The Semismooth Newton Method for the Solution of Reactive Transport Problems Including Mineral Precipitation-Dissolution Reactions

Hannes Buchholzer

Dissertation
University of Würzburg
Department of Mathematics

The Semismooth Newton Method for the Solution of Reactive Transport Problems Including Mineral Precipitation-Dissolution Reactions

Dissertation zur Erlangung
des naturwissenschaftlichen Doktorgrades
der Julius-Maximilians-Universität Würzburg

vorgelegt von

Hannes Buchholzer

aus
Hermannstadt

Eingereicht am: 2. Februar 2011
1. Gutachter: Prof. Dr. Christian Kanzow, Universität Würzburg
2. Gutachter: Dr. habil. Serge Kräutle, Universität Erlangen-Nürnberg

Acknowledgment

This present doctoral thesis is the result of over 3 years of research that I conducted at the Department of Mathematics at the University of Würzburg. I would like to seize this opportunity to express my thanks to some people who supported me during this time either scientifically or personally.

First of all there is my supervisor Christian Kanzow. He almost always took time (interrupting his present work) to answer questions. Furthermore he provided guidance, discussed ideas, had valuable suggestions and co-authored two articles with me. I could greatly benefit from his experience and knowledge. Also he was (and is) a pleasant person with whom I enjoyed working. Finally he even supported me in my struggle against English grammar and this and other publications. For all of this I want to express my deep gratitude to him.

This research project was a cooperation between the chair of Applied Mathematics II at the University of Würzburg under the head of C. Kanzow and the chair of Applied Mathematics I at the University of Erlangen-Nuremberg under the head of P. Knabner. While the work group in Erlangen provided the application from the field of hydrogeology together with all their knowledge and experience in this area, it was the task of the work group in Würzburg to contribute their experience in the field of optimization to this matter. In particular it was my task to investigate the mentioned application from the viewpoint of modern optimization. My main contact person from the work group in Erlangen was Serge Kräutle. He took the time to answer my questions with very comprehensive Emails that I could understand better than corresponding technical literature. From him I really learned very much about the main application and topics from neighboring fields like physics, chemistry and geoscience. I want to express my sincere gratitude for this and also for being a friendly pleasant person.

Furthermore I would like to thank my father for supporting me financially during my studies, which in the first place enabled me to come this far. Also I thank him for his readiness to help me otherwise as far as he could.

Last but not least I would like to thank my best friend for his support and friendship. This is invaluable to me.

Contents

1 Introduction

In the fundamental and vital papers [39] by Qi and Sun, and [38] by Qi semismooth Newton methods where introduced for the solution of nonlinear systems of equations that are defined by a nonsmooth mapping. Related material can also be found in Kummer [29, 30]. Similar to the classical Newton method these methods can be shown to be locally quadratically convergent under moderate assumptions that are somewhat similar to the assumptions that the classical method requires. Since the introductory publications from Qi and Sun, semismooth Newton methods have been used to solve many classes of mathematical problems that can be formulated in a suitable way as a nonsmooth system of equations, among them are complementary problems, variational inequalities and nonsmooth equation systems themselves, cf. [12a, 12b, 34]. In these days they are widely accepted, used and studied.

In this thesis we consider a particular class of applications from the field of computational geosciences, namely a reactive transport model in the subsurface including mineral precipitation-dissolution reactions. This model involves partial differential equations (PDEs), ordinary differential equations (ODEs), and algebraic equations, together with some complementarity conditions arising from the equilibrium conditions of the minerals. After discretization this results in a mixed complementary problem that can be equivalently written as a nonlinear and nonsmooth system of equations via so-called NCP-functions. This seems to be a new and promising approach in literature for problems of this kind.

In general, the modeling of reactive transport problems in porous media leads to systems involving PDEs and ODEs; the PDEs for the concentration of chemical species which are dissolved in the water (called mobile species), and the ODEs for the concentrations of species which are attached to the soil matrix and which are not subject to transport (called immobile species). In the following we assume that all of the immobile species, from a chemist's point of view, are minerals. Hence, the reactions with minerals are so-called precipitation-dissolution reactions.

In principle, it is possible to model reactions among the mobile and between mobile and immobile species by kinetic rate laws, i.e., the reactive source/sink terms in the PDEs and ODEs are given functions of the concentrations, coupling the PDEs and ODEs. If the reactions are sufficiently fast, then the assumption of local equilibrium instead is reasonable. This equilibrium assumption means that the concentrations of the involved species tend to a certain reaction dependent ratio between reactants and products. For reactions among mobile species, local equilibrium conditions can be described by (nonlinear) algebraic equations (AEs). However for reactions involving minerals, complementary conditions (CCs) are necessary to describe local equilibrium. Because in this case another state of equilibrium is possible: complete dissolution of the mineral. Then the precipitation-dissolution reaction for a particular mineral can only go in the forward direction forcing a different ratio between the reactants and the products (see Sec. 3.1). The resulting system consists of PDEs, ODEs, AEs, and CCs.

For the numerical solution, many publications on reactive transport in porous media suggest to enforce a decoupling between transport and reaction by applying an operator splitting technique. By this, the reaction subproblem is fully local, i.e., it consists only of AEs and CCs, while only the transport subproblem contains the PDEs and ODEs. However, operator splitting either introduces splitting errors or requires a fixed-point type iteration between transport and reaction within each time step. In the first case, accuracy considerations, and in the second case, convergence issues often lead to severe time step restrictions for splitting methods [45].

In the computational geoscience community there is a very popular way to solve the PDE-ODE-AE-CC system [3, 6]: In the current time step, for each mineral and each discretization point, an assumption is made (usually based on the previous time step) whether saturation or complete dissolution will hold. Through this assumptions the complementary conditions are replaced by AEs. Then a Newton iteration is performed. If the result has no physical meaning then the assumptions are modified in some way resulting in new AEs. Then the Newton iteration is repeated, until (hopefully) a physically meaningful solution is obtained. Besides its heuristic motivation, another drawback of this procedure is its lack of efficiency since several Newton iterations have to be executed in one time step. Also this approach needs constant human supervision and intervention.

Other authors from the geosciences community propose to use a formulation as a free boundary problem for front tracking approaches [32]. However, this approach lacks simplicity as soon as more than one space dimension is involved and topology changes of the precipitation-dissolution fronts appear. Another approach is to approximate the equilibrium, i.e., very fast reactions, by a kinetic description with large rate coefficients. Besides the approximative nature of this approach, large rate coefficients may increase the stiffness of the problem to solve.

Modern techniques from the optimization theory for the reactive transport problem are considered in [43, 46] and in [25]. In [43, 46], an operator splitting is performed, and the now fully local reaction problem is replaced by an equivalent constrained minimization problem for the so-called Gibbs free energy. Its KKT conditions are solved with an interior-point algorithm. Numerical test runs are performed without any deeper theoretical investigation. Note that this procedure leads to additional unknowns, the Lagrange multipliers for the equality and inequality constraints.

In [25, Sec. 4], to our knowledge for the first time, the application of a semismooth Newton method to the reactive transport mineral precipitation-dissolution problem is carried out. There the reactive transport problem is tackled fully implicitly, avoiding any operator splitting. The author considers a rather general situation of reactive problems including equilibrium and kinetic reactions, where the equilibrium reactions may be of the aqueous, the sorption, or the mineral precipitation-dissolution type. The implementation of the solution strategy is described and some results on the nonsingularity of the Jacobian of the system are given.

The present thesis propagates and investigates similar solution strategies as in [25, 26], but it focuses on those reactive systems without kinetic reactions, and where all the (equilibrium) reactions are of aquatic and of mineral type, i.e., no so-called sorption reactions are involved. This restriction allows a less technical presentation, to prove stronger theoretical results, and to exploit the overall structure in a better way so that the resulting algorithm can take full advantage of this structure. In contrast to these two publications the subject at hand is tackled more from the viewpoint of modern optimization.

The organization of this thesis is the following: In Chapter 2 theoretical foundations are laid concerning the semismooth Newton method, complementary conditions and related topics. In Chapter 3 background information is given, the problem is formulated and its mathematical model is given. Afterwards an equivalence transformation is applied to the PDE-ODE-AE-CC system (going back to [27, 28, 25]). The motivation for this reformulation is a decoupling of some (linear) PDEs, leading to a smaller nonlinear system. Finally the remaining system is discretized in space and time. The resulting system is a mixed complementarity problem that can be reformulated as a nonlinear (but nonsmooth) system of equations. In Chapter 4 the reformulation of the incorporated complementary problem is done with the minimum function as NCP-function. Then the semismooth Newton method is applied to solve the resulting nonlinear nonsmooth equation system. Afterwards a convergence analysis is performed and some topics related to this algorithm are investigated. Also local existence and uniqueness of a solution of this equation system is tackled. Finally a numerical example is presented. In Chapter 5 the mixed complementary problem is reformulated with the Fischer-Burmeister function resulting in a nonlinear nonsmooth equation system. Again the semismooth Newton method is applied to this equation system and a short convergence analysis is performed along with an investigation of related topics. Then a globalization strategy for this formulation is proposed. Finally a numerical example for the globalized semismooth Newton method is presented. In Chapter 6 we investigate a subproblem that appears in Chapter 4. Its solution plays a crucial role for very many results in both previous chapters (and really for the whole thesis). The matter is the boundedness of a special matrix valued function. In Chapter 7 we introduce a method to estimate the extremal singular values of a class of matrices that come from finite difference discretization of many convection diffusion PDEs. Also this method can determine very accurately whether such a matrix is positive definite. Finally in Chapter 8 we give some concluding remarks concerning this thesis.

2 Preliminaries

In this chapter we want to introduce all necessary theory and tools that we need in the subsequent chapters. The coverage and depth of the topics will be sufficient so that even undergraduate readers will be able to understand the text.

2.1 Notation

The n-dimensional real vector space is denoted by $\mathbb{R}^n$. Every vector $x \in \mathbb{R}^n$ is a column vector unless it is transposed. The components of a vector $x \in \mathbb{R}^n$ we denote with x_i. For a real valued function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ we denote with the column vector $\nabla f(x)$ the gradient of this function in point x where this function is differentiable. Whereas the differential i.e. the Jacobian, which we denote with $f'(x)$, is a row vector and it holds $f'(x) = \nabla f(x)^T$. Similarly , if $G : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, we write G_i for the i-th component function. The symbol $G'(x)$ is used for the Jacobian matrix of G at a differentiable point $x \in \mathbb{R}^n$ and $\nabla G_i(x)$ is the gradient of G_i at x. Also $\nabla G_i(x)^T$ is the i-th row of the Jacobian $G'(x)$.

For a vector $x \in \mathbb{R}^n$ the symbol $||x||$ denotes the Euclidean vector norm which we also denote with $||x||_2$. Furthermore with $||x||_\infty$ we denote the maximum norm of x and with $||x||_1$ we denote the 1-norm of x. With $\mathbb{R}^{m\times n}$ we denote the set of all real $m \times n$ matrices. For a matrix $M \in \mathbb{R}^{m\times n}$ we denote with $||M||$ and $||M||_{sp}$ the spectral norm of a matrix. And with $||M||_R$ we denote the row-sum norm, which is induced by the maximum vector norm, and $||M||_C$ denotes the column-sum norm, which is induced by the 1-norm. With I_n we denote the $n\times n$ identity matrix. If the size of the matrix is not important we leave out the subscript. We use [] to compose a matrix or vector out of scalar or block components, e.g. $[x_1, x_2, x_3]$ would be a 3-dimensional row vector. If the components are not scalar we sometimes use '|' as seperator instead of ',', e.g. $[A \mid B \mid C]$. And finally for column vectors $v_i \in \mathbb{R}^{n_i}$ $(i = 1, \ldots, m)$ the vector $[v_1, v_2, \ldots, v_m]$ shall be the column vector $[v_1^T, v_2^T, \ldots, v_m^T]^T$. For a $n \times m$ matrix $M = (m_{i,j})$ the notation $M^{\mathcal{A},\mathcal{B}}$ denotes the submatrix $(m_{i,j})_{i\in\mathcal{A}, j\in\mathcal{B}}$. Whereas $M^{\mathcal{A}}$ is the row selection $(m_{i,j})_{i\in\mathcal{A}, j=1,\ldots,m}$.

A function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is called a C^k-function if it is k times continually differentiable. In fact C^k is the set of all k-times continually differentiable functions.

With $\mathcal{A} \subset \mathcal{B}$ we denote the subset $\mathcal{A}$ of the set $\mathcal{B}$ where equality $\mathcal{A} = \mathcal{B}$ is included. The product $\mathcal{A} \cdot \mathcal{B} = \mathcal{A}\mathcal{B}$ of these two sets is simply the set $\{a \cdot b \mid a \in \mathcal{A} \wedge b \in \mathcal{B}\}$. Similarly the product of a matrix M with a set of matching vectors $\mathcal{B}$ is defined by $M \cdot \mathcal{B} = \{M \cdot b \mid b \in \mathcal{B}\}$. Likewise the sum of two sets is defined by $\mathcal{A}+\mathcal{B} = \{a + b \mid a \in \mathcal{A} \wedge b \in \mathcal{B}\}$. If $\mathcal{A}$ or $\mathcal{B}$ reduces to a singleton the set braces for this set are conveniently omitted, e.g. $\{1, 4\} + 3 = \{4, 7\}$. For sets we often use calligraphical letter, e.g. $\mathcal{A}$.

The set $B_r(x)$ denotes the open ball with radius r and center x and $\overline{B_r(x)}$ is its closure.

For real vectors $x, y \in \mathbb{R}^n$ the notation $x \geq y$ means that $x_i \geq y_i$ holds for every component. And $x > y$ is also meant component-wise.

2.2 Subdifferentials

In this section we want to introduce the concept of three subdifferentials which are interconnected.

Let $U \subset \mathbb{R}^n$ be an open set and $G : U \longrightarrow \mathbb{R}^m$ be a locally Lipschitz continuous function, i.e. for any $x \in U$ there is an $\epsilon > 0$ so that G reduced on $B_\epsilon(x)$ is Lipschitz continuous. Let $D_G \subset U$ be the set of differential points of G. From Rademacher's theorem [40] we know that the set of nondifferentiable points $U \setminus D_G$ has zero measure, i.e. G is differentiable almost everywhere.

Then the B-subdifferential of G in a point $x \in U$ is defined as

$$\partial_B G(x) := \left\{ H \in \mathbb{R}^{m \times n} \mid \exists \left(x^k\right)_{k \in \mathbb{N}} \subset D_G : \left(x^k\right) \longrightarrow x \text{ and } \left(G'\left(x^k\right)\right) \longrightarrow H \right\} ,$$

as it was defined in Qi [38]. And the generalized Jacobian by Clarke [7] is defined as

$$\partial G(x) := \operatorname{co}\left(\partial_B G(x)\right)$$

the convex hull of the B-subdifferential. In the special case where $m = 1$, we also call $\partial G(x)$ the generalized gradient of G. Hence the generalized gradient is a set of row vectors. Whereas the gradient of a differentiable real function is a column vector. And finally we define the C-subdifferential as

$$\partial_C G(x) := \left[\partial G_1(x) \times \partial G_2(x) \times \ldots \times \partial G_m(x)\right]^T .$$

In the first result we learn an important class of locally Lipschitz continuous functions.

Lemma 2.2.1. *Let $U \subset \mathbb{R}^n$be open and $G : U \longrightarrow \mathbb{R}^m$ be a continuously differentiable function. Then G is locally Lipschitz continuous.*

Proof. Let $x \in U$ be arbitrary. Since U is open there is $r > 0$ such that $\overline{B_r(x)} \subset U$ holds. Thanks to Theorem in [24, Sec. 3.2] with $K = \overline{B_r(x)}$ we can conclude that G is locally Lipschitz-continuous. □

Next we deal with the inclusion relation of these three subdifferentials and with the case of C^1-functions.

Proposition 2.2.2. *Let $G : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be locally Lipschitz continuous. Then:*

(a) $\partial_B G(x) \subseteq \partial G(x) \subseteq \partial_C G(x)$ *for all* $x \in \mathbb{R}^n$.

(b) G is continuously differentiable on an open set $D \subset \mathbb{R}^n$ if and only if $\partial G(x) = \{G'(x)\}$ for all $x \in D$.

Proof. See Clarke [7, Proposition 2.6.2(e)] together with the definitions above for (a) and Clarke [7, Corollary to Proposition 2.2.4] for (b). □

We give an easy example to illustrate the just defined subdifferentials. We consider the function $G : \mathbb{R} \longrightarrow \mathbb{R}^2$ given by $G(x) = (|x|, \min\{0, x\})^T$. This function is obviously Lipschitz-continuous. In every point $x \neq 0$ it is differentiable and therefore

$$\partial_B G(x) = \partial G(x) = \partial_C G(x) = \{G'(x)\}$$

holds. But for $x = 0$ the B-subdifferential for the component functions is

$$\partial_B G_1(0) = \{-1, +1\} \quad , \quad \partial_B G_2(0) = \{0, +1\}$$

and for the whole function

$$\partial_B G(0) = \left\{ \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\} .$$

Therefore the generalized Jacobian is the line segment

$$\partial G(0) = \left\{ \lambda \begin{pmatrix} -1 \\ 1 \end{pmatrix} + (1 - \lambda) \begin{pmatrix} 1 \\ 0 \end{pmatrix} \mid 0 \leq \lambda \leq 1 \right\}$$

while the C-subdifferential is

$$\partial_C G(0) = [-1, +1] \times [0, +1] ,$$

which is a rectangle.

A second and more complicated example is the Euclidean norm function $G : \mathbb{R}^n \longrightarrow \mathbb{R}$, $G(x) := \|x\|_2 = \sqrt{x_1^2 + \ldots + x_n^2}$. We list the generalized Jacobian and the B-subdifferential in the next result.

Lemma 2.2.3. *Let G be the Euclidean norm function as defined above. Then*

$$\begin{aligned} \partial_B G(0) &= \{x \in \mathbb{R}^n \mid \|x\|_2 = 1\} = \partial B_1(0) = S_{n-1} & (2.1) \\ \partial G(0) &= \{x \in \mathbb{R}^n \mid \|x\|_2 \leq 1\} = \overline{B_1(0)} & (2.2) \end{aligned}$$

and

$$\partial_B G\,(x) = \partial G\,(x) = \left\{ \frac{1}{\|x\|_2} \cdot x^T \right\}$$

for every vector $x \in \mathbb{R}^n \setminus \{0\}$.

Proof. The well known inequality

$$\big| \|x\|_2 - \|y\|_2 \big| \leq \|x - y\|_2$$

shows that the norm function is Lipschitz continuous on $\mathbb{R}^n$. If $x \in \mathbb{R}^n$ is nonzero, then both sets only contain the Jacobian, which is $\frac{1}{\|x\|_2} \cdot x^T$. So we now consider the case $x = 0$. It suffices to prove equation (2.1) because by taking the convex hull on both sides of this equation, we obtain equation (2.2).

First we show that $\partial_B G(0)$ is subset of the sphere $S_{n-1} = \{x \in \mathbb{R}^n \mid \|x\|_2 = 1\}$. Since $\|G'(x)\| = 1$ holds for every $x \neq 0$ we can conclude that $\left\|G'\left(x^k\right)\right\| \longrightarrow 1$ for every sequence $\left(x^k\right) \subset \mathbb{R}^n \setminus \{0\}$. From the definition of the B-subdifferential it follows that $\partial_B G(0) \subset S_{n-1}$ is true.

In order to verify the other inclusion let $x \in \mathbb{R}^n$ be an arbitrary vector with $\|x\| = 1$. Then the sequence $x^k := \frac{1}{k}x$ obviously converges to the zero vector. Whereas the corresponding sequence of the Jacobians $G'\left(x^k\right) = \frac{1}{k}x / \left\|\frac{1}{k}x\right\| = x$ obviously converges to x. That means that $S_{n-1} \subset \partial_B G(0)$ holds. Thus everything is proved. □

Now we are ready to calculate the subdifferentials for two important examples.

Example 2.2.4. The first example is the minimum function

$$\varphi_M : \mathbb{R}^2 \longrightarrow \mathbb{R},\ \varphi_M(a,b) = \min\{a,b\}\ .$$

It is obviously globally Lipschitz continuous and continuously differentiable for $a \neq b$. With Proposition 2.2.2 we have

$$\partial_B\varphi_M(a,b) = \partial\varphi_M(a,b) = \partial_C\varphi_M(a,b) = \begin{cases} \{(1,0)\} & \text{for } a < b \\ \{(0,1)\} & \text{for } a > b \end{cases} .$$

Now we study a point (a,b) with $a = b$. Let $\left(x^k\right) \subset \mathbb{R}^2$ be a sequence that converges to (a,b) with $x_1^k > x_2^k$ for almost all k. Then

$$\lim_{k\to\infty} \nabla\varphi_M(x_1^k, x_2^k)^T = (0,1)\,.$$

Similarly

$$\lim_{k\to\infty} \nabla\varphi_M(x_1^k, x_2^k)^T = (1,0)\,,$$

holds if $x_1^k < x_2^k$ for almost all k. For all other sequences $\left(x^k\right) \subset D_{\varphi_M}$ that converge to (a,b) the associated sequence $\nabla\varphi_M(x_1^k, x_2^k)^T$ does not converge. So we get as B-subdifferential

$$\partial_B\varphi_M(a,a) = \{(1,0),\ (0,1)\}$$

and consequently

$$\partial\varphi_M(a,a) = \partial_C\varphi_M(a,a) = \{(\lambda, 1-\lambda) \mid\ 0 \leq \lambda \leq 1\}\ .$$

The second example is the Fischer-Burmeister (FB) function

$$\varphi_F : \mathbb{R}^2 \longrightarrow \mathbb{R},\ \varphi_F(a,b) := \sqrt{a^2 + b^2} - a - b\,.$$

This function is continuously differentiable for $(a,b) \neq (0,0)$. And it is globally Lipschitz continuous, because it is the sum of a norm and a linear term.

For the differentiable points $[a, b] \neq [0, 0]$ we get the subdifferentials

$$\partial_B \varphi_F(a, b) = \partial \varphi_F(a, b) = \partial_C \varphi_F(a, b) = \left\{ \left[\frac{a}{\sqrt{a^2 + b^2}} - 1, \; \frac{b}{\sqrt{a^2 + b^2}} - 1 \right] \right\} .$$

The nondifferentiable point $[a, b] = [0, 0]$ is more difficult. First we notice that the Euclidean norm function $x \mapsto \|x\|_2$ is differentiable in $\mathbb{R}^2 \setminus \{0\}$, too. From Lemma 2.2.3 we know that its B-subdifferential in $[0, 0]$ is the unit sphere S_1. With the definition of the B-subdifferential we can see that the B-subdifferential of the Fischer-Burmeister function is the unit sphere translated by the vector $[-1, -1]^T$, i.e.

$$\partial_B \varphi_F(0, 0) = S_1 + \begin{bmatrix} -1 \\ -1 \end{bmatrix} = \partial B_1(-1, -1) .$$

Then the generalized Jacobian is the convex hull

$$\partial \varphi_F(0, 0) = \overline{B_1(-1, -1)} .$$

Since φ_F is a scalar function the C-subdifferential coincides with the generalized Jacobian.

These examples were not too complex. Here we could calculate the B-subdifferential directly by using its definition. In order to calculate the B-subdifferential and generalized Jacobian of more complex combined functions there are some handy rules to divide this task to simpler calculations. Hereby often equality is lost and the result is a superset of the desired set (see Clarke [7]). Here we introduce just one of these rules, for later use.

