Habilitation Thesis

General Multi-Species Reactive Transport Problems in Porous Media: Efficient Numerical Approaches and Existence of Global Solutions

Serge Kräutle

University of Erlangen–Nuremberg
Department of Mathematics

March 2008

Reviewers:
Prof. Dr. Dr. h.c. mult. Willi Jäger, IWR, Universität Heidelberg
Prof. Dr. Michael Böhm, Universität Bremen
Prof. Dr. Alain Bourgeat, Université de Lyon, France

Mentors of the habilitation procedure
at the University of Erlangen-Nuremberg:
Prof. Dr. Peter Knabner
Prof. Dr. Günter Leugering
Prof. Dr. Heinz Jürgen Tobschall

Contents

Chapter 1

Introduction

In this work multispecies reactive transport problems in porous media are considered. We are regarding a rather general model containing (highly nonlinear) kinetic and equilibrium reactions and mobile and immobile chemical species, including minerals. This problem class covers many applications and plays a central role in the hydrogeologists' and environmental scientists' literature of the last decades.

This work is intended to contribute to this research field in two directions. The first direction is the *development and investigation of efficient numerical techniques* to handle the mathematical model. The second direction is to contribute to the *analysis of the mathematical model* by proving existence and uniqueness of solutions which are global in time.

The model. On a microscopic level, a porous medium consists of a pore space, fully or partially filled with one (or several immiscible) fluid(s), and a solid, the soil matrix. In the fluid, many chemical species (ions, molecules) can be present which may interact with each other, and which also may interact with chemical species (or microbia) which are attached to the soil matrix. While a microscopic view (see for example [25, 65]) is usefull to enlighten and investigate the processes taking place, the numerical simulation of large areas on the field scale (tens, hundreds, thousands of metres) requires the treatment of the porous medium on a macroscopic level. Techniques for the upscaling from the microscopic to the macroscopic scale (or: continuum model) are volume averaging and homogenization [46, 79].

The *macroscopic model* will be the basis of all considerations in this work. It consists of partial differential equations (PDEs) for the 'mobile' species

concentrations (dissolved in the fluid), and ordinary differential equations (ODEs) for 'immobile' species concentrations (attached to the soil matrix). The entities such as the porosity, the water content, the species concentrations, are functions defined on $[0,T]\times\Omega =: Q_T$ where Ω is a spatial domain, and $[0,T]$ is a time interval. The ODEs and PDEs are coupled through (bio)chemical reactions. These reactions can be modelled by kinetic rate laws, which are (often highly nonlinear) functions of the species concentrations. In practice, the time scales on which the reactions take place often differ by many powers of ten, which makes it impossible to resolve the very fast reactions numerically. This leads to a modelling of the fast reactions by equilibrium conditions, which are (nonlinear) algebraic equations (AEs). Hence, the system is a differential-algebraic equation (DAE) consisting of PDEs, ODEs, and AEs. Reactions with minerals usually have to be modelled by equilibrium conditions which consist of AEs combined with inequalities. This extension of the model is considered in the last chapter of this work; our particular treatment of these equilibrium conditions is to regard them as *complementarity conditions* and solve the resulting system by the Semismooth Newton method.
Many applications of the equations under consideration in this work are found in the environmental sciences, such as the fate and prognosis of contaminations in the soil [13, 28, 49, 60, 63, 4] and natural attenuation processes, but also, in a multi-phase setting, the sequestration and storage of carbon dioxide in the subsurface – a field which gains more and more attention among geoscientists – (e.g. [38]), or the compositional modelling of reservoirs.

The intention and the content of this work. As mentioned in the beginning, the intention of this article is twofold:

1. Because of the huge amount of computational resources which are required for 2-D, and even more for 3-D simulations of multi-species reactive transport problems in porous media, efficient numerical techniques for the treatment of the equations are essential. Here, our focus lies on two points:

   (a) We propose a systematic approach for a model-preserving reformulation of the given nonlinear differential or differential-algebraic system such that the resulting formulation is much smaller and better suited for efficient numerical simulation. While operator splitting techniques, which are frequently applied to large reactive

transport problems, may (depending on the specific application) suffer from limitations with respect to time step size, accuracy, or robustness [71, 74, 77, 78], the proposed reformulation technique treats the problem in the sense of a one step method (also called global implicit approach).
The reduction of the problem size is achieved by concentrating the couplings and the nonlinearities in a smaller number of equations (if compared to the original formulation), while some linear equations decouple from the system, and while algebraic constraints are eliminated. The means are linear transformations of the differential equations, linear variable transformation, and elimination of variables by using the algebraic constraints. The difficulty lies in the fact that both mobile and immobile species and both kinetic and equilibrium reactions may be present in the model. State-of-the-art reformulations work pretty well with respect to a decoupling of equations, if all species are mobile (or immobile) or if all reactions are kinetic (or equilibrium); however, they face certain difficulties when the full model is considered. We point out the new proposal in Chapter 2. The main results are published in [50, 51].

(b) The extension of the model to equilibrium reactions (precipitation/dissolution) of minerals leads to a description of the equilibrium state which has to distinguish between a saturated state (where the fluid is saturated with respect to certain ions) and an undersaturated state (where the mineral is fully dissolved) with respect to any mineral. Numerical strategies which are currently proposed and/or applied in the geoscientists' community are (i) combinatorial strategies (requiring some trial-and-error strategy to obtain the correct mineral assemblage) [11], and (ii) the introduction of subdomains of saturation and undersaturation (with respect to each mineral) and the treatment of the problem as a moving boundary problem [54, 55].
In this work a different handling coming from the field of optimization is proposed: It is possible to reformulate the equilibrium conditions as complementarity conditions (CCs). Problems coupled to CCs can be tackled by the Semismooth Newton Method. Such a handling is well known in the optimization community,

but it seems that it has not yet been applied to the mineral problem in reactive transport in this form. The approach is free of Lagrangian multipliers, which are introduced if the equilibrium problem is instead treated as a minimization problem for the so-called Gibbs free energy [69]. We handle the extension of the model, the transformation into CCs, and the application of the Semismooth Newton Method in Chapter 4.
The work is being supported by the DFG (joint DFG project "Efficient numerical methods for large partial differential complementarity systems arising in multispecies reactive transport with minerals in porous media" by P. Knabner, C. Kanzow, S. Kräutle).

The strategies proposed in Chapter 2 were successfully implemented in the course of a diploma thesis by J. Hoffmann [43] in a software running on parallel computers to allow practical tests. The strategies of Chapter 4 were included into the software in the framework of a Ph.D. thesis [44]. Both chapters conclude with some applications.

2. In order to contribute to the *analysis* of the mathematical model, proofs of existence and uniqueness of a global solution, i.e., a solution over arbitrarily large time intervals, are given. We focus on reactions described by the law of mass action. The main difficulty here lies in the high nonlinearity of the reaction rate terms. The proofs are based on fixed point methods (Schaefer's fixed point theorem) and Lyapunov techniques for the derivation of a priori estimates ('method of a priori estimates'). We start by investigating a pure ODE model, then we extend the methods to the PDE model and finally to a coupled PDE-ODE model; see Chapter 3. We conclude the chapter by an extension of some existence und uniqueness results to the DAE case. For this we identify the DAE problem as the limit of a sequence of kinetic problems with reaction rate coefficients going to infinity.

The chapter on the analysis (Ch. 3) on the one hand and the two chapters on the scientific computation aspects (Ch. 2 and the extension in Ch. 4) on the other hand can be read rather independently from each other.

Chapter 2

The reduction method for multicomponent reactive transport problems in porous media

2.1 Introduction

Detailed modelling of reactive transport processes in the underground often requires the consideration of a large number of reactions and reacting species, often combined with the requirement of (locally) fine spatial discretization. A prominent example is natural attenuation, that is the assessment and monitoring of microbially catalysed degradation processes of organic contaminants in the subsurface or aquifer. The reactions often exhibit a wide range of time scales, which advises to model those reactions being much faster than the time scale of the transport processes in a quasistationary manner, i.e., as (algebraically described) equilibrium processes. Additionally, not only mobile species (in solution) appear, but also immobile ones (attached to the porous skeleton). In summary, the resulting system consists of partial differential equations (PDEs), ordinary differential equations (ODEs) and algebraic equations (AEs).

An often used approach to make a numerical solution of these large problems feasible is *operator splitting*, by which transport and reaction become decoupled. This procedure either introduces a further consistency error (in the

non-iterative version, called sequential non-iterative approach, SNIA) which can only be controlled by the time stepping, or applies a fixed point type iteration of unclear convergence properties and time step limitations due to stability problems (sequential iterative approach, SIA).
To circumvent these possible disadvantages of operator splitting, we propose, in the spirit of a global implicit approach (GIA), a model-preserving reformulation of the full given PDE-ODE-AE system in such a way that as many unknowns as possible are determined by linear scalar decoupled transport equations or by local AEs. Our reformulation technique [50, 51, 48] is based on taking linear combinations of the given set of PDEs/ODEs, on the choice of new variables, and on elimination of some of the variables in the sense of a direct substitutional approach (DSA).
The chapter is structured as follows. In Sec. 2.2 the model equations for multispecies reactive transport in porous media with mixed kinetic/equilibrium reactions and mobile/immobile species are given. In Sec. 2.3 some state-of-the-art approaches are considered: Reformulations eliminating the equilibrium reaction rates and treating the system in the sense of a GIA or SNIA, SIA. Limitations of these formulation give the motivation for a new approach, which is described in Sec. 2.4.1-2.4.2, which is the *centrepiece* of Chapter 2. The new approach leads to a formulation with a rigorously reduced size compared to the original system. The first step in deriving this formulation consists of a variable transformation and a linear combination of the PDEs and the ODEs (Sec. 2.4.1) which results in a decoupling of some linear parabolic equations. A second and a third step then remove the unknown rates of the equilibrium reactions from the remaining nonlinear system and substitute the local AEs into the PDEs and ODEs, reducing the system size again (Sec. 2.4.2).
In the remaining Sections 2.4.3-2.4.5, the algorithmic solution by a global and a nested local Newton iteration is described, the existence of a resolution function for the local problem and the structure of the linearized global problem is investigated. In Sec. 2.5 some numerical applications and some remarks on the numerical efficiency of the reformulation are given. Finally, Sec. 2.6 proposes some generalizations of the reduction method to multiphase problems.

2.2 The governing equations

We are considering the nonlinear system

$$\begin{aligned} \partial_t(\theta c_i) + L\, c_i &= \theta \sum_{j=1}^{J} s_{ij} R_j(c_1, ..., c_I, \bar{c}_{I+1}, ..., \bar{c}_{I+\bar{I}}), \quad i = 1, ..., I, \\ \partial_t(\rho \bar{c}_i) &= \theta \sum_{j=1}^{J} s_{ij} R_j(c_1, ..., c_I, \bar{c}_{I+1}, ..., \bar{c}_{I+\bar{I}}), \quad i = I+1, ..., I+\bar{I}, \end{aligned} \tag{2.1}$$

on $[0,T] \times \Omega$, $\Omega \subset \mathbb{R}^2$ or $\mathbb{R}^3$ being a domain, together with appropriate initial and boundary conditions for the unknowns $c = (c_i)_{i=1,...,I}$ being the concentrations of the species dissolved in the groundwater (in moles per fluid volume) and $\bar{c} = (\bar{c}_i)_{i=I+1,...,I+\bar{I}}$ being the surface concentrations of the species attached to the soil matrix (in moles per surface area). $R = (R_1, ..., R_J)$ is a vector of reaction rates for J reactions, and the s_{ij}, $i = 1, ..., I+\bar{I}$, $j = 1, ..., J$, are real numbers (often integers), called stoichiometric coefficients.
The linear differential operator L is required to be independent of $i = 1, ..., I$, i.e., L acts in the same way on all mobile species. Usually, the transport operator L (modeling advection and dispersion) reads

$$Lc_i = -\nabla \cdot (D(q)\, \nabla c_i - c_i\, q) \tag{2.2}$$

where q, depending on space and time, is the Darcy velocity vector of the groundwater flow, and $D(q)$ is the dispersion tensor. Note that the dispersion is species-independent [8]. The assumption of an additional species-*dependent* diffusion in (2.2) would complicate the manipulations to come considerably. However, in the vast majority of applications in the field of hydrology and the geosciences, the diffusion is much smaller than the dispersion, allowing for its neglect in (2.2), (2.1). We assume that q is given or a priori computed by solving the Richards equation [73, 2].
$\theta > 0$ denotes the fraction of the mobile fluid-phase volume, and $\rho > 0$ is the so-called bulk density, relating surface concentrations to volume concentrations. For convenience we set $\theta := 1$ and $\rho := 1$ in the following; however, *all* the transformations and results of Chapter 2 are easily adapted to the case $\theta, \rho \neq 1$.[1]

[1] Just replace $\partial_t u$, where u is any mobile entity, by $\partial_t(\theta u)$, and $\partial \bar{u}$, where $\bar{u}$ is any immobile entity, by $\partial_t(\rho u)$, and insert a factor θ in front of every kinetic reaction term R_j throughout Chapter 2.

In a compact notation the system reads

$$\begin{aligned} \partial_t\, c + Lc &= S_1\, R(c,\bar{c}) \\ \partial_t\, \bar{c} &= S_2\, R(c,\bar{c}) \end{aligned} \tag{2.3}$$

Concerning the reactions, there are two different situations to distinguish: If we assume that the j-th reaction is described by a kinetic rate law, then $R_j = R_j(c,\bar{c})$ is given as a (nonlinear) function of the unknown concentrations $c_i, \bar{c}_i$. 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. If we have for example a chemical reaction

$$X_1 \longleftrightarrow 2X_2 + X_3, \quad \text{or} \quad 0 \longleftrightarrow -X_1 + 2X_2 + X_3, \tag{2.4}$$

we get a column of matrix S with the entries $-1, 2, 1$. An example for a kinetic rate could be the so-called kinetic mass action law for reversible reactions[2]

$$R_j(c,\bar{c}) = k_j^f \prod_{\substack{s_{ij}<0 \\ i=1,\ldots,I+\bar{I}}} c_i^{|s_{ij}|} \;-\; k_j^b \prod_{\substack{s_{ij}>0 \\ i=1,\ldots,I+\bar{I}}} c_i^{s_{ij}} \tag{2.5}$$

with forward and backward rate constant $k_j^f > 0$, $k_j^b > 0$; for (2.4) we get

$$R_j(c,\bar{c}) = k_j^f c_1 - k_j^b c_2^2 c_3.$$

In real-world problems, the timescales for the different reactions may vary over many powers of ten. Hence, besides kinetically controlled reactions, it is frequently assumed that some of the reactions are such fast that a state of local equilibrium can assumed, i.e., its rates can be modelled as equilibrium conditions, i.e., as algebraic equations depending on the participating species concentrations and holding at each point in space and time. Note that the modelling of very fast reactions by equilibrium conditions can reduce the stiffness of the system. A typical equilibrium condition, at least if the species are not minerals (pure solids), is to assume the so-called (equilibrium) law of mass action, which reads[3]

$$\prod_{i=1}^{I} c_i^{s_{ij}} \prod_{i=I+1}^{I+\bar{I}} \bar{c}_i^{s_{ij}} = k_j, \tag{2.6}$$

[2]For the sake of simplicity the bar atop the immobile species $\bar{c}_i$ has been omitted in (2.5).

[3]For a rigorous investigation how the equilibrium law (2.6) relates to the kinetic law (2.5) see Sec. 3.5.

with an equilibrium constant $k_j > 0$. For the example (2.4) this reads $\frac{c_2^2 c_3}{c_1} = k_j$. Concerning implementation issues, there are several equivalent reformulations of (2.6), which may, nevertheless, affect the Newton iteration, such as multiplication by the negative powers, or (assuming the solution to be strictly positive) the following formulation of (2.6) which is linear with respect to the logarithms of the concentrations:

$$Q_j(c, \bar{c}) := \sum_{i=1}^{I} s_{ij} \ln c_i + \sum_{i=I+1}^{I+\bar{I}} s_{ij} \ln \bar{c}_i - \ln k_j = 0 \tag{2.7}$$

From now on we assume that the reactions $j = 1, ..., J_{eq}$ are equilibrium reactions, and the reactions $j = J_{eq+1}, ..., J$ are kinetic reactions, $0 \le J_{eq} \le J$, $J_{kin} := J - J_{eq}$, and for the sake of simplicity we assume that the initial and boundary conditions are fulfilling the equilibrium conditions.
Note that the assumption of local equilibrium does *not* mean that the corresponding rate R_j in (2.1) is zero or infinity or indeterminate, as it is sometimes stated (e.g., [28]): Since the concentrations change in time due to transport and other reactions, R_j is the rate which is necessary to maintain the equilibrium condition, i.e., the R_j, $j = 1, ..., J_{eq}$, are *unknowns* in our system. See the formula (2.14) and the discussion at the end of Sec. 3.5.
The full system reads

$$\begin{aligned} \partial_t c + Lc &= S^1_{eq} R_{eq} + S^1_{kin} R_{kin}(c, \bar{c}) \\ \partial_t \bar{c} &= S^2_{eq} R_{eq} + S^2_{kin} R_{kin}(c, \bar{c}) \end{aligned} \tag{2.8-a}$$

$$Q(c, \bar{c}) := {S^1_{eq}}^T \ln c + {S^2_{eq}}^T \ln \bar{c} - \ln K_{eq} = 0, \tag{2.8-b}$$

for the $I + \bar{I} + J_{eq}$ unknowns $c, \bar{c}, R_{eq}$, where we have used the notation $R_{eq} = (R_1, ..., R_{J_{eq}})$, $R_{kin} = (R_{J_{eq}+1}, ..., R_J)$, $K_{eq} = (k_1, ..., k_{J_{eq}})$, and

$$S = \left(\frac{S_1}{S_2} \right) = \left(\begin{array}{c|c} S^1_{eq} & S^1_{kin} \\ \hline S^2_{eq} & S^2_{kin} \end{array} \right) = \left(S_{eq} \,\middle|\, S_{kin} \right). \tag{2.9}$$

It is well known that any linear dependence of the chemical reactions (i.e., of the columns of S) indicates a redundance of chemical reactions [5]. Hence, without loss of generality we can assume that (at least) S_{eq} has maximal column rank,

$$\operatorname{rank}(S_{eq}) = J_{eq}. \tag{2.10}$$

In order to allow for the variable transformation applied to the system in Sec. 2.3 and Sec. 2.4, we assume that on each part of the boundary of the domain, all mobile species must fulfil the same type (Dirichlet, Neumann, Robin) of boundary condition. Note that the specific rate law (2.5) and equilibrium condition (2.6)/(2.7) serve as an illustration and are not essential for the reformulation technique descriped in this chapter; for example also irreversible kinetic rates or non-ideal activities [11] of species are allowed. Only the investigations of Sec. 2.4.4 and 2.4.5 exploit the structure (2.6)/(2.7).
The system (2.8) (or simplifications/restrictions of it) and possible reformulations are considered by many authors during the past two decades, e.g. [15, 3, 20, 32, 33, 45, 54, 55, 63, 70, 74, 77, 82] due to its relevance in the geohydrologists' and environmental scientists' community.

2.3 State-of-the-art solution methods, motivation for the new decoupling method

In many applications, 10-20 or more chemical species are considered. Hence, after a discretization in space with a finite element or finite volume method and in time with an implicit scheme, the nonlinear discrete problem corresponding to (2.8) is very expensive to solve. Often inhomogenities of the soil or multiple thin soil layers require a fine spatial resolution, leading to a large nonlinear system already in 2-D, and even more in 3-D.
Therefore frequently operator splitting techniques [17, 57, 75, 29] are applied, splitting the computation of a time step into a subproblem consisting of only transport and another subproblem consisting only of the chemistry problem. The advantage is that the solution of the transport subproblem decouples into scalar problems for each species, and the chemistry subproblem decouples into small nonlinear problems for each discretization point. Such a (noniterative) splitting (SNIA) has the drawback that (depending on the specific problem) due to accuracy considerations or stability problems, rather small time steps are required [74, 77]. The problem of accuracy can be overcome by fixed-point-type *iterative* operator splitting techniques SIA [74, 70, 82], where within one time step, transport and chemistry steps are solved alternately. Besides larger costs per time step, if compared to SNIA, the problem of stability and time step limitation remains [71, 78] a severe

restriction.
To avoid the potential disadvantages of the noniterative and the iterative splitting (SNIA/SIA), also one step or globally implicit methods (GIA) are used, since they are considered to be more robust. However, any chance to reduce the number of nonlinearly coupled equations should be exploited to reduce cpu times. One simple possibility to reduce the number of unknowns is to eliminate the unknowns R_{eq}. To do so, one computes a $(I+\bar{I})\times(I+\bar{I}-J_{eq})$ matrix $S_{eq}^{\perp}$ with the property that the columns of $S_{eq}^{\perp}$ are orthogonal to the columns of S_{eq}, i.e.,

$$(S_{eq}^{\perp})^T S_{eq} = 0 \tag{2.11}$$

holds. Equation (2.8-a) can be written as

$$\partial_t \tilde{c} + L\, M\tilde{c} = S_{eq} R_{eq} + S_{kin} R_{kin}(\tilde{c}), \tag{2.12}$$

with M being a diagonal matrix with entries 1 for the mobile and 0 for the immobile species and with $\tilde{c} = (c, \bar{c})$. Multiplication of (2.12) by $(S_{eq}^{\perp})^T$ eliminates the R_{eq}:

$$(S_{eq}^{\perp})^T \left(\partial_t \tilde{c} + LM\tilde{c}\right) = (S_{eq}^{\perp})^T S_{kin} R_{kin}(\tilde{c}), \tag{2.13}$$

and a multiplication of (2.12) by the matrix $(S_{eq}^T S_{eq})^{-1} S_{eq}^T$, i.e., the generalized inverse of S_{eq}^T, leads to the equation

$$R_{eq} = (S_{eq}^T S_{eq})^{-1} S_{eq}^T \left(\partial_t \tilde{c} + L\, M\tilde{c} - S_{kin} R_{kin}(\tilde{c})\right), \tag{2.14}$$

by which R_{eq} can be computed a posteriori[4]. Hence, a decoupling of the computation of $c, \bar{c}$ by (2.13), (2.8-b) and the computation of R_{eq} by (2.14) was achieved. The still coupled system (2.13), (2.8-b) consists of $I + \bar{I} - J_{eq}$ PDEs and J_{eq} AEs. If the AEs (2.8-b) are solved for some J_{eq} components of $\tilde{c}$ (the so-called *secondary species* or local species) and substituted in (2.13) ('DSA'), a nonlinear coupled system of PDEs of size $I + \bar{I} - J_{eq}$ remains ('canonical form', see [54] for the case $J_{kin} = 0$). Note that this substitution leads to coupling terms and nonlinear terms under the transport operator, which is unfavourable concerning (a) the robustness of numerical simulations and (b) the number on nonzero entries in the Jacobian, affecting memory and cpu time requirements.

[4] See also [20] for explicit expressions for the equilibrium rates.

Also the standard formulation which is often denoted by 'Morel tableau' [64, 3, 13] is quite similar: It is assumed that the equilibrium reactions are given by a stoichiometric matrix S_{eq} of the structure

$$S_{eq} = \left(\begin{array}{c|c} C & A \\ \hline -\mathrm{Id} & 0 \\ \hline\hline 0 & B \\ \hline 0 & -\mathrm{Id} \end{array}\right)$$

consisting of J_{mob} equilibrium reactions among mobile species and J_{sorp} equilibrium reactions among mobile and immobile species; $J_{mob} + J_{sorp} = J_{eq}$. Such a structure can always be obtained, if (2.10) holds, by column based Gaussian elimination and renumbering of the mobile and of the immobile species. A matrix consisting of a maximal set of columns fulfilling (2.11) is obviously

$$S_{eq}^{\perp} = \left(\begin{array}{c|c} \mathrm{Id} & 0 \\ \hline C^T & 0 \\ \hline\hline 0 & \mathrm{Id} \\ \hline A^T & B^T \end{array}\right).$$

By multiplying (2.8-a) by $(S_{eq}^{\perp})^T$, the R_{eq} vanishes due to the orthogonality relation (2.11) and one obtains, after a splitting into so-called primary and secondary variables

$$c = \begin{pmatrix} c_{prim} \\ c_{sec} \end{pmatrix}, \quad c = \begin{pmatrix} \bar{c}_{prim} \\ \bar{c}_{sec} \end{pmatrix}$$

the formulation

$$\begin{aligned} \begin{pmatrix} (\partial_t + L)\,(c_{prim} + C c_{sec}) + \partial_t A \bar{c}_{sec} \\ \partial_t\,(\bar{c}_{prim} + B \bar{c}_{sec}) \end{pmatrix} &= (S_{eq}^{\perp})^T S_{kin} R_{kin}(c_{prim}, c_{sec}, \bar{c}_{prim}, \bar{c}_{sec}) \\ &=: f(c_{prim}, c_{sec}, \bar{c}_{prim}, \bar{c}_{sec}) \end{aligned} \tag{2.15}$$

with the equilibrium conditions (2.8-b)

$$\begin{aligned} \ln c_{sec} &= C^T \ln c_{prim} - \ln k_1 \\ \ln \bar{c}_{sec} &= A^T \ln c_{prim} + B^T \ln \bar{c}_{prim} - \ln k_2. \end{aligned} \tag{2.16}$$

After the definition of the mobile fraction $T_M := c_{prim} + C c_{sec}$ and the immobile fractions $T_F := A \bar{c}_{sec}$, $T_S := \bar{c}_{prim} + B \bar{c}_{sec}$, one gets the formulation

$$\begin{aligned} \partial_t (T_M + T_F) + L\, T_M &= f_1(c_{prim}, c_{sec}, \bar{c}_{prim}, \bar{c}_{sec}) \\ \partial_t\, T_S &= f_2(c_{prim}, c_{sec}, \bar{c}_{prim}, \bar{c}_{sec}) \end{aligned} \tag{2.17}$$

together with the 'local' equations

$$\begin{array}{ll} T_M = c_{prim} + C c_{sec}, & \ln c_{sec} = C^T \ln c_{prim} - \ln k_1 \\ T_F = A\bar{c}_{sec}, & \ln \bar{c}_{sec} = A^T \ln c_{prim} + B^T \ln \bar{c}_{prim} - \ln k_2 \\ T_S = \bar{c}_{prim} + B\bar{c}_{sec}; & \end{array} \quad (2.18)$$

a formulation which is used for example in the benchmark [13]. Possible solution strategies may be for example (a) to solve (2.17)-(2.18) directly, or (b) to substitute the secondary variables $c_{sec}, \bar{c}_{sec}$ (2.16) into (2.15), or (c) to solve the large local problem (2.18) for $T_F, c_{prim}, c_{sec}, \bar{c}_{prim}, \bar{c}_{sec}$ and thus eliminate T_F in (2.17), or (d) to use iterative or non-iterative splitting techniques on (2.17)-(2.18). Strategy (a) would lead to a very large Jacobian (even compared to the original formulation (2.8)), making any *efficient* numerical solution questionable. Strategy (b), which is called direct substitutional approach (DSA), leads to a smaller system, but the substitution leads to couplings and nonlinearities in the transport operator. Strategy (c) has, compared to (b), the advantage that the substitution leaves the transport operator linear, only the local problem is larger; it is applied in [3] for $J_{kin} = 0$. However, all the three one-step strategies (a)-(c) for (2.15)-(2.16) and (2.17)-(2.18) lead to $I - J_{mob}$ many nonlinearly coupled PDEs together with $\bar{I} - J_{sorp}$ ODEs (that are, all in all, $I + \bar{I} - J_{eq}$ nonlinearly coupled differential equations).

Of course, a further reduction is desirable. Many different versions of reformulations of the system under consideration were published in the hydrogeologists' community, see the citations at the end of Sec. 2.2. Some of these authors try to find PDEs which are void of any kinetic reaction term (since these are the coupling terms), and to decouple these equations by introduction of new variables in such a way that some linear scalar advection diffusion equations for the new variables decouple from the system. A possible approach to achieve such a decoupling could be to multiply (2.12) by the matrix $(S^\perp)^T$, where $S^\perp$ consists of $I + \bar{I} - J$ columns orthogonal to the columns of S. This cancels both reactive terms R_{eq} and R_{kin} in (2.12):

$$(S^\perp)^T \left(\partial_t \tilde{c} + LM\tilde{c}\right) = 0 \quad (2.19)$$

In the *special case* where all species are mobile ($\bar{I} = 0$), M is the unity matrix, and since $(S^\perp)^T$ commutes with this and with the differential operators, a substitution $\eta := (S^\perp)^T \tilde{c}$ leads to a decoupling of these linear advection diffusion equations

$$\partial_t \eta + L\eta = 0. \quad (2.20)$$

Within each time step, these equations can be solved a priori before the smaller nonlinear problem for the remaining unknowns is tackled. However, all published reformulation/decoupling algorithms lack generality. They all assume certain restrictions on the model, such as all species being mobile or all equilibrium reactions being restricted to only mobile species, or a certain structure of the stoichiometric matrix, in order to obtain a decoupling of equations of type (2.20). In [63], the required assumption is that if there are equilibrium reactions with immobile species, these immobile species must have constant activities.

The purpose of Sec. 2.4 is to propose a new reformulation, which requires no specific restrictions of the generality of the model and where the remaining coupled set of equations is as small as possible. An analysis of the mentioned reformulation reveals that the reason why a decoupling of equations of type (2.20) in general does not take place is the fact that the transformation builds linear combinations of mobile *and* immobile species concentrations, such as $T_M + T_F$ in (2.17) and $(S^\perp)^T M\tilde{c}$ in (2.19). The algorithm of Sec 2.4 will try to avoid such a mixing-up of mobile and immobile entities as far as possible. A *full* avoidance of a combination of mobile and immobile species is, however, not possible, if the removal of R_{eq} from the system is desired; see Sec. 2.4.2.

2.4 The model-preserving a priori decoupling technique

The main difference of the new method compared to those mentioned in the previous section is that the main transformation step (see Sec. 2.4.1) treats the block of PDEs and the block of ODEs of (2.3)/(2.8-a) *separately*. The formulation obtained in Sec. 2.4.1 is only preliminary in the sense that it is only useful for numerical simulation if no equilibrium reactions are present. A general formulation also applicable for mixed kinetic/equilibrium problems is derived in Sec. 2.4.2, based upon the results of Sec. 2.4.1.

2.4.1 The variable transformation and the decoupling of conservation equations

Let J_1, J_2 be the number of linearly independent columns in the matrix S_1, S_2, respectively. For S_i, $i = 1, 2$, we define S_i^* as a matrix consisting of a

maximum system of linearly independent columns of S_i, and A_i such that[5]

$$S_i = S_i^* A_i \tag{2.21}$$

holds. S_1 is a matrix of size $I \times J_1$, S_2 has the size $\bar{I} \times J_2$, A_1 is of the size $J_1 \times J$, and A_2 of size $J_2 \times J$. For S_i^* we define $S_i^\perp$ consisting of a maximum set of linearly independent columns that are orthogonal to each column of S_i^*:

$$(S_i^\perp)^T S_i^* = 0, \quad i = 1, 2; \tag{2.22}$$

$S_1^\perp \in \mathbb{R}^{I \times (I - J_1)}$, $S_2^\perp \in \mathbb{R}^{\bar{I} \times (\bar{I} - J_2)}$. Since S_i^*, $S_i^\perp$ have maximal column rank, the matrices $S_i^{*T} S_i^*$ and $S_i^{\perp T} S_i^\perp$ are regular. By using (2.21) in (2.3)/(2.8-a) and then multiplication of the i-th block, $i = 1, 2$, of (2.3)/(2.8-a) by $(S_i^{*T} S_i^*)^{-1} S_i^{*T}$ and by $(S_i^{\perp T} S_i^\perp)^{-1} S_i^{\perp T}$, we derive the following four blocks:

$$\begin{aligned}
(S_1^{\perp T} S_1^\perp)^{-1} S_1^{\perp T} \, (\partial_t c + Lc) &= 0 \\
(S_2^{\perp T} S_2^\perp)^{-1} S_2^{\perp T} \, \partial_t \bar{c} &= 0 \\
(S_1^{*T} S_1^*)^{-1} S_1^{*T} \, (\partial_t c + Lc) &= A_1 \, R(c, \bar{c}) \\
(S_2^{*T} S_2^*)^{-1} S_2^{*T} \, \partial_t \bar{c} &= A_2 \, R(c, \bar{c})
\end{aligned} \tag{2.23}$$

This manipulation corresponds to the forming of linear combinations within each of the two blocks of equations in (2.3). No linear combinations between mobile and immobile species were formed yet. The number of equations in (2.23) is the same as the number of equations in (2.3) or (2.8-a). Hence, the reduction is still to come.

Since the matrices and the differential operators in system (2.23) commute (which is in general *not* the case in (2.13), (2.19) where mobile and immobile species are combined), we can substitute

$$\begin{aligned}
\eta &= (S_1^{\perp T} S_1^\perp)^{-1} S_1^{\perp T} \, c, \qquad \xi = (S_1^{*T} S_1^*)^{-1} S_1^{*T} \, c, \\
\bar{\eta} &= (S_2^{\perp T} S_2^\perp)^{-1} S_2^{\perp T} \, \bar{c}, \qquad \bar{\xi} = (S_2^{*T} S_2^*)^{-1} S_2^{*T} \, \bar{c}
\end{aligned} \tag{2.24}$$

in (2.23). It is easy to see that the inversion of relation (2.24) reads [32, 67]

$$c = S_1^* \xi + S_1^\perp \eta, \qquad \bar{c} = S_2^* \bar{\xi} + S_2^\perp \bar{\eta}. \tag{2.25}$$

[5]The entries of the matrices A_i have the following meaning: The r-th column of A_i contains the coefficients which are required to express the r-th column of S_i as a linear combination of the columns of S_i^*. An explicit formula for A_i can be obtained from (2.21) by multiplying it by the generalized inverse $(S_i^{*T} S_i^*)^{-1} S_i^{*T}$ of S_i^*.

The vectors $(\xi, \eta) \in \mathbb{R}^I, (\bar{\xi}, \bar{\eta}) \in \mathbb{R}^{\bar{I}}$ are representations of the vectors c, $\bar{c}$, respectively, with respect to another basis. ξ and η are linear combinations of only mobile species, and $\bar{\xi}$, $\bar{\eta}$ are linear combinations of only immobile species. Equations (2.25) represent the decomposition of a vector in $\mathbb{R}^I$ ($\mathbb{R}^{\bar{I}}$, resp.) with respect to two orthogonal subspaces, spanned by the columns of S_i^*, $S_i^\perp$, respectively.
System (2.23) becomes

$$\begin{aligned}
\partial_t \eta + L\eta &= 0 && (I - J_1 \text{ eqs.}) && \text{(2.26-a)}\\
\partial_t \bar{\eta} &= 0 && (\bar{I} - J_2 \text{ eqs.}) && \text{(2.26-b)}\\
\partial_t \xi + L\xi &= A_1 R \begin{pmatrix} S_1^*\xi + S_1^\perp \eta \\ S_2^*\bar{\xi} + S_2^\perp \bar{\eta} \end{pmatrix} && (J_1 \text{ eqs.}) && \text{(2.26-c)}\\
\partial_t \bar{\xi} &= A_2 R \begin{pmatrix} S_1^*\xi + S_1^\perp \eta \\ S_2^*\bar{\xi} + S_2^\perp \bar{\eta} \end{pmatrix} && (J_2 \text{ eqs.}). && \text{(2.26-d)}
\end{aligned}$$

The advantage of (2.26) compared to other possible formulations is that the 'reaction invariants', i.e., the equations for the components η and $\bar{\eta}$, are decoupled from the equations for $\xi, \bar{\xi}$. This is caused by the fact that our components $\eta, \bar{\eta}$ do not consist of mixed linear combinations of mobile and immobile species, as it is usually the case for other formulation (cf. Sec. 2.3). A disadvantage of formulation (2.26) is that, if some of the reactions are equilibrium, these are not yet eliminated from the system, but this will be handled in Sec. 2.4.2. Note that the decoupled equations for η are linear and scalar which makes their solution rather fast. The components $\bar{\eta}$ are even constant throughout the whole simulation. The evolution of both of these quantities is completely determined by the initial and boundary values for $c, \bar{c}$ and is independent of the reactions. In long-term simulations with boundary conditions constant in time, even the η may become stationary, saving even more cpu time. The $\xi, \bar{\xi}$ differ from the classical definition of 'reaction extents' for closed systems [67] and also from the definition in [32, 33] for open systems, since on the right-hand sides of the equations (2.26-c), (2.26-d), linear combinations of the rates instead of single rates occur.
The number of remaining coupled nonlinear equations for representation (2.26) is $J_1 + J_2$. If all reactions are kinetic ($J_{eq} = 0$), then the formulation (2.26) could directly be used for simulation. However, it is possible to further reduce the system size: After discretization in time, (2.26-d) consists of local AEs for $\bar{\xi}$. These equations can be solved for $\bar{\xi}$ (at least if the time

step τ is not too large) and substituted in the right-hand side of (2.26-c). This reduces the number of remaining coupled nonlinear PDEs to J_1.
In the general case where some reactions are at equilibrium ($J_{eq} > 0$), the system (2.26-c)-(2.26-d) would have to be solved together with (2.8-b)[6] for the unknowns $\xi, \bar{\xi}, R_{eq}$. The size and the structure of this problem is not yet satisfying. The following section explains how to eliminate R_{eq} from (2.26).

2.4.2 The general algorithm capable of handling (homogeneous and heterogeneous) equilibrium reactions

If some of the reactions are assumed to be at equilibrium ($J_{eq} > 0$), then formulation (2.26), (2.8-b) should not be used directly. It is reasonable to eliminate the equilibrium rates R_{eq} from as many equations of system (2.26-c)-(2.26-d) as possible and use the remaining R_{eq}-free equations together with the J_{eq} equilibrium conditions (2.8-b). The way how to eliminate the equilibrium reactions from the system (2.26) such that the component equations for $\eta, \bar{\eta}$ remain decoupled will be shown in this section.
Let us sort the vector of equilibrium reactions in the following way: We start with those equilibrium reactions whose participants are all mobile, then we take the heterogeneous equilibrium reactions (i.e., those having participants both in the mobile and in the immobile phase) and at last those equilibrium reactions in the immobile phase:

$$\begin{aligned} &R_{eq} = (R_{mob}, R_{het}, R_{immo}), \\ &R = (R_{eq}, R_{kin}) = (R_{mob}, R_{het}, R_{immo}, R_{kin}) \end{aligned}$$

For the kinetic reactions R_{kin}, no special order is required. The size of the subvectors is $J_{mob}, J_{het}, J_{immo}, J_{kin}$; $J_{mob} + J_{het} + J_{immo} = J_{eq}$, $J_{eq} + J_{kin} = J$. Due to this order, the stoichiometric matrix has the following shape:

$$\begin{aligned} S &= \left(\frac{S_1}{S_2}\right) = \left(\, S_{eq} \,\middle|\, S_{kin} \,\right) = \left(\begin{array}{c|c} S^1_{eq} & S^1_{kin} \\ \hline S^2_{eq} & S^2_{kin} \end{array}\right) \\ &= \left(\begin{array}{c|c|c\|c} S^1_{mob} & S^1_{het} & 0 & S^1_{kin} \\ \hline 0 & S^2_{het} & S^2_{immo} & S^2_{kin} \end{array}\right) \end{aligned} \tag{2.27}$$

[6] with the $c, \bar{c}$ in (2.8-b) expressed in terms of the $\eta, \xi, \bar{\eta}, \bar{\xi}$ via (2.25)

Note that for representation (2.27) we made no assumptions on the stoichiometry; every reactive system has such a representation. For later use let us state that if the equilibrium reactions follow the mass action law (2.8-b), then the structure (2.27) leads to a representation

$$\begin{array}{lcllll} Q_{mob}(c) & := & {S^1_{mob}}^T \ln c & & - \ln K_{mob} & = 0, \\ Q_{het}(c,\bar{c}) & := & {S^1_{het}}^T \ln c & + {S^2_{het}}^T \ln \bar{c} & - \ln K_{het} & = 0, \\ Q_{immo}(\bar{c}) & := & & {S^2_{immo}}^T \ln \bar{c} & - \ln K_{immo} & = 0 \end{array} \tag{2.28}$$

with equilibrium constants $K_{mob} \in \mathbb{R}_+^{J_{mob}}$, $K_{het} \in \mathbb{R}_+^{J_{het}}$, $K_{immo} \in \mathbb{R}_+^{J_{immo}}$; $(K_{mob}, K_{het}, K_{immo}) = K_{eq}$,

$$\mathbb{R}_+ := \{x \in \mathbb{R} \,|\, x > 0\}.$$

For the following we will additionally need that the matrices $(S^1_{mob}|S^1_{het})$ and $(S^2_{het}|S^2_{immo})$ have maximal column rank, i.e.,

$$\operatorname{rank}(S^1_{mob}|S^1_{het}) = J_{mob} + J_{het}, \quad \operatorname{rank}(S^2_{het}|S^2_{immo}) = J_{het} + J_{immo}. \tag{2.29}$$

For an arbitrary matrix S with property (2.10), with the columns arranged as in (2.27), the property (2.29) is usually not met. However, one can show that a Gauss type column based algorithm transforms (2.27) into a matrix of property (2.29) (for a proof see Lemma 5.9 in the appendix; for an example see [51] Sec. 4). It is well known that the replacement of columns in the stoichiometric matrix by nontrivial linear combinations of other columns does not change the chemistry; for example building the sum of two columns corresponds to summing up two chemical equations and to the multiplication of the corresponding equilibrium conditions (2.6).
Hence, without loss of generality we assume that S has the structure (2.27) and the property (2.29). As a consequence of (2.29), we can choose a maximum set of linearly independent columns from S_1 containing $(S^1_{mob}|S^1_{het})$, and a maximum set of linearly independent columns from S_2 containing $(S^2_{het}|S^2_{immo})$. As in Sec. 2.4.1, we denote these matrices by S^*_1, S^*_2 again. We get the block structure

$$\begin{array}{lcl} S^*_1 & = & (S^1_{mob} \,|\, S^1_{het} \,|\, S^{1'}_{kin}), \\ S^*_2 & = & (S^2_{het} \,|\, S^2_{immo} \,|\, S^{2'}_{kin}), \end{array} \tag{2.30}$$

where $S^{1'}_{kin}, S^{2'}_{kin}$ consist of columns taken from S^1_{kin}, S^2_{kin}, respectively, such that S^*_1, S^*_2 consist of a maximal set of linear independent columns of S_1, S_2.

Let $J^{1'}_{kin}, J^{2'}_{kin}$ be the number of columns of $S^{1'}_{kin}, S^{2'}_{kin}$. Then for J_1, J_2 (defined in the beginning of Sec 2.4.1) the equations

$$J_1 = J_{mob} + J_{het} + J^{1'}_{kin}, \quad J_2 = J_{het} + J_{immo} + J^{2'}_{kin} \tag{2.31}$$

hold.
The selection process is described by (2.21), where A_1, A_2 have, thanks to (2.29) and (2.30) and possibly a reordering of the columns of S^1_{kin} and S^2_{kin}, the block structure

$$A_1 = \left(\begin{array}{c|c|c|c} \mathrm{Id}_{J_{mob}} & 0 & 0 & A^1_{mob} \\ \hline 0 & \mathrm{Id}_{J_{het}} & 0 & A^1_{het} \\ \hline 0 & 0 & 0 & A^1_{kin} \end{array}\right), \tag{2.32}$$

$$A_2 = \left(\begin{array}{c|c|c|c} 0 & \mathrm{Id}_{J_{het}} & 0 & A^2_{het} \\ \hline 0 & 0 & \mathrm{Id}_{J_{immo}} & A^2_{immo} \\ \hline 0 & 0 & 0 & A^2_{kin} \end{array}\right), \tag{2.33}$$

and where Id_n is the $n \times n$ identity matrix. Exactly as in (2.21)-(2.26) we define $S^\perp_1, S^\perp_2$, multiply the two blocks of the given system (2.8-a) by the matrices $(S^{*T}_i S^*_i)^{-1} S^{*T}_i$ and by $(S^{\perp T}_i S^\perp_i)^{-1} S^{\perp T}_i$, and substitute $c, \bar{c}$ by the new variables $\eta, \xi, \bar{\eta}, \bar{\xi}$. Additionally we make use of the block structure (2.32), (2.33) and we introduce the splitting into subvectors

$$\xi = (\xi_{mob}, \xi_{het}, \xi_{kin}), \quad \overline{\xi} = (\overline{\xi}_{het}, \overline{\xi}_{immo}, \overline{\xi}_{kin})$$

of size $J_{mob}, J_{het}, J^{1'}_{kin}$; $J_{het}, J_{immo}, J^{2'}_{kin}$. We get:

$$\begin{aligned}
(\partial_t + L)\,\eta &= 0 && \text{(2.34-a)}\\
\partial_t \quad \overline{\eta} &= 0 && \text{(2.34-b)}\\
(\partial_t + L)\,\xi_{mob} &= R_{mob} + A^1_{mob}\, R_{kin}(c, \overline{c}) && \text{(2.34-c)}\\
(\partial_t + L)\,\xi_{het} &= R_{het} + A^1_{het}\, R_{kin}(c, \overline{c}) && \text{(2.34-d)}\\
(\partial_t + L)\,\xi_{kin} &= A^1_{kin}\, R_{kin}(c, \overline{c}) && \text{(2.34-e)}\\
\partial_t \quad \overline{\xi}_{het} &= R_{het} + A^2_{het}\, R_{kin}(c, \overline{c}) && \text{(2.34-f)}\\
\partial_t \quad \overline{\xi}_{immo} &= R_{immo} + A^2_{immo} R_{kin}(c, \overline{c}) && \text{(2.34-g)}\\
\partial_t \quad \overline{\xi}_{kin} &= A^2_{kin}\, R_{kin}(c, \overline{c}) && \text{(2.34-h)}
\end{aligned}$$

This formulation corresponds completely to the representation (2.26); the only difference is that we used the block structure of $S, A_1, A_2, \xi, \bar{\xi}$.

To prepare the elimination of the equilibrium reaction rates $R_{mob}, R_{het}, R_{immo}$ from (2.34), we replace (2.34-d) by the difference of (2.34-d) and (2.34-f)[7] and obtain

$$\begin{aligned}
(\partial_t + L)\,\eta &= 0 && \text{(2.35-a)}\\
\partial_t \quad \overline{\eta} &= 0 && \text{(2.35-b)}\\
(\partial_t + L)\,\xi_{mob} &= R_{mob} + A^1_{mob}\, R_{kin}(c,\overline{c}) && \text{(2.35-c)}\\
\partial_t(\xi_{het} - \bar{\xi}_{het}) + L\,\xi_{het} &= A_{het}\, R_{kin}(c,\overline{c}) && \text{(2.35-d)}\\
(\partial_t + L)\,\xi_{kin} &= A^1_{kin}\, R_{kin}(c,\overline{c}) && \text{(2.35-e)}\\
\partial_t \quad \overline{\xi}_{het} &= R_{het} + A^2_{het}\, R_{kin}(c,\overline{c}) && \text{(2.35-f)}\\
\partial_t \quad \overline{\xi}_{immo} &= R_{immo} + A^2_{immo}\, R_{kin}(c,\overline{c}) && \text{(2.35-g)}\\
\partial_t \quad \overline{\xi}_{kin} &= A^2_{kin}\, R_{kin}(c,\overline{c}), && \text{(2.35-h)}
\end{aligned}$$

where we have set $A_{het} := A^1_{het} - A^2_{het}$. Now, we can consider the equations (2.35-c), (2.35-f), (2.35-g) as explicit equations for the equilibrium reactions $(R_{mob}, R_{het}, R_{immo}) = R_{eq}$; they are eliminated from all the other blocks of (2.35). We can drop these J_{eq} equations together with the J_{eq} unknowns R_{eq}, or use these equations for an a posteriori reconstruction of R_{eq},[8] and consider the rest of (2.35) together with (2.8-b) as equations for $\eta, \bar{\eta}, \xi, \bar{\xi}$:

$$\begin{aligned}
\partial_t\,\eta + L\,\eta &= 0 && (I - J_1 \text{ eqs.}) && \text{(2.36-a)}\\
\partial_t \quad \overline{\eta} &= 0 && (\bar{I} - J_2 \text{ eqs.}) && \text{(2.36-b)}\\
\partial_t(\xi_{het} - \bar{\xi}_{het}) + L\,\xi_{het} &= A_{het} R_{kin}(c,\overline{c}) && (J_{het} \text{ eqs.}) && \text{(2.36-c)}\\
\partial_t\,\xi_{kin} + L\,\xi_{kin} &= A^1_{kin}\, R_{kin}(c,\overline{c}) && (J^{1'}_{kin} \text{ eqs.}) && \text{(2.36-d)}\\
\partial_t\,\overline{\xi}_{kin} &= A^2_{kin} R_{kin}(c,\bar{c}) && (J^{2'}_{kin} \text{ eqs.}) && \text{(2.36-e)}\\
Q(c,\bar{c}) &= 0 && (J_{eq} \text{ eqs.}) && \text{(2.36-f)}
\end{aligned}$$

where $c, \bar{c}$ is short for (see (2.25), (2.27))

$$\begin{aligned}
c &= S^*_1\xi + S^\perp_1\eta\\
&= S^1_{mob}\xi_{mob} + S^1_{het}\xi_{het} + S^{1'}_{kin}\xi_{kin} + S^\perp_1\eta,\\
\overline{c} &= S^*_2\overline{\xi} + S^\perp_2\overline{\eta}\\
&= S^2_{het}\overline{\xi}_{het} + S^2_{immo}\overline{\xi}_{immo} + S^{2'}_{kin}\overline{\xi}_{kin} + S^\perp_2\overline{\eta}.
\end{aligned} \qquad (2.37)$$

[7] This is the only step where a combination of mobile and immobile entities takes place.
[8] such as (2.14)

The system (2.36) consists of the decoupled linear problems (2.36-a) for η, a block of 'local' problems (2.36-e), (2.36-f), and a system of PDEs (2.36-c), (2.36-d). As the next section shows, the blocks of local equations (2.36-e), (2.36-f) can (after discretization of (2.36-e) in time) be solved for the unknowns $\xi_{loc} := (\xi_{mob}, \bar{\xi}_{het}, \bar{\xi}_{immo}, \bar{\xi}_{kin})$, and substituted into the remaining PDEs (2.36-c)-(2.36-d) in the sense of a DSA, and a system with the structure

$$\begin{aligned} \partial_t(\xi_{het} - f_1(\xi_{het}, \xi_{kin})) + L\,\xi_{het} &= f_2(\xi_{het}, \xi_{kin}) \\ \partial_t\ \xi_{kin} \qquad\qquad\quad + L\,\xi_{kin} &= f_3(\xi_{het}, \xi_{kin}) \end{aligned} \tag{2.38}$$

for the 'global' variables $\xi_{glob} := (\xi_{het}, \xi_{kin})$ results. Note that (different from the scheme arising from other methods of GIA-DSA type, see Sec. 2.3), the introduction of the resolution functions for $\xi_{mob}, \bar{\xi}_{het}, \bar{\xi}_{immo}, \bar{\xi}_{kin}$ in (2.36-c), (2.36-d) does not take place under the transport operator, but only under the time derivative and the right-hand sides. Hence, the sparsity pattern for the linear problems arising from the discretization of (2.38) is more convenient for efficient numerical solution with iterative linear solvers, and the transport operator stays linear with respect to the unknowns.
After the local equations (2.36-e), (2.36-f) are substituted into the PDEs (2.36-c), (2.36-d), the size of remaining coupled nonlinear equations (2.38) is $J_{het} + J^{1'}_{kin}$. Because of (2.31), $J_{het} + J^{1'}_{kin} = J_1 - J_{mob}$ holds. Since J_1 is the rank of matrix S_1, and I is the number of rows of S_1, $J_1 \leq I$ holds, i.e., the number of nonlinearly coupled PDEs is smaller than or equal to $I - J_{mob}$, which is the number of coupled PDEs when using classical formulations such as in Sec. 2.3 (not counting the additional $\bar{I} - J_{sorp}$ ODEs appearing in the classical formulations).

2.4.3 Solution of the resulting nonlinear problem by nested Newton iteration

As already indicated, we tackle our problem (2.36-c)-(2.36-f), which can be written in the abstract formulation

$$f_{glob}(\xi_{glob}, \xi_{loc}) = 0 \tag{2.39-a}$$

$$f_{loc}(\xi_{glob}, \xi_{loc}) = 0, \tag{2.39-b}$$

in the form (2.38), i.e.,

$$f_{glob}(\xi_{glob}, \xi_{loc}(\xi_{glob})) = 0; \tag{2.40}$$

the ξ_{loc} and the ξ_{glob} can be regarded as the secondary and the primary variables, respectively. For the decomposition of (2.36-c)-(2.36-f) into f_{glob} and f_{loc}, and, correspondingly, the decomposition of $(\xi, \bar{\xi})$ into ξ_{glob}, ξ_{loc}, there are two possibilities obvious: One is to choose

$$\xi_{glob} := (\xi_{het}, \xi_{kin}, \bar{\xi}_{kin}), \quad \xi_{loc} := (\xi_{mob}, \bar{\xi}_{het}, \bar{\xi}_{immo}) \tag{2.41}$$

with the global problem (2.39-a) corresponding to (2.36-c)-(2.36-e) and the local problem (2.39-b) corresponding to the AEs (2.36-f). The other is, as already indicated in Sec. 2.4.2,

$$\xi_{glob} := (\xi_{het}, \xi_{kin}), \quad \xi_{loc} := (\xi_{mob}, \bar{\xi}_{het}, \bar{\xi}_{immo}, \bar{\xi}_{kin}) \tag{2.42}$$

with the global problem (2.39-a) corresponding to (2.36-c)-(2.36-d) and the local problem (2.39-b) corresponding to the AEs and the ODEs (2.36-e)-(2.36-f).

In both cases the resulting nonlinear problem (2.40) is solved by Newton's method or a modified Newton (Newton-Armijo) iteration. Note that into the Jacobian of (2.40), the partial derivatives of the *local* equations obviously enter:

$$J_{glob} = \frac{\partial f_{glob}}{\partial \xi_{glob}} - \frac{\partial f_{glob}}{\partial \xi_{loc}} \left(\frac{\partial f_{loc}}{\partial \xi_{loc}} \right)^{-1} \frac{\partial f_{loc}}{\partial \xi_{glob}} \tag{2.43}$$

However, a Newton step for (2.40) only gives an update for the *global* variables ξ_{glob}. The update for the local variables is, since the resolution function $\xi_{loc}(\xi_{glob})$ is not explicitely known, computed by a *nested Newton iteration* for (2.39-b) with fixed ξ_{glob}. Note that this nested Newton iteration consists of small local problems at each discretization point or control volume being decoupled from each other. The Jacobian of the local problem simply reads

$$J_{loc} = \frac{\partial f_{loc}}{\partial \xi_{loc}}.$$

Computational tests [43, 50] showed that the amount of cpu time for these local problems is negligible compared to the solution of the global problem.

In order to reduce the cpu time for assembling the global problem, it turned out that it is reasonable to omit all terms $O(\tau^2)$ entering the global matrix through the substitution of the time-discretized ODEs (2.36-e) into the global equations.

Note that, if $J'^{2}_{kin} = 0$ (which *not* necessarily requires $J_{kin} = 0$), then the two cases (2.41), (2.42) coincide, since the ODE block (2.36-e) then is non-existent.

2.4.4 Existence of the resolution function

Local resolution functions

The resolution function $\xi_{loc} = \xi_{loc}(\xi_{glob})$ due to (2.41) or (2.42) cannot be given explicitely. However, one can prove (see Lemma 2.1) that a local resolution function exists around each point $p^0 = (\eta^0, \bar{\eta}^0, \xi^0_{glob}, \xi^0_{loc}) \in P$, where P is the set of all points for which the corresponding[9] vector $(c^0, \bar{c}^0)$ is strictly positive, i.e., $c^0_i > 0$, $\bar{c}^0_i > 0$ for all i. In this section we require that the equilibrium reactions are given by the mass action relation (2.7)/(2.28). For the proof of existence by the implicit function theorem, the regularity of the Jacobian $\partial f_{loc}/\partial \xi_{loc}$ has to be checked. Note that this regularity also implies the solvability of the local Newton step.

Lemma 2.1 *Let (2.28) hold. For each point $p^0 = (\eta^0, \bar{\eta}^0, \xi^0_{glob}, \xi^0_{loc}) \in P$, there is a local resolution function in the sense of (2.41). If the ODE (2.36-e) is discretized with the implicit Euler method*[10] *then there is a local resolution function in the sense of (2.42) if the time step size is sufficiently small.*

Proof. Using (2.28), (2.37), one gets

$$J_{loc} = \frac{\partial Q}{\partial(\xi_{mob}, \bar{\xi}_{het}, \bar{\xi}_{immo})} = \left(\begin{array}{c|c|c} {S^1_{eq}}^T \Lambda S^1_{mob} & {S^2_{eq}}^T \bar{\Lambda}\, S^2_{het} & {S^2_{eq}}^T \bar{\Lambda}\, S^2_{immo} \end{array} \right)$$

$$= \left(\begin{array}{c|c|c} {S^1_{mob}}^T \Lambda\, S^1_{mob} & 0 & 0 \\ \hline {S^1_{het}}^T \Lambda\, S^1_{mob} & {S^2_{het}}^T \bar{\Lambda}\, S^2_{het} & {S^2_{het}}^T \bar{\Lambda}\, S^2_{immo} \\ \hline 0 & {S^2_{immo}}^T \bar{\Lambda}\, S^2_{het} & {S^2_{immo}}^T \bar{\Lambda}\, S^2_{immo} \end{array} \right) \tag{2.44}$$

$$= \left(\begin{array}{c|c} {S^1_{mob}}^T \Lambda\, S^1_{mob} & 0 \\ \hline * & (S^2_{het}|S^2_{immo})^T \bar{\Lambda}\, (S^2_{het}|S^2_{immo}) \end{array} \right) \tag{2.45}$$

with the diagonal matrices $\Lambda = \mathrm{diag}\left(\frac{1}{c_i}\right) \in \mathbb{R}^{I\times I}$, $\bar{\Lambda} = \mathrm{diag}\left(\frac{1}{\bar{c}_i}\right) \in \mathbb{R}^{\bar{I}\times\bar{I}}$. Using (2.29), the two blocks on the diagonal are symmetric positive definite; hence the whole matrix (2.45) is regular, and the local resolution function of (2.36-f) with respect to $\xi_{loc} = (\xi_{mob}, \bar{\xi}_{het}, \bar{\xi}_{immo})$ exists. For the existence of the local resolution function for (2.42) we have to check the regularity of the Jacobian

$$\tilde{J}_{loc} = \frac{\partial(Q, Q_{kin})}{\partial(\xi_{mob}, \bar{\xi}_{het}, \bar{\xi}_{immo}, \bar{\xi}_{kin})}$$

[9] by virtue of the variable transform (2.25)

[10] Similar results should hold for other time stepping schemes.

with $Q_{kin} := \bar{\xi}_{kin} - \bar{\xi}_{kin}^{old} - \tau A_{kin}^2 R_{kin}(c, \bar{c})$; $\bar{\xi}_{kin}^{old}$ from the old time step. We obtain

$$\tilde{J}_{loc} = \left(\begin{array}{c|c} J_{loc} & \frac{\partial Q}{\partial \bar{\xi}_{kin}} \\ \hline \frac{\partial Q_{kin}}{\partial(\xi_{mob}, \bar{\xi}_{het}, \bar{\xi}_{immo})} & \frac{\partial Q_{kin}}{\partial \bar{\xi}_{kin}} \end{array} \right) = \left(\begin{array}{c|c} J_{loc} & {S_{eq}^2}^T \bar{\Lambda} \, S_{kin}^2 \\ \hline O(\tau) & \mathrm{Id}_{J_{kin}^{2'}} - O(\tau) \end{array} \right)$$

This matrix is regular for $\tau = 0$ and, using a continuity argument, also for sufficiently small τ.[11] ∎

Global resolution functions

It is well known that in general, the existence of a local resolution function at each point of a domain does *not* imply the existence of a global resolution function,[12] i.e., the existence and uniqueness of a ξ_{loc} for a given ξ_{glob} (and given $\eta, \bar{\eta}$), such that (2.39-b) holds, is not obvious. However, if we consider the local problem consisting only of the equilibrium conditions, i.e., (2.41), then we can prove the existence and uniqueness of a $\xi_{loc} = \xi_{loc}(\xi_{glob})$.
We proceed as follows: We show that the solution ξ_{loc} of the local problem (2.39-b) for given ξ_{glob} is equivalent to the solution of the following couple of minimization problems. These minimization problems are strictly convex and lower semicontinuous and the levelsets are bounded; hence, the solution exists and is unique. The minization problems are based on functionals which are closely related to the so-called Gibbs free energy, which will be frequently used in Chapter 3. Note that the use of certain convex minimization problems for proving existence of equilibria for chemical kinetic mass action ODE problems also appears in Feinberg's reactive network theory [30], and for example in [69] Sec. 2.5.4 and the papers cited there.
For $c \in \overline{\mathbb{R}}_+^I$, let us define the 'energy functional'

$$G_{mob}(c) := \sum_{i=1}^{I} \mu_i(c_i)\, c_i + \exp(1 - \mu_i^0), \quad \mu_i(c_i) := \mu_i^0 - 1 + \ln c_i,$$

where the constant vector $\mu_i^0 \in \mathbb{R}^I$ is defined as a solution of the linear system

$${S_{mob}^1}^T \mu^0 = -\ln K_{mob}, \tag{2.46}$$

[11] The proof does not state a *uniform* τ with respect to the point $p^0 \in P$. However, for certain reactive networks, such as S_{eq}^2 being a zero matrix and S_{kin}^2 being a diagonal matrix, and R_{kin} complies with the mass action law, one can show the existence even for arbitrary large $\tau > 0$ independent of $p^0 \in P$.

