linear and acts in the same way on all mobile species, i.e. $L_1 = \cdots = L_I$.

The constant $\theta \in (0, 1)$ denotes the fraction of the mobile fluid-phase volume. With $\bar{c} = (\bar{c}_{I+1}, \ldots, \bar{c}_{I+\bar{I}})^T$ we denote the concentrations of the $\bar{I}$ mineral species. These concentrations

are also variable in time and space. They are attached to the soil matrix and therefore neither subject to convection nor diffusion. But they are like the mobile species involved in chemical reactions with other mobile or mineral species. In this paper we restrict ourselves to equilibrium reactions, i.e. reactions that are actually in the condition of equilibrium or equations which are sufficiently fast to be approximately considered to be in equilibrium. $R = (R_1, \ldots, R_J)$ denotes the vector of reaction rates that are necessary to keep the chemical system in equilibrium. Together with c and $\bar{c}$ they form the unknowns of the system to be considered here.

The $I + \bar{I}$ mass balance equations are

$$\begin{aligned} \frac{\partial}{\partial t}\theta c + Lc &= S_1 R\,, \\ \frac{\partial}{\partial t}\bar{c} &= S_2 R\,, \end{aligned} \tag{1}$$

given on the domain $[0,T] \times \Omega \subset \mathbb{R}^2$ or $\mathbb{R}^3$ together with given initial and boundary conditions. The matrix $(s_{ij}) = S = \begin{pmatrix} S_1 \\ S_2 \end{pmatrix} \in \mathbb{R}^{(I+\bar{I})\times J}$ is the matrix of stoichiometric coefficients, where J is the number of chemical reactions. If we have, for example, an equilibrium reaction

$$X_1 + 2X_2 \longleftrightarrow X_3$$

we shift all species to the right side

$$0 \longleftrightarrow -X_1 - 2X_2 + X_3$$

and get a column of matrix S with entries $-1, -2, 1$ in the corresponding positions. It is well known that any linear dependence of the chemical reactions (i.e. the columns of S) indicates a redundancy of chemical reactions [1]. Hence, without loss of generality, we can assume that S has full column rank,

$$\operatorname{rank}(S) = J\,. \tag{2}$$

Additionally, we demand that the columns of S_1 are linearly independent

$$\operatorname{rank}(S_1) = J\,. \tag{3}$$

Furthermore, we assume that each mineral is participating in one and only one mineral reaction, and that in each mineral reaction, exactly one mineral is involved. Hence, in this paper, a mineral reaction is a reaction with one mineral and one or more mobile species involved. With mobile reactions we indicate reactions in which only mobile species participate. By J_{mob} we denote the number of mobile reactions and with J_{min} the number of mineral reactions. It follows that $J_{min} = \bar{I}$. Since in our model we have only mineral or mobile reactions, it holds $J = J_{mob} + J_{min}$. The stoichiometric matrix then reads

$$S = \left(\frac{S_1}{S_2}\right) = \left(\begin{array}{c|c} S^1_{mob} & S^1_{min} \\ \hline 0 & -I \end{array}\right), \quad \text{with} \quad S^1_{mob} \in \mathbb{R}^{I\times J_{mob}},\ S^1_{min} \in \mathbb{R}^{I\times J_{min}},$$

where, for simplicity of notation, we have replaced the diagonal matrix representing the mineral participation in the mineral reactions by $-I$, the negative identity matrix. Therefore, reactions

$1, \ldots, J_{mob}$ are mobile and reactions $J_{mob}+1, \ldots, J$ are mineral, because the columns of S refer to chemical reactions.

The equilibrium conditions for the reactions not involving minerals are modeled by algebraic equations

$$\prod_{i=1}^{I} c_i^{s_{i,j}} - k_j = 0 \qquad (j = 1, \ldots, J_{mob}) \,, \tag{4}$$

where k_j are given constants. They hold in each point of space and time. Since we expect the solutions to be positive, equation (4) can equivalently be written as

$$Q_j(c) := \sum_{i=1}^{I} s_{i,j} \cdot \ln c_i - \ln k_j = 0 \qquad (j = 1, \ldots, J_{mob}) \,.$$

In matrix notation, the vector $Q_{mob} = (Q_1, Q_2, \ldots, Q_{J_{mob}})$ then becomes

$$Q_{mob}(c) = \left(S^1_{mob}\right)^T \ln c - K_1 \,,$$

where $K_1 = (\ln k_1, \ldots, \ln k_{J_{mob}})^T$ is the vector of equilibrium constants in logarithmic form and $\ln c$ is a vector where the logarithm is applied separately to every component of the vector c.

For the mineral equilibrium reactions, we have the complementary conditions

$$E_j(c) \cdot \bar{c}_j \;=\; 0 \;\wedge\; \bar{c}_j \geq 0 \;\wedge\; E_j(c) \geq 0 \qquad (j = J_{mob}+1, \ldots, J) \,,$$

where $E_j(c) := \ln k_j - \sum_{i=1}^{I} s_{i,j} \cdot \ln c_i$. The case $E_j(c) = 0, \bar{c}_j \geq 0$ corresponds to a saturation of the fluid with respect to this mineral reaction, and the case $E_j(c) \geq 0, \bar{c}_j = 0$ corresponds to the total dissolution of the mineral and an undersaturation of the fluid. Again we can write $E = (E_{J_{mob}+1}, \ldots, E_J)^T$ in matrix notation as

$$E(c) = K_2 - \left(S^1_{min}\right)^T \ln c$$

with $K_2 = (\ln k_{J_{mob}+1}, \ldots, \ln k_J)^T$. The constant $1/K_2$ (componentwise) is the so-called *solubility product*.

We decompose the reaction vector R into

$$R = \begin{pmatrix} R_{mob} \\ R_{min} \end{pmatrix}$$

with R_{mob} and $R_{\min}$ being of size J_{mob} and J_{min}, respectively. Utilizing the structure of S, the full system reads

$$\frac{\partial}{\partial t}\theta c + Lc = S^1_{mob}R_{mob} + S^1_{min}R_{min} = S_1 R\,, \tag{5}$$

$$\frac{\partial}{\partial t}\bar{c} = -R_{min}\,, \tag{6}$$

$$E_j(c)\cdot\bar{c}_j = 0\ (j = J_{mob}+1,\dots,J)\,, \tag{7}$$

$$\bar{c}_j \geq 0\ (j = J_{mob}+1,\dots,J)\,, \tag{8}$$

$$E_j(c) \geq 0\ (j = J_{mob}+1,\dots,J)\,, \tag{9}$$

$$Q_{mob}(c) = 0\,, \tag{10}$$

for the $I + \bar{I} + J$ unknowns $c, \bar{c}$ and R. Note that this is a differential-algebraic system of ordinary and partial differential equations coupled with complementary conditions arising from the mineral equilibrium reactions.

3. Decoupling and reformulation of the complementary conditions

The aim of this section is to reduce the size of the overall system (5)–(10) by using suitable decouplings and reformulations. Since these techniques are already known from [11, 12] (but strictly needed for our subsequent analysis), we will keep this section as short as possible.

First, we apply the decoupling technique proposed in [11, 12] to the PDE-ODE system (5)–(6). This will lead to a decoupling of some linear PDEs. The remaining PDE-system will then be significantly smaller than the original PDE-system. To this end, we define $S_1^\perp$ as a matrix consisting of a maximum set of linearly independent columns that are orthogonal to each column of S_1, i.e. $(S_1)^T S_1^\perp = 0$. Recall that the columns of S_1 were assumed to be linearly independent, cf. (3). Hence the pseudo-inverses of S_1 and $S_1^\perp$ are given by $\left(S_1^T S_1\right)^{-1} S_1^T$ and $\left(\left(S_1^\perp\right)^T S_1^\perp\right)^{-1}\left(S_1^\perp\right)^T$, respectively. Multiplying (5) with these two pseudo-inverses, we obtain

$$\left(\left(S_1^\perp\right)^T S_1^\perp\right)^{-1}\left(S_1^\perp\right)^T\left(\frac{\partial}{\partial t}\theta c + Lc\right) = 0\,, \tag{11}$$

$$\left(S_1^T S_1\right)^{-1} S_1^T\left(\frac{\partial}{\partial t}\theta c + Lc\right) = R\,, \tag{12}$$

$$\frac{\partial}{\partial t}\bar{c} = -R_{min}\,. \tag{13}$$

We now substitute

$$\eta := \left(\left(S_1^\perp\right)^T S_1^\perp\right)^{-1}\left(S_1^\perp\right)^T c,\quad \xi := \left(S_1^T S_1\right)^{-1} S_1^T c, \tag{14}$$

and partition the vector ξ into

$$\xi = (\xi_{mob}, \xi_{min})$$

of size J_{mob}, J_{min}. Then splitting equation (12) into two parts and adding the third block to the second part, we get

$$\begin{aligned} \frac{\partial}{\partial t}\theta\eta + L\eta &= 0\,, \\ \frac{\partial}{\partial t}\theta\xi_{mob} + L\xi_{mob} &= R_{mob}\,, \\ \frac{\partial}{\partial t}\left(\theta\xi_{min} + \bar{c}\right) + L\xi_{min} &= 0\,, \\ \frac{\partial}{\partial t}\bar{c} &= -R_{min}\,. \end{aligned}$$

We may consider the second and fourth equations as a definition for R_{min} resp. R_{mob}. Since we are not directly interested in R, we drop both equations (but can use them to compute R a posteriori).

It is well known that complementary conditions can be expressed equivalently via NCP-functions, also called C-functions, cf. [6, 7]. Let $\varphi(a,b) = \min\{a,b\}$ be the minimum function. This function is an NCP-function, i.e. it has the defining property that

$$\varphi(a,b) = 0 \iff a \geq 0\,,\ b \geq 0,\ a \cdot b = 0\,.$$

Using this minimum function, we can write the complementary conditions (7)–(9) as

$$\varphi\left(E_j(c)\,, \bar{c}_j\right) = 0 \qquad (j = 1, \ldots, J_{min})\ .$$

In vector notation, this becomes

$$\varphi\left(E(c)\,, \bar{c}\right) = 0\,, \tag{15}$$

where φ is applied to each component of $E(c)$ and $\bar{c}$.

The resulting system now reads

$$\frac{\partial}{\partial t}\theta\eta + L\eta = 0\,, \tag{16}$$

$$\frac{\partial}{\partial t}\left(\theta\xi_{min} + \bar{c}\right) + L\xi_{min} = 0\,, \tag{17}$$

$$-\varphi\left(E(c)\,, \bar{c}\right) = 0\,, \tag{18}$$

$$Q_{mob}(c) = 0\,, \tag{19}$$

where c can be represented as

$$c = c\left(\xi_{min},\, \xi_{mob},\, \eta\right) = S^1_{min} \cdot \xi_{min} + S^1_{mob} \cdot \xi_{mob} + S^{\perp}_1 \eta, \tag{20}$$

cf. (14). Note that (16) is now linear with respect to η and it is decoupled from the other equations (η is not contained in the other equations). The remaining nonlinearly coupled system (17)–(19) is reduced in size from $I + J + J_{min}$ rows to $I + J_{min}$ rows compared to the original system (5)–(10). Together with the size reduction of J rows, the J unknowns R could be dropped. They can be computed a posteriori. We now discretize the system in space and time. To keep the notation simple, we suppress subscripts indicating the discretization (except we denote L_h as the discretization of L). For the sake of simplicity, we assume the implicit

Euler time stepping scheme. We further mention that equation (16) in its discretized version can be solved for η directly (say, by a linear system solver like GMRES). Hence η is not viewed as a variable any longer. We therefore write $c = c(\xi_{min}, \xi_{mob})$ for the discretized function c.