Proposition 2.2.5. *Let $f = g \circ G$, where $G : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is locally Lipschitz continuous and $g : \mathbb{R}^m \longrightarrow \mathbb{R}$ is continuous differentiable. Then*

$$\partial f(x) = g'(G(x))\, \partial G(x) .$$

Proof. This assertion is contained in Theorem 2.6.6 of Clarke [7]. □

Finally we cite two properties which we don't need explicitly. But they are useful for understanding the generalized Jacobian.

Proposition 2.2.6. *Let $G : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be locally Lipschitz continuous. Then:*

(a) $\partial G(x)$ is a nonempty, convex and compact subset of $\mathbb{R}^{m \times n}$ for every vector $x \in \mathbb{R}^n$.

(b) the mapping $x \mapsto \partial G(x)$ is upper semicontinuous at any $x \in \mathbb{R}^n$, i.e. for every $\epsilon > 0$ there is a $\delta > 0$ such that for every y with $\|y - y\| \leq \delta$, it holds

$$\partial G(y) \subset \partial G(x) + \epsilon \mathcal{B}_{m \times n}$$

where $\mathcal{B}_{m \times n}$ denotes the unit ball in $\mathbb{R}^{m \times n}$.

Proof. see Clarke [7, Theorem 2.6.2] □

The reader might be familiar with the notion of the convex subdifferential for a convex function $G : \mathbb{R}^n \longrightarrow \mathbb{R}$ which is usually also denoted with $\partial G(x)$ (see e.g. Rockafellar [41]). If G is additionally locally Lipschitz continuous then the generalized Jacobian $\partial G(x)$ is also defined. In [7] Clarke showed that in this case the generalized Jacobian coincides with the subdifferential in the sense of convex analysis.

2.3 Semismooth Functions

For the problem we study in this thesis we need the notion of semismooth functions. The set of these functions is a subset of the set of locally Lipschitz continuous functions and a superset of the set of continuously differentiable functions.

At first we remind the reader that a function $G : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is called directionally differentiable in a point $x \in \mathbb{R}^n$ if the limit

$$G'(x; d) := \lim_{t \downarrow 0} \frac{G(x + td) - G(x)}{t}$$

exists for all directions $d \in \mathbb{R}^n$. And G is called directionally differentiable on a open set $U \subset \mathbb{R}^n$ if it is directionally differentiable for all $x \in U$.

Definition 2.3.1. Let $U \subset \mathbb{R}^n$ be open and $G : U \longrightarrow \mathbb{R}^m$ be a locally Lipschitz continuous and directional differentiable function. Then G is called

1. *semismooth* in $x \in U$, if
$$\lim_{\substack{d \to 0 \\ H \in \partial G(x+d)}} \frac{Hd - G'(x; d)}{||d||} = 0$$
holds.

2. *strongly semismooth* in $x \in U$, if
$$\limsup_{\substack{d \to 0 \\ H \in \partial G(x+d)}} \frac{Hd - G'(x; d)}{||d||^2} < \infty$$
holds.

3. semismooth or strongly semismooth on U, if it is semismooth or strongly semismooth for all $x \in U$.

With the Landau symbols we can write the first and second definitions as

$$||Hd - G'(x; d)|| = o(||d||) \quad \text{and} \quad ||Hd - G'(x; d)|| = O(||d||^2) \tag{2.3}$$

for $d \to 0$ and all $H \in \partial G(x + d)$. A vector valued function is C^1 if and only if all component functions are C^1. The first Lemma gives an analogous result for semismooth functions.

Lemma 2.3.2. *Let $U \subset \mathbb{R}^n$ be open and $G : U \longrightarrow \mathbb{R}^m$. Then G is (strongly) semismooth in $x \in U$ if and only if every component function G_i is (strongly) semismooth in x.*

Proof. " $\Longleftarrow$ ". The definition of the C-subdifferential and the infinity norm give us this implication.

" $\Longrightarrow$ ". With the chain rule from Theorem 2.3.7 applied to the function $g := f \circ G$ with the C^1 function $f : \mathbb{R}^m \longrightarrow \mathbb{R}$, $f(x) := x_i$ we get the generalized gradient

$$\begin{aligned} \partial g(x) &= \partial f(G(x))\, \partial G(x) \\ &= e_i^T \cdot \partial G(x) \\ &= \left\{ h_i^T \mid h_i^T \text{ is the } i - \text{th row for some } H \in \partial G(x) \right\}, \end{aligned}$$

where e_i is the i-th canonical unit vector. Let $d \in \mathbb{R}^n$ be an arbitrary direction and $i \in \{1, \ldots, m\}$ an arbitrary index. For $h_i^T \in \partial G_i(x)$ there is a $H \in \partial G(x)$ so that its i-th row $e_i^T H$ equals h_i^T. Then we have

$$\begin{aligned} \left| h_i^T d - G_i^{'}(x; d) \right| &= \left| (Hd - G'(x; d))_i \right| \\ &\leq \|Hd - G'(x; d)\|_\infty \end{aligned}$$

and with (2.3) for G it follows that G_i is (strongly) semismooth. □

The following Lemma introduces an important class of semismooth functions.

Lemma 2.3.3. *Let $U \subset \mathbb{R}^n$be open, $x \in U$ and $G : U \longrightarrow \mathbb{R}^m$ be a function. Then*

1. *If G is continuously differentiable around x, then G is semismooth in x.*
2. *If G is differentiable around x and G′ is Lipschitz continuous around x, then G is strongly semismooth in x.*

Proof. With [7, Propositon 2.2.4 and next Corollary] we have that $\partial G(x) = \{G'(x)\}$ and since G is continuously differentiable around x (i.e. in an open ball $B_\epsilon(x)$) it holds $G'(x) \cdot d = G'(x; d)$ for all $d \in \mathbb{R}^n$. Then we have

$$\begin{aligned} \lim_{\substack{d \to 0 \\ H \in \partial G(x+d)}} \frac{Hd - G'(x; d)}{\|d\|} &= \lim_{d \to 0} \frac{G'(x + d) \cdot d - G'(x) \cdot d}{\|d\|} \\ &= \lim_{d \to 0} (G'(x + d) - G'(x)) \frac{d}{\|d\|} = 0 \end{aligned}$$

since G' is continuous around x. This proves the first assertion. To see the second assertion we mention that there is a Lipschitz constant K and $\epsilon > 0$ so that

$$\|G'(y) - G'(z)\| \leq K \cdot \|y - z\|$$

for all $y, z \in B_\epsilon(x)$. For $d \in \mathbb{R}^n$ with $x + d \in B_\epsilon(x)$ and $H \in \partial G(x + d)$ we can conclude

$$\begin{aligned} \frac{\|Hd - G'(x; d)\|}{\|d\|^2} &= \frac{\|G'(x + d) - G'(x)\| \cdot \|d\|}{\|d\|^2} \\ &\leq K \end{aligned}$$

and therefore

$$\limsup_{\substack{d \to 0 \\ H \in \partial G(x+d)}} \frac{Hd - G'(x; d)}{\|d\|^2} < \infty$$

and G is strongly semismooth. □

For the sake of convenience we derive a Corollary from this result.

Corollary 2.3.4. *Let $U \subset \mathbb{R}^n$ be open and $G : U \longrightarrow \mathbb{R}^m$ be a C^2-function. Then G is strongly semismooth.*

Proof. The differential $x \mapsto G'(x)$ is continuously differentiable. Let $x \in U$ be arbitrary. Since U is open there is $r > 0$ such that $\overline{B_r(x)} \subset U$ holds. Thanks to the Theorem in [24, Sec. 3.2] applied with $K = \overline{B_r(x)}$ we can conclude that the differential is locally Lipschitz-continuous. Now the previous Lemma gives the assertion. □

Another important class of semismooth functions are convex functions.

Lemma 2.3.5. *Let $U \subset \mathbb{R}^n$ be open and convex and let $f : U \longrightarrow \mathbb{R}$ be a convex function. Then f is semismooth on U.*

Proof. see Mifflin [35, Proposition 3] □

Now it is time to study two examples. We continue with the functions from Example 2.2.4.

Example 2.3.6. The first example is the minimum function φ_M. We show that it is strongly semismooth. It is obvious that φ_M is even globally Lipschitz continuous. In every point $[a, b]$ with $a \neq b$ the minimum function is continuously differentiable with a locally Lipschitz continuous derivative and therefore strongly semismooth. In the following we only consider points $[a, b]$ with $a = b$. Let $d = [d_1, d_2]^T \in \mathbb{R}^2$ be a direction vector. By means of a simple calculation one can see that

$$\varphi_M' ([a, b]; [d_1, d_2]) = \min \{d_1, d_2\}$$

holds. Now we have to make a distinction of cases.
Firstly we consider the case $d_1 < d_2$. Then φ_M is differentiable in $[a + d_1,\ b + d_2]$ and therefore the generalized Jacobian is $\partial\varphi_M (a + d_1,\ b + d_2) = \{[1, 0]\}$. Hence for $H \in \partial\varphi_M (a + d_1,\ b + d_2)$ we have

$$\begin{aligned} \varphi_M' ([a, b]; [d_1, d_2]) - H \cdot d &= \min \{d_1, d_2\} - [1, 0] \cdot \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} \\ &= d_1 - d_1 = 0\,. \end{aligned}$$

Secondly we study the case $d_1 > d_2$ which goes very similar to the previous one. So for $H \in \partial\varphi_M (a + d_1,\ b + d_2) = \{[1, 0]\}$ we have

$$\begin{aligned} \varphi_M' ([a, b]; [d_1, d_2]) - H \cdot d &= \min \{d_1, d_2\} - [1, 0] \cdot \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} \\ &= d_2 - d_2 = 0\,. \end{aligned}$$

Thirdly we study the more difficult case of $d_1 = d_2$. From Example 2.2.4 we know that $\partial_B\varphi_M (a + d_1,\ b + d_2)$ equals $\{[1, 0]\,,\ [0, 1]\}$ and therefore

$$\partial G (a + d_1,\ b + d_2) = \{[\lambda,\ 1 - \lambda] \mid \lambda \in [0, 1]\}$$

holds. For an arbitrary $H \in \partial\varphi_M(a + d_1, b + d_2)$ there is a unique $\lambda \in [0, 1]$ with $H = [\lambda, 1 - \lambda]$ and we get

$$\begin{aligned}\partial\varphi_M(a + d_1, b + d_2) &= \min\{d_1, d_2\} - [\lambda, 1] - \lambda \cdot \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} \\ &= d_1 - \lambda d_1 - d_2 + \lambda d_2 = 0 .\end{aligned}$$

In all three cases it holds

$$\limsup_{\substack{d\to 0\\ H\in\partial\varphi_M(x+d)}} \frac{Hd - \varphi_M'(x; d)}{\|d\|^2} = 0$$

and we have shown directly with the definition that G is strongly semismooth.

The second example is the Fischer-Burmeister function $\varphi_F : \mathbb{R}^2 \longrightarrow \mathbb{R}$, $\varphi_F(a, b) = \sqrt{a^2 + b^2} - a - b$ .We show that this function is strongly semismooth, too. By means of a simple calculation one can directly verify that φ_F is a convex function (it is also a well known fact in the optimization community). Then φ_F is semismooth according to Lemma 2.3.5.

We show now that it is even strongly semismooth. In every point $[a, b] \neq [0, 0]$ is φ_F differentiable and φ_F' is locally Lipschitz continuous. Let $d = [d_1, d_2]^T$ be a *nonzero* real vector. Then the directional derivative is

$$\varphi_F'([0, 0]; [d_1, d_2]) = \lim_{t\downarrow 0} \frac{\varphi_F(td_1, td_2) - \varphi_F(0, 0)}{t} = \varphi_F(d_1, d_1)$$

and the generalized Jacobian is

$$\partial\varphi_F(0 + d) = \{\varphi_F'(d)\} = \left\{\left[\frac{d_1}{\sqrt{d_1^2 + d_2^2}} - 1,\ \frac{d_2}{\sqrt{d_1^2 + d_2^2}} - 1\right]\right\} .$$

With $H \in \partial\varphi_F(d)$ we have

$$\begin{aligned}H \cdot d &= \left[\frac{d_1}{\sqrt{d_1^2 + d_2^2}} - 1,\ \frac{d_2}{\sqrt{d_1^2 + d_2^2}} - 1\right] \cdot \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} \\ &= \sqrt{d_1^2 + d_2^2} - d_1 - d_2 \\ &= \varphi_F(d_1, d_2)\end{aligned}$$

and therefore it holds for $d \neq 0$

$$Hd - \varphi_F'([0, 0]; [d_1, d_2]) = \varphi_F(d_1, d_2) - \varphi_F(d_1, d_2) = 0 .$$

With this we calculate the limit

$$\limsup_{\substack{d \to 0\\ H \in \partial\varphi_F(x + d)}} \frac{Hd - \varphi_F'(x; d)}{\|d\|^2} = 0$$

$$\limsup_{\substack{d\to 0\\ H\in\partial\varphi_F(x+d)}} \frac{Hd-\varphi_F'(x;d)}{\|d\|^2}=0$$

and φ_F is strongly semismooth.

If a function is more complex it might be quite difficult to show directly with the definition that a function is semismooth. Therefore we will introduce a chain rule property for semismooth functions.

Theorem 2.3.7. *Let $G:\mathbb{R}^m \longrightarrow \mathbb{R}^p$ and $F:\mathbb{R}^n \longrightarrow \mathbb{R}^m$ be directional differentiable and locally Lipschitz continuous functions and $H:=G\circ F:\mathbb{R}^n \longrightarrow \mathbb{R}^p$. Then*

(a) If F is semismooth in $x\in\mathbb{R}^n$ and G is semismooth in $F(x)\in\mathbb{R}^m$ then H is semismooth in x.

(a) If F is strongly semismooth in $x\in\mathbb{R}^n$ and G is strongly semismooth in $F(x)\in\mathbb{R}^m$ then H is strongly semismooth in x.

Proof. see Fischer [11, Theorem 19] □

From this theorem we can immediately derive the following Corollary. It shows that the sum, product and quotient of (strongly) semismooth functions is again (strongly) semismooth if it is defined.

Corollary 2.3.8. *Let $G_1,G_2:\mathbb{R}^n \longrightarrow \mathbb{R}$ be (strongly) semismooth in $x\in\mathbb{R}^n$. Then*

(a) The linear combination $\alpha_1 G_1+\alpha_2 G_2$ is (strongly) semismooth in x for arbitrary numbers $\alpha_1,\alpha_2\in\mathbb{R}$.

(b) The product $G_1\cdot G_2$ is (strongly) semismooth in x.

(c) The quotient $\frac{G_1}{G_2}$ is (strongly) semismooth in x if $G_2(x)\neq 0$ holds.

Proof. The functions $(x,y)\mapsto \alpha_1\cdot x+\alpha_2\cdot y$, $(x,y)\mapsto x\cdot y$ and $(x,y)\mapsto \frac{x}{y}, y\neq 0$ are all C^2 functions and therefore strongly semismooth (see Corollary 2.3.4). Thanks to Lemma 2.3.2 is the vector valued function $x\mapsto [G_1(x),G_2(x)]^T$ (strongly) semismooth. Therefore Theorem 2.3.7 yields that the concatenations are (strongly) semismooth. □

With this result we have some simple but powerful tools to show that a function is (strongly) semismooth. Of course statement (a) holds even for vector valued functions $G_1,G_2:\mathbb{R}^n \longrightarrow \mathbb{R}^m$ because of Corollary 2.3.2. The other statements are in the first instance only defined for scalar valued functions.

2.4 Complementary Problems and NCP-Functions

In this section we introduce the complementary problem, which plays an important part in this thesis. Then we will show how to reformulate this problem via NCP-functions. We start with the definition.

Definition 2.4.1. Let $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be a function. A (nonlinear) *complementary problem* is the task to find a vector $x^* \in \mathbb{R}^n$ that it is a solution of the following system of inequalities and equalities

$$x \geq 0,\ F(x) \geq 0,\ x^T F(x) = 0\,.$$

We denote this problem with the abbreviation NCP or $NCP(F)$.

This problem formulation can be written equivalently as

$$x_i \geq 0,\ F_i(x) \geq 0,\ x_i \cdot F_i(x) = 0 \qquad , i = 1, \ldots, n\,.$$

If F is a affine function we speak of a linear complementary problem.

It is important to discriminate two types of solutions for NCPs. In the *degenerate* solution x^* is a component i_0 such that $x^*_{i_0} = F_{i_0}(x^*) = 0$ holds. For the *nondegenerate* solution x^* no such component exists, i.e. for $i = 1, \ldots, n$ it holds $x^*_i + F_i(x^*) > 0$. The degenerate case is numerically more difficult.

Examples which lead to complementary problems are the Nash equilibrium problem, the barrier problem and KKT conditions. We show a simple case of the last one.

Example 2.4.2. Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a convex and continuously differentiable function. Then the simple optimization problem

$$\min f(x) \qquad s.t.\, x \geq 0$$

is equivalent to the KKT-conditions

$$x \geq 0,\ \nabla f(x) \geq 0,\ x^T \nabla f(x) = 0$$

which obviously form a complementary problem.

We do not solve the NCP directly. Instead we reformulate it equivalently into a well known problem. One way to do this is with NCP-functions.

Definition 2.4.3. A function $\varphi : \mathbb{R}^2 \longrightarrow \mathbb{R}$ with the property

$$\varphi(a, b) = 0 \quad \Longleftrightarrow \quad a \geq 0,\ b \geq 0,\ a \cdot b = 0$$

is called NCP-function.

The already mentioned Fischer-Burmeister function φ_F and the minimum function φ_M are NCP-functions. This is obvious for the minimum function. We verify this for φ_F. From squaring $\varphi_F(a, b) = 0$ it follows that

$$a^2 + b^2 = (a + b)^2$$

and from this we conclude that

$$a \cdot b = 0 \, .$$

And $\varphi_F(a,b) = 0$ is equivalent with

$$a + b = \sqrt{a^2 + b^2} \geq 0$$

together with $a \cdot b = 0$ this means that either $a = 0,\ b \geq 0$ or $a \geq 0,\ b = 0$ holds. This is the first implication. The other implication can be directly verified with the same case distinction $a = 0,\ b \geq 0$ and $a \geq 0,\ b = 0$. Both functions φ_F and φ_M are not differentiable but strongly semismooth (see example 2.3.6).

We give a few examples of differentiable NCP-functions which are not important for this thesis. We leave the verification to the reader.

Example 2.4.4.

(a) $\varphi(a,b) := -ab + \frac{1}{2}\min^2\{0, a+b\}$

(b) $\varphi(a,b) := -ab + \frac{1}{2}\min^2\{0, a\} + \frac{1}{2}\min^2\{0, b\}$

Let $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be a function and $\varphi : \mathbb{R}^2 \longrightarrow \mathbb{R}$ be a NCP-function. Then we define the vector-valued function $\Phi : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ as

$$\Phi(x) := \begin{pmatrix} \varphi\,(x_1,\ F_1(x)) \\ \varphi\,(x_2,\ F_2(x)) \\ \vdots \\ \varphi\,(x_n,\ F_n(x)) \end{pmatrix} . \tag{2.4}$$

The following result describes the connection of Φ to the complementary problem.

Theorem 2.4.5. *A vector $x^* \in \mathbb{R}^n$ is solution of the complementary problem NCP(F) if and only if $x^*$is a solution of the nonlinear equation system $\Phi(x) = 0$.*

Proof. From the definition of Φ and the property of NCP-functions it follows immediately

$$\begin{aligned} \Phi(x) = 0 \quad &\Longleftrightarrow \quad \varphi\,(x_i,\ F_i(x)) \quad \forall i = 1, \ldots, n \\ &\Longleftrightarrow \quad x_i \geq 0,\ F_i(x) \geq 0,\ x_i \cdot F_i(x) = 0 \quad \forall i = 1, \ldots, n \end{aligned}$$

which is the assertion of this theorem.

With this theorem we have reduced the complementary problem to the well known problem of solving a nonlinear system of equations. If F and the NCP-function φ are continuously differentiable then Φ is also continuously differentiable and we can solve the equation system $\Phi(x) = 0$ e.g. with Newton's method. A further requirement for Newton's method is that the Jacobian $\Phi'\,(x^*)$ in the solution x^* has to be nonsingular. The next result shows that this might not be fulfilled in the given context. □

Theorem 2.4.6. *Let $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ and $\varphi : \mathbb{R}^2 \longrightarrow \mathbb{R}$ be differentiable and x^* be a degenerate solution of NCP(F). Then the Jacobian $\Phi^{'}\,(x^*)$ contains a zero row i, where $x_i^* = F_i(x^*) = 0$ holds, and is therefore singular.*

Proof. Since φ and F are differentiable the composed function $\varphi(x_i, F_i(x))$ if differentiable and we calculate the Jacobian of $x \mapsto \varphi(x_i, F_i(x))$ with the chain rule. This clearly gives

$$\frac{\partial\varphi(x_i, F_i(x))}{\partial x} = \frac{\partial\varphi(x_i, F_i(x))}{\partial a} \cdot \frac{\partial F_i(x)}{\partial x} + \frac{\partial\varphi(x_i, F_i(x))}{\partial b} \cdot e_i^T$$

where e_i is the i-th unit vector. Since x^* is degenerate, there is an index i with $F_i(x^*) = x_i^* = 0$. Now let i be such an index. Since $\varphi(x_i, F_i(x))$ is the i-th component function of Φ the row vector $\frac{\partial\varphi(x_i,F_i(x))}{\partial x}$ is the i-th row of the Jacobian $\Phi'(x^*)$. The proof is complete if we show that $\nabla\varphi(0,0) = (0,0)^T$ holds. For the first partial derivative we have

$$\frac{\partial\varphi(0,0)}{\partial a} = \lim_{t\downarrow 0} \frac{\varphi(t,0) - \varphi(0,0)}{t} = \lim_{t\downarrow 0} \frac{0-0}{t} = 0$$

which follows from the NCP-function definition. In the same way we have

$$\frac{\partial\varphi(0,0)}{\partial b} = 0\,.$$

□

If one deals with large complementary problems, i.e. n is a large number, then it is quite likely that a solution contains a component that is at least numerically degenerate. With numerically degenerate we mean that there is a component i so that the absolute value of $F_i(x^*)$ and x_i^* is very small.

Remark 2.4.7. We want to show here that the assertion in Theorem 2.4.6 holds for a certain more general problem class than NCP(F), which will be important later. Let n_1, n_2 be integers with $n := n_1 + n_2$. Furthermore let $F : \mathbb{R}^{n_1} \longrightarrow \mathbb{R}^{n_2}$ and $G : \mathbb{R}^n \longrightarrow \mathbb{R}^{n_1}$ be continuously differentiable functions. For the variables $x \in \mathbb{R}^{n_1}$ and $y \in \mathbb{R}^{n_2}$ we seek solutions of the problem

$$\begin{aligned} G(x,y) &= 0 \\ x \cdot F(y) &= 0 \\ x \geq 0 \quad &, \quad F(y) \geq 0\,. \end{aligned}$$

We reformulate this problem equivalently as the nonlinear equation $\Psi(x,y) = 0$ by defining the component functions of $\Psi : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ as

$$\begin{aligned} \Psi_i(x,y) &:= G_i(x,y) && i = 1,\ldots,n_1 \\ \Psi_i(x,y) &:= \varphi(x_{i-n_1},\, F_{i-n_1}(y)) && i = n_1+1,\ldots,n_1+n_2\,. \end{aligned}$$

In a degenerate solution (x^*, y^*) there is an index j with $x_j^* = F_j(y)^* = 0$. We have seen in the last proof that $\varphi'(0,0) = (0,0)$ holds for every NCP-function. Via the chain rule one can see immediately that the Jacobian of $(x,y) \mapsto \varphi\big(x_j,\, F_j(y)\big)$ vanishes in (x^*, y^*). Therefor the $j + n_1$ row of the Jacobian $\Psi'(x^*, y^*)$ is zero and the whole matrix singular.