[12] See [41] Sec. 171 or [66] for counter-examples.

$K_{mob} \in \mathbb{R}_+^{J_{mob}}$ being the vector of equilibrium constants for the reactions R_{mob}. For $\bar{c} \in \overline{\mathbb{R}}_+^{\bar{I}}$ and fixed $c \in \overline{\mathbb{R}}_+^{I}$, let us define the energy functional

$$G_{het,immo}(\bar{c}) := \sum_{i=I+1}^{\bar{I}} \bar{\mu}_i(\bar{c}_i)\, \bar{c}_i + \exp(1 - \bar{\mu}_i^0), \quad \bar{\mu}_i(\bar{c}_i) := \bar{\mu}_i^0 - 1 + \ln \bar{c}_i$$

where $\bar{\mu}_i^0 \in \mathbb{R}^{\bar{I}}$ is defined as a solution of the linear system

$$(S_{het}^2 | S_{immo}^2)^T \bar{\mu}^0 = -\ln \begin{pmatrix} K_{het} \\ K_{immo} \end{pmatrix} + \begin{pmatrix} {S_{het}^1}^T \ln c \\ 0 \end{pmatrix}, \tag{2.47}$$

$K_{het} \in \mathbb{R}_+^{J_{het}}$, $K_{immo} \in \mathbb{R}_+^{J_{immo}}$ being the vector of equilibrium constants for the reactions R_{het}, R_{immo}. The existence (not the uniqueness) of solutions μ^0 and $\bar{\mu}^0$ of the linear systems (2.46) and (2.47) follows from the fact that the columns of S_{mob}^1 and of $(S_{het}^2 | S_{immo}^2)$ are linearly independend (cf. the rank conditions (2.29))[13]. For later use we state that the estimates

$$G_{mob}(c) \geq \sum_{i=1}^{I} c_i =: \|c\|_1, \quad G_{het,immo}(\bar{c}) \geq \sum_{i=I+1}^{\bar{I}} \bar{c}_i =: \|\bar{c}\|_1 \tag{2.48}$$

for all $c \in \overline{\mathbb{R}}_+^{I}$, $\bar{c} \in \overline{\mathbb{R}}_+^{\bar{I}}$ hold; for a justification see (3.12), (3.15)-(3.16) and Footnote 4 on page 71.

Via (2.37), G_{mob} can be considered as a function of the mobile entities $\eta, \xi_{mob}, \xi_{het}, \xi_{kin}$, and $G_{het,immo}$, for fixed $\eta, \xi_{mob}, \xi_{het}, \xi_{kin}$, as a function of the immobile entities $\bar{\eta}, \bar{\xi}_{het}, \bar{\xi}_{immo}, \bar{\xi}_{kin}$. For given $\eta, \bar{\eta}$ and $\xi_{glob} = (\xi_{het}, \xi_{kin}, \bar{\xi}_{kin})$, we want to find $\xi_{loc} = (\xi_{mob}, \bar{\xi}_{het}, \bar{\xi}_{immo})$ by the following minimization problems. We are going to show that this problem is equivalent to the local problem (2.39-b)/(2.36-f):

(a) Let $\eta \in \mathbb{R}^{I-J_1}$, $\xi_{het} \in \mathbb{R}^{J_{het}}$, $\xi_{kin} \in \mathbb{R}^{J_{kin}^{1'}}$ be given such that $M := \{\xi_{mob} \in \mathbb{R}^{J_{mob}} \mid c(\eta, \xi_{mob}, \xi_{het}, \xi_{kin}) \in \mathbb{R}_+^I\} \neq \emptyset$.

Find $\xi_{mob} \in M$ solving

$$G_{mob}(c(\eta, \xi_{mob}, \xi_{het}, \xi_{kin})) \longrightarrow \min_{\xi_{mob} \in M}.$$

(b) Let $c \in \mathbb{R}_+^I$, $\bar{\eta} \in \mathbb{R}^{\bar{I}-J_2}$, $\bar{\xi}_{kin} \in \mathbb{R}^{J_{kin}^{2'}}$ be given such that $\bar{M} := \{(\bar{\xi}_{het}, \bar{\xi}_{immo}) \in \mathbb{R}^{J_{het}+J_{immo}} \mid \bar{c}(\bar{\eta}, \bar{\xi}_{het}, \bar{\xi}_{immo}, \bar{\xi}_{kin}) \in \mathbb{R}_+^{\bar{I}}\} \neq \emptyset$.

[13] for example choose $\mu^0 := -S_{mob}^1 ({S_{mob}^1}^T S_{mob}^1)^{-1} \ln K_{mob}$

Find $(\bar{\xi}_{het}, \bar{\xi}_{immo}) \in \bar{M}$ solving

$$G_{het,immo}(\bar{c}(\bar{\eta}, \bar{\xi}_{het}, \bar{\xi}_{immo}, \bar{\xi}_{kin})) \longrightarrow \min_{(\bar{\xi}_{het}, \bar{\xi}_{immo}) \in \bar{M}}.$$

Lemma 2.2 *There is a unique solution to the previous problems (a), (b), and this solution is the only solution of the local problem (2.39-b) $Q = 0$.*

Proof.
(i) Derivatives of $\mathbf{G_{mob}}, \mathbf{G_{het,immo}}$. The gradients of the energy functionals are

$$\nabla_c G_{mob}(c) = \mu(c), \quad \nabla_{\bar{c}} G_{het,immo}(\bar{c}) = \bar{\mu}(\bar{c}).$$

With (2.37) and the chain rule we obtain

$$\begin{aligned} \frac{\partial G_{mob}}{\partial \xi_{mob}} &= \left[\frac{\partial c}{\partial \xi_{mob}}\right]^T \nabla_c G_{mob}(c) = {S^1_{mob}}^T(\mu^0 + \ln c) \\ &= -\ln K_{mob} + {S^1_{mob}}^T \ln c = Q_{mob}(c). \end{aligned}$$

Analogously,

$$\begin{aligned} \frac{\partial G_{het,immo}}{\partial(\bar{\xi}_{het}, \bar{\xi}_{immo})} &= \left[\frac{\partial \bar{c}}{\partial(\bar{\xi}_{het}, \bar{\xi}_{immo})}\right]^T \nabla_{\bar{c}} G_{het,immo}(\bar{c}) \\ &= (S^2_{het}|S^2_{immo})^T(\bar{\mu}^0 + \ln \bar{c}) \\ &= -\ln \begin{pmatrix} K_{het} \\ K_{immo} \end{pmatrix} + \begin{pmatrix} {S^1_{het}}^T \ln c \\ 0 \end{pmatrix} + \begin{pmatrix} {S^2_{het}}^T \ln \bar{c} \\ {S^2_{immo}}^T \ln \bar{c} \end{pmatrix} \\ &= \begin{pmatrix} Q_{het}(c, \bar{c}) \\ Q_{immo}(\bar{c}) \end{pmatrix}. \end{aligned}$$

(ii) Uniqueness. Both mappings

$$\begin{aligned} &\xi_{mob} \longmapsto G_{mob}(c(\eta, \xi_{mob}, \xi_{het}, \xi_{kin})), \\ &(\bar{\xi}_{het}, \bar{\xi}_{immo}) \longmapsto G_{het,immo}(\bar{c}(\bar{\eta}, \bar{\xi}_{het}, \bar{\xi}_{immo}, \bar{\xi}_{kin})) \end{aligned} \tag{2.49}$$

are strictly convex on M, $\bar{M}$, respectivly, since

$$\frac{\partial^2 G_{mob}}{\partial \xi^2_{mob}} = {S^1_{mob}}^T \Lambda \, S^1_{mob},$$

$$\frac{\partial^2 G_{het,immo}}{\partial \bar{\xi}^2_{loc}} = (S^2_{het}|S^2_{immo})^T \bar{\Lambda} \, (S^2_{het}|S^2_{immo}),$$

where $\Lambda, \bar{\Lambda}$ are positive definite diagonal matrices (defined in the proof of Lemma 2.1), and where S^1_{mob} and $(S^2_{het}|S^2_{immo})$ have maximal column rank, i.e., both Hessians are positive definite. We conclude that both functionals are strictly convex on the convex domain of definition M, $\bar{M}$, respectively. Hence, any possible solution of (a), (b) is unique.

(iii) Existence. Extending the mappings (2.49) by the value $+\infty$ on $\mathbb{R}^{J_{mob}} \backslash M$ and $\mathbb{R}^{J_{het}+J_{immo}} \backslash \bar{M}$, respectively, both functionals are lower semicontinuous with values in $\mathbb{R} \cup \{\infty\}$. For lower semicontinuous funcionals $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, the levelsets $L_k := \{x \in \mathbb{R}^n \mid f(x) \leq k\}$ are closed sets.[14] From the estimate (2.48) and the estimates

$$\|\xi_{mob}\| \leq \text{const}\, \|c\|, \quad \|(\bar{\xi}_{het}, \bar{\xi}_{immo})\| \leq \text{const}\, \|\bar{c}\|$$

(which are obvious due to the linear relation (2.24)) we obtain that the levelsets are also bounded. Hence, they are compact. As a consequence we obtain the existence of solutions of (a), (b).[14]

(iv) Correspondence of minimization problem and local problem. Both mappings (2.49), extended by $+\infty$ outside M, $\bar{M}$, respectively, are convex. Hence, they attain their minimum (which exists and is unique due to (ii), (iii)) if and only if zero is an element of their subdifferential. Due to (i), this is the case if and only if $(\xi_{mob}, \bar{\xi}_{het}, \bar{\xi}_{immo})$ is a solution of the local problem (2.36-f). ∎

Remark. Note the reason why we considered two problems (a) and (b) instead of the *single* minimization problem

$$G(c, \bar{c}) \longrightarrow \min_{\xi_{mob}, \bar{\xi}_{het}, \bar{\xi}_{immo}}$$

where G is defined by

$$G(c) := \sum_{i=1}^{I} \mu_i(c_i)\, c_i + \sum_{i=I+1}^{I+\bar{I}} \mu_i(\bar{c}_i)\, \bar{c}_i + \sum_{i=1}^{I+\bar{I}} \exp(1 - \mu_i^0),$$

$$\mu_i(c_i) := \mu_i^0 - 1 + \ln c_i, \quad \mu_i(\bar{c}_i) := \mu_i^0 - 1 + \ln \bar{c}_i,$$

and where $\mu_i^0 \in \mathbb{R}^{I+\bar{I}}$ is, let us say, defined such that

$$S_{eq}^T \mu^0 = -\ln K_{eq}$$

[14] See for example the textbook [34].

holds: For *this* minimization problem,

$$\frac{\partial G}{\partial(\xi_{mob}, \bar{\xi}_{het}, \bar{\xi}_{immo})} = \tilde{S}_{eq}^T \left(\mu^0 + \begin{pmatrix} \ln c \\ \ln \bar{c} \end{pmatrix} \right) \tag{2.50}$$

with

$$\tilde{S}_{eq} := \left(\begin{array}{c|c|c} S_{mob}^1 & 0 & 0 \\ \hline 0 & S_{het}^2 & S_{immo}^2 \end{array} \right) \neq S_{eq}$$

holds; the proof that (2.50) is equal to $Q(c, \bar{c})$ would require that μ^0 is both a solution of $\tilde{S}_{eq}^T \mu^0 = -\ln K_{eq}$ *and* a solution of $S_{eq}^T \mu^0 = -\ln K_{eq}$, which is problematic.

2.4.5 The structure of the global problem

Questions may arise if the global problem in the formulation (2.38) could be degenerated, or, regarded on a linearized and discretized level, if each Newton step for (2.38) is well defined in the sense that the system matrix (2.43) is regular. For illustration, let us imagine that the resolution function of the local problem simply was $\bar{\xi}_{het}(\xi_{het}, \xi_{kin}) = \xi_{het}$ [or $\bar{\xi}_{het}(\xi_{het}, \xi_{kin}) = 2\xi_{het}$]. If this was possible, the PDE system was no longer parabolic [forward in time], and the part of the Jacobian which comes from the accumulation term, which is

$$\begin{pmatrix} \xi_{het} - \bar{\xi}_{het}(\xi_{het}, \xi_{kin}) \\ \xi_{kin} \end{pmatrix} \tag{2.51}$$

in situation (2.42) and

$$\begin{pmatrix} \xi_{het} - \bar{\xi}_{het}(\xi_{het}, \xi_{kin}, \bar{\xi}_{kin}) \\ \xi_{kin} \\ \bar{\xi}_{kin} \end{pmatrix} \tag{2.52}$$

in situation (2.41), was not positive definite.
To tackle this complex of questions, one first result *on the Jacobian of the accumulation term* can be given (Lemma 2.3). If we consider the advection-dispersion and the kinetic reactions as a *perturbation* of (2.51)/(2.52), we may derive some results on the regularity of the Jacobian of the full problem (2.38), provided that the perturbation is sufficiently small. Another benefit from the investigation of the Jacobian of (2.51)/(2.52) is that it gives some insight into the influence of the resolution function on the *condition number* of the global problem. Note that the investigation of the Jacobian of (2.51)/(2.52) is a *local* consideration, though problem (2.38)/(2.40) is global.

Lemma 2.3 *Let all $c_i, \bar{c}_i$ be strictly positive. Let us solve the local problem (2.36-f) in the sense of (2.41) for the local variables $\xi_{loc} = (\xi_{mob}, \bar{\xi}_{het}, \bar{\xi}_{immo})$.*

(a) Then the derivative of the resolution function $\bar{\xi}_{het}(\xi_{het}, \xi_{kin}, \bar{\xi}_{kin})$ with respect to ξ_{het} has the shape

$$\frac{\partial \bar{\xi}_{het}}{\partial \xi_{het}} = -AB$$

where both $A, B \in \mathbb{R}^{J_{het} \times J_{het}}$ are symmetric positive definite. As a consequence, the product AB is similar to a diagonal matrix with all diagonal entries being strictly positive (i.e., there is a complete set of J_{het} eigenvectors for AB with all J_{het} eigenvalues being positive, and the matrix $\partial \bar{\xi}_{het} / \partial \xi_{het}$ is regular).

(b) As a consequence, the derivative of (2.52) with respect to $\xi_{glob} = (\xi_{het}, \xi_{kin}, \bar{\xi}_{kin})$ has the structure

$$\begin{aligned} \frac{d}{d\,\xi_{glob}} \begin{pmatrix} \xi_{het} - \bar{\xi}_{het}(\xi_{glob}) \\ \xi_{kin} \\ \bar{\xi}_{kin} \end{pmatrix} &= \left(\begin{array}{c|c} Id_{J_{het}} + AB & -\frac{\partial \bar{\xi}_{het}}{\partial (\xi_{kin}, \bar{\xi}_{kin})} \\ \hline 0 & Id_{J^{1'}_{kin} + J^{2'}_{kin}} \end{array} \right) \\ &= Id_{J_{het} + J^{1'}_{kin} + J^{2'}_{kin}} + \left(\begin{array}{c|c} AB & * \\ \hline 0 & 0 \end{array} \right) \end{aligned} \tag{2.53}$$

Thus, matrix (2.53) has a complete set of $J_{het} + J^{1'}_{kin} + J^{2'}_{kin}$ eigenvectors with all eigenvalues being greater than or equal to 1, and therefore is regular.

Proof. **(a)** Let us denote the resolution function $\xi_{glob} \longmapsto \xi_{loc}(\xi_{glob})$ by f. Then the local problem reads

$$Q(f(\xi_{het}, \xi_{kin}, \bar{\xi}_{kin}), \xi_{het}, \xi_{kin}, \bar{\xi}_{kin}) = 0.$$

By differentiation of this equation we obtain

$$\frac{\partial Q}{\partial (\xi_{mob}, \overline{\xi}_{het}, \overline{\xi}_{immo})} \, \frac{\partial (\xi_{mob}, \overline{\xi}_{het}, \overline{\xi}_{immo})}{\partial (\xi_{het}, \xi_{kin}, \bar{\xi}_{kin})} + \frac{\partial Q}{\partial (\xi_{het}, \xi_{kin}, \bar{\xi}_{kin})} = 0.$$

Since matrix $\frac{\partial Q}{\partial(\xi_{mob},\overline{\xi}_{het},\overline{\xi}_{immo})} = J_{loc}$ is known to be regular (cf., (2.45)) we get

$$\begin{aligned}\frac{\partial(\xi_{mob},\overline{\xi}_{het},\overline{\xi}_{immo})}{\partial\xi_{het}} &= -\left(\frac{\partial Q}{\partial(\xi_{mob},\overline{\xi}_{het},\overline{\xi}_{immo})}\right)^{-1}\frac{\partial Q}{\partial\xi_{het}} \\ &= -\begin{pmatrix} M_{11} & 0 \\ M_{21} & M_{22}\end{pmatrix}^{-1}\begin{pmatrix} {S^1_{mob}}^T \\ \hline\hline {S^1_{het}}^T \\ \hline 0\end{pmatrix}\Lambda S^1_{het} \end{aligned} \tag{2.54}$$

with the blocks M_{ij} defined through (2.45). As already stated in the proof of Lemma 2.1, $M_{11} = {S^1_{mob}}^T\Lambda S^1_{mob}$ and $M_{22} = (S^2_{het}|S^2_{immo})^T\bar{\Lambda}(S^2_{het}|S^2_{immo})$ are regular; hence (cf. (5.11)-(5.14))

$$\begin{pmatrix} M_{11} & 0 \\ M_{21} & M_{22}\end{pmatrix}^{-1} = \begin{pmatrix} M_{11}^{-1} & 0 \\ -M_{22}^{-1}M_{21}M_{11}^{-1} & M_{22}^{-1}\end{pmatrix} \tag{2.55}$$

holds. From (2.54), (2.55) we extract by dropping those rows of (2.54) which are related to ξ_{mob}:

$$\begin{aligned}\frac{\partial(\overline{\xi}_{het},\overline{\xi}_{immo})}{\partial\xi_{het}} &= -\left[-M_{22}^{-1}M_{21}M_{11}^{-1}\,{S^1_{mob}}^T\Lambda S^1_{het} + M_{22}^{-1}\begin{pmatrix} {S^1_{het}}^T \\ 0\end{pmatrix}\Lambda S^1_{het}\right] \\ &= -M_{22}^{-1}\left[-\begin{pmatrix} {S^1_{het}}^T\Lambda S^1_{mob} \\ 0\end{pmatrix}M_{11}^{-1}\,{S^1_{mob}}^T\Lambda S^1_{het} + \begin{pmatrix} {S^1_{het}}^T \\ 0\end{pmatrix}\Lambda S^1_{het}\right] \\ &= -M_{22}^{-1}\begin{pmatrix} -{S^1_{het}}^T\Lambda S^1_{mob}({S^1_{mob}}^T\Lambda S^1_{mob})^{-1}\,{S^1_{mob}}^T\Lambda S^1_{het} + {S^1_{het}}^T\Lambda S^1_{het} \\ 0\end{pmatrix}\end{aligned} \tag{2.56}$$

If we denote the blocks in M_{22}^{-1} by

$$M_{22}^{-1} = \begin{pmatrix} M'_{11} & M'_{12} \\ M'_{21} & M'_{22}\end{pmatrix}, \tag{2.57}$$

we can extract from (2.56) by dropping the rows related to $\bar{\xi}_{immo}$

$$-\frac{\partial\overline{\xi}_{het}}{\partial\xi_{het}} = M'_{11}\left[-{S^1_{het}}^T\Lambda S^1_{mob}({S^1_{mob}}^T\Lambda S^1_{mob})^{-1}\,{S^1_{mob}}^T\Lambda S^1_{het} + {S^1_{het}}^T\Lambda S^1_{het}\right]. \tag{2.58}$$

It remains to verify that the two factors on the right-hand side are symmetric positive definite. The matrix

$$\begin{aligned} M_{22} &= \left(\begin{array}{c|c} {S^2_{het}}^T \bar{\Lambda} S^2_{het} & {S^2_{het}}^T \bar{\Lambda} S^2_{immo} \\ \hline {S^2_{immo}}^T \bar{\Lambda} S^2_{het} & {S^2_{immo}}^T \bar{\Lambda} S^2_{immo} \end{array}\right) \\ &= (S^2_{het}|S^2_{immo})^T \bar{\Lambda} (S^2_{het}|S^2_{immo}) \end{aligned} \tag{2.59}$$

is symmetric positive definite, as already stated before. Hence, also the inverse matrix M_{22}^{-1} is symmetric positive definite. From this we can conclude that the subblock M'_{11} of M_{22}^{-1} is also symmetric positive definite.
Let us denote the term in brackets in (2.58) by B. We will prove that B is symmetric positive definite:
We observe that the matrix

$$\begin{aligned} H &:= \left(\begin{array}{c|c} {S^1_{het}}^T \Lambda S^1_{het} & {S^1_{het}}^T \Lambda S^1_{mob} \\ \hline {S^1_{mob}}^T \Lambda S^1_{het} & {S^1_{mob}}^T \Lambda S^1_{mob} \end{array}\right) \\ &= (S^1_{het}|S^1_{mob})^T \Lambda (S^1_{het}|S^1_{mob}) \end{aligned} \tag{2.60}$$

is symmetric and positive definite, since the columns of $(S^1_{het}|S^1_{mob})$ are linearly independent. Therefore also H^{-1} is symmetric positive definite. Computing $H^{-1} = \begin{pmatrix} H'_{11} & H'_{12} \\ H'_{21} & H'_{22} \end{pmatrix}$ by the formula (5.11)-(5.12), also H'_{11} and $(H'_{11})^{-1}$ are symmetric positive definite. We obtain that the matrix $(H'_{11})^{-1}$, computed with the first line of (5.12), with H_{ij} from (2.60), is equal to B. Therefore B is symmetric positive definite.
Hence, we found a representation

$$-\frac{\partial \bar{\xi}_{het}}{\partial \xi_{het}} = M'_{11} B$$

where both M'_{11} and B are symmetric positive definite. By Lemma 5.10 it follows that the product has a complete system of eigenvectors with positive eigenvalues.
(b) The matrix

$$\left(\begin{array}{c|c} AB & * \\ \hline 0 & 0 \end{array}\right) \in \mathbb{R}^{(J_{het}+J^{1'}_{kin}+J^{2'}_{kin})\times(J_{het}+J^{1'}_{kin}+J^{2'}_{kin})} \tag{2.61}$$

obviously has the same J_{het} positive eigenvalues as the matrix $AB \in \mathbb{R}^{J_{het}\times J_{het}}$ and a kernel of dimension $J^{1'}_{kin}+J^{2'}_{kin}$, i.e., a basis of eigenvectors with all eigenvalues being nonnegative. ∎

We can derive the following conclusions from Lemma 2.3:

1. The matrices depend on the concentrations c_i, $\bar{c}_i$ in terms of the diagonal matrices $\Lambda = \mathrm{diag}(1/c_i)$, $\bar{\Lambda} = \mathrm{diag}(1/\bar{c}_i)$. If there are strictly positive upper *and* lower bounds for the c_i, $\bar{c}_i$, then the condition numbers stay bounded. However, not only for concentrations going to infinity,[15] but also for concentrations going to zero, the boundedness of the condition numbers is not obvious. Indeed, numerical application of the reduction method to the artificial and very challenging benchmark problem [13][16] confirmed this difficulty. The analysis of Lemma 2.3 may be used as a starting point for the development of variants of the reduction scheme for those problems for which zero concentrations occur. It turns out that a version where another variable $\tilde{\xi}_{het} = \xi_{het} - \bar{\xi}_{het}$ is introduced may reduce the condition number for such problems. However, for problems with concentrations being bounded from below by positive constants, the original algorithm of this Chapter 2.4 is favorable and should be used in order to avoid any unnecessary increase of the problem size by additional variables.

2. The Lemma 2.3 may be used as a starting point for some investigations on the regularity of the whole Jacobian J_{glob}. If we focus on the advective-free case where L represents an elliptic differential operator, it is still not obvious that J_{glob} is always regular, since unfortunately it is of course in general *not* true that *the sum* of two matrices, each being similar to a positive diagonal matrix, is regular. At least under some quite restrictive assumptions, some preliminary ad hoc statements on the regularity of J_{glob} can be made, such as:

   (a) Assume that for a given concentration field the time step size $\tau > 0$ is sufficiently small. Since J_{glob} is a perturbation of (2.52), then J_{glob} is regular.

   (b) Assume that the concentrations are *constant in space* and L (discretized) is symmetric positive definite. Then the eigenvectors of J_{glob} minus the mass matrix are cartesian products of eigenvectors

[15] Note that under reasonable assumptions unbounded concentrations cannot occur, since in the framework of Chapter 3, boundedness of solutions in $C([0,T]\times\overline{\Omega})$ is proven.

[16] This benchmark has the specific difficulty that locally extremely small values of concentrations are imposed by zero initial/boundary conditions for some species.

of matrix (2.61) and eigenvectors of L. Adding the mass matrix one obtains that J_{glob} has a full system of eigenvectors with all eigenvalues being bounded from below by a positive constant.

3. Formula (2.58) reveals that the second factor, denoted by B, can be written as

$$\begin{aligned} & {S^1_{het}}^T \Lambda S^1_{het} - {S^1_{het}}^T \Lambda S^1_{mob} ({S^1_{mob}}^T \Lambda S^1_{mob})^{-1} {S^1_{mob}}^T \Lambda S^1_{het} \\ &= {S^1_{het}}^T \Lambda^{1/2} \left[\mathrm{Id}_I - \Lambda^{1/2} S^1_{mob} ({S^1_{mob}}^T \Lambda S^1_{mob})^{-1} {S^1_{mob}}^T \Lambda^{1/2} \right] \Lambda^{1/2} S^1_{het} \\ &= {S^1_{het}}^T \Lambda^{1/2} \, P_{(\Lambda^{1/2} S^1_{mob})^\perp} \, \Lambda^{1/2} S^1_{het} \end{aligned}$$

where $P_{(\Lambda^{1/2} S^1_{mob})^\perp} : \mathbb{R}^I \longrightarrow \mathbb{R}^I$ is the orthogonal projection onto the orthogonal complement of the space spanned by the columns of matrix $\Lambda^{1/2} S^1_{mob}$.
The first factor M'_{11} or A in (2.58) can be written as

$$\begin{aligned} M'_{11} &= \left[{S^2_{het}}^T \bar{\Lambda} S^2_{het} - {S^2_{het}}^T \bar{\Lambda} S^2_{immo} ({S^2_{immo}}^T \bar{\Lambda} S^2_{immo})^{-1} {S^2_{immo}}^T \bar{\Lambda} S^2_{het} \right]^{-1} \\ &= \left[{S^2_{het}}^T \bar{\Lambda}^{1/2} \left(\mathrm{Id}_{\bar{I}} - \bar{\Lambda}^{1/2} S^2_{immo} ({S^2_{immo}}^T \bar{\Lambda} S^2_{immo})^{-1} {S^2_{immo}}^T \bar{\Lambda}^{1/2} \right) \bar{\Lambda}^{1/2} S^2_{het} \right]^{-1} \\ &= \left[{S^2_{het}}^T \bar{\Lambda}^{1/2} \, P_{(\bar{\Lambda}^{1/2} S^2_{immo})^\perp} \, \bar{\Lambda}^{1/2} S^2_{het} \right]^{-1}, \end{aligned} \tag{2.62}$$

where we have used (5.11)-(5.12) and (2.57), (2.59). The operator $P_{(\bar{\Lambda}^{1/2} S^2_{immo})^\perp} : \mathbb{R}^{\bar{I}} \longrightarrow \mathbb{R}^{\bar{I}}$ is the orthogonal projection onto the complement of the space spanned by the columns of the matrix $\bar{\Lambda}^{1/2} S^2_{immo}$. Note that despite the projection, the matrix M'_{11} is regular, what has been proven before.[17] We obtain

$$\begin{aligned} & -\frac{\partial \overline{\xi}_{het}}{\partial \xi_{het}} = \\ & \left[(\bar{\Lambda}^{1/2} S^2_{het})^T \, P_{(\bar{\Lambda}^{1/2} S^2_{immo})^\perp} \, \bar{\Lambda}^{1/2} S^2_{het} \right]^{-1} (\Lambda^{1/2} S^1_{het})^T \Lambda^{1/2} \, P_{(\Lambda^{1/2} S^1_{mob})^\perp} \, \Lambda^{1/2} S^1_{het}. \end{aligned}$$

[17] The (already proven) regularity of $M := {S^2_{het}}^T \bar{\Lambda}^{1/2} \, P_{(\bar{\Lambda}^{1/2} S^2_{immo})^\perp} \, \bar{\Lambda}^{1/2} S^2_{het}$ can be recovered in formulation (2.62): The regularity is related to the linear independence of the columns of $(S^2_{het} | S^2_{immo})$: Assume that there is an $x \neq 0$ in the kernel of M. Then $\langle Mx, x \rangle = 0$, $\langle P_{(\bar{\Lambda}^{1/2} S^2_{immo})^\perp} y, P_{(\bar{\Lambda}^{1/2} S^2_{immo})^\perp} y \rangle = 0$ with $y := \bar{\Lambda}^{1/2} S^2_{het} x$ holds. Hence, $y \in \mathrm{range}(\bar{\Lambda}^{1/2} S^2_{immo})$ and $P_{(\bar{\Lambda}^{1/2} S^2_{immo})^\perp} y = 0$. Hence, $y \in \mathrm{range}(\bar{\Lambda}^{1/2} S^2_{het}) \cap \mathrm{range}(\bar{\Lambda}^{1/2} S^2_{immo})$. The linear independence results in $y = 0$, from which $x = 0$ follows again by linear independence.

2.5 Implementation and numerical results

2.5.1 Implementation

The algorithm of Sec. 2.4 was implemented for 2-D computations by J. Hoffmann [43] using the programming environment M++ [80]. M++ is a finite element tool, designed to run on parallel computers, with a lean data structure, based on C++. The structure is modular; in particular, solvers, preconditioners, FE ansatz spaces, time-stepping schemes, etc. can be exchanged by simple switches in a script file.

The linear solver used for the following applications is GMRES(k), the nonlinear solver is Newton-Armijo. Parallelization is done automatically. The partitioning of the computational domain among the CPUs is done based on the coordinates of the barycenters of the elements. Mass lumping is applied to the reactive terms and to the accumulation terms, and, correspondingly, the ODEs and AEs are evaluated pointwise.[18]

For the PDEs of the transport-reaction problem, standard linear finite elements on triangles (P_1), bilinear elements on quadrilaterals (Q_1), and second order finite elements on triangles are implemented. In order to maintain compatibility to mass lumping, not *standard* second order elements were implemented, but the elements proposed in [19], i.e., the polynomial space P_2 extended by the 'bubble' function. This is necessary since the quadrature rule that would normally be applied for standard P_2 elements on triangles has zero weights, leading to unstable schemes with mass lumping.

The flow field can be user-defined, or it can be computed using the Richards' equation. For the solution of the Richards' equation, mixed finite elements (BDM0 and RT1[19]) are used in order to guarantee local mass conservation of the flow field.

The whole chemistry, the flow parameters, the ansatz space, discretization parameters, stopping criteria etc. are defined by the user in a script file.

For comparational reasons (cf. Tables 2.2, 2.3, 2.5), also a 'naive' approach based on the formulation (2.3), where equilibrium reactions are treated as kinetic reactions with large rate coefficients, was implemented with the same

[18] The pointwise evaluation of the ODEs and AEs seems reasonable since otherwise the decoupling of the local problems fails. The mass lumping for the reaction terms in the PDEs is reasonable in order to match the evaluation of the corresponding terms in the ODEs.

[19] Brezzi–Douglas–Marini, Raviart–Thomas

numerical kernel.

2.5.2 Numerical examples, efficiency

The efficiency of the transformation/size reduction technique can be estimated in two different ways:

1. By counting the number of remaining coupled nonlinear PDEs or

2. by cpu measurements.

Concerning the number of remaining nonlinearly coupled PDEs, it was already shown at the end of Sec. 2.4.2 that the number is always less than or equal to the number of coupled PDEs for classical formulations (while classical formulations also have additional ODEs which are not eliminated). In order to quantify the specific amount of the reduction of the problem size, four examples of reactive transport problems, mostly taken from the literature, are analysed in the following.
Example 1 is the biodegradation chain of an organic pollutant. Two kinetic biodegradation reactions, one aerobic, one anaerobic, are considered ([43], Sec. 6.2). Example 2 is the biodegradation of the pollutant EDTA considered in [28], Sec. 6.1. Two kinetic microbially catalyzed decay reactions, five equilibrium sorption reactions, and three further kinetic reactions are considered here. Example 3 is the hydrolysis of a trivalent metal in water, where the equilibrium between the metal ions and several mobile complexes and a kinetic sorption reaction are considered (stoichiometry as in the example [76] p.267). Example 4 is the benchmark problem [13].
The following table demonstrates the size reduction for these scenarios. For example, Scenario 2 contains $I = 6$ mobile and $\bar{I} = 8$ immobile species and $J_{eq} = 5$ equilibrium and $J_{kin} = 5$ kinetic reactions ($J_{het} = J_{mob} = 0$). The number of unknown species concentrations is $6 + 8$; the number of unknowns in the formulation (2.8) is $I+\bar{I}+J_{eq}=6+8+5$. For the canonical formulation, $I+\bar{I}-J_{eq}=9$ PDEs/ODEs remain (from which $I-J_{mob}=6$ are PDEs), and the number of remaining coupled PDEs with the newly proposed reformulation is $J_{het}+J_{kin}^{1'}=6$. In all the examples, the number of remaining coupled PDEs is significantly smaller than for the initial formulation (2.8) and for standard formulations (Sec. 2.3); see Table 2.1. Example 4 demonstrates that also in the presence of equilibrium sorption reactions, a true reduction of the

Scenario	mob./immob. species	equ. reactions	kin. reactions	orig. size PDEs+ODEs +AEs	size std. form. (PDEs+ODEs)	size for new approach (PDEs)
Ex. 1	5+2= 7	0	2	5+2+0	5+0=5	2
Ex. 2	6+8=14	5	5	6+8+5	6+3=9	6
Ex. 3	9+1=10	6	1	9+1+6	3+0=3	1
Ex. 4	9+3=12	7	0	9+3+7	4+1=5	2

Table 2.1: Overview over the reduction in terms of the number of equations.

problem size can be achieved, while e.g. the method [63] has to exclude such problems.
See also Sec. 5 in [51] for a methodological comparison of the new reduction scheme to classical formulations [54] and to recently developed methods [63].
For the simulations, a 36 processor Linux cluster was available. However, the CPU time measurements (Tables 2.2, 2.3, 2.5) refer to computations on *one* CPU, Q_1-elements are used, and a *given* flow field.

2.5.3 Example 1: A kinetic biodegradation chain/redox zoning

We consider the decay of a pollutant A by two biological reactions, one aerobic and one anaerobic:

$$\begin{aligned} \mathrm{A} + \mathrm{O_2} &\xrightarrow{X_1} \mathrm{B} \\ \mathrm{B} + \mathrm{NO_3^-} &\xrightarrow{X_2} \mathrm{C} \quad \text{(inhib. by } \mathrm{O_2}\text{)} \end{aligned} \tag{2.63}$$

The two reactions are catalyzed by two different populations of microbia, $\mathrm{X_1}$ and $\mathrm{X_2}$. We get the following system of differential equations:

$$\begin{aligned} \mathrm{A}: &\quad \partial_t c_1 + L\, c_1 = -R_1 \\ \mathrm{O_2}: &\quad \partial_t c_2 + L\, c_2 = -R_1 \\ \mathrm{B}: &\quad \partial_t c_3 + L\, c_3 = +R_1 - R_2 \\ \mathrm{NO_3^-}: &\quad \partial_t c_4 + L\, c_4 = -R_2 \\ \mathrm{C}: &\quad \partial_t c_5 + L\, c_5 = +R_2 \\ \mathrm{X_1}: &\quad \partial_t \bar{c}_1 = R_3 \\ \mathrm{X_2}: &\quad \partial_t \bar{c}_2 = R_4 \end{aligned} \tag{2.64}$$

For the kinetic rates we assume the Monod model [60, 7] with inhibition terms, with biomass growth limitations and death rates to hold:

$$\begin{aligned} R_1(c_1, c_2, \bar{c}_1) &= \mu_1 \tfrac{c_1}{\alpha_1+c_1} \tfrac{c_2}{\alpha_2+c_2} \bar{c}_1 \\ R_2(c_2, c_3, c_4, \bar{c}_2) &= \mu_2 \tfrac{c_3}{\alpha_3+c_3} \tfrac{c_4}{\alpha_4+c_4} \bar{c}_2 \tfrac{\alpha_5}{\alpha_5+c_2} \\ R_3(c_1, c_2, \bar{c}_1) &= R_1 \cdot (1 - \bar{c}_1/\bar{c}_{1,max}) - d_1 \bar{c}_1 \\ R_4(c_2, c_3, c_4, \bar{c}_2) &= R_2 \cdot (1 - \bar{c}_2/\bar{c}_{2,max}) - d_2 \bar{c}_2 \end{aligned}$$

with given rate parameters $\mu_1 = \mu_2 = 1$, death rates $d_1 = d_2 = 0.1$, biomass growth limitations $\bar{c}_{1,max} = \bar{c}_{2,max} = 1$, and Monod parameters $\alpha_1 = ... = \alpha_5 = 10^{-3}$. The last factor in R_2 is an inhibition term. It models that the the anaerob reaction is suppressed in the presence of O_2.
The dispersion tensor is, following Bear-Scheidegger,

$$D(q) = \beta_t \, |q|_2 \, \mathrm{Id}_2 + (\beta_l - \beta_t) \, \frac{qq^T}{|q|_2}$$

with longitudinal and transversal dispersion length $\beta_l = 0.10$, $\beta_t = 0.02$. The domain is $(0, 10) \times (0, 6)$, the time step size is $\tau = 0.2$, the total time $T = 40$. At the left boundary ($x = 0$) the plume of width 2 enters with the groundwater flow. We assume a concentration of pollutant A of 1 where the plume enters and zero otherwise. We assume that oxygen and nitrate are already exhausted in the plume when it enters the domain, so we prescribe an oxygen and nitrate concentration of zero where the plume enters and 1 elsewhere:

$$c_1 = \begin{cases} 1, & \text{on } \{0\} \times [2, 4] \\ 0, & \text{on } \{0\} \times ((0, 6) \setminus [2, 4]) \end{cases}, \quad c_2 = c_4 = \begin{cases} 0, & \text{on } \{0\} \times [2, 4] \\ 1, & \text{on } \{0\} \times ((0, 6) \setminus [2, 4]) \end{cases}$$

All the other mobile concentrations are set to zero on the inflow boundary:

$$c_3 = c_5 = 0 \quad \text{on } \{0\} \times [0, 6]$$

For the discretisation $2145, 8385, 33155$ elements (Q1) were used.

The stoichiometric matrix arising from (2.64) reads

$$S = \begin{pmatrix} S^1_{kin} \\ \hline S^2_{kin} \end{pmatrix} = \left(\begin{array}{cccc} -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right)$$

The entities required for the transformation are

$$S_1^* = \begin{pmatrix} -1 & 0 \\ -1 & 0 \\ 1 & -1 \\ 0 & -1 \\ 0 & 1 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad S_2^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A_2 = S_2,$$

$I = 5,\ \bar{I} = 2,\ J = J_{kin} = 4,\ J_{eq} = J_{mob} = J_{het} = J_{immo} = 0;$ the rank of S_1 and S_2 is $J_1 = 2,\ J_2 = 2$, respectively. Hence, the size reduction technique leads to $I - J_1 = 3$ decoupled scalar η-equations, $\bar{I} - J_2 = 0$ $\bar{\eta}$-equations, $J_{kin}^{1'} = J_1 - J_{mob} - J_{het} = 2$ ξ_{kin}-equations, and $J_{kin}^{2'} = J_2 - J_{het} - J_{immo} = 2$ $\bar{\xi}_{kin}$-equations:

$$\begin{array}{lcl} \partial_t \eta_1 + L\, c_1 & = & 0 \\ \partial_t \eta_2 + L\, c_2 & = & 0 \\ \partial_t \eta_3 + L\, c_3 & = & 0 \\ \hline \partial_t \xi_{kin,1} + L\, \xi_{kin,1} & = & R_1 \\ \partial_t \xi_{kin,2} + L\, \xi_{kin,2} & = & R_2 \\ \hline \partial_t \bar{\xi}_{kin,1} & = & R_1 \\ \partial_t \bar{\xi}_{kin,2} & = & R_2 \end{array}$$

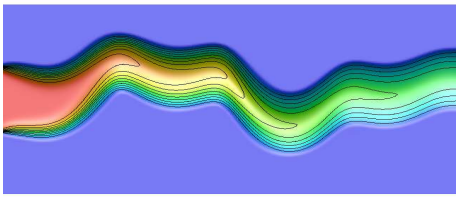
where

$$\begin{array}{lcl} \eta_1 & = & c_1 + c_3 + c_5 \\ \eta_2 & = & c_2 + c_3 + c_5 \\ \eta_3 & = & c_4 + c_5 \\ \xi_{kin,1} & = & \frac{1}{8}\,(-3c_1 - 3c_2 + 2c_3 - c_4 + c_5) \\ \xi_{kin,2} & = & \frac{1}{8}\,(-c_1 - c_2 - 2c_3 - 3c_4 + 3c_5). \end{array}$$

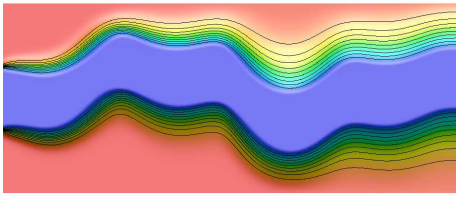
After temporal discretization and elimination of the two local equations, 2 nonlinearly coupled PDEs and 3 linear scalar PDEs remain, while the original problem (2.64) consisted of 7 nonlinearly coupled PDEs/ODEs.
Fig. 2.1 shows that two narrow zones establish at the edge of the plume where the reactants A and O_2 mix through dispersion and react, allowing the microbia population X_1 to grow. The product of the first reaction, B, reacts with NO_3^- in a second reaction zone, which is accompanied by a growth of the population X_2.

Reaction 1:

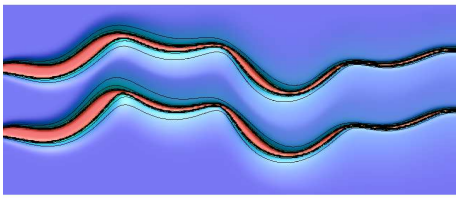
pollutant A (educt):



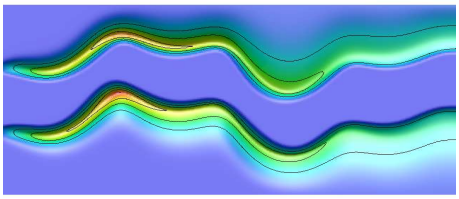
oxygen (educt):



biomass 1 (catalyzer):

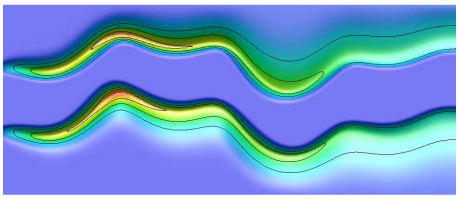


product B:

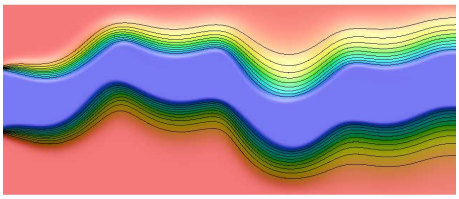


Reaction 2:

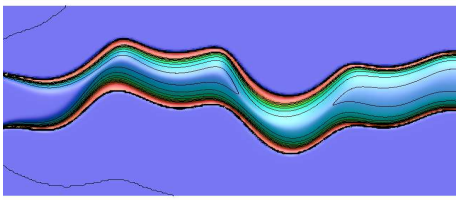
pollutant B (educt):



nitrate (educt):



biomass 2 (catalyzer):



product C:

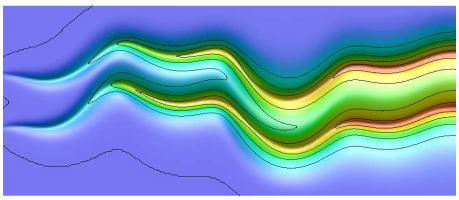


Figure 2.1: Concentrations at time $T = 40$, computation with 33 155 elements (Q_1); 232 085 D.o.F.; $h \approx 0.042$, and with a given (constant in time) flow field q.

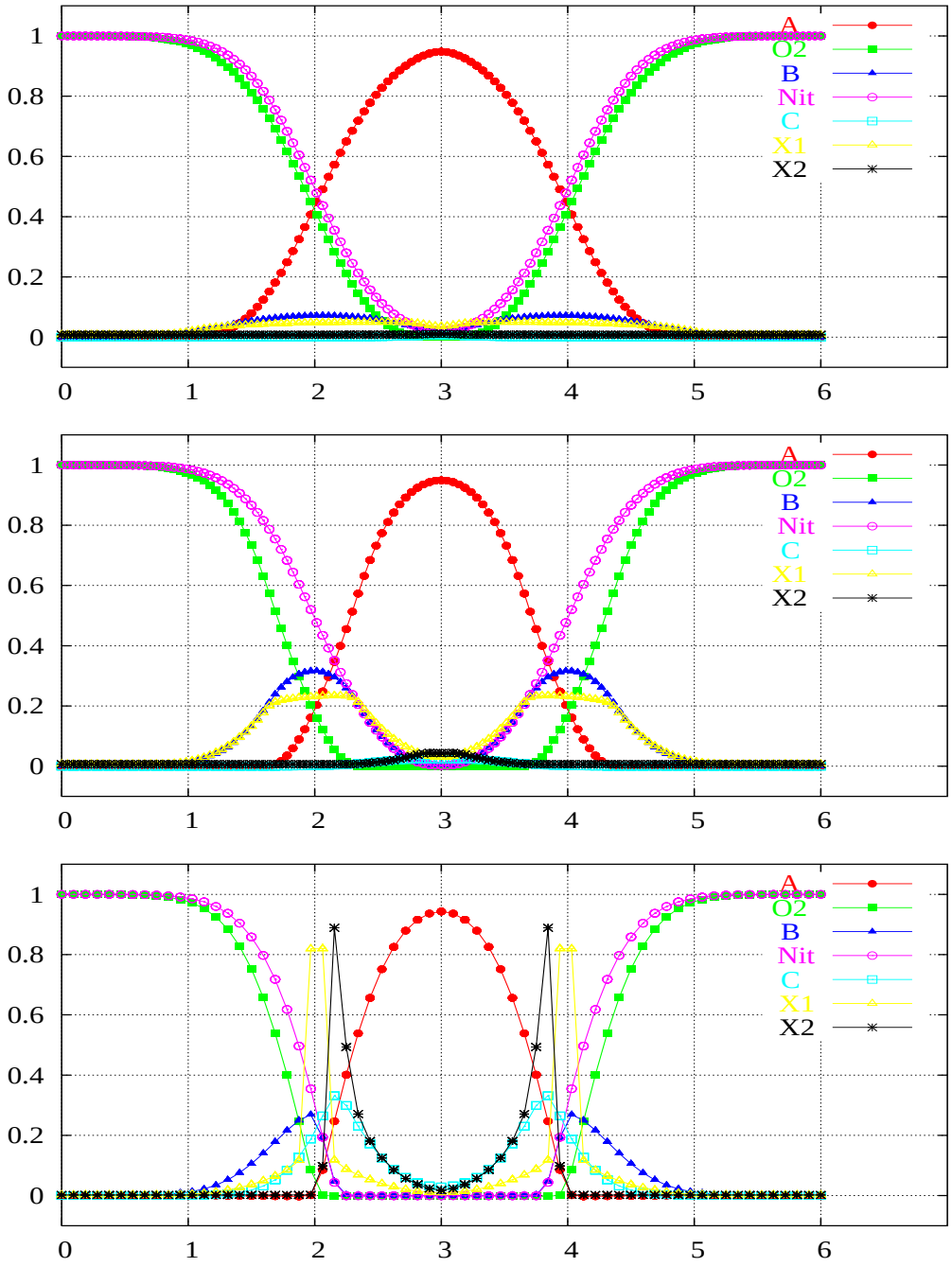


Figure 2.2: Transversal cut through plume at $x=5$ at times $T=5$, $T=10$, $T=40$ (top to bottom). 33 155 elements (Q_1); 232 085 D.o.F.; $h \approx 0.042$, Darcy velocity $q=(1,0)^T$.

reduced scheme:

elements/d.o.f. elements/d.o.f.	Newton steps glob. (loc.)	total cpu	cpu lin. sol.	cpu assem.	cpu loc. probl.
2 145/ 15 015	3.645 (0.635)	29	7	20	2
8 385/ 58 695	2.620 (0.777)	121	48	64	6
33 155/232 085	2.065 (0.906)	636	377	223	23

nonreduced scheme:

elements/d.o.f. elements/d.o.f.	Newton steps glob. (loc.)	total cpu	cpu lin. sol.	cpu assem.	cpu loc. probl.
2 145/ 15 015	4.163 (—)	74	36	35	—
8 385/ 58 695	4.493 (—)	425	265	155	—
33 155/232 085	7.341 (—)	3305	1977	1299	—

Table 2.2: CPU time in seconds for the reduced (upper table) and for the non-reduced formulation (lower table). All computations on a single processor.

elements	time red./time nonred.
2 145	0.39
8 385	0.28
33 155	0.19

Table 2.3: The acceleration by the reduction: a factor of 2.5 to 5, depending on the problem size.

Fig. 2.2 shows transversal cuts through the plume at different time levels. (Deviating from the simulation of Fig. 2.1, a standardised flow field $q = (1, 0)^T$ was used for Fig. 2.2.)
Table 2.2 shows the required CPU time and the number of Newton steps for the code using the reformulation and the code without the reformulation (i.e., based on formulation (2.3)). The given number of global Newton steps is the average number of Newton steps per time step for the global problem. The given local number of Newton steps is the average number of Newton steps for the local problem per mesh point and per time step.
A first observation is that the reformulation, which implies a 'concentration' of the nonlinearities in fewer equations, does *not* increase the number of (global) Newton steps. The number is even reduced (see Tab. 2.2)! This might be understood by the fact that to a large amount the nonlinearities

of the reactions are handled in the *local* problem, so that the *first* solution of the *local* problem, which precedes the first *global* Newton step, leads to a really good initial guess for the global Newton iteration.
The surprisingly small number of *local* Newton steps per global Newton step is due to the fact that in large parts of the computational domain almost stationary states are soon reached (which is not atypical for reactive transport problems in the subsurface).
Another observation is that the cpu time used for the local problems is only a few percent of the total cpu time. Furthermore we state that (though the assembling requires the computation of some additional terms as seen in (2.43), thanks to the smaller problem size, not only the time for the linear solver, but also the time for the assembling of the linear system is diminished by the reduction scheme.
The factor by which the cpu time is diminished through the reduction scheme depends on the size of the linear system, i.e., on the discretization parameter; see Table 2.3. The reason for this behaviour is the following: Since the complexity of the chosen linear solver is a bit worse than $O(n)$, while the complexity of the matrix assembling is $O(n)$, the *portion* of cpu time required for the linear solver increases with increasing mesh point number n. On large problems, where the time for linear solving is dominant, the impact of the reduction scheme is stronger.

2.5.4 Example 2: Kinetic biodegradation and equilibrium sorption of EDTA

In this section we consider a reactive network from *Chilakapati, Ginn, Szecsody* [18] which was also considered in *Fang, Yeh, Burgos* [28], and which models the fate of the contaminant EDTA[20]. 14 species are involved: $I=6$ mobile species ($Co(II)_{(aq)}$, Co(II)EDTA, Fe(III)EDTA, EDTA, Co(III)EDTA, O_2), and $\bar{I}=8$ immobile species (S^-Co, S^-, S^+Co(II)EDTA, S^+, S^+Fe(III)EDTA, S^+EDTA, S^+Co(III)EDTA, and the biospecies X). There are 10 reactions considered: $J_{eq} = J_{het} = 5$ equilibrium reactions, which are all heterogeneous, 3 kinetic chemical reactions and 2 biological binary Monod reactions

$$R = \mu\, c_X \, \frac{c_A}{\alpha_A + c_A} \, \frac{c_D}{\alpha_D + c_D}.$$

[20] ethylenediamine tetraacetic acid

R_1:	$\mathrm{Co(II)(aq)} + \mathrm{S^-}$	$\leftrightarrow$	$\mathrm{S^-Co}$	$k_f = 1200$	$k_b = 100$
R_2:	$\mathrm{Co(II)EDTA(aq)} + \mathrm{S^+}$	$\leftrightarrow$	$\mathrm{S^+Co(II)EDTA}$	$k_f = 2500$	$k_b = 100$
R_3:	$\mathrm{Fe(III)EDTA(aq)} + \mathrm{S^+}$	$\leftrightarrow$	$\mathrm{S^+Fe(III)EDTA}$	$k_f = 900$	$k_b = 100$
R_4:	$\mathrm{EDTA(aq)} + \mathrm{S^+}$	$\leftrightarrow$	$\mathrm{S^+EDTA}$	$k_f = 2500$	$k_b = 100$
R_5:	$\mathrm{Co(III)EDTA(aq)} + \mathrm{S^+}$	$\leftrightarrow$	$\mathrm{S^+Co(III)EDTA}$	$k_f = 250$	$k_b = 100$
R_6:	$\mathrm{S^+Co(II)EDTA}$	$\leftrightarrow$	$\mathrm{Co(II)(aq)+S^+EDTA}$	$k_f = 1$	$k_b = 10^{-3}$
R_7:	$\mathrm{S^+EDTA}$	$\rightarrow$	$\mathrm{Fe(III)EDTA(aq)+S^+}$	$k_f = 2.5$	$k_b = 0$
R_8:	$\mathrm{Co(II)EDTA(aq)}$	$\rightarrow$	$\mathrm{Co(III)EDTA(aq)}$	$k_f = 10^{-3}$	$k_b = 0$
R_9:	$\mathrm{EDTA(aq)} + 6\mathrm{O_2}$	$\rightarrow$	$3\mathrm{CO_2} + \mathrm{Biomass}$	$\mu = 0.025$	$\alpha_D = \alpha_A = 10^{-5}$
R_{10}:	$\mathrm{Fe(III)EDTA(aq)} + 6\mathrm{O_2}$	$\rightarrow$	$3\mathrm{CO_2} + \mathrm{Biomass}$	$\mu = 0.00025$	$\alpha_D = \alpha_A = 10^{-5}$

Since this example contains many sorption (i.e., heterogeneous) equilibrium reactions, it may be considered as a 'hard' test case for the reduction scheme (see the 'failure' of the decoupling method [63] for such problems).
As already in Example 1, each Monod reaction leads to two kinetic rates in our differential system. Hence, our system contains $R_{kin} = 7$ kinetic rate terms. Let us apply the reduction mechanism of Section 2.4 to this complex chemical reaction network with equilibrium/non-equilibrium reactions and mobile/immobile species, including biological decay:
The stoichiometric matrix has $I + \bar{I} = 14$ rows (each row corresponding to one species, in the above order of species) and $J = 12$ columns and reads as follows:

$$S = \left(\frac{S_1}{S_2}\right) = \left(\begin{array}{c|c} S^1_{eq} & S^1_{kin} \\ \hline S^2_{eq} & S^2_{kin} \end{array}\right)$$

$$= \left(\begin{array}{cccccc|cccccc} 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 6 & 0 & 0 \\ \hline -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 \end{array}\right)$$

Comparing this stoichiometric matrix with the scheme (2.27) we see that the blocks S^1_{mob}, S^2_{immo} are empty (i.e., $J_{mob} = J_{immo} = 0$, $J_{eq} = J_{het} = 5$). We observe that all columns in $S^1_{eq} = S^1_{het}$ and all columns in $S^2_{eq} = S^2_{het}$ are

linearly independent. Hence, the condition (2.29) for the applicability of our reduction method is met without any further transformation of S_{eq} in the sence of Lemma 5.9.
Let us determine the size of the different blocks of the reduced scheme. The column rank of S_1, S_2 is $J_1 = J_2 = 6$, each. Therefore S_1^*, S_2^* have 6 columns each. To get a complete basis of the space of the c_i, $\bar{c}_i$, which is $\mathbb{R}^I = \mathbb{R}^6$, $\mathbb{R}^{\bar{I}} = \mathbb{R}^8$, respectively, we have to add $I - J_1 = 0$ columns to S_1 and $\bar{I} - J_2 = 2$ columns to S_2. Hence, there is no η-block and an $\bar{\eta}$-block of size 2. The 5 linearly independent columns in S_{het}^1, S_{het}^2 yield 5 generalized sorption equations for ξ_{het} and 5 local algebraic equations for $\bar{\xi}_{het}$. The difference between the column ranks of S_1 and of S_{eq}^1 is $J_{kin}^{1'} = 1$, therefore the ξ_{kin}-block of PDEs has size 1. Analogously, The difference between the column rank of S_2 and of S_{eq}^2 is $J_{kin}^{2'} = 1$, therefore the $\bar{\xi}_{kin}$-block of ODEs has size 1. After elimination of the local equations, the reduced system consists of $J_{het} + J_{kin}^{1'} = 5 + 1 = 6$ coupled nonlinear equations for ξ_{het}, ξ_{kin}, while the given problem consists of 14 coupled PDEs/ODEs and 8 AEs. The resulting structure, after the reduction method has been applied, is given in Table 2.4: After elimination of the local equations, 6 nonlinearly coupled PDEs are left.