The remaining discrete system in the variables $(\xi_{min}, \xi_{mob}, \bar{c})$ then reads

$$G_1 := \theta\xi_{min} + \bar{c} + \tau L_h \xi_{min} - \theta\xi_{min}^{old} - \bar{c}^{old} = 0\,, \tag{21}$$

$$G_2 := -\varphi\left(E\left(c\left(\xi_{min}, \xi_{mob}\right)\right), \bar{c}\right) = 0\,, \tag{22}$$

$$G_3 := Q_{mob}\left(c\left(\xi_{min}, \xi_{mob}\right)\right) = 0\,. \tag{23}$$

The superscript 'old' indicates the previous time-step. The time-step size is τ. We assume the domain Ω has been discretized into the grid set Ω_h with $|\Omega_h|$ grid points. Then $\xi_{min}, \xi_{mob}, \bar{c}$ are vectors with $J_{min} \cdot |\Omega_h|, J_{mob} \cdot |\Omega_h|, J_{min} \cdot |\Omega_h|$ components. These vectors are concatenations of the function values in every node of the grid. L_h is a linear mapping which is the discretization of the PDE operator L. In (22) and (23), the functions Q_{mob}, φ, E, c are to be applied to (the discretizations of) $\xi_{min}, \xi_{mob}, \bar{c}$ in every node separately. For example, a more detailed way to represent $c(\xi_{min}, \xi_{mob})$ is

$$c\left(\xi_{min}, \xi_{mob}\right) = \left[c\left(\xi_{min}^1, \xi_{mob}^1\right)^T, c\left(\xi_{min}^2, \xi_{mob}^2\right)^T, \ldots, c\left(\xi_{min}^{|\Omega_h|}, \xi_{mob}^{|\Omega_h|}\right)^T\right]^T,$$

where ξ_{min}^i, ξ_{mob}^i are our variables in one grid point. For the sake of simplicity, we define the abbreviations

$$\begin{aligned} \tilde{E}\left(\xi_{min}, \xi_{mob}\right) &:= E\left(c\left(\xi_{min}, \xi_{mob}\right)\right), \\ \tilde{Q}_{mob}\left(\xi_{min}, \xi_{mob}\right) &:= Q_{mob}\left(c\left(\xi_{min}, \xi_{mob}\right)\right). \end{aligned}$$

Let

$$G = \begin{pmatrix} G_1 \\ G_2 \\ G_3 \end{pmatrix}.$$

Then we have to solve the nonlinear system of equations

$$G\left(\xi_{min}, \xi_{mob}, \bar{c}\right) = 0\,.$$

Note that this is a nonsmooth system due to the definition of G_2 via the minimum function.

4. The Generalized Jacobian and Semismooth Functions

In this section, we will shortly review the definition for the generalized Jacobian and introduce an interesting result concerning the vector field G. More detailed statements and examples can be found in [4, 6, 7, 14, 15, 16].

Definition 1. *Let $F : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be locally Lipschitz continuous and $w \in \mathbb{R}^n$ be arbitrarily given. Let $D_F \subset \mathbb{R}^n$ be the set of differentiable points of F. Then the set*

$$\partial_B F(w) := \left\{H \in \mathbb{R}^{m\times n} \mid \exists \{w_k\} \subseteq D_F,\, w_k \longrightarrow w \text{ and } JF(w_k) \longrightarrow H\right\}$$

*is called the B-*subdifferential *of F in w, where JF is the Jacobian of F. The convex hull*

$$\partial F(w) := \operatorname{conv}\left(\partial_B F(w)\right)$$

is Clarke's generalized Jacobian *of F in w. Finally, the* C-subdifferential *is defined by*

$$\partial_C F(w) := \left(\partial F_1(w)^T \times \partial F_2(w)^T \times \cdots \times \partial F_m(w)^T\right)^T .$$

If $m = 1$ then $\partial F(w)$ is also called the *generalized gradient* of F. Note that if F is continuously differentiable in a neighborhood of w, both sets $\partial F(w)$ and $\partial_B F(w)$ contain the Jacobian of F as their only element.

Note that $\partial F(w) \subseteq \partial_C F(w)$ always holds . The C-subdifferential of F can be computed very easily, which is often not the case for the generalized Jacobian of F. Since $\operatorname{conv}(A \times B) = \operatorname{conv}(A) \times \operatorname{conv}(B)$ for any sets A, B, it follows that the C-subdifferential can also be represented as

$$\partial_C F(w) = \operatorname{conv}\left(\left(\partial_B F_1(w)^T \times \partial_B F_2(w)^T \times \cdots \times \partial_B F_m(w)^T\right)^T\right) . \tag{24}$$

Now let G be the mapping from the previous chapter. Each part $G_i\,(i = 1, 2, 3)$ of G is itself a multidimensional mapping. With $G_{i,j}$ we denote the components of the mapping $G_i (i = 1, 2, 3)$.

Lemma 1. *Let G be the nonlinear mapping that was introduced in (21)–(23) and let $p := |\Omega_h|$, say $\Omega_h = \{x_1, x_2, \ldots, x_p\}$. Furthermore, let $w = (\xi_{min}, \xi_{mob}, \bar{c})$ be an arbitrary element of $\mathbb{R}^{J_{min} \cdot p} \times \mathbb{R}^{J_{mob} \cdot p} \times \mathbb{R}^{J_{min} \cdot p}$ with components $\xi_{min} = \left(\xi_{min}(x_1)^T, \xi_{min}(x_2)^T, \ldots, \xi_{min}(x_p)^T\right)^T$ and $\xi_{mob}, \bar{c}$ defined in a similar way. Suppose that $c(\xi_{min}, \xi_{mob}) > 0$ componentwise. Then the following statements hold:*

(1) *The B-subdifferential of G can be written as the cross product*

$$\partial_B G(w) = \partial_B G_1(w) \times \partial_B G_2(w) \times \partial_B G_3(w)$$

with $\partial_B G_1(w) = \{JG_1(w)\}$ and $\partial_B G_3(w) = \{JG_3(w)\}$, where JG_1 and JG_3 are the Jacobians of G_1 and G_3, respectively.

(2) *The B-subdifferential of G_2 can be broken down into*

$$\partial_B G_2(w) = \partial_B G_2(w_1) \times \partial_B G_2(w_2) \times \ldots \times \partial_B G_2(w_p) ,$$

where $w_i = (\xi_{min}(x_i), \xi_{mob}(x_i), \bar{c}(x_i))$.

(3) *Let $x_i \in \Omega_h$, $a = (\xi_{min}(x_i), \xi_{mob}(x_i))$ and $b = \bar{c}(x_i)$. Then we have*

$$\partial_B G_2(w_i) = -\partial_B \varphi\left(\tilde{E}_1(a), b_1\right) \times -\partial_B \varphi\left(\tilde{E}_2(a), b_2\right) \times \ldots \times -\partial_B \varphi\left(\tilde{E}_{\bar{I}}(a), b_{\bar{I}}\right) .$$

(4) *Let x_i, a and b be as before. Then*

$$\partial_B \varphi\left(\tilde{E}_j(a), b_j\right) = \begin{cases} \left\{\left(\frac{\partial \tilde{E}_j(a)}{\partial \xi_{min}}, \frac{\partial \tilde{E}_j(a)}{\partial \xi_{mob}}, 0\right), \left(0, 0, e_l^T\right)\right\}, & if\ \tilde{E}_j(a) = b_j, \\ \left\{\left(0, 0, e_l^T\right)\right\}, & if\ \tilde{E}_j(a) > b_j, \\ \left\{\left(\frac{\partial \tilde{E}_j(a)}{\partial \xi_{min}}, \frac{\partial \tilde{E}_j(a)}{\partial \xi_{mob}}, 0\right)\right\}, & if\ \tilde{E}_j(a) < b_j, \end{cases}$$

where e_l is a unit vector, with all components vanishing and component $l = i \cdot J_{min} + j$ being one.

Proof. It is easy to see that G is locally Lipschitz continuous, since G_1, G_3 and $\tilde{E}$ are continuously differentiable and the minimum function φ is (globally) Lipschitz continuous.

(1) This statement follows directly from the observation that the two block components G_1 and G_3 are continuously differentiable, so that $\partial_B G_1(w) = \{JG_1(w)\}$ and $\partial_B G_3(w) = \{JG_3(w)\}$.

(2+3) These two statements are direct consequences of the definition of the corresponding B-subdifferentials, taking into account that the second argument $\bar{c}$ of the NCP-function φ can vary independently in every component. Note that statement (2) expresses the B-subdifferential $\partial_B G_2(w)$ as a Cartesian product of the B-subdifferentials at each of the p vectors w_i (which itself is still a vector in $\mathbb{R}^{J_{min}}$ for all $i = 1, \ldots, p$), whereas statement (3) gives the structure of the B-subdifferentials for each of these block components.

(4) The two cases $\tilde{E}_j(a) > b_j$ and $\tilde{E}_j(a) < b_j$ are obvious since φ is continuously differentiable in these cases, so that the B-subdifferential reduces to the existing gradient which can be calculated directly from (22). The remaining case $\tilde{E}_j(a) = b_j$ can be verified by choosing suitable sequences $\{b^k\}$ converging to b. □

Note the fact that G_1 and G_3 are continuously differentiable means that their B-subdifferential equals the cross product of the B-subdifferential of their components. Therefore, an immediate consequence of this lemma and (24) is the following

Corollary 1. *Let G be the nonlinear mapping that was introduced in* (21)–(23)*, and let $w = (\xi_{min}, \xi_{mob}, \bar{c})$ be an arbitrary element of $\mathbb{R}^{J_{min}\cdot|\Omega_h|} \times \mathbb{R}^{J_{mob}\cdot|\Omega_h|} \times \mathbb{R}_+^{J_{min}\cdot|\Omega_h|}$. Then it holds*

$$\partial G(w) = \partial_C G(w)\,.$$

5. Newton's Method and Active Set Strategy

Here we describe our Newton-type method applied to the nonlinear system of equations (21)–(23) and its relation to an active-set strategy. Some parts of this section is taken from the Habilitation Thesis [10], whereas the relationship between our Newton-type method and an active set strategy is, in principle, known [8, 9], although it has not been discussed within our context. The formulas to be derived in this section will, in particular, be needed in the subsequent sections.

The linearization of (21)–(23) via Newton's method leads to the linear system

$$H \begin{pmatrix} \Delta\xi_{min} \\ \Delta\xi_{mob} \\ \Delta\bar{c} \end{pmatrix} = - \begin{pmatrix} G_1 \\ G_2 \\ G_3 \end{pmatrix}, \tag{25}$$

with $H \in \partial_B G(w)$. Recall that G is not differentiable everywhere due to the nondifferentiability of the minimum function φ. In the points where G is (continuously) differentiable, H coincides with the Jacobian of G and the formula above is equal to the formula of the classical Newton

method. For the non-differentiable case, we have replaced the Jacobian in a suitable way following the theory of the semismooth Newton method from [15], see also [6, 7, 16, 14] for related material. In the following, we will construct one particular element of the B-subdifferential of G at w. The construction also shows how the other elements from the B-subdifferential can be obtained. In the differentiable case, our element is simply the Jacobian of G at w.