This problem can be avoided if we take a non-differentiable function as NCP-function in the definition of Φ. Then we need an algorithm that can solve $\Phi(x) = 0$ even for a certain class of non-smooth functions. Such an algorithm exists for semismooth functions. For the rest of this section let φ be either the minimum or the Fischer-Burmeister function and let Φ be defined accordingly.

Theorem 2.4.8. *Let $\varphi \in \{\varphi_M, \varphi_F\}$ and let F be two times continuously differentiable in the definition of Φ in (2.4). Then Φ is strongly semismooth.*

Proof. For all $i = 1, \dots, n$ the function $x \mapsto (x_i, F_i(x))$ is obviously a C^2-function and therefore strongly semismooth (see Corollary 2.3.4). Example 2.3.6 shows that the Fischer-Burmeister and the minimum functions are strongly semismooth. With the chain rule Theorem 2.3.7 we conclude that the i-th component $\Phi_i(x) = \varphi(x_i, F_i(x))$ is strongly semismooth for all $i = 1, \dots, n$. Then Lemma 2.3.2 gives the assertion. □

In the next section we answer the question how nonlinear equation systems with a semismooth left hand side can be solved efficiently.

2.5 Newton Method for Semismooth Functions

In this section we will introduce the semismooth Newton method, which is a version of Newton's method, for solving nonlinear equation systems under weaker requirements. The theory of the semismooth Newton method was developed by Qi [38], see also [12a, 12b, 39, 36] for related material.

Let $G : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a given function and consider the problem of finding a solution $x^* \in \mathbb{R}^n$ for

$$G(x) = 0 \, .$$

If G is differentiable we can try to solve this with the classical Newton method. It produces a sequence $\left(x^i\right) \subset \mathbb{R}^n$ according to the rule

$$x^{i+1} = x^i - G'\left(x^i\right)^{-1} \cdot G\left(x^i\right) \qquad i = 0, 1, 2, \dots$$

for a starting vector $x^0 \in \mathbb{R}^n$. If G is not differentiable then the Jacobian $G'\left(x^i\right)$ might not exist and the next iterate x^i is not defined. With the theory of subdifferentials from subsection 2.2 it presents itself to generalize Newton's method for locally Lipschitz continuous functions G as follows

$$x^{i+1} = x^i - H_i^{-1} \cdot G\left(x^i\right) \qquad i = 0, 1, 2, \dots$$

where $H_i \in \partial G\left(x^i\right)$. In the following algorithm we restrain ourselves to the B-subdifferential but the generalized Jacobian would be equally possible.

Algorithm 2.5.1 (Semismooth Newton method)**.**

(S.0) (Initialization)
Choose $x^0 \in \mathbb{R}^n$, $\epsilon \geq 0$ *and set* $k := 0$.

(S.1) (Termination Criterion)
If $\left\|G\left(x^k\right)\right\| \leq \epsilon$, *stop.*

(S.2) (Newton Direction Calculation)
Choose a matrix $H_k \in \partial_B G\left(x^k\right)$ *and find a solution* d^k *of the linear system*

$$H_k d = -G\left(x^k\right).$$

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

In the termination criterion one can use any norm. But it is sometimes useful to choose a certain norm. For differentiable functions G this algorithm reduces to the classical one since $\partial_B G\left(x^i\right) = \left\{G'\left(x^i\right)\right\}$ holds for such functions.

In the rest of this section we will show that this algorithm has the same local convergence properties like the classical Newton method if certain requirements are met. Note that the use of the generalized Jacobian would not alter the convergence properties but a little the requirements thereof. We introduce this algorithm in the form in which it is applied in the following chapters. One requirement for convergence is the following regularity condition.

Definition 2.5.2. Let $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be Lipschitz continuous in $x \in \mathbb{R}^n$. Then x is called *BD-regular* if all matrices $H \in \partial_B G(x)$ are nonsingular.

Again this is a not a surprising generalization of the classical Newton's method. For differentiable functions it reduces to the well known condition that $G'(x)$ is nonsingular. This regularity condition is of course necessary to ensure the solvability of the linear system in the algorithm. We regularity condition in a simple example. We consider the scalar function $G(x) = |x|$. In the solution $x^* = 0$ we have $\partial_B G(0) = \{-1, 1\}$. Therefore $x = 0$ is BD-regular for this function. On the other hand does $\partial G(0) = [-1, 1]$ contain zero. So the generalized Jacobian does not fulfill a similar regularity condition and therefore does not meet an important requirement for convergence. This is an important advantage of the formulation of the semismooth Newton method with the B-subdifferential.

The next step toward a convergence result is the following Lemma.

Lemma 2.5.3. *Let* $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ *be locally Lipschitz continuous and let* $x^* \in \mathbb{R}^n$ *be a BD-regular point of G. Then there are numbers* $\epsilon > 0$ *and* $c > 0$ *so that all matrices* $H \in \partial_B G(x)$ *for all points* $x \in B_\epsilon(x^*)$ *are nonsingular and*

$$\left\|H^{-1}\right\|_{sp} \leq c \qquad \forall H \in \partial_B G(x) \quad \forall x \in B_\epsilon(x^*).$$

Proof. see [39, Proposition 3.1] □

Again this result reduces to a well known Lemma for differentiable functions. Finally we can state the main theorem.

Theorem 2.5.4. *Let $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be semismooth and let x^* be a BD-regular solution for $G(x) = 0$. Then there is a $\epsilon > 0$ so that for every starting vector $x^0 \in B_\epsilon(x^*)$ the following holds:*

(a) Algorithm 2.5.1 is well defined and produces a sequence $\left(x^i\right)_{i\in\mathbb{N}_0}$ which converges to x^*.

(b) The convergence rate is superlinear, i.e. $\left\|x^{i+1} - x^*\right\| = o\left(\left\|x^i - x^*\right\|\right)$

(c) If G is strongly semismooth then the convergence rate is quadratic, i.e. $\left\|x^{i+1} - x^*\right\| = O\left(\left\|x^i - x^*\right\|^2\right)$

Proof. see [39, Theorem 3.2] □

With this amazing result, we can solve nonlinear systems with nondifferentiable functions if they are semismooth. In particular we can solve nonlinear systems that stem from complementary problems. If these problems were reformulated with the Fischer-Burmeister or the minimum function then the resulting left hand side is even strongly semismooth (provided that F is sufficiently smooth).

3 Problem Formulation and Transformation

In this chapter we introduce the problem class that we investigate in this thesis. We start with background information and from this derive a mathematical model consisting of PDEs, ODEs, AEs (algebraic equations) and CCs (complementary conditions). Then we will apply some transformations to this system and discretize the resulting system. Then we have finally arrived to the finite dimensional model which we will study in the following chapters. For the sake of readers who are not familiar with the geological background we will be quite comprehensive and descriptive.

Please note that neither is the author an expert in hydrology resp. hydrogeology nor are you the reader expected to be one. Therefore the use of special terminology will be simplified, generalized and more descriptive but not always scientifically exact. Note also that the further chapters deal only with the equation system that is derived in this chapter. Therefore it is not necessary to understand all the hydrogeological background fully in order to understand the following chapters. In fact the results in the further chapters are true even if they are derived from a totally different background.

3.1 Background

3.1.1 Overall Scenario

A porous medium (typically an aquifer) consists of pores which are fully or partially filled with water and a solid (in which the pores are embedded) that is called the soil matrix. The pores are sufficiently interconnected so that water can flow through the porous medium. In the water are many chemical substances (ions or molecules) present which interact with each other in chemical reactions. Also they are transported with the water flow and they are also subject to dispersion and diffusion (which we will neglect in this thesis). We refer to them as mobile species. We assume that the occurring reactions are “sufficiently” fast. To the soil matrix attached are minerals (solids). They engage in precipitation-dissolution reactions, i.e. they are dissolved in the water by decomposing into ions. In their unsolved solid state they are not transported with the water and are therefore called immobile species. The unknowns are the concentrations of the dissolved species in the water (in moles per water volume) and the concentrations of the minerals (pure solids) attached to the soil matrix (in moles per surface area of the soil matrix). This microscopic view helps to understand what processes are taking place. But for investigating large areas in the realm of tens through hundreds of meters is this view not suitable. Through volume averaging and homogenization is this model upscaled to a macroscopic view. Homogenization means to treat the porous medium as if every part of it would have the same composition. The qualities of the homogenized medium are the averaged qualities of the actual porous medium. For example we will postulate that a chemical equilibrium equation holds on every point of the domain although it refers to a chemical process which only takes place in the water in a pore.

The macroscopic model will be the basis of this thesis.

3.1.2 Index of Technical Terms

In this subsection we explain some technical terms. The reader who is only interested in mathematics can skip this part and can later look them up as he needs them. This subsection will provide more insight for the reader who is interested to understand the processes involved.

advection As fluids flow through porous media some contaminants in the fluid are transported with it. This transport of these contaminants (mobile *species*) is called advection (at least in our context). While other chemicals are somehow attached to the solid medium and are therefore not transported with the fluid flow (immobile *species*). The speed of the transportation is the same as that of the fluid flow (*Darcy velocity*).

aquifer phases in the subsurface which are porous. The pores are interconnected and are therefore permeable for liquids especially water. Groundwater flows in aquifers. The scope of our model is not limited to aquifers but inspired from them. Aquifers can be permeable rocks, sediments or soil.

bulk density is a constant and property ϱ of the porous medium. It is defined as

$$\varrho = m_p/V$$

where m_p is the mass of the solid part (*soil matrix*) of a porous medium without water in a unit volume V.

Darcy velocity On a microscopical level there are many different velocity vectors of water-flow in a porous medium. Even the velocity in one pore is not homogeneous. Therefore the Darcy velocity (named after Darcy who invented it) is an averaged velocity. It is defined as the volume discharge of the fluid per time (dimension is volume/time) divided through a cross-sectional flat area (dimension is area). It is not a true velocity because the area is partially blocked with solid material (from the porous medium). But the dimension of the Darcy velocity is that of a velocity namely length/time.

diffusion a physical process which takes place in aqueous solutions. On a macroscopical level it describes movement of molecules or ions of one kind to achieve the same concentration in the whole solution. Therefore the movement is from areas of high concentration to areas of low concentration. In the hydrogeological context that we study diffusion is almost always negligible compared to dispersion which takes place too. We therefore omit this process for the sake of simplicity (diffusion is quite complex; its rate is species dependent).

dispersion a physical process which takes place in porous media. As a contaminated fluid (a chemical in solution) flows through a porous medium, it will mix with not contaminated water (water with a smaller concentration of that chemical). The result will be a dilution of the contaminant by a process known as dispersion. The mixing that occurs along the streamline of fluid flow is called longitudinal dispersion. Dispersion that occurs normal to the pathway of fluid flow is transversal dispersion. Dispersion is a mechanical phenomenon. Longitudinal dispersion is caused by fluid flow through pores. Some of the fluid will take longer pathways than other. The flow velocity in the center of a pore is faster than on the edges. Fluid flow is faster in larger pores than in smaller pores. Transversal dispersion is caused by the fact that the flow paths of contaminated fluid will split and branch out to the side. Longitudinal dispersion is usually 10 to 100 times greater than transversal dispersion. Both types of dispersion are proportional to the velocity of the fluid flow.

minerals are naturally occurring solids with a defined chemical composition and a certain physical crystal structure. Examples are calcite and saltpeter. In our work they are only involved in precipitation-dissolution reactions in water, which are *reversible reactions*. An example of a precipitation-dissolution reaction would be the dissolution of salt in water. In general a mineral, which is a pure solid, dissolves in water by decomposing in ions. If there are more ions in the water than the water can hold the solution is supersaturated and the ions will precipitate as minerals through crystallization. But if there are less ions present in the water than the water can hold, more minerals will dissolve as ions. The special thing about these mineral precipitation-dissolution reactions is that the direction of the reaction (dissolution or precipitation) is only determined by the concentration of the ions in the water and not by the amount of unsolved solid. At least as long as there is still solid mineral present. This fact leads to complementary conditions in the mathematical model.

mole is a unit of measurement for the amount of chemical substance. It specifies the amount of pure substance that contains as many elementary particles (atoms, molecules or ions) as there are atoms in 12 g of the isotope carbon-12 (^{12}C). One mole of a chemical substance equates about $6.022 \cdot 10^{23}$ particles of this substance.

pore the void space in a porous medium. It is usually fully or partially filled with a fluid, usually water.

porosity is a constant $n = V_v/V$, where V_v is the void space in a unit volume V of a porous medium. It is the fraction of void volume through total volume.

water content is a constant that we denote with θ. It is a property of a porous medium.

It is defined as

$$\theta = \frac{V_w}{V}$$

where V_w is the volume of water in a unit volume V of the porous medium.

reversible reaction is a chemical reaction with can go in both directions. It aims to reach an equilibrium state concerning the concentrations of the participants. Usually the reaction goes in both directions at the same time even if equilibrium is reached. But the rates of the forward and backward reactions may differ (in equilibrium these rates are the same) depending on the concentrations of the participants in proportion to their concentrations in equilibrium. So the overall direction of the reaction is determined by the difference of forward and backward reaction rate. A simple example is the dissolution of salt in water:

$$NaCl \rightleftharpoons Na^+ + Cl^-$$

If the solution is under-saturated more salt will dissolve. If the solution is supersaturated salt will crystallize. The antidote is an irreversible reaction. We will only consider reversible reactions, which most reactions are.

soil matrix the solid part of a porous medium.

species is a technical term for chemical compounds (like Calcite or Carbondioxid CO_2) or ions in a solution like H^+ or HCO_3^-. In a porous medium we distinguish between mobile species which are dissolved in the fluid (usually water) which is in the medium. They are transported in the fluid-flow and with its *Darcy velocity*. Whereas immobile species are attached to the solid part of the medium (*soil matrix*) through some chemical, electrical or mechanical process e.g. crystallization. They are not transported with the water flow. In this work the only immobile species we consider are minerals.

3.1.3 Physical Background

In this subsection we want to develop very briefly the physical laws and equations, which describe the reactive transport of chemical substances in solution in a porous medium. The interested reader can find more details in [23, sec. 0.3], from which this section is inspired too. There are three basic physical principles, which are important in this context. The basic equation comes from the law of conservation of mass

$$\partial_t (\theta c(t,x)) + \nabla \cdot (q(t,x) \cdot c(t,x)) + \nabla \cdot J^{(1)}(t,x) + \nabla \cdot J^{(2)}(t,x) = \theta f(t,x) \qquad (3.1)$$

where $c(t,x)$ $[mol/m^3]$ is the concentration of a *mobile* species in solution, $q(t,x)$ $[m/s]$ is the mass averaged velocity, the so called Darcy-velocity and $f(t,x)$ is a source or sink term. Before we continue with the physics we have to explain some symbols of vector analysis. The operator $\nabla\cdot$ applied to a vector field $a : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$, $a(x) = [a_1(x), a_2(x), a_3(x)]^T$ yields

$$\nabla \cdot a(x) = \frac{\partial}{\partial x_1} a_1(x) + \frac{\partial}{\partial x_2} a_2(x) + \frac{\partial}{\partial x_3} a_3(x)$$

where x is the spatial variable and the gradient ∇ of a scalar valued function $b : \mathbb{R}^3 \longrightarrow \mathbb{R}$ is the vector

$$\nabla b(x) = \left[\frac{\partial}{\partial x_1} b(x),\ \frac{\partial}{\partial x_2} b(x),\ \frac{\partial}{\partial x_2} b(x)\right]^T$$

and combining this two operators yields

$$\nabla \cdot (\nabla b(x)) = \frac{\partial}{\partial x_1^2} b(x) + \frac{\partial}{\partial x_2^2} b(x) + \frac{\partial}{\partial x_3^2} b(x)\,.$$

This law basically says that the mass of species in solution in a desired domain Ω and time frame $t \in [0, T]$ consists of the mass which flows out or in over the border of the domain $\partial\Omega$ and of the mass which is produced from a source or absorbed from a sink (this is the term $f(t, x)$ on the right hand side). Such an equation is also called mass balance equation. The term $f(t, x)$ $[mol/m^2/s]$ is called volumetric production rate. In our context this will be the reaction rate function of a chemical reaction (see next section). That means that the source is a chemical reaction where the considered species is produced (sign of $f(t, x)$ is positive) and the sink is a reaction where the species is the reactant (sign of $f(t, x)$ is negative). The water content constant $\theta \in (0, 1)$ takes into account that the water volume is only part of the total volume. The function $J^{(1)}(t, x)$ $[mol/m^2/s]$ is called diffusive mass flow and $J^{(2)}(t, x)$ $[mol/m^2/s]$ is called dispersive mass flow. The second term $\nabla\,(q(t, x) \cdot c(t, x))$ on the left hand side describes the forced mass transport through the water flow. This is usually called convection or advection. The third term $\nabla J^{(1)}(t, x)$ describes the molecular diffusion between the mobile species and the water. And the fourth term $\nabla J^{(2)}(t, x)$ on the left hand side describes mass flow through mechanical dispersion. Dispersion is in its effects (not its cause) very similar to diffusion. That is why dispersion and diffusion are both summarized as diffusion. Since diffusion is species dependent and in our context notedly smaller than dispersion, we will leave it out.

The second principle is Fick's law, which describes diffusion. Based on this law the dispersion mass flow $J^{(2)}(t, x)$ $[mol/m^2/s]$ can be written as

$$J^{(2)}(t, x) = -\theta \tilde{D}_{mech}(q)\nabla c(t, x) = -D_{mech}(q)\nabla c(t, x) \tag{3.2}$$

with a symmetric positive definite matrix of mechanical dispersion $\tilde{D}_{mech}$ which depends on the Darcy velocity q. In D_{mech} is the water content θ already incorporated, which is the only difference to $\tilde{D}_{mech}$. According to the Bear-Scheidegger dispersion model (cf. [2, Chapter 10]) we can write the dispersion matrix as

$$D_{mech}(q) = \theta\, \|q\|_2\, (\beta_t I + \frac{(\beta_l - \beta_t)}{\|q\|_2^2} q q^T)$$

where β_l is the longitudinal dispersion length and β_t is the transversal dispersion length.

Inserting (3.2) in (3.1) yields the final diffusion-convection PDE for *mobile* species

$$\partial_t\,(\theta c(t, x)) + \nabla \cdot (q(t, x) \cdot c(t, x) - D_{mech}(q) \cdot \nabla c(t, x)) = \theta f(t, x)\,. \tag{3.3}$$

Species that are attached to the solid part of the porous medium (soil matrix) are not subject to convection, diffusion and dispersion. They are called *immobile* species. The only immobile species we consider are minerals. They are forming crystals and adhere to the soil matrix. For this species, equation (3.3) simplifies to

$$\partial_t \left(\varrho\bar{c}(t,x)\right) = \theta f(t,x)\,, \tag{3.4}$$

where $\bar{c}(t,x)$ $[mol/m^2]$ is the surface concentration of a mineral species and the constant ϱ is called (volumetric) bulk density of the porous medium. It takes into account that the soil matrix is only part of a unit volume (see Subsection 3.1.2).

With additional assumptions, (3.3) can be simplified. If we assume that the water-flow is constant in time and space and parallel to the $x-$axis, i.e. $q(t,x) = [q_0, 0]^T$ then (3.3) simplifies to

$$\partial_t \left(\theta c(t,x)\right) + q_0 \cdot \frac{\partial}{\partial x_1} c(t,x) - \beta_l q_o \frac{\partial^2}{\partial x_1^2} c(t,x) - \beta_t q_0 \frac{\partial^2}{\partial x_2^2} c(t,x) = \theta f(t,x)\,, \tag{3.5}$$

because the dispersion tensor becomes

$$D_{mech}(q) = \begin{bmatrix} \beta_l q_0 & 0 \\ 0 & \beta_t q_0 \end{bmatrix}$$

whereβ_l is the longitudinal dispersion constant andβ_t is the transversal dispersion constant.

The third physical principle involved is Darcy's law. It gives a formula for calculating the already mentioned Darcy velocity q. The velocity of the water in a porous medium is not even the same in one pore. And there are many different velocity vectors due to the structure of the solid part of the porous medium (soil matrix). The Darcy velocity models the effects of pore-scale fluctuations of the flow field in our macro-scale flow model. So the Darcy velocity is not a true velocity but is an averaged velocity. It is really defined as the water discharge $[m^3/s]$ through a cross-sectional area $[m^2]$, which then has the unit of a velocity. This velocity is calculated by Darcy's law

$$q(t,x) = -K(x)\left(\nabla p(t,x) + \rho(t,x) g e_z\right)$$

where p $[N/m^2]$ is the averaged water pressure, the matrix K describes the permeability of the porous medium and the viscosity of water, ρ $[kg/m^3]$ is the density of water, g $[m/s^2]$ is the gravity constant and e_z is the unity vector of the $z-$axis. We do not use Darcy's law directly but we consider the Darcy velocity as given. It can be calculated in advance with Darcy's law.

3.1.4 Chemical Background

As mentioned before we will only consider reversible reactions. And as simplifying assumption we will assume that all reactions are in local equilibrium. And we will not consider sorption reactions. This is quite a restriction but will make the basic structure of the mathematical model clearer.

The *stoichiometric matrix* $(s_{i,j}) = S \in \mathbb{R}^{(I+\bar{I})\times J}$ is the matrix of *stoichiometric coefficients* of I *non-mineral* and $\bar{I}$ *mineral* species in J chemical reactions. We always assume that there are more species than reactions, i.e. $I \geq J$. If we have for example, an equilibrium reaction

$$2X_2 + 1X_7 \longleftrightarrow 1X_4 + 3X_6$$

we shift as a convention all species to the right side

$$0 \longleftrightarrow 0X_1 - 2X_2 + 0X_3 + 1X_4 + 0X_5 + 3X_6 - 1X_7$$

and get a column of matrix S with entries $0, -2, 0, 1, 0, 3, -1$ in the rows that correspond to the species. Entries in this column are zero for species which do not participate in this reaction. The stoichiometric coefficient of a reactant is always negative and of a product always positive. The stoichiometric coefficients are mostly (but not always) integers. It is well known that any linear dependence of the chemical reactions (i.e. the columns of S) indicates a redundancy of chemical reactions [1]. Therefore it is permissible to assume that the columns of S are linear independent.

The *law of mass action* is a mathematical model that explains and predicts the behavior of solutions. It states that reversible reactions in solutions always strive to reach an equilibrium state, i.e. a certain proportion of the concentrations of reactant and product. It predicts the reaction rate (speed). This law comes in a *kinetic* and an *equilibrium* version. For more information on the law of mass action see [33].