$\bar{\eta}_1$		$=$	const
$\bar{\eta}_2$		$=$	const
$\partial_t\, \xi_{kin,1}$	$+L\, \xi_{kin,1}$	$=$	$f_0(\bar{\eta}, \xi_{kin}, \bar{\xi}_{kin}, \xi_{het}, \bar{\xi}_{het})$
$\partial_t(\xi_{het1} - \bar{\xi}_{het1})$	$+L\, \xi_{het1}$	$=$	$f_1(\bar{\eta}, \xi_{kin}, \bar{\xi}_{kin}, \xi_{het}, \bar{\xi}_{het})$
$\partial_t(\xi_{het2} - \bar{\xi}_{het2})$	$+L\, \xi_{het2}$	$=$	$f_2(\bar{\eta}, \xi_{kin}, \bar{\xi}_{kin}, \xi_{het}, \bar{\xi}_{het})$
$\partial_t(\xi_{het3} - \bar{\xi}_{het3})$	$+L\, \xi_{het3}$	$=$	$f_3(\bar{\eta}, \xi_{kin}, \bar{\xi}_{kin}, \xi_{het}, \bar{\xi}_{het})$
$\partial_t(\xi_{het4} - \bar{\xi}_{het4})$	$+L\, \xi_{het4}$	$=$	$f_4(\bar{\eta}, \xi_{kin}, \bar{\xi}_{kin}, \xi_{het}, \bar{\xi}_{het})$
$\partial_t(\xi_{het5} - \bar{\xi}_{het5})$	$+L\, \xi_{het5}$	$=$	$f_5(\bar{\eta}, \xi_{kin}, \bar{\xi}_{kin}, \xi_{het}, \bar{\xi}_{het})$
$\partial_t\, \bar{\xi}_{kin,1}$		$=$	$f_6(\bar{\eta}, \xi_{kin}, \bar{\xi}_{kin}, \xi_{het}, \bar{\xi}_{het})$
$\partial_t\, \bar{\xi}_{kin,2}$		$=$	$f_7(\bar{\eta}, \xi_{kin}, \bar{\xi}_{kin}, \xi_{het}, \bar{\xi}_{het})$
$Q_1(\bar{\eta}, \xi_{kin}, \bar{\xi}_{kin}, \xi_{het}, \bar{\xi}_{het})$		$=$	0
$Q_2(\bar{\eta}, \xi_{kin}, \bar{\xi}_{kin}, \xi_{het}, \bar{\xi}_{het})$		$=$	0
$Q_3(\bar{\eta}, \xi_{kin}, \bar{\xi}_{kin}, \xi_{het}, \bar{\xi}_{het})$		$=$	0
$Q_4(\bar{\eta}, \xi_{kin}, \bar{\xi}_{kin}, \xi_{het}, \bar{\xi}_{het})$		$=$	0
$Q_5(\bar{\eta}, \xi_{kin}, \bar{\xi}_{kin}, \xi_{het}, \bar{\xi}_{het})$		$=$	0

Table 2.4: Example 2 after the transformation.

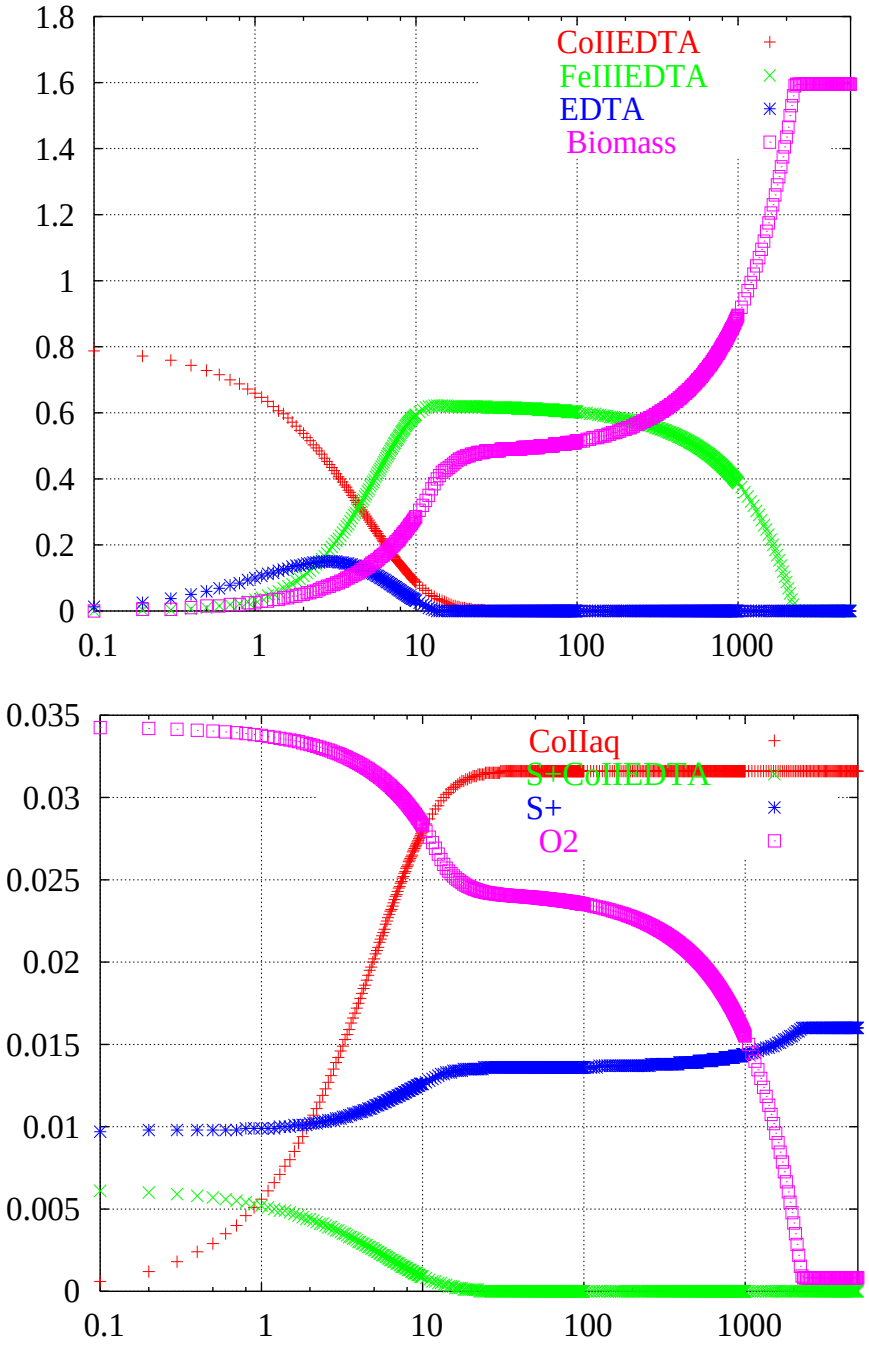


Figure 2.3: Evolution of some of the concentrations of Example 2.

The 2-D code was used to simulate the effectively 1-D problem with 561 grid points and 540 time-steps with a variable (preadapted) time-step size. The ratio of cpu time with/without reduction is 0.55. A gain of cpu time is achieved *both* for the linear solver and the matrix assembling (Table 2.5); the *relative* gain is, as expected, bigger for the linear solver, but the absolute

method	tot. cpu	cpu lin.sol.	cpu ass.	cpu loc. prob.
reduced	40.13	4.32	25.43	5.76
nonred.	73.07	16.08	51.61	—

Table 2.5: CPU time in seconds for the reduced and the nonreduced algorithm for Example 2.

gain is even bigger for the assembling time.

2.5.5 Ongoing tests

Currently comparisons of the presented reduction method with codes from many other research groups, some of them using splitting methods, some of them using one step methods, are going on for the benchmark problem [13]. Preliminary results on the efficiency are very promising [44].
Besides, in the framework of a diploma thesis an iterative splitting method is currently implemented in M++. This will allow direct cpu time comparisons of the size reduction method with established splitting methods, both using the *same* numerical kernel, in the next future.

2.6 Extensions to three-phase flows

In this section we will discuss how the reformulation method can be generalized to reactive flow problems where more than one mobile phase is present. A first observation is that an application to a two-phase problem, where *both* phases are mobile (equation (2.65)), is simple, at least if we assume that the flow regime decouples from the transport-reaction equations, i.e., the reactions have no back coupling to the flow regime.[21] The problem

$$\begin{aligned} \partial_t c + L_1 c &= S_1 R(c, \bar{c}), \\ \partial_t \bar{c} + L_2 \bar{c} &= S_2 R(c, \bar{c}) \end{aligned} \tag{2.65}$$

with a vector of species $c \in \mathbb{R}^{I_1}$ dissolved in phase 1 and a vector of species $\bar{c} \in \mathbb{R}^{I_2}$ dissolved in a phase 2, with an advection-dispersion operator L_1 for phase 1 and a different advection-dispersion operator L_2 for phase 2, is treated analogue to Sec. 2.4 . The resulting system (after the reformulation

[21] Such an assumption is often made, see e.g. [4]. Problems comprising a back coupling of the reactions to the flow are discussed in Sec. 2.6.2.

of Sec. 2.4) looks the same as system (2.36), with the main difference that (2.36-e) is no longer an ODE, but a PDE:

$$\begin{aligned}
\partial_t \eta + L_1 \eta &= 0 && \text{(2.66-a)}\\
\partial_t \overline{\eta} + L_1 \bar{\eta} &= 0 && \text{(2.66-b)}\\
\partial_t(\xi_{het} - \bar{\xi}_{het}) + L_1 \xi_{het} - L_2 \bar{\xi}_{het} &= A_{het} R_{kin}(c, \overline{c}) && \text{(2.66-c)}\\
\partial_t \xi_{kin} + L_1 \xi_{kin} &= A^1_{kin} R_{kin}(c, \overline{c}) && \text{(2.66-d)}\\
\partial_t \overline{\xi}_{kin} + L_2 \overline{\xi}_{kin} &= A^2_{kin} R_{kin}(c, \bar{c}) && \text{(2.66-e)}\\
Q(c, \bar{c}) &= 0 && \text{(2.66-f)}
\end{aligned}$$

As a result, the local system consists only of (2.66-f), and the PDEs (2.66-c)-(2.66-e) form the global problem.
We can proceed to *three* phases, from which either two or all three are mobile. Let us focus on the situation where two phases are mobile, since the generalization to mobile phases is, as seen above, trivial.
We will distinguish two different cases: In the first case (Sec. 2.6.1) we assume that the reactions do not affect the flow of the fluids. Note that this assumption was already made for the two-phase situation of Sec. 2.2-2.5. In the second case (Sec. 2.6.2) we allow that the chemical reactions have a non-negligible feedback on the flow regime.

2.6.1 Three-phase problems with decoupled fluid regime

We assume that the computation of the saturations (and the pressure) can be computed separately from the reactive transport equation. This is always the case when the reactions have only negligible effects on the fluid properties (viscosity, volume, ...) and the porosity. This assumption is usually justified, if in each phase there is a 'master species' which is hardly affected by reactions and which dominates its phase (such as the species H_2O in the water phase).
We are extending the model (2.8) to the following system (2.67)-(2.72):
The flow of the two mobile (fluid or gas) phases is governed by

$$\begin{aligned}
&\partial_t(\theta \rho^1 s^1) + \nabla \cdot (\rho^1 q^1) = 0,\\
&\partial_t(\theta \rho^2 s^2) + \nabla \cdot (\rho^2 q^2) = 0,
\end{aligned} \qquad (2.67)$$

with Darcy's law

$$\begin{aligned}
q^1 &= -\tfrac{K_a k^1_r(s^1)}{\mu^1} (\nabla p^1 - \rho^1 g \nabla z),\\
q^2 &= -\tfrac{K_a k^2_r(s^2)}{\mu^2} (\nabla p^2 - \rho^2 g \nabla z).
\end{aligned} \qquad (2.68)$$

Here, θ is the porosity, p^j the phase pressure, K_a the absolute permeability, k_r^j the phase relative permeability, μ^j the phase viscosity, s^j the phase saturation, g the gravitational constant, z the depth. Besides this we need two equations of state, relating ρ^j to p^j, a capillary pressure law $p^2 - p^1 = p_c(s_2)$, and the volume balance

$$s_1 + s_2 = 1. \tag{2.69}$$

The transport and the reactions of the species is governed by

$$\begin{array}{rcl} \partial(\theta c^1) + L^1 c^1 & = & \theta\, S_1 R(c^1, c^2, c^3), \\ \partial(\theta c^2) + L^2 c^2 & = & \theta\, S_2 R(c^1, c^2, c^3) \\ \partial(\rho c^3) & = & \theta\, S_3 R(c^1, c^2, c^3). \end{array} \tag{2.70}$$

where ρ is the bulk density,

$$L^j c_i^j = -\nabla \cdot (D^j(q^j)\nabla c_i^j - c_i^j q^j), \quad j = 1, 2, \;\; i = 1, ..., I_j,$$

where I_j is the number of species in phase j, i.e., $c^j \in \mathbb{R}^{I_j}$. Any mass exchange between the phases (such as the exchange between $CO_{2(aq)}$ and $CO_{2(gas)}$) is modeled by *reactions* (kinetic or equilibrium) between the species. If (2.70) contains equilibrium reactions, it has to be solved together with a set of algebraic equilibrium conditions

$$Q_{eq}(c) := S_{eq}^T \ln c - \ln K_{eq} = 0 \tag{2.71}$$

for $c = (c^1, c^2, c^3)$, resulting from the submatrix S_{eq} of $S = \left(\begin{array}{c} S_1 \\ \hline S_2 \\ \hline S_3 \end{array}\right)$. The problem (2.67)-(2.71) can be considered as the model [4] extended by equilibrium reactions and by reactive species being present in *every* phase. The subproblem (2.67)-(2.69) is well known for example in the field of reservoir modeling ('black oil model'); it can be solved for example by IMPES (implicit pressure explicit saturations) or IMPSAT (implicit pressure implicit saturations) schemes. Subproblem (2.70)-(2.71) is a transport-reaction problem.
The question now is whether we need any restrictive assumption on matrix S such that the transfromation of Sec. 2.4 can be applied to (2.70)-(2.71). It turns out that it is sufficient that

> all equilibrium conditions are acting between phase 1 and phase 2, or between phase 2 and phase 3, or within a single phase, cf. (2.73), and all equilibrium reactions are linearly independent. (2.72)

The assumption (2.72) obviously implies the following structure (2.73) for S_{eq}. The following lemma shows that the assumption (2.72) can be exploited to assume a rank condition (2.74), playing the role of (2.29) in the setting of Sec. 2.4:

Lemma 2.4 *Let S_{eq} be a stoichiometric matrix corresponding to the requirement (2.72), i.e., S_{eq} has the shape*[22]

$$S_{eq} = \left(\begin{array}{c|c|c|c|c} S^1_{gas} & S^1_{ex} & 0 & 0 & 0 \\ \hline 0 & S^2_{ex} & S^2_{mob} & S^2_{sorp} & 0 \\ \hline 0 & 0 & 0 & S^3_{sorp} & S^3_{immo} \end{array}\right). \tag{2.73}$$

Assume that S_{eq} has maximal column rank. Then there is a transformation of S_{eq} by elementary column operations such that the matrix, again denoted by S_{eq}, has the shape (2.73) where

$$(S^1_{gas}|S^1_{ex}), (S^2_{ex}|S^2_{mob}), \text{ and } (S^3_{sorp}|S^3_{immo}) \text{ have maximal column rank.} \tag{2.74}$$

For later use let us denote the number of reactions of the different types by $J_{gas}, J_{ex}, J_{mob}, J_{sorp}, J_{immo}$.

Proof. Applying Lemma 5.9 to the submatrices

$$\left(\begin{array}{c|c|c} S^1_{mob} & S^1_{ex} & 0 \\ \hline 0 & S^2_{ex} & S^2_{mob} \end{array}\right) \quad \text{and} \quad \left(\begin{array}{c|c} S^2_{sorp} & 0 \\ \hline S^3_{sorp} & S^3_{immo} \end{array}\right)$$

we transform S_{eq} into the shape

$$\left(\begin{array}{ccc|ccc} A & B & 0 & 0 & 0 & 0 \\ 0 & C & D & E & F & 0 \\ 0 & 0 & 0 & 0 & G & H \end{array}\right),$$

where $(A|B), (C|D), (E|F), (G|H)$ have maximal column rank. Possibly the columns of $(C|D|E)$ are linearly dependent. In this case we apply Lemma 5.9 to the submatrix

$$\left(\begin{array}{cccc} A & B & 0 & 0 \\ 0 & C & D & E \end{array}\right)$$

and thus obtain the structure of Lemma 2.4. ∎

[22]The notation 'S^1_{gas}' is not meant to indicate that the first phase is necessarily a gas phase.

Hence, it is justified to assume that without loss of generality the stoichiometric matrix

$$S = \left(\begin{array}{c|c} S_{eq} & S_{kin} \end{array} \right) = \left(\begin{array}{c|c} S_{eq}^1 & S_{kin}^1 \\ \hline S_{eq}^2 & S_{kin}^2 \\ \hline S_{eq}^3 & S_{kin}^3 \end{array} \right)$$

with S_{eq} fulfilling the condition (2.72) has the structure (2.73)-(2.74). Under this assumptions, the transformation of Sec. 2.4 can be extended to the three-phase problem (2.70)-(2.71) as follows. We perform the decomposition $S_i = S_i^* A_i$, $i = 1, 2, 3$ such as in (2.21). Thanks to the structure (2.73) and the linear independencies (2.74) we can choose the decomposition such that

$$\begin{aligned}
S_1^* &= (S_{gas}^1 | S_{ex}^1 | S_{kin}^{1'}), & A_1 &= \begin{pmatrix} \mathrm{Id}_{J_{gas}} & 0 & 0 & 0 & 0 & A_{gas}^1 \\ 0 & \mathrm{Id}_{J_{ex}} & 0 & 0 & 0 & A_{ex}^1 \\ 0 & 0 & 0 & 0 & 0 & A_{kin}^1 \end{pmatrix} \\
S_2^* &= (S_{ex}^2 | S_{mob}^2 | S_{sorp}^{2'} | S_{kin}^{2'}), & A_2 &= \begin{pmatrix} 0 & \mathrm{Id}_{J_{ex}} & 0 & A_{ex}^{sorp} & 0 & A_{ex}^2 \\ 0 & 0 & \mathrm{Id}_{J_{mob}} & A_{mob}^{sorp} & 0 & A_{mob}^2 \\ 0 & 0 & 0 & A_{sorp}^{sorp} & 0 & A_{sorp}^2 \\ 0 & 0 & 0 & A_{kin}^{sorp} & 0 & A_{kin}^2 \end{pmatrix} \\
S_3^* &= (S_{sorp}^3 | S_{immo}^3 | S_{kin}^{3'}), & A_3 &= \begin{pmatrix} 0 & 0 & 0 & \mathrm{Id}_{J_{sorp}} & 0 & A_{sorp}^3 \\ 0 & 0 & 0 & 0 & \mathrm{Id}_{J_{immo}} & A_{immo}^3 \\ 0 & 0 & 0 & 0 & 0 & A_{kin}^3 \end{pmatrix}
\end{aligned} \tag{2.75}$$

such that the S_i^* consist of a maximal system of linearly independent columns of S_i, $i = 1, 2, 3$. Herein, $S_{kin}^{1'} \in \mathbb{R}^{I_1 \times J_{kin}^{1'}}$, $S_{sorp}^{2'} \in \mathbb{R}^{I_2 \times J_{sorp}^{2'}}$, $S_{kin}^{2'} \in \mathbb{R}^{I_2 \times J_{kin}^{2'}}$, $S_{kin}^{3'} \in \mathbb{R}^{I_3 \times J_{kin}^{3'}}$ are submatrices of S_{kin}^1, S_{sorp}^2, S_{kin}^2, S_{kin}^3, respectively. We also define the matrices $S_i^\perp$ fulfilling the condition (2.22) for $i = 1, 2, 3$. Following Sec. 2.4 we multiply the i-th block of equations of system (2.70) by matrix $(S_i^{\perp T} S_i^\perp)^{-1} S_i^{\perp T}$ and by matrix $(S_i^{* T} S_i^*)^{-1} S_i^{* T}$, $i = 1, 2, 3$; we introduce the new vectors of unknowns $\eta^1 \in \mathbb{R}^{I_1 - J_{gas} - J_{ex} - J_{kin}^{1'}}$, $\xi^1 \in \mathbb{R}^{J_{gas} + J_{ex} + J_{kin}^{1'}}$, $\eta^2 \in \mathbb{R}^{I_2 - J_{ex} - J_{mob} - J_{sorp}^{2'} - J_{kin}^{2'}}$, $\xi^2 \in \mathbb{R}^{J_{ex} + J_{mob} + J_{sorp}^{2'} + J_{kin}^{2'}}$, $\eta^3 \in \mathbb{R}^{I_3 - J_{sorp} - J_{immo} - J_{kin}^{3'}}$, $\xi^3 \in \mathbb{R}^{J_{sorp} + J_{immo} + J_{kin}^{3'}}$ fulfilling

$$c^j = S_j^* \xi^j + S_j^\perp \eta^j, \quad j = 1, 2, 3. \tag{2.76}$$

We split the reaction vector

$$R = (R_{gas}, R_{ex}, R_{mob}, R_{sorp}, R_{immo}, R_{kin})$$

with the subvectors having the size $J_{gas}, J_{ex}, J_{mob}, J_{sorp}, J_{immo}, J_{kin}$, and we split the vectors ξ^i:

$$\xi^1 = \begin{pmatrix} \xi^1_{gas} \\ \xi^1_{ex} \\ \xi^1_{kin} \end{pmatrix}, \quad \xi^2 = \begin{pmatrix} \xi^2_{ex} \\ \xi^2_{mob} \\ \xi^2_{sorp} \\ \xi^2_{kin} \end{pmatrix}, \quad \xi^3 = \begin{pmatrix} \xi^3_{sorp} \\ \xi^3_{immo} \\ \xi^3_{kin} \end{pmatrix}$$

The resulting system (corresponding to (2.34)) reads

$$\begin{array}{llcl}
(\partial_t\theta + L_1) & \eta^1 & = & 0 \\
(\partial_t\theta + L_2) & \eta^2 & = & 0 \\
\partial_t\theta_3 & \eta^3 & = & 0 \\
\hline
(\partial_t\theta + L_1) & \xi^1_{gas} & = & \theta R_{gas} \qquad\qquad\qquad +\theta A^1_{gas} R_{kin} \\
(\partial_t\theta + L_1) & \xi^1_{ex} & = & \theta R_{ex} \qquad\qquad\qquad +\theta A^1_{ex} R_{kin} \\
(\partial_t\theta + L_1) & \xi^1_{kin} & = & \qquad\qquad\qquad\qquad\quad \theta A^1_{kin} R_{kin} \\
\hline
(\partial_t\theta + L_2) & \xi^2_{ex} & = & \theta R_{ex} + \theta A^{sorp}_{ex} R_{sorp} +\theta A^2_{ex} R_{kin} \\
(\partial_t\theta + L_2) & \xi^2_{mob} & = & \theta R_{mob} + \theta A^{sorp}_{mob} R_{sorp} +\theta A^2_{mob} R_{kin} \\
(\partial_t\theta + L_2) & \xi^2_{sorp} & = & \theta A^{sorp}_{sorp} R_{sorp} +\theta A^2_{sorp} R_{kin} \\
(\partial_t\theta + L_2) & \xi^2_{kin} & = & \theta A^{sorp}_{kin} R_{sorp} +\theta A^2_{kin} R_{kin} \\
\hline
\partial_t\rho & \xi^3_{sorp} & = & \theta R_{sorp} +\theta A^3_{sorp} R_{kin} \\
\partial_t\rho & \xi^3_{immo} & = & R_{immo} +\theta A^3_{immo} R_{kin} \\
\partial_t\rho & \xi^3_{kin} & = & \qquad\qquad\qquad\qquad\quad \theta A^3_{kin} R_{kin} \\
\hline
Q_{eq}(c_1, c_2, c_3) & & = & 0
\end{array}$$

Note that the assumptions (2.74) determine the presence/absence of coefficient matrices in front of the reaction terms. In order to eliminate the equilibrium rate terms $R_{gas}, R_{ex}, R_{mob}, R_{sorp}, R_{immo}$, we multiply the equation for ξ^3_{sorp} by A^{sorp}_{ex}, A^{sorp}_{mob}, A^{sorp}_{sorp}, A^{sorp}_{kin}, and substract it from the equation for ξ^2_{sorp}. Furthermore, we substract the ξ^1_{ex}-equation from the ξ^2_{ex}-equation. Then the J_{eq} equilibrium rate terms are concentrated in J_{eq} equations:

$$
\begin{array}{lll}
(\partial_t\theta + L_1) & \eta^1 & = 0 \\
(\partial_t\theta + L_2) & \eta^2 & = 0 \\
\partial_t\,\rho & \eta^3 & = 0 \\
\hline
\partial_t(\theta\xi^2_{ex} - \theta\xi^1_{ex} - \rho A^{sorp}_{ex}\xi^3_{sorp}) + L_2\xi^2_{ex} - L_1\xi^1_{ex} & & = \theta(A^2_{ex} - A^1_{ex} - A^{sorp}_{ex}A^3_{sorp})R_{kin} \\
(\partial_t\theta + L_1) & \xi^1_{kin} & = \theta A^1_{kin}R_{kin} \\
\partial_t\rho & \xi^3_{kin} & = \theta A^3_{kin}R_{kin} \\
\partial_t(\theta\xi^2_{sorp} - \rho A^{sorp}_{sorp}\xi^3_{sorp}) + L_2\xi^2_{sorp} & & = \theta(A^2_{sorp} - A^{sorp}_{sorp}A^3_{sorp})R_{kin} \\
\partial_t(\theta\xi^2_{kin} - \rho A^{sorp}_{kin}\xi^3_{sorp}) + L_2\xi^2_{kin} & & = \theta(A^2_{kin} - A^{sorp}_{kin}A^3_{sorp})R_{kin} \\
\hline
(\partial_t\theta + L_1) & \xi^1_{gas} & = \theta(R_{gas} + A^1_{gas}R_{kin}) \\
(\partial_t\theta + L_1) & \xi^1_{ex} & = \theta(R_{ex} + A^1_{ex}R_{kin}) \\
(\partial_t\theta + L_2) & \xi^2_{mob} & = \theta(R_{mob} + A^{sorp}_{mob}R_{sorp} + A^2_{mob}R_{kin}) \\
\partial_t\rho & \xi^3_{sorp} & = \theta(R_{sorp} + A^3_{sorp}R_{kin}) \\
\partial_t\rho & \xi^3_{immo} & = \theta(R_{immo} + A^3_{immo}R_{kin}) \\
\hline
 & Q_{eq}(c_1, c_2, c_3) & = 0
\end{array}
$$

We drop the J_{eq} equations containing the J_{eq} unknowns $R_{eq} = (R_{gas}, ..., R_{immo})$ and get

$$
\begin{aligned}
\partial_t\theta\eta^1 + L_1\eta^1 &= 0 && \text{(2.77-a)}\\
\partial_t\theta\eta^2 + L_2\eta^2 &= 0 && \text{(2.77-b)}\\
\partial_t\rho\eta^3 &= 0 && \text{(2.77-c)}\\
\partial_t\theta\xi^1_{kin} + L_1\xi^1_{kin} &= \theta A^1_{kin}R_{kin} && \text{(2.77-d)}\\
\partial_t(\theta\xi^2_{sorp} - \rho A^{sorp}_{sorp}\xi^3_{sorp}) + L_2\xi^2_{sorp} &= \theta A_{sorp}R_{kin} && \text{(2.77-e)}\\
\partial_t(\theta\xi^2_{kin} - \rho A^{sorp}_{kin}\xi^3_{sorp}) + L_2\xi^2_{kin} &= \theta A_{kin}R_{kin} && \text{(2.77-f)}\\
\partial_t(\theta\xi^2_{ex} - \theta\xi^1_{ex} - \rho A^{sorp}_{ex}\xi^3_{sorp}) + L_2\xi^2_{ex} - L_1\xi^1_{ex} &= \theta A_{ex}R_{kin} && \text{(2.77-g)}\\
\partial_t\rho\xi^3_{kin} &= \theta A^3_{kin}R_{kin} && \text{(2.77-h)}\\
Q_{eq}(c_1, c_2, c_3) &= 0. && \text{(2.77-i)}
\end{aligned}
$$

with $A_{ex} := A^2_{ex} - A^1_{ex} - A^{sorp}_{ex}A^3_{sorp}$, $A_{sorp} := A^2_{sorp} - A^{sorp}_{sorp}A^3_{sorp}$, $A_{kin} := A^2_{kin} - A^{kin}_{sorp}A^3_{sorp}$.
The equations for η^1, η^2, η^3 are linear and decoupled. The local problem consisting of the (time-discretized) ODEs and the AEs (2.77-h)-(2.77-i) can be solved for the local variables $\xi_{loc} := (\xi^1_{gas}, \xi^2_{ex}, \xi^2_{mob}, \xi^3_{sorp}, \xi^3_{immo}, \xi^3_{kin})$, and ξ_{loc} can be eliminated in the remaining PDEs; see the following Lemma 2.5.

The remaining nonlinear system for $\xi_{glob} := (\xi^1_{ex}, \xi^1_{kin}, \xi^2_{sorp}, \xi^2_{kin})$ consists of $J_{ex}+J^{1'}_{kin}+J^{2'}_{sorp}+J^{2'}_{kin}$ PDEs.
We close this section with a proof of the existence of a local resolution function of the local problem. With the implicit function theorem, it is sufficient to prove that the derivative of the local equations with respect to the local variables is invertible. Similar to Lemma 2.1 we obtain the following Lemma 2.5. Of course an analysis of the structure of the accumulation term of the global problem, similar to Lemma 2.3, should follow up in order to guarantee the well-posedness of the discretized global problem at least for small time steps.

Lemma 2.5 *(i) The matrix* $\frac{\partial Q_{eq}}{\partial \xi_{loc}}$ *with* $\xi_{eqloc} := (\xi^1_{gas}, \xi^2_{ex}, \xi^2_{mob}, \xi^3_{sorp}, \xi^3_{immo})$ *is regular.*

(ii) If the time step size $\tau > 0$ *is sufficiently small, then the matrix*

$$\frac{\partial Q}{\partial \xi_{loc}}$$

with

$$\begin{aligned} &\xi_{loc} = (\xi_{eqloc}, \xi^3_{kin}), \quad Q = (Q_{eq}, Q_{kin}), \\ &Q_{kin} = \xi^3_{kin} - \xi^3_{kin,old} - \tau A^3_{kin} R_{kin}(c_1, c_2, c_3), \end{aligned}$$

is regular.

Proof. (i). We compute

$$\frac{\partial Q_{eq}}{\partial (c^1, c^2, c^3)} = S^T_{eq}\Lambda = \left(\begin{array}{c|c|c|c|c} S^1_{gas} & S^1_{ex} & 0 & 0 & 0 \\ \hline 0 & S^2_{ex} & S^2_{mob} & S^2_{sorp} & 0 \\ \hline 0 & 0 & 0 & S^3_{sorp} & S^3_{immo} \end{array}\right)^T \begin{pmatrix} \Lambda_1 & & \\ & \Lambda_2 & \\ & & \Lambda_3 \end{pmatrix}$$

where $\Lambda_j \in \mathbb{R}^{I_j \times I_j}$ is the diagonal matrix with entries $1/c^j_i$. Using (2.76) and (2.75) we obtain

$$\frac{\partial (c^1, c^2, c^3)}{\partial (\xi^1_{gas}, \xi^2_{ex}, \xi^2_{mob}, \xi^3_{sorp}, \xi^3_{immo})} = \left(\begin{array}{c|c|c|c|c} S^1_{gas} & 0 & 0 & 0 & 0 \\ \hline 0 & S^2_{ex} & S^2_{mob} & 0 & 0 \\ \hline 0 & 0 & 0 & S^3_{sorp} & S^3_{immo} \end{array}\right)$$

By multiplication of the matrices we obtain

$$\frac{\partial Q_{eq}}{\partial \xi_{eqloc}} =$$

$$
\left(\begin{array}{c|c|c||c|c}
S_{gas}^{1T}\Lambda_1 S_{gas}^1 & 0 & 0 & 0 & 0 \\ \hline
S_{ex}^{1T}\Lambda_1 S_{gas}^1 & S_{ex}^{2T}\Lambda_2 S_{ex}^2 & S_{ex}^{2T}\Lambda_2 S_{mob}^2 & 0 & 0 \\ \hline
0 & S_{mob}^{2T}\Lambda_2 S_{ex}^2 & S_{mob}^{2T}\Lambda_2 S_{mob}^2 & 0 & 0 \\ \hline\hline
0 & S_{sorp}^{2T}\Lambda_2 S_{ex}^2 & S_{sorp}^{2T}\Lambda_2 S_{mob}^T & S_{sorp}^{3T}\Lambda_3 S_{sorp}^3 & S_{sorp}^{3T}\Lambda_3 S_{immo}^3 \\ \hline
0 & 0 & 0 & S_{immo}^{3T}\Lambda_3 S_{sorp}^3 & S_{immo}^{3T}\Lambda_3 S_{immo}^3
\end{array}\right)
$$

$$
= \left(\begin{array}{c|c|c}
S_{gas}^{1T}\Lambda_1 S_{gas}^1 & 0 & 0 \\ \hline
* & (S_{ex}^2|S_{mob}^2)^T\Lambda_2(S_{ex}^2|S_{mob}^2) & 0 \\ \hline
0 & * & (S_{sorp}^3|S_{immo}^3)^T\Lambda_3(S_{sorp}^3|S_{immo}^3)
\end{array}\right)
$$

Since each of the matrices S_{gas}, $(S_{ex}^2|S_{mob}^2)$, $(S_{sorp}^3|S_{immo}^3)$ has maximal column rank and all $c_i^j > 0$, this matrix is regular.

(ii). We get

$$
\frac{\partial Q}{\partial \xi_{loc}} = \left(\begin{array}{c|c}
\frac{\partial Q_{eq}}{\partial \xi_{eqloc}} & \frac{\partial Q_{eq}}{\partial \xi_{kin}^3} \\ \hline
O(\tau) & \mathrm{Id}_{J_{kin}^{3\prime}} + O(\tau)
\end{array}\right)
$$

Due to (i), the upper left part is regular. For $\tau = 0$, also the whole matrix is regular. With a continuity argument, the whole matrix is regular for τ sufficiently small. ∎

2.6.2 A three-phase problem with feedback of reactions to the fluid regime

The assumption of Sec. 2.6.1 that the flow regime on the one hand and the transport/reaction of species on the other hand are decoupled is not always justified. As an example may serve the problem of CO_2 storage in the subsurface. Let us apply a model with a water phase, a gas phase, and an immobile phase. Chemical reactions (including mineral precipitation/dissolution) influence the CO_2 concentration in the aqueous phase. The CO_2 concentration in the aqueous phase directly affects the amount of CO_2 in the gas phase (usually, an equilibrium between the two entities is assumed). If, due to injection, CO_2 is the dominant (or even the only) constituent of the gas phase, and gaseous CO_2 is coupled to aqueous CO_2, it follows that reactions can massively affect the amount of mass and the volume of the gas phase; chemical reactions may even cause a (local) disappearence/reappearance of the gas phase, and the assumption of reactions being decoupled from the fluid regime is unrealistic. Due to the size and the structure of the problem,

some splitting techniques seem inevitable. The question is which subproblems should be solved coupled and which can be splitted, and if the splitting can be non-iterative (SNIA) or should be iterative (SIA). In [38, 37, 39], an SNIA computation is proposed as follows: The computation of the pressure (and hence the Darcy velocities) is performed in a first subproblem. Then, with known velocities, the transport of all species, including H_2O, the 'master species' of the aqueous phase, is performed. This yields new saturations for all phases. After that, the chemistry problem, including exchange between the water and the gas phase, is solved. This last subproblem affects the saturation s^1 of the gas phase, hence the conservation of volume (2.69) is not exactly met after a time step is completed. This may cause instabilities [53].

In this section a method is proposed by which the reduction technique of Sec. 2.4 can be applied to the problem of subsurface CO_2-storage, after an iterative splitting between pressure and saturation computation on the one hand and transport-reaction including interphase mass exchange on the other hand, has been applied.

To handel this problem we extend the model (2.67)-(2.72) as follows: The flow is governed by

$$\begin{aligned} &\partial_t(\theta\rho^1 s^1) + \nabla\cdot(\rho^1 q^1) = f^1, \\ &\partial_t(\theta\rho^2 s^2) + \nabla\cdot(\rho^2 q^2) = f^2[=0], \\ &s^1 + s^2 = 0 \end{aligned} \tag{2.78}$$

with Darcy's law (2.68), equations of state, and a capillary pressure equation. The equations (2.78) contain a mass density source term f^1 for phase 1, the gas phase, and a mass density source term f^2 for phase 2, the water phase, modelling reactions including interphase mass exchange. We may set $f^2=0$, since the mass of the water phase is usually hardly affected by reactions. If we assume that $CO_{2(gas)}$ is the *only* constituent of the gas phase, we do not have to solve any transport-reaction equations within the gas phase and (2.70)-(2.71) becomes

$$\begin{aligned} \partial(\theta c^2) + L^2 c^2 &= r^1(s^1, c^2, c^3) \\ \partial(\rho c^3) &= r^2(s^1, c^2, c^3) \\ Q(s^1, c^2, c^3) &= 0 \end{aligned} \tag{2.79}$$

The term f^1 in (2.78) (which depends on the unknown concentrations c_i^j and possibly on s^1) contains the back coupling of the reactive transport system

(2.79) to the flow regime. r^1 and r^2 denote chemical source terms (chemical reactions and interphase exchanges), and Q denotes equilibrium conditions due to reactions and interphase exchanges.
For the numerical solution I suggest to use a sequential operator splitting (SIA). We may adapt the following scheme (e.g., [74]): Solve a time step for the problem

$$\partial_t u = A(u) + B(u) \tag{2.80}$$

by the fixed point iteration, consisting of two semi-steps,

$$\begin{aligned} \tfrac{u_{n+1,k+1}-u_n}{\tau} &= A(u_{n+1,k+1}) + B(u_{n+1,k}) \\ \tfrac{u_{n+1,k+2}-u_n}{\tau} &= A(u_{n+1,k+1}) + B(u_{n+1,k+2}) \end{aligned}$$

where n denotes the time step index and k the fixed point iterate. We adapt this splitting to problem (2.78)-(2.79) with the unknowns s^1, s^2, p^2 and c_i^2, $i = 1, ..., I_2$ and c_i^3, $i = 1, ..., I_3$. Let operator A contain the convection in the saturation equations, and let operator B contain the advection-dispersion in the transport-reaction equations and all reactions and interphase exchange terms. We obtain the following splitting scheme: In the first semi-step, solve a system of type

$$\begin{aligned} &\tfrac{\theta(\rho^1(p)s^1_{n+1,k+1}-\rho^1_n s^1_n)}{\tau} + \nabla\cdot(\rho^1 q^1) = f^1(s^1_{n+1,k}, c^2_{n+1,k}, c^3_{n+1,k}), \\ &\tfrac{\theta(\rho^2(p)s^2_{n+1,k+1}-\rho^2_n s^2_n)}{\tau} + \nabla\cdot(\rho^2 q^2) = 0, \\ &s^1_{n+1,k+1} + s^2_{n+1,k+1} = 0, \\ &\tfrac{\theta\,(c^2_{i,n+1,k+1}-c^2_{i,n})}{\tau} = -L^2 c^2_{i,n+1,k} + r^2_i(s^1_{n+1,k}, c^2_{n+1,k}, c^3_{n+1,k}), \quad i = 1, ..., I_2 \\ &\tfrac{\rho\,(c^3_{i,n+1,k+1}-c^3_{i,n})}{\tau} = r^3_i(s^1_{n+1,k}, c^2_{n+1,k}, c^3_{n+1,k}), \qquad\qquad i = 1, ..., I_3 \end{aligned} \tag{2.81}$$

together with Darcy's law, for $s^1_{n+1,k+1}$, $s^2_{n+1,k+1}$, $p^2_{n+1,k+1}$, q^1, q^2, $c^1_{i,n+1,k+1}$, $c^2_{i,n+1,k+1}$ where f^1, r^2, r^3 are discretized *explicitely*, i.e., depending on the known values $c^j_{i,n+1,k}$ and being independent of the $c^j_{i,n+1,k+1}$. The update of the c^j_i is obviously trivial, so (2.81) amounts to solve a problem of the structure (2.67)-(2.69) for the saturations and the pressure, with additional (explicit) source terms.
The second semi-step reads

$$\begin{aligned} &\tfrac{\theta(\rho^1 s^1_{n+1,k+2}-\rho^1_n s^1_n)}{\tau} + \nabla\cdot(\rho^1 q^1) = f^1(s^1_{n+1,k+2}, c^2_{n+1,k+2}, c^3_{n+1,k+2}), \\ &\tfrac{\theta(\rho^2 s^2_{n+1,k+2}-\rho^2_n s^2_n)}{\tau} + \nabla\cdot(\rho^2 q^2) = 0, \\ &\tfrac{\theta(c^2_{i,n+1,k+2}-c^2_{i,n})}{\tau} = -L^2 c^2_{i,n+1,k+2} + r^2_i(s^1_{n+1,k+2}, c^2_{n+1,k+2}, c^3_{n+1,k+2}), \quad i=1, ..., I_2 \\ &\tfrac{\rho(c^3_{i,n+1,k+2}-c^3_{i,n})}{\tau} = r^3_i(s^1_{n+1,k+2}, c^2_{n+1,k+2}, c^3_{n+1,k+2}) \quad i = 1, ..., I_3 \end{aligned} \tag{2.82}$$

where now the old Darcy velocities q^j and densities $\rho^j(p)$ from (2.81) in the saturation equations are used, the advective term $L^2 c^2_{i,n+1,k+2}$ is taken implicitely, and the reaction and phase exchange terms f^1, r^2, r^3 are taken implicitely. The equation for $s^2_{n+1,k+2}$ is explicit and decouples. The remaining system for the unknowns $s^1_{n+1,k+2}$ and the $c^2_{i,n+1,k+2}$, $i = 1, ..., I_2$ and the $c^3_{i,n+1,k+2}$, $i = 1, ..., I_3$, has, if all the exchange terms r^j, f^1 are modeled by chemical reactions, the structure

$$\begin{aligned} \begin{pmatrix} \frac{\theta\rho_1(s^1_{n+1,k+2} - s^1_n)}{\tau} \\ \frac{\theta\,(c^2_{i,n+1,k+2} - c^2_{i,n})}{\tau} + L^2 c^2_{i,n+1,k+2} \\ \frac{\rho\,(c^3_{i,n+1,k+2} - c^3_{i,n})}{\tau} \end{pmatrix} &= \theta SR(s^1_{n+1,k+2}, c^2_{n+1,k+2}, c^2_{n+1,k+2}) \\ &\quad + \text{explicit terms} \qquad (2.83) \\ Q(s^1_{n+1,k+2}, c^2_{n+1,k+2}, c^3_{n+1,k+2}) &= 0 \end{aligned}$$

The interesting observation is that in the second semi-step (2.83) of the SIA only *one* advection-dispersion operator L^2 is present (the transport of s^1 enters as an *explicit* term); from a structural point of view s^1 appears like an immobile species. Hence, this problem is of the same structure as the model treated in Sec. 2.4, and not of the structure of the model of Sec. 2.6.1. So the problem proposed in the beginning of this section may be tackled by the strategies of Sec. 2.4, after an SIA has been applied.

Note the the question of how to handle the potential (local) vanishing and reappearing of the gas phase is discussed in Sec. 4.7 in the framework of complementarity problems.

Also note that only the first semi-step complies with the postulate of volume conservation. Hence, it seems indeed reasonable to prefer an SIA scheme, as proposed here, to a non-iterative scheme, where (2.80) would be approximated by the two semi-steps

$$\begin{aligned} \tfrac{\tilde{u} - u_n}{\tau} &= A(\tilde{u}) \\ \tfrac{u_{n+1} - \tilde{u}}{\tau} &= B(u_{n+1}), \end{aligned}$$

where, compared to the SIA method proposed here, just the explicit terms are missing. When the SIA method has converged, the solution of the time step fulfills the volume conservation (2.69).

2.7 Summary of Chapter 2

We have proposed a new reduction scheme for multicomponent reactive transport problems, which is able to handle coupled mobile and immobile species and mixed equilibrium and kinetic reactions. The basic philosophy is to figure out how much the size of the nonlinear system of differential equations can be reduced *without* using splitting techniques or imposing restrictions on the stoichiometry.

A special property of the scheme is that even within those equations which remain coupled, these coupling terms do not occur under the transport operator, which would cause many nonzero entries in the Jacobian. The sparsity of the system matrix can be exploited if iterative linear solver are used. The transport operator remains linear. Under rather general conditions, assuming mass action law for the equilibrium reactions, the existence of the resolution functions can be theoretically confirmed.

The number of coupled ODEs and PDEs for standard formulations (e.g., [55, 64]) is the number of species minus the number of equilibrium reactions $I+\bar{I}-J_{eq}$; from these, $I-J_{mob}$ are PDEs. The number of coupled nonlinear PDEs for our reduction scheme is $J_{het}+J_{kin}^{1'}$, which lies between J_{het} and $J_{het}+J_{kin}$, where J_{het} is the number of heterogeneous equilibrium reactions and J_{kin} is the number of kinetic reactions. This number is always less than or equal to $I-J_{mob}$. The reformulation method has been implemented and applied to several test problems. Extensions to multiphase settings are possible.

An extension of the model to mineral precipitation-dissolution reactions and the adaption of the reformulation technique to this model are investigated in Chapter 4. The difficulty with mineral precipitation-dissolution is that the equilibrium condition, if total dissolution of minerals is not excluded by some a priori knowledge, is no longer a mere algebraic equation $Q_j(c,\bar{c})=0$ as it has been the case in this chapter (cf. (2.7)).

Chapter 3

Existence of global solutions

3.1 Introduction

In this chapter proofs of existence and uniqueness of a *global* solution, i.e., a solution on arbitrarily large time intervals, of multicomponent reactive transport problems in porous media with mass action reactions are given. The function space in which our global solution is found is $W_p^{2,1}(Q_T)$, $p > n+1$, (see Sec. 3.3.1 for the definition) for the mobile species concentrations. A recent result on global existence for the PDE-case in a higher regularity class $C^2(Q_T) \cap C^1(\overline{Q}_T)$, requiring higher regularity of initial values ($u_0 \in C^{2+\alpha}(\overline{\Omega})$) and of the domain ($\partial\Omega \in C^{2+\alpha}$) and with missing advection ($q \equiv 0$) and missing immobile species can be found in [61]. The technique which will be used in the following requires less regularity; it is based on Schaefer's fixed point theorem and Lyapunov techniques to derive the required a priori estimates. The main difficulty lies in the *high nonlinearity* of the mass action terms.

As a preparation to the PDE case, we start in Sec. 3.2 by giving an introduction into this field by considering the *ODE* version of the problem, since the ODE model can be analysed in a much less technical way than the PDE- (PDE-ODE-, DAE-) problem. We use this simple ODE model to explain how the specific structure of the source term can be exploited, which Lyapunov functional is suited for the ODE case, and how positivity/nonnegativity (see also [62, 9, 10]) of the solution can be assured.

The centrepiece of this chapter is Sec. 3.3 where the PDE case is investigated. The main task, after nonnegativity is settled, is the proof of a priori bounds

for solutions of the PDE problem. Together with a fixed point theorem, they lead to existence results, and the Gronwall lemma leads to uniqueness.
In Sec. 3.4, a generalization to coupled PDE-ODE systems (with moderate couplings) is given. In Sec. 3.5, some extensions of the existence and uniqueness results of the PDE problem to DAE problems including also equilibrium reactions are given.

3.2 The batch problem ('ODE case')

Let us consider the initial value problem related to the batch problem of a multispecies problem with kinetic rates. We assume the reactions to be *reversible*, i.e., for each 'forward' reaction, there is a corresponding 'backward' reaction:

$$\begin{aligned} u' &= SR(u), \\ u(0) &= u_0 \in \mathbb{R}^I_+ \end{aligned} \tag{3.1}$$

with forward-backward mass aktion kinetics

$$R_j(u) = R^f_j(u) - R^b_j(u), \tag{3.2}$$

$$R^f_j(u) = k^f_j \prod_{\substack{m=1...I \\ s_{mj}<0}} u_m^{-s_{mj}}, \quad R^b_j(u) = k^b_j \prod_{\substack{m=1...I \\ s_{mj}>0}} u_m^{s_{mj}}, \tag{3.3}$$

$j = 1, ..., J$, with $k^f_j > 0, k^b_j > 0, s_{mj} \in \mathbb{Z}$. The stoichiometric coefficients s_{mj} form an $I \times J$ matrix S. We will further need the assumption that all reactions are linearly independent:

$$\text{rank}(S) = J \tag{3.4}$$

It is well known that already very simple nonlinear initial value problems such as $u' = u^\alpha$, $u(0)=1$, $\alpha>1$, have no global solution. Hence, any attempt to prove the existence of a global solution for (3.1) has to exploit information of the special structure of this problem. One important fact is that, for a fixed species i, the source term $\sum_{j=1}^J s_{ij} R_j(u)$ is a polynomial in $u \in \mathbb{R}^I$ with coefficients being negative (positive) if and only if the corresponding monomial depends (does not depend) on u_i. This can be illustrated in a following simple example: consider the reactive network consisting of the

species A, B, C, D and the single reaction $A + B \longleftrightarrow 2C + D$. We get $R^f = k^f u_A u_B$, $R^b = k^b u_C^2 u_D$, $R = R^f - R^b$, and

$$\frac{d}{dt}\begin{pmatrix} u_A \\ u_B \\ u_C \\ u_D \end{pmatrix} = \begin{pmatrix} -R \\ -R \\ 2R \\ R \end{pmatrix} = \begin{pmatrix} -k^f \underline{\underline{u_A}} u_B + k^b u_C^2 u_D \\ -k^f u_A \underline{\underline{u_B}} + k^b u_C^2 u_D \\ -2k^b u_C^2 \underline{\underline{u_D}} + 2k^f u_A c_B \\ -k^b \underline{\underline{u_C^2}} u_D + k^f u_A u_B \end{pmatrix}; \tag{3.5}$$

note that any source term for u_i depending on u_i has a negative sign, and any source term independend of u_i has a positive sign.
In a general setting we can prove this interrelation as follows: Introducing the splitting

$$s_{ij} = s_{ij}^+ - s_{ij}^-, \quad s_{ij}^+ = \max\{s_{ij}, 0\} \geq 0, \quad s_{ij}^- = \max\{-s_{ij}, 0\} \geq 0, \tag{3.6}$$

we can write the rates (3.3), ordering the terms by their sign, as

$$R_j^f(u) = k_j^f \prod_{m=1..I} u_m^{s_{mj}^-}, \quad R_j^b(u) = k_j^b \prod_{m=1..I} u_m^{s_{mj}^+}, \tag{3.7}$$

and get for the source term of the i-th species equation by using (3.6)

$$\begin{aligned} f_i(u) &= \sum_{j=1}^{J} s_{ij} \left(R_j^f(u) - R_j^b(u) \right) \\ &= + \left(\sum_{j=1}^{J} s_{ij}^+ R_j^f(u) + \sum_{j=1}^{J} s_{ij}^- R_j^b(u) \right) \\ &\quad - \left(\sum_{j=1}^{J} s_{ij}^- R_j^f(u) + \sum_{j=1}^{J} s_{ij}^+ R_j^b(u) \right) \\ &= \left(\sum_{j=1}^{J} s_{ij}^+ k_j^f \prod_{m=1}^{I} u_m^{s_{mj}^-} + \sum_{j=1}^{J} s_{ij}^- k_j^b \prod_{m=1}^{I} u_m^{s_{mj}^+} \right) \qquad \text{(3.8-a)} \\ &\quad - \left(\sum_{j=1}^{J} s_{ij}^- k_j^f \prod_{m=1}^{I} u_m^{s_{mj}^-} + \sum_{j=1}^{J} s_{ij}^+ k_j^b \prod_{m=1}^{I} u_m^{s_{mj}^+} \right) \qquad \text{(3.8-b)} \end{aligned}$$

Considering $m = i$ in (3.8) we realize that, since always '$s_{ij}^+ = 0$ or $s_{ij}^- = 0$' holds, the term (3.8-a) is independent of u_i, while each summand in (3.8-b),

which is non-vanishing (i.e., $s_{ij} \neq 0$), contains a factor u_i^r with $r = s_{ij}^- \geq 1$ or $r = s_{ij}^+ \geq 1$.
Hence, for *given* $u_j = u_j(t) \geq 0$, $j \neq i$, the source term f_i considered as a function of u_i is of the following polynomial shape:

$$f_i(u) = \alpha_0(t) - \sum_{r=1}^{m} \alpha_r(t)\, u_i^r, \quad \alpha_i \geq 0, \tag{3.9}$$

where the α_i depend on the u_j, $j \neq i$, and on the stoichiometric coefficients, and are independent of u_i. It is important to note that the nonnegativity of the α_r is based on the nonegativity of the u_j, $j \neq i$. [1]
Nonnegativity of solutions. The first step, before the question of existence of a global solution can be tackled, is to prove that any (local) solution of (3.1) is *positive* or at least nonnegative.

Lemma 3.1 *For a strict positive initial value $u_0 \in \mathbb{R}_+^I$, the (local) solution of (3.1) is strictly positive and unique.*

Proof. Because of Lipschitz-continuity, problem (3.1) has a unique local solution. Let u be a solution of (3.1) on an arbitrary interval of existence $[0, T)$. Assume that there is a $t \in [0, T)$ and an index i with $u_i(t) \leq 0$. Then the set $N := \{t \in [0, T) \mid \exists j : u_j(t) = 0\}$ is nonempty, and since $u_j(0) > 0$ for all j, N has a minimal element $t_{min} \in (0, T)$; $u_i(t_{min}) = 0$ for an index i, and $u_j(t) > 0$ for all $t \in [0, t_{min})$ and all j follows. Hence, all the $\alpha_0(t)$, $\alpha_r(t)$ in the representation (3.9) are nonnegative on $[0, t_{min}]$. We recall that $\alpha_r(t) \geq 0$ are of C^0-regularity, since the u_j as functions of time have C^0-regularity and are nonnegative. On the compact interval $[0, t_{min}]$ the solution u_i is bounded, hence there is an interval $[0, C]$, $C > 0$, such that

$$0 \leq \frac{u_i(t)}{C} \leq 1, \quad \text{and therefore} \quad 0 \leq \left(\frac{u_i(t)}{C}\right)^r \leq \frac{u_i(t)}{C}$$

for all $r \in \mathbb{N}$, $t \in [0, t_{min}]$. Hence, $u_i(t)^r \leq C^{r-1} u_i(t)$. Applying this estimate on (3.9) we get the inequality[2]

$$u_i'(t) \geq -\sum_{r=1}^{m} C^{r-1} \alpha_r(t)\, u_i(t), \quad t \in [0, t_{min}]. \tag{3.10}$$

[1] For visualization see the example (3.5) which clearly meets this shape.

[2] Note that the essential property that is exploited for the representation (3.10) (and already for (3.9)) is that the rates contain no fractional exponents (i.e., the stoichiometric numbers $|s_{ij}|$) between zero and one.

Since the continuous functions α_r must have an upper bound on the compact interval $[0, t_{min}]$, and since $u_i(t)$ is positive on $[0, t_{min})$, we obtain the estimate

$$\frac{u_i'(t)}{u_i(t)} \geq -\tilde{C}, \quad t \in [0, t_{min}),$$

for a constant $\tilde{C} \geq 0$. By integration,

$$u_i(t) \geq u_i(0)\, e^{-\tilde{C}\, t}$$

follows for all $t \in [0, t_{min})$, which is a contradiction to the continuity of u_i on $[0, T)$ and the fact that $u_i(t_{min}) = 0$. So the assumption was wrong and u is strictly positive on $[0, T)$. ∎

Proof of existence of a global solution of (3.1). It is well known that for the existence of a solution of an initial value problem on the whole positive axis $[0, \infty)$, it is sufficient to prove that there is a function $f : [0, \infty) \longrightarrow \mathbb{R}$ such that for any (local) solution $u(t)$ and any t taken from the interval of existence, $|u(t)| \leq f(t)$ holds (see e.g. [36]). The construction of such a bound f can be done in two different ways. The first method is very simple, but requires a certain additional assumption on the stoichiometry. The second method is more general, but requires the definition of a Lyapunov functional.

Method A: Assumption of conservation of the number of atoms. Let us assume that there is a vector $s^{\perp}$ being orthogonal to all columns of S such that all components of $s^{\perp}$ are strictly positive:

$$\exists\, s^{\perp} \in S^{\perp} \cap \mathbb{R}^I_+ \neq \emptyset \tag{3.11}$$

This assumption is true for many reactive networks: Let us assume that there is a (finite) number of 'elementary entities' (='atoms') from which the species (='molecules') are built. Let $s_i^{\perp} > 0$, $i = 1, ..., I$, be the total number of atoms (which is assumed to be a strictly positive number) of which a molecule of the i-th species consists. If the j-th chemical reaction respects the conservation of the number of atoms, then $s^{\perp} \cdot s_j = 0$ holds; $s_j \in \mathbb{R}^I$ being the j-th column of S. Hence, if all reactions respect this conservation rule, then (3.11) holds. Note that this is the only place in this thesis where the species' composition of atoms is exploited; all the rest of this paper works without this assumption. (On the other hand method A does not require the specific structure (3.2)-(3.3) of the reaction rate, which is required by the Lyapunov technique of method B; see (3.18).)

Now multiplication of the ODE (3.1) by $s^\perp$ and definition of $\eta(t) := \langle s^\perp, u(t)\rangle_I$ yields

$$\eta'(t) = 0, \quad \eta(0) = \langle s^\perp, u(0)\rangle_I.$$

Hence, $\eta(t) \equiv \eta(0) = \langle s^\perp, u(0)\rangle_I$ for all t; thus $\langle s^\perp, u(t)\rangle_I = \eta(0) = \text{const}$. Since $s_i^\perp > 0$ we can solve this last equation for any $u_i(t)$:

$$u_i(t) = \frac{1}{s_i^\perp}\left(\eta(0) - \sum_{j\neq i} s_j^\perp u_j(t)\right) \leq \frac{\eta(0)}{s_i^\perp},$$

where we have used the nonnegativity result on u_j from the previous paragraph. Hence,

$$0 \leq u_i(t) \leq \frac{\langle s^\perp, u(0)\rangle_I}{s_i^\perp}$$

holds for all t for which the local solution exists. This bound is sufficient for the existence of a global solution $u(t)$, $t \in [0,\infty)$.

Method B: Lyapunov functional based on Gibbs free energy. The method A has the drawback that the assumption (3.11) is required. However, despite of the practical interpretation of this assumption given in the last paragraph, it is often *not* fulfilled even for 'reasonable' reactive networks having the claim to model real world problems. A reason could be that (reasonable) simplifications have been applied to a network, such as the following: Since often some species (like H_2O; in saline aquifers also Na^+, Cl^-, etc.) are present in abundance, with concentrations hardly affected by reactions, the evolution equations of these species are often dropped and the corresponding concentrations are assumed to be constant. The resulting symplified system is again of mass action structure. However, the corresponding reactions are void of the eliminated species, i.e., a chemical reaction $H_2O \longleftrightarrow OH^- + H^+$ becomes $0 \longleftrightarrow OH^- + H^+$; i.e., a column of S may be $(1,1,0)^T$, for which obviously no positive orthogonal vector $(s_1^\perp, s_2^\perp, s_3^\perp)^T \in \mathbb{R}^+_3$ exists. Such simplified systems do not meet the conservation of atoms, and they not necessarily meet condition (3.11). This motivates the following procedure to find a bound for $u(t)$ *without* assuming (3.11) to hold:
We define the functionals

$$\begin{aligned} &g_i : \overline{\mathbb{R}}_+ \longrightarrow \mathbb{R}, \quad g_i(u_i) = (\mu_i^0 - 1 + \ln u_i)\, u_i + \exp(1-\mu_i^0), \\ &g : \overline{\mathbb{R}}_+^I \longrightarrow \mathbb{R}, \quad g(u) = \sum_{i=1}^I g_i(u_i). \end{aligned} \tag{3.12}$$

The constants μ_i^0, $i = 1, ..., I$, are defined as follows: Let $\mu^0 \in \mathbb{R}^I$ be a solution of the linear system

$$S^T \mu^0 = -\ln K \tag{3.13}$$

where $K \in \mathbb{R}^J$ is the vector of equilibrium constants $K_j = k_j^f / k_j^b$ related to the J (kinetic) reactions. Note that, due to the assumption (3.4) that S has maximal column rank, the range of S^T is the whole $\mathbb{R}^J$. Hence, a solution μ^0 of (3.13) exists.[3] In general, μ^0 is not unique. The definition of g is inspired by the Gibbs free energy [35, 76, 69, 81]

$$\sum_{i=1}^{I} (\mu_i^0 + RT \ln u_i)\, u_i, \tag{3.14}$$

where the term in brackets is called the chemical potential of species i. An important property of g_i, g is that

$$\begin{aligned} g_i'(u_i) &= \mu_i^0 + \ln u_i, \\ \nabla g(u) &= \mu^0 + \ln u. \end{aligned}$$

Note that g_i, g are also well defined for zero concentrations (but g_i', ∇g are not).
A simple computation[4] shows that

$$g_i(u_i) \geq u_i \quad \text{and} \quad g_i(u_i) \geq e^{-\mu_i^0}(e-1) > 0 \quad \text{for all } u_i \in \overline{\mathbb{R}}_+; \tag{3.15}$$

in fact, the additive constant $\exp(1-\mu_i^0)$ incorporated in the definition (3.12) of g_i was chosen in such a way that g_i is strictly positive and dominates u_i. By (3.15), any bound for $g(u)$ immediately leads to a bound for u:

$$g(u) \geq g_i(u_i) \geq u_i \quad \forall u \in \overline{\mathbb{R}}_+^I \tag{3.16}$$

We can show that the mapping $t \longmapsto g(u(t))$, where $u(t)$ is a (local) solution of (3.1), is nonincreasing: We get[5]

$$\frac{d}{dt} g(u(t)) = \langle \nabla g(u(t)), u'(t) \rangle_I = \langle \mu^0 + \ln u(t), SR(u(t)) \rangle_I$$

[3] See also Footnote 13 on p. 31.