To this end, we introduce an active set strategy. As already mentioned, the vectors ξ_{min} and $\bar{c}$ each contain $|\Omega_h| \cdot J_{min}$ components (and the vector ξ_{mob} contains $|\Omega_h| \cdot J_{mob}$ components), where $|\Omega_h|$ is the number of grid points. We partition the set $\{1, \ldots, J_{min}\} \times \Omega_h$ into

$$\mathcal{A} = \left\{(i,x) \in \{1,\ldots,J_{min}\} \times \Omega_h \mid \tilde{E}_i\left(\xi_{min}(x), \xi_{mob}(x)\right) > \bar{c}_i(x)\right\}, \tag{26}$$

$$\mathcal{I} = \left\{(i,x) \in \{1,\ldots,J_{min}\} \times \Omega_h \mid \tilde{E}_i\left(\xi_{min}(x), \xi_{mob}(x)\right) \leq \bar{c}_i(x)\right\}. \tag{27}$$

Note that this partition is somewhat artificial. Alternatively, we could have defined the sets $\mathcal{A}$ and $\mathcal{I}$ in a different way by putting all index pairs satisfying $\tilde{E}_i\left(\xi_{min}(x), \xi_{mob}(x)\right) > \bar{c}_i(x)$ as well as an arbitrary subset of the index pairs satisfying $\tilde{E}_i\left(\xi_{min}(x), \xi_{mob}(x)\right) = \bar{c}_i(x)$ into the set $\mathcal{A}$, whereas the remaining index pairs belong to the index set $\mathcal{I}$. We will come back to this point at a later stage. For the moment, we use the two particular index sets $\mathcal{A}$ and $\mathcal{I}$ as defined in our previous formula. In contrast to the more general case, this simplifies to some extent our notation; moreover, it corresponds to our actual implementation of the nonsmooth Newton-type method.

For reasons that will become clear soon, the set $\mathcal{A}$ will be called the set of active indices, whereas its complement $\mathcal{I}$ will be called the set of inactive indices. We emphasize that this partitioning into active and inactive indices has to be computed in each Newton step, since ξ_{min}, ξ_{mob} and $\bar{c}$ change in each Newton iteration. Restricted to one species i, we can define the set of active and inactive indices as

$$\begin{aligned} \mathcal{A}_i &= \{x \in \Omega_h \mid (i,x) \in \mathcal{A}\}, \\ \mathcal{I}_i &= \{x \in \Omega_h \mid (i,x) \in \mathcal{I}\} \end{aligned}$$

for $i = 1, \ldots, J_{min}$. With these sets, we have

$$\varphi\left(\tilde{E}_i\left(\xi_{min}(x), \xi_{mob}(x)\right), \bar{c}_i(x)\right) = \begin{cases} \bar{c}_i(x), & \text{for } x \in \mathcal{A}_i, \\ \tilde{E}_i\left(\xi_{min}(x), \xi_{mob}(x)\right), & \text{for } x \in \mathcal{I}_i. \end{cases} \tag{28}$$

For an index $(i,x) \in \mathcal{I}$ with $\tilde{E}_i\left(\xi_{min}(x), \xi_{mob}(x)\right) = \bar{c}_i(x)$, the function $\varphi\left(\tilde{E}_i(\cdot,\cdot), \cdot\right)$ is not differentiable. As a replacement for the Jacobian, we take an element of its B-subdifferential, namely

$$\left(\frac{\partial \tilde{E}_i\left(\xi_{min}(x), \xi_{mob}(x)\right)}{\partial \xi_{min}}, \frac{\partial \tilde{E}_i\left(\xi_{min}(x), \xi_{mob}(x)\right)}{\partial \xi_{mob}}, 0\right)$$

(which is consistent with the previous definition of the active and inactive index sets, cf. (28)). For the indices in $\mathcal{A}$, we always have the differentiable case due to the definition of this index set. Due to Lemma 1, it follows that this particular element belongs to the B-subdifferential

of G at w. In our subsequent analysis, we will mainly work with this particular element from $\partial_B G(w)$ and therefore call it J. In particular, we consider the nonsmooth Newton iteration in (25) using this particular element J rather than an arbitrary element $H \in \partial_B G(w)$.

We now want to exploit the special structure of the particular matrix J in order to decompose the linear system (25). To this end, we reorder the entries of ξ_{min} and $\bar{c}$ in the following way

$$\xi_{min} = \begin{pmatrix} \xi_{min}^{\mathcal{A}} \\ \xi_{min}^{\mathcal{I}} \end{pmatrix}, \; \bar{c} = \begin{pmatrix} \bar{c}^{\mathcal{A}} \\ \bar{c}^{\mathcal{I}} \end{pmatrix}.$$

We apply the same reordering to our function G. Additionally, we reorder the rows of G_1 and G_2. Altogether, this corresponds to reordering the rows and columns of J. We perform the following decompositions:

$$G_1 = \begin{pmatrix} G_1^{\mathcal{A}} \\ G_1^{\mathcal{I}} \end{pmatrix}, L_h = \begin{pmatrix} L_h^{\mathcal{A}} \\ L_h^{\mathcal{I}} \end{pmatrix}, \; \tilde{E} = \begin{pmatrix} \tilde{E}_{\mathcal{A}} \\ \tilde{E}_{\mathcal{I}} \end{pmatrix}, \; S_{min}^1 = \left(S_{min,\mathcal{A}}^1 \mid S_{min,\mathcal{I}}^1 \right),$$

etc. Similar to the partition of ξ_{min}, we split the discrete differential operator L_h in

$$\begin{aligned} L_h^{\mathcal{A}} \xi_{min} &:= L_h^{\mathcal{A},\mathcal{A}} \xi_{min}^{\mathcal{A}} + L_h^{\mathcal{A},\mathcal{I}} \xi_{min}^{\mathcal{I}}, \\ L_h^{\mathcal{I}} \xi_{min} &:= L_h^{\mathcal{I},\mathcal{A}} \xi_{min}^{\mathcal{A}} + L_h^{\mathcal{I},\mathcal{I}} \xi_{min}^{\mathcal{I}}. \end{aligned}$$

With this restructuring, the linear system (25) reads

$$J \begin{pmatrix} \Delta\xi_{min}^{\mathcal{A}} \\ \Delta\xi_{min}^{\mathcal{I}} \\ \hline \Delta\xi_{mob} \\ \hline \Delta\bar{c}^{\mathcal{A}} \\ \Delta\bar{c}^{\mathcal{I}} \end{pmatrix} = - \begin{pmatrix} G_1^{\mathcal{A}} \\ G_1^{\mathcal{I}} \\ \hline -\bar{c}^{\mathcal{A}} \\ \hline -\tilde{E}_{\mathcal{I}} \\ \hline G_3 \end{pmatrix}, \tag{29}$$

with

$$J = \begin{pmatrix} \left(\theta I_{|\mathcal{A}|} + \tau L_h^{\mathcal{A},\mathcal{A}}\right) & \tau L_h^{\mathcal{A},\mathcal{I}} & 0 & I_{|\mathcal{A}|} & 0 \\ \tau L_h^{\mathcal{I},\mathcal{A}} & \left(\theta I_{|\mathcal{I}|} + \tau L_h^{\mathcal{I},\mathcal{I}}\right) & 0 & 0 & I_{|\mathcal{I}|} \\ 0 & 0 & 0 & -I_{|\mathcal{A}|} & 0 \\ -\frac{\partial \tilde{E}_{\mathcal{I}}}{\partial \xi_{min}^{\mathcal{A}}} & -\frac{\partial \tilde{E}_{\mathcal{I}}}{\partial \xi_{min}^{\mathcal{I}}} & -\frac{\partial \tilde{E}_{\mathcal{I}}}{\partial \xi_{mob}} & 0 & 0 \\ \frac{\partial \tilde{Q}_{mob}}{\partial \xi_{min}^{\mathcal{A}}} & \frac{\partial \tilde{Q}_{mob}}{\partial \xi_{min}^{\mathcal{I}}} & \frac{\partial \tilde{Q}_{mob}}{\partial \xi_{mob}} & 0 & 0 \end{pmatrix}. \tag{30}$$

From the third set of equations, we immediately obtain

$$-\Delta\bar{c}^{\mathcal{A}} = \bar{c}^{\mathcal{A}}. \tag{31}$$

There is no need to compute $\Delta\bar{c}^{\mathcal{A}}$, because of (31) we can simply set the *new* Newton iterate as

$$\bar{c}^{\mathcal{A},new} := 0$$

(this explains why $\mathcal{A}$ is called the active set). Furthermore, the unknowns $\Delta\bar{c}^{\mathcal{I}}$ only appear in the second set of equations. These equations can be solved for $\Delta\bar{c}^{\mathcal{I}}$:

$$\Delta\bar{c}^{\mathcal{I}} = -G_1^{\mathcal{I}} - \tau L_h^{\mathcal{I},\mathcal{A}} \cdot \Delta\xi_{min}^{\mathcal{A}} - \left(\theta I_{|\mathcal{I}|} + \tau L_h^{\mathcal{I},\mathcal{I}}\right) \cdot \Delta\xi_{min}^{\mathcal{I}} \,.$$

By these equations, $\Delta\bar{c}^{\mathcal{I}}$ can be computed a posteriori. After these two reductions, the resulting system reads

$$\tilde{J} \begin{pmatrix} \Delta\xi_{min}^{\mathcal{A}} \\ \Delta\xi_{min}^{\mathcal{I}} \\ \hline \Delta\xi_{mob} \end{pmatrix} = - \begin{pmatrix} G_1^{\mathcal{A}} - \bar{c}^{\mathcal{A}} \\ -\tilde{E}_{\mathcal{I}} \\ \hline G_3 \end{pmatrix} \tag{32}$$

with

$$\tilde{J} := \begin{pmatrix} \left(\theta I_{|\mathcal{A}|} + \tau L_h^{\mathcal{A},\mathcal{A}}\right) & \tau L_h^{\mathcal{A},\mathcal{I}} & 0 \\ -\frac{\partial \tilde{E}_{\mathcal{I}}}{\partial \xi_{min}^{\mathcal{A}}} & -\frac{\partial \tilde{E}_{\mathcal{I}}}{\partial \xi_{min}^{\mathcal{I}}} & -\frac{\partial \tilde{E}_{\mathcal{I}}}{\partial \xi_{mob}} \\ \frac{\partial \tilde{Q}_{mob}}{\partial \xi_{min}^{\mathcal{A}}} & \frac{\partial \tilde{Q}_{mob}}{\partial \xi_{min}^{\mathcal{I}}} & \frac{\partial \tilde{Q}_{mob}}{\partial \xi_{mob}} \end{pmatrix} . \tag{33}$$

This linear system is smaller than the original linear system (29), and it is solvable if and only if (29) is solvable. More precisely, the absolute values of the determinants of J and $\tilde{J}$ coincide. To see this, note that, by using elementary row and column additions as well as row interchanges, we can transform J into

$$J_1 := \begin{pmatrix} 0 & 0 & 0 & 0 & I_{|\mathcal{I}|} \\ 0 & 0 & 0 & -I_{|\mathcal{A}|} & 0 \\ \left(\theta I_{|\mathcal{A}|} + \tau L_h^{\mathcal{A},\mathcal{A}}\right) & \tau L_h^{\mathcal{A},\mathcal{I}} & 0 & 0 & 0 \\ -\frac{\partial \tilde{E}_{\mathcal{I}}}{\partial \xi_{min}^{\mathcal{A}}} & -\frac{\partial \tilde{E}_{\mathcal{I}}}{\partial \xi_{min}^{\mathcal{I}}} & -\frac{\partial \tilde{E}_{\mathcal{I}}}{\partial \xi_{mob}} & 0 & 0 \\ \frac{\partial \tilde{Q}_{mob}}{\partial \xi_{min}^{\mathcal{A}}} & \frac{\partial \tilde{Q}_{mob}}{\partial \xi_{min}^{\mathcal{I}}} & \frac{\partial \tilde{Q}_{mob}}{\partial \xi_{mob}} & 0 & 0 \end{pmatrix} .$$

Of course, J_1 is nonsingular if and only if J is nonsingular, and their determinants are the same except for possibly the factor -1. The same holds for J_1 and $\tilde{J}$, because J_1 results from $\tilde{J}$ by erasing the first rows and last columns, which belong to the block with the unity matrix and the negative unity matrix. Altogether, it follows that

$$\det J = \pm \det \tilde{J} \,. \tag{34}$$

In the following section, we will show that this determinant is nonzero.