The kinetic version describes the total reaction rate (speed) and the overall direction of the reaction. It reads

$$R_j(a) = k_j^f \prod_{\substack{s_{i,j}<0 \\ i=1,\ldots,I+\bar{I}}} a_i^{-s_{i,j}} - k_j^b \prod_{\substack{s_{i,j}>0 \\ i=1,\ldots,I+\bar{I}}} a_i^{s_{i,j}}, \qquad j = 1, \ldots, J$$

with forward and backward rate constant $k_j^f > 0, k_j^b > 0$ and $a = [a_1, \ldots, a_{I+\bar{I}}]$ an activity vector with a_i the *activity* of the i-th species. The activity of a chemical species is closely related to its concentration and is sometimes called effective concentration of a species. For the example above, here denoted as reaction j, this equation reads

$$R_j(a) = k_j^f a_2^2 a_7 - k_j^b a_4 a_6^3 .$$

If $R_j(a)$ is positive, the reaction will go (mainly) in forward direction, i.e.

$$2X_2 + 1X_7 \longrightarrow 1X_4 + 3X_6,$$

and if $R_j(a)$ is negative the reaction will go (mainly) in backward direction. Dynamic equilibrium is reached if $R_j(a) = 0$ (forward and backward reactions are still going on but at equal rate). For non-minerals, usually the concentration is a good approximation, i.e. $a_i \approx c_i$, where c_i is the concentration of non-mineral species i. But for minerals usually constant activity is assumed, i.e. $a_i = 1$ (the actual constant is incorporated into the rate constant). If X_7 was a mineral and X_2, X_4, X_6 were non-minerals the rate function would read

$$R_j(c, \bar{c}) = k_j^f c_2^2 - k_j^b c_4 c_6^3$$

with $c = [c_1, \dots, c_I]$ and $\bar{c} = [\bar{c}_{I+1}, \dots, \bar{c}_{I+\bar{I}}]$ the vectors of the concentrations of the non-mineral and mineral species (with a possibly changed forward rate constant k_j^f). This rate function will later in the mathematical model appear as source/sink term depending on its sign. The equilibrium version of the mass action law reads

$$\prod_{\substack{s_{i,j}<0 \\ i=1,\dots,I+\bar{I}}} a_i^{s_{i,j}} \cdot \prod_{\substack{s_{i,j}>0 \\ i=1,\dots,I+\bar{I}}} a_i^{s_{i,j}} = k_j, \qquad {\scriptstyle j=1,\dots,J} \tag{3.6}$$

where a_i here denotes the activity of species i when the reaction is in dynamic equilibrium and $k_j = k_j^f / k_j^b > 0$ is called equilibrium constant. Like before the activities in equilibrium can be replaced by the concentrations in equilibrium for non-minerals or by 1 for minerals. For our example reaction this equation reads

$$\frac{c_4 \cdot c_6^3}{c_2^2} = k_j \, .$$

In this thesis minerals shall only be involved in precipitation-dissolution reactions where exactly one mineral dissolves into ions. We call these mineral reactions. That means that we have exactly $\bar{I}$ minerals and mineral reactions. For such reactions is the equilibrium equation or the corresponding kinetic formulation not sufficient to describe all states of equilibrium. Lets consider the above example again, where X_7 is a mineral. Then $R_j(c) = 0$ and $\frac{c_4 \cdot c_6^3}{c_2^2} = k_j$ respectively is only true, if there is still solid mineral present i.e. $\bar{c}_7 > 0$. Since minerals are solids attached to the soil matrix they can completely diminish in a part of the porous medium. In this case, i.e. $\bar{c}_7 = 0$, a reasonable model would be to demand $R_j(c) = k_j^f c_2^2 - k_j^b c_4 c_6^3 \geq 0$. Since mineral X_7 is not present any more the forward reaction $2X_2 + 1X_7 \longrightarrow 1X_4 + 3X_6$ can not go on anymore (the backward reaction is still possible). Therefore the concentrations c_4 and c_6 are (possibly) smaller than in dynamic equilibrium. Conversely the concentration c_2 of the reactant X_2 is (possibly) greater than in dynamic equilibrium. So the case $\bar{c}_7 = 0$ yields another (forced) equilibrium state. The complete model of equilibrium for mineral reactions would be

$$\left(\bar{c}_7 = 0 \ \wedge \ k_j^f c_2^2 - k_j^b c_4 c_6^3 \geq 0\right) \vee \left(\bar{c}_7 \geq 0 \ \wedge \ k_j^f c_2^2 - k_j^b c_4 c_6^3 = 0\right) .$$

This can be written as the *complementary condition*

$$\bar{c}_7 \cdot \left(k_j^f c_2^2 - k_j^b c_4 c_6^3\right) = 0 \, , \bar{c}_7 \geq 0, \; k_j^f c_2^2 - k_j^b c_4 c_6^3 \geq 0,$$

which can not be expressed well with a smooth function (see Section 2.4). This is the particular difficulty when minerals are involved. For mineral precipitation-dissolution reactions the equilibrium constant $k_j = k_j^f / k_j^b$ is also called solubility product.

In real-word problems, the timescales for the different reactions can vary over many powers of ten. Hence it is frequently assumed that some of the reactions are so fast that a state of dynamic equilibrium can be assumed always and everywhere. Even if this equilibrium is distorted by water flow and dispersion these reactions are so fast that they can balance this distortion almost immediately (compared to the speed of the water flow). In our model we will only consider such fast reactions, because this allows us to focus on the main structure of the problem.

There are several equivalent formulations of the equilibrium equation (3.6), which may affect the structure of our problem. For a reaction j where no minerals are involved (non-mineral reaction) a favorable reformulation (assuming all concentrations are positive) would be

$$Q_j(c) := \sum_{i=1}^{I} s_{i,j} \cdot \ln c_i - \ln k_j = 0, \tag{3.7}$$

with $k_j = k_j^f / k_j^b$. Now let us consider the mineral example above. The condition $k_j^f c_2^2 - k_j^b c_4 c_6^3 \geq 0$ can be equivalently (provided all mobile concentrations are positive) written as

$$\ln k_j - \ln\left(\frac{c_4 c_6^3}{c_2^2}\right) \geq 0,$$

since the logarithm is monotone. Using the calculation rules for the logarithm we can write the equilibrium complementary condition for mineral reaction j as

$$\bar{c}_{I+j} \cdot E_j(c) = 0,\ \bar{c}_{I+j} \geq 0,\ E_j(c) \geq 0 \tag{3.8}$$

with

$$E_j(c) = \ln k_j - \left[s_{1,j}, \ldots, s_{I,j}\right] \cdot \ln c \qquad \text{and} \qquad (\ln c)^T = [\ln(c_1), \ldots, \ln(c_I)]^T\,. \tag{3.9}$$

The case $E_j(c) = 0, \bar{c}_j \geq 0$ corresponds to a saturation of the water with respect to this mineral reaction, and the case $E_j(c) \geq 0, \bar{c}_j = 0$ corresponds to the total dissolution of the mineral and an under-saturation of the water.

3.2 Mathematical Modeling

This section gives a precise formulation of the mathematical model for the application that was outlined in the introduction. This formulation will be the basis for our subsequent theoretical and numerical investigations. The reader who is less interested in the derivation of the model may skip most of this section and continue to read at formulas (3.14)–(3.19). In fact, sketching at least some of the notation introduced in this section and Section 3.3, the reader may alternatively continue reading with formulas (3.31)–(3.33) if he wants to skip also the model reduction and wants to concentrate on the mathematical convergence analysis. The notation introduced in this and the next section will be valid and used throughout the whole thesis.

To this end, let us consider the vector $c = [c_1, c_2, \ldots, c_I]^T$ of concentrations of I species dissolved in the groundwater (mobile species). These concentrations are time- and space-dependent. The transport of a species consists of forced advection by the given flow field q of the groundwater (Darcy velocity) and dispersion. The advection-dispersion operator for these species is given by

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

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

With $\bar{c} = [\bar{c}_{I+1}, \dots, \bar{c}_{I+\bar{I}}]^T$ we denote the concentrations of the $\bar{I}$ mineral species. These concentrations are also variable in time and space. They are attached to the soil matrix and therefore neither subject to advection nor dispersion. But they are involved in precipitation-dissolution reactions, where they dissolve into mobile species (ions). In this paper we restrict ourselves to equilibrium reactions, i.e. reactions that are actually in the condition of equilibrium or equations which are sufficiently fast to be approximately considered to be in equilibrium. $R = [R_1, \dots, R_J]$ denotes the vector of reaction rates that are necessary to keep the chemical system in equilibrium, because the chemical equilibrium is constantly distorted through the advection. Together with c and $\bar{c}$ they form the unknowns of the system to be considered here.

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

$$\frac{\partial}{\partial t}\theta c + Lc = \theta S_1 R, \tag{3.10}$$

$$\frac{\partial}{\partial t}\bar{c} = \theta S_2 R, \tag{3.11}$$

given on the domain $[0, T] \times \Omega \subset \mathbb{R}^n$ with $n = 2$ or $n = 3$ together with given initial and boundary and equilibrium conditions. The constant $\theta \in (0, 1)$ denotes water content and the matrix $\left(s_{ij}\right) = S = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \in \mathbb{R}^{(I+\bar{I})\times J}$ is the matrix of stoichiometric coefficients, where J is the number of chemical reactions. Without loss of generality (see Subsection 3.1.4), we can assume that S has full column rank,

$$\operatorname{rank}(S) = J. \tag{3.12}$$

Remember here that we have at least as many species $I + \bar{I}$ as we have reactions J. Additionally, we demand that the columns of S_1 are linearly independent

$$\operatorname{rank}(S_1) = J. \tag{3.13}$$

Because we are only considering precipitation-dissolution reactions for minerals, we assume that each mineral is participating in one and only one mineral reaction (in these reactions are usually two or more mobile ions involved). With mobile reactions we indicate reactions in which only mobile species participate. By J_{mob} we denote the number of mobile reactions and with J_{min} the number of mineral reactions. It follows that $J_{min} = \bar{I}$. Since in our model we have only mineral or mobile reactions, it holds $J = J_{mob} + J_{min}$. The stoichiometric matrix then reads

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

where, for simplicity of notation, we have replaced the diagonal matrix representing the mineral participation in the mineral reactions by $-I$, the negative identity matrix. The submatrix S^1_{min} is adjusted accordingly. Therefore reactions $1, \dots, J_{mob}$ are mobile and reactions $J_{mob} + 1, \dots, J$ are mineral, because the columns of S refer to chemical reactions.

In the following we assume that all mobile concentrations c_i $(i = 1, \dots, I)$ are positive. Negative concentrations make no sense but it is a priori not guaranteed that they stay positive in numerical computations.

Building on (3.7) we can write the equilibrium equations for the mobile species in matrix notation as

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

where $K_1 = (\ln k_1, \dots, \ln k_{J_{mob}})^T$ is the vector of equilibrium constants in logarithmic form and

$$\ln c = [\ln c_1, \dots, \ln c_I]^T$$

is a vector of the logarithmic concentrations. These equations hold in every point of space and time (after homogenization).

Building on (3.8-3.9) we write the mineral equilibrium complementary conditions as

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

where

$$E(c) = K_2 - \left(S^1_{min}\right)^T \ln c , \quad E = [E_{J_{mob}+1}, \dots, E_J]^T .$$

Hereby

$$K_2 = [\ln k_{J_{mob}+1}, \dots, \ln k_J]^T$$

is the vector of solubility products in logarithmic notation.

We decompose the reaction vector R into

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

with R_{mob} and R_{min} being of size J_{mob} and J_{min}, respectively. The reaction rate R_j for reaction j in dynamic equilibrium should be zero. But in our model water flow is involved, which changes the concentrations of reactants and/or products in a fixed point. With this the reaction rate R_j is changed, too. Therefore we treat these reaction rates as unknowns. They reflect the reaction rates that are necessary to maintain the chemical equilibrium.

Utilizing the structure of S, the full system reads

$$\begin{aligned}
\frac{\partial}{\partial t}\theta c + Lc &= \theta S^1_{mob} R_{mob} + \theta S^1_{min} R_{min} = \theta S_1 R , && (3.14)\\
\frac{\partial}{\partial t}\bar{c} &= -\theta R_{min} , && (3.15)\\
E_j(c) \cdot \bar{c}_j &= 0 \; (j = J_{mob} + 1, \dots, J) , && (3.16)\\
\bar{c}_j &\geq 0 \; (j = J_{mob} + 1, \dots, J) , && (3.17)\\
E_j(c) &\geq 0 \; (j = J_{mob} + 1, \dots, J) , && (3.18)\\
Q_{mob}(c) &= 0 , && (3.19)
\end{aligned}$$

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

3.3 Transformation of the Dynamic System

The aim of this section is to reduce the size of the overall system (3.14)–(3.19) by using suitable decouplings and reformulations. Since these techniques are already known from [27, 28] (but strictly needed for our subsequent analysis), we will keep this section as short as possible.

First, we apply the decoupling technique proposed in [27, 28] to the PDE-ODE system (3.14)–(3.15). This will lead to a decoupling of some linear PDEs. The remaining PDE-system will then be significantly smaller than the original PDE-system. To this end, we define $S_1^\perp$ as a matrix consisting of a maximum set of linearly independent columns that are orthogonal to each column of S_1, i.e. $(S_1)^T S_1^\perp = 0$. Recall that the columns of S_1 were assumed to be linearly independent, cf. (3.13). Hence the pseudo-inverses of S_1 and $S_1^\perp$ are given by $\left(S_1^T S_1\right)^{-1} S_1^T$ and $(\left(S_1^\perp\right)^T S_1^\perp)^{-1} \left(S_1^\perp\right)^T$, respectively. Multiplying (3.14) with these two pseudo-inverses, we obtain

$$\begin{aligned} \left((S_1^\perp)^T S_1^\perp\right)^{-1} (S_1^\perp)^T \left(\frac{\partial}{\partial t}\theta c + Lc\right) &= 0\,, && (3.20)\\ \left(S_1^T S_1\right)^{-1} S_1^T \left(\frac{\partial}{\partial t}\theta c + Lc\right) &= \theta R\,, && (3.21)\\ \frac{\partial}{\partial t}\bar{c} &= -\theta R_{min}\,. && (3.22) \end{aligned}$$

We now substitute

$$\eta := \left(\left(S_1^\perp\right)^T S_1^\perp\right)^{-1} \left(S_1^\perp\right)^T c, \quad \xi := \left(S_1^T S_1\right)^{-1} S_1^T c, \tag{3.23}$$

and partition the vector ξ into

$$\xi = [\xi_{mob}, \xi_{min}]$$

of size J_{mob}, J_{min}. Then splitting equation (3.21) into two parts and adding the third block to the second part, we get

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

We have obtained a decoupling of the computation of the concentrations and of the rates, i.e., we may drop now the second and the fourth block of equations and solve the system consisting of the first and the third block and the equilibrium conditions (3.16)–(3.19) for the concentrations. The rates can be computed a posteriori, if desired.

Now we reformulate the complementary conditions with the NCP-function φ. For now φ stands for an arbitrary function of this class. In Chapter 4 we will use the minimum function φ_M and in Chapter 5 the Fischer-Burmeister function φ_F.

Using this NCP-function, we can write the complementarity conditions (3.16)–(3.18) as

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

In vector notation, this becomes

$$\varphi\left(E(c), \bar{c}\right) = 0 \, , \tag{3.24}$$

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

The resulting system now reads

$$\begin{aligned} \frac{\partial}{\partial t}\theta\eta + L\eta &= 0 \, , && (3.25) \\ \frac{\partial}{\partial t}\left(\theta\xi_{min} + \bar{c}\right) + L\xi_{min} &= 0 \, , && (3.26) \\ -\varphi\left(E(c), \bar{c}\right) &= 0 \, , && (3.27) \\ Q_{mob}(c) &= 0 \, , && (3.28) \end{aligned}$$

where c can be represented as

$$c = c\left(\xi_{min}, \xi_{mob}, \eta\right) = S^{1}_{min} \cdot \xi_{min} + S^{1}_{mob} \cdot \xi_{mob} + S^{\perp}_{1}\eta, \tag{3.29}$$

cf. (3.23). Note that (3.25) is now linear with respect to η and it is decoupled from the other equations ($\xi_{min}, \xi_{mob}, \bar{c}$ are not contained in (3.25)). The remaining non-linearly coupled system (3.26)–(3.28) is reduced in size from $I + J + J_{min}$ rows to $I + J_{min}$ rows compared to the original system (3.14)–(3.19). Together with the size reduction of J rows, the J unknowns R could be dropped. They can be computed a posteriori.

3.4 Discretization of the Dynamic System

In this section we perform the discretization of (3.25)–(3.28) in space and time. To keep the notation simple, we suppress subscripts indicating the discretization (except we denote L_h as the discretization of L). For the sake of simplicity, we assume the implicit Euler time stepping scheme.

We start with the discretization of the decoupled equation (3.25), which reads

$$\left(\theta I + \tau L_h\right) \cdot \eta = \theta\eta^{old} \, , \tag{3.30}$$

where η^{old} denotes the solution vector of the previous time step. This linear system can be solved for η directly (say, by a linear system solver like GMRES). Hence η is not viewed as a variable any longer, because it can be computed a priori in every time step. We therefore write $c = c\left(\xi_{min}, \xi_{mob}\right)$ for the discretized function c.

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

$$\begin{aligned} G_1(\xi_{min}, \xi_{mob}, \bar{c}) &:= \quad \theta\xi_{min} + \bar{c} + \tau L_h \xi_{min} - \theta\xi_{min}^{old} - \bar{c}^{old} &= 0\,, \qquad (3.31)\\ G_2(\xi_{min}, \xi_{mob}, \bar{c}) &:= \quad -\varphi\left(E\left(c\left(\xi_{min}, \xi_{mob}\right)\right), \bar{c}\right) &= 0\,, \qquad (3.32)\\ G_3(\xi_{min}, \xi_{mob}, \bar{c}) &:= \quad Q_{mob}\left(c\left(\xi_{min}, \xi_{mob}\right)\right) &= 0\,. \qquad (3.33) \end{aligned}$$

The superscript 'old' indicates the previous time-step. The time-step size is τ. We assume the domain Ω has been discretized into the grid set Ω_h with $p = |\Omega_h|$ grid points. Then $\xi_{min}, \xi_{mob}, \bar{c}$ are vectors with $J_{min} \cdot p$, $J_{mob} \cdot p$, $J_{min} \cdot p$ components. These vectors are concatenations of the function values in every node of the grid. L_h is a linear mapping which is the discretization of the PDE operator L. In (3.32) and (3.33), the functions Q_{mob}, φ, E, c are to be applied to (the discretizations of) $\xi_{min}, \xi_{mob}, \bar{c}$ in every node separately. For example, a more detailed way to represent $c\left(\xi_{min}, \xi_{mob}\right)$ is

$$c\left(\xi_{min}, \xi_{mob}\right) = \left[c(\xi_{min}(x^1), \xi_{mob}(x^1))^T, \dots, c(\xi_{min}(x^p), \xi_{mob}(x^p))^T\right]^T\,, \qquad (3.34)$$

where $\xi_{min}(x^i), \xi_{mob}(x^i)$ are our variables in one grid point $x^i \in \Omega_h$. And for example for $Q_{mob}(c)$ this means

$$Q_{mob}\left(c(\xi_{min}, \xi_{mob}, \bar{c})\right) = \left[Q_{mob}\left(c(\xi^1_{min}, \xi^1_{mob})\right)^T, \dots, Q_{mob}\left(c(\xi^p_{min}, \xi^p_{mob})\right)^T\right]$$

and for the other functions E and φ likewise. Often we will also use the following notation: We enumerate the set of grid points as $\Omega_h = \{x_1, x_2, \dots\}$ and then we write $\xi_{min}(x_i) = \xi^i_{min}$ etc.

For the sake of simplicity, we define the abbreviations

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

And we define the open and convex set (c in (3.34) is linear)

$$\mathcal{D} := \left\{[\xi_{min}, \xi_{mob}, \bar{c}] \in \mathbb{R}^{J_{min}p} \times \mathbb{R}^{J_{mob}p} \times \mathbb{R}^{J_{min}p} \mid c\left(\xi_{min}, \xi_{mob}\right) > 0\right\} \qquad (3.35)$$

on which

$$G\left(\xi_{min}, \xi_{mob}, \bar{c}\right) = \left[\begin{array}{c} G_1\left(\xi_{min}, \xi_{mob}, \bar{c}\right)\\ G_2\left(\xi_{min}, \xi_{mob}, \bar{c}\right)\\ G_3\left(\xi_{min}, \xi_{mob}, \bar{c}\right) \end{array}\right] \qquad (3.36)$$

is defined, i.e.

$$G : \mathcal{D} \longrightarrow \mathbb{R}^{(2J_{min}+J_{mob})p}\,.$$

Then we have to find a solution in $\mathcal{D}$ of the nonlinear system of equations

$$G\left(\xi_{min}, \xi_{mob}, \bar{c}\right) = 0\,. \qquad (3.37)$$

From now on we will only deal with the discretized system. All variables and functions shall from now on be the discretized versions, i.e. long column vectors.

4 The Minimum Function Approach

In this chapter we apply the minimum function as NCP-function to the system (3.31)–(3.33), i.e.

$$G_2(\xi_{min}, \xi_{mob}, \bar{c}) := -\varphi_M\left(\tilde{E}(\xi_{min}, \xi_{mob}), \bar{c}\right). \tag{4.1}$$

Unless mentioned otherwise G_2 shall be defined with the minimum function in this chapter. The resulting total function is then denoted as

$$G_M : \mathcal{D} \longrightarrow \mathbb{R}^{(2J_{min}+J_{mob})p}.$$

We continue to use the notation from the last chapter with the same meaning, in particular $\mathcal{D}, \Omega_h, p = |\Omega_h|, \varphi_F, \varphi_M, J_{min}, J_{mob}, J$ and I . Our aim is to prove local convergence for the semismooth Newton method applied to G_M and to show the local existence and uniqueness of a solution of

$$G_M(w) = 0, \; w \in \mathcal{D}. \tag{4.2}$$

In the first section we study the B-subdifferential and generalized Jacobian of G_M, especially the structure of $\partial_B G_M$. In Section 4.2 we apply the semismooth Newton method to solve the nonlinear equation system (4.2). We show how the resulting linear systems can be simplified and solved efficiently. In the following Section 4.3 we introduce problem specific generalizations of $\partial_B G_M$ and ∂G_M. And we show some properties of their elements. The next section deals with the nonsingularity of the B-subdifferential of G_M, which is essential for the execution of the semismooth Newton method. Also the local convergence result for this algorithm is stated. In Section 4.5 we present a new and improved proof for the nonsingularity of the elements of $\partial_B G_M$. This proof contains a couple of advantages, in particular it enables further results. Section 4.6 introduces a method to solve the linear equation systems arising from the semismooth Newton method in a more efficient way. In the next section we show local existence and uniqueness of a solution of (5.2). And finally in the last section we bring a numerical example.

4.1 Study of Subdifferentials of G_M

In this section we study the properties of G_M related to subdifferentials and the subdifferentials themselves.

We start by showing that G_M is a strongly semismooth function on its domain $\mathcal{D}$. The first block component function G_1 of G_M defined in (3.31) is a linear function. It is therefore a C^2-function and thanks to Corollary 2.3.4 it is strongly semismooth. The last block component function G_3 equals

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

where c is a linear transformation of ξ_{min} and ξ_{mob} (and of η which is a constant in this context). Clearly G_3 is a C^2-function on its domain $\{(\xi_{min}, \xi_{mob}) \in \mathbb{R}^{(J_{min}+J_{mob})p} \mid c\,(\xi_{min}, \xi_{mob}) > 0\}$. Again with Corollary 2.3.4 we get that G_3 is strongly semismooth on its domain. Now we deal with the second block component function G_2 of G_M defined in (4.1). The inner function $\tilde{E}$ in G_2 is

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

and because its obvious similarity to $\tilde{Q}_{mob}$ it, too, is strongly semismooth. Thanks to Lemma 2.3.2 the whole inner function

$$(\xi_{min}, \xi_{mob}, \bar{c}) \mapsto \left[\tilde{E}\,(\xi_{min}, \xi_{mob}), \bar{c}\right]$$

of G_2 is strongly semismooth. From Example 2.3.6 we know that φ_M is strongly semismooth. With the chain rule Theorem 2.3.7 we conclude that G_2 is strongly semismooth.