[4] Computing the minimum of the two functions $u_i \mapsto g_i(u_i) - u_i$ and $u_i \mapsto g_i(u_i)$, $\overline{\mathbb{R}}_+ \to \mathbb{R}$, we see that the minimum of the first is zero and the minimum of the second is $e^{-\mu_i^0}(e-1)$, from which (3.15) follows.

[5] The first step in (3.17) and also (3.18) exploit strict positivity of u (∇g is undefined otherwise) given by Lemma 3.1. In the PDE case (Sec. 3.3) the requirement of *strict* positivity is circumvented by considering $u(t)+\delta$, $\delta > 0$ and letting $\delta \longrightarrow 0$. That shifting procedure could also be applied to the ODE case, but we can do without.

$$
\begin{aligned}
&= \langle S^T \mu^0 + S^T \ln u(t), R(u(t)) \rangle_J \\
&= \langle -\ln K + S^T \ln u(t), R(u(t)) \rangle_J \leq 0. \qquad (3.17)
\end{aligned}
$$

The last estimate in (3.17) results from the fact that for each component $j = 1, ..., J$ of the inner product, the one factor is positive/zero/negative if and only if the other factor is negative/zero/positive: In fact, for $j = 1, ..., J$,

$$
\begin{aligned}
R_j(u) \gtreqless 0 \iff R_j^f(u) \gtreqless R_j^b(u) \qquad (3.18)
\end{aligned}
$$

$$
\begin{aligned}
&\iff \ln k_j^f + \sum_{s_{ij}<0} |s_{ij}| \ln u_i \gtreqless \ln k_j^b + \sum_{s_{ij}>0} s_{ij} \ln u_i \\
&\iff \ln \frac{k_j^f}{k_j^b} \gtreqless \sum_{i=1}^{I} s_{ij} \ln u_i \iff 0 \gtreqless -\ln K_j + \sum_{i=1}^{I} s_{ij} \ln u_i,
\end{aligned}
$$

which completes the proof of (3.17). The monotonicity of $t \mapsto g \circ u(t)$ (3.17) expresses the attempt of the system to reach chemical equilibrium.
Estimates (3.16) and (3.17) lead to the bound

$$
0 \leq u_i(t) \leq g(u(t)) \leq g(u(0)) = \text{const} \qquad (3.19)
$$

for every i and for all t for which the (local) solution u exists. Hence, a global solution of (3.1) exists.

3.3 The transport reaction problem ('PDE case')

In this section we prove the existence of a global solution to *transport*-reaction problems. We assume that all species are mobile. Note that in this section and also in Sec. 3.4, the expression 'global solution' is meant in the following way:
For arbitrarity $T > 0$, there is a solution existing on the time interval $[0, T]$. The ideas of (3.12)-(3.19) will play a central role.

3.3.1 Notation & definition of the problem

Let us define some notation.

$T>0$ is a fixed time given arbitrarily large.
$\Omega \subset \mathbb{R}^n$, $n \geq 2$, is a bounded domain with $\mathcal{O}^2$-regularity (see Def. 5.3)
ν is the outside unit normal vector on $\partial\Omega$.
$Q_T = (0,T) \times \Omega$ is the space-time domain.
$\mathbb{R}_+ = \{x \in \mathbb{R} \,|\, x > 0\}$ and $\overline{\mathbb{R}}_+ = \{x \in \mathbb{R} \,|\, x \geq 0\}$ are the sets of positive (nonnegative, respectively) real numbers.
$L^p_+(M) := \{u \in L^p(M) \,|\, u \geq 0 \text{ a.e.}\}$ is the set of nonnegative L^p-functions.
$|\cdot|_I$, $\langle \cdot, \cdot \rangle_I$: Euclidian norm/inner product in $\mathbb{R}^I$. Besides this notation of the scalar product, we sometimes make use of the symbol "$\cdot$" where this is common practise, e.g. in expressions like "$\nabla_x \cdot u$".
$W^m_p(M) := \{u \,|\, D^\alpha u \in L^p(M)\ \forall\, |\alpha| \leq m\}$, $1 \leq p < \infty$ and
$W^m_\infty(M) := \{u \,|\, \text{ess sup}_{x \in M} |D^\alpha u| < \infty \forall |\alpha| < m\}$ are the well known Sobolev spaces with the usual norm, denoted by $\|\cdot\|_{W^m_p(M)}$, $\|\cdot\|_{W^m_\infty(M)}$, $1 \leq p < \infty$, $m \in \mathbb{N}_0$.
The Banach space

$$W^{2,1}_p(Q_T) := \{u \,|\ u, \partial_t u, \partial_{x_i} u, \partial_{x_i}\partial_{x_j} u \in L^p(Q_T) \forall\, i,j = 1, ..., n\}$$

with the norm

$$\|u\|_{W^{2,1}_p(Q_T)} = \left(\|u\|^p_{L^p(Q_T)} + \|\partial_t u\|^p_{L^p(Q_T)} + \sum_{i=1}^n \|\partial_{x_i} u\|^p_{L^p(Q_T)} + \sum_{i,j=1}^n \|\partial_{x_i}\partial_{x_j} u\|^p_{L^p(Q_T)} \right)^{1/p}$$

is the space in which we will search for a solution $u_i \in W^{2,1}_p(Q_T)$.
Vector-valued function spaces are denoted by a superscipt indicating the dimension, that means for example $L^p(\Omega)^3 = L^p(\Omega) \times L^p(\Omega) \times L^p(\Omega)$; the corresponding norms are denoted by $\|v\|_{L^p(M)^I} := (\sum_{i=1}^I \|v_i\|^p_{L^p(M)})^{1/p}$,
$\|v\|_{L^\infty(M)^I} := \max_{i=1...I} \|v_i\|_{L^\infty(M)}$; analogously we define $W^{2,1}_p(Q_T)^I$, etc.
$\nabla = \nabla_x$ is the derivative of a scalar function with respect to the vector $x \in \Omega$;
∂_t, ∂_{x_i} denote partial differentiation.
D_x is the derivative (Jacobian) of a vector-valued function with respect to the vector $x \in \Omega$.
Inequalities such as $u \geq 0$ are also used for expressions in $\mathbb{R}^n$, if they hold for each component.
c is a generic positive constant depending only on the problem parameters, but not on the solution.

It is well known that even for moderate nonlinearities, such as quadratic source terms, may cause a 'blow up' of solutions in finite times, i.e., the non-existence of global solutions ([27] Sec 9.4, see also [68] Theorem 5). A main challenge in the proof of existence of global solutions lies in the high nonlinearity of the reactive terms. Hence, concerning the coefficients in the PDE system, we restrict our investigation to parameters constant in space and time (water content $\theta := 1$, scalar constant viscosity $\mathcal{K} > 0$) in order not to increase the complexity of the following. We also assume that

$$\operatorname{div} q = 0 \tag{3.20}$$

holds.[6]
Let p be fixed with

$$n + 1 < p < \infty. \tag{3.21}$$

Find a solution $u \in W_p^{2,1}(Q_T)^I$ of the problem

$$\begin{array}{lcll} \partial_t u + Lu & = & SR(u) & \text{on } Q_T, \\ u(0,x) & = & u_0 & \text{on } \Omega, \\ \frac{\partial u_i}{\partial \nu} & = & 0 & \text{on } (0,T] \times \partial\Omega,\ i = 1, ..., I \end{array} \tag{3.22}$$

with

$$S \in \mathbb{R}^{I \times J}, \quad |s_{ij}| \in \{0\} \cup [1, \infty),$$

with a vector of reaction rates R as in Sec. 3.2, with the transport operator

$$Lu = (L_1 u_1, ..., L_I u_I), \quad L_i u_i = \operatorname{div}(-\mathcal{K}\nabla u_i + q\, u_i) = -\mathcal{K}\Delta u_i + q \cdot \nabla u_i$$

for all $i = 1, ..., I$, with

$$u_0 \in W_p^{2-2/p}(\Omega), \quad u_0 \geq 0 \tag{3.23}$$

and with

$$q \in L^\infty(Q_T), \quad q \cdot \nu \in L^\infty((0,T] \times \partial\Omega). \tag{3.24}$$

Let us denote this problem by **(P)**.

[6] The assumption (3.20) is reasonable in the sense that, since usually q is modelled to be the solution of the Richards equation, the solution of this equation is indeed solenoidal for time-independent water content θ.

Note that, concerning the *boundary conditions* in (3.22), some remarks on the generalization to more realistic boundary conditions (inhomogeneous Robin on inflow, homogeneous Neumann on outflow part of the boundary) is made at the end of this section (cf. (3.61)).
For later use we state that $W_p^{2,1}(Q_T)$ is compactly embedded in the solution space $C^0(\overline{Q}_T)$ with p from (3.21):

$$W_p^{2,1}(Q_T) \hookrightarrow\hookrightarrow C^0(\overline{Q}_T) \tag{3.25}$$

This is a consequence of the embedding theorem 5.2 for $W_p^1(Q_T)$ (note that $Q_T \subset \mathbb{R}^{n+1}$).
For technical reasons, we first consider a *modified* problem instead of **(P)** itself:
We define the modified rate function $\tilde{R} : \mathbb{R}^I \longrightarrow \mathbb{R}^J$,

$$\tilde{R}(u) := R(u^+), \tag{3.26}$$

where u^+ is the positive part of u defined componentwise by

$$u_i^+ = \max\{u_i, 0\}, \quad u_i^- = \max\{-u_i, 0\} = -\min\{u_i, 0\}, \quad u_i = u_i^+ - u_i^-.$$

For later use let us state that (3.26) as a function of t, x is continuous if u is continuous. Let us denote the problem **(P)** which is modified in the sense that the rate vector $R(u)$ is replaced by the vector $\tilde{R}(u)$, by **($\mathbf{P}^+$)**.
In the next sections, we are going to prove that there is a (global) solution $u \in W_p^{2,1}(Q_T)^I$ of the modified problem **($\mathbf{P}^+$)**. For this, we will use Schaefer's fixed point theorem (Theorem 5.1). Since we will be able to show that any solution u of **($\mathbf{P}^+$)** is nonnegative, it also solves problem **(P)**. Finally, we will show the uniqueness of the solution of **(P)**.

3.3.2 Nonnegativity of solutions of problem ($\mathbf{P}^+$)

The first step of proving the existence of a global solution is the proof that any local solution of problem **($\mathbf{P}^+$)**, defined in the previous paragraph, is nonnegative.

Lemma 3.2 *Assume that there is a $\tau > 0$ and a function $u \in W_p^{2,1}(Q_\tau)$, $n+1<p<\infty$ being a solution of* **($\mathbf{P}^+$)** *on the time interval $[0, \tau]$. Then*

$$u \geq 0$$

holds on Q_τ.

Proof. Let $\Omega_i^- = \Omega_i^-(t)$ be the support of $u_i^-(t)$. Testing the i-th PDE from problem $(\mathbf{P}^+)$ by $-u_i^-$ yields

$$\frac{1}{2}\frac{d}{dt}\int\limits_{\Omega_i^-} |u_i^-|^2\,dx + \int\limits_{\Omega_i^-} (\mathcal{K}|\nabla u_i^-|^2 + u_i^-\,\nabla u_i^-\cdot q)\,dx = -\sum_{j=1}^{J} s_{ij}\int\limits_{\Omega_i^-} R_j(u^+)\,u_i^-\,dx \tag{3.27}$$

for $i = 1, ..., I$.
Analogue to (3.8) we divide the reactive terms into positive and negative terms, and by estimating all negative terms on the right hand side of (3.27) by zero (this corresponds to the *positive* terms (3.8-a) in (3.8)), the terms of (3.8-a) vanish and only the terms of (3.8-b) remain. We obtain the inequality

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int\limits_{\Omega_i^-} |u_i^-|^2\,dx + \int\limits_{\Omega_i^-} (\mathcal{K}|\nabla u_i^-|^2 + u_i^-\,\langle\nabla u_i^-, q\rangle_n)\,dx\\ &\quad\le \int\limits_{\Omega_i^-}\left(\sum_{j=1}^{J} s_{ij}^- k_j^f \prod_{m=1}^{I}(u_m^+)^{s_{mj}^-} + \sum_{j=1}^{J} s_{ij}^+ k_j^b \prod_{m=1}^{I}(u_m^+)^{s_{mj}^+}\right) u_i^-\,dx\end{aligned} \tag{3.28}$$

From the discussion of the terms in (3.8-b) we recall that each of the reaction terms in the right hand side of (3.28), which is nonvanishing, i.e., $s_{ij}^- > 0$ or $s_{ij}^+ > 0$, respectively, contains a factor $(u_i^+)^r$ with nonvanishing exponent $r = s_{ij}^- \ge 1$ or $r = s_{ij}^+ \ge 1$. Since $u_i^+ \equiv 0$ on the domain of integration Ω_i^-, the right hand side of (3.28) is equal to zero. Proceeding by applying standard estimates to the advective term, we get[7]

$$\frac{1}{2}\frac{d}{dt}\int\limits_{\Omega_i^-} |u_i^-|^2\,dx + \int\limits_{\Omega_i^-} \mathcal{K}|\nabla u_i^-|^2\,dx \le \frac{Q^2}{2\mathcal{K}}\int\limits_{\Omega_i^-} |u_i^-|^2\,dx + \frac{\mathcal{K}}{2}\int\limits_{\Omega_i^-} |\nabla u_i^-|^2\,dx,$$

where $Q := \|q\|_{L^\infty(Q_T)^n}$. Hence,

$$\frac{d}{dt}\int\limits_{\Omega_i^-} |u_i^-|^2\,dx + \int\limits_{\Omega_i^-} \mathcal{K}|\nabla u_i^-|^2\,dx \le \frac{Q^2}{\mathcal{K}}\int\limits_{\Omega} |u_i^-|^2\,dx;$$

[7] even without exploiting (3.20) here

in particular,

$$\frac{d}{dt}\int\limits_{\Omega_i^-} |u_i^-|^2\,dx \le \frac{Q^2}{\mathcal{K}}\int\limits_{\Omega} |u_i^-|^2\,dx.$$

Due to the initial condition, the solution of this ODE is $\int_{\Omega_i^-} |u_i^-|^2\,dx \equiv 0$ for all $t \ge 0$, and therefore $u_i \ge 0$ holds a.e. in Q_τ for all $i = 1, ..., I$. ∎

3.3.3 The fixed point operator

Our objective is to prove the existence of a solution of problem $(\mathbf{P}^+)$ using Schaefer's fixed point theorem. Let us define the fixed point operator.
Let $n+1 < p < \infty$ and $T>0$ be fixed. Let us define the operator $\mathcal{Z}$,

$$\begin{array}{rcl} \mathcal{Z} : W_p^{2,1}(Q_T)^I & \longrightarrow & W_p^{2,1}(Q_T)^I, \\ v & \longmapsto & u = \mathcal{Z}(v) \end{array} \tag{3.29}$$

by u being the solution of the linear problem

$$\begin{array}{lcll} \partial_t u + Lu & = & S\tilde{R}(v) & \text{on } Q_T, \\ u(0,x) & = & u_0 & \text{on } \Omega, \\ \frac{\partial u_i}{\partial \nu} & = & 0 & \text{on } (0,T]\times\partial\Omega,\ i = 1, ..., I \end{array} \tag{3.30}$$

with $u_0 \in W_p^{2-2/p}(\Omega)$.
Clearly, every fixed point of $\mathcal{Z}$ is a solution of problem $(\mathbf{P}^+)$.
Let us verify that $\mathcal{Z}$ is well defined:
For $v \in W_p^{2,1}(Q_T)^I$ with $p > n+1$, i.e., $v \in W_p^1(Q_T)^I$, the incidence $v \in C^0(\overline{Q}_T)^I$ holds, as stated in (3.25). Hence, $v^+ \in C^0(\overline{Q}_T)^I$ and $S\tilde{R}(v) = SR(v^+) \in C^0(\overline{Q}_T)^I$. In particular, $S\tilde{R}(v) \in L^q(Q_T)^I$ holds for all $1 \le q \le \infty$. Due to the linear parabolic theory (see Lemma 5.4), the existence of a unique solution u of problem (3.30) lying in $W_q^{2,1}(Q_T)^I$ for all $p \le q < \infty$ follows. Choose $q = p$.

3.3.4 A priori estimates

Some auxiliary functions, and the basic ideas

Let $\mu^0 \in \mathbb{R}^I$ be defined as in (3.13). In order to construct a Lyapunov functional for problem $\mathbf{(P^+)}$, let us define the following functions:

$$\begin{aligned} &g_i : \overline{\mathbb{R}}_+ \longrightarrow \mathbb{R}, && g_i(u_i) = (\mu_i^0 - 1 + \ln u_i)\, u_i + \exp(1-\mu_i^0),\ i = 1, ..., I \\ &g : \overline{\mathbb{R}}_+^I \longrightarrow \mathbb{R}, && g(u) = \sum_{i=1}^{I} g_i(u_i) \\ &G : L^2_+(\Omega)^I \longrightarrow \mathbb{R}, && G(u) = \int_\Omega g(u(x))\, dx \end{aligned} \tag{3.31}$$

(g_i and g are already known from Sec. 3.2) and, for $r \in \mathbb{N}$, the following powers of g:

$$\begin{aligned} &f_r : \overline{\mathbb{R}}_+^I \longrightarrow \mathbb{R} && f_r(u) = [g(u)]^r, \\ &F_r : L^\infty_+(\Omega)^I \longrightarrow \mathbb{R}, && F_r(u) = \int_\Omega f_r(u(x))\, dx \end{aligned} \tag{3.32}$$

We recall that

$$g_i(u_i) \geq u_i, \quad g_i(u_i) \geq 0 \quad g(u) \geq g_i(u_i) \geq u_i, \quad g(u) \geq 0 \tag{3.33}$$

for all $u_i \in \overline{\mathbb{R}}_+$. As a consequence,

$$f_r(u) = g(u)^r \geq [g_i(u_i)]^r \geq u_i^r, \quad f_r(u) \geq 0 \tag{3.34}$$

for all $u \in \overline{\mathbb{R}}_+^I$, $i = 1, ..., I$, $r \in \mathbb{N}$.
Upper bounds for g_i and f_r are obviously

$$g_i(u_i) \leq c\,(1+u_i^{1+\epsilon}), \quad g(u) \leq c\,(1+|u|_I^{1+\epsilon}), \quad f_r(u) \leq c\,(1+|u|_I^{r+\epsilon}) \tag{3.35}$$

for all $\epsilon > 0$, $u_i \in \overline{\mathbb{R}}_+$, $u \in \overline{\mathbb{R}}_+^I$, with $c = c(\epsilon)$ independent of u_i, u. If $u = u(x)$, integration over $x \in \Omega$ leads to corresponding bounds for F_r and $G = F_1$:

$$F_r(u) \geq \int_\Omega u^r\, dx, \quad F_r(u) \geq 0, \quad F_r(u) \leq c \int_\Omega (1 + |u|_I)^{r+\epsilon}\, dx \tag{3.36-a}$$

$$G(u) \geq \int_\Omega u\, dx \geq 0 \tag{3.36-b}$$

for all $u \in L^p_+(\Omega)$ for $1 \leq r < p$, where c may depend on r and ϵ, but not on u. Thus, in order to find $L^\infty(0,T;L^r(\Omega))$-bounds for solutions u, it is sufficient to estimate the value of the functional F_r along u!
We will frequently make use of the (classical) derivative of $f_r : \overline{\mathbb{R}}^I_+ \longrightarrow \mathbb{R}$,

$$\frac{df_r}{du} : \mathbb{R}^I_+ \longrightarrow \mathbb{R}^I, \frac{df_r}{du}(u) = r\,[g(u)]^{r-1}\nabla_u\, g(u) = r\, f_{r-1}(u)\,(\mu^0 + \ln u),$$

$r \in \mathbb{N}$. Note the different behaviour of f_r and df_r/du for u approaching the boundary of the positive cone: For $u \longrightarrow u_0 \in \partial\mathbb{R}^I_+$, f_r is bounded since $(\mu^0_i - 1 + \ln u_i)u_i \to 0$, while df_r/du is unbounded, since $\mu^0_i + \ln u_i$ is unbounded.
We will deal with the question, for which functions $u : [t_1,t_2]\times\Omega \longrightarrow \mathbb{R}^I$ the function $\partial f_r(u) : [t_1,t_2] \times \Omega \longrightarrow \mathbb{R}^I$ defined by

$$\begin{aligned} \partial f_r(u) : (t,x) \;\longmapsto\; & (\partial f_r)(u(t,x)) \\ & \left[\;:= \frac{df_r}{du}(u(t,x)) \right] \\ & = r\, f_{r-1}(u(t,x))\,(\mu^0 + \ln u(t,x)) \end{aligned} \tag{3.37}$$

is well defined and in $L^2(t_1,t_2;H^1(\Omega))^I$ later (see Lemma 3.6).
While G corresponds to the Gibbs free energy in the domain Ω, the function g is the corresponding energy density. A function like G is sometimes used as a Lyapunov function for reactive transport problems [59]. If one is able to show that G is bounded along solutions of $(\mathbf{P}^+)$, i.e., $G(u(t_2)) \leq G(u(t_1))$ for $t_2 > t_1$, or, a bit weaker, $G(u(t)) \leq c(t)G(u(0))$, where u is any solution of $(\mathbf{P}^+)$, then the boundedness of u in L^1 can easily be concluded through (3.36-b). However, from this information, the boundedness in *stronger* norms can only be derived if the nonlinearities in the rate functions are limited by small exponents. Since this is not the case for our problem, we have to modify the procedure: We consider F_r instead of G as a Lyapunov function (with the corresponding 'energy density' $f_r = g^r$, i.e., the density g to the r-th power). If the boundedness of F_r along solutions of problem $(\mathbf{P}^+)$ can be proven, then the boundedness of the solutions themselves in $L^p(Q_T)^I$ (as indicated in (3.36-a) with $r := p$), and finally, by the linear parabolic theory, in $W^{2,1}_p(Q_T)^I$, can be concluded (Cor. 3.8, 3.9).
The usage of F_r instead of G has major drawbacks: While G is convex and lower semicontinuous on the Hilbert space $L^2(\Omega)^I$ [59] (after extension by values $+\infty$ for arguments $u \in L^2(\Omega)^I \setminus L^2_+(\Omega)^I$), the semicontinuity of F_r on $L^2(\Omega)^I$ for $r > 1$ cannot be shown on this Hilbert space. As a consequence

of the convexity and semicontinuity of G on a Hilbert space, a result from convex analysis (Brezis' Lemma [14], p. 73) could be used, as in [59], serving as a chain rule for the differentiation of $G \circ u$ with respect to time;

$$G(u(t_2)) - G(u(t_1)) = \int_{t_1}^{t_2} \int_{\Omega} \langle \partial g(u(t,x)), \partial_t u(t,x) \rangle_I \, dx \, dt, \tag{3.38}$$

where ∂g is a subgradient of g. Since F_r for $r > 1$ is not semicontinuous, we have to use a different technique: Regularization. Let us lay out the main idea. If all functions F_r, u, $\partial f_r(u)$ were sufficiently smooth, an equation like (3.38) would also hold with G, g replaced by F_r, f_r:

$$F_r(u(t_2)) - F_r(u(t_1)) = \int_{t_1}^{t_2} \int_{\Omega} \langle \partial f_r(u(t,x)), \partial_t u(t,x) \rangle_I \, dx \, dt$$

with ∂f_r from (3.37). Hence, we approximate u by a *smooth* function $\bar{u} \in C^\infty(\overline{Q_T})^I$ with respect to the $W_p^{2,1}(Q_T)^I$ norm. We may expect an equation

$$F_r(\bar{u}(t_2)) - F_r(\bar{u}(t_1)) = \int_{t_1}^{t_2} \int_{\Omega} \langle \partial f_r(\bar{u}(t,x)), \partial_t \bar{u}(t,x) \rangle_I \, dx \, dt$$

to hold. However, there is another technical difficulty, since $\partial f_r \bar{u}(t,x))$ becomes unbounded if $\bar{u}(t,x) \geq 0$ approaches zero. (Note that such a behaviour is *not* the case for $F_r(\bar{u}(t,x))$ itself; compare the different behaviour of g_i' and g_i for $u_i \to 0$). Since only the nonnegativity, but not any strict positivity of u is known, we approximate u by $u_\delta := u + \delta$,[8] where $\delta > 0$, and apply the previously mentioned regularization to u_δ instead of u. The regularization of u_δ is denoted by $\overline{u_\delta}$.

The procedure can be visualized by the following diagram:

In (i) and in (ii), some continuity of F_r is required (see Lemma 3.4). In (iii), we use the smoothness of $\overline{u_\delta}$ and $\partial f_r(\overline{u_\delta})$ (see (3.56) for details). In (iv), some continuity of $\partial f_r(u_\delta)$ (for strictly positive arguments u_δ) is used (see Lemma 3.5). In (v), the trivial equality $\partial_t u_\delta = \partial_t u$ and then the PDE system

[8] An alternative to this approximation might be the construction of a maximum or comparison principle [52] to get a positive lower bound.

$$
\begin{array}{ccccc}
F_r(u(t_2)) - F_r(u(t_1)) & & & & \\
\text{(i)} \wr \ \text{Lem.3.4} & & & & \\
F_r(u_\delta(t_2)) - F_r(u_\delta(t_1)) & & \int_{t_1}^{t_2}\int_\Omega \langle \partial f_r(u_\delta), \partial_t u_\delta\rangle_I \, dx\, dt & \overset{\text{(v)}}{=} \overset{\text{L.3.7}}{\ldots} \leq & h(\delta,u) \\
\text{(ii)} \wr \ \text{Lem.3.4} & & \text{(iv)} \wr \ \text{Lem.3.5} & & \\
F_r(\overline{u_\delta}(t_2)) - F_r(\overline{u_\delta}(t_1)) & \overset{\text{(iii)}}{=} & \int_{t_1}^{t_2}\int_\Omega \langle \partial f_r(\overline{u_\delta}), \partial_t \overline{u_\delta}\rangle_I \, dx\, dt & &
\end{array}
$$

Figure 3.1: Outline of the proof of Theorem 3.3.

in its weak formulation is used, i.e., $\partial f_r(u_\delta)$ is used as a test function (see Lemma 3.6 for a justification).
The resulting terms are estimated in Lemma 3.7 by a term $h(\delta, u)$ with the property $h(\delta, u) \to 0$ for $\delta \to 0$ and fixed u, plus some boundary term. The boundary term vanishes if the condition[9]

$$q \cdot \nu = 0 \quad \text{on } \partial\Omega \times (0,T] \tag{3.39}$$

is assumed. Hence, we will simply call the case when (3.39) holds for short *the closed system case*[10], since no fluid enters or leaves the domain.
If (3.39) does not hold, then the occuring boundary term is estimated in the proof of Lemma 3.7, part (v). Finally, letting $\delta \to 0$ in the proof of Theorem 3.3 gives the desired bound for F_r.
The details of the proof visualized by Fig. 3.1 are given in the following. Lemmas 3.4, 3.5, 3.6 are rather technical. The main estimate, showing that the chosen functional F_r is, in spite of the highly nonlinear reaction terms, bounded along solutions, is contained in the proof of Lemma 3.7; see estimate (3.50), which is based on (3.18). Besides, the proof of Lemma 3.7 shows that the functional F_r 'behaves well' concerning the diffusive and the convective part of the equation in the sense that these terms do not lead to an increase of F_r along solutions (except a boundary term reflecting the effect of 'energy' entering the domain through advection). The Corollars 3.8 and 3.9 show that from the boundedness of F_r the boundedness of $\|u(t)\|_{L^r(\Omega)^I}$ for arbitrarily large $r > n+1$ follows, from which we obtain an upper bound for $\|u\|_{W_p^{2,1}(Q_T)^I}$ (i.e., the a priori estimate).

[9] In the proof, only $q \cdot n \geq 0$ is required; however, due to assumption (3.20), this is equivalent.

[10] This covers at least the non-advective case where $q \equiv 0$ on Q_T.

Boundedness of the Lyapunov functional

Since we want to apply Schaefer's Fixed Point Theorem (Theorem 5.1) to the operator $\mathcal{Z}$ defined in (3.29)-(3.30). Hence, we have to show that there is a constant $c > 0$ such that arbitrary solutions of the problem

$$u = \lambda \mathcal{Z}(u), \quad u \in W_p^{2,1}(Q_T)^I, \quad \lambda \in [0,1], \tag{3.40}$$

fulfil

$$\|u\|_{W_p^{2,1}(Q_T)^I} \leq c$$

with c independent of u and λ.
The equality (3.40) reads:

$$\begin{array}{lcll} \partial_t u + Lu & = & \lambda\, S\tilde{R}(u) & \text{on } Q_T, \\ u(0,x) & = & \lambda\, u_0(x) & \text{on } \Omega, \\ \frac{\partial u_i}{\partial \nu} & = & 0 & \text{on } (0,T] \times \Omega,\ i = 1, ..., I \end{array} \tag{3.41}$$

with $\Omega, p, u_0, L, \tilde{R}$ as in Sec. 3.3.1. Let us denote this problem (3.40)/(3.41) by $(\mathbf{P}_\lambda^+)$. Note that in the light of Lemma 3.2 we may replace $\tilde{R}(u)$ by $R(u)$ in (3.41) for solutions u of $(\mathbf{P}_\lambda^+)$.
Let us start by stating the main result of this section, expressing the monotonicity/limited growth of the functional F_r along solutions:

Theorem 3.3 *Let $0 \leq t_1 < t_2 \leq T$ and $p > n+1$ and $0 \leq \lambda \leq 1$ be given. Let $u \in W_p^{2,1}(Q_T)^I$ be a solution of $(\mathbf{P}_\lambda^+)$. Then the estimate*

$$F_r(u(t_2)) \leq e^{c\,(t_2 - t_1)}\, F_r(u(t_1)) \tag{3.42}$$

holds for all $r \in \mathbb{N}$, $r \geq 2$, where the constant $c > 0$ depends only on the problem parameters $\Omega, \mathcal{K}, q$ and on r, but not *on u, λ. If additionally (3.39) holds, then estimate (3.42) holds with $c = 0$ (even for all $r \in \mathbb{N}$ including $r = 1$); i.e., F_r is monotone along solutions:*[11]

$$F_r(u(t_2)) \leq F_r(u(t_1)). \tag{3.43}$$

The proof consists, as indicated in Fig. 3.1, of several steps, which are elaborated in the following lemmas. The proof of the Theorem 3.3 is given thereafter.

[11] If some fluid enters the domain (i.e., $q \cdot \nu < 0$), it may carry some 'chemical energy' into the domain. If there is no such inflow, then a monotonicity condition (3.43) holds.

Lemma 3.4 *Let $p > n+1$ and $r \in \mathbb{N}$. Then the mapping $F_r : L^\infty_+(\Omega)^I \longrightarrow \mathbb{R}$ is continuous.*

For the sake of completeness, we give a proof of this elementary statement:
Proof. Let $v_0 \in L^\infty_+(\Omega)^I$ and $\epsilon > 0$ be given. The mapping $f_r : \overline{\mathbb{R}}^I_+ \to \mathbb{R}$ is obviously continuous. Hence, it is uniformly continuous on the compact quadrilateral $\prod_{i=1}^{I}[0, 1+\text{ess sup}_{x\in\Omega}|v_0(x)|_I] \subset \overline{\mathbb{R}}^I_+$ which covers the range of v_0. So there is a $\delta > 0$ such that $|f_r(\xi) - f_r(\xi_0)| < \epsilon$ for all $\xi, \xi_0 \in [0, 1+$ ess $\sup_{x\in\Omega}|v_0(x)|_I]$ with $|\xi - \xi_0|_I < \delta$. Without loss of generality let us assume $\delta < 1$.
Now let $v \in L^\infty_+(\Omega)^I$ be arbitrarily chosen with $\|v - v_0\|_{L^\infty(\Omega)^I} < \delta$, i.e., $|v(x) - v_0(x)|_I < \delta$ for all $x \in \Omega$. From the construction of δ, $|f_r(v(x)) - f_r(v_0(x))|_I < \epsilon$ for all $x \in \Omega$ follows (note that $v(x), v_0(x) \in [0, \delta + \text{ess sup}_{x\in\Omega}|v_0(x)|_I] \subset [0, 1+\text{ess sup}_{x\in\Omega}|v_0(x)|_I]$). Integrating over Ω we get $|F_r(v) - F_r(v_0)| \le |\Omega|\,\epsilon$.
∎

For the following two lemmas, where the mapping (3.37) is investigated, it is essential that we consider only functions having a positive lower bound $\delta > 0$.

Lemma 3.5 *The mapping*

$$v \longmapsto \partial f_r(v) = r\, f_{r-1}(v)\,(\mu^0 + \ln v), \quad W^{2,1}_{p,\delta}(t_1, t_2; \Omega)^I \longrightarrow L^\infty([t_1, t_2]\times\Omega)^I,$$

where $p > n + 1$, $\delta > 0$, and

$$W^{2,1}_{p,\delta}(t_1, t_2; \Omega) := \{v \in W^{2,1}_p(t_1, t_2; \Omega) \,|\, v \ge \delta\},$$

is continuous.

Proof. Let $v \in W^{2,1}_{p,\delta}(t_1, t_2; \Omega)$. The space $W^{2,1}_p(t_1, t_2; \Omega)$ is continuously embedded in $L^\infty([t_1, t_2]\times\Omega)$ (see (3.25)). Similarly to the proof of Lemma 3.4 we see that the mapping $(t, x) \to r\, f_{r-1}(v(t, x))\,(\mu^0 + \ln v(t, x))$ with values in $\mathbb{R}^I$ is continuous (hence, uniformly continuous) on the compact quadrilateral

$$[t_1, t_2] \times \prod_{i=1}^{I}[\delta, 1 + \text{ess sup}_{(t,x)\in[t_1,t_2]\times\Omega}|v(t, x)|_I] \subset \mathbb{R} \times \mathbb{R}^I_+.$$

Similar as in the proof of Lemma 3.4, the continuity of the mapping

$$v \longmapsto \partial f_r(v), \quad W^{2,1}_{p,\delta}(t_1,t_2;\Omega)^I \longrightarrow \text{Ł}^\infty([t_1,t_2]\times\Omega)^I$$

follows. ∎

We now investigate the derivative of the mapping $(t,x) \longmapsto \partial f_r(v(t,x))) = r\, f_{r-1}(v(t,x))\,(\mu^0 + \ln v(t,x))$ with respect to $x \in \Omega$.

Lemma 3.6 (Derivative of the vector function $x \mapsto \partial f_r(v(t,x))$ w.r.t. $x \in \Omega$.)
Let $r \in \mathbb{N}$ and $p > n+1$. Let $v \in W^{2,1}_{p,\delta}(0,T;\Omega)^I$. Let us define the function $W(v) : Q_T \longrightarrow \mathbb{R}^{I\times n}$ by[12]

$$W(v) := [r(r-1)\, f_{r-2}(v) M_\mu(v) + r\, f_{r-1}(v)\Lambda_{1/v}]\, D_x v,$$

where $M_\mu(v)$ is the symmetric $I\times I$ matrix with entries $(\mu^0_i + \ln v_i)(\mu^0_j + \ln v_j)$, and $\Lambda_{1/v}$ is the $I\times I$ diagonal matrix with entries $1/v_i$. Then the identity

$$D_x(\partial f_r(v)) = W(v) \tag{3.44}$$

holds in $L^2(0,T;L^2(\Omega))^{I\times n}$; i.e.,

$$W(v) \in L^2(0,T;L^2(\Omega))^{I\times n}, \tag{3.45}$$

$$\partial f_r(v) \in L^2(0,T;H^1(\Omega))^I. \tag{3.46}$$

As a consequence of (3.46), $\partial f_r(v)$ can be used as a test function in problem $(\mathbf{P}^+_\lambda)$, if $v \in W^{2,1}_{p,\delta}(0,T;\Omega)^I$, i.e., if, in particular, $v \geq \delta$.

Proof. For the given function $v \in W^{2,1}_{p,\delta}(Q_T)^I$, obviously $v \in L^\infty(Q_T)^I$ and $v \geq \delta$ hold. Hence, from the definition of ∂f_r, $M_\mu(v)$, $\Lambda_{1/v}$,

$$r(r-1)\, f_{r-2} M_\mu(v) + r\, f_{r-1}\Lambda_{1/v} \in L^\infty(Q_T)^{I\times I}$$

follows. Since $D_x v \in L^2(Q_T)^{I\times n}$, (3.45) follows.

[12] Note that the first term of W vanishes for $r=1$.

For *smooth* functions v, e.g., $v \in C^1(\overline{Q}_T)^I \cap W^{2,1}_{p,\delta}(0,T;\Omega)^I$, classical calculus (see Lemma 5.8) shows that

$$\begin{aligned} D_x\,\partial f_r(v) &= r\,D_x[f_{r-1}(v)\,(\mu^0+\ln v)] \\ &= [r(r-1)\,f_{r-2}(v)M_\mu(v) + r\,f_{r-1}(v)\Lambda_{1/v}]\,D_x v \\ &= W(v) \in C^0(\overline{Q}_T)^{I\times n} \end{aligned} \tag{3.47}$$

holds (where in particular $v \geq \delta$ was used); so (3.45)-(3.44) are valid for such v. By approximating a given $v \in W^{2,1}_{p,\delta}(0,T;\Omega)^I$ by a sequence of smooth functions $(v^{(k)})$ w.r.t. the $W^{2,1}_{p,\delta}(0,T;\Omega)^I$-norm, the property is easily generalized to functions $v \in W^{2,1}_{p,\delta}(0,T;\Omega)^I$. Note that this makes use of the embedding $W^{2,1}_p(0,T;\Omega) \subset L^\infty(Q_T)$ (since this allows the approximating sequence with $v^{(k)} \geq \delta/2$ to exist) and of the continuity result of Lemma 3.5. ∎

Lemma 3.7 *Let $u \in W^{2,1}_p(Q_T)^I$ be a solution of problem $(\mathbf{P}^+_\lambda)$ and $\delta > 0$, $u_\delta := u+\delta$, $r \in \mathbb{N}$. In the closed system case (3.39) we have*

$$\int_{t_1}^{t_2}\int_\Omega \langle \partial f_r(u_\delta), \partial_t u_\delta\rangle_I\,dx\,dt \leq h(\delta,u),$$

and without assumption (3.39) we have

$$\int_{t_1}^{t_2}\int_\Omega \langle \partial f_r(u_\delta), \partial_t u_\delta\rangle_I\,dx\,dt \leq c\int_{t_1}^{t_2} F_r(u_\delta(t))\,dt + h(\delta,u), \tag{3.48}$$

if $r \geq 2$, where $c>0$ is a constant independent of u, δ and where $h(\delta,u) \longrightarrow 0$ for $\delta \longrightarrow 0$.

Proof. (i). Due to Lemma 3.2[13], $u \geq 0$; hence, $u_\delta \geq \delta$ holds:

$$u_\delta \in W^{2,1}_{p,\delta}(Q_T)^I$$

[13]The proof/assertion of Lemma 3.2 for solutions of $(\mathbf{P}^+)$ obviously also applies to solutions of the problem $(\mathbf{P}^+_\lambda)$

Therefore, due to Lemma 3.6, $\partial f_r(u_\delta) \in L^2(0,T;H^1(\Omega))^I$. As a consequence, we can use $\partial f_r(u_\delta)$ as a test function in the weak formulation of $(\mathbf{P}_\lambda^+)$ and get

$$\begin{aligned}
\int_{t_1}^{t_2}\int_\Omega \langle \partial f_r(u_\delta), \partial_t u_\delta\rangle_I \, dx\, dt &= \int_{t_1}^{t_2}\int_\Omega \langle \partial f_r(u_\delta), \partial_t u\rangle_I \, dx\, dt \\
&= \lambda \int_{t_1}^{t_2}\int_\Omega \langle \partial f_r(u_\delta), S\tilde{R}(u)\rangle_I \, dx\, dt \\
&\quad -\mathcal{K}\sum_{k=1}^n \int_{t_1}^{t_2}\int_\Omega \langle \partial_{x_k}(\partial f_r(u_\delta)), \partial_{x_k} u\rangle_I \, dx\, dt \\
&\quad -\int_{t_1}^{t_2}\int_\Omega \sum_{i=1}^I (\partial f_r(u_\delta))_i \, \langle \nabla u_i, q\rangle_n \, dx\, dt \\
&\quad +\mathcal{K}\int_{t_1}^{t_2}\int_{\partial\Omega}\sum_{i=1}^I [\partial f_r(u_\delta)]_i \frac{\partial u_i}{\partial \nu}\, ds\, dt \\
&=: \mathcal{I}_{reac} + \mathcal{I}_{diff} + \mathcal{I}_{adv} + 0
\end{aligned}$$

since the boundary condition lets the boundary term vanish.
(ii). For the reactive term $\mathcal{I}_{reac}$ we get

$$\mathcal{I}_{reac} = \lambda r \int_{t_1}^{t_2}\int_\Omega f_{r-1}(u_\delta)\, \langle \mu^0 + \ln u_\delta, SR(u)\rangle_I \, dx\, dt$$

where we used $\tilde{R}(u) = R(u)$, since $u \geq 0$. We split the inner product into a main part and a remainder,

$$\begin{aligned}
\langle \mu^0 + \ln u_\delta, SR(u)\rangle_I &= \langle \mu^0 + \ln u_\delta, SR(u_\delta)\rangle_I \\
&\quad + \langle \mu^0 + \ln u_\delta, SR(u) - SR(u_\delta)\rangle_I, \qquad (3.49)
\end{aligned}$$

where the main part is nonpositive:

$$\begin{aligned}
\langle \mu^0 + \ln u_\delta, SR(u_\delta)\rangle_I &= \langle S^T(\mu^0 + \ln u_\delta), R(u_\delta)\rangle_J \\
&= \langle -\ln K + S^T \ln u_\delta, R(u_\delta)\rangle_J \leq 0 \qquad (3.50)
\end{aligned}$$

In (3.50) we used the relation (3.13) between the vector μ^0 and the vector K. The last estimate in (3.50) results from the computation (3.18) stating that for each component $j = 1, ..., J$ of the inner product, the one factor is positive/zero/negative if and only if the other factor is negative/zero/positive. (3.49) and (3.50) yield

$$\langle \mu^0 + \ln u_\delta, SR(u) \rangle_I \leq \langle \mu^0 + \ln u_\delta, SR(u) - SR(u_\delta) \rangle_I,$$

and, multiplying this estimate by the (nonnegative) factor $\lambda r f_{r-1}(u_\delta)$, the estimate

$$\mathcal{I}_{reac} \leq \lambda r \int_{t_1}^{t_2} \int_\Omega f_{r-1}(u_\delta) \, \langle \mu^0 + \ln u_\delta, SR(u) - SR(u_\delta) \rangle_I \, dx \, dt$$

follows. Expanding[14] the term $SR(u) - SR(u_\delta)$, we obtain a representation as a sum in which each summand contains a factor $u_\delta - u = \delta$; the remaining products (i.e., after the separation of the factor δ) are in $C^0(\overline{Q}_T)$ and bounded in $L^\infty(Q_T)$ for fixed u, $\lambda \in [0,1]$, and $\delta \longrightarrow 0$. The factor $f_{r-1}(u_\delta)$ is also bounded in $L^\infty(Q_T)$ for $\delta \to 0$. It remains to show that the product consisting of the factor δ and the factor $\mu^0 + \ln u_\delta$ goes to zero in $L^1((t_1, t_2) \times \Omega)^I$ for $\delta \to 0$ and fixed u. For this, it is sufficient to show that $\delta \ln u_{i,\delta} \to 0$ in $L^1((t_1, t_2) \times \Omega)$, for $i = 1, ..., I$. We derive

$$\begin{aligned}
\int_{t_1}^{t_2} \int_\Omega |\delta \ln u_{i,\delta}| \, dx \, dt &= \int_{t_1}^{t_2} \int_{\ln u_\delta \leq 0} |\delta \ln u_{i,\delta}| \, dx \, dt + \int_{t_1}^{t_2} \int_{\ln u_\delta > 0} \delta \ln u_{i,\delta} \, dx \, dt \\
&\leq \int_{t_1}^{t_2} \int_{\ln u_\delta \leq 0} |\delta \ln \delta| \, dx \, dt + \int_{t_1}^{t_2} \int_{\ln u_\delta > 0} \delta \, u_{i,\delta} \, dx \, dt \\
&\leq T \, |\Omega| \, \delta \, |\ln \delta| + \delta \int_{t_1}^{t_2} \int_\Omega (u_i + \delta) \, dx \, dt
\end{aligned}$$

where both terms go to zero for $\delta \to 0$. Hence

$$\mathcal{I}_{reac} \leq h(\delta, u) \overset{\delta \to 0}{\longrightarrow} 0.$$

[14] in the sense $a_1 \cdot ... \cdot a_s - \bar{a}_1 \cdot ... \cdot \bar{a}_s = \sum_{l=1}^{s} a_1 \cdot ... \cdot a_{l-1} \cdot (a_l - \bar{a}_l) \cdot \bar{a}_{l+1} \cdot ... \cdot \bar{a}_s$

(iii). With the representations (3.44)-(3.47), (5.6)-(5.7) the diffusive term $\mathcal{I}_{diff}$ equals

$$\begin{aligned}\mathcal{I}_{diff} &= -\mathcal{K}r\sum_{k=1}^{n}\int_{t_1}^{t_2}\int_{\Omega}\langle\,[(r-1)\,f_{r-2}(u_\delta)M_\mu(u_\delta)+f_{r-1}(u_\delta)\Lambda_{1/u_\delta})]\,\partial_{x_k}u_\delta,\\ &\qquad\qquad ,\partial_{x_k}u\rangle_I\,dx\,dt\\ &= -r\,(r-1)\,\mathcal{K}\int_{t_1}^{t_2}\int_{\Omega}f_{r-2}(u_\delta)\sum_{k=1}^{n}\left|\left\langle\mu^0+\ln u_\delta,\partial_{x_k}u\right\rangle_I\right|^2\,dx\,dt \quad (3.51)\\ &\quad -\mathcal{K}r\int_{t_1}^{t_2}\int_{\Omega}f_{r-1}(u_\delta)\sum_{i=1}^{I}\sum_{k=1}^{n}\frac{1}{u_{\delta,i}}(\partial_{x_k}u_i)^2\end{aligned}$$

where the trivial equality $\nabla u_\delta = \nabla u$ was used. With the nonnegativity of f_{r-1}, f_{r-2} [15] we obviously can conclude that $\mathcal{I}_{diff}\le 0$.

(iv). Because of the complexity of the test function $\partial f_r(u_\delta)$ we want to avoid to shift the spatial derivatives (by integration by parts) within the advective term to this factor for the moment. Instead we treat the advective term $\mathcal{I}_{adv}$ as follows:

$$\begin{aligned}&\sum_{i=1}^{I}(\partial f_r(u_\delta))_i\,\langle\nabla u_i,q\rangle_n=\sum_{k=1}^{n}\sum_{i=1}^{I}(\partial f_r(u_\delta))_i\,\partial_{x_k}u_{\delta,i}\,q_k\\ &\quad=\langle\partial f_r(u_\delta),D_xu_\delta\,q\rangle_I=\langle(D_xu_\delta)^T\,\partial f_r(u_\delta),q\rangle_n=\left\langle\frac{d}{dx}\,f_r(u_\delta(x)),q\right\rangle_n.\end{aligned}$$

Now integration by parts yields

$$\mathcal{I}_{adv}=\int_{t_1}^{t_2}\int_{\Omega}f_r(u_\delta(x))\,\operatorname{div}q\,dx\,dt-\int_{t_1}^{t_2}\int_{\partial\Omega}f_r(u_\delta(x))\,q\cdot\nu\,ds\,dt.$$

Due to assumption (3.20) only the boundary term remains. The contribution of the outflow boundary $\partial\Omega_{out}$ can be estimated by zero; hence

$$\mathcal{I}_{adv}=-\int_{t_1}^{t_2}\int_{\partial\Omega}f_r(u_\delta)\,q\cdot\nu\,ds\,dt\le\int_{t_1}^{t_2}\int_{\partial\Omega_{in}}f_r(u_\delta)\,|q\cdot\nu|\,ds\,dt. \quad (3.52)$$

[15] Remember that the term containing f_{r-2} is immaterial for $r=1$.

The estimation of this boundary term follows in part (v). Note that in the case of a closed system (condition (3.39)), we have $\mathcal{I}_{adv} = 0$.

(v). The rest of the proof is devoted to the estimation of the boundary integral (3.52). We use (3.24) to obtain

$$-\int\limits_{\partial\Omega} f_r(u_\delta)\; q\cdot\nu\, ds \le c\,\|f_r(u_\delta)\|_{L^1(\partial\Omega)}$$

We note that $f_r = (f_{r/2})^2$, i.e,

$$\|f_r(u_\delta)\|_{L^1(\partial\Omega)} = \|f_{r/2}(u_\delta)\|^2_{L^2(\partial\Omega)}, \quad \|f_r(u_\delta)\|_{L^1(\Omega)} = \|f_{r/2}(u_\delta)\|^2_{L^2(\Omega)},$$

and we apply Lemma 5.6 to $f_{r/2}(u_\delta)$. We obtain

$$\begin{aligned}
-\int\limits_{\partial\Omega} f_r(u_\delta)\; q\cdot\nu\, ds &\le c\,\|f_r(u_\delta)\|_{L^1(\partial\Omega)} = c\,\|f_{r/2}(u_\delta)\|^2_{L^2(\partial\Omega)} \\
&\le c\,(\,\|\nabla_x f_{r/2}(u_\delta)\|_{L^2(\Omega)^n}\,\|f_{r/2}(u_\delta)\|_{L^2(\Omega)} + \|f_{r/2}(u_\delta)\|^2_{L^2(\Omega)}) \\
&\le c\,(\epsilon\,\|\nabla_x f_{r/2}(u_\delta)\|^2_{L^2(\Omega)^n} + \Lambda_\epsilon\,\|f_{r/2}(u_\delta)\|^2_{L^2(\Omega)}) \\
&= c\left(\epsilon\int\limits_\Omega \left|\nabla_x f_{r/2}(u_\delta)\right|^2_n\, dx + \Lambda_\epsilon\int\limits_\Omega f_r(u_\delta)\, dx\right)
\end{aligned}$$

where c only depends on q and on the domain, and where Λ_ϵ is a constant depending only on $\epsilon > 0$, both being independent of u, δ. The value of ϵ will be chosen later. An evaluation of the derivative yields

$$\begin{aligned}
\left|\nabla_x f_{r/2}(u_\delta)\right|^2_n &= \left|\frac{r}{2}\, f_{\frac{r}{2}-1}(u_\delta)(D_x u_\delta)^T(\mu^0+\ln u_\delta)\right|^2_n \\
&= \frac{r^2}{4}\, f_{r-2}(u_\delta)\;|(D_x u_\delta)^T(\mu^0+\ln u_\delta)|^2_n
\end{aligned}$$

Putting all together and integrating over time we get

$$\begin{aligned}
\mathcal{I}_{adv} \;\le\; & c\,\Lambda_\epsilon\int\limits_{t_1}^{t_2}\int\limits_\Omega f_r(u_\delta)\, dx\, dt \\
& +\epsilon\, c\,\frac{r^2}{4}\int\limits_{t_1}^{t_2}\int\limits_\Omega f_{r-2}(u_\delta)\;|(D_x u_\delta)^T(\mu^0+\ln u_\delta)|^2_n\, dx\, dt \qquad (3.53)
\end{aligned}$$

where c only depends on q, Ω. Now let us choose $\epsilon \leq \frac{2\mathcal{K}}{c} \leq \frac{4\mathcal{K}(r-1)}{cr}$, where the assumption $r \geq 2$ was used. Now Λ_ϵ only depends on $\mathcal{K}, \Omega, q$. Due to this choice of ϵ, the second of the two integrals of (3.53) can be absorbed by $\mathcal{I}_{diff}$ (see (3.51)); i.e.,

$$\mathcal{I}_{diff} + \mathcal{I}_{adv} \leq c\,\Lambda_\epsilon \int_{t_1}^{t_2} \int_{\Omega} f_r(u_\delta)\,dx\,dt = c\,\Lambda_\epsilon \int_{t_1}^{t_2} F_r(u_\delta(t))\,dt.$$

■

Now we can combine the lemmas and prove the main result of this section, following the strategy outlined in Fig. 3.1:

Proof of Theorem 3.3. Let $u \in W_p^{2,1}(Q_T)^I$ be a solution of $(\mathbf{P}_\lambda^+)$. Due to Lemma 3.2, we know[16], that $u \geq 0$ on $\overline{Q}_T$. Let $\delta > 0$ be fixed and $u_\delta := u + \delta$. Let $\epsilon = \epsilon(u, \delta) > 0$ be chosen such that

$$\epsilon\,\|\partial f_r(u_\delta)\|_{L^\infty(Q_T)^I} + \epsilon\,\|\partial_t u_\delta\|_{L^1(Q_T)^I} + \epsilon^2 \leq \delta.$$

Using the continuity of F_r (Lemma 3.4) and the continuity of ∂f_r (Lemma 3.5), we can choose a smooth function $\overline{u_\delta} \in C^\infty(\overline{Q}_T)^I$ sufficiently close to u_δ with respect to the $W_p^{2,1}(Q_T)^I$-norm, such that

$$\overline{u_\delta} \geq \frac{\delta}{2} \tag{3.54-a}$$

and

$$\|\partial_t \overline{u_\delta} - \partial_t u_\delta\|_{L^1(Q_T)^I} \leq \epsilon \tag{3.54-b}$$

and

$$|[F_r(u_\delta(t_2)) - F_r(u_\delta(t_1))] - [F_r(\overline{u_\delta}(t_2)) - F_r(\overline{u_\delta}(t_1))]| \leq \delta \tag{3.54-c}$$

and

$$\|\partial f_r(u_\delta) - \partial f_r(\overline{u_\delta})\|_{L^\infty([t_1,t_2]\times\Omega)^I} \leq \epsilon. \tag{3.54-d}$$

[16] see footnote 13 on p. 85

Note that for (3.54-a) and (3.54-c) we have used the embedding (3.25). From (3.54-d) and (3.54-b) and Hölder's inequality we obtain

$$\begin{aligned}
&\left| \int_{t_1}^{t_2} \int_{\Omega} \langle \partial f_r(u_\delta), \partial_t u_\delta \rangle_I \, dx \, dt - \int_{t_1}^{t_2} \int_{\Omega} \langle \partial f_r(\overline{u_\delta}), \partial_t \overline{u_\delta} \rangle_I \, dx \, dt \right| \\
&\quad \leq \quad \epsilon \, \| \partial f_r(u_\delta) \|_{L^\infty(Q_T)^I} + \epsilon \, \| \partial_t \overline{u_\delta} \|_{L^1(Q_T)^I} \\
&\quad \leq \quad \epsilon \, \| \partial f_r(u_\delta) \|_{L^\infty(Q_T)^I} + \epsilon \, (\, \| \partial_t u_\delta \|_{L^1(Q_T)^I} + \epsilon) \\
&\quad \leq \quad \delta. \qquad (3.55)
\end{aligned}$$

For the smooth function $\overline{u_\delta}$, we have

$$\begin{aligned}
F_r(\overline{u_\delta}(t_2)) - F_r(\overline{u_\delta}(t_1)) &= \int_{t_1}^{t_2} \partial_t F_r(\overline{u_\delta}(t)) \, dt = \int_{t_1}^{t_2} \partial_t (\int_{\Omega} f_r(\overline{u_\delta}(t,x)) \, dx) \, dt \\
&= \int_{t_1}^{t_2} \int_{\Omega} \frac{d}{dt} f_r(\overline{u_\delta})(t,x) \, dx \, dt = \int_{t_1}^{t_2} \int_{\Omega} \langle \partial f_r(\overline{u_\delta}(t,x)), \partial_t \overline{u_\delta}(t,x) \rangle_I \, dx \, dt, \qquad (3.56)
\end{aligned}$$

where the fundamental theorem of calculus for the smooth function $\int_\Omega f_r \circ \overline{u_\delta} \, dx : [0,T] \to \mathbb{R}$, and the exchange of ∂_t and $\int_\Omega$ for $f_r(\overline{u_\delta}) \in C^\infty(\overline{Q}_T)$, as well as the chain rule for differentiation for the smooth function $(t,x) \mapsto \partial f_r(\overline{u_\delta}))(t,x)$ were used. Putting the estimate (3.54-c) and equation (3.56) and (3.55) together yields

$$\left| [F_r(u_\delta(t_2)) - F_r(u_\delta(t_1))] - \int_{t_1}^{t_2} \int_{\Omega} \langle \partial f_r(u_\delta), \partial_t u_\delta \rangle_I \, dx \, dt \right| \leq 2\delta. \qquad (3.57)$$

Applying Lemma 3.7 we obtain

$$[F_r(u_\delta(t_2)) - F_r(u_\delta(t_1))] \leq 2\delta + h(\delta, u) \qquad (3.58)$$

in the closed system case (3.39), and

$$[F_r(u_\delta(t_2)) - F_r(u_\delta(t_1))] \leq c \int_{t_1}^{t_2} F_r(u_\delta(t)) \, dt + 2\delta + h(\delta, u) \qquad (3.59)$$

else, with c independent of u, δ and with $h(\delta, u) \longrightarrow 0$ for $\delta \longrightarrow 0$. In the closed system case, letting $\delta \longrightarrow 0$ in (3.58), leads to (3.43), since we have $F_r(u_\delta(t_i)) \longrightarrow F_r(u(t_i))$ (Lemma 3.4).
In the general case we get, passing to the limit $\delta \longrightarrow 0$ in (3.59),

$$F_r(u(t_2)) - F_r(u(t_1)) \leq c \int_{t_1}^{t_2} F_r(u(t))\, dt$$

Applying the Gronwall Lemma we get (3.42). ∎

Corollary 3.8 (preliminary a priori estimate) *Let $p > n+1$. For arbitrary solutions $u \in W_p^{2,1}(Q_T)$ of $(\mathbf{P}_\lambda^+)$, $\lambda \in [0,1]$, and arbitrary $r \in \mathbb{N}$, there is a constant $C = C(u_0, q, \mathcal{K}, \Omega, T)$ independent of u and t and λ, such that*

$$\|u(t)\|_{L^r(\Omega)^I} \leq C$$

for all $t \in [0,T]$.

Proof. Obviously it is sufficient to consider $r \geq 2$. We use the initial condition $u(0) = \lambda u_0$, then Theorem 3.3 for $t_1 = 0$, and finally property (3.34) to get

$$\begin{aligned} e^{ct}\, F_r(\lambda\, u_0) &= e^{ct}\, F_r(u(0)) \geq F_r(u(t)) = \int_\Omega f_r(u(t,x))\, dx \\ &= \int_\Omega [g(u(t,x))]^r\, dx \geq \int_\Omega u_i(t,x)^r\, dx. \end{aligned} \tag{3.60}$$

Let $\epsilon > 0$. Using (3.35) there is a $\tilde{c} > 0$ independent of r, u, λ such that

$$f_r(\lambda u_0) \leq \tilde{c}\,(1 + |\lambda u_0|_I^{r+\epsilon}) \leq \tilde{c}\,(1 + |u_0|_I^{r+\epsilon})$$

for all $u \in \overline{\mathbb{R}}_+^I$, $0 \leq \lambda \leq 1$. By integration and multiplication by e^{ct} and application of this to the left hand side of estimate (3.60), we obtain

$$\tilde{c}\, e^{ct} \int_\Omega (1 + |u_0(x)|_I^{r+\epsilon})\, dx \geq e^{ct}\, F_r(\lambda u_0) \geq \|u_i(t)\|_{L^r(\Omega)}^r.$$

Since (3.23) implies $u_0 \in L^\infty(\Omega)^I$, the expression on the right is bounded for $t \in [0,T]$. ∎

Corollary 3.9 (a priori estimate). *Let $p > n+1$ and $1 \leq s < \infty$ be fixed. Then there is a constant $c > 0$ (depending on s and u_0 and other parameters of problem $(\mathbf{P}_\lambda^+)$, but independent of u and $\lambda \in [0,1]$) such that for an arbitrary solution $u \in W_p^{2,1}(Q_T)^I$ of problem $(\mathbf{P}_\lambda^+)$*

$$\|u\|_{W_s^{2,1}(Q_T)^I} \leq c$$

holds for all $1 \leq s < \infty$. In particular, $u \in W_s^{2,1}(Q_T)^I$ holds.