6. Convergence of the Newton-type Algorithm

Now we want to study the nonsingularity of the matrix $\tilde{J}$ from the previous section (at the arbitrary point w considered so far which is not necessarily assumed to be a solution of our problem). The nonsingularity of the matrix J and therefore of $\tilde{J}$ was first shown in [10, Section 4.4.5] even in a more general setting. The proof given here, however, is different and the statement is stronger. The nonsingularity of this matrix is essential both for the solvability of the linear system (32) and for the local rate of convergence of our nonsmooth Newton method.

First, let us examine the submatrix

$$B := \begin{bmatrix} -\frac{\partial \tilde{E}_{\mathcal{I}}(\xi_{min},\xi_{mob})}{\partial \xi^{\mathcal{I}}_{min}} & -\frac{\partial \tilde{E}_{\mathcal{I}}(\xi_{min},\xi_{mob})}{\partial \xi_{mob}} \\ \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi^{\mathcal{I}}_{min}} & \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{mob}} \end{bmatrix}$$

of the matrix J. The nonsingularity of this matrix is shown even more generally in [10, Section 4.4.5]. Note that every entry of B is a block diagonal matrix. For example,

$$\frac{\partial \tilde{Q}_{mob}\left(\xi_{min},\xi_{mob}\right)}{\partial \xi^{\mathcal{I}}_{min}} = \operatorname{diag}\left(\frac{\partial \tilde{Q}_{mob}\left(\xi_{min}\left(x_1\right),\xi_{mob}\left(x_1\right)\right)}{\partial \xi^{\mathcal{I}}_{min}},\ldots,\frac{\partial \tilde{Q}_{mob}\left(\xi_{min}\left(x_p\right),\xi_{mob}\left(x_p\right)\right)}{\partial \xi^{\mathcal{I}}_{min}}\right),$$

with $p = |\Omega_h|$. By column and row interchanges, we can transform B into a block diagonal matrix, further denoted by C, so that every block corresponds to one grid point $x \in \Omega_h$ and has the form

$$\tilde{B} = \begin{bmatrix} -\frac{\partial \tilde{E}_{\mathcal{I}}(\xi_{min}(x),\xi_{mob}(x))}{\partial \xi^{\mathcal{I}}_{min}} & -\frac{\partial \tilde{E}_{\mathcal{I}}(\xi_{min}(x),\xi_{mob}(x))}{\partial \xi_{mob}} \\ \frac{\partial \tilde{Q}_{mob}(\xi_{min}(x),\xi_{mob}(x))}{\partial \xi^{\mathcal{I}}_{min}} & \frac{\partial \tilde{Q}_{mob}(\xi_{min}(x),\xi_{mob}(x))}{\partial \xi_{mob}} \end{bmatrix}.$$

With the definitions from Section 2 and the representation (20) of c, we can easily see that

$$\tilde{B} = \left(S^1_{min,\mathcal{I}} \mid S^1_{mob}\right)^T \Lambda_c \left(S^1_{min,\mathcal{I}} \mid S^1_{mob}\right)$$

holds, where $\Lambda_c = \operatorname{diag}\left(\frac{1}{c_1},\ldots,\frac{1}{c_I}\right)$. Since we postulated that all c_i should be positive on the whole domain Ω_h, the block $\tilde{B}$ is always symmetric positive definite in view of our rank condition (3). Therefore, C is symmetric positive definite. In particular, C is nonsingular. Since column and row interchanges do not change the rank of a matrix, it follows that B is also nonsingular. To prove the nonsingularity of the global matrix $\tilde{J}$ we deviate from the strategy used in [10].

The columns of B form a basis of its column space. Consequently, there exist unique matrices D_1 and D_2 such that

$$\begin{pmatrix} -\frac{\partial \tilde{E}_{\mathcal{I}}(\xi_{min},\xi_{mob})}{\partial \xi^{\mathcal{I}}_{min}} \\ \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi^{\mathcal{I}}_{min}} \end{pmatrix} D_1 + \begin{pmatrix} -\frac{\partial \tilde{E}_{\mathcal{I}}(\xi_{min},\xi_{mob})}{\partial \xi_{mob}} \\ \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{mob}} \end{pmatrix} D_2 = -\begin{pmatrix} -\frac{\partial \tilde{E}_{\mathcal{I}}(\xi_{min},\xi_{mob})}{\partial \xi^{\mathcal{A}}_{min}} \\ \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi^{\mathcal{A}}_{min}} \end{pmatrix}$$

or, equivalently,

$$B \cdot \begin{pmatrix} D_1 \\ D_2 \end{pmatrix} = -\begin{pmatrix} -\frac{\partial \tilde{E}_{\mathcal{I}}(\xi_{min},\xi_{mob})}{\partial \xi^{\mathcal{A}}_{min}} \\ \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi^{\mathcal{A}}_{min}} \end{pmatrix}.$$

Next we post-multiply $\tilde{J}$ in (33) with the block matrix

$$X := \begin{pmatrix} I & 0 & 0 \\ D_1 & I & 0 \\ D_2 & 0 & I \end{pmatrix}$$

from the right hand side and obtain

$$\tilde{J}_1 := \tilde{J} \cdot X = \left(\begin{array}{c|cc} \theta I_{|\mathcal{A}|} + \tau L_h^{\mathcal{A},\mathcal{A}} + \tau L_h^{\mathcal{A},\mathcal{I}} \cdot D_1 & \tau L_h^{\mathcal{A},\mathcal{I}} & 0 \\ \hline 0 & -\frac{\partial \tilde{E}_{\mathcal{I}}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{I}}} & -\frac{\partial \tilde{E}_{\mathcal{I}}(\xi_{min},\xi_{mob})}{\partial \xi_{mob}} \\ 0 & \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{I}}} & \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{mob}} \end{array} \right).$$

Since the determinant of X is obviously 1, it follows that

$$\det \tilde{J} = \det \tilde{J}_1 .$$

On the other hand, the determinant of $\tilde{J}_1$ is given by

$$\det\left(\tilde{J}_1\right) = \det\left(\theta I_{|\mathcal{A}|} + \tau L_h^{\mathcal{A},\mathcal{A}} + \tau L_h^{\mathcal{A},\mathcal{I}} \cdot D_1\right) \cdot \det B.$$

Therefore, in view of the previous discussion, $\tilde{J}_1$ is nonsingular if and only if $H := \theta I_{|\mathcal{A}|} + \tau L_h^{\mathcal{A},\mathcal{A}} + \tau L_h^{\mathcal{A},\mathcal{I}} \cdot D_1$ is nonsingular.

Now we apply Lemma 2 from the appendix to the matrix H and obtain

$$\det H = \sum_{\beta} \det \theta I_{\beta,\beta} \cdot \det\left(\tau L_h^{\mathcal{A},\mathcal{A}} + \tau L_h^{\mathcal{A},\mathcal{I}} \cdot D_1\right)_{\bar{\beta},\bar{\beta}},$$

where $\beta \subset \{1,\ldots,|\mathcal{A}|\}$ and $\bar{\beta} := \{1,\ldots,|\mathcal{A}|\}\backslash\beta$. The matrices $\theta I_{\beta,\beta}$ and $\left(\tau L_h^{\mathcal{A},\mathcal{A}} + \tau L_h^{\mathcal{A},\mathcal{I}} \cdot D_1\right)_{\bar{\beta},\bar{\beta}}$ are submatrices of $\theta I_{|\mathcal{A}|}$ resp. $\left(\tau L_h^{\mathcal{A},\mathcal{A}} + \tau L_h^{\mathcal{A},\mathcal{I}} \cdot D_1\right)$. Since the determinant of a 0×0 matrix is defined as 1, we get

$$\begin{aligned} \det H &= \sum_{\beta} \theta^{|\beta|} \cdot \det\left(\tau L_h^{\mathcal{A},\mathcal{A}} + \tau L_h^{\mathcal{A},\mathcal{I}} \cdot D_1\right)_{\bar{\beta},\bar{\beta}} \\ &= \theta^{|\mathcal{A}|} + \sum_{|\beta|<|\mathcal{A}|} \theta^{|\beta|} \cdot \tau^{|\bar{\beta}|} \det\left(L_h^{\mathcal{A},\mathcal{A}} + L_h^{\mathcal{A},\mathcal{I}} \cdot D_1\right)_{\bar{\beta},\bar{\beta}} . \end{aligned}$$

For the next theorem, we assume that L_h is an arbitrary discretization of the PDE operator L. This discretization might depend on h but not on τ. Furthermore, we assume that the spatial step size h is given and fixed. Then our theorem states the dependence of the nonsingularity of J on the time step size τ.

Theorem 1. *For sufficiently small time steps τ, the system matrix J is nonsingular. Furthermore, there are at most $J_{min} \cdot |\Omega_h|$ time steps τ such that J is singular.*

Proof. Note that the determinant of H is a polynomial in τ. The degree of this polynomial is $|\mathcal{A}|$, where $|\mathcal{A}| \leq J_{min} \cdot |\Omega_h|$ always holds by definition of the active set $\mathcal{A}$. So this polynomial has a maximum degree of $J_{min} \cdot |\Omega_h|$. It is not the zero polynomial since it has $\theta^{|\mathcal{A}|}$ as constant term. So $J_{min} \cdot |\Omega_h|$ is also the maximum number of its roots. Hence either all roots are complex, or there exists a smallest positive root which is our smallest time step. Since $\det B \neq 0$ always holds, and since we have $\det \tilde{J}_1 = \det \tilde{J} = \pm \det J$ according to (34), the statement follows. □

Additionally, we now assume that our PDE operator L_h emerged from a difference scheme of first or second order. In fact, the subsequent discussion would hold for any PDE operator that contains $\frac{1}{h}$ in every nonvanishing entry. The variable h is the spatial grid width of our discretization. Hence every entry of L_h that does not vanish contains the factor $\frac{1}{h}$. We therefore conclude that every non-vanishing entry of $L_h^{\mathcal{A},\mathcal{A}} + L_h^{\mathcal{A},\mathcal{I}} \cdot D_1$ contains the factor $\frac{1}{h}$ (some entries may contain $\frac{1}{h^2}$). Hence, for every index subset δ, there exists a matrix L_δ such that

$$\left(L_h^{\mathcal{A},\mathcal{A}} + L_h^{\mathcal{A},\mathcal{I}} \cdot D_1\right)_{\delta,\delta} = \frac{1}{h} \cdot L_\delta$$

holds.

In contrast to the previous theorem, we study in our next result the correlation of the nonsingularity of J for variable space step size h, while we assume that the time step size τ is given and fixed.

Theorem 2. *Let the PDE operator L_h result from a difference scheme of first or second order. Then the system matrix J is nonsingular for all sufficiently small space steps h. Furthermore, there are at most $2 \cdot J_{min} \cdot |\Omega_h|$ space steps h such that J is singular.*

Proof. Every non-vanishing entry of L_h is a polynomial in $\frac{1}{h}$ of first or second order. The same holds for $L_h^{\mathcal{A},\mathcal{A}} + L_h^{\mathcal{A},\mathcal{I}} \cdot D_1$ and all its submatrices. With the Leibniz formula, we conclude that $\det\left(L_h^{\mathcal{A},\mathcal{A}} + L_h^{\mathcal{A},\mathcal{I}} \cdot D_1\right)_{\bar{\beta},\bar{\beta}}$ is a polynomial in $\frac{1}{h}$ of maximal degree $2 \cdot |\mathcal{A}|$ with a zero constant term. Therefore, $\det H$ is always a polynomial in $\frac{1}{h}$ of degree at most $2 \cdot J_{min} \cdot |\Omega_h|$. Again, $\theta^{|\mathcal{A}|}$ is the constant term of this polynomial, hence it is not the zero polynomial. Therefore it has at most $2 \cdot J_{min} \cdot |\Omega_h|$ roots.