Together with Lemma 2.3.2 we have proven

Lemma 4.1.1. *The function $G_M : \mathcal{D} \longrightarrow \mathbb{R}^{(2J_{min}+J_{mob})p}$ is strongly semismooth.*

By definition (at least as we defined it) every semismooth function is also locally Lipschitz continuous. So our function G_M is locally Lipschitz continuous. This justifies the use of the terms B-subdifferential, generalized Jacobian and C-subdifferential as we have defined them in Chapter 2 with respect to G_M. The next result gives important information about the structure of $\partial_B G_M(w)$.

Lemma 4.1.2. *Let $\Omega_h = \left\{x_1, x_2, \ldots, x_p\right\}$. Furthermore let $w = [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$ be arbitrary. Then the following statements hold:*

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

$$\partial_B G_M\,(w) = \partial_B G_1\,(w) \times \partial_B G_2\,(w) \times \partial_B G_3\,(w)$$

with $\partial_B G_1\,(w) = \left\{G_1'(w)\right\}$ and $\partial_B G_3\,(w) = \left\{G_3'(w)\right\}$.

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

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

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

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

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

(4) Let a and b be as before. Then for $j = 1, \ldots, \bar{I}$ we have

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

where $[0 \mid 0 \mid e_l^T] = \frac{\partial\bar{c}_j(x_i)}{\partial w}$ and e_l is a unit vector, with all components vanishing and component $l = i \cdot J_{min} + j$ being one.

Proof. We have already seen in Lemma 4.1.1 that G_M is strongly semismooth. It is therefore locally Lipschitz continuous.

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

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

(4) The two cases $\tilde{E}_j(a) > b_j$ and $\tilde{E}_j(a) < b_j$ are obvious since φ_M is continuously differentiable in these cases, so that the B-subdifferential reduces to the existing gradient which can be calculated directly from (3.32). The remaining case $\tilde{E}_j(a) = b_j$ can be verified by choosing suitable sequences $\{b^k\}$ converging to b. □

Note the fact that G_1 and G_3 are continuously differentiable means that their B-subdifferential equals the cross product of the B-subdifferentials of their components. Therefore we can immediately deduce the following Corollary.

Corollary 4.1.3. *The B-subdifferential of G_M is a cross product of the B-subdifferentials of its scalar components, i.e. with $w = [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$ we have*

$$\partial_B G_M(w) = \partial_B G^1(w) \times \partial_B G^2(w) \times \ldots \times \partial_B G^n(w)$$

where $n = (2J_{min} + J_{mob}) \cdot p$ and G^i is a scalar component function of G_M spanning over all functions G_1, G_2 and G_3.

This result is very useful because it simplifies the calculation of $\partial_B G_M$ very much. This is especially significant since G_M comes from discretization and can therefore have a large number of components. The interesting components of course are those where φ_M is involved. The B-subdifferential of these scalar component functions can have two elements if evaluated in a non-differentiable point (see (4) in the Lemma above).

We will now show, how to construct subsets $\mathcal{A}$ of $\{1, \ldots, J_{min}\} \times \Omega_h$ that can be used to identify an element of $\partial_B G_M$ in a unique way. This construction is based on Lemma 4.1.2(4). Let $w = [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$ be arbitrary and let H be an arbitrary element of $\partial_B G_M(w)$. Then a pair $(i, x) \in \{1, \ldots, J_{min}\} \times \Omega_h$ shall be an element of $\mathcal{A}$ if the row of H corresponding to the component function $-\varphi_M\left(\tilde{E}_i(\xi_{min}(x), \xi_{mob}(x)), \bar{c}_i(x)\right)$ in $G_2(w)$ equals $-\frac{\partial \bar{c}_i(x)}{\partial w}$. This is mandatory if $\tilde{E}_i(\xi_{min}(x), \xi_{mob}(x)) > \bar{c}_i(x)$ and possible if $\tilde{E}_i(\xi_{min}(x), \xi_{mob}(x)) = \bar{c}_i(x)$. Using the index sets

$$\mathcal{P} = \left\{(i, x) \in \{1, \ldots, J_{min}\} \times \Omega_h \mid \tilde{E}_i(\xi_{min}(x), \xi_{mob}(x)) > \bar{c}_i(x)\right\} \quad (4.3)$$

and

$$\mathcal{Q} = \Big\{(i,x) \in \{1,\dots,J_{min}\} \times \Omega_h \mid \tilde{E}_i(\xi_{min}(x), \xi_{mob}(x)) = \bar{c}_i(x)\Big\} \tag{4.4}$$

we always have

$$\mathcal{P} \subseteq \mathcal{A} \subseteq (\mathcal{P} \cup \mathcal{Q}) \ . \tag{4.5}$$

And conversely every set $\mathcal{A}$ with $\mathcal{P} \subseteq \mathcal{A} \subseteq (\mathcal{P} \cup \mathcal{Q})$ determines one and only one element $H \in \partial_B G_M(w)$. We call this set $\mathcal{A}$ active set (or set of active indices) for a reason which will become clear later. It is dependent on the point of evaluation w and we denote its complement with $\mathcal{I} = (\{1,\dots,J_{min}\} \times \Omega_h) \setminus \mathcal{A}$ and call it inactive set (or set of inactive indices). Finally we denote with $H^{\mathcal{A}}$ the element of $\partial_B G_M(w)$ that is determined by $\mathcal{A}$.

Before we can continue we need a small technical Lemma.

Lemma 4.1.4. *Let V be a real vector space and let $M = M_1 \times M_2 \times \dots \times M_n \subset V$. Then*

$$co\,(M) = co\,(M_1) \times co\,(M_2) \times \dots \times co\,(M_n)$$

where "co" denotes the convex hull.

Proof. The inclusion $\mathrm{co}\,(M) \subset \mathrm{co}\,(M_1) \times \mathrm{co}\,(M_2) \times \dots \times \mathrm{co}\,(M_n)$ is clear. We show the other inclusion for $n = 2$. Then the general case follows with a simple induction.

Let $v = [v_1, v_2] \in \mathrm{co}\,(M_1) \times \mathrm{co}\,(M_2)$. For $i = 1,2$ there are integers n_i and real numbers $\alpha_{i,j} \geq 0$ $(j = 1,\dots,n_i)$ such that $\sum_{j=1}^{n_i} \alpha_{i,j} = 1$ and there are n_i elements $m_{i,j} \in M_i$ $(j = 1,\dots,n_i)$ such that

$$\sum_{j=1}^{n_i} \alpha_{i,j} m_{i,j} = v_i$$

holds. Then we have

$$\begin{aligned}
[v_1, v_2] &= \left[\sum_{j=1}^{n_1} \alpha_{1,j} m_{1,j},\ \sum_{k=1}^{n_2} \alpha_{2,k} m_{2,k}\right] \\
&= \left[\sum_{j=1}^{n_1} (\sum_{k=1}^{n_2} \alpha_{2,k}\alpha_{1,j} m_{1,j}),\ \sum_{k=1}^{n_2} (\sum_{j=1}^{n_1} \alpha_{1,j}\alpha_{2,k} m_{2,k})\right] \\
&= \left[\sum_{j=1}^{n_1} \sum_{k=1}^{n_2} \alpha_{1,j}\alpha_{2,k} m_{1,j},\ \sum_{j=1}^{n_1} \sum_{k=1}^{n_2} \alpha_{1,j}\alpha_{2,k} m_{2,k}\right] \\
&= \sum_{j=1}^{n_1} \sum_{k=1}^{n_2} \alpha_{1,j}\alpha_{2,k} \cdot \left[m_{1,j},\ m_{2,k}\right] .
\end{aligned}$$

The vectors $\left[m_{1,j},\ m_{2,k}\right]$ are elements of $M_1 \times M_2$. It holds $\sum_{j=1}^{n_1} \sum_{k=1}^{n_2} \alpha_{1,j}\alpha_{2,k} = 1$ and $\alpha_{1,j}\alpha_{2,k} \geq 0$ for all j, k. So $[v_1, v_2]$ can be written as convex combination of elements in $M_1 \times M_2$, i.e. $[v_1, v_2] \in \mathrm{co}\,(M_1 \times M_2)$. □

As an immediate consequence of this Lemma and of Corollary 4.1.3 we get the last result in this section, which greatly simplifies the calculation of the generalized Jacobian of G_M.

Corollary 4.1.5. *Let $w = (\xi_{min}, \xi_{mob}, \bar{c}) \in \mathcal{D}$ be arbitrary. Then*

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

4.2 Newton's Method and Active Set Strategy

In this section we apply the semismooth Newton method defined in Section 2.5 to solve the equation

$$G_M(w) = 0 \tag{4.6}$$

for $w \in \mathcal{D}$. This equation is of course equivalent to (3.31)–(3.33). We will also see the relation of this method to an active-set strategy and the justification of the name. Some parts of this section are inspired from the Habilitation Thesis [25], whereas the relationship between our semismooth Newton method and an active set strategy is, in principle, known [21, 16], although it has not been discussed within our context. The formulas to be derived in this section will, in particular, be needed in the subsequent sections.

The linearization of (4.6) via Newton's method leads to the linear system

$$H^{\mathcal{A}} \begin{bmatrix} \Delta\xi_{min} \\ \Delta\xi_{mob} \\ \Delta\bar{c} \end{bmatrix} = - \begin{bmatrix} G_1(w) \\ G_2(w) \\ G_3(w) \end{bmatrix}, \tag{4.7}$$

with $H^{\mathcal{A}} \in \partial_B G_M(w)$, $\mathcal{P} \subseteq \mathcal{A} \subseteq (\mathcal{P} \cup \mathcal{Q})$ and $w = [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$ as defined in Section 4.1. Such linear systems must be solved in every iteration step of Newton's method. Note that the choice of $H^{\mathcal{A}}$ is arbitrary in every step. And note that $\mathcal{P}$ and $\mathcal{Q}$ can (and probably will) change in every step and $\mathcal{A}$ must then be chosen accordingly. We note the resulting algorithm.

Algorithm 4.2.1 (Semismooth Newton Method)**.**

(S.0) (Initialization)
Choose $w^0 = [\xi^0_{min}, \xi^0_{mob}, \bar{c}^0] \in \mathcal{D}$, $\epsilon \geq 0$ *and set* $k := 0$.

(S.1) (Termination Criterion)
If $\left\| G_M\left(w^k\right) \right\|_\infty \leq \epsilon$ *or* $w^k \notin \mathcal{D}$, *stop.*

(S.2) (Newton Direction Calculation)
Choose $H^{\mathcal{A}} \in \partial_B G_M(w^k)$. *Find a solution* d^k *of the linear system*

$$H^{\mathcal{A}} d = -G_M(w^k)\,.$$

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

Note that this algorithm is applied to a restricted problem, because G_M is only defined on $\mathcal{D}$. So if an iterate w^k is not in $\mathcal{D}$ the algorithm can't proceed and terminates without a solution. But if we start sufficiently close to a solution in $\mathcal{D}$ then we have no problems with infeasible points. Anyhow Newton's method is only convergent locally. Also it is important to take the infinity norm in the termination criterion of this algorithm. The 1-norm or Euclidean norm take sums which involve all components of a vector. So the larger the vector the more restrictive is a termination criterion with these norms.

For the following analysis we choose an arbitrary but fixed element $H^{\mathcal{A}} \in \partial_B G_M(w)$ which we denote with J. We now want to exploit the special structure of J in order to decompose the linear system (4.7). To this end, we reorder the entries of ξ_{min} and $\bar{c}$ in the following way

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

We apply the same reordering to our component functions G_1 and G_2. Altogether, this corresponds to reordering the rows and columns of J. We perform the following decompositions:

$$G_1 = \begin{bmatrix} G_1^{\mathcal{A}} \\ G_1^{\mathcal{I}} \end{bmatrix}, L_h = \begin{bmatrix} L_h^{\mathcal{A}} \\ L_h^{\mathcal{I}} \end{bmatrix}, \; \tilde{E} = \begin{bmatrix} \tilde{E}_{\mathcal{A}} \\ \tilde{E}_{\mathcal{I}} \end{bmatrix}, \; S^1_{min} = \left[S^1_{min,\mathcal{A}} \mid S^1_{min,\mathcal{I}} \right],$$

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

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

With this restructuring, the linear system (4.7) reads

$$J \left[\begin{array}{c} \Delta\xi^{\mathcal{A}}_{min} \\ \Delta\xi^{\mathcal{I}}_{min} \\ \hline \Delta\xi_{mob} \\ \hline \Delta\bar{c}^{\mathcal{A}} \\ \Delta\bar{c}^{\mathcal{I}} \end{array} \right] = - \left[\begin{array}{c} G_1^{\mathcal{A}} \\ G_1^{\mathcal{I}} \\ \hline -\bar{c}^{\mathcal{A}} \\ -\tilde{E}_{\mathcal{I}} \\ \hline G_3 \end{array} \right], \tag{4.8}$$

with

$$J = \begin{bmatrix} \left(\theta I_{|\mathcal{A}|} + \tau L_h^{\mathcal{A},\mathcal{A}}\right) & \tau L_h^{\mathcal{A},\mathcal{I}} & 0 & I_{|\mathcal{A}|} & 0 \\ \tau L_h^{\mathcal{I},\mathcal{A}} & \left(\theta I_{|\mathcal{I}|} + \tau L_h^{\mathcal{I},\mathcal{I}}\right) & 0 & 0 & I_{|\mathcal{I}|} \\ 0 & 0 & 0 & -I_{|\mathcal{A}|} & 0 \\ -\frac{\partial \tilde{E}_{\mathcal{I}}}{\partial \xi^{\mathcal{A}}_{min}} & -\frac{\partial \tilde{E}_{\mathcal{I}}}{\partial \xi^{\mathcal{I}}_{min}} & -\frac{\partial \tilde{E}_{\mathcal{I}}}{\partial \xi_{mob}} & 0 & 0 \\ \frac{\partial \tilde{Q}_{mob}}{\partial \xi^{\mathcal{A}}_{min}} & \frac{\partial \tilde{Q}_{mob}}{\partial \xi^{\mathcal{I}}_{min}} & \frac{\partial \tilde{Q}_{mob}}{\partial \xi_{mob}} & 0 & 0 \end{bmatrix}. \tag{4.9}$$

From the third set of equations, we immediately obtain

$$-\Delta\bar{c}^{\mathcal{A}} = \bar{c}^{\mathcal{A}}. \tag{4.10}$$

There is no need to compute $\Delta\bar{c}^{\mathcal{A}}$, because of (4.10) we can simply set the *new* Newton iterate as

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

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

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

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

$$\tilde{J}\left[\begin{array}{c}\Delta\xi_{min}^{\mathcal{A}}\\ \Delta\xi_{min}^{\mathcal{I}}\\ \hline \Delta\xi_{mob}\end{array}\right] = -\left[\begin{array}{c}G_1^{\mathcal{A}}-\bar{c}^{\mathcal{A}}\\ \hline -\tilde{E}_{\mathcal{I}}\\ \hline G_3\end{array}\right] \tag{4.11}$$

with

$$\tilde{J} := \begin{bmatrix} \left(\theta I_{|\mathcal{A}|}+\tau L_h^{\mathcal{A},\mathcal{A}}\right) & \tau L_h^{\mathcal{A},\mathcal{I}} & 0\\ -\frac{\partial\tilde{E}_{\mathcal{I}}}{\partial\xi_{min}^{\mathcal{A}}} & -\frac{\partial\tilde{E}_{\mathcal{I}}}{\partial\xi_{min}^{\mathcal{I}}} & -\frac{\partial\tilde{E}_{\mathcal{I}}}{\partial\xi_{mob}}\\ \frac{\partial\tilde{Q}_{mob}}{\partial\xi_{min}^{\mathcal{A}}} & \frac{\partial\tilde{Q}_{mob}}{\partial\xi_{min}^{\mathcal{I}}} & \frac{\partial\tilde{Q}_{mob}}{\partial\xi_{mob}} \end{bmatrix}. \tag{4.12}$$

This linear system is smaller than the original linear system (4.8), and it is solvable if and only if (4.8) is solvable. More precisely, the absolute values of the determinants of J and $\tilde{J}$ coincide. We will verify this statement in the next section.

4.3 More Subdifferentials and their Transformations

For the semismooth Newton method it is important that all elements in $\partial_B G_M$ are nonsingular, especially around and in a solution point. In this section we will actually study the matrices in a superset of $\partial_B G_M$, which will bring great advantages later.

Let $w = [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$ be an arbitrary point. In Section 4.1 we began with an element $H \in \partial_B G_M(w)$ and constructed the corresponding active set $\mathcal{A} \subset (\{1,\dots,J_{min}\} \times \Omega_h)$. And we have seen in (4.3)–(4.5) that these sets fulfill the inclusions $\mathcal{P} \subseteq \mathcal{A} \subseteq (\mathcal{P} \cup \mathcal{Q})$.

Now we start from a an arbitrary set

$$\mathcal{B} \subset \{1,\dots,J_{min}\} \times \Omega_h$$

and its complement

$$\mathcal{J} := (\{1,\dots,J_{min}\} \times \Omega_h) \setminus \mathcal{B}$$

and construct a $((2\cdot J_{min}+J_{mob})\cdot p) \times ((2\cdot J_{min}+J_{mob})\cdot p)$ matrix $H^{\mathcal{B}}$. To this end let $\omega : \{1,\dots,J_{min}\cdot p\} \longrightarrow (\{1,\dots,J_{min}\} \times \Omega_h)$ be a bijective function, which enumerates the elements of $\{1,\dots,J_{min}\} \times \Omega_h$ according to the lexicographical ordering

$$[1,x_1],[2,x_1],\dots,[J_{min},x_1],[1,x_2],[2,x_2],\dots,[J_{min},x_2],\dots,\left[J_{min},x_p\right].$$

For example it is $\omega(2) = [2,x_1]$. We construct the $J_{min}p \times (2\cdot J_{min}+J_{mob})p$ matrix $Q^{\mathcal{B}}$ by defining its k-th row $Q^{\mathcal{B}}(k)$ as

$$Q^{\mathcal{B}}(k) := \begin{cases} \left[\begin{array}{c|c|c} 0 & 0 & -\frac{\partial\bar{c}_i(x)}{\partial\bar{c}} \end{array}\right] & \text{if } \omega(k)\in\mathcal{B}\\ \left[\begin{array}{c|c|c} -\frac{\partial\tilde{E}_i(x)}{\partial\xi_{min}} & -\frac{\partial\tilde{E}_i(x)}{\partial\xi_{mob}} & 0 \end{array}\right] & \text{if } \omega(k)\in\mathcal{J}\,. \end{cases} \tag{4.13}$$

It holds $-\frac{\partial\bar{c}_i(x)}{\partial\bar{c}} = [0,\dots,0,-1,0\dots,0]$ with -1 at the k-th position. We also write $Q^{\mathcal{B}} := \left[Q_1^{\mathcal{B}} \mid Q_2^{\mathcal{B}} \mid Q_3^{\mathcal{B}}\right]$ where this breaking down corresponds to the definition of $Q^{\mathcal{B}}(k)$.

Now we can finally define the matrix

$$H^{\mathcal{B}} := \begin{bmatrix} \theta I + \tau L_h & 0 & I \\ Q_1^{\mathcal{B}} & Q_2^{\mathcal{B}} & Q_3^{\mathcal{B}} \\ \frac{\partial Q_{mob}}{\partial \xi_{min}} & \frac{\partial Q_{mob}}{\partial \xi_{mob}} & 0 \end{bmatrix}$$

and the matrix class

$$\partial_D G(w) = \left\{ H^{\mathcal{B}} \mid \mathcal{B} \subset (\{1, \ldots, J_{min}\} \times \Omega_h) \right\} .$$

Note that $\partial_D G(w)$ is not a set that occurs in literature like the B-subdifferential or the generalized Jacobian. It is just useful for our purposes. For a subset $\mathcal{B} \subset (\{1, \ldots, J_{min}\} \times \Omega_h)$ with $\mathcal{P} \subset \mathcal{B} \subset (\mathcal{P} \cup \mathcal{Q})$ the corresponding matrix $H^{\mathcal{B}}$ coincides with the one defined in the previous section. So $\partial_D G(w)$ is a superset of $\partial_B G_M(w)$ but it has no inclusion relation to $\partial G_M(w)$ or $\partial_C G_M(w)$.

Now we define an extension set of $\partial_D G(w)$. Let $a, b \in \mathbb{R}^{J_{min} p}$. We construct the $J_{min} p \times (2 \cdot J_{min} + J_{mob}) p$ matrix $Q^{a,b} = [Q_1^{a,b} \mid Q_2^{a,b} \mid Q_3^{a,b}]$ by defining its k-th row $Q^{a,b}(k)$ as

$$\begin{aligned} Q^{a,b}(k) &:= \left[\; -a_k \frac{\partial \tilde{E}_i(x)}{\partial \xi_{min}} \;\middle|\; -a_k \frac{\partial \tilde{E}_i(x)}{\partial \xi_{mob}} \;\middle|\; -b_k \frac{\partial \bar{c}_i(x)}{\partial \bar{c}} \; \right] \\ &= \left[\; Q_1^{a,b}(k) \;\middle|\; Q_2^{a,b}(k) \;\middle|\; Q_3^{a,b}(k) \; \right] . \end{aligned} \tag{4.14}$$

That means we take linear combinations of both choices in (4.13). Building on this we define the matrix

$$H^{a,b} := \begin{bmatrix} \theta I + \tau L_h & 0 & I \\ Q_1^{a,b} & Q_2^{a,b} & Q_3^{a,b} \\ \frac{\partial Q_{mob}}{\partial \xi_{min}} & \frac{\partial Q_{mob}}{\partial \xi_{mob}} & 0 \end{bmatrix}$$

and finally the matrix set

$$\partial_E G(w) := \left\{ H^{a,b} \mid a, b \in \mathbb{R}^{J_{min} p},\ a \geq 0,\ b \geq 0,\ (a_i > 0 \ \vee \ b_i > 0)\ \forall i \right\} . \tag{4.15}$$

Please note that we take only nonnegative linear combinations in $\partial_E G(w)$. This set is an extension and generalization of $\partial_C G_M(w)$ similarly like the generalized Jacobian is an extension and generalization of the B-subdifferential. The inclusion relation to the known subdifferentials is

$$\partial_B G(w) \subseteq \partial G_M(w) = \partial_C G_M(w) \subset \partial_E G(w) .$$

With these sets $\partial_D G(w)$ and $\partial_E G(w)$ we are able to unify and simplify our theory. Note while these sets depend on $w \in \mathcal{D}$ it makes no sense to apply ∂_E or ∂_D to other functions. Rather $\partial_D G(w)$ and $\partial_E G(w)$ are defined as is.

Now we will study the determinant of an arbitrary element $H^{\mathcal{B}} \in \partial_D G(w)$, with $w \in \mathcal{D}$. We will do this by transforming $H_1 := H^{\mathcal{B}}$ in several steps. After each step we will record the effect of the transformation on the determinant.