Proof. Let $p > n+1$ and $1 \leq s < \infty$ be given. Let $u \in W_p^{2,1}(Q_T)$ be a solution of $(\mathbf{P}_\lambda^+)$. The Hölder inequality and application of Cor. 3.8 for sufficiently[17] large r leads to the fact that the right hand side of our problem, $S\tilde{R}(u(t))$, is bounded in $L^s(\Omega)$, by a constant c depending on s and u_0 and other parameters, but independent of t, u, λ. Integrating over time, the existence of a constant c depending on s, u_0, and T, but independent of u, λ follows, such that

$$\|S\tilde{R}(u)\|_{L^s(Q_T)^I} \leq c$$

for arbitrary s. From the theory of linear parabolic equations[18] (Lemma 5.4), the existence of a constant c with

$$\|u\|_{W_s^{2,1}(Q_T)^I} \leq c$$

follows. ∎

Other boundary conditions

Summarizing the procedure of this section we state that the proof of a priori bounds in $W_p^{2,1}(Q_T)$ consists of two main steps:
1. The construction of an $L^p(Q_T)$ a priori bound by estimating the functional F_r along solutions (Lemma 3.7, Theorem 3.3, Cor. 3.8).
2. An assertion from the linear parabolic theory that a right-hand side bounded in $L^p(Q_T)$ leads to a solution bounded in $W_p^{2,1}(Q_T)$ (Lemma 5.4, Cor. 3.9) is used.
In problems $(\mathbf{P})$,$(\mathbf{P}^+)$,$(\mathbf{P}_\lambda^+)$ homogeneous Neumann boundary conditions were assumed all over the boundary $(0,T)\times\partial\Omega$. More realistic boundary conditions would be

$$\begin{array}{lcll} (\mathcal{K}\nabla u_i - u_i q)\cdot\nu & = & b_i \geq 0 & \text{on } (0,T)\times\partial\Omega_{in} \\ \frac{\partial u_i}{\partial \nu} & = & 0 & \text{on } (0,T)\times\partial\Omega_{out} \end{array} \tag{3.61}$$

[17] The required value of r obviously depends on s and the stoichiometric coefficients s_{ij}.
[18] We do not claim here that problem $(\mathbf{P}_\lambda^+)$ is linear!

where the boundary of Ω is decomposed into

$$\partial\Omega_{in} \cup \partial\Omega_{out} = \partial\Omega.$$

The rest of this section is devoted to the proof that under this boundary condition, provided $b \in \text{Ł}^\infty(\partial\Omega_1)^I$ and $q \cdot \nu \le 0$ on $\partial\Omega_{in}$, *the $L^p(Q_T)$ a priori estimate remains valid.* (Hence, if the data are such that step 2 holds for the linear parabolic problem, the $W_p^{2,1}(Q_T)$-estimate for $(\mathbf{P}_\lambda^+)$ follows.)
The only deviation from the proof of Lemma 3.7 is that we have to reconcider the boundary terms. Since we decided to treat the advective term (part (iv) of the proof of Lemma 3.7) in a different way than the diffusive term (part (iii) of the proof), the boundary term of the one part contains the factor $f_r(u_\delta)$, while the other term contains the factor $\partial f_r(u_\delta)$. In fact we have the boundary terms

$$-\int\limits_{\partial\Omega} f_r(u_\delta)\, q\cdot\nu\, ds + \sum_{i=1}^{I} \int\limits_{\partial\Omega} [\partial f_r(u_\delta)]_i \mathcal{K}\, \frac{\partial u_i}{\partial \nu}\, ds.$$

The first integral was already estimated in part (v) of the proof of Lemma 3.7. The second integral can be written, by exploiting the boundary condition (3.61) applied to problem $(\mathbf{P}_\lambda^+)$, as

$$I_{bdry} = \sum_{i=1}^{I} I_{bdry,i}\,, \quad I_{bdry,i} = \int\limits_{\partial\Omega_{in}} [\partial f_r(u_\delta)]_i\, (\lambda b_i + u_i\, q\cdot\nu)\, ds.$$

We obtain

$$\begin{aligned}
I_{bdry,i} &= \int\limits_{\partial\Omega_{in}} [\partial f_r(u_\delta)]_i\, (\lambda b_i - u_i\, |q\cdot\nu|)\, ds \\
&= \int\limits_{\partial\Omega_{in}} [\partial f_r(u_\delta)]_i\, (\lambda b_i + (\delta - u_{\delta,i})\, |q\cdot\nu|)\, ds \\
&= \int\limits_{\partial\Omega_{in}} r f_{r-1}(u_\delta)\, (\mu_i^0 + \ln u_{\delta,i})\, (\lambda b_i + \delta\, |q\cdot\nu|)\, ds \\
&\quad - \int\limits_{\partial\Omega_{in}} r f_{r-1}(u_\delta)\, (\mu_i^0 + \ln u_{\delta,i})\, u_{\delta,i}\, |q\cdot\nu|)\, ds \\
&=: I_{bdry,i,1} + I_{bdry,i,2}\,.
\end{aligned}$$

The last integral $I_{bdry,i,2}$ can be estimated, since $-(\mu_i^0+\ln u_{\delta,i})\, u_{\delta,i}$ is obviously bounded from above by a constant depending only on the data μ^0, as

$$I_{bdry,i,2} \le c\, \|f_{r-1}(u_\delta)\|_{L^1(\partial\Omega)} \le c\, \|f_r(u_\delta)\|_{L^1(\partial\Omega)}.$$

It remains to estimate $I_{bdry,i,1}$. For this, let us decompose the boundary $\partial\Omega_{in}$ into

$$\partial\Omega_{in^+,i} := \{x \in \partial\Omega_{in} : \mu_i^0+\ln u_{\delta,i} \ge 0\}$$
$$\partial\Omega_{in^-,i} := \{x \in \partial\Omega_{in} : \mu_i^0+\ln u_{\delta,i} < 0\}.$$

On $\partial\Omega_{in^-,i}$ the integrand of $I_{bdry,i,1}$ is nonpositive and can be estimated by zero. (Here the nonnegativity of b_i was used.) We obtain

$$I_{bdry,i,1} \le \int\limits_{\partial\Omega_{in^+,i}} (b_i+\delta|q\cdot\nu|)\, r f_{r-1}(u_\delta)\, (\mu_i^0+\ln u_{\delta,i})\, ds. \tag{3.62}$$

By using (3.15) we obtain that

$$\ln u_{\delta,i} \le u_{\delta,i} \le g_i(u_{\delta,i}) \quad \text{and} \quad 1 \le e^{\mu_i^0}(e-1)^{-1}\, g_i(u_{\delta,i});$$

hence, with $c := \max\limits_{i=1,...,I} 1+|\mu_i^0|\, e^{\mu_i^0}(e-1)^{-1}$,

$$\mu_i^0 + \ln u_{\delta,i} \le c\, g_i(u_{\delta,i}) \le c\, g(u_\delta)$$

follows. We obtain

$$I_{bdry,i,1} \le c \int\limits_{\partial\Omega_{in^+,i}} (b_i+\delta|q\cdot\nu|)\, r f_{r-1}(u_\delta) g(u_\delta)\, ds, \quad i = 1, ..., I.$$

Since $f_{r-1}g = f_r$ and since b_i and $q \cdot \nu$ were assumed to be bounded in $L^\infty(\partial\Omega_{in})$, we obtain the estimate

$$I_{bdry,i,1} \le c\, \|f_r(u_\delta)\|_{L^1(\partial\Omega_{in})},$$

hence,

$$I_{bdry,i} \le c\, \|f_r(u_\delta)\|_{L^1(\partial\Omega)},$$

which can be estimated such as the boundary term in the proof of Lemma 3.7, part (v).

3.3.5 Continuity and compactness of the fixed point operator

Lemma 3.10 *The mapping $\mathcal{Z}$ defined in (3.29)-(3.30) is continuous and compact.*

Proof. Let us focus on the proof of compactness, since the proof of continuity is analogue.
Let (v^n) be a bounded sequence in $W_p^{2,1}(Q_T)^I$, $p>n+1$. From the compact embedding (3.25) we know that (v^n) has a subsequence which is convergent in $C(\overline{Q}_T)^I$. Let us denote this subsequence again by (v^n). Hence, also $SR(v^n)$ is convergent in $C(\overline{Q}_T)^I$ and therefore also in $L^p(Q_T)^I$. From the linear parabolic theory (Lemma 5.4) we conclude that the sequence $u^n = \mathcal{Z}(v^n)$ is convergent in $W_p^{2,1}(Q_T)^I$. ∎

3.3.6 Existence and Uniqueness of the global solution

Theorem 3.11 *There is a unique solution of problem* **(P)** *(i.e., (3.22)).*

Proof. Applying Schaefer's fixed point theorem (Theorem 5.1) we get, thanks to Cor. 3.9 with $s = p$ and Lemma 3.10 the existence of (at least) one solution of problem $\mathbf{(P^+)}$. This solution is, due to Lemma 3.2, obviously also a solution of problem **(P)**. Let us show now that this solution of **(P)** is *the only* solution of **(P)**.
Let $u^1, u^2 \in W_p^{2,1}(Q_T)^I$ be two solutions of **(P)** and $\tilde{u} = u^1 - u^2$. We know that the u_i^k and all powers $(u_i^k)^s$, $k = 1, 2$, $i = 1, ..., I$, $s \in \mathbb{N}_0$ and all products of such powers are in $L^\infty(Q_T)$. We test the equations for u_i^1 and u_i^2 with $\tilde{u}_i$ on $[0, t] \times \Omega$ and take the difference:

$$\frac{1}{2}\frac{d}{dt}\|\tilde{u}_i\|^2_{L^2(Q_t)^I} + \mathcal{K}\int_0^t\int_\Omega |\nabla\tilde{u}_i|_n^2\,dx\,dt$$
$$= -\sum_{i=1}^I \int_0^t\int_\Omega q\cdot\nabla\tilde{u}_i\;\tilde{u}_i + \sum_{j=1}^J s_{ij}\int_0^t\int_\Omega [R_j(u^1) - R_j(u^2)]\,\tilde{u}_i\,dx\,dt$$

The advective term is treated in a standard way (as in the proof of Lemma 3.2):

$$-\sum_{i=1}^I \int_0^t\int_\Omega q\cdot\nabla\tilde{u}_i\;u_i \le \sum_{i=1}^I Q\,\|\nabla\tilde{u}_i\|_{L^2(\Omega)}\|\tilde{u}_i\|_{L^2(\Omega)} \le \frac{Q^2}{2\mathcal{K}}\,\|\tilde{u}\|^2_{L^2(\Omega)} + \frac{\mathcal{K}}{2}\,\|\nabla\tilde{u}\|^2_{L^2(\Omega)};$$

hence,

$$\frac{1}{2}\frac{d}{dt}\|\tilde{u}_i\|^2_{L^2(Q_t)} + \frac{\mathcal{K}}{2}\int_0^t\int_\Omega |\nabla\tilde{u}_i|^2_n\,dx\,dt$$

$$\leq \frac{Q^2}{2\mathcal{K}}\,\|\tilde{u}\|^2_{L^2(\Omega)} + \sum_{j=1}^{J}|s_{ij}|\int_0^t\int_\Omega |R_j(u^1) - R_j(u^2)|\,\tilde{u}_i\,dx\,dt$$

Expanding[19] all the terms of $R_j(u^1) - R_j(u^2)$, every term in $R_j(u^1) - R_j(u^2)$ contains a factor of the structure $u^1_l - u^2_l = \tilde{u}_l$, while the remaining factor $\Lambda(t,x)$ is in $L^\infty(Q_T)$ since all the u^1_k, u^2_k are in $L^\infty(Q_T)$. Applying to each term the extended Hölder inequality with $p = \infty$, $q = 2$, $r = 2$, and estimating $\|\tilde{u}_l\|_{L^2(Q_t)}$ by $\|\tilde{u}\|_{L^2(Q_t)^I}$ we get

$$\frac{1}{2}\frac{d}{dt}\|\tilde{u}_i\|^2_{L^2(Q_t)} + \frac{\mathcal{K}}{2}\int_0^t\int_\Omega |\nabla\tilde{u}_i|^2_n\,dx\,dt \leq \frac{Q^2}{2\mathcal{K}}\,\|\tilde{u}\|^2_{L^2(\Omega)} + c\,\|\tilde{u}_i\|_{L^2(Q_t)}\|\tilde{u}\|_{L^2(Q_t)^I}$$

where c depends on u^1, u^2. Summing up over $i = 1, ..., I$ and dropping the dissipative term we obtain

$$\frac{1}{2}\frac{d}{dt}\|\tilde{u}\|^2_{L^2(Q_t)^I} \leq c\,\|\tilde{u}\|^2_{L^2(Q_t)^I}.$$

The Gronwall Lemma and the initial value $\tilde{u}(0) = 0$ lead to $\|\tilde{u}(t)\|^2_{L^2(Q_t)^I} = 0$ for all $t \in [0,T]$. ∎

3.4 The coupled problem of mobile and immobile species (PDE-ODE-problem)

In this section we will extend the model of Sec. 3.3 to coupled problems consisting of PDEs for mobile species and ODEs for immobile species. The Sec. 3.4.1 illustrates the difficulties which are caused by this coupling, concerning the existence proof. This will motivate the assumption of Sec. 3.4.2 on the heterogeneous reactions (i.e., the reactions between the mobile and immobile species).

[19] again in the sense of Footnote 14, p. 87

3.4.1 Introduction and a first attempt

We are considering a problem with I mobile species with concentrations $u_1, ..., u_I$ and $\bar{I}$ immobile species with concentration $\bar{u}_1, ..., \bar{u}_{\bar{I}}$ and arbitrary mass action kinetics between all mobile and immobile species

$$\begin{pmatrix} \partial_t u + Lu \\ \partial_t \bar{u} \end{pmatrix} = SR(u, \bar{u}).$$

As in the pure PDE case of Sec. 3.3 we have to define a fixed point operator and to prove an a priori estimate by using a Lyapunov functional. Let us for the moment consider the construction of the Lyapunov functional in the situation that we allow rather *arbitrary* kinetic mass action reactions between the mobile and the immobile species, like in Chapter 2, i.e., the stoichiometrix coefficients s_{ij} are arbitrary large numbers for all $i = 1, ..., I+\bar{I}$, $j = 1, ..., J$. A first try could be to define it as the space integral of the r-th power of the energy density, only now with respect to all the mobile *and* the immobile species, i.e.,

$$\begin{aligned} &f_r(u, \bar{u}) := g(u, \bar{u})^r, \qquad g(u, \bar{u}) := \sum_{i=1}^{I} g_i(u_i) + \sum_{i=1}^{\bar{I}} \bar{g}_i(\bar{u}_i), \\ &g_i(u_i) := (\mu_i^0 - 1 + \ln u_i) + \exp(1 - \mu_i^0), \quad i = 1, ..., I, \\ &\bar{g}_i(\bar{u}_i) = (\mu_{I+i}^0 - 1 + \ln \bar{u}_i) + \exp(1 - \mu_{I+i}^0), \quad i = 1, ..., \bar{I}, \end{aligned} \tag{3.63}$$

the vector $\mu^0 \in \mathbb{R}^{I+\bar{I}}$ being a solution of the linear system $S^T \mu^0 = -\ln K$. Following the proof of Lemma 3.7, we would obviously get

$$\int_{t_1}^{t_2} \int_{\Omega} \langle \partial f_r(u_\delta), \partial_t u_\delta \rangle_{I+\bar{I}} \, dx \, dt = \int_{t_1}^{t_2} \int_{\Omega} \langle \partial f_r(u_\delta), \partial_t u \rangle_{I+\bar{I}} \, dx \, dt = \mathcal{I}_{diff} + \mathcal{I}_{adv} + \mathcal{I}_{react}$$

where the diffusive and the advective term obviously covers only the mobile species and the reactive terms covers both mobile and immobile species:

$$\begin{aligned} \mathcal{I}_{diff} &= -\mathcal{K} \sum_{k=1}^{n} \int_{t_1}^{t_2} \int_{\Omega} \langle \partial_{x_k} (\partial f_r(u_\delta)), \partial_{x_k} u \rangle_I \, dx \, dt, \\ \mathcal{I}_{adv} &= -\int_{t_1}^{t_2} \int_{\Omega} \sum_{i=1}^{I} (\partial f_r(u_\delta))_i \, \langle \nabla_x u_i, q \rangle_n \, dx \, dt \end{aligned}$$

$$\mathcal{I}_{react} \;=\; \lambda \int_{t_1}^{t_2} \int_{\Omega} \langle \partial f_r(u_\delta), S\tilde{R}(u) \rangle_{I+\bar{I}} \, dx \, dt.$$

Concerning the reactive term $\mathcal{I}_{react}$, exactly the same treatment as in part (iii) of that proof can be used here again. However, the treatment of the diffusive term $\mathcal{I}_1$ becomes problematic. Formally, we get

$$\mathcal{I}_{diff} = -\mathcal{K} \sum_{k=1}^{n} \sum_{i=1}^{I} \int_{t_1}^{t_2} \int_{\Omega} \partial_{x_k} u_i \, \partial_{x_k} \left[\partial f_r(u_\delta) \right]_i \, dx \, dt$$

$$= -\mathcal{K} \sum_{k=1}^{n} \sum_{i=1}^{I} \int_{t_1}^{t_2} \int_{\Omega} \partial_{x_k} u_i \; \partial_{x_k} [r g(u_\delta)^{r-1} (\mu_i^0 + \ln u_{\delta,i})] \, dx \, dt$$

$$= -\mathcal{K} \left(\sum_{k=1}^{n} \sum_{i=1}^{I} \int_{t_1}^{t_2} \int_{\Omega} r g(u_\delta)^{r-1} \, \partial_{x_k} u_i \, \frac{1}{u_i + \delta} \, \partial_{x_k} u_i \, dx \, dt \right.$$

$$\left. + \sum_{k=1}^{n} \sum_{i=1}^{I} \sum_{j=1}^{I+\bar{I}} \int_{t_1}^{t_2} \int_{\Omega} r \, (r-1) \, g(u_\delta)^{r-2} \, \partial_{x_k} u_i \, (\mu_i^0 + \ln u_{\delta,i}) \; (\mu_j^0 + \ln u_{\delta,j}) \, \partial_{x_k} u_j \, dx \, dt \right)$$

We note: On the one hand, $\mathcal{I}_{diff}$ only contains viscosity terms for the mobile species, on the other hand $\partial f_r(u)$ is a function depending also on the immobile species, having the consequence that the differentiation of the test function $\partial f_r(u)$ with respect to x leads to partial derivatives with respect to mobile and immobile species! Besides the fact that this seems quite unnatural (the function space for the immobile species concentrations would have to include some regularity with respect to space), the most disturbing result is that the last term is nonsymmetric, i.e., we cannot be sure that $\mathcal{I}_{diff}$ is nonpositive. Let us note that for $r = 1$, the nonsymmetric term vanishes. This means that an a priori estimate using $f_1 = g$ is possible. This, however, leads only to very weak $L^1(\Omega)$, $L^\infty(0,T;L^1(\Omega))$ a priori bounds for u_i, $\bar{u}_i$ and for $u_\delta \ln u_\delta$, $\bar{u}_\delta \ln \bar{u}_\delta$, which are not sufficient for problem with highly nonlinear reactive terms. [20]

[20] The latter bounds can be used by a generalized kind of Gagliardo-Nirenberg inequality (see [59] Satz A.8.7. and [12]) to get an $L^\infty(0,T;L^2(\Omega))$-estimate of u_i and $\bar{u}_i$, which is, however, still not sufficient for problems with *highly* nonlinear reactive terms.

Let us mention that the same difficulty occurs when we consider only mobile species u_i (as in Sec. 3.3), but with *individual* viscosities $\mathcal{K}_i$ for each species, treated with the Lyapunov functional f_r from (3.63). Then the advective term would read

$$\mathcal{I}_{diff} = -\sum_{k=1}^{n}\sum_{i=1}^{I}\int_{t_1}^{t_2}\int_{\Omega} r g(u)^{r-1}\,\partial_{x_k}u_i\,\frac{\mathcal{K}_i}{u_{\delta,i}}\,\partial_{x_k}u_i\,dx\,dt$$

$$-\sum_{k=1}^{n}\sum_{i=1}^{I}\sum_{j=1}^{I}\int_{t_1}^{t_2}\int_{\Omega} r\,(r-1)\,g(u)^{r-2}\,\partial_{x_k}u_i\,(\mu_i^0+\ln u_{\delta,i})\,\mathcal{K}_i\,(\mu_j^0+\ln u_{\delta,j})\,\partial_{x_k}u_j\,dx\,dt$$

which is again nonsymmetric and of unknown sign as soon as $r>1$.
The physical meaning of these difficulties seems to be as follows: The chemical energy measures how far the species distribution is from the equilibrium at each point of the computational domain. The mass action law reactions (fortunately) have the tendency to decrease this chemical energy. As long as there is only a species-independent spatial diffusion/advection, this property is not much disturbed, as seen in the proof of Lemma 3.7. However, as soon as there is a *species-dependent* spatial transport of concentrations, this transport may locally increase the distance of the species assemblage from equilibrium, which might lead to a non-monotoneous 'energy' in the sense of F_r with $r>1$, requiring a different functional.

3.4.2 Definition of the problem and the Lyapunov functional

To avoid the difficulties explained in Section 3.4.1 we are going to use the following stategy: We define two different energy functions, one correlated to the mobile species and to the reactions among them, and one correlated to the immobile species and the reactions among them. Heterogeneous reactions, i.e., reactions between mobile and immobile species, are ignored concerning the definition of the two energy functionals. The separate treatment of mobile and immobile species circumvents the difficulties explained in Sec. 3.4.1. However, the fact that heterogeneous reactions are ignored for the construction of the Lyapunov functionals has the following drawback: With this strategy we can only allow for heterogeneous reactions with *linear* rates. Linear reactions $\mathrm{A_i} \longleftrightarrow \bar{\mathrm{A}}_\mathrm{j}$, between a mobile species $\mathrm{A_i}$ and an immobile

species $\bar{\mathrm{A}}_\mathrm{j}$ would cause, following mass action kinetics, source terms

$$\begin{aligned} f_i &= -\alpha u_i + \beta \bar{u}_j \quad \text{for } \mathrm{A_i} \text{ and} \\ \bar{f}_j &= -\beta \bar{u}_j + \alpha u_i \quad \text{for } \bar{\mathrm{A}}_\mathrm{j}. \end{aligned} \tag{3.64}$$

with $\alpha, \beta \geq 0$. To cover these linear heterogeneous reactions we allow for the following source terms:

$$X_i = X_i(u_i, \bar{u}) = a_0 - a_i u_i + \sum_{j=1}^{\bar{I}} a_{ij} \bar{u}_j \tag{3.65-a}$$

for u_i, $i = 1, ..., I$;

$$\bar{X}_i = \bar{X}_i(\bar{u}_i, u) = \bar{a}_0 - \bar{a}_i \bar{u}_i + \sum_{j=1}^{I} \bar{a}_{ij} u_j \tag{3.65-b}$$

for $\bar{u}_i$, $i = 1, ..., \bar{I}$, with given constants[21] $a_0, \bar{a}_0, a_i, \bar{a}_i, a_{ij}, \bar{a}_{ij} \in \overline{\mathbb{R}}^+$.
Let the problem parameters such as Ω, q, u_0, p, L, etc., be defined as in Section 3.3.1. Let us define the space

$$\mathcal{C}(\overline{Q}_T) := \{v \in C(\overline{Q}_T) \,|\, \partial_t v \in C(\overline{Q}_T)\}$$

with the norm

$$\|v\|_{\mathcal{C}(\overline{Q}_T)} := \|v\|_{L^\infty(Q_T)} + \|\partial_t v\|_{L^\infty(Q_T)}.$$

Additionally, let initial values $u_0 \in W_p^{2-2/p}(\Omega)^I$, $u_0 \geq 0$, for the mobile species and $\bar{u}_0 \in C^0(\bar{\Omega})^{\bar{I}}$, $\bar{u}_0 > 0$, for the immobile species be given. We are considering the following problem, denoted by $(\overline{\mathbf{P}})$:
Find a solution $(u, \bar{u}) \in W_p^{2,1}(Q_T)^I \times \mathcal{C}(\overline{Q}_T)^{\bar{I}}$ of the problem

$$\begin{aligned} \partial_t u + Lu &= S^{mob} R^{mob}(u) + X(u, \bar{u}) && \text{on } Q_T, \\ \partial_t \bar{u} &= S^{immo} R^{immo}(\bar{u}) + \bar{X}(\bar{u}, u) && \text{on } Q_T, \\ u(0, x) &= u_0 && \text{on } \Omega, \\ \bar{u}(0, x) &= \bar{u}_0 && \text{on } \bar{\Omega}, \\ \tfrac{\partial u}{\partial \nu} &= 0 && \text{on } (0, T] \times \partial\Omega \end{aligned} \tag{3.66}$$

[21] The nonnegative sign of the constants a_0, a_{ij}, $\bar{a}_0$, $\bar{a}_{ij}$, which cause $X_i(0, \bar{u}) \geq 0$, $\bar{X}_i(0, u) \geq 0$ for all u, $\bar{u}$, is important in order to guarantee the nonnegativity of solutions, Sec. 3.4.3. The exchange terms $X, \bar{X}$ do not necessarily have to comply with mass conservation rules for what will follow.

with the homogeneous reactions $R^{mob}(u)$, $R^{immo}(\bar{u})$ and the coupling terms $X, \bar{X}$ defined by (3.65). S^{mob}, S^{immo} are the $I \times J_{mob}$-, $\bar{I} \times J_{immo}$-matrices of (integer) stoichiometric constants for J_{mob} linearly independent reactions among the mobile species and J_{immo} linearly independent reactions among the immobile species;

$$S = \left(\begin{array}{c|c} S^{mob} & 0 \\ \hline 0 & S^{immo} \end{array}\right), \quad R = \begin{pmatrix} R^{mob} \\ R^{immo} \end{pmatrix}.$$

For later use let us define the Lyapunov functionals. For the mobile subsystem, we define the functional with respect to the reactions R^{mob} and the stoichiometry S^{mob}; for the immobile subsystem, we define a functional with respect to R^{immo} and S^{immo}:
As in Sec. 3.3.4, let $\mu^0 \in \mathbb{R}^I$ be a solution of the linear system

$$(S^{mob})^T \mu^0 = -\ln K,$$

$K \in \mathbb{R}_+^{J_{mob}}$ being the vector of equilibrium constants for the J_{mob} reactions R^{mob}. Analogously, let $\bar{\mu}^0 \in \mathbb{R}^{\bar{I}}$ be a solution of the linear system

$$(S^{immo})^T \bar{\mu}^0 = -\ln \bar{K},$$

$\bar{K} \in \mathbb{R}_+^{J_{immo}}$ being the vector of equilibrium constants for the J_{immo} reactions R^{immo}. Now we define g_i, g, f_r, G, F_r exactly as in Sec. 3.3.4, i.e., all these functions are related to the mobile subsystem. Then we define in analogy to (3.31), (3.32) the functions

$$\begin{aligned}
&\bar{g}_i : \overline{\mathbb{R}}_+ \longrightarrow \mathbb{R}, && \bar{g}_i(\bar{u}_i) = (\bar{\mu}_i^0 - 1 + \ln \bar{u}_i)\, \bar{u}_i + \exp(1 - \bar{\mu}_i^0), \\
&\bar{g} : \overline{\mathbb{R}}_+^I \longrightarrow \mathbb{R}, && \bar{g}(\bar{u}) = \sum_{i=1}^{\bar{I}} \bar{g}_i(\bar{u}_i); \\
&\bar{f}_r : \overline{\mathbb{R}}_+^I \longrightarrow \mathbb{R} && \bar{f}_r(\bar{u}) = [\bar{g}(\bar{u})]^r, \\
&\bar{F}_r : L_+^\infty(\Omega)^I \longrightarrow \mathbb{R}, && \bar{F}_r(\bar{u}) = \int\limits_\Omega \bar{f}_r(\bar{u}(x))\, dx,
\end{aligned} \tag{3.67}$$

$r \in \mathbb{N}$. The bars atop the entities indicate their relation to the immobile subsystem. The properties (3.15) and (3.16) of g also apply to $\bar{g}$, of course. To get an a priori estimate, we are going to consider $F_r(u) + \bar{F}_r(\bar{u})$ in Sec. 3.4.5.

3.4.3 Nonnegativity of solutions

As in the PDE case of Sec. 3.3, we also consider problem $(\overline{\mathbf{P}})$ with modified rates $\tilde{R}(u) = R(u^+)$, and analogously $\tilde{X}(u,\bar{u}) = X(u^+,\bar{u}^+)$, $\tilde{\bar{X}}(\bar{u},u) = \bar{X}(\bar{u}^+,u^+)$ instead of $R, X, \bar{X}$. The so modified problem is called problem $(\overline{\mathbf{P}^+})$. The proofs of positivity/nonnegativity of the previous sections directly apply:

Lemma 3.12 *Assume that there is a $\tau>0$ and a function $(u,\bar{u}) \in W_p^{2,1}(Q_\tau)^I \times \mathcal{C}(Q_\tau)^{\bar{I}}$, $n+1<p<\infty$ being a solution of $(\overline{\mathbf{P}^+})$ on the time interval $[0,\tau)$. Then*

$$u \geq 0, \quad \bar{u} > 0$$

holds on $\overline{Q}_\tau$.

Proof. The positivity of the *immobile* entities $\bar{u}_i$ for each $x \in \bar{\Omega}$ follows like in the proof of Lemma 3.1 (the α_0, α_r now depending on the $\bar{u}_j$, $j \neq i$, and the u_j). The nonnegativity of the mobile entities u_i follows as in the proof of Lemma 3.2; the immobile $\bar{u}_j^+(t,x)$ enter as factors of known nonnegative sign. ∎

3.4.4 The fixed point operator

Our objective is to prove the existence of a solution of problem $(\overline{\mathbf{P}^+})$, defined in Sec. 3.4.3, using again Schaefer's fixed point theorem. Let us define the fixed point operator.
Let $n+1<p<\infty$ and $T>0$ be fixed. Let us define the operator $\bar{\mathcal{Z}}$,

$$\begin{array}{rcl} \bar{\mathcal{Z}} : W_p^{2,1}(Q_T)^I & \longrightarrow & W_p^{2,1}(Q_T)^I, \\ v & \longmapsto & u = \bar{\mathcal{Z}}(v) \end{array} \tag{3.68}$$

by u being the solution of the problem

$$\begin{array}{lcll} \partial_t u + Lu & = & S^{mob}\tilde{R}^{mob}(v) + \tilde{X}(v,\bar{w}) & \text{on } Q_T, \\ \partial_t \bar{w} & = & S^{immo}\tilde{R}^{immo}(\bar{w}) + \tilde{\bar{X}}(\bar{w},v) & \text{on } Q_T, \\ u(0,x) & = & u_0 & \text{on } \Omega, \\ \bar{w}(0,x) & = & \bar{u}_0 & \text{on } \bar{\Omega}, \\ \frac{\partial u}{\partial \nu} & = & 0 & \text{on } (0,T]\times\partial\Omega, \end{array} \tag{3.69}$$

with $u_0 \in W_p^{2-2/p}(\Omega)^I$, $\bar{u}_0 \in C^0(\bar{\Omega})^{\bar{I}}$, $u_0 \geq 0$, $\bar{u} > 0$.

Let us verify that $\bar{\mathcal{Z}}$ is well defined:
The problem (3.69) obviously decouples into an ODE system for $\bar{w}$, and, with known $\bar{w}$, a PDE system for u.
Let us consider the ODE system first. Since $v^+ \in C(\overline{Q}_T)^I$ and $v^+ \geq 0$, the application of the following Lemma 3.13 (with v^+ playing the role of v in the lemma) proves that $\bar{w}$ exists on the whole interval $[0,T]$, i.e., $\bar{w}(\cdot,x) \in C([0,T])^{\bar{I}}$ for each fixed $x \in \bar{\Omega}$. We have to argue that $\bar{w} \in C(\overline{Q}_T)^{\bar{I}}$, i.e., that the solution $\bar{w}$ also depends continuously on $x \in \bar{\Omega}$. For this we start by stating that the right hand side of the ODE of (3.69) can be written as a sum $\psi_1(\bar{w}^+)+\psi_2(v^+(t,x))$, where ψ_2 is linear due to the definition of $\bar{X}$, and where ψ_1 is a polynomial with coefficients independent of x, t. Since $\bar{w}$ is bounded in $L^\infty(Q_T)^{\bar{I}}$ due to Lemma 3.13, $\bar{w} \mapsto \psi_1(\bar{w})$ (although being a polynomial) has a global Lipschitz constant for all values of $\bar{w}^+(t,x)$ which are possible arguments. v^+ is continuous, hence uniformly continuous on the compact set $\overline{Q}_T$, hence $v^+(\cdot,x)$ as an element of $L^\infty([0,T])^I$ depends continuously on x. The initial value $\bar{u}_0(x)$ also depends continuously on $x \in \bar{\Omega}$. It is well known (see for example [40] Chapter 3 Satz 13.1) that with the existence of a Lipschitz constant for the right hand side $\psi_1(\bar{w}^+) + \psi_2(v^+(t,x))$ with respect to $\bar{w}$, independent of x, t, the solution of an ODE on $[0,T]$ depends continuously on perturbations of the right hand side and on perturbations of the initial value. As a result, $\bar{w} \in C(\overline{Q}_T)^{\bar{I}}$ holds.
We conclude that the whole right hand side of the ODE, $S^{immo}\tilde{R}^{immo}(\bar{w}) + \tilde{\bar{X}}(\bar{w},v)$, lies in $C(\overline{Q}_T)^{\bar{I}}$. Hence,

$$\bar{w} \in \mathcal{C}(\overline{Q}_T)^{\bar{I}}.$$

With the stated regularity of $\bar{w}$ and v, the right hand side of the *PDE* subsystem, $S^{mob}\tilde{R}(v)+\tilde{X}(v,\bar{w})$, is a $C(\overline{Q}_T)^I$-function of (t,x). In particular, it lies in $L^p(\overline{Q}_T)$ for arbitrary p. Using the linear parabolic theory as in Sec. 3.3.3, we get a solution $u \in W_p^{2,1}(Q_T)^I$.
Hence, $\bar{Z}$ is well defined.
Further we state that obviously every fixed point of $\bar{\mathcal{Z}}$ is a solution of problem $(\overline{\mathbf{P}^+})$.
It remains to prove the boundedness of the solution $\bar{w}$ of the ODE subproblem for *given* v:

Lemma 3.13 *Let $[0,\tau)$ be an interval of existence of a solution of the ODE subsystem of (3.69), i.e.*

$$\partial_t \bar{w} = S^{immo}\tilde{R}^{immo}(\bar{w}) + \tilde{\bar{X}}(\bar{w},v), \quad w(0) = \bar{u}_0,$$

for given $v \in C(\overline{Q}_T)$, $v \geq 0$, $\bar{u}_0 \in C(\bar{\Omega})$, $\bar{u}_0 > 0$. *Then there are constants* $c_1, c_2 > 0$, *independent of* $t, x, v, \bar{w}, u_0$ *such that*

$$\bar{w}(t,x) \leq \max_{x \in \bar{\Omega}} \bar{g}(\bar{u}_0(x)) \, \exp((c_1 + c_2 \|v\|_{L^\infty(\overline{Q}_\tau)^I}) \, t)$$

holds for all $(t,x) \in \overline{Q}_\tau$, *i.e.,* $\bar{w}$ *is bounded on* $\overline{Q_\tau}$. *In particular, the solution* $\bar{w}$ *is global, i.e., it exists on* $\overline{Q}_T$.

Proof. As in the proof of Lemma 3.12 and Lemma 3.1, $\bar{w} > 0$ is shown.[22] We use the Lyapunov functional $\bar{g}$ defined in (3.67) and follow the treatment of (3.17), but now taking into account the additional term $\bar{X}$ from (3.65). For each fixed $x \in \bar{\Omega}$, we get the estimate

$$\begin{aligned} \frac{d}{dt} \bar{g}(\bar{w}(t)) &= \langle -\ln \bar{K} + (S^{immo})^T \ln \bar{w}(t), R^{immo}(\bar{w}(t)) \rangle_J \\ &\quad + \langle \bar{X}(\bar{w}(t), v), \bar{\mu}^0 + \ln \bar{w}(t) \rangle_{\bar{I}} \end{aligned} \tag{3.70}$$

where the first term, as shown in (3.18), is nonpositive. With (3.65-b), it remains to estimate terms of the structure

$$\begin{aligned} \mathcal{I}_1 &:= \bar{a}_0 \, (\bar{\mu}_i^0 + \ln \bar{w}_i(t)), \\ \mathcal{I}_2 &:= -\bar{a}_i \, (\bar{\mu}_i^0 + \ln \bar{w}_i(t)) \, \bar{w}_i(t), \\ \mathcal{I}_3 &:= \bar{a}_{ij} \, (\bar{\mu}_i^0 + \ln \bar{w}_i(t)) \, v_j(t) \end{aligned}$$

where $\bar{a}_0, \bar{a}_i, \bar{a}_{ij} \geq 0$. Let us prove that there are constants c_1, c_2 independent of x, t such that

$$\mathcal{I}_k \leq (c_1 + c_2 \, \|v\|_{L^\infty(Q_T)^{\bar{I}}}) \, \bar{g}(\bar{w}(t)), \quad k = 1,2,3 : \tag{3.71}$$

If $\bar{\mu}_i^0 + \ln \bar{w}_i < 0$ then $\mathcal{I}_1 \leq 0 \leq \bar{g}_i(\bar{w}_i)$ follows obviously, since $\bar{g}_i \geq 0$. Otherwise, i.e., if $\exp(-\bar{\mu}_i^0) \leq \bar{w}_i$, then

$$\mathcal{I}_1 \leq \bar{a}_0 \, (\bar{\mu}_i^0 + \ln \bar{w}_i) \, \bar{w}_i \, \exp(\bar{\mu}_i^0) \leq 2 \bar{a}_0 \exp(\bar{\mu}_i^0) \, \bar{g}_i(\bar{w}_i),$$

where Lemma 5.7 was used. Hence, (3.71) holds for $k = 1$ (with $c_2 = 0$). Now we use the lower bound[23]

$$(\bar{\mu}_i^0 + \ln \bar{w}_i) \, \bar{w}_i \geq -c \tag{3.72}$$

[22] Hence, we can write $R^{immo}, \bar{X}$ instead of $\tilde{R}^{immo}, \tilde{\bar{X}}$ within this proof.

[23] (3.72) is nothing but the nonnegativity of $\bar{g}_i(\bar{w}_i)$

for a c depending only on the data $\bar{\mu}_i^0$ to get

$$\mathcal{I}_2 \leq c\bar{a}_i \leq c\bar{a}_i\, e^{\mu_i^0}(e-1)^{-1}\,\bar{g}_i(\bar{w}_i),$$

where the lower bound for $\bar{g}_i$, cf. (3.15), was used in the last step.
$\mathcal{I}_3$ is treated like $\mathcal{I}_1$ with $\bar{a}_{ij}v_j$ playing the role of $\bar{a}_0$.
So we have proven (3.71).
Applying the estimates to (3.70) we obtain the differential inequality

$$\frac{d}{dt}\bar{g}(\bar{w}(t)) \leq c\;\bar{g}(\bar{w}(t))$$

with

$$c = c_1 + c_2\|v\|_{L^\infty(Q_T)^{\bar{I}}} \tag{3.73}$$

for $c_1, c_2 > 0$ being independent of x, t, v, $\bar{w}$. By using the Gronwall lemma we get

$$\bar{g}(\bar{w}(t,x)) \leq \bar{g}(\bar{u}_0(x))\, e^{ct},$$

hence

$$\bar{w}_i(t,x) \leq \bar{g}(\bar{w}(t,x)) \leq \bar{g}(\bar{u}_0(x))\, e^{ct}. \tag{3.74}$$

Hence, under the assumptions on $\bar{u}_0$, $\bar{w}_i$ is bounded and thus exists globally on the whole interval of existence of the given v. ∎

3.4.5 A priori estimates

We have to construct a bound holding for arbitrary solutions $u \in W_p^{2,1}(Q_T)^I$ of the equation

$$u = \lambda\bar{\mathcal{Z}}(u), \tag{3.75}$$

$\lambda \in [0,1]$, $\bar{\mathcal{Z}}$ defined in (3.68)-(3.69). Solving (3.75) is equivalent to finding a solution of the PDE-/ODE-system

$$\begin{array}{lcll}
\partial_t u + Lu & = & \lambda S^{mob}R^{mob}(u) + \lambda X(u,\bar{u}) & \text{on } Q_T, \\
\partial_t \bar{u} & = & S^{immo}R^{immo}(\bar{u}) + \bar{X}(\bar{u},u) & \text{on } Q_T, \\
u(0,x) & = & \lambda\, u_0 & \text{on } \Omega, \\
\bar{u}(0,x) & = & \bar{u}_0 & \text{on } \bar{\Omega}, \\
\frac{\partial u}{\partial \nu} & = & 0 & \text{on } (0,T]\times\partial\Omega,
\end{array} \tag{3.76}$$

$\lambda \in [0,1]$. Let us denote this problem by $(\overline{\mathbf{P}_\lambda^+})$.
We want to generalize the proceedings of Sec. 3.3.4, where a priori estimates are derived for the pure PDE system, to the coupled PDE-/ODE-system $(\overline{\mathbf{P}^+})$. We start by generalizing the Lemma 3.7, which is (somehow) central in Sec. 3.3.4:

Lemma 3.14 *Let $(u,\bar{u}) \in W_p^{2,1}(Q_T)^I \times \mathcal{C}(\overline{Q}_T)^{\bar{I}}$ be a solution of problem $(\overline{\mathbf{P}_\lambda^+})$ and $\delta > 0$, $u_\delta := u + \delta$, $r \in \mathbb{N}$. Then the estimate*

$$\begin{aligned} &\int_{t_1}^{t_2}\int_\Omega \langle \partial f_r(u_\delta), \partial_t u_\delta \rangle_I \, dx\, dt + \int_{t_1}^{t_2}\int_\Omega \langle \partial \bar{f}_r(\bar{u}), \partial_t \bar{u} \rangle_{\bar{I}} \, dx\, dt \\ &\quad \leq \quad c \left[\int_{t_1}^{t_2} F_r(u_\delta(t))\, dt + \bar{F}_r(\bar{u}(t))\, dx\, dt \right] + h(\delta, u), \end{aligned} \tag{3.77}$$

holds where $c > 0$ (depending on the data) is a constant independent of $u, \bar{u}, \delta$ and where $h(\delta, u) \longrightarrow 0$ for $\delta \longrightarrow 0$.

Proof. Part (a). We know from Lemma 3.12 that $u_i \geq 0$, $\bar{u}_i > 0$.[24] Let us consider the first of the two terms on the left hand side of (3.77). We get exactly the same terms $\mathcal{I}_{react}$, $\mathcal{I}_{diff}$, $\mathcal{I}_{adv}$ as in the proof of Lemma 3.7 which can be estimated by the right hand side of (3.48). Besides, we have the coupling term

$$\int_{t_1}^{t_2}\int_\Omega \langle \partial f_r(u_\delta), X(u,\bar{u}) \rangle_I \, dx\, dt \tag{3.78}$$

which has to be estimated. Using the definition of X, it remains to estimate the space-time-integral over expressions of type

$$\begin{aligned} \mathcal{I}_1 &:= a_0 r f_{r-1}(u_\delta)\,(\mu_i^0 + \ln u_{\delta,i}), \\ \mathcal{I}_2 &:= -a_i r f_{r-1}(u_\delta)\,(\mu_i^0 + \ln u_{\delta,i})\, u_i, \\ &= -a_i r f_{r-1}(u_\delta)\,(\mu_i^0 + \ln u_{\delta,i})\, u_{\delta,i} \\ &\quad + a_i r f_{r-1}(u_\delta)\,(\mu_i^0 + \ln u_{\delta,i})\, \delta, \\ &=: \mathcal{I}_3 + \mathcal{I}_4 \\ \mathcal{I}_5 &:= a_{ij} r f_{r-1}(u_\delta)\,(\mu_i^0 + \ln u_{\delta,i})\, \bar{u}_j, \end{aligned}$$

[24] Hence, we write 'X' instead of '$\tilde{X}$', etc., in (3.78), (3.80).

$a_0, a_i, a_{ij} \geq 0$. Term $\mathcal{I}_4$ can be estimated in the same way as $\mathcal{I}_1$, $a_i\delta$ playing the role of a_0. It remains to estimate $\mathcal{I}_1$, $\mathcal{I}_3$, $\mathcal{I}_5$. Terms $\mathcal{I}_1$ and $\mathcal{I}_3$ can be estimated in the same way as the terms $\mathcal{I}_1$, $\mathcal{I}_2$ in Lemma 3.13. We get for $\mathcal{I}_k$, $k=1,2,3,4$:

$$\mathcal{I}_k \leq c\, r f_{r-1}(u_\delta)\, g(u_\delta) = c\, f_r(u_\delta).$$

with constant $c > 0$ depending on a_0, a_i and the μ_i^0, but independent of $u, \bar{u}, \delta, x$.
Term $\mathcal{I}_5$ has to be treated in a different way than the corresponding term $\mathcal{I}_3$ in the proof of Lemma 3.13 since its dependence on $\bar{u}_j$ is now cruxial, while in the previous proof, v_j had been a *fixed* function. We estimate $\mathcal{I}_5$ as follows:
If $\mu_i^0 + \ln u_{\delta,i} \leq 0$, we may estimate $\mathcal{I}_5 \leq 0$. Therefore we can focus on the case $\mu_i^0 + \ln u_{\delta,i} > 0$ in the following. In the case $\bar{u}_j \leq u_{\delta,i}$ we estimate

$$\mathcal{I}_5 \leq a_{ij} r f_{r-1}(u_\delta)\, (\mu_i^0 + \ln u_{\delta,i})\, u_{\delta,i} \leq 2a_{ij} r f_{r-1}(u_\delta)\, g_i(u_{\delta,i}) \leq 2a_{ij} r f_r(u_\delta)$$

where Lemma 5.7 was used. In the case $u_{\delta,i} < \bar{u}_j$ we define the constant

$$C := \max_{i,j} |\bar{\mu}_i^0 - \mu_j^0|$$

(independent of $u, \bar{u}$), so we obviously can estimate $\mu_i^0 \leq C + \bar{\mu}_j^0$. Hence,

$$\begin{aligned} \mathcal{I}_5 &\leq a_{ij}\, r f_{r-1}(u_\delta)\, (C + \bar{\mu}_j^0 + \ln \bar{u}_j)\, \bar{u}_j \leq a_{ij}\, r f_{r-1}(u_\delta)\, (C\, \bar{u}_j + 2\bar{g}_j(\bar{u}_j)) \\ &\leq a_{ij}\, r f_{r-1}(u_\delta)\, C\, \bar{g}_j(\bar{u}_j) \end{aligned}$$

where in the second step we have used Lemma 5.7 again, and in the last step we have used (3.15).
Summarizing the estimates for $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_5$, we can estimate the term (3.78) by

$$cr \int_{t_1}^{t_2} \int_\Omega f_{r-1}(u_\delta)\, (g(u_\delta) + \bar{g}(\bar{u}))\, dx\, dt,$$

with c only depending on parameters, but independent of $u, \bar{u}, \delta$. Using $f_{r-1} g = g^r = f_r$ and $f_{r-1}\bar{g} = g^{r-1}\bar{g} \leq \max\{g^r, \bar{g}^r\} \leq g^r + \bar{g}^r$ we obtain the bound

$$cr \int_{t_1}^{t_2} \int_\Omega (f_r(u_\delta) + \bar{f}_r(\bar{u}))\, dx\, dt \tag{3.79}$$

for (3.78).

Part (b). It remains to estimate the $\bar{f}_r(\bar{u})$-part of (3.77), which reads

$$\int_{t_1}^{t_2}\int_{\Omega} r\bar{f}_{r-1}(\bar{u})\,\langle -K_{immo} + (S^{immo})^T \ln \bar{u}\,,\, R^{immo}(\bar{u})\rangle_{J_{immo}}\,dx\,dt$$
$$+\int_{t_1}^{t_2}\int_{\Omega} r\bar{f}_{r-1}(\bar{u})\,\langle \bar{\mu}^0+\ln \bar{u}\,,\, \bar{X}(\bar{u},u)\rangle_{\bar{I}}\,dx\,dt. \tag{3.80}$$

As in the proof of (3.17) we know that the first term can be estimated by zero. Following part (a) of this proof we can estimate the second term of (3.80) by a term of the shape (3.79). ∎

As a consequence of the previous lemma, we get a result which corresponds to Theorem 3.3 in the pure PDE case:

Lemma 3.15 *Let $0 \le t_1 < t_2 \le T$ and $p > n+1$ and $0 \le \lambda \le 1$ be given. Let $u \in W_p^{2,1}(Q_T)^I$ be a solution of $(\overline{\mathbf{P}_\lambda^+})$. Then the estimate*

$$F_r(u(t_2)) + \bar{F}_r(\bar{u}(t_2)) \le e^{c\,(t_2-t_1)}\,(F_r(u(t_1)) + \bar{F}_r(\bar{u}(t_1))) \tag{3.81}$$

holds for all $r \in \mathbb{N}$, where the constant $c > 0$ depends only on the problem parameters $\Omega, \mathcal{K}, u_0, \bar{u}_0$, etc., and on r, but not *on $u, \bar{u}, \lambda$. If additionally condition (3.39) holds, then estimate (3.81) holds with $c = 0$; i.e., $F_r(u) + \bar{F}_r(\bar{u})$ is monotone along solutions $(u, \bar{u})$.*

Proof. Exactly as in the proof of Theorem 3.3 we state that (3.57) holds. The application of Lemma 3.7 is replaced by the application of Lemma 3.14 we get estimate (3.59), but with $F_r(u_\delta)$ replaced by $F_r(u_\delta)+\bar{F}_r(\bar{u})$. Following the rest of the proof of Theorem 3.3 we get the desired result. ∎

As a consequence, a priori estimates like in Cor. 3.8 and Cor. 3.9 hold:

Corollary 3.16 (A priori estimate) *Let $p > n+1$ and $1 \le s < \infty$ be fixed. Then there is a constant $c > 0$ (depending on s and u_0 and other parameters of problem $(\overline{\mathbf{P}_\lambda^+})$, but independent of u, $\bar{u}$ and $\lambda \in [0,1]$) such that for an arbitrary solution $(u, \bar{u}) \in W_p^{2,1}(Q_T)^I \times \mathcal{C}(\overline{Q}_T)^{\bar{I}}$ of problem $(\overline{\mathbf{P}_\lambda^+})$ the estimates*

$$\|u\|_{L^\infty(0,T;L^s(\Omega))^I} + \|\bar{u}\|_{L^\infty(0,T;L^s(\Omega))^{\bar{I}}} \le c, \tag{3.82}$$

$$\|u\|_{W_s^{2,1}(Q_T)^I} \le c \tag{3.83}$$

holds for all $1 \le s < \infty$. In particular, $u \in W_s^{2,1}(Q_T)^I$ holds.

Proof. The proof of estimate (3.82) is a straight-forward generalization of the proof of Cor. 3.8.
The bounds for u and for $\bar{u}$ from (3.82) are used to prove the estimate (3.83) exactly as Cor. 3.9. ∎

3.4.6 Compactness of the fixed point operator and existence of a global solution

Theorem 3.17 *The operator $\bar{\mathcal{Z}}$ defined in (3.68) is continuous and compact.*

Proof. Let (v_n) be a sequence bounded in $W_p^{2,1}(Q_T)^I$. We have to prove that there is a subsequence such that $u_n = \bar{\mathcal{Z}}(v_n)$ is convergent in $W_p^{2,1}(Q_T)^I$. Due to the compact embedding (3.25) there is a subsequence, again denoted by (v_n), which is convergent in $C(\overline{Q}_T)^I$. In the following we want to prove that the solution $\bar{w}_n$ of the ODE subproblem

$$\partial_t \bar{w}_n = S^{immo}\tilde{R}^{immo}(\bar{w}_n) + \tilde{\bar{X}}(\bar{w}_n, v_n), \quad \bar{w}_n(0) = \bar{u}_0, \tag{3.84}$$

depends continuously on v_n in the $C(\overline{Q}_T)$-sense, because this continuous dependence is sufficient for the convergence of $(\bar{w}_n)$ in $C(\overline{Q}_T)^{\bar{I}}$. For this, we start by stating that the right hand side of (3.84) can be written as a sum $\psi_1(\bar{w}_n^+) + \psi_2(v_n^+)$, where ψ_2 is linear due to the definition of $\bar{X}$, and where ψ_1 is a polynomial with coefficients independent of x, t. For the desired continuous dependence of $\bar{w}_n$ on v_n, it is sufficient that ψ_1 fulfills a Lipschitz condition with Lipschitz constant independent of v_n, $\bar{w}_n$, x, t (see for example [40] Chapter 3 Satz 13.1). Indeed, such a Lipschitz constant for ψ_1 can be given (although ψ_1 is a polynomial), since all possible arguments $\bar{w}_n(t,x)$ of ψ_1 are, since the sequence $(\bar{w}_n)$ is bounded in $L^\infty(Q_T)^{\bar{I}}$, taken from a compact set. The boundedness of the sequence $(\bar{w}_n)$ in $L^\infty(Q_T)^{\bar{I}}$ follows from Lemma 3.13, in particular from (3.74) with (3.73), and with the boundednes of (v_n) in $L^\infty(Q_T)^I$.
All in all, we have proven that $(\bar{w}_n)$ is convergent in $C(\overline{Q}_T)$. Since now both (v_n) and $(\bar{w}_n)$ converge in $C(\overline{Q}_T)$, so does the right hand side of the PDE

$$\partial_t u_n + Lu_n = S^{mob}\tilde{R}^{mob}(v_n) + \tilde{X}(v_n, \bar{w}_n).$$

In particular, the right hand side converges in $L^p(\overline{Q}_T)$ for arbitrary p. From the linear parabolic theory we know that the sequence of solutions (u_n) then converges in $W_p^{2,1}(Q_T)^I$, $n+1<p<\infty$. ∎

Theorem 3.18 *The problem* $(\overline{\mathbf{P}^+})$ *has a solution, and this solution also solves* $(\overline{\mathbf{P}})$.

Proof. We apply Schaefer's Fixed Point Theorem 5.1 to the operator $\bar{\mathcal{Z}}$, using the a priori estimate (3.83) with $s = p$ and Theorem 3.17, to obtain a solution of $(\overline{\mathbf{P}^+})$, which is nonnegative due to Lemma 3.12, hence, solves $(\overline{\mathbf{P}})$. ∎

The uniqueness can be tackled as in Sec. 3.3.6.

3.5 Extension to problems with equilibrium conditions ('DAE case')

As already mentioned before, often some of the reactions are such fast (compared to the transport or to slower reactions in the system) that they cannot be resolved properly by numerical simulations. Often, the ratio of reaction rate coefficients varies by many powers of ten, covering characteristic time scales from seconds to years. Besides the impossibility to resolve the fast reactions by a time stepping scheme, large rate coefficients lead to very stiff problems. Therefore it is common to express 'sufficently fast' reactions by equilibrium conditions.

The intention of this section is the analysis of such coupled kinetic-equilibrium reactive transport problems, based on the existence and regularity results of the PDE (i.e., purely kinetic) case of Sec. 3.3. The usual procedure to derive equilibrium conditions is the following: Let us assume that all reactions have a forward-backward mass action shape. The forward and the backward rate coefficient of each fast reaction are assumed to go to infinity in such a way that its ratio remains constant. This can be expressed in the following way: Let us split the rate vector $R(u)$ into a part of slow (kinetic) reactions $R_{kin}(u)$ and a part of fast (equilibrium) reactions $R_{eq}(u)$:

$$R = \begin{pmatrix} R_{eq} \\ R_{kin} \end{pmatrix}$$

All rate coefficients k_j^f, k_j^b in R_{kin} and R_{eq} are fixed, and the chemical source term for the system is assumed to be

$$S_{kin} R_{kin}(u) + k\, S_{eq} R_{eq}(u),$$

where $k \to \infty$. We will show that the sequence of solutions u^k will converge weakly in $L^p(Q_T)^I$, and that the limit function fulfils a certain limit problem which consists of partial differential equations and algebraic equilibrium conditions. Let us formalize this procedure:
Let us consider a sequence of real positive numbers (k_m) with $\lim_{m\to\infty} k_m = \infty$. Let us consider the following sequence of problems: Let $T > 0$, $p > n+1$, $u_0 \in W_p^{2-2/p}(\Omega)^I$, $u_0 \geq 0$, and let the columns of $S = (S_{eq}|S_{kin})$ be linearly independent, and let R be as in (3.2)-(3.3).
Find $u \in W_p^{2,1}(Q_T)^I$ such that

$$\begin{aligned} \partial_t u^m + Lu^m &= k_m\, S_{eq} R_{eq}(u^m) + S_{kin} R_{kin}(u^m) \quad \text{on } Q_T, && \text{(3.85-a)} \\ \frac{\partial u^m}{\partial \nu} &= 0 \quad \text{on } (0,T]\times\partial\Omega && \text{(3.85-b)} \\ u^m(0) &= u_0 \quad \text{on } \{0\}\times\Omega && \text{(3.85-c)} \end{aligned}$$

Due to Sec. 3.3, for each $m \in \mathbb{N}$ there is a solution $u^m \in W_p^{2,1}(Q_T)^I$, $u^m \geq 0$. Now let us state the main results of this section. We will prove that the equilibrium problem (3.86) has at least one solution. This solution is the weak limit of a subsequence of (u^m). We will state two different formulations of the equilibrium problem, (3.86) and (3.90):

Theorem 3.19 *Let $T>0$, $p>n+1$, $v_0 \in W_p^{2-2/p}(\Omega)^I$. Let $S_{eq}^\perp \in \mathbb{R}^{I\times(I-J_{eq})}$ be a matrix consisting of linearly independent columns begin orthogonal to the columns of matrix S_{eq}, i.e., $(S_{eq}^\perp)^T S_{eq} = 0$.*

(a) The problem

$$\begin{aligned} \partial_t (S_{eq}^\perp)^T u + L(S_{eq}^\perp)^T u &= (S_{eq}^\perp)^T S_{kin} R_{kin}(u) \quad \textit{on } Q_T, && \text{(3.86-a)} \\ R_{eq}(u) &= 0 \quad \textit{on } Q_T, && \text{(3.86-b)} \\ \frac{\partial}{\partial \nu}(S_{eq}^\perp)^T u &= 0 \quad \textit{on } (0,T]\times\partial\Omega && \text{(3.86-c)} \\ (S_{eq}^\perp)^T u(0) &= v_0 \quad \textit{on } \{0\}\times\Omega && \text{(3.86-d)} \end{aligned}$$

has a solution $u \in L^p(Q_T)^I$ with $(S_{eq}^\perp)^T u \in W_p^{2,1}(Q_T)^{I-J_{eq}}$, and $u \geq 0$ holds. If the initial conditions (3.86-d) and (3.85-c) coincide in the sense that[25]

$$v_0 = (S_{eq}^\perp)^T u_0, \tag{3.87}$$

[25] Or, what is equivalent to $(S_{eq}^\perp)^T u(0) = (S_{eq}^\perp)^T u_0 (= v_0)$, if $P_{S_{eq}^\perp} u(0) = P_{S_{eq}^\perp} u_0$.

then u is the weak $L^p(Q_T)^I$-limit of a subsequence of (u^m), and $(S_{eq}^\perp)^T u$ is the weak $W_p^{2,1}(Q_T)^{I-J_{eq}}$-limit of a subsequence of $((S_{eq}^\perp)^T u^m)$.
At all $(t,x) \in Q_T$ with $u(t,x) > 0$, u fulfils

$$Q_{eq}(u) = 0, \tag{3.88}$$

where

$$Q_{eq}(u) := S_{eq}^T \ln u - \ln K$$

$K = (k_j^f / k_j^b)_{j=1,\dots,J_{eq}} \in \mathbb{R}_+^{J_{eq}}$ being the vector of equilibrium constants.