Let z_∞ be the largest real root of this polynomial. Then there exists a corresponding smallest positive space step h_0 with $z_\infty = \frac{1}{h_0}$. So $\det H \neq 0$ holds for all $h \in (0, h_0)$. Since $\det B \neq 0$ always holds, and because $\det \tilde{J}_1 = \det \tilde{J} = \pm \det J$, according to (34), we have proved everything. □

We now generalize the previous two theorems slightly.

Corollary 2. *Let $w^* := (\xi^*_{min}, \xi^*_{mob}, \bar{c}^*) \in \mathbb{R}^{J_{min} \cdot |\Omega_h|} \times \mathbb{R}^{J_{mob} \cdot |\Omega_h|} \times \mathbb{R}_+^{J_{min} \cdot |\Omega_h|}$ be a grid vector. Then the following statements hold:*

1. *Let h be given. Then all $H \in \partial_B G(w^*)$ are nonsingular for all sufficiently small time steps τ. Furthermore, there is only a finite number of time steps τ such that at least one element in $\partial_B G(w^*)$ is singular.*
2. *Let τ be given and let L_h be as in Theorem 2. Then all $H \in \partial_B G(w^*)$ are nonsingular for all sufficiently small space steps h. Furthermore, there are only a finite number of space steps h such that at least one element in $\partial_B G(w^*)$ is singular.*

Proof. So far, we have shown the two statements for the particular element J from the B-subdifferential. However, as outlined after the definitions of the active and inactive index sets $\mathcal{A}$ and $\mathcal{I}$ in (26) and (27), respectively, the other elements from $\partial_B G(w^*)$ can be obtained by a minor change of these definitions where, basically, some of the index pairs from $\mathcal{I}$ are moved

to the index set $\mathcal{A}$. The nonsingularity of the corresponding element can then be shown in essentially the same way as we proved the nonsingularity of the particular element J. Hence the desired statements follow from Theorems 1 and 2, respectively, taking into account that the number of matrices in $\partial_B G(w^*)$ is finite, cf. Lemma 1. □

Note that all the previous nonsingularity results hold at an arbitrary point w (or w^*). Hence all iterations of our Newton-type method are (not only locally) well-defined. But it should be mentioned that the minimal time step size in two different grid points may differ. So this value could decrease constantly during a Newton iteration.

We next give an exact statement of our Newton-type method for the solution of the nonlinear system of equations from (29).

Algorithm 1. *(Nonsmooth Newton Method)*

(S.0) *Let* $w^0 \in \mathbb{R}^{J_{min}\cdot|\Omega_h|} \times \mathbb{R}^{J_{mob}\cdot|\Omega_h|} \times \mathbb{R}^{J_{min}\cdot|\Omega_h|}$, *and set* $k := 0$.

(S.1) *If* $G\left(w^k\right) = 0$, *stop.*

(S.2) *Let* $J_k \in \partial_B G\left(w^k\right)$ *be the element defined in Section 5. Find a solution* d^k *of the linear system*

$$J_k d = -G\left(w^k\right) .$$

(S.3) *Set* $w^{k+1} := w^k + d^k$, $k \leftarrow k+1$, *and go to (S.1).*

The following is the main local convergence result for this Newton-type method.

Theorem 3. *Let* $w^* := (\xi^*_{min}, \xi^*_{mob}, \bar{c}^*) \in \mathbb{R}^{J_{min}\cdot|\Omega_h|} \times \mathbb{R}^{J_{mob}\cdot|\Omega_h|} \times \mathbb{R}^{J_{min}\cdot|\Omega_h|}$ *be a grid vector such that* w^* *is a solution of the nonlinear system* $G(w) = 0$ *and* H *is nonsingular for all* $H \in \partial_B G(w^*)$. *Then there exists an* $\epsilon > 0$ *such that for every starting point* $w^0 \in B_\epsilon(w^*)$, *the following assertions hold:*

(1) *The Newton-type iteration defined in Algorithm 1 is well-defined and produces a sequence* $\left\{w^k\right\}$ *that converges to* w^*.

(2) *The rate of convergence is quadratic.*

Proof. The assertion follows from [15] as soon as we have shown that the equation operator G is a strongly semismooth function, see [6, 7, 16, 14] and references therein for further details on (strongly) semismooth functions. We apply several known results from these papers in order to verify the strong semismoothness of G.

First note that the strong semismoothness of G is equivalent to the strong semismoothness of all component functions of G. Now, the functions E_j, Q_{mob}, G_1 and the linear transformations $(\xi_{min}, \xi_{mob}, \eta) \mapsto c(\xi_{min}, \xi_{mob}, \eta)$ are continuous differentiable with derivatives that are locally Lipschitz-continuous on their domains. Therefore, these functions are strongly semismooth according to [7]. Moreover, the minimum function is known to be strongly semismooth, and the composition of strongly semismooth functions is again strongly semismooth. Hence also the remaining components of the mapping G are strongly semismooth. □

Unfortunately, we do not know a priori whether the requirement of Theorem 3 regarding the nonsingularity of all elements from the B-subdifferential of G holds. However, Corollary 2 guarantees that it is at least very unlikely to hit a point where this requirement is not satisfied.

Moreover, it shows that we can change this situation by changing the time step size τ or the spatial step size h (for practical reasons, it is easier to change τ). But after changing the time step size τ, the Newton iteration has to be restarted. So the previous statement is of more theoretic nature, because is is unlikely to stumble across the same iterate with this changed time step size. In our computational test runs, we never had problems with singular matrices from $\partial_B G$.

7. Schur Complement Approach

In this section, we want to discuss how the linear system (32) can be transformed in such a way that it can be solved more efficiently. To this end, we utilize a Schur complement approach. We begin by introducing some abbreviations to keep the formulas clear:

$$A := \left(\theta I_{|\mathcal{A}|} + \tau L_h^{\mathcal{A},\mathcal{A}}\right), \qquad B := [B_1 \mid 0] := \left[\tau L_h^{\mathcal{A},\mathcal{I}} \mid 0\right],$$

$$C := \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} := \begin{bmatrix} -\frac{\partial \tilde{E}_{\mathcal{I}}}{\partial \xi_{min}^{\mathcal{A}}} \\ \frac{\partial \tilde{Q}_{mob}}{\partial \xi_{min}^{\mathcal{A}}} \end{bmatrix}, \quad D := \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} := \begin{bmatrix} -\frac{\partial \tilde{E}_{\mathcal{I}}}{\partial \xi_{min}^{\mathcal{I}}} & -\frac{\partial \tilde{E}_{\mathcal{I}}}{\partial \xi_{mob}} \\ \frac{\partial \tilde{Q}_{mob}}{\partial \xi_{min}^{\mathcal{I}}} & \frac{\partial \tilde{Q}_{mob}}{\partial \xi_{mob}} \end{bmatrix}.$$

With these abbreviations, (32) reads

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{pmatrix} \Delta\xi_{min}^{\mathcal{A}} \\ \Delta\xi_{min}^{\mathcal{I}} \\ \Delta\xi_{mob} \end{pmatrix} = - \begin{pmatrix} G_1^{\mathcal{A}} - \bar{c}^{\mathcal{A}} \\ -\tilde{E}_{\mathcal{I}} \\ G_3 \end{pmatrix}.$$

We begin by writing this linear system in detail

$$A \cdot \Delta\xi_{min}^{\mathcal{A}} + B_1 \cdot \Delta\xi_{min}^{\mathcal{I}} = -G_1^{\mathcal{A}} + \bar{c}^{\mathcal{A}}, \tag{35}$$

$$C_1 \cdot \Delta\xi_{min}^{\mathcal{A}} + D_{11} \cdot \Delta\xi_{min}^{\mathcal{I}} + D_{12} \cdot \Delta\xi_{mob} = \tilde{E}_{\mathcal{I}}, \tag{36}$$

$$C_2 \cdot \Delta\xi_{min}^{\mathcal{A}} + D_{21} \cdot \Delta\xi_{min}^{\mathcal{I}} + D_{22} \cdot \Delta\xi_{mob} = -G_3. \tag{37}$$

Similar to the previous section, D_{11} is a block diagonal matrix, where each block has the form $\left(S_{min,\mathcal{I}}^1\right)^T \Lambda_c \left(S_{min,\mathcal{I}}^1\right)$. Likewise, D_{22} is a block diagonal matrix, where each block has the form $\left(S_{mob}^1\right)^T \Lambda_c \left(S_{mob}^1\right)$. Recall that $S_{min,\mathcal{I}}^1$ and S_{mob}^1 have full column rank, $\Lambda_c = \operatorname{diag}\left(\frac{1}{c_1}, \frac{1}{c_2}, \ldots, \frac{1}{c_I}\right)$, and that all c_i are assumed to be positive. Hence D_{11} and D_{22} are positive definite and therefore nonsingular.

We now rewrite (36) to obtain

$$D_{11} \cdot \Delta\xi_{min}^{\mathcal{I}} = \tilde{E}_{\mathcal{I}} - D_{12} \cdot \Delta\xi_{mob} - C_1 \cdot \Delta\xi_{min}^{\mathcal{A}}. \tag{38}$$

Furthermore, we transform (37) into

$$\Delta\xi_{mob} = -\left(D_{22}\right)^{-1} \cdot G_3 - \left(D_{22}\right)^{-1} \cdot C_2 \cdot \Delta\xi_{min}^{\mathcal{A}} - \left(D_{22}\right)^{-1} \cdot D_{21} \cdot \Delta\xi_{min}^{\mathcal{I}}. \tag{39}$$

Now we insert $\Delta\xi_{mob}$ into (38) and obtain

$$\Delta\xi_{min}^{\mathcal{I}} = \tilde{D}^{-1}\tilde{E}_{\mathcal{I}} + \tilde{D}^{-1}D_{12}D_{22}^{-1} \cdot G_3 - \tilde{D}^{-1}\left(C_1 - D_{12} \cdot D_{22}^{-1} \cdot C_2\right) \cdot \Delta\xi_{min}^{\mathcal{A}} \tag{40}$$

with $\tilde{D} = \left(D_{11} - D_{12}D_{22}^{-1}D_{21}\right)$. $\tilde{D}$ can be obtained from D through a block Gauss elimination step. It is a Schur complement of D. Since D is positive definite, $\tilde{D}$ is also positive definite, cf. [19]. In particular, $\tilde{D}$ is nonsingular.

Finally, we insert $\Delta\xi_{min}^{\mathcal{I}}$ in (35) and obtain

$$\left[A - B_1 \cdot \tilde{D}^{-1}\tilde{C}\right] \cdot \Delta\xi_{min}^{\mathcal{A}} = -G_1^{\mathcal{A}} + \bar{c}^{\mathcal{A}} - B_1\tilde{D}^{-1}\tilde{E}_{\mathcal{I}} - B_1\tilde{D}^{-1}D_{12} \cdot D_{22}^{-1} \cdot G_3, \tag{41}$$

with $\tilde{C} = \left(C_1 - D_{12}D_{22}^{-1}C_2\right)$.