We start with reordering the columns and rows of the first block row and block column in the same way according to the sets $\mathcal{B}$ and $\mathcal{J}$. That means we do exactly the same reordering in the columns and in the rows. Then we do the same reordering in the second block row and in the third block column of H_1. This is the same restructuring that we did in the previous section with J. This yields the matrix

$$H_2 := \begin{bmatrix} \theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} & \tau L_h^{\mathcal{B},\mathcal{J}} & 0 & I_{|\mathcal{B}|} & 0 \\ \tau L_h^{\mathcal{J},\mathcal{B}} & \theta I_{|\mathcal{J}|} + \tau L_h^{\mathcal{J},\mathcal{J}} & 0 & 0 & I_{|\mathcal{J}|} \\ 0 & 0 & 0 & -I_{|\mathcal{B}|} & 0 \\ -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi_{min}^{\mathcal{B}}} & -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi_{min}^{\mathcal{J}}} & -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi_{mob}} & 0 & 0 \\ \frac{\partial Q_{mob}}{\partial \xi_{min}^{\mathcal{B}}} & \frac{\partial Q_{mob}}{\partial \xi_{min}^{\mathcal{J}}} & \frac{\partial Q_{mob}}{\partial \xi_{mob}} & 0 & 0 \end{bmatrix} .$$

We did two times the same number of column and row swaps, so we have

$$\det(H_2) = \det(H_1) .$$

Now we do some column and row additions to zero the entries $\tau L_h^{\mathcal{J},\mathcal{B}}$, $\theta I_{|\mathcal{J}|} + \tau L_h^{\mathcal{J},\mathcal{J}}$ and $I_{|\mathcal{B}|}$. We do this by adding linear combinations of the fifth block column of H_2 to the first and second block columns and by adding the third block row to the first block row. This yields the matrix

$$H_3 := \begin{bmatrix} \theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} & \tau L_h^{\mathcal{B},\mathcal{J}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{|\mathcal{J}|} \\ 0 & 0 & 0 & -I_{|\mathcal{B}|} & 0 \\ -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi_{min}^{\mathcal{B}}} & -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi_{min}^{\mathcal{J}}} & -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi_{mob}} & 0 & 0 \\ \frac{\partial Q_{mob}}{\partial \xi_{min}^{\mathcal{B}}} & \frac{\partial Q_{mob}}{\partial \xi_{min}^{\mathcal{J}}} & \frac{\partial Q_{mob}}{\partial \xi_{mob}} & 0 & 0 \end{bmatrix} .$$

These changes do not affect the determinant, because adding multiples of a row or column to another row or column does not change the determinant. So we have

$$\det(H_3) = \det(H_2) .$$

We swap the second and fourth block row of H_3, which have the same number of rows. This yields the matrix

$$H_4 := \begin{bmatrix} \theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} & \tau L_h^{\mathcal{B},\mathcal{J}} & 0 & 0 & 0 \\ -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi_{min}^{\mathcal{B}}} & -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi_{min}^{\mathcal{J}}} & -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi_{mob}} & 0 & 0 \\ 0 & 0 & 0 & -I_{|\mathcal{B}|} & 0 \\ 0 & 0 & 0 & 0 & I_{|\mathcal{J}|} \\ \frac{\partial Q_{mob}}{\partial \xi_{min}^{\mathcal{B}}} & \frac{\partial Q_{mob}}{\partial \xi_{min}^{\mathcal{J}}} & \frac{\partial Q_{mob}}{\partial \xi_{mob}} & 0 & 0 \end{bmatrix} .$$

This change can be done by $|\mathcal{J}|$ row swaps. So we have

$$\det(H_4) = (-1)^{|\mathcal{J}|} \det(H_3) .$$

Now we multiply the third block row of H_4 with -1. That means we multiply $|\mathcal{B}|$ rows with -1. This yields the matrix

$$H_5 := \begin{bmatrix} \theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} & \tau L_h^{\mathcal{B},\mathcal{J}} & 0 & 0 & 0 \\ -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi^{\mathcal{B}}_{min}} & -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi^{\mathcal{J}}_{min}} & -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi_{mob}} & 0 & 0 \\ 0 & 0 & 0 & I_{|\mathcal{B}|} & 0 \\ 0 & 0 & 0 & 0 & I_{|\mathcal{J}|} \\ \frac{\partial Q_{mob}}{\partial \xi^{\mathcal{B}}_{min}} & \frac{\partial Q_{mob}}{\partial \xi^{\mathcal{J}}_{min}} & \frac{\partial Q_{mob}}{\partial \xi_{mob}} & 0 & 0 \end{bmatrix} .$$

This of course changes the determinant to

$$\det(H_5) = (-1)^{|\mathcal{B}|} \det(H_4) .$$

Now we switch the third and fourth block row with the fifth block row. This yields the matrix

$$H_6 := \begin{bmatrix} \theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} & \tau L_h^{\mathcal{B},\mathcal{J}} & 0 & 0 & 0 \\ -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi^{\mathcal{B}}_{min}} & -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi^{\mathcal{J}}_{min}} & -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi_{mob}} & 0 & 0 \\ \frac{\partial Q_{mob}}{\partial \xi^{\mathcal{B}}_{min}} & \frac{\partial Q_{mob}}{\partial \xi^{\mathcal{J}}_{min}} & \frac{\partial Q_{mob}}{\partial \xi_{mob}} & 0 & 0 \\ 0 & 0 & 0 & I_{|\mathcal{B}|} & 0 \\ 0 & 0 & 0 & 0 & I_{|\mathcal{J}|} \end{bmatrix} .$$

The effects of this operation on the determinant are according to Theorem A.2.3 in the appendix. So we have

$$\det(H_6) = (-1)^{(|\mathcal{B}|+|\mathcal{J}|)(J_{mob}\cdot p)} \cdot \det(H_5) .$$

In the last step we just drop the lower right corner and yield the matrix

$$H_7 := \begin{bmatrix} \theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} & \tau L_h^{\mathcal{B},\mathcal{J}} & 0 \\ -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi^{\mathcal{B}}_{min}} & -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi^{\mathcal{J}}_{min}} & -\frac{\partial \tilde{E}_{\mathcal{J}}}{\partial \xi_{mob}} \\ \frac{\partial Q_{mob}}{\partial \xi^{\mathcal{B}}_{min}} & \frac{\partial Q_{mob}}{\partial \xi^{\mathcal{J}}_{min}} & \frac{\partial Q_{mob}}{\partial \xi_{mob}} \end{bmatrix} .$$

Since the block we left out was the identity matrix we have

$$\det(H_7) = \det(H_6) .$$

Now we put all the changes of the determinant together in one formula

$$\begin{aligned} \det(H_7) &= (-1)^{(|\mathcal{B}|+|\mathcal{J}|)(J_{mob}\cdot p+1)} \det(H_1) && (4.16) \\ &= (-1)^{(J_{min}\cdot p)(J_{mob}\cdot p+1)} \det\left(H^{\mathcal{B}}\right) . && (4.17) \end{aligned}$$

If $\mathcal{B} = \mathcal{A}$ with $\mathcal{A}$ from the previous section, then $H_1 = J$ and $H_7 = \tilde{J}$, again with J and $\tilde{J}$ from the previous section. Therefore the absolute value of the determinant of J and $\tilde{J}$ coincide as was claimed at the end of that section. The factor $(J_{min} \cdot p)(J_{mob} \cdot p + 1)$ will be in many cases an even number. However it does *not* depend on the choice of $\mathcal{B}$.

Now let us examine a submatrix of H_7, namely

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

The nonsingularity of this matrix is shown even more generally in [25, Section 4.4.5]. Every block of B is itself a block diagonal matrix. For example,

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

with $x_i \in \Omega_h$. By column and row interchanges, we can transform B into a block diagonal matrix

$$\tilde{B} = \operatorname{diag}\left(\tilde{B}(x_1), \tilde{B}(x_2), \ldots, \tilde{B}(x_p)\right)$$

where a Block $\tilde{B}(x_i)$ has the form

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

Since every block of $-\frac{\partial \tilde{E}_{\mathcal{J}}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{J}}}$ and $\frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{mob}}$ is square this transformation requires exactly the same row and column exchanges. This means that there is an orthogonal matrix O (a permutation matrix) such that

$$B = O^T \tilde{B} O \tag{4.18}$$

holds. This implies that B is positive definite if and only if $\tilde{B}$ is positive definite. But $\tilde{B}$ is positive definite if and only if every block $\tilde{B}(x_i)$ is positive definite.

With the definitions of $\tilde{Q}_{mob}$ and $\tilde{E}$ from Chapter 3 and the representation (3.29) of c, we can easily see that

$$\tilde{B}(x_i) = \left[S^{\,1}_{min,\mathcal{J}(x_i)} \mid S^{\,1}_{mob}\right]^T \Lambda_{c(x_i)} \left[S^{\,1}_{min,\mathcal{J}(x_i)} \mid S^{\,1}_{mob}\right] \tag{4.19}$$

holds, where $\Lambda_{c(x_i)} = \operatorname{diag}\left(\frac{1}{c_1(x_i)}, \ldots, \frac{1}{c_I(x_i)}\right)$, $\mathcal{J}(x_i) := \{j \mid (j, x_i) \in \mathcal{J}\} \subset \{1, \ldots, J_{min}\}$ is the projection of $\mathcal{J}$ on $x_i \in \Omega_h$ and $S^{\,1}_{min,\mathcal{J}(x_i)}$ is the submatrix of $S^{\,1}_{min}$ consisting of the columns of $\mathcal{J}(x_i)$. Our matrix $H^{\mathcal{B}}$ is an element of $\partial_D G(w)$ with $w \in \mathcal{D}$. It follows that $c(x_i) > 0$ for all $x_i \in \Omega_h$ from the definition of $\mathcal{D}$ in (3.35). Therefore the blocks $\tilde{B}(x_i)$ are always symmetric positive definite in view of our rank condition (3.13) for all $x_i \in \Omega_h$. So $\tilde{B}$ and B are symmetric positive definite. In particular they are nonsingular. Note that $\tilde{B}$ and B depend on the index sets $\mathcal{B}$ and $\mathcal{J}$. But regardless of the choice of these index sets they are always symmetric positive definite.

Since B is nonsingular, the columns of B form a basis of its column space. Consequently, there are unique matrices D_1 and D_2 such that

$$B \cdot \begin{bmatrix} D_1 \\ D_2 \end{bmatrix} = -\begin{bmatrix} -\frac{\partial \tilde{E}_{\mathcal{J}}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{B}}} \\ \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{B}}} \end{bmatrix} \tag{4.20}$$

or, more detailed

$$\begin{bmatrix} -\frac{\partial \tilde{E}_{\mathcal{J}}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{J}}} \\ \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{J}}} \end{bmatrix} D_1 + \begin{bmatrix} -\frac{\partial \tilde{E}_{\mathcal{J}}(\xi_{min},\xi_{mob})}{\partial \xi_{mob}} \\ \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{mob}} \end{bmatrix} D_2 = -\begin{bmatrix} -\frac{\partial \tilde{E}_{\mathcal{J}}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{B}}} \\ \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{B}}} \end{bmatrix}.$$

Note that D_1 and D_2 also depend upon $\mathcal{B}$ and $\mathcal{J}$. Next we multiply H_7 in (4.12) with the block matrix

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

from the right hand side and obtain

$$H_8 := H_7 \cdot X = \left[\begin{array}{c|cc} \theta I_{|\mathcal{J}|} + \tau L_h^{\mathcal{B},\mathcal{B}} + \tau L_h^{\mathcal{B},\mathcal{J}} \cdot D_1 & \tau L_h^{\mathcal{B},\mathcal{J}} & 0 \\ \hline 0 & \partial \xi_{min}^{\mathcal{J}} & -\frac{\partial \tilde{E}_{\mathcal{J}}(\xi_{min},\xi_{mob})}{\partial \xi_{mob}} \\ 0 & \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{J}}} & \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{mob}} \end{array}\right].$$

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

$$\det H_8 = \det H_7 \,.$$

On the other hand, the determinant of H_8 is given by

$$\det H_8 = \det\left(\theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} + \tau L_h^{\mathcal{B},\mathcal{J}} \cdot D_1\right) \cdot \det B.$$

Therefore H_7 is nonsingular if and only if $\theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} + \tau L_h^{\mathcal{B},\mathcal{J}} \cdot D_1$ is nonsingular. Together with the representation for the determinant of H_7 in (4.16) we have proven the following result.

Lemma 4.3.1. *Let $H^{\mathcal{B}} \in \partial_D G(w)$ be arbitrary and $w \in \mathcal{D}$.Then it holds*

$$\det\left(H^{\mathcal{B}}\right) = (-1)^{(J_{min} \cdot p)(J_{mob} \cdot p+1)} \cdot \det\left(\theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} + \tau L_h^{\mathcal{B},\mathcal{J}} \cdot D_1\right) \cdot \det B \,.$$

In particular $H^{\mathcal{B}}$ is nonsingular if and only if $\theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} + \tau L_h^{\mathcal{B},\mathcal{J}} \cdot D_1$ is nonsingular.

The next theorem is one of the reasons why we introduced $\partial_E G(w)$. If we know something about the sign of the determinant of the elements of $\partial_D G(w)$ we can transfer this to the elements of $\partial_E G(w)$. This will turn out useful later.

Theorem 4.3.2. *Let $w \in \mathcal{D}$. Then*

$$\begin{aligned} \det H > 0 \ \forall H \in \partial_D G(w) &\iff \det H > 0 \ \forall H \in \partial_E G(w) \\ \det H < 0 \ \forall H \in \partial_D G(w) &\iff \det H < 0 \ \forall H \in \partial_E G(w) \end{aligned}$$

Proof. We only show the first equivalence, because the second one can be proved in the same way.

The implication

$$\det(H) > 0 \ \forall H \in \partial_E G(w) \quad \Longrightarrow \quad \det(H) > 0 \ \forall H \in \partial_D G(w)$$

is trivial since $\partial_D G(w) \subset \partial_E G(w)$ holds. So we will only show the other implication

$$\det(H) > 0 \ \forall H \in \partial_D G(w) \quad \Longrightarrow \quad \det(H) > 0 \ \forall H \in \partial_E G(w)\,.$$

Before we can do this we need some additional notation building on the notation we introduced in the beginning of this section.

Let $H^{a,b} \in \partial_E G(w)$ be arbitrary and fixed for the whole proof, with $a, b \in \mathbb{R}^{J_{min} p}$. It is a $q \times q$ matrix with $q := (2 \cdot J_{min} + J_{mob}) \cdot p$ and $p = |\Omega_h|$. The vectors a, b have the properties as defined in $\partial_E G(w)$, see (4.15). Let $\Phi_k := \{\omega(1), \omega(2), \ldots, \omega(k)\} \subset \{1, \ldots, J_{min}\} \times \Omega_h$. Let $m := J_{min} \cdot p$. For $\mathcal{B} \subset \Phi_k$ and $k \in \{1, \ldots, m\}$ we define

$$c_k^{\mathcal{B}} := \prod_{i=1}^{k} c^i \text{ with } c^i := \begin{cases} a_i & \text{for } \omega(i) \in \mathcal{B} \\ b_i & \text{for } \omega(i) \notin \mathcal{B} \end{cases}.$$

From the properties of a, b it follows that

$$c_k^{\mathcal{B}} \geq 0$$

holds for all $\mathcal{B} \subset \Phi_k$.

For abbreviation and for $(i, x) := \omega(k)$ we define the row vectors

$$\begin{aligned} g^k &:= \left[-\frac{\partial \tilde{E}_i(x)}{\partial \xi_{min}} \;\middle|\; -\frac{\partial \tilde{E}_i(x)}{\partial \xi_{mob}} \;\middle|\; 0 \right] \\ h^k &:= \left[\quad 0 \quad \middle| \quad 0 \quad \middle| \; -\frac{\partial \bar{c}_i(x)}{\partial \bar{c}} \right] \end{aligned}$$

both with q components. We define the $(m-k) \times q$ matrix $Q_k^{a,b}$ as the $m-k$ *last* rows of the $m \times q$ submatrix $Q^{a,b}$ of $H^{a,b}$ defined in (4.14). Then the i-th row $Q_k^{a,b}(i)$ of $Q_k^{a,b}$ looks like

$$Q_k^{a,b}(i) = a_{i+k} g^{i+k} + b_{i+k} h^{i+k}\,.$$

Similarly for $\mathcal{B} \subset \Phi_k$ we define the $k \times q$ matrix $Q_k^{\mathcal{B}}$ as the *first* k rows of the matrix $Q^{\mathcal{B}}$ defined in (4.13). Then the i-th row $Q_k^{\mathcal{B}}(i)$ of $Q_k^{\mathcal{B}}$ looks like

$$Q_k^{\mathcal{B}}(i) = \begin{cases} g^i & \text{for } \omega(i) \in \mathcal{B} \\ h^i & \text{for } \omega(i) \notin \mathcal{B} \end{cases}.$$

Finally we define the $(m+k) \times q$ matrix

$$S_k^{\mathcal{B}} := \begin{bmatrix} [\theta I + \tau L_h \mid 0 \mid I] \\ Q_k^{\mathcal{B}} \end{bmatrix}$$

and the $(m - k + J_{mob}p) \times q$ matrix

$$T^k := \left[\begin{array}{c} Q_k^{a,b} \\ \left[\frac{\partial Q_{mob}}{\partial \xi_{min}} \mid \frac{\partial Q_{mob}}{\partial \xi_{mob}} \mid 0\right] \end{array} \right].$$

With this notation it holds

$$H^{a,b} = \left[\begin{array}{c} S_0^{\emptyset} \\ T^0 \end{array} \right].$$

And for $\mathcal{B} \subset \Phi_m = \{1, \ldots, J_{min}\} \times \Omega_h$ the matrix

$$\left[\begin{array}{c} S_m^{\mathcal{B}} \\ T^m \end{array} \right]$$

is an element of $\partial_D G(w)$. Now we can start the rather simple induction proof. The induction assertion for $i \in \{1, \ldots, m\}$ is

$$\det H^{a,b} = \sum_{\mathcal{B} \subset \Phi_i} c_i^{\mathcal{B}} \det \left[\begin{array}{c} S_i^{\mathcal{B}} \\ T^i \end{array} \right],$$

and there is (at least) one set $\mathcal{J} \subset \Phi_i$ with $c_i^{\mathcal{J}} > 0$.

Induction start for $i = 1$: We can write $H^{a,b}$ as

$$H^{a,b} = \left[\begin{array}{c} S^{\emptyset} \\ T^0 \end{array} \right] = \left[\begin{array}{c} S_0^{\emptyset} \\ a_1 g^1 + b_1 h^1 \\ T^1 \end{array} \right].$$

With the linearity of the determinant in every row it follows

$$\begin{aligned} \det H^{a,b} &= a_1 \det \left[\begin{array}{c} S_0^{\emptyset} \\ g^1 \\ T^1 \end{array} \right] + b_1 \det \left[\begin{array}{c} S_0^{\emptyset} \\ h^1 \\ T^1 \end{array} \right] \\ &= \sum_{\mathcal{B} \subset \Phi_1} c_1^{\mathcal{B}} \det \left[\begin{array}{c} S_1^{\mathcal{B}} \\ T^1 \end{array} \right]. \end{aligned}$$

From the definition of a, b either $a_1 > 0$ or $b_1 > 0$ holds.

Induction step $i \leadsto i + 1$ for $i < m$: We assume that our induction assertion is true for i, i.e.

$$\det H^{a,b} = \sum_{\mathcal{B} \subset \Phi_i} c_i^{\mathcal{B}} \det \left[\begin{array}{c} S_i^{\mathcal{B}} \\ T^i \end{array} \right],$$

and there is a set $\mathcal{J} \subset \Phi_i$ with $c_i^{\mathcal{J}} > 0$. Then

$$\begin{aligned}
\det H^{a,b} &= \sum_{\mathcal{B}\subset\Phi_i} c_i^{\mathcal{B}} \det\begin{bmatrix} S_i^{\mathcal{B}} \\ T^i \end{bmatrix} \\
&= \sum_{\mathcal{B}\subset\Phi_i} c_i^{\mathcal{B}} \det\begin{bmatrix} S_i^{\mathcal{B}} \\ a_{i+1}g^{i+1} + b_{i+1}h^{i+1} \\ T^{i+1} \end{bmatrix} \\
&= \sum_{\mathcal{B}\subset\Phi_i} c_i^{\mathcal{B}} \left(a_{i+1} \det\begin{bmatrix} S_i^{\mathcal{B}} \\ g^{i+1} \\ T^{i+1} \end{bmatrix} + b_{i+1} \det\begin{bmatrix} S_i^{\mathcal{B}} \\ h^{i+1} \\ T^{i+1} \end{bmatrix} \right) \\
&= \sum_{\mathcal{B}\subset\Phi_i} \left(c_{i+1}^{\hat{\mathcal{B}}} \det\begin{bmatrix} S_i^{\mathcal{B}} \\ g^{i+1} \\ T^{i+1} \end{bmatrix} + c_{i+1}^{\mathcal{B}} \det\begin{bmatrix} S_i^{\mathcal{B}} \\ h^{i+1} \\ T^{i+1} \end{bmatrix} \right)
\end{aligned}$$

with $\hat{\mathcal{B}} \subset \Phi_{i+1}$ defined as $\hat{\mathcal{B}} := \mathcal{B} \cup \{\omega(i+1)\}$. Now let $\mathcal{J} \subset \Phi_i$ with $c_i^{\mathcal{J}} > 0$. Either $a_{i+1} > 0$ or $b_{i+1} > 0$ holds. Therefore we have either $c_{i+1}^{\hat{\mathcal{J}}} > 0$ or $c_{i+1}^{\mathcal{J}} > 0$, where $\hat{\mathcal{J}} := \mathcal{J} \cup \{\omega(i+1)\}$. For every set $\mathcal{B} \subset \Phi_{i+1}$ it holds either $\mathcal{B} \subset \Phi_i$ or $\mathcal{B} \setminus \{\omega(i+1)\} \subset \Phi_i$. Therefore we can resume our equation chain with

$$\begin{aligned}
&= \sum_{\mathcal{B}\subset\Phi_i} \left(c_{i+1}^{\hat{\mathcal{B}}} \det\begin{bmatrix} S_{i+1}^{\hat{\mathcal{B}}} \\ T^{i+1} \end{bmatrix} + c_{i+1}^{\mathcal{B}} \det\begin{bmatrix} S_{i+1}^{\mathcal{B}} \\ T^{i+1} \end{bmatrix} \right) \\
&= \sum_{\mathcal{B}\subset\Phi_{i+1}} c_{i+1}^{\mathcal{B}} \det\begin{bmatrix} S_{i+1}^{\mathcal{B}} \\ T^{i+1} \end{bmatrix} .
\end{aligned}$$

This ends the induction proof. For $k = m$ we have

$$\det H^{a,b} = \sum_{\mathcal{B}\subset\Phi_m} c_m^{\mathcal{B}} \det\begin{bmatrix} S_m^{\mathcal{B}} \\ T^m \end{bmatrix}, \tag{4.21}$$

where $\begin{bmatrix} S_m^{\mathcal{B}} \\ T^m \end{bmatrix}$ is an element of $\partial_D G(w)$. It is

$$\det\begin{bmatrix} S_m^{\mathcal{B}} \\ T^m \end{bmatrix} > 0$$

the assumption of the implication we are proving. Together with $c_m^{\mathcal{B}} \geq 0$ we have

$$c_m^{\mathcal{B}} \det\begin{bmatrix} S_m^{\mathcal{B}} \\ T^m \end{bmatrix} \geq 0$$

for all $\mathcal{B} \subset \Phi_m$. And there is at least one subset $\mathcal{J} \subset \Phi_m$ with

$$c_m^{\mathcal{J}} \det\begin{bmatrix} S_m^{\mathcal{J}} \\ T^m \end{bmatrix} > 0 .$$

So we have

$$\det H^{a,b} > 0 .$$

□

4.4 First Proof of Nonsingularity

In this section we continue to study $\tilde{J}$ from Section 4.2. The aim is to prove the nonsingularity of J and $\tilde{J}$, which were defined in (4.9) and (4.12) respectively. We have already seen in the previous section that

$$\det \tilde{J} = (-1)^{(J_{min}\cdot p)(J_{mob}\cdot p+1)} \det J \tag{4.22}$$

holds and therefore the absolute values of the determinants of J and $\tilde{J}$ coincide. There we have also verified the identity

$$\det \tilde{J} = \det K \cdot \det B \tag{4.23}$$

with

$$K := \theta I_{|\mathcal{A}|} + \tau L_h^{\mathcal{A},\mathcal{A}} + \tau L_h^{\mathcal{A},\mathcal{I}} \cdot D_1$$

and

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

We already know that B is symmetric positive definite and therefore $\det B > 0$ holds. So $\tilde{J}$ is nonsingular if and only if K is nonsingular.