(b) If a solution u of (a) fulfils $u_i \geq \epsilon$ on Q_T for all $i = 1, ..., I$ for an $\epsilon > 0$, then the improved regularity

$$u \in W_p^{2,1}(Q_T)^I \tag{3.89}$$

holds, and u solves the problem

$$\begin{aligned} \partial_t u + Lu &= S_{eq} r^{eq} + S_{kin} R_{kin}(u) \quad \text{on } Q_T, \\ Q_{eq}(u) &= 0 \quad \text{on } Q_T, \\ \frac{\partial}{\partial \nu}(S_{eq}^\perp)^T u &= 0 \quad \text{on } (0,T] \times \partial\Omega \\ (S_{eq}^\perp)^T u(0) &= v_0 \quad \text{on } \{0\} \times \Omega \end{aligned} \tag{3.90}$$

where

$$r^{eq} \in L^p(Q_T)^{J_{eq}}. \tag{3.91}$$

(c) Assume $J_{kin} = 0$. Then the set

$$\{u \text{ is solution of (3.86)-(3.87), and } u > 0 \text{ holds on } Q_T\}$$

contains at most one element.

Remark. Part (a) means that there is (at least) one nonnegative solution of the equilibrium problem (3.86) in $L^p(Q_T)^I$, where $(S_{eq}^\perp)^T u$, or, equivalently[26],

[26] Note that any regularity of $(S_{eq}^\perp)^T u$ is obviously equivalent to the corresponding regularity of the orthogonal projection $P_{S_{eq}^\perp} u = S_{eq}^\perp [(S_{eq}^\perp)^T S_{eq}^\perp]^{-1} (S_{eq}^\perp)^T u$.

$P_{S_{eq}^{\perp}}u$, has high regularity. Part (b) says that under an additional positivity condition, we can be sure that not only $(S_{eq}^{\perp})^T u$, but u itself, has $W_p^{2,1}$-regularity, i.e., the formulation (3.90) makes sense.
Part (c) says that, at least if *all* reactions are at equilibrium, among the strictly positive functions the solution of (3.86) is unique. The existence of a *strictly* positive solution is not shown here. This question could be tackled by using comparison principles on $\eta = ((S_{eq}^{\perp})^T S_{eq}^{\perp})^{-1}(S_{eq}^{\perp})^T$ (η then fulfills $\partial_t \eta + L\eta = 0$) to show that $\eta(t)$ remains in the open convex set

$$\mathcal{E} = \{\eta = [(S_{eq}^{\perp})^T S_{eq}^{\perp}]^{-1}(S_{eq}^{\perp})^T u \,|\, u \in \mathbb{R}_+^I\} \tag{3.92}$$

provided $\eta(0) \in \mathcal{E}$; then proceed as in the proof of part (c).

Proof of Theorem 3.19.

Part (a).

Step (1).

Let p, q with $n+1 < p, q < \infty$ be arbitrarily given. Let $u^m \in W_p^{2,1}(Q_T)^I$ be the solution of problem (3.85) with (3.87) holding for the initial condition. We note that the Lyapunov function F_r constructed in Sec. 3.3.4 depends only on the equilibrium constants of the reactions, i.e., on the ratios of the forward and the backward rate coefficients $K_j = k_j^f / k_j^b$ and is otherwise independent of the rate coefficients k_j^f, k_j^b. Hence, the definition of F_r for the sequence of problems (3.85) is independent of k_m and therefore of m. So with Cor. 3.8, there is a constant $c = c(u_0, r)$ independent of m such that

$$\|u^m\|_{L^r(Q_T)^I} \leq c \tag{3.93}$$

for all $m \in \mathbb{N}$, $1 \leq r < \infty$. Choosing r sufficiently large[27], we get the estimate

$$\|S_{eq} R_{eq}(u^m) + \frac{1}{k_m} S_{kin} R_{kin}(u^m)\|_{L^q(Q_T)^I} \leq c$$

for a constant $c = c(u_0, q)$, for all $m \in \mathbb{N}$, $1 \leq q < \infty$. The linear parabolic theory (Lemma 5.4) applied to (3.85) (after division of PDE and initial condition by k_m) tells us that

$$\|\frac{1}{k_m} u^m\|_{W_q^{2,1}(Q_T)^I} \leq c$$

for a $c = c(u_0, q)$, for all $m \in \mathbb{N}$, $1 \leq q < \infty$. We conclude that (u^m) has a subsequence, which will be denoted by (u^m) again, such that $(\frac{1}{k_m} u^m)$ is weakly

[27] How large r has to be obviously depends on q and the s_{ij}.

convergent in $W_q^{2,1}(Q_T)^I$. On the other hand, due to (3.93), $\frac{1}{k_m}u^m \longrightarrow 0$ in $L^q(Q_T)^I$. Hence,

$$\frac{1}{k_m}u^m \rightharpoonup 0 \quad \text{in } W_q^{2,1}(Q_T)^I \text{ weakly.} \tag{3.94}$$

From (3.94) and (3.93) we can conclude, for r, q sufficiently large, that all the terms $\frac{1}{k_m}S_{kin}R_{kin}(u^m)$, $\frac{1}{k_m}\partial_t u^m$, $\frac{1}{k_m}Lu^m$ converge weakly in $L^p(Q_T)^I$ to zero. Hence, the remaining term of (3.85-a), $S_{eq}R_{eq}(u^m)$, also has to converge to zero weakly in $L^p(Q_T)^I$. Multiplication by $(S_{eq}^T S_{eq})^{-1}S_{eq}^T$ yields

$$R_{eq}(u^m) \rightharpoonup 0 \quad \text{in } L^p(Q_T)^{J_{eq}} \text{ weakly.} \tag{3.95}$$

Another consequence of (3.93) is that there is a subsequence of (u^m), which will be denoted by (u^m) again, which is weakly convergent in $L^r(Q_T)^I$. r will be chosen soon. Let $u^* \in L^r(Q_T)^I$ be the weak limit:

$$u^m \rightharpoonup u^* \quad \text{in } L^r(Q_T)^I \text{ weakly.} \tag{3.96}$$

Using a decomposition of the rate terms like in Footnote 14 on p. 87 and choosing $r = r(q)$ sufficiently large we can conclude that

$$R_{eq}(u^m) \rightharpoonup R_{eq}(u^*) \quad \text{in } L^q(Q_T)^{J_{eq}} \text{ weakly,} \tag{3.97}$$
$$R_{kin}(u^m) \rightharpoonup R_{kin}(u^*) \quad \text{in } L^q(Q_T)^{J_{kin}} \text{ weakly,} \tag{3.98}$$

for arbitrary $1 \le q < \infty$. From (3.97) and (3.95),

$$R_{eq}(u^*) = 0 \tag{3.99}$$

follows.

Step (2). It remains to show that $(S_{eq}^\perp)^T u^*$ is in $W_p^{2,1}(Q_T)^{J_{eq}}$ and fulfils the differential equation (3.86-a) and the boundary and initial conditions (3.86-c), (3.86-d), and that

$$(S_{eq}^\perp)^T u^m \rightharpoonup (S_{eq}^\perp)^T u^* \quad \text{in } W_p^{2,1}(Q_T)^{I-J_{eq}} \text{ weakly.} \tag{3.100}$$

We perform a decomposition like in Section 2.4: We introduce

$$\begin{aligned} \xi^* &:= (S_{eq}^T S_{eq})^{-1}S_{eq}^T u^*, & \eta^* &:= ({S_{eq}^\perp}^T S_{eq}^\perp)^{-1}{S_{eq}^\perp}^T u^*, \\ \xi^m &:= (S_{eq}^T S_{eq})^{-1}S_{eq}^T u^m, & \eta^m &:= ({S_{eq}^\perp}^T S_{eq}^\perp)^{-1}{S_{eq}^\perp}^T u^m, \end{aligned} \tag{3.101}$$

for all $m \in \mathbb{N}$, and for $m = 0$, denoting the initial value. Note that the inverse matrices exist, since the columns both of S_{eq} and of $S_{eq}^{\perp}$ are linear independent. We know from Section 2.4 that

$$\begin{aligned} u^m &= S_{eq}\xi^m + S_{eq}^{\perp}\eta^m, \\ u^* &= S_{eq}\xi^* + S_{eq}^{\perp}\eta^*. \end{aligned} \tag{3.102}$$

For later use we state that from (3.102) the relation

$$(S_{eq}^{\perp})^T S_{eq}^{\perp}\eta^* = (S_{eq}^{\perp})^T u^* \tag{3.103}$$

follows. Multiplying equation (3.85-a) by $(S_{eq}^T S_{eq})^{-1} S_{eq}^T$ and by $(S_{eq}^{\perp\,T} S_{eq}^{\perp})^{-1} S_{eq}^{\perp\,T}$, and using the property $S_{eq}^{\perp\,T} S_{eq} = 0$ and the fact that the matrices commute with the differential operators, we see that ξ^m and η^m meet the equations

$$\partial_t \eta^m + L\eta^m = (S_{eq}^{\perp\,T} S_{eq}^{\perp})^{-1} S_{eq}^{\perp\,T} S_{kin} R_{kin}(u^m) \quad \text{on } Q_T, \tag{3.104-a}$$

$$\partial_t \xi^m + L\xi^m = k_m\, R_{eq}(u^m) + (S_{eq}^T S_{eq})^{-1} S_{eq}^T S_{kin} R_{kin}(u^m) \quad \text{on } Q_T, \tag{3.104-b}$$

$$\frac{\partial \eta^m}{\partial \nu} = 0 \quad \text{on } (0,T]\times\partial\Omega \tag{3.104-c}$$

$$\frac{\partial \xi^m}{\partial \nu} = 0 \quad \text{on } (0,T]\times\partial\Omega \tag{3.104-d}$$

$$\eta^m(0) = \eta_0 \quad \text{on } \{0\}\times\Omega \tag{3.104-e}$$

$$\xi^m(0) = \xi_0 \quad \text{on } \{0\}\times\Omega, \tag{3.104-f}$$

with

$$\eta_0 = (S_{eq}^{\perp\,T} S_{eq}^{\perp})^{-1} S_{eq}^{\perp\,T}\, u_0, \quad \xi_0 = (S_{eq}^T S_{eq})^{-1} S_{eq}^T\, u_0,$$

$m \in \mathbb{N}$. From (3.96) it is clear that

$$\eta^m \rightharpoonup \eta^*, \quad \xi^m \rightharpoonup \xi^* \quad \text{ weakly in } L^q(Q_T)^{I-J_{eq}}, L^q(Q_T)^{J_{eq}}, \text{ resp.} \tag{3.105}$$

Since we know from (3.93) that the right hand side of (3.104-a) is bounded in $L^q(Q_T)^{I-Jeq}$ for arbitrary $n+1 < q < \infty$ for $m \to \infty$, the linear parabolic theory yields that the sequence (η^m) is bounded in $W_q^{2,1}(Q_T)^{I-J_{eq}}$. Therefore there is a subsequence of (u_m) such that the corresponding subsequence (η^m) converges weakly in $W_q^{2,1}(Q_T)^{I-J_{eq}}$. Due to (3.105), the weak limit can only be η^*, which lies in $W_p^{2,1}(Q_T)^{I-J_{eq}}$, then. Together with (3.98) the equalities

$$\begin{aligned} \partial_t \eta^* + L\eta^* &= (S_{eq}^{\perp\,T} S_{eq}^{\perp})^{-1} S_{eq}^{\perp\,T} S_{kin} R_{kin}(u^*) \\ \frac{\partial \eta^*}{\partial \nu} &= 0 \quad \text{on } (0,T]\times\partial\Omega \\ \eta^*(0) &= \eta_0 \quad \text{on } \{0\}\times\Omega \end{aligned} \tag{3.106}$$

follow from (3.104) for $m \to \infty$. Multiplication of (3.106) by $(S_{eq}^{\perp})^T S_{eq}^{\perp}$ and the η^*-u^*-relation (3.103) lead to the desired result.[28]
Equation $Q_{eq}(u^*) = 0$ holds since this equation and (3.99) are equivalent for $u^* > 0$.
The pending convergence (3.100) follows from the convergence result $\eta^m \rightharpoonup \eta^*$ in $W_p^{2,1}(Q_T)^{I-Jeq}$ by multiplication with $(S_{eq}^{\perp})^T S_{eq}^{\perp}$.
Part (b). Let u be a solution of (a) and η, ξ as in (3.101)-(3.102)[29]. We know from part (a) that the corresponding $\eta \in W_p^{2,1}(Q_T)^{I-Jeq}$. We have to improve the regularity of $\xi \in L^p(Q_T)^{J_{eq}}$; then the relation $u = S_{eq}\xi + S_{eq}^{\perp}\eta$ leads to a higher regularity of u.
Since u is assumed to be strictly positive on Q_T, the local algebraic equilibrium equation (3.99) can be expressed by

$$Q(\xi, \eta) := S_{eq}^T \ln(S_{eq}\xi + S_{eq}^{\perp}\eta) - \ln K = 0$$

at each point $(t, x) \in Q_T$. The mapping $Q : \mathbb{R}_+^I \to \mathbb{R}^{J_{eq}}$ at each fixed point $(t, x) \in Q_T$ is smooth, which allows us to apply the Implicit Function Theorem.

We note that $\partial Q/\partial \xi = S_{eq}^T \Lambda_{1/u} S_{eq}$, where $\Lambda_{1/u}$ is the diagonal matrix with entries $1/u$, and where S_{eq} consists of linearly independent columns. Hence, matrix $\partial Q/\partial \xi$ is symmetric positive definite, hence regular. Due to the Implicit Function Theorem, in a neighbourhood of η there is a C^1 resolution function $\xi = \xi(\eta)$. One easily computes the derivative

$$\frac{\partial \xi}{\partial \eta} = -\left(\frac{\partial Q}{\partial \xi}\right)^{-1} \frac{\partial Q}{\partial \eta} = (S_{eq}^T \Lambda_{1/u} S_{eq})^{-1} S_{eq}^T \Lambda_{1/u} S_{eq}^{\perp}.$$

Let us estimate the 2-norm

$$\|A\| = \max_{x \neq 0} \frac{|Ax|_{J_{eq}}}{|x|_{I-J_{eq}}}$$

of this $J_{eq} \times (I - J_{eq})$-matrix. At first we consider the factor $S_{eq}^T \Lambda_{1/u} S_{eq}^{\perp}$: Obviously

$$\|S_{eq}^T \Lambda_{1/u} S_{eq}^{\perp}\| \leq c(S) \max_{i=1,\ldots,I} \frac{1}{u_i} \leq \frac{c(S)}{\epsilon} \tag{3.107}$$

[28] Note that $(S_{eq}^{\perp})^T u^*$ has the required regularity for meeting boundary and initial values. Also note that in general, $\xi^*(0) = \xi_0$, $u^*(0) = u_0$ do *not* hold.
[29] We omit the '$*$' of u^*, η^*, ξ^* from now on.

holds, where $c(S) \geq 0$ is a constant depending only on the entries of the matrices S_{eq}, $S_{eq}^{\perp}$, and which is independent of t, x, and with $\epsilon > 0$ from the theorem.
Let us estimate the smallest eigenvalue of the symmetric positive definite matrix $S_{eq}^T \Lambda_{1/u} S_{eq}$, denoted by λ_{min}. We write

$$\begin{aligned} \lambda_{min} &= \min_{0 \neq w \in \mathbb{R}^{J_{eq}}} \frac{\langle S_{eq}^T \Lambda_{1/u} S_{eq} w, w \rangle_{J_{eq}}}{\langle w, w \rangle_{J_{eq}}} = \min_{w} \frac{\langle \Lambda_{1/u} S_{eq} w, S_{eq} w \rangle_I}{\langle w, w \rangle_{J_{eq}}} \\ &\geq \min_{i=1,...,I} \frac{1}{u_i} \min_{w} \frac{\langle S_{eq} w, S_{eq} w \rangle_I}{\langle w, w \rangle_{J_{eq}}} = \tilde{c}(S) \min_{i=1,...,I} \frac{1}{u_i} \end{aligned}$$

where $\tilde{c}(S)$, using the linear indepdendence of the columns of S_{eq}, is a strictly positive constant independent of u. Hence, we get the following estimate for the 2-norm of $(S_{eq}^T \Lambda_{1/u} S_{eq})^{-1}$:

$$\|(S_{eq}^T \Lambda_{1/u} S_{eq})^{-1}\| \leq \frac{1}{\lambda_{min}} \leq \frac{1}{\tilde{c}(S)} \max_{i=1,...,I} u_i.$$

Together with (3.107) and with $u \geq \epsilon$ we get the following estimate for the 2-norm of matrix $\frac{\partial \xi}{\partial \eta}$, holding globally in Q_T:

$$\left\| \frac{\partial \xi}{\partial \eta} \right\| \leq \tilde{\tilde{c}}(S) \frac{\max_{i=1,...,I} u_i}{\min_{i=1,...,I} u_i} \leq \frac{\tilde{\tilde{c}}(S)}{\epsilon} \max_{i=1,...,I} u_i,$$

and in particular, for each matrix entry,

$$\left| \frac{\partial \xi_i}{\partial \eta_j} \right| \leq \frac{c}{\epsilon} |u|_I \quad \forall\, i = 1, ..., J_{eq}, \quad j = 1, ..., I - J_{eq},$$

where we have used the equivalence of norms in the matrix space $\mathbb{R}^{J_{eq} \times (I - J_{eq})}$ on the left side of the estimate and the equivalence of norms in $\mathbb{R}^I$ on the right side.[30]
Now let us investigate the regularity of

$$\frac{\partial \xi_i}{\partial x_k} = \sum_{j=1}^{I - J_{eq}} \frac{\partial \xi_i}{\partial \eta_j} \frac{\partial \eta_j}{\partial x_k}.$$

[30]Remember that $|\cdot|_I$ denotes the Eucledian norm in $\mathbb{R}^I$.

By Cauchy-Schwarz and Hölder we get, since $p \geq 2$, with generic constants c,

$$|\partial_{x_k} \xi_i| \leq \frac{c}{\epsilon} |u|_I \, |\partial_{x_k} \eta|_{I-J_{eq}} \leq \frac{c}{\epsilon} \left(\sum_{i=1}^{I} {u_i}^p \right)^{1/p} \left(\sum_{i=1}^{I-J_{eq}} |\partial_{x_k} \eta|^p \right)^{1/p}$$

and

$$\begin{aligned} \int_{Q_T} |\partial_{x_k} \xi_i|^p \, d(t,x) &\leq \frac{c}{\epsilon^p} \left[\int_{Q_T} \left(\sum_{i=1}^{I} {u_i}^p \right)^2 d(t,x) \right]^{\frac{1}{2}} \left[\int_{Q_T} \left(\sum_{i=1}^{I-J_{eq}} |\partial_{x_k} \eta|^p \right)^2 d(t,x) \right]^{\frac{1}{2}} \\ &\leq \frac{c}{\epsilon^p} \left[\int_{Q_T} \sum_{i=1}^{I} {u_i}^{2p} d(t,x) \right]^{\frac{1}{2}} \left[\int_{Q_T} \sum_{i=1}^{I-J_{eq}} |\partial_{x_k} \eta|^{2p} d(t,x) \right]^{\frac{1}{2}}, \end{aligned}$$

hence

$$\|\partial_{x_k} \xi_i\|_{L^p(Q_T)^{J_{eq}}} \leq \frac{c}{\epsilon} \|u\|_{L^{2p}(Q_T)^I} \|\partial_{x_k} \eta\|_{L^{2p}(Q_T)^{I-J_{eq}}} < \infty$$

where the finiteness of the $2p$-norms follows, since if the assumptions of Theorem 3.19 hold for a p, then they also hold for every $\tilde{p} \geq p$. We repeat this procedure for the temporal derivative and for the second order spatial derivates and get

$$\xi \in W_p^{2,1}(Q_T)^{J_{eq}}.$$

Hence, $u \in W_p^{2,1}(Q_T)^I$. Setting

$$r^{eq} := (S_{eq}^T S_{eq})^{-1} S_{eq}^T \left(\partial_t u + Lu - S_{kin} R_{kin}(u) \right), \tag{3.108}$$

which clearly lies in $L^p(Q_T)^{J_{eq}}$, we get the desired result.

Part (c). Let u^1, u^2 be two solutions with $u^i > 0$, $\eta^i := ((S_{eq}^{\perp})^T S_{eq}^{\perp})^{-1} (S_{eq}^{\perp})^T u^i$, $i = 1, 2$. Since $J_{kin} = 0$, the η^i fulfil

$$\partial_t \eta^i + L \eta^i = 0$$

with identical initial and boundary conditions. Hence, $\eta^1 = \eta^2 =: \eta$.
At each point (t, x) both functions $\xi^i := (S_{eq}^T S_{eq})^{-1} (S_{eq})^T u^i$, $i = 1, 2$, have to fulfil

$$Q(S_{eq} \xi^i + S_{eq}^{\perp} \eta) = 0. \tag{3.109}$$

Since $\eta = \eta(t, x) \in \mathcal{E} \subset \mathbb{R}^{I-J_{eq}}$ (cf. (3.92)), we will be able to prove that the (global!) implicit solution $\xi(t, x)$ of (3.109) is uniquely defined (from which

$\xi^1 = \xi^2$ and thus $u^1 = u^2$ follows). For this we proceed like in the proof of Lemma 2.2:
Let g be the function defined in (3.31). For fixed $\eta \in \mathcal{E}$ let us define $\tilde{g} : \mathbb{R}^{J_{eq}} \longrightarrow \mathbb{R} \cup \{\infty\}$ by

$$\tilde{g}(\xi) := \begin{cases} g(S_{eq}\xi + S_{eq}^{\top}\eta), & S_{eq}\xi + S_{eq}^{\top}\eta \geq 0, \\ +\infty, & \text{else.} \end{cases}$$

Note that $\tilde{g}$ is proper, i.e., $\mathrm{dom}(\tilde{g}) := \{\xi \,|\, \tilde{g}(\xi) < \infty)\} \neq \emptyset$, since $\eta \in \mathcal{E}$. We state that $\mathrm{dom}(\tilde{g})$ is convex, the function $\tilde{g}$ is convex, since the Hessian $H\tilde{g}(\xi) = S_{eq}^T \Lambda_{1/u} S_{eq}$ is symmetric positive definite on the inner of $\mathrm{dom}(\tilde{g})$, and $\tilde{g}$ is lower semicontinuous. It is well known that under these conditions, all levelsets $\mathcal{L}(c) := \{\xi \in \mathbb{R}^{J_{eq}} \,|\, \tilde{g}(\xi) \leq c\}$ are closed sets. We know that $\tilde{g}(\xi) \geq |S_{eq}\xi + S_{eq}^{\top}\eta|_I$ holds (see (3.15)) and that $\xi = (S_{eq}^T S_{eq})^{-1} S_{eq}^T (S_{eq}\xi + S_{eq}^{\top}\eta)$ (see (3.102)), hence $|\xi|_{J_{eq}} \leq c(S)\,|S_{eq}\xi + S_{eq}^{\top}\eta|_I \leq c(S)\,\tilde{g}(\xi)$ for a constant $c(S) > 0$. Hence, $\tilde{g}(\xi) \geq \frac{1}{c(S)}\,|\xi|_{J_{eq}}$. That means that the levelsets of $\tilde{g}$ are bounded, hence compact. From this it follows by standard arguments from the finite dimensional optimization [34], that $\inf_{\xi \in \mathbb{R}^{J_{eq}}} \tilde{g}(\xi) = \min_{\xi \in \mathbb{R}^{J_{eq}}} \tilde{g}(\xi)$ exists, i.e., the minimization problem

$$\min \tilde{g}(\xi), \quad \xi \in \mathbb{R}^{J_{eq}} \tag{3.110}$$

has a solution. Since the Hessian is positive definite, $\tilde{g}$ is *strictly* convex on $\mathrm{dom}(\tilde{g})$, hence the solution of the minimization problem is unique. This unique solution ξ must lie in $\mathcal{E}$, since for $\xi \in \partial\mathcal{E}$ the subdifferential $\partial\tilde{g}(\xi)$ is the empty set. On $\mathcal{E}$, $\tilde{g}$ is differentiable, and $\xi \in \mathcal{E}$ is a solution of $Q_{eq}(S_{eq}\xi + S_{eq}^{\perp}\eta) = 0$ if and only if ξ solves (3.110), since

$$\nabla\tilde{g}(\xi) = S_{eq}^T(\mu^0 + \ln c) = -\ln K_{mob} + S_{eq}^T \ln c = Q_{eq}(c)$$

with $c = S_{eq}\xi + S_{eq}^{\perp}\eta$, i.e.,

$$\nabla\tilde{g}(\xi) = 0 \Longleftrightarrow Q_{eq}(c) = 0.$$

■

Some remarks on 'equilibrium rates'.
If we name the term r^{eq} from problem (3.90) 'equilibrium rates' then we have proven that (at least strictly positive) solutions u of the equilibrium problem (3.90) have *finite* equilibrium rates in the $L^p(Q_T)$-sense.

Let us recall that we have *not* proven that

$$u^m \rightharpoonup u^* \quad \text{in } W_p^{2,1}(Q_T)^I \ \text{ weakly.} \tag{3.111}$$

To get such a result and avoid a singularity of u^m at $t = 0$ for $m \to \infty$, it is probably required that u_0 (which is not uniquely determined by (3.87)) additionally meets the equilibrium condition $Q_{eq}(u_0) = 0$. The problem might be tackled by Lyapunov techniques. If(!) a convergence result (3.111) holds, then from (3.85-a) follows that r^{eq} is the weak $L^p(Q_T)^{J_{eq}}$-limit of $k_m R_{eq}(u^m)$:

$$k_m R_{eq}(u^m) \rightharpoonup r^{eq} \quad \text{in } L^p(Q_T)^{J_{eq}} \ \text{ weakly;} \tag{3.112}$$

i.e., the weak limit of the source term from the fast reactions in (3.85-a) exists and is *finite*.
In contrast to this, in the hydrogeochemical community it is sometimes believed that the equilibrium rates are 'infinity'; see [28]. In that publication '$R_{eq} = \infty$' is claimed, but this is quite misleading if we compare it with our results (3.86-b) and (3.91); it is only is true (in some sense) for initial time $t = 0$ and if the initial conditions are inconsistent with the equilibrium condition, or if the claim of being infinite is applied to the rates of *semi*reactions $k_m R^f_{eq}(u^m)$, $k_m R^b_{eq}(u^m)$ instead of $k_m R_{eq}(u^m)$, $m \to \infty$.
In [55], p. 18, the equilibrium rates are called 'indeterminate', by which the author betokens the fact that in the rate term $k_m R_{eq}(u^m)$ of the fast reactions in (3.85-a) consists of a factor k_m that goes to infinity, while the other factor $R_{eq}(u^m)$ goes to zero, for $m \to \infty$. However, if we define equilibrium rates as it is done in (3.108), this leads to a well defined, neither indeterminate nor infinite value, describing the amount of moles per time and volume taking part in equilibrium reactions. Obviously my conjecture (3.111), which would imply (3.112), is opposed to a claim of indetermination also for $k_m R_{eq}(u^m)$, $m \to \infty$.

Chapter 4

Reactions with minerals: A complementarity problem solved by the semismooth Newton method

In this chapter we extend the algorithms of Chapter 2 to problems with minerals. We start with an introduction on the specific difficulties of problems with minerals (Sec. 4.1). We proceed by recalling some solution strategies which are presently used in the geoscientists' community (Sec. 4.2). We discuss some disadvantages of these methods to motivate the introduction of modern mathematical methods in this field of application. These modern methods are the formulation of the mineral problem as a (nonlinear) complementarity problem (coupled to PDEs, ODEs, AEs) and its solution by the Semismooth Newton method. This technique is well known in the field of mathematical optimization. Its central idea is laid out in (4.10)-(4.13). Its application to the mineral precipitation-dissolution reactive transport problem in this form seems novel and is pointed out in Sec. 4.3-4.4. Two applications of the proposed algorithm are presented in Sec. 4.5.

4.1 Introduction: Reactions with minerals

Let us point out the specific difficulty that complicate reactive transport problems when reactions *with minerals* (pure solids) are considered. Actually,

the mass action laws in the formulations (2.5), (2.6), are only approximations of the preciser versions

$$R_j(u,\bar{u}) = k_j^f \prod_{\substack{s_{ij}<0 \\ i=1,\dots,I+\bar{I}}} a_i^{|s_{ij}|} \;-\; k_j^b \prod_{\substack{s_{ij}>0 \\ i=1,\dots,I+\bar{I}}} a_i^{s_{ij}}, \tag{4.1}$$

$$\prod_{i=1}^{I} a_i^{s_{ij}} \prod_{i=I+1}^{I+\bar{I}} \bar{a}_i^{s_{ij}} = k_j, \tag{4.2}$$

where $a_i = a_i(u)$ is the *activity* of the i-th species. The activity depends on the vector of concentrations u; usually a dependence

$$a_i = \gamma_i(u)\, u_i$$

is assumed,[1] where $\gamma_i(u)$ is called the *activity coefficient* of species i (depending on the whole concentration vector).
For *non-minerals*, usually the concentration is a good approximation for the activity (i,.e., $\gamma_i \approx 1$, at least if concentrations are not too large), giving the justification for (2.5), (2.6) in Chapter 2. However, for *minerals*, usually *constant* activity is assumed, i.e., (by incorporating any activity constant into the rate coefficients/equilibrium constants),

$$a_i = 1.$$

A kinetic mass action law, for the example

$$\mathrm{C} \longleftrightarrow \mathrm{A} + 2\mathrm{B} \tag{4.3}$$

reads

$$R(u) = k^f a_C - k^b a_A a_B^2 \tag{4.4}$$

which is simplified by $a_i \stackrel{!}{=} u_i$ to

$$R(u) = k^f u_C - k^b u_A u_B^2$$

throughout Chapters 2 and 3. However, if we assume that species C is a mineral, then (4.4) reads

$$R(u) = k^f - k_b u_A u_B^2, \tag{4.5}$$

[1] Detailed activity models can be found in [76] Sec. 3.4.

and the corresponding equilibrium condition would read

$$u_A u_B^2 = \frac{k^f}{k^b} =: K,$$

i.e.,

$$Q(u) := u_A u_B^2 - K = 0; \tag{4.6}$$

the constant $K>0$ is called the *solubility product* of the reaction.
The specific mass action form of the equilibrium condition (4.6) is not essential in what will follow in this chapter (modifications may be considered instead); the main difficulty is that *a smooth, algebraic equation such as (4.6) is usually not sufficient to describe all states of local equilibrium*:
Considering the corresponding kinetic differential system

$$\begin{aligned} u_A' &= R(u) \\ u_B' &= 2R(u) \\ u_C' &= -R(u) \end{aligned} \tag{4.7}$$

it is obvious that the rate law (4.5) cannot be valid for

$$u_C = 0 \ \ \wedge \ \ k^f - k_b u_A u_B^2 > 0,$$

since a positive rate $R(u)$ would lead to negative concentrations u_C (see Fig. 4.1). A reasonable model, complying with the nonnegativity of mineral concentrations, is [23, 24]

$$R(u) \in \begin{cases} [0, k^f - k^b u_A u_B^2], & \text{if } u_C = 0 \ \wedge \ k^f - k^b u_A u_B^2 > 0, \\ \{k^f - k^b u_A u_B^2\}, & \text{if } u_C > 0 \ \vee \ k^f - k^b u_A u_B^2 \le 0. \end{cases} \tag{4.8}$$

The system (4.7)-(4.8) is a *differential inclusion* (a set-valued differential equation) [6].
Our main interest lies in the *equilibrium problem* compatible to the kinetic law (some further remarks on the handling of the kinetic problem are given in Sec. 4.6). The equilibrium state corresponding to the kinetic law (4.8) is obviously (see also Fig. 4.1)

$$(u_C \ge 0 \ \wedge \ k^f - k^b u_A u_B^2 = 0) \ \vee \ (u_C = 0 \ \wedge \ k^f - k^b u_A u_B^2 \ge 0). \tag{4.9}$$

One branch describes the state of a saturated solution, and the other branch describes the state of a totally dissolved mineral.

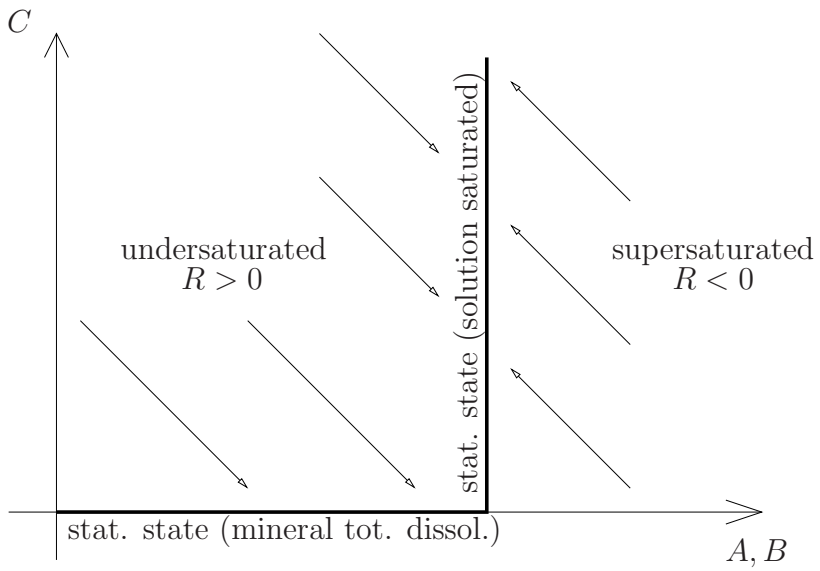


Figure 4.1: Schematic view on a mineral precipitation-dissolution reaction. The arrows indicate the behaviour in the *kinetic* situation; the alignment of the arrows is determined by mass conservation (if no spatial transport and no further reactions are considered). The bold line shows the corresponding stationary states. In the vertical part of the bold line, the solution is saturated, left from it it is undersaturated, and right from it, the solution is supersaturated. In the horizontal part of the bold line, the mineral is fully dissolved *and* the solution is undersaturated.

The main difficulty, when a system with minerals is considered, is the fact that besides AEs, describing equilibria between non-minerals, and the ODEs and PDEs, describing mass conservation, we have to consider *conditions consisting of equations and inequalities* such as (4.9). In many chemical systems arising in the geosciences, there is a multitude of minerals involved, i.e., there are several conditions of type (4.9) to be considered. The retrieval of the correct mineral assemblage at each part of the computational domain is often nontrivial. In the next section, we revisit some available solution strategies, before we propose the method of Complementarity Formulations and Semismooth Newton Method.

4.2 State of the art: Available numerical strategies, and motivation of the proposed method

A classical strategy to determine the correct mineral assemblage, at least as far as batch problems are considered, is the following [11]:
Make an assumption on the mineral assemblage, i.e., fix a set of minerals having $u_i = 0$. Under this assumption, solve the equilibrium conditions (beeing pure *equations*, each being either of type $u_i = 0$ or of type (4.6)) together with the mass conservation equations for the remaining species concentrations. In general, the solution will be nonphysical, i.e., the solution contains negative mineral concentrations or supersaturated states. This indicates that an incorrect assumption that has been made. Hence, modify the assumption, e.g., take that mineral which has the 'most negative' concentration into the set of minerals with $u_i = 0$, and then repeat the computation of an equilibrium state. This is done until a physical solution is gained. It seems that this algorithm, though working pretty well in practice, at least for batch problems and for operator splitting approaches, lacks general convergence proofs. It is currently widespread in the geosciences (e.g., [16]). The main disadvantage of this startegy is that it requires *a sequence* of Newton iterations instead of *one* Newton interation to find a solution. Especially if we want to adapt this algorithm from the batch problem to the fully coupled transport-reaction problem, where assumptions on the mineral assemblage at each mesh point/cell have to be made, it can be expected that convergence takes unrealistically long (if it converges at all, since a change of the assumption at one point may influence the negativity/nonnegativity of the solution at another point).
A refinement of the method might be strategies to change the assumption on the mineral assemblage *within* the Newton iteration, thus avoiding the sequence of iterations. This would resemble an *active set method*, thus coming close to the resulting algorithm of Sec. 4.3-4.4.
Another approach, propagated in [54, 55] and in [58], is to consider reactive transport problems with minerals as *moving boundary problems*. The computational domain Ω is, with respect to *each* mineral species i, divided into a subdomain Ω_i where $u_i = 0$ and a subdomain $\Omega \backslash \bar{\Omega}_i$ on which the equilibrium condition of type (4.6) holds. On the interface between the subdomains, generalized Rankine-Hugoniot conditions, relating the propagation speed of the

interfaces with the unknowns on the subdomains, are set up. The resulting system could be tackled by a front-tracking algorithm. While this treatment seems interesting for 1-D problems [55] where well-distinguished fronts move in flow direction or maybe in 2-D/3-D problems where some a priori knowledge on the topology of the fronts is available [58], the practical algorithmic difficulties in general 2-D and 3-D computations can be considerable (possible changes of topology of subdomains and interfaces within a time-step).
A third way to handle reactions with minerals is proposed in [69]. There, after a splitting of transport from reactions, the chemistry problem is formulated as a minimization problem for the Gibbs free energy (see (3.14), but, due to the constant activity, with constant potentials μ_i^0 instead of $\mu_i^0 + RT \ln u_i$ for every mineral). This minimization problem is considered (in the setting of an operator splitting approach) together with (algebraic) equality constraints representing mass conservation and with inequality constraints due to the nonnegativity requirement for the mineral concentrations. For this constrained minimization problem, the Karush-Kuhn-Tucker (KKT) first order optimality conditions are solved. As a numerical solver, an *interior point method* (central path method, see, e.g., [26]) is proposed in [69]. Interior point methods are in particular popular for monotone problems. Note that since we are considering fully coupled transport-reaction problmes with kinetic reactions, we cannot assume monotonicity.

The method we are proposing is the following: The equilibrium conditions of type (4.9) is equivalently transformed into

$$(k^f - k^b u_A u_B^2)\, u_C = 0 \quad \wedge \quad k^f - k^b u_A u_B^2 \geq 0 \quad \wedge \quad u_C \geq 0. \tag{4.10}$$

A condition of this type

$$E_i(u)\, u_i = 0 \ \wedge \ E_i(u) \geq 0 \ \wedge u_i \geq 0. \tag{4.11}$$

is called a *complementarity condition* (CC). A *general* mineral precipitation-dissolution reaction

$$\mathrm{s_{1i}A_1} + ... + \mathrm{s_{Ii}A_I} \longleftrightarrow \mathrm{A_i}$$

with $\mathrm{A_1}, ..., \mathrm{A_I}$ being non-minerals, A_i, $i \in \{I+1, ..., I+\bar{I}\}$ being a mineral, is described by (4.11) with

$$E_i(u) = K_i - \prod_{k=1}^{I} u_k^{s_{ki}}$$

or

$$E_i(u) = \ln K_i - \sum_{k=1}^{I} s_{ki} \ln u_k.$$

Hence, it is a *nonlinear* CC (in the sense that E_i is in general nonlinear). Note that the KKT-conditions mentioned in the third approach of Sec. 4.2 has also the shape of a complementarity problem (CP)[2]. However, note that our approach does not require the introduction of Lagrangian multiplies as additional variables.

Our proposal concerning the algorithmic solution of the CP is the Semismooth Newton Method [21, 42, 47] for the fully implicit reactive transport system. As a first step, the CP, consisting of equations and inequalities, is equivalently transformed into a pure set of equations. For this, a Lipschitz continuous function

$$\varphi : \mathbb{R}^2 \longrightarrow \mathbb{R} \tag{4.12-a}$$

with the property

$$\varphi(a,b) = 0 \quad \Longleftrightarrow \quad ab = 0 \ \wedge \ a \geq 0 \ \wedge \ b \geq 0 \tag{4.12-b}$$

has to be chosen. Such a function fulfilling (4.12) is called a *complementarity function*. Then, the CC (4.11) is obviously equivalent to the *equation*(!)

$$\varphi(E_i(u), u_i) = 0. \tag{4.13}$$

Hence, the resulting system consisting of mass conservation (differential) equations and equilibrium conditions *is fully free of inequalities*.

Typical representatives of φ are the Fischer-Burmeister function [31] and the minimum function:

$$\varphi_{FB}(a,b) = a + b - \sqrt{a^2 + b^2}, \tag{4.14-a}$$

$$\varphi_{MIN}(a,b) = \min\{a, b\} \tag{4.14-b}$$

The only drawback of the introduction of a complementarity problem is that the problem is no longer C^1, since $\varphi \notin C^1(\mathbb{R}^2)$, while in textbooks on Newton's method [22] the typical assumption for local quadratic convergence is "C^1 with Lipschitz-continuous derivative" (or similar). However, it is well known, especially in the community of optimization, that the assumptions can be weakened to the assumption of *strong semismoothness*.

We give here a definition of semismoothness and strong semismoothness and a convergence theorem from [21]:

[2] i.e, a system containing a CC

Definition 4.1 (Semismoothness, strong semismoothness). *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be locally Lipschitz continuous. Then, by Rademacher's Theorem, there is a dense subset $D \subset \mathbb{R}^n$ on which f is differentiable. We define the* B-subdifferential *of f at $x \in \mathbb{R}^n$ by*

$$\partial_B f(x) := \{M \in \mathbb{R}^{n\times n} \mid M = \lim_{x_k \in D, x_k \to x} f'(x_k)\}.$$

We call the convex hull $\partial f(x) := conv(\partial_B f(x))$ the generalized Jacobian *of f at x. f is said to be* semismooth *at x if f is directionally differentiable at x and*

$$Md - f'(x; d) = o(\|d\|),$$

for any $d \to 0$ and any $M \in \partial f(x+d)$, where $f'(x; d)$ denotes the directional derivative of f at x in the direction of d. Analogously, f is called strongly semismooth*, if*

$$Md - f'(x; d) = O(\|d\|^2).$$

f is called (strongly) semismooth if f is (strongly) semismooth at any point $x \in \mathbb{R}^n$.

It is well known that the functions (4.14) are strongly semismooth. Note that the above definition of the B-subdifferential does not require the assumption of convexity of f.

Definition 4.2 (Semismooth Newton Method, BD-regularity). *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be locally Lipschitz continuous.*

(i) The semismooth Newton Method is defined such as the classical Newton method, but with the classical Jacobian replaced by an arbitrary B-subdifferential.

(ii) A solution x of $f(x) = 0$ is called BD-regular*, if all elements $M \in \partial_B G(x)$ are non-singular.*

Theorem 4.3 ((Local, superlinear/quadratic) convergence of Semismooth Newton). *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be semismooth (strongly semismooth) and let x^* be a BD-regular solution of $f(x) = 0$. Then, for any initial value sufficiently close to x^*, the Semismooth Newton Method generates a sequence which converges superlinearly (quadratically, resp.) to x^*.*

In order to apply this finite-dimensional result to our situation, we consider our problem in a *discretized* version.

4.3 Definition of the model equations for reactive transport problems with minerals

Let us extend the model of Chapter 2 in the following way: We assume that the $\bar{I}$ immobile species are separated into $\bar{I}_{nm}$ non-minerals and $\bar{I}_{min}$ minerals; $\bar{I} = \bar{I}_{nm} + \bar{I}_{min}$. We assume that there are J_{sorp} equilibrium sorption equations, acting between the mobile and the immobile non-minerals, and there are $J_{min} = \bar{I}_{min}$ equilibrium mineral precipitation-dissolution reactions between the mobile and the immobile mineral species. In particular, we assume that there is a one-to-one mapping between the minerals an the mineral reactions in the sense that each mineral is participating in one and only one mineral reaction, and in every mineral reaction, one and only one mineral is involved. We assume that the immobile non-minerals do not participate in mineral reactions. In order not to complicate the representation too much, we assume that there are no reactions which are restricted to immobile species, i.e., $J_{immo} = 0$. This seems to be not very restictive; most examples lack of such reactions anyway. The equilibrium part of the stoichiometric matrix then reads

$$S_{eq} = \left(\begin{array}{c} S^1_{eq} \\ \hline S^2_{eq} \end{array}\right) = \left(\begin{array}{c|c|c} S^1_{mob} & S^1_{sorp} & S^1_{min} \\ \hline\hline 0 & S^2_{sorp} & 0 \\ \hline 0 & 0 & -\mathrm{Id} \end{array}\right).$$

Note that without loss of generality we had replaced the diagonal matrix, expressing the mineral participation in the mineral reactions, by the identity matrix.

Furthermore we restrict ourselves to kinetic reactions among the mobile and the immobile non-mineral species, i.e, minerals do not participate in kinetic reactions. The stoichiometric matrix then reads

$$S = (S_{eq} \,|\, S_{kin}) = \left(\begin{array}{c|c} S^1_{eq} & S^1_{kin} \\ \hline S^2_{eq} & S^2_{kin} \end{array}\right) = \left(\begin{array}{c|c|c||c} S^1_{mob} & S^1_{sorp} & S^1_{min} & S^1_{kin} \\ \hline\hline 0 & S^2_{sorp} & 0 & S^2_{kin} \\ \hline 0 & 0 & -\mathrm{Id} & 0 \end{array}\right). \quad (4.15)$$

We assume that both the columns of $(S^1_{mob}|S^1_{sorp}|S^1_{min})$ and S^2_{sorp} are linearly independent.

We decompose the reaction vector

$$R = (R_{eq}, R_{kin}) = (R_{mob}, R_{sorp}, R_{min}, R_{kin})$$

of size $J_{mob}, J_{sorp}, J_{min}, J_{kin}$. Then we can write the differential system as

$$
\begin{aligned}
\partial_t c + Lc &= S^1_{eq} R_{eq} + S^1_{kin} R_{kin}(c, \bar{c}_{nm}) \\
&= S^1_{mob} R_{mob} + S^1_{sorp} R_{sorp} + S^1_{min} R_{min} + S^1_{kin} R_{kin}(c, \bar{c}_{nm}) \\
\partial_t \bar{c}_{nm} &= S^2_{sorp} R_{sorp} + S^2_{kin} R_{kin}(c, \bar{c}_{nm}) \\
\partial_t \bar{c}_{min} &= -R_{min}
\end{aligned}
\tag{4.16-a}
$$

which has to be solved for the unknowns $c \in \mathbb{R}^I$, $\bar{c}_{nm} \in \mathbb{R}^{\bar{I}_{nm}}$, $\bar{c}_{min} \in \mathbb{R}^{\bar{I}_{min}}$, $R_{eq} \in \mathbb{R}^{J_{eq}}$ together with the equilibrium conditions

$$
\begin{aligned}
&Q_{mob}(c) := {S^1_{mob}}^T \ln c - \ln K_{mob} = 0, \\
&Q_{sorp}(c, \bar{c}_{nm}) := {S^1_{sorp}}^T \ln c + {S^2_{sorp}}^T \ln \bar{c}_{nm} - \ln K_{sorp} = 0, \\
&\varphi(E(c), \bar{c}_{min}) = 0, \quad \text{where } E(c) := \ln K_{min} - {S^1_{min}}^T \ln c,
\end{aligned}
\tag{4.16-b}
$$

where $K_{mob} \in \mathbb{R}_+^{J_{mob}}$, $K_{sorp} \in \mathbb{R}_+^{J_{sorp}}$, $K_{min} \in \mathbb{R}_+^{J_{min}}$ are the equilibrium constants/solubility products, $J_{min} = \bar{I}_{min}$. Note that in (4.16-b), we have already introduced a complementarity function φ (see Sec. 4.2).[3] Also note that if we had not made the assumption that the mineral reactions are independent of the sorbed concentrations $\bar{c}_{nm}$, then E would also depend on $\bar{c}_{nm}$ besides c.

Now we apply the reformulation technique of Chapter 2 to the system (4.16). We obtain the system

$$
\begin{array}{llll}
\partial_t \eta & +L\eta & = 0 & \left.\begin{array}{l}\\ \\ \end{array}\right\} \text{decoupled} \\
\partial_t \bar{\eta} & & = 0 & \\
\partial_t \xi_{kin} & +L\xi_{kin} \quad -f_1(\xi, \bar{\xi}) & = 0 & \left.\begin{array}{l}\\ \\ \\ \end{array}\right\} \begin{array}{l}\text{coupled} \\ \text{PDEs}\end{array} \\
\partial_t (\xi_{sorp} - \bar{\xi}_{sorp}) & +L\xi_{sorp} \quad -f_2(\xi, \bar{\xi}) & = 0 & \\
\partial_t (\xi_{min} - \bar{\xi}_{min}) & +L\xi_{min} \quad -f_3(\xi, \bar{\xi}) & = 0 & \\
\partial_t \bar{\xi}_{kin} & -f_4(\xi, \bar{\xi}) & = 0 & \left.\begin{array}{l}\\ \\ \end{array}\right\} \text{local eq.} \\
Q_{sorp}(c, \bar{c}_{nm}) = Q_{sorp}(\xi_{mob}, \xi_{sorp}, \xi_{min}, \xi_{kin}, \bar{\xi}_{sorp}, \bar{\xi}_{kin}) & & = 0 & \\
\varphi(E(c), \bar{c}_{min}) & & = 0 & \text{CCs (local)} \\
Q_{mob}(c) = Q_{mob}(\xi_{mob}, \xi_{sorp}, \xi_{min}, \xi_{kin}) & & = 0 & \text{local eq.}
\end{array}
\tag{4.17}
$$

The terms f_i contain the kinetic rates:

$$
\begin{aligned}
f_1 &= A^1_{kin} R_{kin}(c, \bar{c}_{nm}), \quad f_2 = A_{sorp} R_{kin}(c, \bar{c}_{nm}), \\
f_3 &= A_{min} R_{kin}(c, \bar{c}_{nm}), \quad f_4 = A^2_{kin} R_{kin}(c, \bar{c}_{nm}),
\end{aligned}
$$

[3] The application of φ to vectors $E(c), \bar{c}_{nm}$ is meant componentwise: $\varphi(E(c), \bar{c}_{min}) = (\varphi(E_1(c), \bar{c}_{min,1}), ..., \varphi(E_{\bar{I}_{min}}(c), \bar{c}_{min,\bar{I}_{min}}))$

where the matrices $A^1_{kin}, A^2_{kin}, A_{sorp}, A_{min}$ can be computed from S following Chapter 2. The arguments c, $\bar{c}_{nmin}$ are expressed by (4.18). Any dependence of the f_i on $\eta, \bar{\eta}$ is suppressed, since the computation of $\eta, \bar{\eta}$ is decoupled so that these entities can be considered to be *known*. The remaining unknowns in (4.17) are

$$\begin{aligned} \xi &= (\xi_{mob}, \xi_{sorp}, \xi_{min}, \xi_{kin}) && \in \mathbb{R}^{J_{mob}} \times \mathbb{R}^{J_{sorp}} \times \mathbb{R}^{J_{min}} \times \mathbb{R}^{J^{1'}_{kin}} \\ \bar{\xi} &= (\bar{\xi}_{sorp}, \bar{\xi}_{min}, \bar{\xi}_{kin}) && \in \mathbb{R}^{J_{sorp}} \times \mathbb{R}^{J_{min}} \times \mathbb{R}^{J^{2'}_{kin}} \end{aligned}$$

The transformation between c and $\bar{c} = (\bar{c}_{nm}, \bar{c}_{min})$ on the one hand and $\eta, \xi, \bar{\eta}, \bar{\xi}$ on the other hand, reads

$$\begin{aligned} c &= S_1^* \xi + S_1^\top \eta = S^1_{mob}\xi_{mob} + S^1_{sorp}\xi_{sorp} + S^1_{min}\xi_{min} + S^{1'}_{kin}\xi_{kin} + S_1^\perp \eta, \\ \bar{c} = \begin{pmatrix} \bar{c}_{nm} \\ \bar{c}_{min} \end{pmatrix} &= S_2^* \bar{\xi} + S_2^\top \bar{\eta} = \begin{pmatrix} S^2_{sorp}\bar{\xi}_{sorp} & + S^{2'}_{kin}\bar{\xi}_{kin} \\ & -\bar{\xi}_{min} \end{pmatrix} + S_2^\perp \bar{\eta}, \end{aligned} \tag{4.18}$$

where the specific structure (4.15) has been used. In particular, $\bar{\xi}_{min} = -\bar{c}_{min}$ holds. Since R_{kin} has been assumed to be independed of $\bar{c}_{min}$, the f_j only depend on ξ and $\bar{\xi}_{nmin} := (\bar{\xi}_{sorp}, \bar{\xi}_{kin})$ and are independent of $\bar{\xi}_{min}$.

4.4 Semismooth Newton Method for reactive transport problems with minerals in porous media

4.4.1 General statements

In Sec. 4.4 we assume that we have discretized the system (4.17) in space and time. To keep the notation clear, we suppress subscripts indicating the discretization (except L_h for the discretization of L). For the sake of simplicity, we assume the Implicite Euler time stepping scheme; however, other temporal discretizations are possible. Furthermore, we omit the decoupled

η-, $\bar{\eta}$-equations from now on. The discrete system reads

$$\begin{aligned}
G_1 &:= \xi_{kin} &&+\tau L_h \xi_{kin} &&-\tau f_1(\xi_{mob},\xi_{sorp},\xi_{min},\xi_{kin},\bar{\xi}_{sorp},\bar{\xi}_{kin}) - \xi_{kin}^{old} &&= 0\\
G_2 &:= \xi_{sorp} - \bar{\xi}_{sorp} &&+\tau L_h \xi_{sorp} &&-\tau f_2(\xi_{mob},\xi_{sorp},\xi_{min},\xi_{kin},\bar{\xi}_{sorp},\bar{\xi}_{kin}) - \xi_{sorp}^{old} + \bar{\xi}_{sorp}^{old} &&= 0\\
G_3 &:= \xi_{min} - \bar{\xi}_{min} &&+\tau L_h \xi_{min} &&-\tau f_3(\xi_{mob},\xi_{sorp},\xi_{min},\xi_{kin},\bar{\xi}_{sorp},\bar{\xi}_{kin}) - \xi_{min}^{old} + \bar{\xi}_{min}^{old} &&= 0\\
G_4 &:= \bar{\xi}_{kin} &&&&-\tau f_4(\xi_{mob},\xi_{sorp},\xi_{min},\xi_{kin},\bar{\xi}_{sorp},\bar{\xi}_{kin}) - \bar{\xi}_{kin}^{old} &&= 0\\
G_5 &:= Q_{sorp}(\xi_{mob},\xi_{sorp},\xi_{min},\xi_{kin},\bar{\xi}_{sorp},\bar{\xi}_{kin}) &&&&&&= 0\\
G_6 &:= -\varphi(E(\xi_{mob},\xi_{sorp},\xi_{min},\xi_{kin}),-\bar{\xi}_{min}) &&&&&&= 0\\
G_7 &:= Q_{mob}(\xi_{mob},\xi_{sorp},\xi_{min},\xi_{kin}) &&&&&&= 0
\end{aligned} \tag{4.19}$$

The superscript 'old' indicates values from the previous time-step. The minus in front of φ is chosen in order to simplify the proofs of the following Lemmas 4.4, 4.5.
Now, we want to apply Newton's method. At the same time we want to reduce the size of the system by elimination of equations and unknowns. This elimination can take place on the nonlinear or on the linear level. Hence, there are (at least) two different possibilities how to proceed:

1. Apply the Semismooth Newton method to system (4.19). In the resulting linear problem, solve the local equations for some 'local' variables and eliminate these unknowns from the remaining linear system. This elimination is a Schur complement method.

2. Solve the local equations (including the CP-equations) of (4.19) for some local unknowns, eliminate these unknowns in the remaining PDE-/ ODE-system, and apply Newton's method to the remaining system.

Intermediate versions are thinkable, such as:

3. Eliminate the local equations (exclusive of the CP-equations) on the nonlinear level, and the CP-equations on the linear level.

The different procedures lead to very similar problems in the sense that the resulting system matrix has the same (block) structure. However, the methods are not equivalent in the sense that they would generate identical Newton iterates. Let us focus on version 1.

The linearization by Newton's method yields the linear system

$$J \begin{pmatrix} \Delta\xi_{kin} \\ \Delta\xi_{sorp} \\ \Delta\xi_{min} \\ \Delta\bar{\xi}_{kin} \\ \Delta\bar{\xi}_{sorp} \\ \Delta\bar{\xi}_{min} \\ \Delta\xi_{mob} \end{pmatrix} = - \begin{pmatrix} G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \\ G_6 \\ G_7 \end{pmatrix} \tag{4.20}$$

for the Newton update, where J is given in Table 4.1, left part.

4.4.2 The choice of the CP-function

If we choose the Fischer-Burmeister function (4.14-a) for φ, then the partial derivatives read

$$\varphi_a(a,b) = 1 - \frac{a}{\sqrt{a^2+b^2}}, \quad \varphi_b(a,b) = 1 - \frac{b}{\sqrt{a^2+b^2}},$$

for $(a,b) \neq (0,0)$. Unfortunately, when the Newton iterates approach the solution of the nonlinear problem, then a or b approaches zero ($a = E_i(u) \longrightarrow 0$ in the saturated case, $b = \bar{c}_{min,i} \longrightarrow 0$ in the undersaturated case). Hence, some entries (possibly diagonal entries!) in the Jacobian approach zero, while other entries in the same row of the Jacobian converge to nonzero values. Therefore we can expect a deteriorating condition number of the linear system (4.20) when the iteration approaches the solution. A similar behaviour is described in [47].
If we choose the min-function for ϕ, then

$$\varphi_a(a,b) = \begin{cases} \{0\}, & a > b \\ \{1\}, & a \leq b \\ \{0,1\}, & a = b \end{cases}, \qquad \varphi_b(a,b) = \begin{cases} \{1\}, & a > b \\ \{0\}, & a \leq b \\ \{0,1\}, & a = b \end{cases}, \tag{4.21}$$

holds. The problem of matrix (diagonal) entries close to zero, but different from zero, disappears. Instead, the appearence of zero entries in the Jacobian can be exploited to decouple some equations and reduce the size of the remaining system; see the following Section 4.4.3.