To obtain the solution of the initial linear system (35)–(37), we first solve (41) for $\Delta\xi_{min}^{\mathcal{A}}$. Subsequently, we compute $\Delta\xi_{min}^{\mathcal{I}}$ from (40) which essentially requires some matrix-vector multiplications. Finally, we get $\Delta\xi_{mob}$ from (39) again by matrix-vector multiplications and additions.

The main computational cost is, on the one hand, in solving the linear system (41) and, on the other hand, in the computation of the inverses needed in (39)-(41).

We now want to take a closer look at the computation of the required inverses. To be more precise, we do not really need the inverses themselves, but we need their effect on several matrices resp. vectors. For the purpose of clarifying the computational cost, we introduce the variables $X_1, X_2, x_3, Y_1, y_2, y_3, z_3$, which we define subsequently. Now we recapitulate the transformation.

First we solve the linear system

$$D_{22} \cdot [X_1 \mid X_2 \mid x_3] = [D_{21} \mid C_2 \mid G_3] \ .$$

The matrices D_{22}, D_{21} as well as C_2 are block diagonal matrices. The dimensions of the blocks of all three matrices match up in a way that this linear system can be broken down in $|\Omega_h|$ totally independent linear systems of size $J_{mob} \times J_{mob}$. We already mentioned that all the blocks of D_{22} are positive definite. So we can solve these small systems by the Cholesky decomposition. Note that all of these have multiple right hand sides. However, this does not increase the computational cost significantly, since we need only one decomposition. The resulting matrices X_1 and X_2 are again block diagonal matrices.

Now we compute

$$\tilde{D} = D_{11} - D_{12} \cdot X_1 \ , \ \tilde{C} = C_1 - D_{12} \cdot X_2 \ , \ z_3 := D_{12} \cdot x_3.$$

Again this can be done block-wise. Therefore, $\tilde{D}$ and $\tilde{C}$ have block diagonal form, too.

Next we solve the linear system

$$\tilde{D} \cdot [Y_1 \mid y_2 \mid y_3] = \left[\tilde{C} \mid z_3 \mid \tilde{E}_{\mathcal{I}}\right] \ .$$

For this system, the same applies as for the previous one. Here $\tilde{C}$ and $\tilde{D}$ have a matching block diagonal form. Therefore, Y_1 is a block diagonal matrix, whereas $z_3, \tilde{E}_{\mathcal{I}}$ are just vectors. Again, the small systems have multiple right-hand sides. This time, however, the square blocks of $\tilde{D}$ have variable sizes from 0×0 to $J_{min} \times J_{min}$.

Using this notation, our transformed system reads

$$[A - B_1 \cdot Y_1] \cdot \Delta\xi^{\mathcal{A}}_{min} = -G^{\mathcal{A}}_1 + \bar{c}^{\mathcal{A}} - B_1 \cdot [y_2 + y_3] \tag{42}$$

$$\Delta\xi^{\mathcal{I}}_{min} = y_2 + y_3 - Y_1 \cdot \Delta\xi^{\mathcal{A}}_{min} \tag{43}$$

$$\Delta\xi_{mob} = -x_3 - X_1 \cdot \Delta\xi^{\mathcal{I}}_{min} - X_2 \cdot \Delta\xi^{\mathcal{A}}_{min} . \tag{44}$$

Through this transformation of the original system (32), we could exploit especially the structure of D and its submatrices, which would have been unused otherwise.

Since B_1 is sparse and Y_1 is block diagonal, the product $B_1 \cdot Y_1$ again is sparse. Its structure is similar to the structure of A. Therefore, the matrix $A - B_1 \cdot Y_1$ in the linear system (42) is sparse, too. It can be solved by a linear solver like GMRES.

Finally, it should be mentioned that we really have only one Newton-type algorithm and that is the one which was introduced in Algorithm 1. The Schur-complement approach and the simplifications in (32) and (33) are only different ways to solve the resulting linear systems efficiently.

8. Numerical Example

The reactive transport problem introduced in Section 2 was implemented in two versions using MATLAB®. One version uses the Schur-complement approach from Section 7, whereas the other version utilizes the whole system (25) with the special element $J \in \partial_B G$.

For both versions, the discretization of the PDE-operator was done via the same difference scheme of second order. Both versions have to solve the same a priori linear decoupled system, the discretization of (16). This is done through a GMRES iteration in both implementations, since it is a sparse system. In practice, this seems to work very well for this particular linear system. Usually only 2 or 3 steps are needed to calculate a sufficiently accurate solution. Thus we will focus on the Newton iteration.

In our test example (taken from [10]), the interaction of CO_2 with minerals is considered. In these days, we are facing the global warming of the earth which is at least partly due to the CO_2-concentration in the atmosphere. Therefore, techniques have been investigated to inject CO_2 into the subsurface. The long term storage of CO_2 beneath the surface of our planet is the desired goal. This might be more likely if the carbon precipitates would form minerals than the carbon being dissolved in the ground water.

We use the following generic simplified set of chemical reactions to model the desired mechanism:

$$\begin{aligned}
CO_2^{(aq)} + H_2O &\xleftrightarrow{R_1} HCO_3^- + H^+ \\
\text{Calcite} + H^+ &\xleftrightarrow{R_2} Ca^{2+} + HCO_3^- \\
\text{Min A} + 3H^+ &\xleftrightarrow{R_3} Me^{3+} + SiO_2^{(aq)} \\
\text{Min B} + 2H^+ &\xleftrightarrow{R_4} Me^{3+} + HCO_3^-
\end{aligned}$$

It consists of 3 minerals (calcite and mineral B are carbonates, mineral A is a silicate) and 6 species which are dissolved in the ground water and one aqueous tracer. More details and

insights for this example, especially its internal functionality, can be found in [10, Subsection 4.5.2].

The technical details for this example are: domain $\Omega = (0, 10) \times (0, 6)$, Darcy velocity $q = (0.015, 0)^T$, water content $\theta = 0.3$, (i.e. pore velocity $\|q\|/\theta = 0.05$), longitudinal/transversal dispersion length $(\beta_l, \beta_t)^T = (0.3, 0.03)^T$, time step size $\tau = 0.1$. The equilibrium constant of the first reaction is $K_1 = 0.1$, where the activity of H_2O is already incorporated; i.e. $c_{H^+} c_{HCO_e^-}/c_{CO_2} = 0.1$. The solubility products of the three mineral reactions are $K_2 = 100$, $K_3 = 10$, $K_4 = 1.25$; i.e. $c_{Ca^{2+}} c_{HCO_3^-}/c_{H^+} = 100$ (if $c_{Calcite} > 0$), etc. The initial values are $c_{CO_2} = c_{HCO_3^-} = c_{SiCO_2} = 1$, $c_{H^+} = 0.1$, $c_{Me^{3+}} = 0.01$, $c_{Ca^{2+}} = 10$ (constant within Ω), and $c_A = 0.2$ for $x \geq 6$, $c_{Calcite} = 0.2$ for $1 < x < 6$, and zero else. The Dirichlet boundary values for the mobile species are $c_{CO_2} = 3.787$, $c_{H^+} = 0.3124\,, c_{HCO_3^-} = 1.212$, $c_{Me^{3+}} = 0.01$, $c_{SiO_2} = 1$, $c_{Ca^{2+}} = 10$ on $\{0\} \times [1.5, 4.5]$, whereas we use the initial values on $(0, y)$ with $y < 1.5$, $y > 4.5$. For the other three borders, the homogeneous Neumann boundary condition is given.

In the following calculation, we set the spatial and the time step to $h = \tau = 0.1$. With this setting, we get 6100 grid nodes for an equidistant quadratic grid. The discretization was done via a second-order finite difference method. With the Schur complement implementation we calculate the resulting concentrations for the 10 species for 3600 time steps, i.e. a time span of 360 seconds. The results have been checked to match the results from [10].

Figures 1–3 visualize the numerical results. Note that the differences to the results given in [10] are only due to a different color scaling. There is a slow water flow in horizontal direction from the left to the right. With it enters dissolved CO_2 into the computational domain. This decreases the pH value (the negative common logarithm of the concentration of H^+ ions in the water). The water stream of low pH value dissolves Mineral A and Calcite, when it reaches those areas. Moreover, the dissolution of Mineral A leads to an immediate precipitation of Mineral B.

Table 1 shows the quadratic convergence for both implementations of our Newton-type methods as predicted in the previous theory. The third column contains the errors of the Schur complement method, whereas the fourth column gives the errors of the full Jacobian method. The good consistency of these errors shows that these two methods realize the same Newton method where only the linear systems are solved differently. Usually these two methods need the same number of Newton iterations to get below the termination condition of $2 \cdot 10^{-6}$. With time step size $\tau = 0.1$, they both need almost always only two Newton iterations after about 10 time iterations.

In Table 2 we compare the linear systems which arise in these two methods. Both of these sparse systems are solved with the GMRES(30) method. The numbers in the last two columns show the total number of inner GMRES iterations which are needed in both methods. The fifth and sixth columns display the condition numbers of the linear systems of both methods. Finally, we present in the third and fourth columns the dimensions of these linear systems. Of course, the linear system of the full Jacobian method has always the same size, since the arising Jacobians always stem from the same function. While the linear system of the Schur

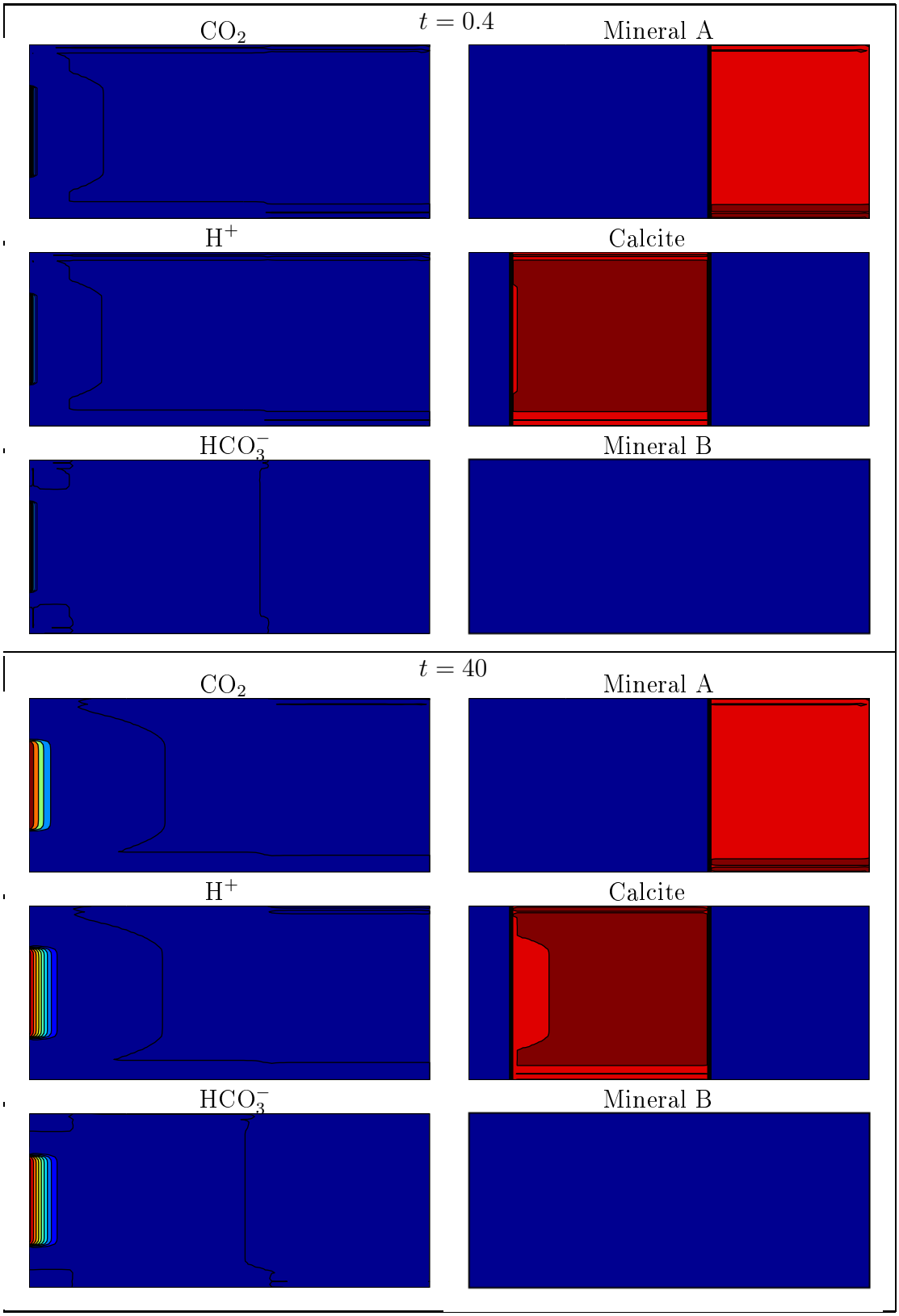

FIGURE 1. Results obtained after $t = 0.4$ seconds. (The graphics are compressed by a factor 1.5 in vertical direction.)