The matrix K can be written as a sum of the diagonal matrix $\theta I_{|\mathcal{A}|}$ and the square matrix $\tau L_h^{\mathcal{A},\mathcal{A}} + \tau L_h^{\mathcal{A},\mathcal{I}} \cdot D_1$. Therefore Theorem A.1.1 is applicable for K and its determinant can be calculated with the formula

$$\det K = \sum_{\alpha \subset \mathcal{L}} \det \theta I^{\alpha,\alpha} \cdot \det \left(\tau L_h^{\mathcal{A},\mathcal{A}} + \tau L_h^{\mathcal{A},\mathcal{I}} \cdot D_1\right)^{\bar{\alpha},\bar{\alpha}},$$

where $\mathcal{L} = \{1, \ldots, |\mathcal{A}|\}$ and $\bar{\alpha} := \mathcal{L} \setminus \alpha$. Since the determinant of a 0×0 matrix is defined as 1, we get

$$\begin{aligned} \det K &= \sum_{\alpha \subset \mathcal{L}} \theta^{|\alpha|} \cdot \det \left(\tau L_h^{\mathcal{A},\mathcal{A}} + \tau L_h^{\mathcal{A},\mathcal{I}} \cdot D_1\right)^{\bar{\alpha},\bar{\alpha}} \\ &= \theta^{|\mathcal{A}|} + \sum_{\substack{\alpha \subset \mathcal{L} \\ |\alpha| < |\mathcal{A}|}} \theta^{|\alpha|} \cdot \tau^{|\bar{\alpha}|} \det \left(L_h^{\mathcal{A},\mathcal{A}} + L_h^{\mathcal{A},\mathcal{I}} \cdot D_1\right)^{\bar{\alpha},\bar{\alpha}}. \end{aligned} \tag{4.24}$$

For the next theorem, we assume that our PDE matrix L_h depends on h but not on τ. This should be always the case, because L_h comes from the discretization of derivatives for spatial variables. Furthermore, we assume that the spatial step size h is given and fixed. Then our theorem states the dependence of the nonsingularity of J on the time step size τ.

Theorem 4.4.1. *Let h be an arbitrarily given spatial step size. Then, for all sufficiently small time steps τ, the system matrices J and $\tilde{J}$ are nonsingular. Furthermore, there are at most $J_{min} \cdot p$ time steps τ such that J and $\tilde{J}$ are singular.*

Proof. Equation (4.24) shows that the determinant of K is a polynomial in τ. The degree of this polynomial is $|\mathcal{A}|$, where $|\mathcal{A}| \leq J_{min} \cdot p$ always holds by definition of the active set $\mathcal{A}$. So this polynomial has a maximum degree of $J_{min} \cdot p$. It is not the zero polynomial since it has $\theta^{|\mathcal{A}|}$ as constant term. So $J_{min} \cdot p$ is also the maximum number of its roots. Hence either all roots are complex, or there exists a smallest positive root which is our smallest time step. Since $\det B \neq 0$ always holds, and since we have $\det K \cdot \det B = \det \tilde{J} = \pm \det J$ according to (4.23) and (4.22), the assertion follows. □

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

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

holds.

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

Theorem 4.4.2. *Let the PDE operator L_h result from a difference scheme of first or second order. Then the system matrices J and $\tilde{J}$ are nonsingular for all sufficiently small space steps h. Furthermore, there are at most $2 \cdot J_{min} \cdot p$ space steps h such that J and $\tilde{J}$ are singular.*

Proof. Every non-vanishing entry of L_h is a polynomial in $\frac{1}{h}$ of first or second order. The same holds for $L_h^{\mathcal{A},\mathcal{A}} + L_h^{\mathcal{A},\mathcal{I}} \cdot D_1$ and all its submatrices. Thanks to the Leibniz formula it holds that the factors $\det(L_h^{\mathcal{A},\mathcal{A}} + L_h^{\mathcal{A},\mathcal{I}} \cdot D_1)^{\bar{\alpha},\bar{\alpha}}$ in (4.24) are polynomials in $\frac{1}{h}$ of maximal degree $2\,|\bar{\alpha}| \leq 2 \cdot |\mathcal{A}|$ with a zero constant term. Since $|\mathcal{A}| \leq J_{min}p$ holds, we can conclude with (4.24) that $\det K$ is always a polynomial in $\frac{1}{h}$ of degree at most $2J_{min}p$. Again, $\theta^{|\mathcal{A}|}$ is the constant term of this polynomial, hence it is not the zero polynomial. Therefore it has at most $2J_{min}p$ roots.

Let z_∞ be the largest real root of this polynomial (if it has no real roots, we are finished). Then there exists a corresponding smallest positive space step h_0 with $h_0 = \frac{1}{z_\infty}$. So $\det K \neq 0$ holds for all $h \in (0, h_0)$. Since $\det B \neq 0$ always holds, and because $\det K \cdot \det B = \det \tilde{J} = \pm \det J$ according to (4.23) and (4.22), we have proven everything. □

We now generalize the previous two theorems slightly to all elements of the B-subdifferential of G_M.

Corollary 4.4.3. *Let $w := [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$. Then the following statements hold:*

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

(2) *Let τ be given and let L_h be as in Theorem 4.4.2. Then all $H \in \partial_B G_M(w)$ are nonsingular for all sufficiently small space steps h. Furthermore, there are only a finite number of space steps h such that at least one element in $\partial_B G_M(w)$ is singular.*

Proof. We have already shown the two statements for an arbitrary element J from the B-subdifferential. Hence the desired statements follow from Theorems 4.4.1 and 4.4.2, respectively, taking into account that the number of matrices in $\partial_B G_M(w)$ is finite, cf. Lemma 4.1.2. □

Note that all the previous nonsingularity results hold at an arbitrary point $w \in \mathcal{D}$. Hence all iterations of our Newton-type method are (not only locally) well-defined as long as they stay in $\mathcal{D}$. But it should be mentioned that the minimal time step size τ in two different grid points $w \in \mathcal{D}$ and $w' \in \mathcal{D}$ may differ. So this value could decrease constantly during a Newton iteration. This is unsatisfactory. In the next section we will improve this result with a different approach.

After we have ensured that the semismooth Newton method in Algorithm 4.2.1 is well defined almost everywhere, we formulate its main local convergence result.

Theorem 4.4.4. *Let $w^* \in \mathcal{D}$ be a BD-regular point of G_M. Then there exists an $\epsilon > 0$ such that for every starting point $w^0 \in B_\epsilon(w^*)$, the following assertions hold:*

(1) *The Newton-type iteration defined in Algorithm 4.2.1 is well-defined and produces a sequence $\left\{w^k\right\}$ that converges to w^*.*

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

Proof. From the definition of $\mathcal{D}$ in (3.35) it is easy to see that $\mathcal{D}$ is an open and convex set. So we can choose $\epsilon > 0$ so that $B_\epsilon(w^*) \subset \mathcal{D}$ holds. In Lemma 4.1.1 we have seen that G_M is strongly semismooth. With Theorem 2.5.4 the assertion follows. □

Unfortunately, we do not know before whether the requirement of Theorem 4.4.4 regarding the nonsingularity of all elements from the B-subdifferential of G_M is fulfilled. However, Corollary 4.4.3 guarantees that it is at least very unlikely to hit a solution point w^* where this requirement is not satisfied. Moreover, it shows that we can change this situation by changing the time step size τ or the spatial step size h (for practical reasons, it is easier to change τ). But after changing the time step size τ, the Newton iteration has to be restarted. In our computational test runs, we never had problems with singular matrices from $\partial_B G_M$.

4.5 New Proof of Nonsingularity

The result from the previous section is somewhat unsatisfactory. For a given step size h, we know that there is only a finite number of τ so that there is a singular matrix in $\partial_B G_M(w)$.

So if $\tau > 0$ is small enough $\partial_B G_M(w)$ has maximal rank, i.e. all elements are nonsingular. But we do not know how small τ has to be chosen. And it could decrease from time step to time step or even during the Newton iteration. In this section we will develop a threshold $\tau_{max} > 0$ such that even $\partial_D G(w)$ has maximal rank, if τ is chosen in the interval $0 \leq \tau < \tau_{max}$. Remember that $\partial_D G(w)$ is a superset of $\partial_B G_M(w)$. Working with $\partial_D G(w)$ instead of $\partial_B G_M(w)$ will not complicate the matter, but will prove very useful later.

We start with the result from Lemma 4.3.1. Let $H^{\mathcal{B}} \in \partial_D G(w)$ be arbitrary, then

$$\det\left(H^{\mathcal{B}}\right) = (-1)^{(J_{min}\cdot p)(J_{mob}\cdot p+1)} \cdot \det\left(\theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} + \tau L_h^{\mathcal{B},\mathcal{J}} \cdot D_1\right) \cdot \det B$$

holds. We already know that B is always symmetric positive definite for all $w \in \mathcal{D}$ and for all index set $\mathcal{B}$ and that $\det B > 0$ always holds. Our aim is now to find τ_{max} such that $\det\left(\theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} + \tau L_h^{\mathcal{B},\mathcal{J}} \cdot D_1\right)$ is positive for $\tau < \tau_{max}$. First we have

$$K := \theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} + \tau L_h^{\mathcal{B},\mathcal{J}} \cdot D_1 = \theta I_{|\mathcal{B}|} + \tau \left[L_h^{\mathcal{B},\mathcal{B}} \mid L_h^{\mathcal{B},\mathcal{J}}\right] \cdot \begin{bmatrix} I \\ D_1 \end{bmatrix}.$$

It is well known that the determinant of a square matrix is equal to the product of its eigenvalues. Our matrix K is a real matrix and its characteristic polynomial has therefore only real coefficients. Hence if the smallest real eigenvalue is positive, then the determinant of this matrix will be positive. The eigenvalues in $\mathbb{C} \setminus \mathbb{R}$ do not matter, because they only occur in complex conjugate pairs, i.e. numbers $a+bi$, $a-bi$. The product of these conjugate pairs is $a^2 + b^2 > 0$. This product $a^2 + b^2$ can't be zero, because that would mean that $a + bi = 0$, which is a real number. If there is no real eigenvalue the determinant must be positive.

If we can show that the smallest real eigenvalue of K is positive then we know that its determinant is positive, too. With [17, Corollary 6.3.4] we obtain a crude estimate for its smallest real eigenvalue

$$\begin{aligned} \lambda_1\left(\theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} + \tau L_h^{\mathcal{B},\mathcal{J}} \cdot D_1\right) &\geq \lambda_1\left(\theta I_{|\mathcal{B}|}\right) - \left\|\tau \left[L_h^{\mathcal{B},\mathcal{B}} \mid L_h^{\mathcal{B},\mathcal{J}}\right] \cdot \begin{bmatrix} I \\ D_1 \end{bmatrix}\right\|_{sp} \\ &\geq \theta - \tau \cdot \left\|\left[L_h^{\mathcal{B},\mathcal{B}} \mid L_h^{\mathcal{B},\mathcal{J}}\right]\right\|_{sp} \left\|\begin{bmatrix} I \\ D_1 \end{bmatrix}\right\|_{sp}. \end{aligned}$$

Interchanging the columns of a matrix does not change its spectral norm, because this can be done by multiplication with an orthogonal matrix. Therefore we have $\left\|\left[L_h^{\mathcal{B},\mathcal{B}} \mid L_h^{\mathcal{B},\mathcal{J}}\right]\right\|_{sp} = \left\|L_h^{\mathcal{B}}\right\|_{sp}$. Adding rows to a matrix can only increase its spectral norm, which can be seen by its definition as lub-norm. So we have $\left\|L_h^{\mathcal{B}}\right\|_{sp} \leq \|L_h\|_{sp}$. For an arbitrary vector $a \in \mathbb{R}^{|\mathcal{B}|}$ with $\|a\|_2 = 1$ we have

$$\left\|\begin{bmatrix} I \\ D_1 \end{bmatrix} \cdot a\right\|_2^2 = 1 + \|D_1 a\|_2^2$$

and can therefore conclude that

$$\left\|\begin{bmatrix} I \\ D_1 \end{bmatrix}\right\|_{sp} = \sqrt{1 + \|D_1\|_{sp}^2}.$$

With this we can resume our estimation chain with

$$\theta - \tau \cdot \left\| \left[L_h^{\mathcal{B},\mathcal{B}} \mid L_h^{\mathcal{B},\mathcal{J}} \right] \right\|_{sp} \left\| \begin{bmatrix} I \\ D_1 \end{bmatrix} \right\|_{sp} \quad \geq \quad \theta - \tau \cdot \|L_h\|_{sp} \cdot \sqrt{1 + \|D_1\|_{sp}^2} \,.$$

Now we take a closer look at D_1. Remember that D_1 and D_2 were defined in (4.20) as the unique matrices that fulfille

$$\begin{bmatrix} -\frac{\partial \tilde{E}_{\mathcal{J}}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{J}}} \\ \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{J}}} \end{bmatrix} D_1 + \begin{bmatrix} -\frac{\partial \tilde{E}_{\mathcal{J}}(\xi_{min},\xi_{mob})}{\partial \xi_{mob}} \\ \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{mob}} \end{bmatrix} D_2 = - \begin{bmatrix} -\frac{\partial \tilde{E}_{\mathcal{J}}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{B}}} \\ \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{B}}} \end{bmatrix} .$$

This means that D_1 depends on the variables $[\xi_{min}, \xi_{mob}]$ and on the index sets $\mathcal{B}$ and $\mathcal{J}$ in every evaluation point $w = [\xi_{min}, \xi_{mob}, \bar{c}]$. Lemma 6.4.1 gives a constant $s > 0$ such that $\|D_1\|_{sp} < s$ holds in all points $w \in \mathcal{D}$ and for all index sets $\mathcal{B}$ and $\mathcal{J}$. This result is not so easy to prove. It requires a small theory in itself. It is done in Chapter 6. So together we have shown that

$$\lambda_1 \left(\theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} + \tau L_h^{\mathcal{B},\mathcal{J}} \cdot D_1 \right) \geq \theta - \tau \cdot \|L_h\|_{sp} \cdot \sqrt{1 + s^2}$$

holds. The right hand side is positive if and only if

$$\tau < \frac{\theta}{\|L_h\|_{sp} \cdot \sqrt{1 + s^2}} =: \tau_{max} \,. \tag{4.25}$$

Note that this constant τ_{max} depends on the step size h.

Lemma 4.5.1. *Let* $w = [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$*. Furthermore let* $0 \leq \tau < \tau_{max}$*. Then*

$$\lambda_1 \left(\theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} + \tau L_h^{\mathcal{B},\mathcal{J}} \cdot D_1 \right) > 0$$

and

$$\det \left(\theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} + \tau L_h^{\mathcal{B},\mathcal{J}} \cdot D_1 \right) > 0 \,.$$

For a given step size h one can actually calculate the maximal time step size τ_{max}. Note that the estimations in this section are on the one hand very crude but on the other hand do not require any additional knowledge about the PDE matrix L_h. If we calculate τ_{max} for an actual numerical example it would probably be much smaller then what is actually possible. That means that one could choose τ a lot bigger without having to worry about the nonsingularity of $\theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} + \tau L_h^{\mathcal{B},\mathcal{J}} \cdot D_1$.

With this Lemma and Lemma 4.3.1 we have proven the following

Theorem 4.5.2. *Let* $w = [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$ *and let* $H^{\mathcal{B}} \in \partial_D G(w)$ *be arbitrary. Furthermore let* $0 \leq \tau < \tau_{max}$*. Then* $H^{\mathcal{B}}$ *is nonsingular, more precisely,*

$$(-1)^{(J_{min} \cdot p)(J_{mob} \cdot p + 1)} \det H^{\mathcal{B}} > 0 \,.$$

With Theorem 4.3.2 we can immediately extend this assertion to the set $\partial_E G(w)$.

Theorem 4.5.3. *Let $w = [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$ and let $H^{a,b} \in \partial_E G(w)$ be arbitrary. Furthermore let $0 \leq \tau < \tau_{max}$. Then $H^{a,b}$ is nonsingular, more precisely,*

$$(-1)^{(J_{min} \cdot p)(J_{mob} \cdot p+1)} \det H^{a,b} > 0 \,.$$

Since $\partial_B G_M(w) \subset \partial_D G(w)$ and $\partial G_M(w) \subset \partial_E G(w)$ hold, we can immediately note the following corollary.

Corollary 4.5.4. *Let $w = [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$ and let $J \in \partial_B G_M(w)$ or $J \in \partial G_M(w)$ be arbitrary. Furthermore let $0 \leq \tau < \tau_{max}$. Then J is nonsingular, more precisely,*

$$(-1)^{(J_{min} \cdot p)(J_{mob} \cdot p+1)} \det J > 0 \,.$$

Finally we bring two results which are not directly linked with the nonsingularity of J but come as a byproduct of the arguments we have brought in this section. First we consider the matrix $\theta I + \tau L_h$. This matrix appears in the decoupled linear equation system (3.30). With the same argument as above we first estimate its smallest eigenvalue λ_1 as

$$\lambda_1 \left(\theta I + \tau L_h\right) \geq \theta - \tau \, ||L_h||_{sp}$$

then we can conclude that $\det\left(\theta I + \tau L_h\right)$ is positive if $0 < \tau < \frac{\theta}{||L_h||_{sp}}$ holds. We have proven

Lemma 4.5.5. *Let $0 \leq \tau < \frac{\theta}{||L_h||_{sp}}$. Then $\theta I + \tau L_h$ is nonsingular, more precisely*

$$\det\left(\theta I + \tau L_h\right) > 0 \,.$$

Note that $\tau_{max} \leq \frac{\theta}{||L_h||_{sp}}$ holds. So for τ in $0 \leq \tau < \tau_{max}$ both $\theta I + \tau L_h$ and K are nonsingular. In fact if the discretization for L_h is done with finite differences then one can easily see with Gersgorin's Theorem that $\theta I + \tau L_h$ must be always nonsingular (the absolute value of the center element in each row is bigger then the sum of the absolute values of the other elements). With similar arguments we are now able to give upper bounds for the condition number of K.

Theorem 4.5.6. *Let $0 \leq \tau < \tau_{max}$. Then the condition number κ_2 of $\theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} + \tau L_h^{\mathcal{B},\mathcal{J}} \cdot D_1$ with respect to the spectral norm is bounded by*

$$\kappa_2 \left(\theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}} + \tau L_h^{\mathcal{B},\mathcal{J}} \cdot D_1\right) \leq \frac{\theta + \tau \cdot \rho}{\theta - \tau \cdot \rho}$$

with $\rho = ||L_h||_{sp} \cdot \sqrt{1 + s^2}$ and s is the upper bound for D_1 from Lemma 6.4.1.

Proof. Let $z \in \mathbb{R}^{|\mathcal{B}|}$ with $||z||_2 = 1$. Then we have with similar arguments as above

$$\begin{aligned}
||K \cdot z||_2 &= \left\| \left(\theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}}\right) \cdot z + \tau \cdot L_h^{\mathcal{B},\mathcal{J}} \cdot D_1 \cdot z \right\|_2 \\
&\leq ||\theta z||_2 + \left\| \tau L_h^{\mathcal{B},\mathcal{B}} \cdot z + \tau \cdot L_h^{\mathcal{B},\mathcal{J}} \cdot D_1 \cdot z \right\|_2 \\
&= \theta + \left\| \left[\tau L_h^{\mathcal{B},\mathcal{B}} \mid \tau L_h^{\mathcal{B},\mathcal{J}}\right] \cdot \begin{bmatrix} I \\ D_1 \end{bmatrix} z \right\|_2 \\
&\leq \theta + \left\| \left[\tau \cdot L_h^{\mathcal{B},\mathcal{B}} \mid \tau \cdot L_h^{\mathcal{B},\mathcal{J}}\right] \right\|_{sp} \cdot \left\| \begin{bmatrix} I \\ D_1 \end{bmatrix} \right\|_{sp} \\
&\leq \theta + \tau \cdot ||L_h||_{sp} \cdot \sqrt{1 + s^2} \,.
\end{aligned}$$

The estimation chain in the other direction goes like above

$$\begin{aligned}
\|K \cdot z\|_2 &= \left\|\left(\theta I_{|\mathcal{B}|} + \tau L_h^{\mathcal{B},\mathcal{B}}\right) \cdot z + \tau \cdot L_h^{\mathcal{B},\mathcal{J}} \cdot D_1 \cdot z\right\|_2 \\
&= \left\|\theta z - \tau\left(-L_h^{\mathcal{B},\mathcal{B}} \cdot z - L_h^{\mathcal{B},\mathcal{J}} \cdot D_1 \cdot z\right)\right\|_2 \\
&\geq \left|\|\theta z\|_2 - \tau\left\|L_h^{\mathcal{B},\mathcal{B}} \cdot z + L_h^{\mathcal{B},\mathcal{J}} \cdot D_1 \cdot z\right\|_2\right| \\
&\geq \theta - \tau\left\|L_h^{\mathcal{B},\mathcal{B}} \cdot z + L_h^{\mathcal{B},\mathcal{J}} \cdot D_1 \cdot z\right\|_2 \\
&= \theta - \tau\left\|\left[L_h^{\mathcal{B},\mathcal{B}} \mid L_h^{\mathcal{B},\mathcal{J}}\right]\right\|_{sp} \cdot \left\|\begin{bmatrix} I \\ D_1 \end{bmatrix}\right\|_{sp} \\
&\geq \theta - \tau \cdot \|L_h\|_{sp} \cdot \sqrt{1+s^2}\,.
\end{aligned}$$

For $\tau < \tau_{max}$ is $\theta - \tau \cdot \|L_h\|_{sp} \cdot \sqrt{1+s^2}$ positive as we have already seen above. The condition number with respect to the spectral norm for a nonsingular square matrix A is defined as $\kappa_2(A) = \|A\|_{sp} \|A^{-1}\|_{sp}$, see [17, p. 336]. With an easy transformation one can see that $\|A^{-1}\|_{sp} = 1/(\min_{\|x\|_2=1} \|Ax\|_2)$ holds. Then we can estimate the condition number of K with respect to the spectral norm as

$$\kappa_2(K) = \frac{\max_{\|z\|_2=1} \|K \cdot z\|_2}{\min_{\|z\|_2=1} \|K \cdot z\|_2} \leq \frac{\theta + \tau\rho}{\theta - \tau\rho}\,.$$

□

Remark 4.5.7. We will shortly discuss a globalization strategy for the minimum formulation. Let $0 \leq \tau < \tau_{max}$. We define the function $F_M : \mathcal{D} \longrightarrow \mathbb{R}$, $F_M(w) := \frac{1}{2}\|G_M(w)\|^2$. This is the merit function that is to be minimized in this approach. Obviously a global minimizer w^* with $F_M(w^*) = 0$ is also a solution of $G_M(w) = 0$.