Table 4.1: Left part: Matrix J from (4.20). Right part: Matrix J from (4.23).

$$
\left(\begin{array}{c|c|c|c|c|c|c}
\mathrm{Id}+\tau L_h-\frac{\tau\,\partial f_1}{\partial \xi_{kin}} & -\frac{\tau\,\partial f_1}{\partial \xi_{sorp}} & -\frac{\tau\,\partial f_1}{\partial \xi_{min}} & -\frac{\tau\,\partial f_1}{\partial \bar{\xi}_{kin}} & -\frac{\tau\,\partial f_1}{\partial \bar{\xi}_{sorp}} & 0 & -\frac{\tau\,\partial f_1}{\partial \xi_{mob}} \\ \hline
-\frac{\tau\,\partial f_2}{\partial \xi_{kin}} & \mathrm{Id}+\tau L_h-\frac{\tau\,\partial f_2}{\partial \xi_{sorp}} & -\frac{\tau\,\partial f_2}{\partial \xi_{min}} & -\frac{\tau\,\partial f_2}{\partial \bar{\xi}_{kin}} & -\mathrm{Id}-\frac{\tau\,\partial f_2}{\partial \bar{\xi}_{sorp}} & 0 & -\frac{\tau\,\partial f_2}{\partial \xi_{mob}} \\ \hline
-\frac{\tau\,\partial f_3}{\partial \xi_{kin}} & -\frac{\tau\,\partial f_3}{\partial \xi_{sorp}} & \mathrm{Id}+\tau L_h-\frac{\tau\,\partial f_3}{\partial \xi_{min}} & -\frac{\tau\,\partial f_3}{\partial \bar{\xi}_{kin}} & -\frac{\tau\,\partial f_3}{\partial \bar{\xi}_{sorp}} & -\mathrm{Id} & -\frac{\tau\,\partial f_3}{\partial \xi_{mob}} \\ \hline
-\frac{\tau\,\partial f_4}{\partial \xi_{kin}} & -\frac{\tau\,\partial f_4}{\partial \xi_{sorp}} & -\frac{\tau\,\partial f_4}{\partial \xi_{min}} & \mathrm{Id}-\frac{\tau\,\partial f_4}{\partial \bar{\xi}_{kin}} & -\frac{\tau\,\partial f_4}{\partial \bar{\xi}_{sorp}} & 0 & -\frac{\tau\,\partial f_4}{\partial \xi_{mob}} \\ \hline
\frac{\partial Q_{sorp}}{\partial \xi_{kin}} & \frac{\partial Q_{sorp}}{\partial \xi_{sorp}} & \frac{\partial Q_{sorp}}{\partial \xi_{min}} & \frac{\partial Q_{sorp}}{\partial \bar{\xi}_{kin}} & \frac{\partial Q_{sorp}}{\partial \bar{\xi}_{sorp}} & 0 & \frac{\partial Q_{sorp}}{\partial \xi_{mob}} \\ \hline
-\Lambda(\varphi_a)\frac{\partial E}{\partial \xi_{kin}} & -\Lambda(\varphi_a)\frac{\partial E}{\partial \xi_{sorp}} & -\Lambda(\varphi_a)\frac{\partial E}{\partial \xi_{min}} & 0 & 0 & \Lambda(\varphi_b) & -\Lambda(\varphi_a)\frac{\partial E}{\partial \xi_{mob}} \\ \hline
\frac{\partial Q_{mob}}{\partial \xi_{kin}} & \frac{\partial Q_{mob}}{\partial \xi_{sorp}} & \frac{\partial Q_{mob}}{\partial \xi_{min}} & 0 & 0 & 0 & \frac{\partial Q_{mob}}{\partial \xi_{mob}}
\end{array}\right)
$$

$$
\left(\begin{array}{c|c|c|c|c|c|c|c}
\mathrm{Id}+\tau L_h-\frac{\tau\,\partial f_1}{\partial \xi_{kin}} & -\frac{\tau\,\partial f_1}{\partial \xi_{sorp}} & -\frac{\tau\,\partial f_1}{\partial \xi^{\mathcal{A}}_{min}} & -\frac{\tau\,\partial f_1}{\partial \xi^{\mathcal{I}}_{min}} & -\frac{\tau\,\partial f_1}{\partial \bar{\xi}_{kin}} & -\frac{\tau\,\partial f_1}{\partial \bar{\xi}_{sorp}} & 0 & -\frac{\tau\,\partial f_1}{\partial \xi_{mob}} \\ \hline
-\frac{\tau\,\partial f_2}{\partial \xi_{kin}} & \mathrm{Id}+\tau L_h-\frac{\tau\,\partial f_2}{\partial \xi_{sorp}} & -\frac{\tau\,\partial f_2}{\partial \xi^{\mathcal{A}}_{min}} & -\frac{\tau\,\partial f_2}{\partial \xi^{\mathcal{I}}_{min}} & -\frac{\tau\,\partial f_2}{\partial \bar{\xi}_{kin}} & -\mathrm{Id}-\frac{\tau\,\partial f_2}{\partial \bar{\xi}_{sorp}} & 0 & -\frac{\tau\,\partial f_2}{\partial \xi_{mob}} \\ \hline
-\frac{\tau\,\partial f_3}{\partial \xi_{kin}} & -\frac{\tau\,\partial f_3}{\partial \xi_{sorp}} & \begin{array}{c}\mathrm{Id}_{\mathcal{A}}+\tau L_h^{\mathcal{A}\mathcal{A}}-\frac{\tau\,\partial f_{3\,\mathcal{A}}}{\partial \xi^{\mathcal{A}}_{min}} \\ \tau L_h^{\mathcal{I}\mathcal{A}}-\frac{\tau\,\partial f_{3\,\mathcal{I}}}{\partial \xi^{\mathcal{A}}_{min}}\end{array} & \begin{array}{c}\tau L_h^{\mathcal{A}\mathcal{I}}-\frac{\tau\,\partial f_{3\,\mathcal{A}}}{\partial \xi^{\mathcal{I}}_{min}} \\ \mathrm{Id}_{\mathcal{I}}+\tau L_h^{\mathcal{I}\mathcal{I}}-\frac{\tau\,\partial f_{3\,\mathcal{I}}}{\partial \xi^{\mathcal{I}}_{min}}\end{array} & -\frac{\tau\,\partial f_3}{\partial \bar{\xi}_{kin}} & -\frac{\tau\,\partial f_3}{\partial \bar{\xi}_{sorp}} & \begin{array}{cc}-\mathrm{Id}_{\mathcal{A}} & 0 \\ 0 & -\mathrm{Id}_{\mathcal{I}}\end{array} & -\frac{\tau\,\partial f_3}{\partial \xi_{mob}} \\ \hline
-\frac{\tau\,\partial f_4}{\partial \xi_{kin}} & -\frac{\tau\,\partial f_4}{\partial \xi_{sorp}} & -\frac{\tau\,\partial f_4}{\partial \xi^{\mathcal{A}}_{min}} & -\frac{\tau\,\partial f_4}{\partial \xi^{\mathcal{I}}_{min}} & \mathrm{Id}-\frac{\tau\,\partial f_4}{\partial \bar{\xi}_{kin}} & -\frac{\tau\,\partial f_4}{\partial \bar{\xi}_{sorp}} & 0 & -\frac{\tau\,\partial f_4}{\partial \xi_{mob}} \\ \hline
\frac{\partial Q_{sorp}}{\partial \xi_{kin}} & \frac{\partial Q_{sorp}}{\partial \xi_{sorp}} & \frac{\partial Q_{sorp}}{\partial \xi^{\mathcal{A}}_{min}} & \frac{\partial Q_{sorp}}{\partial \xi^{\mathcal{I}}_{min}} & \frac{\partial Q_{sorp}}{\partial \bar{\xi}_{kin}} & \frac{\partial Q_{sorp}}{\partial \bar{\xi}_{sorp}} & 0 & \frac{\partial Q_{sorp}}{\partial \xi_{mob}} \\ \hline
\begin{array}{c}0 \\ \frac{-\partial E_{\mathcal{I}}}{\partial \xi_{kin}}\end{array} & \begin{array}{c}0 \\ \frac{-\partial E_{\mathcal{I}}}{\partial \xi_{sorp}}\end{array} & \begin{array}{c}0 \\ \frac{-\partial E_{\mathcal{I}}}{\partial \xi_{min\,\mathcal{A}}}\end{array} & \begin{array}{c}0 \\ \frac{-\partial E_{\mathcal{I}}}{\partial \xi^{\mathcal{I}}_{min}}\end{array} & 0 & 0 & \begin{array}{cc}\mathrm{Id}_{\mathcal{A}} & 0 \\ 0 & 0\end{array} & \begin{array}{c}0 \\ \frac{-\partial E_{\mathcal{I}}}{\partial \xi_{mob}}\end{array} \\ \hline
\frac{\partial Q_{mob}}{\partial \xi_{kin}} & \frac{\partial Q_{mob}}{\partial \xi_{sorp}} & \frac{\partial Q_{mob}}{\partial \xi^{\mathcal{A}}_{min}} & \frac{\partial Q_{mob}}{\partial \xi^{\mathcal{I}}_{min}} & 0 & 0 & 0 & \frac{\partial Q_{mob}}{\partial \xi_{mob}}
\end{array}\right)
$$

4.4.3 Size reduction with the minimum-CP-function: an active set strategy

Let Ω_h the set of all grid points. The variables ξ_{min} and $\bar{\xi}_{min}$ from Sec. 4.3 consist of $J_{min} \times |\Omega_h|$ degrees of freedom each. We partition the set $M := \{1, ..., J_{min}\} \times \Omega_h$ into

$$\begin{aligned} \mathcal{A} &:= \{(i, x_k) \in \{1, ..., J_{min}\} \times \Omega_h \mid E_i(c(x_k)) > \bar{c}_{min,i}(x_k)\} \\ \mathcal{I} &:= \{(i, x_k) \in \{1, ..., J_{min}\} \times \Omega_h \mid E_i(c(x_k)) \leq \bar{c}_{min,i}(x_k)\} \end{aligned} \tag{4.22}$$

$\mathcal{A}$ is the set of *active* degrees of freedom in the sense of an active set method. Restricted to one species i we can define the set of active/inactive degrees of freedom

$$\begin{aligned} \mathcal{A}_i &:= \{x_k \in \Omega_h \,|\, (i, x_k) \in \mathcal{A}\}, \\ \mathcal{I}_i &:= \{x_k \in \Omega_h \,|\, (i, x_k) \in \mathcal{I}\}, \end{aligned}$$

$i = 1, ..., J_{min}$.
Let us emphasize that this partition into active and inactive degrees of freedom due to condition (4.22) is based on the evaluation of the current Newton iterate and is renewed after each Newton step. For the choice $\varphi = \varphi_{MIN}$ we recover the two cases (saturated solution $\leftrightarrow$ fully dissolved mineral) from Sec. 4.1, and the above defined index sets indicate which case for the current Newton iterate locally holds:

$$\varphi(E_i(c), \bar{c}_{min,i}) = \begin{cases} \bar{c}_{min,i}, & x \in \mathcal{A}_i \\ E_i(c), & x \in \mathcal{I}_i \end{cases}$$

The Semismooth Newton method with φ_{MIN} as a CP-function can thus be regarded as an active set strategy, a fact that has been stated in [47] and in [42]. Now we decompose the vector $\bar{\xi}_{min} \in \mathbb{R}^{|M|}$ into a vector $\bar{\xi}^{\mathcal{I}}_{min}$ living on $\mathcal{I}$ and a vector $\bar{\xi}^{\mathcal{A}}_{min}$ living on $\mathcal{A}$. The same decomposition is applied to vector ξ_{min}. By reordering the entries of vector $\bar{\xi}_{min}$, starting with the 'active' components and then proceeding with the inactive, we can write

$$\bar{\xi}_{min} = \begin{pmatrix} \bar{\xi}^{\mathcal{A}}_{min} \\ \bar{\xi}^{\mathcal{I}}_{min} \end{pmatrix}, \quad \xi_{min} = \begin{pmatrix} \xi^{\mathcal{A}}_{min} \\ \xi^{\mathcal{I}}_{min} \end{pmatrix}.$$

The same reordering can be applied to *equations*. The rows within the third and within the sixth set of equations of the linear system (4.20) reordered, starting with those equations acting on active nodes, followed by

those equations acting on inactive nodes. That means that we decompose $f_3, L_h\xi_{min}, G_3, G_6, E$ into the components $f_{3\mathcal{A}}, L_h^{\mathcal{A}}\xi_{min}, G_{3\mathcal{A}}, G_{6\mathcal{A}}, E_{\mathcal{A}}$, and $f_{3\mathcal{I}}, L_h^{\mathcal{I}}\xi_{min}, G_{3\mathcal{I}}, G_{6\mathcal{I}}, E_{\mathcal{I}}$:

$$f_3 = \begin{pmatrix} f_{3\mathcal{A}} \\ f_{3\mathcal{I}} \end{pmatrix}, \quad L_h\xi_{min} = \begin{pmatrix} L_h^{\mathcal{A}}\xi_{min} \\ L_h^{\mathcal{I}}\xi_{min} \end{pmatrix}, \quad E = \begin{pmatrix} E_{\mathcal{A}} \\ E_{\mathcal{I}} \end{pmatrix}, \quad S^1_{min} = (S^1_{min\mathcal{A}} | S^1_{min\mathcal{I}}),$$

etc. The discrete differential operators can be split into a part acting on the variables with $\mathcal{A}$-subscript and a part acting on the variables with $\mathcal{I}$-subscript:

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

With these definitions and restructuring, exploiting (4.21), the linear system (4.20) reads

$$J \begin{pmatrix} \frac{\Delta\xi_{kin}}{\Delta\xi_{sorp}} \\ \frac{\Delta\xi_{min}^{\mathcal{A}}}{\Delta\xi_{min}^{\mathcal{I}}} \\ \frac{\Delta\bar{\xi}_{kin}}{\Delta\bar{\xi}_{sorp}} \\ \frac{\Delta\bar{\xi}_{min}^{\mathcal{A}}}{\Delta\bar{\xi}_{min}^{\mathcal{I}}} \\ \frac{}{\Delta\xi_{mob}} \end{pmatrix} = - \begin{pmatrix} G_1 \\ G_2 \\ G_{3\mathcal{A}} \\ G_{3\mathcal{I}} \\ G_4 \\ G_5 \\ \bar{\xi}_{min}^{\mathcal{A}} \\ -E_{\mathcal{I}} \\ G_7 \end{pmatrix} \tag{4.23}$$

where J is displayed in Table 4.1, right part. We emphasize that the zeros in this matrix and one of the $\mathrm{Id}_{\mathcal{A}}$-blocks are due to the specific choice $\varphi = \varphi_{MIN}$ for the CF-function. The presence of these blocks can be exploited to decouple some equations: The upper part of the sixth block of equations simply reads

$$\Delta\bar{\xi}_{min}^{\mathcal{A}} = -\bar{\xi}_{min}^{\mathcal{A}},$$

where on the right-hand side the *old* Newton iterate is meant; subscripts denoting the Newton iterates are suppressed here and in the following. Hence, the update for the variable $\bar{\xi}_{min}^{\mathcal{A}}$ just reads

$$\bar{\xi}_{min}^{\mathcal{A}} := 0, \tag{4.24}$$

i.e., active mineral degrees of freedom are directly set to zero within a Newton step. Hence, the linear system (4.23) can be diminished by these equations/unknowns. Another decoupling is possible since the unknowns $\Delta\bar{\xi}^{\mathcal{I}}_{min}$ *only* occur in the lower part of the third block of equations. This block of equations can be written as

$$\begin{aligned}\Delta\bar{\xi}^{\mathcal{I}}_{min} := \; & -G_{3\mathcal{I}} \\ & -\tau\frac{\partial f_{3\mathcal{I}}}{\partial\xi_{kin}}\Delta\xi_{kin} - \tau\frac{\partial f_{3\mathcal{I}}}{\partial\xi_{sorp}}\Delta\xi_{sorp} + \left(\mathrm{Id}_{\mathcal{I}} + \tau L_h^{\mathcal{I}} - \tau\frac{\partial f_{3\mathcal{I}}}{\partial\xi_{min}}\right)\Delta\xi_{min} \\ & -\tau\frac{\partial f_{3\mathcal{I}}}{\partial\bar{\xi}_{kin}}\Delta\bar{\xi}_{kin} - \tau\frac{\partial f_{3\mathcal{I}}}{\partial\bar{\xi}_{sorp}}\Delta\bar{\xi}_{sorp} - \tau\frac{\partial f_{3I}}{\partial\xi_{mob}}\Delta\xi_{mob}.\end{aligned}$$

This block of equations can be solved *a posteriori* after the rest of the system has been solved for the other unknowns. That means that the mineral concentrations (in the linear problem) are expressed as a function of the non-mineral concentrations.
After these two decouplings the remaining linear system reads

$$J\begin{pmatrix}\Delta\xi_{kin}\\ \hline \Delta\xi_{sorp}\\ \hline \Delta\xi^{\mathcal{A}}_{min}\\ \Delta\xi^{\mathcal{I}}_{min}\\ \hline \Delta\bar{\xi}_{kin}\\ \hline \Delta\bar{\xi}_{sorp}\\ \hline \Delta\xi_{mob}\end{pmatrix} = -\begin{pmatrix}G_1\\ G_2\\ G_{3\mathcal{A}}\\ G_4\\ G_5\\ -E_{\mathcal{I}}\\ G_7\end{pmatrix} \tag{4.25}$$

with system matrix J as it is displayed in Table 4.2, left part. The decoupling of the mineral concentrations $\bar{\xi}^{\mathcal{A}}_{min}$, $\bar{\xi}^{\mathcal{I}}_{min}$ has an analogon on the *PDE* level (4.17):

$$\bar{\xi}_{min,i} = 0 \quad \vee \quad \partial_t\bar{\xi}_{min,i} = \partial_t\xi_{min,i} + L\xi_{min,i} - f_{3,i}(\xi,\bar{\xi})$$

With the intention to separate local from global unknowns, we reorder the unknowns in (4.25) by moving the position of the entries $\Delta\xi^{\mathcal{I}}_{min}$. This requires to move also the position of the fourth column of the matrix (Table 4.2, left

$$
\left(\begin{array}{c|c|c|c|c|c|c}
\mathrm{Id}+\tau L_h-\frac{\tau\,\partial f_1}{\partial \xi_{kin}} & -\frac{\tau\,\partial f_1}{\partial \xi_{sorp}} & -\frac{\tau\,\partial f_1}{\partial \xi^{\mathcal{A}}_{min}} & -\frac{\tau\,\partial f_1}{\partial \xi^{\mathcal{I}}_{min}} & -\frac{\tau\,\partial f_1}{\partial \bar{\xi}_{kin}} & -\frac{\tau\,\partial f_1}{\partial \bar{\xi}_{sorp}} & -\frac{\tau\,\partial f_1}{\partial \xi_{mob}} \\
\hline
-\frac{\tau\,\partial f_2}{\partial \xi_{kin}} & \mathrm{Id}+\tau L_h-\frac{\tau\,\partial f_2}{\partial \xi_{sorp}} & -\frac{\tau\,\partial f_2}{\partial \xi^{\mathcal{A}}_{min}} & -\frac{\tau\,\partial f_2}{\partial \xi^{\mathcal{I}}_{min}} & -\frac{\tau\,\partial f_2}{\partial \bar{\xi}_{kin}} & -\mathrm{Id}-\frac{\tau\,\partial f_2}{\partial \bar{\xi}_{sorp}} & -\frac{\tau\,\partial f_2}{\partial \xi_{mob}} \\
\hline
-\tau\,\frac{\partial f_{3\,\mathcal{A}}}{\partial \xi_{kin}} & -\tau\,\frac{\partial f_{3\,\mathcal{A}}}{\partial \xi_{sorp}} & \mathrm{Id}_{\mathcal{A}}+\tau L_h^{\mathcal{A}\mathcal{A}}-\frac{\tau\,\partial f_{3\,\mathcal{A}}}{\partial \xi^{\mathcal{A}}_{min}} & \tau L_h^{\mathcal{A}\mathcal{I}}-\frac{\tau\,\partial f_{3\,\mathcal{A}}}{\partial \xi^{\mathcal{I}}_{min}} & -\tau\,\frac{\partial f_{3\,\mathcal{A}}}{\partial \bar{\xi}_{kin}} & -\tau\,\frac{\partial f_{3\,\mathcal{A}}}{\partial \bar{\xi}_{sorp}} & -\frac{\tau\,\partial f_{3\,\mathcal{A}}}{\partial \xi_{mob}} \\
\hline
-\frac{\tau\,\partial f_4}{\partial \xi_{kin}} & -\frac{\tau\,\partial f_4}{\partial \xi_{sorp}} & -\frac{\tau\,\partial f_4}{\partial \xi^{\mathcal{A}}_{min}} & -\frac{\tau\,\partial f_4}{\partial \xi^{\mathcal{I}}_{min}} & \mathrm{Id}-\frac{\tau\,\partial f_4}{\partial \bar{\xi}_{kin}} & -\frac{\tau\,\partial f_4}{\partial \bar{\xi}_{sorp}} & -\frac{\tau\,\partial f_4}{\partial \xi_{mob}} \\
\hline
\frac{\partial Q_{sorp}}{\partial \xi_{kin}} & \frac{\partial Q_{sorp}}{\partial \xi_{sorp}} & \frac{\partial Q_{sorp}}{\partial \xi^{\mathcal{A}}_{min}} & \frac{\partial Q_{sorp}}{\partial \xi^{\mathcal{I}}_{min}} & \frac{\partial Q_{sorp}}{\partial \bar{\xi}_{kin}} & \frac{\partial Q_{sorp}}{\partial \bar{\xi}_{sorp}} & \frac{\partial Q_{sorp}}{\partial \xi_{mob}} \\
\hline
\frac{-\partial E_{\mathcal{I}}}{\partial \xi_{kin}} & \frac{-\partial E_{\mathcal{I}}}{\partial \xi_{sorp}} & \frac{-\partial E_{\mathcal{I}}}{\partial \xi^{\mathcal{A}}_{min}} & \frac{-\partial E_{\mathcal{I}}}{\partial \xi^{\mathcal{I}}_{min}} & 0 & 0 & \frac{-\partial E_{\mathcal{I}}}{\partial \xi_{mob}} \\
\hline
\frac{\partial Q_{mob}}{\partial \xi_{kin}} & \frac{\partial Q_{mob}}{\partial \xi_{sorp}} & \frac{\partial Q_{mob}}{\partial \xi^{\mathcal{A}}_{min}} & \frac{\partial Q_{mob}}{\partial \xi^{\mathcal{I}}_{min}} & 0 & 0 & \frac{\partial Q_{mob}}{\partial \xi_{mob}}
\end{array}\right)
$$

$$
\left(\begin{array}{c|c|c||c|c|c|c}
\mathrm{Id}+\tau L_h-\frac{\tau\,\partial f_1}{\partial \xi_{kin}} & -\frac{\tau\,\partial f_1}{\partial \xi_{sorp}} & -\frac{\tau\,\partial f_1}{\partial \xi^{\mathcal{A}}_{min}} & -\frac{\tau\,\partial f_1}{\partial \bar{\xi}_{kin}} & -\frac{\tau\,\partial f_1}{\partial \bar{\xi}_{sorp}} & -\frac{\tau\,\partial f_1}{\partial \xi^{\mathcal{I}}_{min}} & -\frac{\tau\,\partial f_1}{\partial \xi_{mob}} \\
\hline
-\frac{\tau\,\partial f_2}{\partial \xi_{kin}} & \mathrm{Id}+\tau L_h-\frac{\tau\,\partial f_2}{\partial \xi_{sorp}} & -\frac{\tau\,\partial f_2}{\partial \xi^{\mathcal{A}}_{min}} & -\frac{\tau\,\partial f_2}{\partial \bar{\xi}_{kin}} & -\mathrm{Id}-\frac{\tau\,\partial f_2}{\partial \bar{\xi}_{sorp}} & -\frac{\tau\,\partial f_2}{\partial \xi^{\mathcal{I}}_{min}} & -\frac{\tau\,\partial f_2}{\partial \xi_{mob}} \\
\hline
-\tau\,\frac{\partial f_{3\,\mathcal{A}}}{\partial \xi_{kin}} & -\tau\,\frac{\partial f_{3\,\mathcal{A}}}{\partial \xi_{sorp}} & \mathrm{Id}_{\mathcal{A}}+\tau L_h^{\mathcal{A}\mathcal{A}}-\frac{\tau\,\partial f_{3\,\mathcal{A}}}{\partial \xi^{\mathcal{A}}_{min}} & -\tau\,\frac{\partial f_{3\,\mathcal{A}}}{\partial \bar{\xi}_{kin}} & -\tau\,\frac{\partial f_{3\,\mathcal{A}}}{\partial \bar{\xi}_{sorp}} & \tau L_h^{\mathcal{A}\mathcal{I}}-\frac{\tau\,\partial f_{3\,\mathcal{A}}}{\partial \xi^{\mathcal{I}}_{min}} & -\frac{\tau\,\partial f_{3\,\mathcal{A}}}{\partial \xi_{mob}} \\
\hline
-\frac{\tau\,\partial f_4}{\partial \xi_{kin}} & -\frac{\tau\,\partial f_4}{\partial \xi_{sorp}} & -\frac{\tau\,\partial f_4}{\partial \xi^{\mathcal{A}}_{min}} & \mathrm{Id}-\frac{\tau\,\partial f_4}{\partial \bar{\xi}_{kin}} & -\frac{\tau\,\partial f_4}{\partial \bar{\xi}_{sorp}} & -\frac{\tau\,\partial f_4}{\partial \xi^{\mathcal{I}}_{min}} & -\frac{\tau\,\partial f_4}{\partial \xi_{mob}} \\
\hline
\frac{\partial Q_{sorp}}{\partial \xi_{kin}} & \frac{\partial Q_{sorp}}{\partial \xi_{sorp}} & \frac{\partial Q_{sorp}}{\partial \xi^{\mathcal{A}}_{min}} & \frac{\partial Q_{sorp}}{\partial \bar{\xi}_{kin}} & \frac{\partial Q_{sorp}}{\partial \bar{\xi}_{sorp}} & \frac{\partial Q_{sorp}}{\partial \xi^{\mathcal{I}}_{min}} & \frac{\partial Q_{sorp}}{\partial \xi_{mob}} \\
\hline
\frac{-\partial E_{\mathcal{I}}}{\partial \xi_{kin}} & \frac{-\partial E_{\mathcal{I}}}{\partial \xi_{sorp}} & \frac{-\partial E_{\mathcal{I}}}{\partial \xi^{\mathcal{A}}_{min}} & 0 & 0 & \frac{-\partial E_{\mathcal{I}}}{\partial \xi^{\mathcal{I}}_{min}} & \frac{-\partial E_{\mathcal{I}}}{\partial \xi_{mob}} \\
\hline
\frac{\partial Q_{mob}}{\partial \xi_{kin}} & \frac{\partial Q_{mob}}{\partial \xi_{sorp}} & \frac{\partial Q_{mob}}{\partial \xi^{\mathcal{A}}_{min}} & 0 & 0 & \frac{\partial Q_{mob}}{\partial \xi^{\mathcal{I}}_{min}} & \frac{\partial Q_{mob}}{\partial \xi_{mob}}
\end{array}\right)
$$

Table 4.2: Left: Linear system (4.25) after decoupling of a posteriori equations. – Right: The system after reordering unknowns (4.26).

part) to the right. We get the system

$$J \left(\begin{array}{c} \Delta\xi_{kin} \\ \Delta\xi_{sorp} \\ \Delta\xi_{min}^{\mathcal{A}} \\ \hline \Delta\bar{\xi}_{kin} \\ \Delta\bar{\xi}_{sorp} \\ \Delta\xi_{min}^{\mathcal{I}} \\ \Delta\xi_{mob} \end{array}\right) = - \left(\begin{array}{c} G_1 \\ G_2 \\ G_{3_{\mathcal{A}}} \\ \hline G_4 \\ G_5 \\ -E_{\mathcal{I}} \\ G_7 \end{array}\right) \tag{4.26}$$

with system matrix J as it is displayed in Table 4.2, right part.
The lower right part of the system matrix does not contain any spatial couplings. Hence, a Schur complement technique can be applied to this system, which requires only the solution of *local* problems. It is worthwile to check if a proof of solvability of the local problem can be found. This is done in the following section.

4.4.4 Solvability of the local problem

As already done in Sec. 2.4.4 for the system without minerals, we want to investigate the solvability of the local problem in the extended model. Let us write system (4.26) as

$$\left(\begin{array}{cc} A & B \\ C & D \end{array}\right) \left(\begin{array}{c} \Delta\xi_{glob} \\ \Delta\xi_{loc} \end{array}\right) = \left(\begin{array}{c} \mathcal{G}_{glob} \\ \mathcal{G}_{loc} \end{array}\right) \tag{4.27}$$

with

$$\Delta\xi_{glob} := \begin{pmatrix} \Delta\xi_{kin} \\ \Delta\xi_{sorp} \\ \Delta\xi_{min}^{\mathcal{A}} \end{pmatrix}, \quad \Delta\xi_{loc} := \begin{pmatrix} \Delta\bar{\xi}_{kin} \\ \Delta\bar{\xi}_{sorp} \\ \Delta\xi_{min}^{\mathcal{I}} \\ \Delta\xi_{mob} \end{pmatrix}, \quad \mathcal{G}_{glob} := \begin{pmatrix} G_1 \\ G_2 \\ G_{3_{\mathcal{A}}} \end{pmatrix}, \quad \mathcal{G}_{loc} := \begin{pmatrix} G_4 \\ G_5 \\ -E_{\mathcal{I}} \\ G_7 \end{pmatrix}.$$

and (cf. Table 4.2, right)

$$D = \left(\begin{array}{c|c||c|c} \mathrm{Id} - \frac{\tau\,\partial f_4}{\partial\bar{\xi}_{kin}} & -\frac{\tau\,\partial f_4}{\partial\bar{\xi}_{sorp}} & -\frac{\tau\,\partial f_4}{\partial\xi_{min}^{\mathcal{I}}} & -\frac{\tau\,\partial f_4}{\partial\xi_{mob}} \\ \hline \frac{\partial Q_{sorp}}{\partial\bar{\xi}_{kin}} & \frac{\partial Q_{sorp}}{\partial\bar{\xi}_{sorp}} & \frac{\partial Q_{sorp}}{\partial\xi_{min}^{\mathcal{I}}} & \frac{\partial Q_{sorp}}{\partial\xi_{mob}} \\ \hline\hline 0 & 0 & -\frac{\partial E_{\mathcal{I}}}{\partial\xi_{min}^{\mathcal{I}}} & -\frac{\partial E_{\mathcal{I}}}{\partial\xi_{mob}} \\ \hline 0 & 0 & \frac{\partial Q_{mob}}{\partial\xi_{min}^{\mathcal{I}}} & \frac{\partial Q_{mob}}{\partial\xi_{mob}} \end{array}\right)$$

Lemma 4.4 *If the discrete values of all non-minerals c_i, $\bar{c}_{nmin,i}$ are strictly positive at each mesh point and if the time step size $\tau > 0$ is sufficiently small, then the matrix D is regular, i.e., problem (4.26)/(4.27) is equivalent to the (smaller) problem*

$$(A - BD^{-1}C)\,\Delta\xi_{glob} = \mathcal{G}_{glob} - BD^{-1}\mathcal{G}_{loc}. \tag{4.28}$$

If $J^{2'}_{kin} = 0$ (which in particular holds if all reactions are equilibrium reactions), then the statement even holds for arbitrarily large τ.

Proof. Since D is void of any spatial couplings, the regularity of D can be reduced to the regularity of a block of D at a specific mesh point. Let us denote the *local* matrix at an arbitrary mesh point again by D. We obtain

$$\begin{aligned} D \;=\; & \left(\begin{array}{c|c||c|c} \mathrm{Id} - O(\tau) & O(\tau) & O(\tau) & O(\tau) \\ \hline {S^2_{sorp}}^T \tilde{\Lambda} S^{2'}_{kin} & {S^2_{sorp}}^T \tilde{\Lambda} S^2_{sorp} & {S^1_{sorp}}^T \Lambda S^1_{min\mathcal{I}} & {S^1_{sorp}}^T \Lambda S^1_{mob} \\ \hline\hline 0 & 0 & {S^1_{min\mathcal{I}}}^T \Lambda S^1_{min\mathcal{I}} & {S^1_{min\mathcal{I}}}^T \Lambda S^1_{mob} \\ \hline 0 & 0 & {S^1_{mob}}^T \Lambda S^1_{min\mathcal{I}} & {S^1_{mob}}^T \Lambda S^1_{mob} \end{array}\right) \\ =: \; & \left(\begin{array}{c|c} D_{11} & D_{12} \\ \hline 0 & D_{22} \end{array}\right) \end{aligned} \tag{4.29}$$

with $\tilde{\Lambda} := \mathrm{diag}(1/\bar{c}_{nmin,i}) \in \mathbb{R}^{\bar{I}_{nmin} \times \bar{I}_{nmin}}$ and $\Lambda = \mathrm{diag}(1/c_i) \in \mathbb{R}^{I \times I}$.
It is sufficient to prove the regularity of D_{11} and D_{22}. We can write

$$D_{22} = (S^1_{min\mathcal{I}} | S^1_{mob})^T \, \Lambda \, (S^1_{min\mathcal{I}} | S^1_{mob}).$$

Due to the assumption of linear independence of the columns of matrix $(S^1_{min\mathcal{I}} | S^1_{mob})$ and the positivity of the c_i, matrix D_{22} is symmetric positive definite. The entries of D_{11} depend continuously on the time step size τ, and for $\tau = 0$ matrix D_{11} becomes

$$D_{11_0} = \left(\begin{array}{c|c} \mathrm{Id} & 0 \\ \hline {S^2_{sorp}}^T \tilde{\Lambda} S^{2'}_{kin} & {S^2_{sorp}}^T \tilde{\Lambda} S^2_{sorp} \end{array}\right)$$

Matrix D_{11_0} is obviously regular, since its lower right block is symmetric positive definite. Then there is a neighbourhood around D_{11_0} where all matrices are regular. Hence, for sufficiently small τ, D_{11} is regular.
The τ-depending terms in D_{11} are nonexistent, if $J^{2'}_{kin} = 0$. ∎

Note that the practical solution of the local problem at each mesh point can be split into two even smaller problems, exploiting the block structure, by solving problems related to the blocks D_{22}, D_{11}, i.e., one problem of size $J^{\mathcal{I}}_{min} + J_{mob}$ and one problem of size $J^{2'}_{kin} + J_{sorp}$ per mesh point, if $J^{\mathcal{I}}_{mob} = J^{\mathcal{I}}_{mob}(x_h)$ denotes the number of nonactive degrees of freedom at the specific mesh point x_h.

4.4.5 Structure of the global problem

Similar as in Sec. 2.4.5 for the system without minerals, we can also investigate the structure of the global Jacobian (4.28) for problems with minerals, in particular the fraction of (4.28) that comes from the accumulation terms and that remains for time step $\tau \longrightarrow 0$. Again, the result is that at least for very small τ, where transport and kinetic reactions can be considered as a perturbation of the pure equilibrium problem (without transport and kinetic reactions), the global Jacobian is guaranteed to be regular. In terms of Def. 4.2 and Theorem 4.3, this amounts to a proof of *BD-regularity* for sufficiently small τ, which may serve as a starting point for the investigation of convergence properties etc. for the Semismooth Newton[4] applied to the reactive transport problem with minerals.

Lemma 4.5 *Let the current Newton iterate of the global problem be such that the corresponding value in terms of* c, $\bar{c}_{nmin}$, $\bar{c}_{min}$ *lies in* $\mathbb{R}_+^{I+\bar{I}_{nmin}} \times \mathbb{R}^{\bar{I}_{min}}$ *at each of the* $m \in \mathbb{N}$ *mesh points. The system matrix* $A - BC^{-1}D$ *of the global problem, after the local equations have been eliminated, (4.28), has the structure*

$$\begin{aligned} \mathcal{J}_{glob} &= Id + \mathcal{J}_{equil} + \tau\, \mathcal{J}_{transp} + \tau \mathcal{J}_{bdry} + \tau\, \mathcal{J}_{kinet} \\ &\in \mathbb{R}^{(mJ^{1'}_{kin} + mJ_{sorp} + |\mathcal{A}|) \times (mJ^{1'}_{kin} + mJ_{sorp} + |\mathcal{A}|)} \end{aligned}$$

where $\mathcal{J}_{transp}$ *contains the transport terms* L_h*, where* $\mathcal{J}_{bdry}$ *contains the terms containing* $L_h^{\mathcal{AI}}$*, where* $\mathcal{J}_{kinet}$ *contains terms from kinetic reactions*[5]*, and where the* τ*-independent part (coming from the accumulation term) has the*

[4] Such an investigation is planned by C. Kanzow and H. Buchholz for the very next future in the framework of a DFG-supported joint project.

[5] $\mathcal{J}_{kinet}$ may depend on τ due to the substitution of the local resolution function, but only as $O(1)$ for $\tau \to 0$.

structure

$$Id + \mathcal{J}_{equil} = Id + \left(\begin{array}{c|c|c} 0 & 0 & 0 \\ \hline U_1 & U_2 & U_3 \\ \hline 0 & 0 & 0 \end{array}\right),$$

Here, U_2 is a quadratic matrix with a complete system of eigenvectors with all eigenvalues being strictly positive. As a consequence, $\mathcal{J}_{equil}$ ($Id+\mathcal{J}_{equil}$, resp.) has a complete set of eigenvectors with all eigenvalues being nonnegative (strictly positive, resp.). Hence, for fixed Newton iterate and time step size $\tau>0$ sufficiently small, $\mathcal{J}_{glob}$ is regular.

Proof. The statement on the eigenvectors and eigenvalues of U_2 is given in Lemma 5.11 in the appendix. The consequence for the eigenvectors and eigenvalues of $\mathcal{J}_{equil}$ follows obviously by considering the dimension of the kernel of $\mathcal{J}_{equil}$: The (all positive) eigenvalues of U_2 are obviously also eigenvalues of $\mathcal{J}_{equil}$, and the corresponding eigenvectors of $\mathcal{J}_{equil}$ together with a basis of the kernel of $\mathcal{J}_{equil}$ form a basis of the whole space. The result on the regularity of $\mathcal{J}_{glob}$ for sufficiently small τ follows by a continuity argument. ∎
Explicit formulas for U_1, U_2, U_3, J_{bdry} are given in Lemma 5.11.
Let us recall that the regularity of the reduced global system matrix $\mathcal{J}_{glob}=A–BD^{-1}C$ is of course equivalent to the regularity of the non-reduced system matrix $J = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ from (4.26)/(4.27), since the regularity of D has been settled in Lemma 4.4.

4.4.6 A variant: Elimination on the nonlinear level

As mentioned in 4.4.1 it is thinkable to eliminate the local equations already on the *nonlinear* level. A practical reason to implement an elimination on the nonlinear level is the fact that this can easier be implemented into a code of the algorithms of Chapter 2 for mineral-free problems, which also use elimination on the nonlinear level. Note that a drawback of elimination for the reactive-transport-*mineral* problem on the *non*linear level is that from a theoretical point of view this means to solve in each Newton step a *different* nonlinear equation; in fact, even the *size* of the problem varies from Newton step to Newton step, since the cardinality of the active/inactive sets varies. This might complicate theoretical investigations. However, this version was implemented and showed satisfying results (see Sec. 4.5).

Exploiting the choice $\varphi = \varphi_{MIN}$ (4.14-b), the nonlinear discrete system (4.19) reads

$$\begin{aligned}
G_1 &:= \xi_{kin} + \tau L_h \xi_{kin} - \tau f_1(\xi_{mob}, \xi_{sorp}, \xi_{min}, \xi_{kin}, \bar{\xi}_{sorp}, \bar{\xi}_{kin}) - \xi_{kin}^{old} = 0 \\
G_2 &:= \xi_{sorp} - \bar{\xi}_{sorp} + \tau L_h \xi_{sorp} - \tau f_2(\xi_{mob}, \xi_{sorp}, \xi_{min}, \xi_{kin}, \bar{\xi}_{sorp}, \bar{\xi}_{kin}) - \xi_{sorp}^{old} + \bar{\xi}_{sorp}^{old} = 0 \\
G_3 &:= \xi_{min} - \bar{\xi}_{min} + \tau L_h \xi_{min} - \tau f_3(\xi_{mob}, \xi_{sorp}, \xi_{min}, \xi_{kin}, \bar{\xi}_{sorp}, \bar{\xi}_{kin}) - \xi_{min}^{old} + \bar{\xi}_{min}^{old} = 0 \\
G_4 &:= \bar{\xi}_{kin} - \tau f_4(\xi_{mob}, \xi_{sorp}, \xi_{min}, \xi_{kin}, \bar{\xi}_{sorp}, \bar{\xi}_{kin}) - \bar{\xi}_{kin}^{old} = 0 \\
G_5 &:= Q_{sorp}(\xi_{mob}, \xi_{sorp}, \xi_{min}, \xi_{kin}, \bar{\xi}_{sorp}, \bar{\xi}_{kin}) = 0 \\
G_{6\mathcal{A}} &:= \bar{\xi}_{min}^{\mathcal{A}} = 0 \\
G_{6\mathcal{I}} &:= -E_{\mathcal{I}}(\xi_{mob}, \xi_{sorp}, \xi_{min}, \xi_{kin})) = 0 \\
G_7 &:= Q_{mob}(\xi_{mob}, \xi_{sorp}, \xi_{min}, \xi_{kin}) = 0
\end{aligned} \tag{4.30}$$

The (decoupled) equations for η, $\bar{\eta}$ were omitted again as well as any dependence of functions on these variables. The variables ξ_{min} and $\bar{\xi}_{min}$ have been decomposed into active and inactive degrees of freedom as in Sec. 4.4.3. This decomposition is based on the evaluation of the variables of the current Newton iteration; a subscript indicating the Newton iterate is suppressed. The equation $G_{6\mathcal{A}}$ allows the elimination of the unknown $\bar{\xi}_{min}^{\mathcal{A}}$ from the system. Again, we can split equation G_3 into a part acting on the active/on the inactive degrees of freedom, $G_{3\mathcal{A}}, G_{3\mathcal{I}}$. The equation $G_{3\mathcal{I}}$ can be used for an a posteriori computation of $\bar{\xi}_{min}^{\mathcal{I}}$, since this unknown does not occur in any of the other equations. Hence, we obtain the system

$$\begin{aligned}
&\text{(i)} && \xi_{kin} + \tau L_h \xi_{kin} - \tau f_1(\xi_{mob}, ..., \bar{\xi}_{kin}) - \xi_{kin}^{old} = 0 \\
&\text{(ii)} && \xi_{sorp} - \bar{\xi}_{sorp} + \tau L_h \xi_{sorp} - \tau f_2(\xi_{mob}, ..., \bar{\xi}_{kin}) - \xi_{sorp}^{old} + \bar{\xi}_{sorp}^{old} = 0 \\
&\text{(iii)} && \xi_{min}^{\mathcal{A}} + \tau L_h^{\mathcal{A}\mathcal{A}} \xi_{min}^{\mathcal{A}} + \tau L_h^{\mathcal{A}\mathcal{I}} \xi_{min}^{\mathcal{I}} - \tau f_3^{\mathcal{A}}(\xi_{mob}, ..., \bar{\xi}_{kin}) - \xi_{min}^{old\,\mathcal{A}} + \bar{\xi}_{min}^{old\,\mathcal{A}} = 0 \\
&\text{(iv)} && \bar{\xi}_{kin} - \tau f_4(\xi_{mob}, ..., \bar{\xi}_{kin}) - \bar{\xi}_{kin}^{old} = 0 \\
&\text{(v)} && Q_{mob}(\xi_{mob}, \xi_{sorp}, \xi_{min}, \xi_{kin}) = 0 \\
&\text{(vi)} && Q_{sorp}(\xi_{mob}, \xi_{sorp}, \xi_{min}, \xi_{kin}, \bar{\xi}_{sorp}, \bar{\xi}_{kin}) = 0 \\
&\text{(vii)} && -E_{\mathcal{I}}(\xi_{mob}, \xi_{sorp}, \xi_{min}, \xi_{kin}) = 0
\end{aligned} \tag{4.31}$$

with the decoupled equations

$$\bar{\xi}_{min}^{\mathcal{I}} := \xi_{min} + \tau L_h^{\mathcal{I}} \xi_{min} - \tau f_{3\mathcal{I}}(\xi_{mob}, \xi_{sorp}, \xi_{min}, \xi_{kin}, \bar{\xi}_{sorp}, \bar{\xi}_{kin}) - \xi_{min}^{old,\mathcal{I}} + \bar{\xi}_{min}^{old,\mathcal{I}},$$
$$\bar{\xi}_{min}^{\mathcal{A}} := 0.$$

The ODE (iv) and the AEs (v)-(vii) are local equations. As in Chapter 2, a reduction of the problem size can be achieved by solving the local equation

with respect to local unknowns and to eliminate these unknowns from the remaining equations. As seen in Sec. 2.4.3 the algorithmic treatment leads to a local Newton iteration nested in the global Newton iteration. Concerning the local Newton iteration, the local variables fulfilling (iv)-(vii) provided given values for the global variables, are wanted. Let us check the existence of a local resolution function of this problem.

Lemma 4.6 *Let us fix a point* $(c^0, \bar{c}^0_{nmin}, \bar{c}^0_{min}) \in \mathbb{R}^{I+\bar{I}_{nmin}}_+ \times \mathbb{R}^{\bar{I}_{min}}$, *and let* $\eta^0, \xi^0, \bar{\eta}^0, \bar{\xi}^0$ *be the transformed coordinates of this point.*

(i) There is a unique local resolution function $\xi_{loc} = \xi_{loc}(\xi_{glob})$ *for the equations (v)-(vii) around this point, where*

$$\xi_{loc} = (\xi_{mob}, \xi^{\mathcal{I}}_{min}, \bar{\xi}_{sorp}), \quad \xi_{glob} = (\xi_{kin}, \xi_{sorp}, \xi^{\mathcal{A}}_{min}, \bar{\xi}_{kin}).$$

(ii) For $\tau > 0$ *sufficiently small there is a unique local resolution function* $\xi_{loc} = \xi_{loc}(\xi_{glob})$ *for the equations (iv)-(vii), where*

$$\xi_{loc} = (\xi_{mob}, \xi^{\mathcal{I}}_{min}, \bar{\xi}_{sorp}, \bar{\xi}_{kin}), \quad \xi_{glob} = (\xi_{kin}, \xi_{sorp}, \xi^{\mathcal{A}}_{min}).$$

Proof. See proof of Lemma 4.4, since the local problem has exactly the same structure. ∎

The remaining global nonlinear discrete problem reads

$$\begin{aligned}
&\xi_{kin} && +\tau L_h \xi_{kin} && -\tau f_1(\xi_{glob}, \xi_{loc}(\xi_{glob})) - \xi^{old}_{kin} && = 0\\
&\xi_{sorp} - \bar{\xi}_{sorp}(\xi_{glob}) && +\tau L_h \xi_{sorp} && -\tau f_2(\xi_{glob}, \xi_{loc}(\xi_{glob})) - \xi^{old}_{sorp} + \bar{\xi}^{old}_{sorp} && = 0\\
&\xi^{\mathcal{A}}_{min} && + \tau L^{\mathcal{A}\mathcal{A}}_h \xi^{\mathcal{A}}_{min} && - \tau f^{\mathcal{A}}_3(\xi_{glob}, \xi_{loc}(\xi_{glob})) - \xi^{old\ \mathcal{A}}_{min} + \bar{\xi}^{old\,\mathcal{A}}_{min} + \tau L^{\mathcal{A}\mathcal{I}}_h \xi^{\mathcal{I}}_{min}(\xi_{glob}) && = 0
\end{aligned}$$

if the elimination is performed according to part (ii) of the previous lemma. The statement made for the formulation of Chapter 2, that there are no coupling terms or nonlinear terms under the transport operator, requires a minor modification: At those mesh points lying close to the boundary of the active domain $\mathcal{A}_i$, the 'boundary term' $\tau L^{\mathcal{A}\mathcal{I}}_h \xi^{\mathcal{I}}_{min}(\xi_{glob})$ also contains the resolution function.

4.5 Numerical results

The reactive transport problem in the version of Sec. 4.4.6 was implemented using M++ [80], thus extending the software of Chapter 2. Some numerical experiments were run in order to validate the code and to check that

complementarity formulation/active set strategy reproduces correct numerical solutions, in particular correct speeds of precipitation/dissolution fronts. In Sec. 4.5.1, 4.5.2 we present two of them.

4.5.1 Test example 1: Dislocation of a mineral by dissolution/precipitation

We consider a very simple example consisting of two mobile species A, B and one mineral C. We consider one precipitation/dissolution-reaction (4.3).
The scenario is the following: A concentration of A and B is given on two disjoint, adjacent parts of the computational domain (see Fig. 4.2, upper row). Strong dispersive mixing of A and B leads to precipitation of mineral C in those regions where the solubility product is reached. Then (slow) advective transport of A, B lets the value of $c_A c_B^2$ fall below the solubility product, which causes an immediate dissolution of the precipitated C. A dislocation of the (immobile!) mineral takes places through a dissolution of C into A, B in one place, transport of A, B, and a precipitation of C at another place.
The technical data are as follows: domain $\Omega = (0, 10) \times (0, 6)$, number of elements (P1)= 16384, Darcy velocity $q = (1, 0)^T$, water content $\theta = 0.396$, (i.e., pore velocity $|q|/\theta = 2.525$), longitudinal/transversal dispersion length $(\beta_l, \beta_t) = (0.2, 0.25)^T$, time step size $\tau = 0.1$. The solubility product of the reaction is $K = 1$; i.e., $c_A c_B^2 = 1$ (if $c_C > 0$).
The initial values are $c_A = 4$, $c_B = 2$ in a part of the computational domain, and a background concentration of 0.25 for both A and B and 0.0 for C.
80 time steps of length $\tau = 0.05$ were performed.
Obviously the CP-formulation with the Semismooth Newton Methods handles the change from active to inactive (precipitation) and from inactive to active (dissolution) well. The average number of Newton steps per time step was 4.49; the average number of local Newton steps per global Newton step and per node was 0.44, which is quite satisfying for the chosen time step size (the CFL number, with respect to the pore velocity, larger than one in this program run). Note that the asymmetry between A and B in Fig. 4.2 is caused by the different stoichiometric coefficients for A and B ("$c_A c_B^2 = 1$").

4.5.2 Test example 2: Sequestration of carbon dioxide

In test example 2 the interaction of CO_2 with minerals is considered.

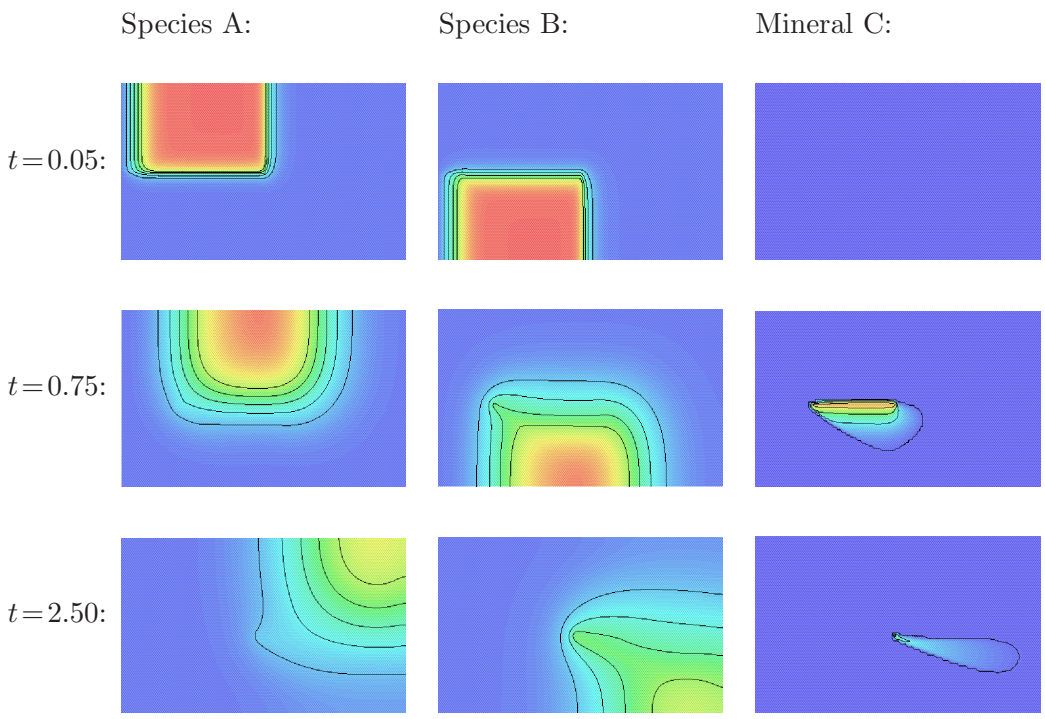


Figure 4.2: A time series of the concentrations of species A, B, C. Advective flow direction from left to right. The colour blue corresponds to the background concentration, which is 0.25 for A and B and 0.0 for C.

During the last years the behaviour of CO_2 in the groundwater has become of great interest, since techniques are developed and investigated to inject CO_2 in the subsurface in order to counteract the greenhouse effect. Suitable storage sites are searched and investigated. A very desirable effect would be if the carbon precipitates forming minerals, since in a solid form a long-term storage of the carbon might be more likely than in its dissolved (mobile) form. A long-term injection takes place in the Utsira formation in the subsurface of the Norwegian North Sea, accompanied by numerical simulations [38, 37, 39]. We use the following generic simplified set of chemical reactions to model the desired mechanism:

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

It consists of 3 minerals (calcite and mineral B are carbonates, mineral A is a silicate) and 6 aqueous reacting species plus H_2O and one (aqueous) tracer. We assume that the CO_2 is dissolved in the ground water and that there is no gas phase present or at least no CO_2 present in the gas phase. The first two reactions allow a transition of the carbon into HCO_3^- (mobile) and calcit (mineral). These reactions affect (and are affected by) the pH, i.e., the concentration of H^+. A silicate (mineral A), present in the initial state, dissolves at high H^+-concentrations. Thereby, it releases metal ions Me^{3+}. These ions can initiate the precipitation of the carbonate (mineral B), storing the carbon.

Having the Utsira storage site in mind, the metal could be aluminium (Al^{3+}), the carbonate could be Dawsonite [37]; however, the problem, including the reaction constants chosen for this simulation) are generic. The most interesting point in this model is that the way the Me^{3+} ions are incorporated, the dissolution of one mineral can cause an *immediate* precipitation of another mineral (i.e., the dissolution front of one mineral coincides with the precipitation front of another mineral).

The parameters are: domain $\Omega = (0,10)\times(0,6)$, number of nodes= 2145, Darcy velocity $q = (0.015, 0)^T$, water content $\theta = 0.3$, (i.e., pore velocity $|q|/\theta = 0.05$), longitudinal/transversal dispersion length $(\beta_l, \beta_t) = (0.3, 0.03)^T$, 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., $c_{H^+}c_{HCO_3^-}/c_{CO_2} = 0.1$. 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 $c_{CO_2} = c_{HCO_3^-} = c_{SiO_2} = 1$, $c_{H^+} = 0.1$, $c_{Me^{3+}} = 0.01$, $c_{Ca^{2+}} = 10$ (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 $\{0\}\times[1.5, 4.5]$ and zero on $(0, y)$ with $y < 1.5$, $y > 4.5$.

The numerical test runs were executed by J. Hoffmann [44].

The reduction method of Sec. 4.4.6 leads to $I - J_1 = 7 - 4 = 3$ decoupled linear scalar PDEs for the η_i, and a nonlinear system of $J_{min} = 3$ coupled PDEs for $\xi^{\mathcal{A}}_{min,i}$, each living on the active domain of the corresponding mineral $\bar{\xi}_{min,i}$; $J_{eq} = J_{mob} + J_{sorp} + J_{min} = 1 + 0 + 3$.

Fig. 4.3-4.5 show the computational results. The CO_2 entering the domain from the left affects the pH, which leads to a dissolution of calcite (Fig. 4.3). In the right part of the computational domain, the dissolution of mineral A

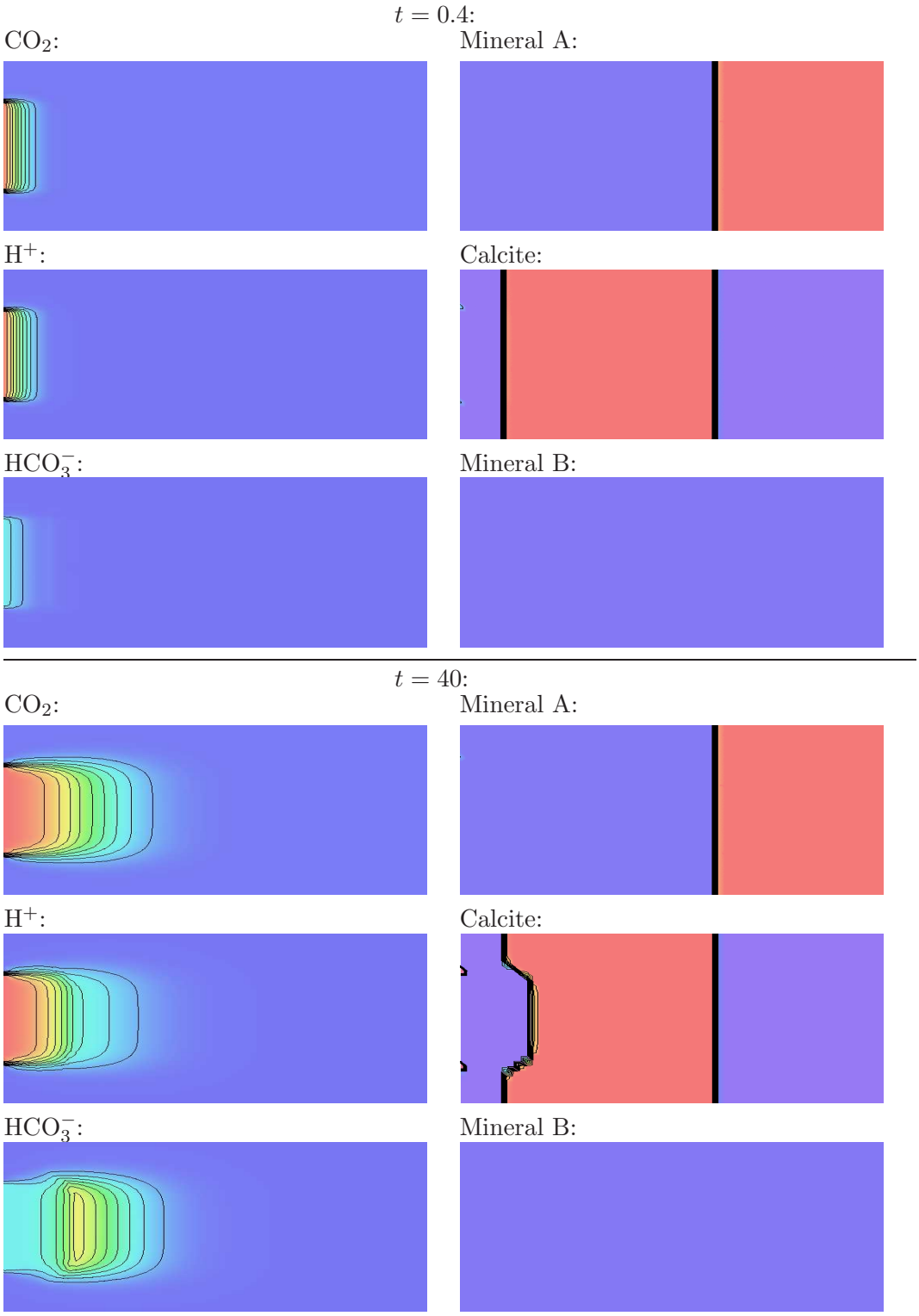


Figure 4.3: The water flow with dissolved CO_2 enters from the left. The dropping pH causes a dissolution of calcite layer. (The graphics are compressed by a factor of 1.5 in vertical direction.)

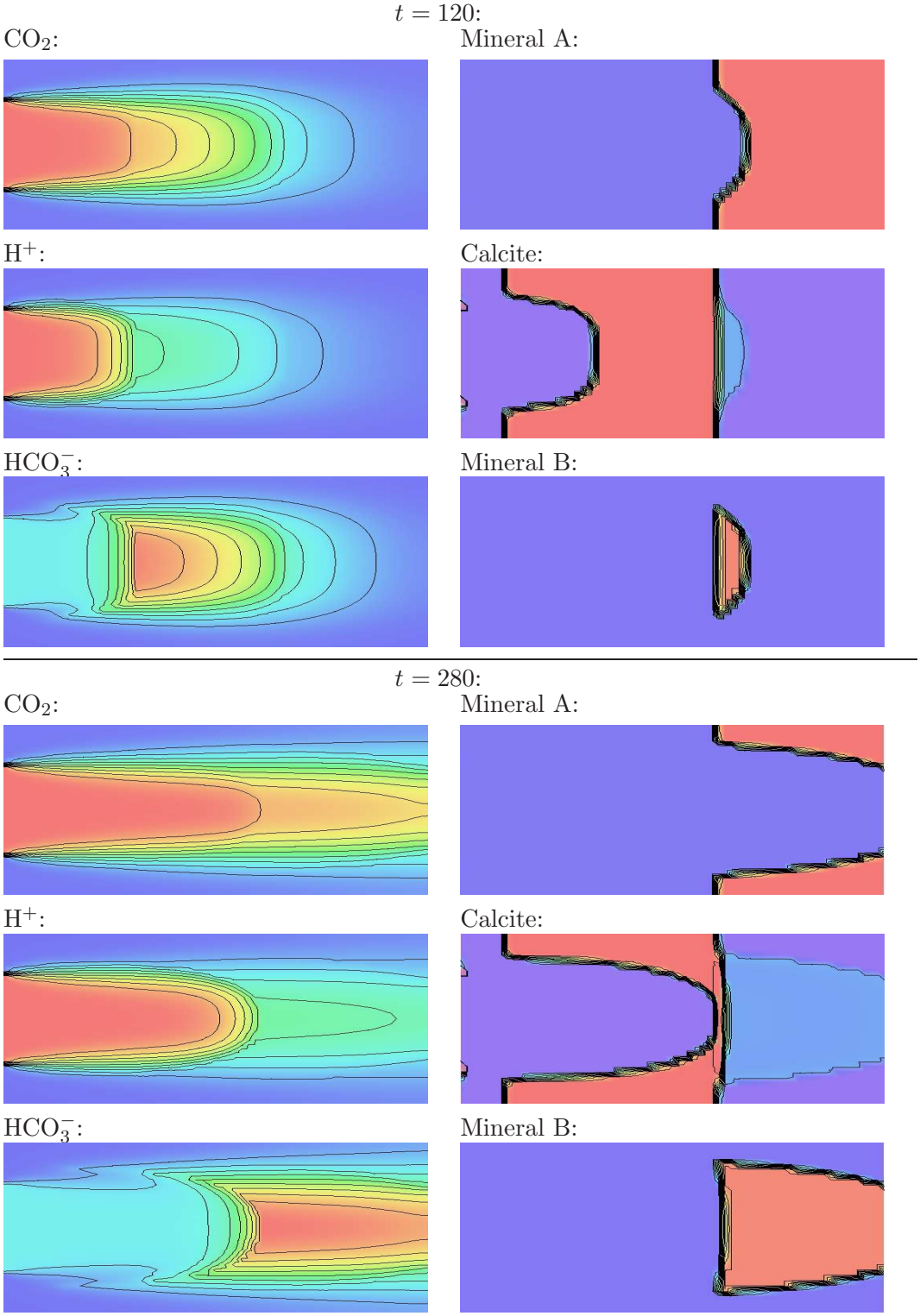


Figure 4.4: Dissolution of mineral A and precipitation of mineral B. (The graphics are compressed by a factor of 1.5 in vertical direction.)