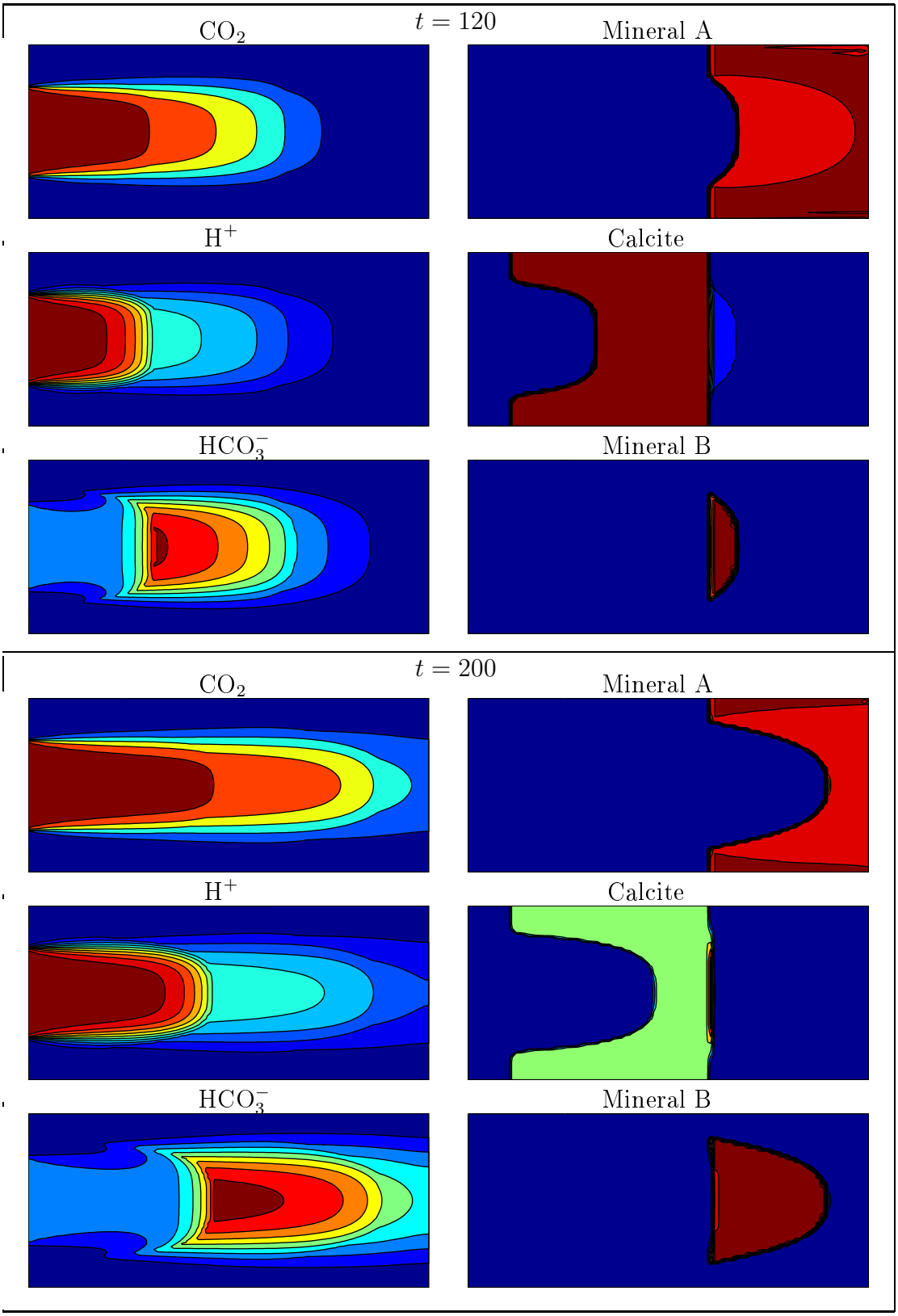

FIGURE 2. Results obtained after $t = 120$ seconds. (The graphics are compressed by a factor 1.5 in vertical direction.)

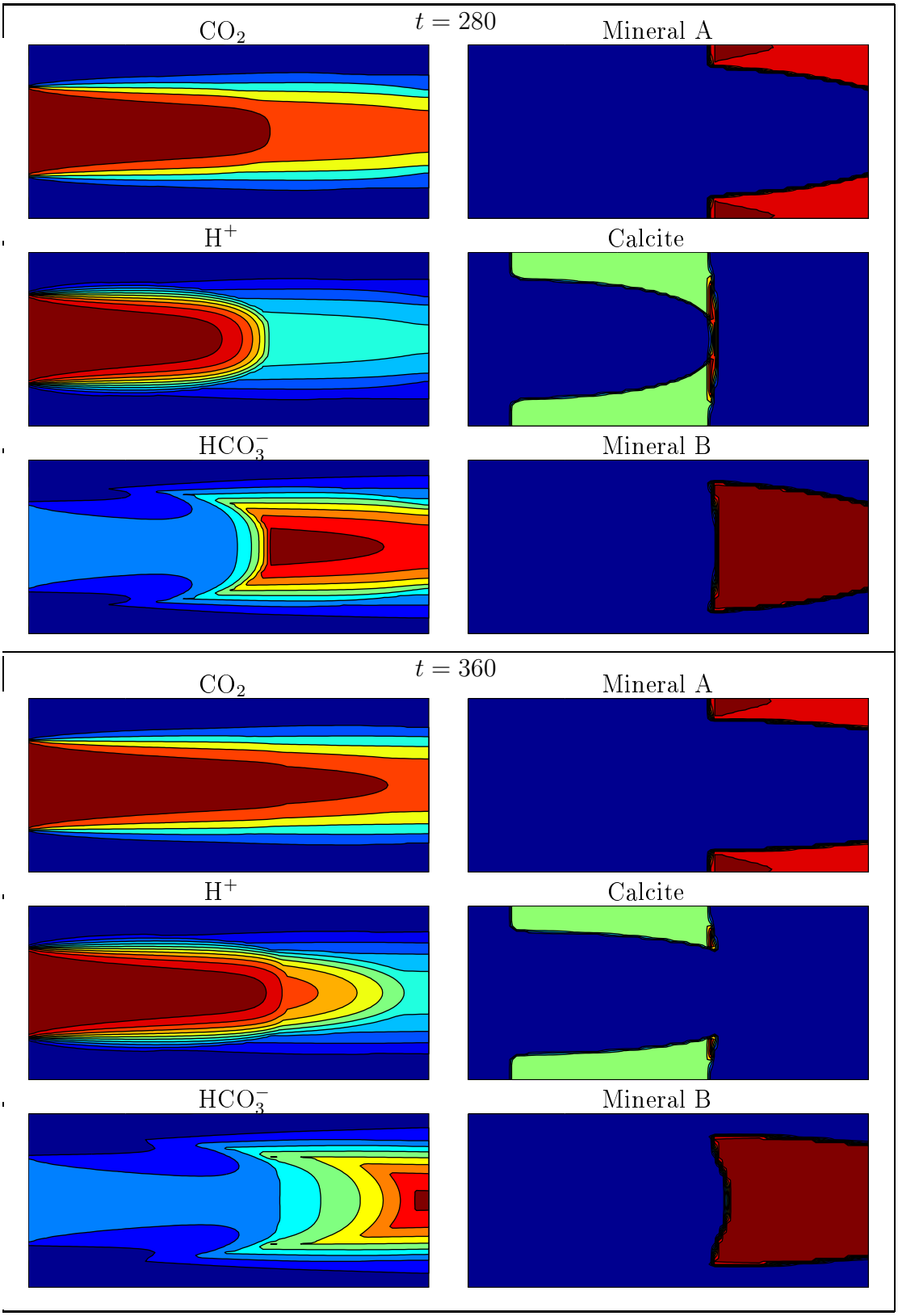

FIGURE 3. Results obtained after $t = 280$ seconds. (The graphics are compressed by a factor 1.5 in vertical direction.)

time step	iteration	method Schur:$\|G(z)\|_2$	method full:$\|G(z)\|_2$
1	0	$3.1753 \cdot 10^{0}$	$3.1753 \cdot 10^{0}$
	1	$2.7353 \cdot 10^{0}$	$2.3026 \cdot 10^{-1}$
	2	$1.6990 \cdot 10^{-2}$	$6.2355 \cdot 10^{-3}$
	3	$2.9673 \cdot 10^{-3}$	$2.3402 \cdot 10^{-6}$
	4	$5.2980 \cdot 10^{-7}$	$3.9298 \cdot 10^{-9}$
2	0	$1.8504 \cdot 10^{0}$	$1.8504 \cdot 10^{0}$
	1	$3.3186 \cdot 10^{-2}$	$3.3186 \cdot 10^{-2}$
	2	$6.9773 \cdot 10^{-4}$	$6.9771 \cdot 10^{-4}$
	3	$2.9498 \cdot 10^{-8}$	$4.2795 \cdot 10^{-8}$
3	0	$1.4602 \cdot 10^{0}$	$1.4602 \cdot 10^{0}$
	1	$2.1604 \cdot 10^{-2}$	$2.1604 \cdot 10^{-2}$
	2	$1.0084 \cdot 10^{-4}$	$1.0084 \cdot 10^{-4}$
	3	$5.9014 \cdot 10^{-10}$	$4.4814 \cdot 10^{-9}$
8	0	$8.1019 \cdot 10^{-1}$	$8.1019 \cdot 10^{-1}$
	1	$5.4402 \cdot 10^{-3}$	$5.4403 \cdot 10^{-3}$
	2	$1.1288 \cdot 10^{-6}$	$1.1334 \cdot 10^{-6}$
	3	$7.4144 \cdot 10^{-14}$	$1.4089 \cdot 10^{-9}$
18	0	$5.0502 \cdot 10^{-1}$	$5.0502 \cdot 10^{-1}$
	1	$1.7743 \cdot 10^{-3}$	$1.7743 \cdot 10^{-3}$
	2	$1.0200 \cdot 10^{-7}$	$3.0794 \cdot 10^{-7}$

TABLE 1. Comparison of errors

complement approach is not the Jacobian of G itself but only a reordered submatrix, whose size depends on the size of the active set.

In this table, we have only listed three time steps since the displayed tendencies always remain unchanged. The Schur complement linear system is almost always four times smaller then the full Jacobian linear system (in the number of rows and in the number of columns). Furthermore, its condition number is usually smaller than 3, while the condition number of the full Jacobian is typically more than 1000 times greater. The last two columns show that the full Jacobian method needs much more total GMRES iterations than the Schur complement method except for the first linear system in each time step.

9. Final Remarks

We have investigated and implemented a solution procedure for reactive transport problems including equilibrium mineral precipitation-dissolution reactions. While currently in the geo-scientists' community often strategies which are time consuming [2, 3] or which are of limited practical applicability [13] are used, our intention was to apply modern mathematical strategies to this problem. We avoid operator splitting techniques because of their well-known potential disadvantages. The PDE-ODE-AE-CC system is solved with the semismooth Newton method. We have shown that this semismooth Newton method is typically quadratically convergent, and

time step	iteration	size Schur	size full	cond. Schur	cond. full	Schur GMRes iterations	full GMRes iterations
1	0	9628	42700	2.8497	$3.9813 \cdot 10^3$	4	4
	1	10070	42700	2.9073	$3.9812 \cdot 10^3$	5	85
	2	10116	42700	2.8551	$4.0279 \cdot 10^3$	5	68
	3	10161	42700	2.8551	$3.9812 \cdot 10^3$	5	64
	4	10167		2.8551		5	
2	0	9670	42700	2.8497	$3.9812 \cdot 10^3$	4	4
	1	10142	42700	2.8554	$4.0278 \cdot 10^3$	5	90
	2	10180	42700	2.8554	$3.9860 \cdot 10^3$	5	98
	3	10180	42700	2.8554	$4.0278 \cdot 10^3$	5	70
3	0	9677	42700	2.9272	$4.0273 \cdot 10^3$	4	4
	1	10156	42700	2.9549	$4.0273 \cdot 10^3$	5	85
	2	10200	42700	2.9549	$4.0278 \cdot 10^3$	5	99
	3	10200	42700	2.9549	$4.0276 \cdot 10^3$	5	90

TABLE 2. comparison of the arising linear systems

have confirmed this by our numerical test runs. Compared to other solvers, our implementation keeps the number of unknowns small, first by using the reformulation/decoupling technique of Sec. 3, and second by using a particular Schur complement technique which exploits the special structure of the resulting linear systems of equations.

The geat reduction of the condition number of the Schur complement approach compared to the full system is an interesting observation in our numerical test runs. A theoretical explanation is currently under investigation.

10. Appendix

The following result was used in Section 6. The result itself can be found in [5, p. 60] but without proof. Since we are not aware of an explicit reference containing the proof, we give the details here.

Lemma 2. *Let $B, D \in \mathbb{R}^{n\times n}$ with D being a diagonal matrix, and let $M = D + B$. Then*

$$\det M = \sum_{\alpha \subset I} \det D_{\alpha,\alpha} \cdot \det B_{\bar{\alpha},\bar{\alpha}}\,,$$

where $I := \{1, \ldots, n\}, \bar{\alpha} := I \setminus \alpha$ denotes the complement of $\alpha \subset I$, and where the determinant of a 0×0 matrix is 1.

Proof. The proof is by induction on n.

Let $n = 1$. Then M, B, D are real numbers and the determinant is a linear mapping. Therefore it holds

$$\det M = \det D + \det B = \det D_{\{1\},\{1\}} \cdot \det B_{\emptyset,\emptyset} + \det D_{\emptyset,\emptyset} \cdot \det B_{\{1\},\{1\}} .$$

Now assume the statement holds for all matrices of dimension $n \times n$ and let $B, D \in \mathbb{R}^{(n+1)\times(n+1)}$ with D diagonal and $M := D+B$. Here we need some specific notation. Let $B_i := B_{J,J}$ with $J = \{1, \ldots, n+1\} \setminus \{i\}$. This is the matrix that emerges from B by cancelling the i-th column and row. Let M_i be defined in an analogous way. Furthermore let $D_{\bar{i}} := \operatorname{diag}(\underbrace{0, \ldots, 0}_{i-1}, d_{i+1}, \ldots, d_{n+1})$ be the matrix that evolves from $D = \operatorname{diag}(d_1, d_2, \ldots, d_{n+1})$ by discarding the i-th row and column and setting the first $i-1$ diagonal entries to zero. With d^i and b^i we denote the i-th column of D and B, respectively. Because of the linearity of the determinant in the first column, we then get

$$\begin{aligned} \det M &= \det \left[d^1 + b^1, d^2 + b^2, \ldots, d^{n+1} + b^{n+1}\right] \\ &= \det \left[d^1, d^2 + b^2, \ldots, d^{n+1} + b^{n+1}\right] + \det \left[b^1, d^2 + b^2, \ldots, d^{n+1} + b^{n+1}\right] \\ &= d_1 \cdot \det M_1 + \det \left[b^1, d^2 + b^2, \ldots, d^{n+1} + b^{n+1}\right] , \end{aligned}$$

where the last equation follows by expanding the determinant in the first column. We repeat this procedure and get

$$\begin{aligned} \det M &= d_1 \cdot \det M_1 + \det \left[b^1, d^2 + b^2, \ldots, d^{n+1} + b^{n+1}\right] \\ &= d_1 \cdot \det \left(D_{\bar{1}} + B_1\right) + d_2 \cdot \det \left(D_{\bar{2}} + B_2\right) \\ &\qquad + \det \left[b^1, b^2, d^3 + b^3, \ldots, d^{n+1} + b^{n+1}\right] . \end{aligned}$$

Now we iterate this and eventually get

$$\det M = \sum_{i=1}^{n+1} d_i \cdot \det \left(D_{\bar{i}} + B_i\right) + \det B. \tag{45}$$

Note that $D_{\bar{i}}$ and B_i are $n \times n$ matrices. Hence we can apply the induction hypothesis to obtain

$$\begin{aligned} d_i \cdot \det \left(D_{\bar{i}} + B_i\right) &= d_i \cdot \sum_{\alpha \subseteq \{1,\ldots,n\}} \det \left(D_{\bar{i}}\right)_{\alpha,\alpha} \cdot \det \left(B_i\right)_{\bar{\alpha},\bar{\alpha}} \\ &= d_i \cdot \sum_{\alpha \subseteq \{i,\ldots,n\}} \det \left(D_{\bar{i}}\right)_{\alpha,\alpha} \cdot \det \left(B_i\right)_{\bar{\alpha},\bar{\alpha}} , \end{aligned}$$

where the last equation holds because of the definition of $D_{\bar{i}}$. Now it is not difficult to see that, given any $i \in \{1, \ldots, n+1\}$, we have

$$d_i \cdot \det \left(D_{\bar{i}} + B_i\right) = \sum_{i \in \alpha, \alpha \subset \{i, i+1, \ldots, n+1\}} \det D_{\alpha,\alpha} \cdot \det B_{\bar{\alpha},\bar{\alpha}} ,$$

since $i \in \alpha$ guarantees, on the one hand, that d_i is always on the diagonal of $D_{\alpha,\alpha}$, and, on the other hand, that the index i does not belong to $\bar{\alpha}$ so that we can replace B_i by B. Now we can insert this result in (45) and get

$$\det M = \sum_{i=1}^{n+1} \left[\sum_{i\in\alpha,\alpha\subset\{i,i+1,\ldots,n+1\}} \det D_{\alpha,\alpha} \cdot \det B_{\bar{\alpha},\bar{\alpha}} \right] + \det B\,.$$

Now it holds that $\cup_{i=1}^{n+1} \{i \in \alpha,\, \alpha \subset \{i, i+1, \ldots, n+1\}\}$ equals the power set of $\{1, 2, \ldots, n+1\}$ off the empty set. Furthermore, for different i, two sets $\{i \in \alpha,\, \alpha \subset \{i, i+1, \ldots, n+1\}\}$ do not have an intersection. Therefore α runs through every subset of $\{1, 2, \ldots, n+1\}$ once except for the empty set. But for the empty set, we have

$$\det D_{\emptyset,\emptyset} \cdot \det B_{\bar{\emptyset},\bar{\emptyset}} = \det B\,.$$

Hence we obtain

$$\det M = \sum_{\alpha} \det D_{\alpha,\alpha} \cdot \det B_{\bar{\alpha},\bar{\alpha}},$$

with $\alpha \subset \{1, \ldots, n+1\}$ and $\bar{\alpha} := \{1, \ldots, n+1\} \setminus \alpha$. That is exactly our assertion for $n+1$. □

References

[1] R. Aris and R.H.S. Mah: *Independence of chemical reactions.* Ind. Eng. Chem. Fundam. 2, 1963, pp. 90-94.

[2] C.M. Bethke: *Geochemical reaction modeling, concepts and applications.* Oxford University Press, 1996.

[3] J. Carrayrou, R. Mosé, and P. Behra: *New efficient algorithm for solving thermodynamic chemistry.* AIChE J. 48, 2002, pp. 894–904.

[4] F.H. Clarke: *Optimization and Nonsmooth Analysis.* John Wiley, New York, 1983 (reprinted by SIAM, Philadelphia, 1990).

[5] R.W. Cottle, J.-S. Pang, and R.E. Stone: *The linear complementarity Problem.* Academic Press, Boston, 1992.

[6] F. Facchinei and J.-S. Pang: *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume I.* Springer, New York, NY, 2003.

[7] F. Facchinei and J.-S. Pang: *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume II.* Springer, New York, NY, 2003.

[8] C. Kanzow: *Inexact semismooth Newton methods for large-scale complementary problems.* Optimization Methods and Software 19, 2004, pp. 309–325.

[9] M. Hintermüller, K. Ito, and K. Kunisch: *The primal-dual active set strategy as a semi-smooth Newton method.* SIAM Journal on Optimization 13, 2002, pp. 865–888

[10] S. Kräutle: *General multi-species reactive transport problems in porous media: Efficient numerical approaches and existence of global solutions.* Habilitation Thesis, University of Erlangen, Germany, 2008.

[11] S. Kräutle and P. Knabner: *A new numerical reduction scheme for fully coupled multicomponent transport-reaction problems in porous media.* Water Resour. Res. 41, W09414, doi:10.1029/2004WR003624, 2005.

[12] S. Kräutle and P. Knabner: *A reduction scheme for coupled multicomponent transport-reaction problems in porous media: Generalization to problems with heterogeneous equilibrium reactions*, Water Resour. Res. 43, W03429, doi:10.1029/2005WR004465, 2007.

[13] P.C. Lichtner: *Continuum formulation of multicomponent-multiphase reactive transport,* in: Reviews in Mineralogy, Vol. 34, P.C. Lichtner, C.I. Steefel, and E.H. Oelkers (eds.), Mineralogical Society of America, 1996, pp. 1-88.

[14] J.-S. Pang and L. Qi: *Nonsmooth equations: motivation and algorithms.* SIAM Journal on Optimization 3, 1993, pp. 443–465.

[15] L. Qi: *Convergence analysis of some algorithms for solving nonsmooth equations.* Mathematics of Operations Research 18, 1993, pp. 227–244

[16] L. Qi and J. Sun: *A nonsmooth version of Newton's method.* Mathematical Programming 58, 1993, pp. 353–367.

[17] F. Saaf: *A study of reactive transport phenomena in porous media*, Doctoral Thesis, Rice University, Houston, 1996.

[18] F. Saaf, R.A. Tapia, S. Bryant, and M.F. Wheeler: *Computing general chemical equilibria with an interior-point method*, in: Computational methods in subsurface flow and transport problems, Aldama et al. (eds.), Computational Mechanics Publications, Southampton, U.K., 1996, pp. 201–209.

[19] F. Zhang: *The Schur complement and its applications.* Springer, New York, 2005.