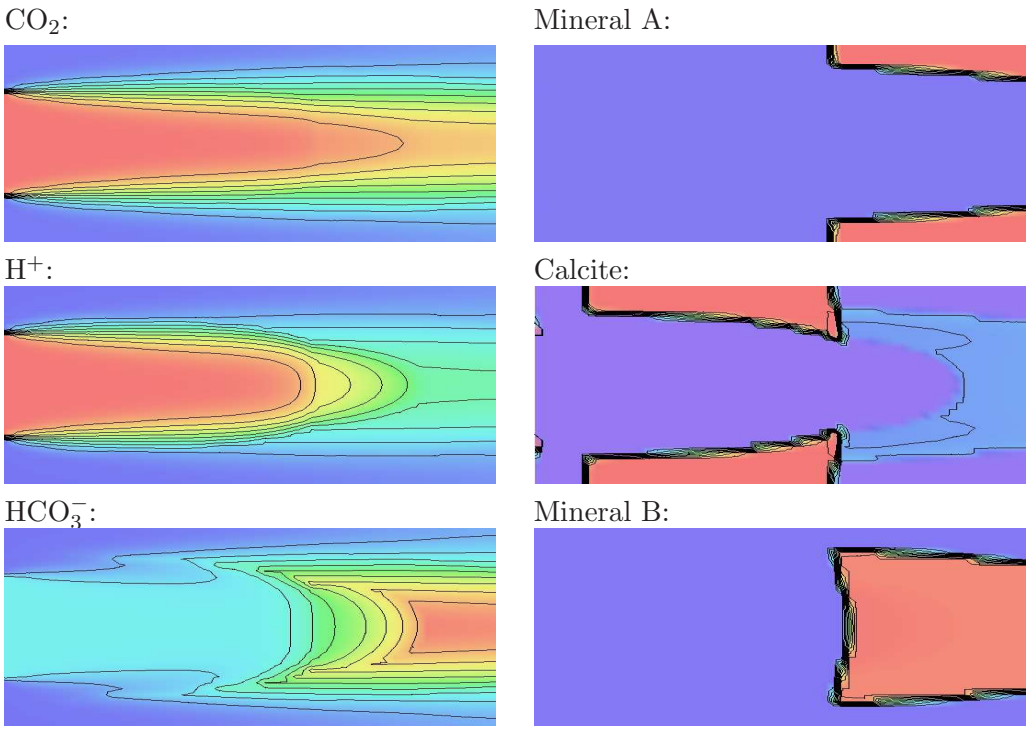


Figure 4.5: (The graphics are compressed by a factor of 1.5 in vertical direction.)

leads to an immediate precipitation of mineral B (Fig. 4.4). The propagation of the fronts, in particular of the precipitation-dissolution front between mineral A and mineral B, is reprodiced well by the complementarity based Semismooth Newton code. Also the discontinuities of the mineral concentrations are handled in a stable and robust way. (The fronts are caused by the equilibrium reactions, the flow is not convection-dominated.) The *discontinuities* in the mineral concentrations are not surprising; even simple example problems for which the exact solution can be given show this behaviour [44].

We observed similar convergence properties of Semismooth Newton similar to classical Newton for this application, such as mesh indepencence of the convergence speed.

Besides the two examples, we also simulated the problem [69] p. 110-114, for which an analytical solution (a good aproximation of the exact solution by an analytically given function, resp.) is available. This served as a validation of the code and showed that our CP/Semismooth Newton/active set method provides a numerical solution of satisfying accuracy; in particular, a correct

propagation speed of the precipitation/dissolution fronts.

4.6 Extensions to kinetic mineral reactions

The usage of a CP formulation is also possible for *kinetic* mineral reactions. Let us revisit the example problem (4.3) with the rate (4.8). Let us set up a discretized model of this kinetic precipitation/dissolution process:
We postulate

$$\begin{aligned} \frac{u_A - u_{A,old}}{\tau} + Lu_A &= -\frac{u_C - u_{C,old}}{\tau} \\ \frac{u_B - u_{B,old}}{\tau} + Lu_B &= -2\,\frac{u_C - u_{C,old}}{\tau}. \end{aligned} \tag{4.32}$$

Defining

$$f(u_A, u_B) := -k^f + k^b u_A u_B^2, \quad f : \mathbb{R}^2 \longrightarrow \mathbb{R}, \tag{4.33}$$

the condition (4.8) can be expressed by

$$\left(\tfrac{u_C - u_{C,old}}{\tau} - f(u_A, u_B) = 0 \ \wedge \ u_C \geq 0\right) \ \vee \ \left(\tfrac{u_C - u_{C,old}}{\tau} - f(u_A, u_B) \geq 0 \ \wedge \ u_C = 0\right). \tag{4.34}$$

The interpretation is: As long as the computation of u_C by $(u_C - u_{C,old})/\tau = -k^f + k^b u_A u_B^2$ leads to a value $u_C \geq 0$ ('left' case in (4.34)), use this formula. Otherwise, set $u_C := 0$, which then automatically means that $(u_C - u_{C,old})/\tau \geq -k^f + k^b u_A u_B^2$ ('right' case in (4.34)).
(4.34) is again a complementarity problem; it can be written[6]

$$\begin{aligned} &\left(\frac{u_C - u_{C,old}}{\tau} - f(u_A, u_B)\right) u_C = 0 \\ &\qquad \wedge \quad \frac{u_C - u_{C,old}}{\tau} - f(u_A, u_B)) \geq 0 \quad \wedge \quad u_C \geq 0 \end{aligned} \tag{4.35}$$

which has to be solved together with (4.32).

[6]It is convenient to apply a scaling by the factor $1/\tau$ to u_C, i.e., to write (4.35) as $\varphi(\frac{u_C - u_{C,old}}{\tau} - f(u_A, u_B), \frac{c_C}{\tau}) = 0$. As long as the CC is written in terms of equations and inequalities, this scaling is meaningless. However, as soon as a CP-function is applied to (4.35), and Semismooth Newton is applied to it, the criterium of activity/inactivity and thus the specific Newton sequence depends on the scaling.

The procedure applied to the example problem could be generalized to problems containing *several* kinetic mineral reactions. This could be integrated into the algorithm to handle minerals/immobile non-minerals/mobile species with kinetic/equilibrium reactions.
Note that if we handle this CP with the minimum CP-function (4.14-b), then the condition (4.35), reading

$$\min\left\{\frac{u_C - u_{C,old}}{\tau} - f(u_A, u_B), \frac{u_C}{\tau}\right\} = 0,$$

reduces (see the interpretation given after (4.34)) to the following rather simple strategy:

> "If the current Newton iterate u_A, u_B matches $u_{C,old} + \tau f(u_A, u_B) \geq 0$, then assemble the equation $\frac{u_C - u_{C,old}}{\tau} = f(u_A, u_B)$. Otherwise assemble the equation $u_C = 0$." (4.36)

However, due to the background of CPs and Semismooth Newton, this strategy now has a more profound justification (local quadratic convergence, if the problem is BD-regular).

4.7 Another Complementarity Problem in the geosciences

Another problem that can be handled through a complementarity formulation is the problem of the vanishing/reappearing of a gas phase in a multiphase problem such as the one discussed in Sec. 2.6.2.
It is usually assumed that the relation between the dissolved fraction of a gas and the its fraction in the gas phase is gouverned by Henry's law. In the setting of Sec. 2.6.2 it reads

$$c_{CO_2(aq)} = \frac{p\, c_{CO_2(gas)}}{H}$$

where H is Henry's law coefficient. If we assume $CO_{2(gas)}$ to be the only constituent of the gas phase, i.e., $c_{CO_2(gas)} \equiv 1$, then it simplifies to

$$c_{CO_2(aq)} = \frac{p}{H}. \tag{4.37}$$

However, the equilibrium condition (4.37) only holds if the gas phase is present: $s^1 > 0$. In this case the water phase is *saturated* with CO_2. Besides the saturated case, also the undersaturated case (in absence of the gas phase)

$$c_{CO_2(aq)} \leq \frac{p}{H} \quad \wedge \quad s^1 = 0$$

is possible. Again, the unified equilibrium condition

$$\left(\frac{p}{H} - c_{CO_2(aq)}\right) s^1 = 0 \quad \wedge \quad \frac{p}{H} - c_{CO_2(aq)} \geq 0 \quad \wedge \quad s^1 \geq 0$$

can be expressed as a complementarity problem; see Sec. 4.2.

4.8 Summary of Chapter 4

In this chapter, the model of reactive transport in porous media from Chapter 2 is extended by equilibrium precipitation-dissolution reactions of minerals. The difficulty is that the corresponding equilibrium conditions in general cannot be described by single smooth algebraic equations. Either a distinction of cases between local saturation of the fluid and local absence of the mineral have to be distinguished, or an equilibrium condition consisting of a combination of equations and inequalities can be used to describe the equilibrium correctly. In this chapter it was shown that the equilibrium condition can be regarded as a complementarity condition. This eliminates all inequalities, and it does not require the introduction of Lagrangian multipliers as additional unknowns. The system of complementarity conditions coupled to PDEs, ODEs, and AEs can be solved by the Semismooth Newton method. If a specific complementarity function, the minimum function, is used, the method can be considered as an active set strategy. The Semismooth Newton strategy provides a theoretical justification (proof of local quadratic convergence, since the BD-regularity of the problem, at least for small time steps, can be shown). The algorithms were combined with the reformulation/size reduction method of Chapter 2 und implemented in a software. The successful acting of the method was demonstrated in two applications.

Appendix

Theorem 5.1 Schaefer's Fixed Point Theorem. *Let X be a real Banach space. Suppose*

$$\mathcal{Z} : X \longrightarrow X$$

is a continuous and compact mapping. Assume further that the set

$$\{u \in X \mid \exists \lambda \in [0,1] : \ u = \lambda \mathcal{Z}(u)\}$$

is bounded. Then $\mathcal{Z}$ has a fixed point.

Proof. Can be derived from Schauder's fixed point theorem. See [72], [27].

Lemma 5.2 *Let $\Omega \subset \mathbb{R}^n$ a bounded domain with locally Lipschitz continuous boundary, $1 \leq p < \infty$, $m \in \mathbb{N}$. For $p > \frac{n}{m}$, the embedding*

$$W_p^m(\Omega) \subset C(\bar{\Omega})$$

is compact.
In particular, by applying this result to $Q_T = (0,T) \times \Omega \subset \mathbb{R}^{n+1}$ instead of Ω, for $p > n+1$ the space $W_p^{2,1}(Q_T)$ (defined in Sec. 3.3.1) is compactly embedded in the space $C(\overline{Q}_T)$.

Proof. See [1], Sec. 6.3, 6.4.1, 4.9. ∎

Definition 5.3 ($\mathcal{O}^l$-boundary). *Let us consider $n, l \in \mathbb{N}$, $n \geq 2$ and a bounded domain $\Omega \subset \mathbb{R}^n$ (i.e., a bounded connected open subsets of $\mathbb{R}^n$). We say that Ω has a $\mathcal{O}^l$-boundary, if there exists a number $r > 0$ such that the intersection of $\partial\Omega$ with the ball $B_r(x_0)$ with radius r and arbitrary centre $x_0 \in \partial\Omega$ is a connected surface, the equation of which in the local coordinate system with origin at x_0 has the form $y_n = \omega(y_1, ..., y_{n-1})$ and is a function with C^{l-1}-regularity, where the derivatives of order l–1 have a first differential at each point, which is bounded on the domain of definition. [56]*

Lemma 5.4 (linear parabolic problem). *Let $\Omega \subset \mathbb{R}^n$, $n \geq 2$, be a domain with $\mathcal{O}^2$-boundary in the sense of Def. 5.3 and $T > 0$ be given. Consider the linear parabolic equation*

$$\begin{aligned} \partial_t u - \sum_{i,j=1}^{n} a_{ij}\, \partial_i \partial_j u + \sum_{i=1}^{n} a_i\, \partial_i u + bu &= f \quad \text{on } Q_T \\ \partial u/\partial \nu &= 0 \quad \text{for } x \in \partial\Omega \\ u &= u_0 \quad \text{for } t = 0. \end{aligned} \tag{5.1}$$

where the a_{ij} fulfil the uniform ellipticity condition

$$c_1 \left|\xi\right|^2 \leq \sum_{i,j=1}^{n} a_{ij}(t,x)\, \xi_i \xi_j \leq c_2 \left|\xi\right|^2$$

for all $\xi \in \mathbb{R}^n$, $(t,x) \in Q_T$ for given constants $c_1, c_2 > 0$. We assume that $a_{ij} \in C(\overline{Q}_T)$, $a_i \in L^r(Q_T)$, $, b \in L^s(Q_T)$, $f \in L^q(Q_T)$, $u_0 \in W_q^{2-2/q}(\Omega)$, $1 < q < \infty$, $q \neq 3$,

$$r = \begin{cases} \max\{q, n+2\}, & q \neq n+2 \\ n+2+\epsilon, & q = n+2 \end{cases}, \quad s = \begin{cases} \max\{q, (n+2)/2\}, & q \neq (n+2)/2 \\ (n+2)/2+\epsilon, & q = (n+2)/2 \end{cases}.$$

Then problem (5.1) has a unique solution $u \in W_q^{2,1}(Q_T)$ and there is a constant $c > 0$ independent of f, u_0, u such that the estimate

$$\|u\|_{W_q^{2,1}(Q_T)} \leq c \left(\|f\|_{L^q(Q_T)} + \|u_0\|_{W_q^{2-2/q}(\Omega)} \right)$$

holds.

See [56] p. 342 ff. (note the remark at the end of § 9, p. 351); and p. 621 ff;[7] see [59] Sec. 8.7.

Lemma 5.5 (boundary estimate). *Let $\Omega \subset \mathbb{R}^n$, $n \geq 2$, be a bounded domain with boundary $\partial\Omega$ piecewise smooth with nonzero interior angles (see below for a definition). Then the estimate*

$$\|u\|_{L^q(\partial\Omega)} \leq c\, \|\nabla u\|_{L^2(\Omega)}^{\alpha}\, \|u\|_{L^2(\Omega)}^{1-\alpha}, \quad \alpha = \frac{n}{2} - \frac{n-1}{q}, \tag{5.2}$$

[7] Note that due to the Neumann boundary conditions for the required regularity no compatibility condition is required.

holds for all $u \in H^1(\Omega)$ *with*

$$\int_\Omega u(x)\,dx = 0,$$

where $q \in [1,\infty)$ *for* $n = 2$, *and* $q \in [2 - 2/n, 2 + 2/(n-2)]$ *for* $n \geq 3$.
Here, the expression 'piecewise smooth with nonzero interior angles' means: $\bar\Omega$ *can be written as* $\bar\Omega = \bar\Omega_1 \cup ... \cup \bar\Omega_N$, $\Omega_i \cap \Omega_j = \emptyset$, *where each* $\bar\Omega_k$ *can be homeomorphically mapped onto the unit ball or onto the unit cube by means of a Lipschitz continuous mapping with a Jacobian which is bounded from below by a positive constant.*

Proof. See [56] Ch. II, § 2, pp. 62, especially (2.21); also p. 9. ∎

We need a result of the type of Lemma 5.5 also for functions which are not meanvalue-free:

Lemma 5.6 (boundary estimate). *Let* $\Omega \subset \mathbb{R}^n$ *be a bounded domain, piecewise smooth with nonzero interior angles. For all* $u \in H^1(\Omega)$ *the estimate*

$$\|u\|^2_{L^2(\partial\Omega)} \leq c\,\|u\|_{H^1(\Omega)}\,\|u\|_{L^2(\Omega)}$$

holds with c *independent of* u.

Proof. We want to trace back the statement to Lemma 5.5. Therefore we introduce the mean value

$$m = \int_\Omega u\,dx.$$

We note that $m^2 \leq c\,\|u\|^2_{L^2(\Omega)}$. We expand

$$\begin{aligned} u^2 &= ((u-m)+m)^2 \\ &\leq 2\,(u-m)^2 + 2m^2 \end{aligned}$$

and get by integration over the trace of u on $\partial\Omega$

$$\|u\|^2_{L^2(\partial\Omega)} \leq 2\,\|u-m\|^2_{L^2(\partial\Omega)} + c\,\|u\|^2_{L^2(\Omega)}$$

where c only depends on Ω. We apply Lemma 5.5 with $q=2$ to the function $u-m$ and get

$$\|u\|^2_{L^2(\partial\Omega)} \le c\,\|\nabla u\|_{L^2(\Omega)}\,\|u-m\|_{L^2(\Omega)} + c\,\|u\|^2_{L^2(\Omega)}.$$

We estimate $\|u-m\|_{L^2(\Omega)} \le \|u\|_{L^2(\Omega)} + \|m\|_{L^2(\Omega)} \le c\,\|u\|_{L^2(\Omega)}$ and obtain

$$\|u\|^2_{L^2(\partial\Omega)} \le c\,(\,\|\nabla u\|_{L^2(\Omega)}\,\|u\|_{L^2(\Omega)} + \|u\|^2_{L^2(\Omega)})$$

and apply Poincaré's inequality. ∎

Lemma 5.7 *Let $m, u \in \mathbb{R}$, $u \ge 0$. Then the estimate*

$$(m+\ln u)\,u \le 2\,[(m-1+\ln u)\,u + e^{1-m}]$$

holds.

Proof. As already stated in the first estimate of (3.15),

$$u \le (m-1+\ln u)\,u + e^{1-m}$$

holds for all $m, u \in \mathbb{R}$, $u \ge 0$. By adding $(m-1+\ln u)\,u$ to both sides of the inequality, we obtain the desired result. ∎

Lemma 5.8 *For smooth functions $u : \Omega \longrightarrow \mathbb{R}^I$ with $u > 0$ on Ω, the derivative $D_x(\partial f_r(u))$ of $\partial f_r(u) : \Omega \longrightarrow \mathbb{R}^I$ (being defined as in (3.37), for fixed t) with respect to $x \in \Omega \subset \mathbb{R}^n$ reads*

$$W(x) = r(r-1) f_{r-2}(u(x))\,M D_x u + r f_{r-1}(u(x)) \Lambda_{1/u} D_x u \;\in\; \mathbb{R}^{I\times n}, \tag{5.3}$$

where $M = M(x)$ is the symmetric $I\times I$-matrix with entries

$$M_{ij} = (\mu_i^0 + \ln u_i)\,(\mu_j^0 + \ln u_j),$$

and where $\Lambda_{1/u} = \Lambda_{1/u}(x)$ is the $I\times I$ diagonal matrix with entries $1/u_i(x)$, and[8] *$r \in \mathbb{N}$.*
Furthermore, the real number $W : D_x u$ defined by

$$W : D_x u := \sum_{i=1}^{I} \sum_{k=1}^{n} \frac{(\partial f_r(u))_i}{\partial x_k}\,\frac{\partial u_i}{\partial x_k} \tag{5.4}$$

is nonnegative for all $x \in \Omega$.

[8] Note that for $r=1$, the term containing f_{r-2} is immaterial.

Proof. The partial derivatives of the i-th component of ∂f_r obviously[9] read

$$\frac{\partial}{\partial x_k}(\partial f_r(u))_i = r(r-1)f_{r-2}\sum_{j=1}^{I}(\mu_j^0+\ln u_j)\,\frac{\partial u_j}{\partial x_k}\,(\mu_i^0+\ln u_i) + r\,f_{r-1}\,\frac{1}{u_i}\,\frac{\partial u_i}{\partial x_k} \tag{5.5}$$

for $i = 1, ..., I$, $k = 1, ..., n$. From this, (5.3) follows directly.
Now let us consider $W : D_x u$ defined in (5.4). Using (5.5) we get

$$\begin{aligned} W : D_x u \;\; &= \;\; r(r-1)\,f_{r-2}(u)\sum_{i=1}^{I}\sum_{k=1}^{n}\sum_{j=1}^{I}(\mu_i^0+\ln u_i)(\mu_j^0+\ln u_j)\,\frac{\partial u_j}{\partial x_k}\,\frac{\partial u_i}{\partial x_k} \\ &\quad + r\,f_{r-1}(u)\sum_{i=1}^{I}\sum_{k=1}^{n}\frac{1}{u_i}\,\left(\frac{\partial u_i}{\partial x_k}\right)^2. \end{aligned} \tag{5.6}$$

The latter term is obviously nonnegative. Considering the first term, we rearrange the summations

$$\begin{aligned} &\sum_{i=1}^{I}\sum_{k=1}^{n}\sum_{j=1}^{I}(\mu_i^0+\ln u_i)(\mu_j^0+\ln u_j)\,\frac{\partial u_j}{\partial x_k}\,\frac{\partial u_i}{\partial x_k} \\ &= \sum_{k=1}^{n}\left[\left(\sum_{j=1}^{I}(\mu_j^0+\ln u_j)\,\frac{\partial u_j}{\partial x_k}\right)\left(\sum_{i=1}^{I}(\mu_i^0+\ln u_i)\,\frac{\partial u_i}{\partial x_k}\right)\right] \\ &= \sum_{k=1}^{n}\left(\sum_{j=1}^{I}(\mu_j^0+\ln u_j)\,\frac{\partial u_j}{\partial x_k}\right)^2 \\ &= \sum_{k=1}^{n}\left\langle \mu^0+\ln u, \frac{\partial u}{\partial x_k}\right\rangle_I^2 \geq 0 \end{aligned} \tag{5.7}$$

Since f_{r-2}, f_{r-1} are nonnegative, we get the desired result. ∎

Lemma 5.9 *Let $S_1 \in \mathbb{R}^{I_1\times J}$ and $S_2 \in \mathbb{R}^{I_2\times J}$ be two matrices and $S = \begin{pmatrix} S_1 \\ \hline S_2 \end{pmatrix}$ an $(I_1+I_2)\times J$ matrix such that the columns of S are linearly independent. Then by elementary column operations the matrix S can be transformed into an $(I_1+I_2)\times J$ matrix*

$$\left(\begin{array}{c|c|c} S_{11} & S_{12} & 0 \\ \hline 0 & S_{22} & S_{23} \end{array}\right)$$

[9] see (3.37)

where both $(S_{11}|S_{12})$ *and* $(S_{22}|S_{23})$ *have maximal column rank, and where the* S_{1i} *and* S_{2i} *have the same number of rows as* S_1, S_2, *respectively.*

Proof. Step 1: By elementary column operations, one can obviously transform S into the shape

$$\left(\begin{array}{c|c} A & 0 \\ \hline B & C \end{array}\right), \tag{5.8}$$

where both the columns of A and also the columns of C are linearly independent. However, the columns of $(B|C)$ may still be linear dependent (see [51] Sec. 4 for more details).
Step 2: The result of step 1 is a matrix of shape

$$S = \left(\begin{array}{c|c|c} A_1 & A_2 & 0 \\ \hline 0 & B & C \end{array}\right) \tag{5.9}$$

with both $(A_1|A_2)$ and C having linearly independent columns (with possibly A_1 having zero columns). While the columns of $(B|C)$ are still linearly dependent, do the following iterative procedure:
Since the columns of C are linearly independent, one column of B must be a linear combination of the (other) columns of $(B|C)$. That means that by a column operation, applied to the whole matrix S (i.e., lower and upper part), this particular column of B can be brought to zero. This manipulation does not affect the linear independence of the columns of C and of $(A_1|A_2)$. The resulting matrix is again of shape (5.9), but with B and A_2 reduced by one column and A_1 enlarged by one column.
The process is repeated until $(B|C)$ consists of linearly independent columns.
■

Lemma 5.10 *Let* $A, B \in \mathbb{R}^{n\times n}$ *be two symmetric matrices.*

(i) If A *(or* B*) is positive definite, and the other matrix* B *(*A*, resp.) is positive semidefinite, then the matrix* AB *has* n *linearly independent eigenvectors with nonnegative eigenvalues, i.e.,* AB *is similar to a diagonal matrix with nonnegative diagonal entries.*

(ii) If A *and* B *are both positive definite, the (i) holds with strictly positive eigenvalues.*[10]

[10]Note that we of course do *not* claim that AB is symmetric positive definite!

Proof. **(i).** Let us assume that B is symmetric positive definite (spd). Then the spd square root $B^{1/2}$ exists. Then matrix $B^{1/2}AB^{1/2}$ is symmetric positive semidefinite, i.e., it has n linearly independent eigenvectors with nonnegative eigenvalues. Matrix $B^{1/2}AB^{1/2}$ is obviously similar to matrix AB. Hence, AB has the required property.
If A is spd, then consider $A^{1/2}BA^{1/2}$ instead of $B^{1/2}AB^{1/2}$.
(ii). Proceed as in (i), but note that $B^{1/2}AB^{1/2}$ is spd, i.e., it has *strictly* positive eigenvalues. ∎

Lemma 5.11 *Let the current Newton iterate of the global problem (4.28) be positive ('physical') in the sense that the corresponding value in terms of c, $\bar{c}_{nmin}$, $\bar{c}_{min}$ lies in $\mathbb{R}_+^{I+\bar{I}_{nmin}} \times \mathbb{R}^{\bar{I}_{min}}$ at each of the $m \in \mathbb{N}$ mesh points. The Jacobian of the global problem, after the local equations have been eliminated, (4.28), has the structure*

$$\begin{aligned} \mathcal{J}_{glob} &= Id + \mathcal{J}_{equil} + \tau\,\mathcal{J}_{transp} + \tau\mathcal{J}_{bdry} + \tau\,\mathcal{J}_{kinet} \\ &\in \mathbb{R}^{(mJ_{kin}^{1'}+mJ_{sorp}+|\mathcal{A}|)\times(mJ_{kin}^{1'}+mJ_{sorp}+|\mathcal{A}|)} \end{aligned}$$

where[11] *the τ-independent part $\mathcal{J}_{equil}$ has the structure*

$$\mathcal{J}_{equil} = \left(\begin{array}{c|c|c} 0 & 0 & 0 \\ \hline U_1 & U_2 & U_3 \\ \hline 0 & 0 & 0 \end{array}\right),$$

and where

$$\mathcal{J}_{transp} = \left(\begin{array}{c|c|c} L_h & 0 & 0 \\ \hline 0 & L_h & 0 \\ \hline 0 & 0 & L_h^{\mathcal{A}\mathcal{A}} \end{array}\right), \quad \mathcal{J}_{bdry} = \left(\begin{array}{c|c|c} 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline L_h^{\mathcal{A}\mathcal{I}}V_1 & L_h^{\mathcal{A}\mathcal{I}}V_2 & L_h^{\mathcal{A}\mathcal{I}}V_3 \end{array}\right)$$

whith the blocks

$$\begin{aligned} U_1 &= H^{-1}{S_{sorp}^1}^T \Lambda^{1/2} P_{(\Lambda^{1/2}(S_{min\mathcal{I}}^1|S_{mob}^1))^\perp} \Lambda^{1/2} S_{kin}^1 \in \mathbb{R}^{(mJ_{sorp})\times(mJ_{kin}^{1'})}, \\ U_2 &= H^{-1}{S_{sorp}^1}^T \Lambda^{1/2} P_{(\Lambda^{1/2}(S_{min\mathcal{I}}^1|S_{mob}^1))^\perp} \Lambda^{1/2} S_{sorp}^1 \in \mathbb{R}^{(mJ_{sorp})\times(mJ_{sorp})}, \\ U_3 &= H^{-1}{S_{sorp}^1}^T \Lambda^{1/2} P_{(\Lambda^{1/2}(S_{min\mathcal{I}}^1|S_{mob}^1))^\perp} \Lambda^{1/2} S_{min\mathcal{A}}^1 \in \mathbb{R}^{(mJ_{sorp})\times|\mathcal{A}|}, \end{aligned}$$

[11] $\mathcal{J}_{kinet}$ may depend on τ due to the substitution of the local resolution function, but only as $O(1)$ for $\tau \to 0$.

where m is the number of mesh points, and $P_{(\Lambda^{1/2}(S^1_{min\mathcal{I}}|S^1_{mob}))^\perp}$ is the orthogonal projection onto the orthogonal complement of the space spanned by the columns of the matrix $M_P := \Lambda^{1/2}(S^1_{min\mathcal{I}}|S^1_{mob})$, i.e.,

$$P := P_{(\Lambda^{1/2}(S^1_{min\mathcal{I}}|S^1_{mob}))^\perp} = Id - P_{\Lambda^{1/2}(S^1_{min\mathcal{I}}|S^1_{mob})} = Id - M_P(M_P^T M_P)^{-1} M_P^T$$

and H is the positive definite matrix

$$H := {S^2_{sorp}}^T \tilde{\Lambda} S^2_{sorp},$$

and Λ, $\tilde{\Lambda}$ are defined as in the proof of Lemma 4.4.
In particular, U_2 has a full system of eigenvectors with eigenvalues all being strictly positive.
V_1, V_2, V_3 can be expressed as

$$(V_1|V_2|V_3) = H_1^{-1} H_2 \, (S^1_{kin}|S^1_{sorp}|S^1_{min\mathcal{A}})$$

with

$$\begin{aligned}
H_1 &:= {S^1_{min\mathcal{I}}}^T \Lambda^{1/2} P_{(\Lambda^{1/2} S^1_{mob})^\top} \Lambda^{1/2} S^1_{min\mathcal{I}} \\
H_2 &:= [{S^1_{min\mathcal{I}}}^T - {S^1_{min\mathcal{I}}}^T \Lambda S^1_{mob} (S^1_{mob}{}^T \Lambda S^1_{mob})^{-1} {S^1_{mob}}^T] \, \Lambda \\
&= {S^1_{min\mathcal{I}}}^T \Lambda^{1/2} \, [Id - \Lambda^{1/2} S^1_{mob} (S^1_{mob}{}^T \Lambda S^1_{mob})^{-1} {S^1_{mob}}^T \Lambda^{1/2}] \, \Lambda^{1/2} \\
&= {S^1_{min\mathcal{I}}}^T \Lambda^{1/2} P_{(\Lambda^{1/2} S^1_{mob})^\top} \Lambda^{1/2}
\end{aligned}$$

Proof. The system (4.26)/(4.27) (see Table 4.2, right part) can be written

$$\left[\left(\begin{array}{ccc||cccc} \mathrm{Id}+\tau L_h & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathrm{Id}+\tau L_h & 0 & 0 & -\mathrm{Id} & 0 & 0 \\ 0 & 0 & \mathrm{Id}_{\mathcal{A}}+\tau L_h^{\mathcal{A}\mathcal{A}} & 0 & 0 & 0 & 0 \end{array} \right) + \tau \left(\begin{array}{ccc||cccc} & & & 0 & 0 & 0 & 0 \\ & 0 & & 0 & 0 & 0 & 0 \\ & & & 0 & 0 & L_h^{\mathcal{A}\mathcal{I}} & 0 \\ \hline 0 & & & & 0 & & \end{array} \right) + \text{kinetic terms} \right] \left(\begin{array}{c} \Delta\xi_{kin} \\ \Delta\xi_{sorp} \\ \Delta\xi^{\mathcal{A}}_{min} \\ \hline \Delta\bar{\xi}_{kin} \\ \Delta\bar{\xi}_{sorp} \\ \Delta\xi^{\mathcal{I}}_{min} \\ \Delta\xi_{mob} \end{array} \right) = \text{r.h.s.}$$

where the kinetic terms (containing derivatives of $f_1, ..., f_4$) are $O(\tau)$. We have to solve the lower part of this system,

$$C\,\Delta\xi_{glob} + D\,\Delta\xi_{loc} + \text{kin.terms} = \text{r.h.s.}, \tag{5.10}$$

for $\Delta\xi_{loc}$, where $\Delta\xi_{glob} = (\Delta\xi_{kin}, \Delta\xi_{sorp}, \Delta\xi^{\mathcal{A}}_{min})$, $\Delta\xi_{loc} = (\Delta\bar{\xi}_{kin}, \Delta\bar{\xi}_{sorp}, \Delta\xi^{\mathcal{I}}_{min}, \Delta\xi_{mob})$, and substitute the local variables (only $\Delta\bar{\xi}_{sorp}$ and $\Delta\xi^{\mathcal{A}}_{min}$ are required) into the remaining system. Hence, the system of reduced size (4.27) reads

$$(\mathrm{Id} + \tau L_h)\,\Delta\xi_{glob} - \begin{pmatrix} 0 \\ \Delta\bar{\xi}_{sorp} \\ 0 \end{pmatrix} + \tau \begin{pmatrix} 0 \\ 0 \\ L_h^{\mathcal{A}\mathcal{I}}\Delta\xi^{\mathcal{I}}_{min} \end{pmatrix} + \text{kin.terms} = \text{r.h.s.}$$

where $\Delta\bar{\xi}_{sorp}$ and $\Delta\xi^{\mathcal{I}}_{min}$ have to be expressed in terms of the global variables by using (5.10).
The general formula for the inversion of a block matrix with quadratic blocks M_{11}, M_{22} (if all the inverse in (5.12) exist) is:

$$M^{-1} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}^{-1} = \begin{pmatrix} M'_{11} & M'_{12} \\ M'_{21} & M'_{22} \end{pmatrix}, \tag{5.11}$$

$$\begin{array}{rcl} M'_{11} &=& (M_{11} - M_{12}M_{22}^{-1}M_{21})^{-1}, \\ M'_{12} &=& -(M_{11} - M_{12}M_{22}^{-1}M_{21})^{-1}M_{12}M_{22}^{-1} \\ M'_{21} &=& -(M_{22} - M_{21}M_{11}^{-1}M_{12})^{-1}M_{21}M_{11}^{-1} \\ M'_{22} &=& (M_{22} - M_{21}M_{11}^{-1}M_{12})^{-1}, \end{array} \tag{5.12}$$

so in particular

$$\left(\begin{array}{c|c} M_{11} & M_{12} \\ \hline 0 & M_{22} \end{array}\right)^{-1} = \left(\begin{array}{c|c} M_{11}^{-1} & -M_{11}^{-1}M_{12}M_{22}^{-1} \\ \hline 0 & M_{22}^{-1} \end{array}\right), \tag{5.13}$$

and

$$\left(\begin{array}{c|c} M_{11} & 0 \\ \hline M_{21} & M_{22} \end{array}\right)^{-1} = \left(\begin{array}{c|c} M_{11}^{-1} & 0 \\ \hline -M_{22}^{-1}M_{21}M_{11}^{-1} & M_{22}^{-1} \end{array}\right). \tag{5.14}$$

The local problem (5.10) reads, computing C from Tab. 4.2, right part:

$$D \begin{pmatrix} \Delta\bar{\xi}_{kin} \\ \Delta\bar{\xi}_{sorp} \\ \Delta\xi^{\mathcal{I}}_{min} \\ \Delta\xi_{mob} \end{pmatrix} = -\left(\frac{0}{(S^1_{sorp}|S^1_{min\mathcal{I}}|S^1_{mob})^T\Lambda(S^1_{kin}|S^1_{sorp}|S^1_{min\mathcal{A}})}\right) \begin{pmatrix} \Delta\xi_{kin} \\ \Delta\xi_{sorp} \\ \Delta\xi^{\mathcal{A}}_{min} \end{pmatrix}$$
$$+\text{kinetic terms} = \text{r.h.s.},$$

where D is the matrix from (4.29), but, as already C, without the kinetic $O(\tau)$-terms. Hence, by inverting D by formula (5.13) we obtain

$$\begin{pmatrix} \Delta\bar{\xi}_{kin} \\ \hline \Delta\bar{\xi}_{sorp} \\ \Delta\xi^{\mathcal{I}}_{min} \\ \Delta\xi_{mob} \end{pmatrix} = -\left(\begin{array}{c|c} D_{11}^{-1} & -D_{11}^{-1}D_{12}D_{22}^{-1} \\ \hline 0 & D_{22}^{-1} \end{array}\right) \cdot$$

$$\cdot \left(\begin{array}{c} 0 \\ \hline (S^1_{sorp}|S^1_{min\mathcal{I}}|S^1_{mob})^T\Lambda(S^1_{kin}|S^1_{sorp}|S^1_{min\mathcal{A}}) \end{array}\right) \begin{pmatrix} \Delta\xi_{kin} \\ \Delta\xi_{sorp} \\ \Delta\xi^{\mathcal{A}}_{min} \end{pmatrix}$$

$$+\text{kin.terms} + \text{r.h.s.} \tag{5.15}$$

We compute the inverse of D_{11} by formula (5.14):

$$D_{11}^{-1} = \left(\begin{array}{c|c} \mathrm{Id} & 0 \\ \hline (D_{11})^{21} & (D_{11})^{22} \end{array}\right)^{-1} = \left(\begin{array}{c|c} \mathrm{Id} & 0 \\ \hline -(D_{11})^{22^{-1}}(D_{11})^{21} & (D_{11})^{22^{-1}} \end{array}\right) \tag{5.16}$$

where the $(D_{11})^{ij}$ are the subblocks of D_{11} (cf. (4.29)), i.e.,

$$(D_{11})^{11} = \mathrm{Id},\ \ (D_{11})^{12} = 0,\ \ (D_{11})^{21} = S^2_{sorp}\tilde{\Lambda}S^{2'}_{kin},\ \ (D_{11})^{22} = S^2_{sorp}\tilde{\Lambda}S^2_{sorp}.$$

By denoting the subblocks of D_{12} by $(D_{12})^{ij}$, i.e.,

$$(D_{12})^{11} = 0,\ (D_{12})^{12} = 0,\ (D_{12})^{21} = {S^1_{sorp}}^T\Lambda S^1_{min\mathcal{I}},\ (D_{12})^{22} = {S^1_{sorp}}^T\Lambda\, S^1_{mob},$$

we obtain

$$\begin{aligned} D_{11}^{-1}D_{12} &= \left(\begin{array}{c|c} \mathrm{Id} & 0 \\ \hline -(D_{11})^{22^{-1}}(D_{11})^{21} & (D_{11})^{22^{-1}} \end{array}\right)\left(\begin{array}{c|c} 0 & 0 \\ \hline (D_{12})^{21} & (D_{12})^{22} \end{array}\right) \\ &= \left(\begin{array}{c|c} 0 & 0 \\ \hline (D_{11})^{22^{-1}}(D_{12})^{21} & (D_{11})^{22^{-1}}(D_{12})^{22} \end{array}\right). \end{aligned}$$

Hence

$$D_{11}^{-1}D_{12}D_{22}^{-1} = \left(\begin{array}{c} 0 \\ \hline \left((D_{11})^{22^{-1}}(D_{12})^{21}|(D_{11})^{22^{-1}}(D_{12})^{22}\right)\ D_{22}^{-1} \end{array}\right). \tag{5.17}$$

Now we omit the equations for $\Delta\xi^{\mathcal{I}}_{min}$, $\Delta\xi_{mob}$ from (5.15) to obtain an equation for $(\Delta\bar{\xi}_{kin}, \Delta\bar{\xi}_{sorp})$. This means to drop the blocks 0 and D_{22}^{-1} from the

first block matrix from the right-hand side of (5.15). Then we also drop the equation for $\Delta\bar{\xi}_{kin}$, which means to drop the *upper* blocks of the matrices D_{11}^{-1} and $-D_{11}^{-1}D_{12}D_{22}^{-1}$, leaving us with only the lower blocks from these two matrices, which are given in (5.16) and (5.17). We get:

$$\begin{aligned}
&\Delta\bar{\xi}_{sorp} = \\
&-\Big(-(D_{11})^{22^{-1}}(D_{11})^{21}\Big|(D_{11})^{22^{-1}}\Big|-\Big((D_{11})^{22^{-1}}(D_{12})^{21}|(D_{11})^{22^{-1}}(D_{12})^{22}\Big)D_{22}^{-1}\Big)\cdot \\
&\quad\cdot\left(\begin{array}{c}0\\ \hline (S^1_{sorp}|S^1_{min\mathcal{I}}|S^1_{mob})^T\Lambda(S^1_{kin}|S^1_{sorp}|S^1_{min\mathcal{A}})\end{array}\right)\begin{pmatrix}\Delta\xi_{kin}\\ \Delta\xi_{sorp}\\ \Delta\xi^{\mathcal{A}}_{min}\end{pmatrix}+O(\tau)+\text{r.h.s.}
\end{aligned}$$

The first block $-D_{11}^{22^{-1}}D_{11}^{21}$ is multiplied with the zero-block of the following matrix and vanishes. From the other three blocks we can extract the factor $(D_{11})^{22^{-1}}$:

$$\begin{aligned}
\Delta\bar{\xi}_{sorp} &= -(D_{11})^{22^{-1}}\cdot\left(\mathrm{Id}|-((D_{12})^{21}|(D_{12})^{22})\,D_{22}^{-1}\right)\cdot\left(S^1_{sorp}|S^1_{min\mathcal{I}}|S^1_{mob}\right)^T\cdot \\
&\quad\cdot\Lambda\,\left(S^1_{kin}|S^1_{sorp}|S^1_{min\mathcal{A}}\right)\begin{pmatrix}\Delta\xi_{kin}\\ \Delta\xi_{sorp}\\ \Delta\xi^{\mathcal{A}}_{min}\end{pmatrix}+O(\tau)+\text{r.h.s.} \\
&= -(D_{11})^{22^{-1}}\cdot\left[{S^1_{sorp}}^T-{S^1_{sorp}}^T\Lambda(S^1_{min\mathcal{I}}|S^1_{mob})D_{22}^{-1}(S^1_{min\mathcal{I}}|S^1_{mob})^T\right]\Lambda\cdot \\
&\quad\cdot\left(S^1_{kin}|S^1_{sorp}|S^1_{min\mathcal{A}}\right)\begin{pmatrix}\Delta\xi_{kin}\\ \Delta\xi_{sorp}\\ \Delta\xi^{\mathcal{A}}_{min}\end{pmatrix}+O(\tau)+\text{r.h.s.} \\
&= -(D_{11})^{22^{-1}}{S^1_{sorp}}^T\Lambda^{1/2}\left[\mathrm{Id}-\Lambda^{1/2}(S^1_{min\mathcal{I}}|S^1_{mob})D_{22}^{-1}(S^1_{min\mathcal{I}}|S^1_{mob})^T\Lambda^{1/2}\right]\cdot \\
&\quad\cdot\Lambda^{1/2}\left(S^1_{kin}|S^1_{sorp}|S^1_{min\mathcal{A}}\right)\begin{pmatrix}\Delta\xi_{kin}\\ \Delta\xi_{sorp}\\ \Delta\xi^{\mathcal{A}}_{min}\end{pmatrix}+O(\tau)+\text{r.h.s.}
\end{aligned}$$

With $D_{22}=(S^1_{min\mathcal{I}}|S^1_{mob})^T\Lambda(S^1_{min\mathcal{I}}|S^1_{mob})$ and $P_{(\Lambda^{1/2}({S^1_{min}}^{\mathcal{I}}|S^1_{mob}))^\perp}$ as defined in the lemma we obtain the desired structure of U_1, U_2, U_3.
$H:=(D_{11})^{22}={S^2_{sorp}}^T\tilde{\Lambda}S^2_{sorp}$ is symmetric positiv definite, since the columns of S^2_{sorp} are linearly independend and $\tilde{\Lambda}$ is a diagonal matrix with positive entries. Hence, also $(D_{11})^{22^{-1}}$ is symmetric positive definite. The other factor of U_2, $W:={S^1_{sorp}}^T\Lambda^{1/2}P_{(\Lambda^{1/2}(S^1_{min\mathcal{I}}|S^1_{mob}))^\perp}\Lambda^{1/2}S^1_{sorp}$, is obviously symmetric positive *semi*definite. By exploiting the linear independence of the columns

of $(S^1_{min\mathcal{I}}|S^1_{mob}|S^1_{sorp})$, the existence of a *zero* eigenvalue of W can be excluded (like in Footnote 17 on p. 39 or like the regularity proof of B on p. 37). Hence, also W is strictly positive definite. With Lemma 5.10, U_2 has a full system of eigenvectors with all eigenvalues being positive.
The formulas for the matrices V_1, V_2, V_3 can be derived in a similar way as the U_i. ∎

Acknowledgments

I like to thank my colleagues of the chair AM1 of the Department of Mathematics, University of Erlangen, in particular Joachim Hoffmann, Wilhelm Merz, Alexander Prechtel, for many fruitful discussions. Furthermore I like to thank the reviewers of this thesis, W. Jäger (University Heidelberg), M. Böhm (University Bremen), A. Bourgeat (University Lyon), and the mentors of the habilitation procedure G. Leugering, H.J. Tobschall, and in particular P. Knabner (University Erlangen-Nuremberg).
I also like to express my happiness about the decision of Joachim Hoffmann to choose the implementation of the algorithms proposed in Chapters 2 and 4 of this work as the topic of his diploma thesis and to continue this work in his Ph.D. thesis. Without such an implementation, the theoretically derived algorithms could not have been tested and confirmed by numerical experiments.
Last but not least I want to express how I enjoyed the invitations by and the discussions with C. Kanzow (University Würzburg), C. Wieners (University Karlsruhe), M. Kern (INRIA Rocquencourt, France), I. Aavatsmark (CIPR, Bergen, Norway), H. Hellevang (University Bergen, Norway).

Bibliography

[1] R. Adams, J. Fournier, *Sobolev spaces*, Elsevier Science, Oxford, (2nd ed.), 2003.

[2] H.W. Alt, S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311-341.

[3] L. Amir, M. Kern, *Newton-Krylov methods for coupling transport with chemistry in porous media*, CMWR XVI - XVI International Conference on Computational Methods in Water Resources, Copenhagen, Denmark, P. Binning, P. Engesgaard, H. Dahle, G. Pinder, W.G. Gray (editors), June 2006.

[4] T. Arbogast, S. Bryant, C. Dawson, F. Saaf, C. Wang, M. Wheeler, *Computational methods for multiphase flow and reactive transport problems arising in subsurface contaminant remediation*, J. Comput. Appl. Math., 74 (1996), pp. 19–32.

[5] R. Aris, R.H.S. Mah, *Independence of chemical reactions*, Ind. Eng. Chem. Fundam., 2 (1963), pp. 90–94.

[6] J.P. Aubin, A. Cellina, *Differential inclusions*, Springer, 1984.

[7] M. Bause, W. Merz, *Higher order regularity and approximation of solutions to the Monod biodegradation model*, Appl. Num. Math., 55 (2005), pp. 154–172.

[8] J. Bear, *Dynamics of fluids in porous media*, Elsevier Publishing, 1972.

[9] D. Bernstein, S. Bhat, *Nonnegativity, reducibility, and semistability of mass action kinetics*, Proceedings of the 38th Conference on Decision & Control, Phoenix, Arizona, 1999.

[10] D. Bernstein, D. Hyland, *Compartimental modeling and second-moment analysis of state space systems*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 880–901.

[11] C.M. Bethke, *Geochemical reaction modeling, concepts and applications*, Oxford Univ. Press, 1996.

[12] P. Biler, W. Hebisch, T. Nadzieja, *The Debye system: Existence and large time behaviour of solutions*, Nonlin. Anal., 23 (1994), pp. 1189–1209.

[13] A. Bourgeat, S. Bryant, J. Carrayrou, A. Dimier, C.V. van Duijn, M. Kern, P. Knabner, *GdR MoMaS - Benchmark Reactive Transport*, 2007 `http://www.gdrmomas.org/Ex_qualif/Geochimie/Documents/Benchmark-MoMAS.pdf`

[14] H. Brezis, *Operateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North-Holland, Amsterdam, 1973.

[15] J. Carrayrou, *Modélisation du transport de solutes réactifs en milieu poreux saturé*, doctoral thesis, Université de Strasbourg, 2001.

[16] J. Carrayrou, R. Mosé, P. Behra, *New efficient algorithm for solving thermodynamic chemistry*, AIChE J. 48 (2002), pp. 894-904.

[17] J. Carrayrou, R. Mosé, P. Behra, *Operator-splitting procedures for reactive transport and comparison of mass balance errors*, J. Cont. Hydrol., 68 (2004), pp. 239–268.

[18] A. Chilakapati, T. Ginn, J. Szecsody, *An analysis of complex reaction networks in groundwater modeling*, Water Resour. Res., 34 (1998), pp. 1767-1780.

[19] G. Cohen, P. Joly, J.E. Roberts, N. Tordjman, *Higher order triangular finite elements with mass lumping for the wave equation*, SIAM J. Numer. Anal. 38 (2001), pp. 2047–2078.

[20] M. De Simoni, J. Carrera, X. Sánchez-Vila, A. Guadagnini, *A procedure for the solution of multicomponent reactive transport problems*, Water Resour. Res., 41 (2005), W11410 doi:10.1029/2005WR004056.

[21] T. De Luca, F. Facchinei, C. Kanzow, *A theoretical and numerical comparison of some semismooth algorithms for complementarity problems*, Computational Optimization and Applications, 16 (2000), 173–205.

[22] P. Deuflhard, *Newton methods for nonlinear problems*, Springer, Berlin, 2004.

[23] C.J. van Duijn, P. Knabner, *Travelling wave behaviour of crystal dissolution in porous media flow*, European J. Appl. Math., 8 (1997), p. 455 ff.

[24] C.J. van Duijn, P. Knabner, R. Schotting, *An analysis of crystal dissolutions fronts of flows through porous media. Part 2: Incompatible boundary conditions*, Adv. Water Resour., 22 (1998), pp. 1–16.

[25] C.J. van Duijn, I.S. Pop, *Crystal dissolution and precipitation in porous media: pore scale analysis*, J. Reine Angew. Math., 577 (2004), 171–211.

[26] A.S. El-Bakry, R.A. Tapia, T. Tsuchiya, Y. Zhang, *On the formulation and theory of the Newton interior-point method for nonlinear programming*, J. Opt. Theory Appl., 89 (1996), 507–541.

[27] L.C. Evans, *Partial differential equations*, American Mathematical Society, Providence, 1998.

[28] Y. Fang, G.-T. Yeh, W.D. Burgos, *A general paradigm to model reaction-based biochemical processes in batch systems*, Water Resour. Res., 39 (2003), 1083, doi:10.1029/2002WR001694.

[29] I. Faragó, J. Geiser, *Iterative operator-splitting methods for linear problems*, International Journal of Computational Sciences, 1 (2005), pp. 64–74.

[30] M. Feinberg, *The existence and uniqueness of steady states for a class of chemical reaction networks*, Arch. Mech. Anal., 132 (1995), pp. 311-370.

[31] A. Fischer, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.

[32] J.C. Friedly, *Extent of reaction in open systems with multiple heterogeneous reactions*, AIChE J., 37, 1991, pp. 687–693.

[33] J.C. Friedly, J. Rubin, *Solute transport with multiple equilibrium-controlled or kinetically controlled chemical reactions*, Water Resour. Res., 28 (1992), pp. 1935–1953.

[34] C. Geiger, C. Kanzow, *Theorie und Numerik restringierter Optimierungsaufgaben*, Springer, 2002.

[35] J.W. Gibbs, *The scientific papers of J. Willard Gibbs, Vol. 1*, Dover Publ., New York, 1961 (1st ed. 1906).

[36] P. Hartman, *Ordinary differential equations*, John Wiley & Sons, 1964.

[37] H. Hellevang, P. Aagaard, E.H. Oelkers, B. Kvamme, *Can Dawsonite permanently trap* CO_2*?*, Environ. Sci. Technol. 39 (2005), pp. 8281–8287.

[38] H. Hellevang, S. Kumar, G. Fladmark, B. Kvamme, *CO2 storage in the Utsira Formation - ATHENA 3D reactive transport simulations*. Submitted to Basin Research.

[39] H. Hellevang, *Interactions between* CO_2*, saline water and minerals during geological storage of* CO_2, Doctoral Thesis, University of Bergen, Norway, 2006.

[40] H. Heuser, *Gewöhnliche Differentialgleichungen*, Teubner, Stuttgart, 1989.

[41] H. Heuser, *Lehrbuch der Analysis, Teil 2*, Teubner, Stuttgart, 1981.

[42] M. Hintermüller, K. Ito, K. Kunisch, *The primal-dual active set strategy as a Semismooth Newton method*, SIAM J. Optim. 13 (2003), pp. 865–888.

[43] J. Hoffmann, *Ein Entkopplungsverfahren für Systeme von Transportreaktionsgleichungen in porösen Medien: Algorithmische Realisierung und Simulation realistischer 2D-Szenarien*, Diploma Thesis, Universität Erlangen-Nürnberg, 2005.
`http://www.am.uni-erlangen.de/am1/en/theses.html`

[44] J. Hoffmann, Doctoral Thesis, University of Erlangen, planned for 2009.

[45] A. Holstad, *A mathematical and numerical model for reactive fluid flow systems*, Comput. Geosciences, 4 (2000), pp. 103–139.

[46] U. Hornung (ed.), *Homogenization and porous media*, Springer, New York, 1997.

[47] C. Kanzow, *Inexact semismooth Newton methods for large-scale complementarity problems*, Optimization Methods and Software, 19 (2004), pp. 309-325.

[48] S. Kräutle, *Reduction of the problem size through linear algebra for the problem of reactive transport in porous media*, in: Electronic Transactions on Numerical Analysis (ETNA), GAMM-SIAM Conference on Applied Linear Algebra, Düsseldorf 2006, A. Frommer, M. Hochbruck, B. Lang (editors), submitted.

[49] S. Kräutle, M. Bause, A. Prechtel, F. Radu, P. Knabner: *ParRichy: Parallel simulation of bioreactive multicomponent transport processes in porous media*, in: High Performance Computing in Science and Engineering, Garching 2004, Transaction of the KONWIHR Result Workshop, October 14-15, 2004, Technical University of Munich; A. Bode, F. Durst (editors), Springer, Berlin, pp. 181-192, 2005.

[50] S. Kräutle, 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.

[51] S. Kräutle, 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.

[52] N.V. Krylov, *Nonlinear elliptic and parabolic equations of the second order*, D. Reidel Publishing Company, 1987.

[53] B. Kvamme, H. Hellevang, University of Bergen, Norway, private communication, 2005-2006.

[54] P.C. Lichtner, *Continuum model for simultaneous chemical reactions and mass transport in hydrothermal systems*, Geochimica et Cosmochimica Acta, 49 (1985), pp. 779–800.

[55] P.C. Lichtner, *Continuum formulation of multicomponent-multiphase reactive transport*, in Reactive transport in porous media, Reviews in Mineralogy, Vol. 34, P.C. Lichtner, C.I. Steefel, E.H. Oelkers (eds.), Mineralogical Society of America, 1996, pp. 1–81.

[56] O.A. Ladyženskaja, V.A. Solonnikov, N.N. Uralceva, *Linear and quasilinear equations of parabolic type*, American Mathematical Society, 1968.

[57] G.I. Marchuk, *Splitting and alternating direction methods*, in Handbook of numerical analysis, Vol. 1, P.G. Ciarlet, J.L. Lions (eds.), North Holland, 1990, pp. 197–462.

[58] A. Muntean, M. Böhm, *A sharp-interface moving-boundary system modeling carbonation penetration in concrete*, Berichte aus der Technomathematik, Report 06-03, Universiy of Bremen, 2006.

[59] W. Merz, *Analysis und numerische Berechnung der Diffusion von Fremdatomen in homogenen Strukturen*, Habilitation Thesis, TU München, 1999.

[60] W. Merz, *Global existence result of the Monod model in bioremediation*, Adv. Math. Sci. Appl., 15 (2005), pp. 709–726.

[61] M. Mincheva, D. Siegel, *Stability of mass action reaction-diffusion systems*, Nonlin. Anal., 56 (2004), pp. 1105–1131.

[62] M. Mincheva, D. Siegel, *Nonnegativity and positiveness of solutions to mass action reaction-diffusion systems*, submitted for publication.

[63] S. Molins, J. Carrera, C. Ayora, M.W. Saaltink, *A formulation for decoupling components in reactive transport problems*, Water Resour. Res., 40 (2004), W10301, doi:10.1029/2003WR002970.

[64] F.M.M. Morel, J.G. Hering, *Principles and applications of aquatic chemistry*, Wiley, 1993.

[65] I.M. Nambi, C.J. Werth, R.A. Sanford, A.J. Valocchi, *Pore-scale analysis of anaerobic halorespiring bacterial growth along the transverse mixing zone of an etched silicon pore network*, Environ Sci Technol., 37 (2003), pp. 5617–5624.

[66] S. Pinchuk, *A counterexample to the strong real Jacobian conjecture*, Math. Z., 217 (1994), pp. 1–4.

[67] I. Prigogine, R. Defay, *Chemical Thermodynamics*, Longmans Green, Toronto, 1954.

[68] F. Rothe, *Global solutions of reaction-diffusion systems, Lecture notes in Mathematics*, Springer, 1984.

[69] F. Saaf, *A study of reactive transport phenomena in porous media*, Doctoral Thesis, Rice Univercity, Houston, Texas, 1996.

[70] M.W. Saaltink, C. Ayora, J. Carrera, *A mathematical formulation for reactive transport that eliminates mineral concentrations*, Water Resour. Res., 34 (1998), pp. 1649–1656.

[71] M.W. Saaltink, J. Carrera, C. Ayora, *A comparison of two approaches for reactive transport modelling*, J Geochem Expl, 69-70 (2000), pp. 97–101.

[72] H. Schaefer, *Über die Methode der a priori-Schranken*, Math. Annalen, 129 (1955), pp. 415–416.

[73] J.S. Selker, C.K. Keller, J.T. McCord, *Vadose zone processes*, Lewis Publishers, 1999.

[74] C.I. Steefel, K.T.B. MacQuarrie, *Approaches to modeling of reactive transport in porous media*, in: Reactive transport in porous media, Reviews in Mineralogy, Vol. 34, P.C. Lichtner, C.I. Steefel, E.H. Oelkers (editors), Mineralogical Society of America, pp. 83–129, 1996.

[75] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), pp. 506–517.

[76] W. Stumm, J.J. Morgan, *Aquatic chemistry*, Wiley Interscience Publication, 1996.

[77] A.J. Valocchi, M. Malmstead, *Accuracy of operator splitting for advection-dispersion-reaction problems*, Water Resour. Res., 28 (1992), pp. 1471–1476.

[78] J. van der Lee, L. De Windt, V. Lagneau, P. Goblet, *Module-oriented modeling of reactice transport with HYTEC*, Computers & Geosciences, 29 (2003), pp. 265–275.

[79] S. Whitaker, *The method of volume averaging*, Kluwer Academic Publishers, Dordrecht, 1998.

[80] W. Wieners, *Distributed point objects, A new concept for parallel finite elements*, in: Domain decomposition methods in science and engineering, Lecture notes in computational science and engineering, Vol. 40, R. Kornhuber, R. Hoppe, J. Périaux, O. Pironneau, O. Widlund, J. Xu (eds.), Springer 2004, pp. 175–183.

[81] D.E. Winterbone, *Advanced thermodynamics for engineers*, Elsevier, 1997.

[82] G.-T. Yeh, V.S. Tripathi, *A critical evaluation of recent developments in hydrogeochemical transport models of reactive multichemical components*, Water Resour. Res., 25 (1989), pp. 93–108.