The function $x \mapsto \frac{1}{2}\|x\|_2^2$ is continuously differentiable. With Lemma 2.2.5 we can conclude

$$\partial F_M(w) = G_M(w)^T \cdot \partial G_M(w)\,.$$

A point $w^* \in \mathcal{D}$ is called Clarke-stationary point of F_M if and only if $0 \in \partial F_M(w^*)$. Then the following statement holds: If w^* is a Clarke-stationary of F_M then w^* is a global minimizer of F_M with

$$F_M(w^*) = 0 \quad \text{and} \quad G_M(w^*) = 0\,.$$

We verify this statement. If there is $0 \in \partial F_M(w^*)$ then there exists a matrix $H \in \partial G_M(w^*)$ such that $H^T G_M(w^*) = 0$. But H is nonsingular according to Corollary 4.5.4. Therefore $G_M(w^*) = 0$ holds. And from the definition of F_M also $F_M(w^*) = 0$ holds.

4.6 Schur Complement Approach

In this section, we want to discuss how the linear system (4.11) can be transformed in such a way that it can be solved even more efficiently. We continue the transformation that was started in Section 4.2. To this end we utilize a Schur complement approach. We begin by introducing some abbreviations to keep the formulas clear:

$$E := \left(\theta I_{|\mathcal{A}|} + \tau L_h^{\mathcal{A},\mathcal{A}}\right) , \qquad A := [A_1 \mid 0] := \left[\tau L_h^{\mathcal{A},\mathcal{I}} \mid 0\right] ,$$

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

With these abbreviations (4.11) reads

$$\tilde{J} \cdot \begin{bmatrix} \Delta\xi^{\mathcal{A}}_{min} \\ \hline \Delta\xi^{\mathcal{I}}_{min} \\ \Delta\xi_{mob} \end{bmatrix} = - \begin{bmatrix} G^{\mathcal{A}}_1 - \bar{c}^{\mathcal{A}} \\ \hline -\tilde{E}_{\mathcal{I}} \\ G_3 \end{bmatrix} , \tag{4.26}$$

where

$$\begin{bmatrix} E & A \\ C & B \end{bmatrix} = \tilde{J} \tag{4.27}$$

from Section 4.2. We begin by writing this linear system in detail

$$E \cdot \Delta\xi^{\mathcal{A}}_{min} + A_1 \cdot \Delta\xi^{\mathcal{I}}_{min} = -G^{\mathcal{A}}_1 + \bar{c}^{\mathcal{A}} , \tag{4.28}$$

$$C_1 \cdot \Delta\xi^{\mathcal{A}}_{min} + B_{11} \cdot \Delta\xi^{\mathcal{I}}_{min} + B_{12} \cdot \Delta\xi_{mob} = \tilde{E}_{\mathcal{I}} , \tag{4.29}$$

$$C_2 \cdot \Delta\xi^{\mathcal{A}}_{min} + B_{21} \cdot \Delta\xi^{\mathcal{I}}_{min} + B_{22} \cdot \Delta\xi_{mob} = -G_3 . \tag{4.30}$$

Similar to the previous section B_{11} is a block diagonal matrix, where each block has the form $\left(S^1_{min,\mathcal{I}}\right)^T \Lambda_c \left(S^1_{min,\mathcal{I}}\right)$. Likewise B_{22} is a block diagonal matrix, where each block has the form $\left(S^1_{mob}\right)^T \Lambda_c \left(S^1_{mob}\right)$. Recall that $S^1_{min,\mathcal{I}}$ and S^1_{mob} have full column rank and that $\Lambda_c = \operatorname{diag}\left(\frac{1}{\mathrm{c}_1}, \frac{1}{\mathrm{c}_2}, \ldots, \frac{1}{\mathrm{c}_I}\right)$. All c_i are positive, because J, from which $\tilde{J}$ was derived, is in $\partial_B G_M(w)$ with $w \in \mathcal{D}$ (cf. Section 4.2). Hence B_{11} and B_{22} are positive definite and therefore nonsingular.

We now rewrite (4.29) to obtain

$$B_{11} \cdot \Delta\xi^{\mathcal{I}}_{min} = \tilde{E}_{\mathcal{I}} - B_{12} \cdot \Delta\xi_{mob} - C_1 \cdot \Delta\xi^{\mathcal{A}}_{min} . \tag{4.31}$$

Furthermore, we transform (4.30) into

$$\Delta\xi_{mob} = -\left(B_{22}\right)^{-1} \cdot G_3 - \left(B_{22}\right)^{-1} \cdot C_2 \cdot \Delta\xi^{\mathcal{A}}_{min} - \left(B_{22}\right)^{-1} \cdot B_{21} \cdot \Delta\xi^{\mathcal{I}}_{min} . \tag{4.32}$$

Now we insert $\Delta\xi_{mob}$ into (4.31) and get

$$\Delta\xi^{\mathcal{I}}_{min} = B_s^{-1}\tilde{E}_{\mathcal{I}} + \tilde{D}^{-1}B_{12}B_{22}^{-1} \cdot G_3 - B_s^{-1}\left(C_1 - B_{12} \cdot B_{22}^{-1} \cdot C_2\right) \cdot \Delta\xi^{\mathcal{A}}_{min} \tag{4.33}$$

with $B_s := \left(B_{11} - B_{12}B_{22}^{-1}B_{21}\right)$. The matrix B_s can be obtained from B through a block Gauss elimination step. It is a Schur complement of B. Since B is positive definite, B_s is also positive definite, cf. [49]. In particular B_s is nonsingular.

Finally, we insert $\Delta\xi^{\mathcal{I}}_{min}$ in (4.28) and obtain

$$(E - A_1 \cdot B_s^{-1} C_s) \cdot \Delta\xi^{\mathcal{A}}_{min} = -G_1^{\mathcal{A}} + \bar{c}^{\mathcal{A}} - A_1 B_s^{-1} \tilde{E}_{\mathcal{I}} - A_1 B_s^{-1} B_{12} \cdot B_{22}^{-1} \cdot G_3, \tag{4.34}$$

with $C_s := \left(C_1 - D_{12} D_{22}^{-1} C_2\right)$. Please note that $B_s^{-1} C_s$ equals D_1 introduced in (4.20).

To obtain the solution of the initial linear system (4.28)–(4.30), we first solve (4.34) for $\Delta\xi^{\mathcal{A}}_{min}$. Subsequently, we compute $\Delta\xi^{\mathcal{I}}_{min}$ from (4.33) which essentially requires some matrix-vector multiplications. Finally, we get $\Delta\xi_{mob}$ from (4.32) again by matrix-vector multiplications and additions.

The main computational cost is, on the one hand, in solving the linear system (4.34) and, on the other hand, in the computation of the inverses needed in (4.32)–(4.34).

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

First we solve the linear system

$$B_{22} \cdot [X_1 \mid X_2 \mid x_3] = [B_{21} \mid C_2 \mid G_3] \ . \tag{4.35}$$

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

Now we compute

$$B_s = B_{11} - B_{12} \cdot X_1, \quad C_s = C_1 - D_{12} \cdot X_2, \quad z_3 := B_{12} \cdot x_3.$$

Again this can be done block-wise. Therefore B_s and C_s have block diagonal form, too.

Next we solve the linear system

$$B_s \cdot [Y_1 \mid y_2 \mid y_3] = \left[C_s \mid z_3 \mid \tilde{E}_{\mathcal{I}}\right] . \tag{4.36}$$

For this system the same applies as for the previous one. Here C_s and B_s have a matching block diagonal form. Therefore Y_1 is a block diagonal matrix, whereas z_3 and $\tilde{E}_{\mathcal{I}}$ are just vectors. Again the block linear systems have multiple right-hand sides. This time, however, the square blocks of B_s have variable sizes from 0×0 to $J_{min} \times J_{min}$. Since B_s and its blocks are positive definite one can solve this linear systems efficiently with Cholesky decompositions.

Using this notation, our transformed system reads

$$\begin{aligned}
(E - A_1 \cdot Y_1) \cdot \Delta\xi^{\mathcal{A}}_{min} &= -G_1^{\mathcal{A}} + \bar{c}^{\mathcal{A}} - A_1 \cdot [y_2 + y_3] && (4.37)\\
\Delta\xi^{\mathcal{I}}_{min} &= y_2 + y_3 - Y_1 \cdot \Delta\xi^{\mathcal{A}}_{min} && (4.38)\\
\Delta\xi_{mob} &= -x_3 - X_1 \cdot \Delta\xi^{\mathcal{I}}_{min} - X_2 \cdot \Delta\xi^{\mathcal{A}}_{min} \ . && (4.39)
\end{aligned}$$

Through this transformation of the original system (4.26)–(4.27), we could exploit especially the structure of B and its submatrices, which would have been unused otherwise.

Since A_1 is sparse and Y_1 is block diagonal, the product $A_1 \cdot Y_1$ again is sparse. Its structure is similar to the structure of E. Therefore, the matrix $E - A_1 \cdot Y_1$ in the linear system (4.37) is sparse, too. It can be solved by a linear solver like GMRES.

One advantage of this Schur complement approach is size reduction. While the cost of computing the Schur complement is not big. In some real world applications the concentrations of species can differ over many powers of ten. So the blocks $(S^1_{mob})^T \Lambda_c (S^1_{mob})$ of B_{22} and B_{22} itself can be very ill conditioned. In Section 4.3 we have seen that B can be orthogonally transformed into a block diagonal matrix where the blocks look like $[S^1_{min,\mathcal{I}} \mid S^1_{mob}]^T \Lambda_c [S^1_{min,\mathcal{I}} \mid S^1_{mob}]$. That means that B and its Schur complement B_s are potentially (very) ill conditioned, too. For a non-iterative solver the condition number certainly does not affect the computational cost (but it affects the accuracy of the result). If B is ill conditioned then the whole matrix $\tilde{J}$ in (4.27) is likely ill conditioned, too. And its condition number depends on the vector of concentrations c. So solving (4.26) with an iterative linear solver like GMRES can be very costly. According to Theorem 4.5.6 the condition number of $E - A_1 \cdot Y_1 = \theta I_{|\mathcal{A}|} + \tau L_h^{\mathcal{A},\mathcal{A}} + \tau L_h^{\mathcal{A},\mathcal{I}} \cdot D_1$ is bounded independently of c if τ is chosen appropriately. Numerical tests even suggest that the condition number of $\theta I_{|\mathcal{A}|} + \tau L_h^{\mathcal{A},\mathcal{A}} + \tau L_h^{\mathcal{A},\mathcal{I}} \cdot D_1$ is nearly as good as the condition number of $\theta I + \tau L_h$. This is the other advantage of this Schur complement approach for solving the emerging linear equation systems. This effect can be seen in the numerical example in Section 4.8.

It should be mentioned that we really have only one semismooth Newton algorithm and that is the one which was introduced in Algorithm 4.2.1. The Schur complement approach and the simplifications in (4.11) and (4.12) are only different ways to solve the resulting linear systems efficiently.

4.7 Existence and Uniqueness of a Local Solution

In this section we want to show that the nonlinear equation system (3.31)-(3.33) together with the decoupled η-system (3.30) has locally a unique solution. Here it is necessary to comprise the decoupled η linear system into the whole equation system, because the nonlinear equation system depends upon η through the function $c = (\xi_{min}, \xi_{mob}, \eta)$. We want to show existence with Clarke's Implicit Function Theorem. To this end we first have to extend the function G_M, in order to write the whole equation system with one function. Therefore we define

$$\begin{aligned} F(\tau, \eta, \xi_{min}, \xi_{mob}, \bar{c}) \; &:= \begin{bmatrix} \theta\eta + \tau L_h \eta - \theta\eta^{old} \\ \theta\xi_{min} + \bar{c} + \tau L_h \xi_{min} - \theta\xi_{min}^{old} - \bar{c}^{old} \\ -\varphi_M\left(E\left(c\left(\xi_{min}, \xi_{mob}, \eta\right)\right), \bar{c}\right) \\ Q_{mob}\left(c\left(\xi_{min}, \xi_{mob}, \eta\right)\right) \end{bmatrix} \\ &=: \begin{bmatrix} F_1(\tau, \eta) \\ F_2(\tau, \xi_{min}, \bar{c}) \\ F_3(\eta, \xi_{min}, \xi_{mob}, \bar{c}) \\ F_4(\eta, \xi_{min}, \xi_{mob}) \end{bmatrix}. \end{aligned}$$

As domain of F we set $\mathcal{D}_F := \mathbb{R} \times \mathcal{P} \times \mathbb{R}^{J_{min}p}$ with

$$\begin{aligned}\mathcal{P} : \; = \; & \Big\{ [\eta, \xi_{min}, \xi_{mob}] \in \mathbb{R}^{(I-J)p} \times \mathbb{R}^{J_{min}p} \times \mathbb{R}^{J_{mob}p} \mid \\ & \qquad c\,(\xi_{min}(x), \xi_{mob}(x), \eta(x)) > 0 \ \forall x \in \Omega_h \Big\}\end{aligned}$$

and its values are in $\mathcal{S} := \mathbb{R}^{(I-J)p} \times \mathbb{R}^{J_{min}p} \times \mathbb{R}^{J_{min}p} \times \mathbb{R}^{J_{mob}p}$ (reminder: $p = |\Omega_h|$). Then solving the equation system (3.30)-(3.33) is equivalent to finding solutions of

$$F\,(\tau, \eta, \xi_{min}, \xi_{mob}, \bar{c}) = 0 \tag{4.40}$$

in $\mathcal{D}_F$. We assume in this section that $(\eta^{old}, \xi_{min}^{old}, \xi_{mob}^{old}) \in \mathcal{P}$ holds according to our general assumption that $c > 0$ holds. Then we already know the trivial solution

$$F\left(0, \eta^{old}, \xi_{min}^{old}, \xi_{mob}^{old}, \bar{c}^{old}\right) = 0\,.$$

In our semismooth Newton iteration applied to G_M this solution $\xi_{min}^{old}, \xi_{mob}^{old}, \bar{c}^{old}$ would be the solution from the previous time step. And η^{old} would be the solution of the decoupled linear system in the previous time step.

The component functions F_1 and F_2 are linear in all variables and therefore they are C^2-functions. So they are strongly semismooth (cf. Corollary 2.3.4). The functions E and Q_{mob} are C^2-functions on $\mathcal{P}$. So they are strongly semismooth too. In Example 2.3.6 it was shown that φ_M is strongly semismooth. Thanks to the chain rule Theorem 2.3.7 we can conclude that F_3 is strongly semismooth. And finally with Lemma 2.3.2 one can conclude that F is strongly semismooth on $\mathcal{D}_F$. This implies that F is locally Lipschitz continuous.

The component function F_3 is the same function as G_2 from G_M (except that we excluded the fact that G_2 also depends on η). The other functions F_1, F_2 and F_4 are continuously differentiable. With the same arguments as for G_M in Section 4.1 we can conclude that the B-subdifferential of F is a cross product of the B-subdifferentials of all its components. With Lemma 4.1.4 this implies that

$$\partial F(\tau, \eta, \xi_{min}, \xi_{mob}, \bar{c}) = \partial_C F(\tau, \eta, \xi_{min}, \xi_{mob}, \bar{c})\,.$$

The elements of the generalized Jacobian of F look like

$$\begin{bmatrix} L_h\eta^{old} & \theta I + \tau L_h & 0 & 0 & 0 \\ L_h\xi_{min}^{old} & 0 & \theta I + \tau L_h & 0 & I \\ 0 & -T_1 & -T_2 & -T_3 & -T_4 \\ 0 & \frac{\partial Q_{mob}}{\partial \eta} & \frac{\partial Q_{mob}}{\partial \xi_{min}} & \frac{\partial Q_{mob}}{\partial \xi_{mob}} & 0 \end{bmatrix}, \tag{4.41}$$

where $-[T_1 \mid T_2 \mid T_3 \mid T_4]$ is an element of the generalized Jacobian of the function $(\eta, \xi_{min}, \xi_{mob}, \bar{c}) \mapsto \varphi_M\,(E\,(c\,(\xi_{min}, \xi_{mob}, \eta))\,, \bar{c})$. Then the block matrix $-\,[T_2 \mid T_3 \mid T_4]$ is an element of $\partial G_2(\xi_{min}, \xi_{mob}, \bar{c})$. And the submatrix

$$\begin{bmatrix} \theta I + \tau L_h & 0 & I \\ -T_2 & -T_3 & -T_4 \\ \frac{\partial Q_{mob}}{\partial \xi_{min}} & \frac{\partial Q_{mob}}{\partial \xi_{mob}} & 0 \end{bmatrix}$$

is an element of $\partial G_M(\xi_{min}, \xi_{mob}, \bar{c})$.

With

$$\tilde{F}_\tau : \mathcal{P} \times \mathbb{R}^{J_{min}p} \longrightarrow \mathcal{S},\ \tilde{F}_\tau(\eta, \xi_{min}, \xi_{mob}, \bar{c}) := F(\tau, \eta, \xi_{min}, \xi_{mob}, \bar{c})$$

we denote the restriction of F for a fixed τ. Since F is continuously differentiable in τ, erasing the first column in (4.41) gives an element of $\partial\tilde{F}_\tau(\eta, \xi_{min}, \xi_{mob}, \bar{c})$. Then the projection

$$\pi_z \partial F(\tau, z),\ z = (\eta, \xi_{min}, \xi_{mob}, \bar{c})$$

that Clarke defines in [7, Section 7.1] coincides with $\partial\tilde{F}_\tau(\eta, \xi_{min}, \xi_{mob}, \bar{c})$, because the first column of the elements of $\partial F(\tau, z)$ is always the same. For local uniqueness we must verify that all elements in $\pi_z \partial F(0, z)$ are nonsingular. An element

$$P := \begin{bmatrix} \theta I + \tau L_h & 0 & 0 & 0 \\ 0 & \theta I + \tau L_h & 0 & I \\ -T_1 & -T_2 & -T_3 & -T_4 \\ \frac{\partial Q_{mob}}{\partial \eta} & \frac{\partial Q_{mob}}{\partial \xi_{min}} & \frac{\partial Q_{mob}}{\partial \xi_{mob}} & 0 \end{bmatrix}$$

of $\partial\tilde{F}_\tau(\eta, \xi_{min}, \xi_{mob}, \bar{c})$ is nonsingular if

$$\theta I + \tau L_h$$

and

$$J := \begin{bmatrix} \theta I + \tau L_h & 0 & I \\ -T_2 & -T_3 & -T_4 \\ \frac{\partial Q_{mob}}{\partial \xi_{min}} & \frac{\partial Q_{mob}}{\partial \xi_{mob}} & 0 \end{bmatrix}$$

are nonsingular. We have already seen that J is an element of $\partial G_M(\xi_{min}, \xi_{mob}, \bar{c})$. Thanks to Corollary 4.5.4 and Lemma 4.5.5 are $\theta I + \tau L_h$ and J nonsingular for $0 \leq \tau < \tau_{max}$ (remember $\tau_{max} > 0$ was defined in (4.25)). We have proven the following Lemma.

Lemma 4.7.1. *Let $\tau \in [0, \tau_{max})$ and $z = (\eta, \xi_{min}, \xi_{mob}, \bar{c}) \in \mathcal{P} \times \mathbb{R}^{J_{min}p}$. Then every element in*

$$\partial\tilde{F}_\tau(z) = \pi_z \partial F(\tau, z)$$

is nonsingular.

Let $\hat{z} := (\eta^{old}, \xi^{old}_{min}, \xi^{old}_{mob}, \bar{c}^{old}) \in \mathcal{P} \times \mathbb{R}^{J_{min}p}$. This implies that $c\left(\eta^{old}, \xi^{old}_{min}, \xi^{old}_{mob}\right) > 0$ holds. Then all elements in $\pi_z \partial F(0, \hat{z})$ are nonsingular and Clarke's Implicit Function Theorem [7, Corollary in Section 7.1] gives a neighborhood U of 0 and a locally Lipschitz continuous function $g : U \longrightarrow \mathcal{P} \times \mathbb{R}^{J_{min}p}$ such that

$$F(\tau, g(\tau)) = 0$$

holds for all $\tau \in U$ and $g(0) = (\eta^{old}, \xi^{old}_{min}, \xi^{old}_{mob}, \bar{c}^{old})$. Strictly speaking Clarke's theorem needs F to be defined on the whole space $\mathbb{R}^{(I+J_{min})p}$ and not only on the open subset $\mathcal{P} \times \mathbb{R}^{J_{min}p}$. But the matter is a local property and by reducing U we can ensure that $g(\tau)$ stays in $\mathcal{P} \times \mathbb{R}^{J_{min}p}$. We are only interested in positive time steps. Therefore let $\tau_s > 0$ be chosen maximal such that $[0, \tau_s) \subset U$ holds and $g(\tau)$ is in $\mathcal{P} \times \mathbb{R}^{J_{min}p}$. We have proven the local existence of a solution.

Theorem 4.7.2. *Let* $\left(\eta^{old}, \xi^{old}_{min}, \xi^{old}_{mob}\right) \in \mathcal{P}$ *and* $\bar{c}^{old} \in \mathbb{R}$*. Then there is a* $\tau_s > 0$ *and a function* $g : [0, \tau_s) \longrightarrow \mathcal{P} \times \mathbb{R}^{J_{min}p}$ *such that*

$$F(\tau, g(\tau)) = 0\,,\ \forall \tau \in [0, \tau_s)$$

and

$$g(0) = (\eta^{old}, \xi^{old}_{min}, \xi^{old}_{mob}, \bar{c}^{old})$$

hold.

And finally we consider uniqueness of a solution $g(\tau)$. Let $\tau \geq 0$ and $\tau < \max\{\tau_s, \tau_{max}\}$ be arbitrary and fixed. From the previous theorem we know that $\tilde{F}_\tau(g(\tau)) = 0$ holds and from Lemma 4.7.1 we know that all elements in $\partial \tilde{F}_\tau(g(\tau))$ are nonsingular. Application of Clarke's Inverse Functions Theorem [7, Theorem 7.1.1] yields a neighborhood V of $g(\tau)$ such that

$$\tilde{F}_\tau(z) = 0,\ z \in V$$

is only fulfilled for $z = g(\tau)$. We note this in the next theorem.

Theorem 4.7.3. *Let* $\left(\eta^{old}, \xi^{old}_{min}, \xi^{old}_{mob}\right) \in \mathcal{P}$ *and* $\bar{c}^{old} \in \mathbb{R}$ *and let* $\tau \in [0, \max\{\tau_s, \tau_{max}\})$*. Then there is a neighborhood* V *of* $g(\tau)$ *such that*

$$F(\tau, z) = 0\,,\ z \in V$$

holds only for $z = g(\tau)$*.*

4.8 Numerical Example

The reactive transport problem introduced in Chapter 3 and formulated with the minimum function was implemented in two versions using MATLAB®. One version uses the Schur-complement approach from Section 4.6, whereas the other version utilizes the whole system (4.7). We will refer to the first version as MinSchur algorithm and to the second version as MinFull algorithm.

For both versions, the discretization of the PDE-operator was done with the standard finite difference scheme of second order on a regular Cartesian mesh. Both versions have to solve the same a priori linear decoupled system, the discretization of (3.25). This is done through a GMRES iteration in both implementations, since it is a sparse system. In practice, this seems to work very well for this particular linear system. Usually only 2 or 3 steps are needed to calculate a sufficiently accurate solution. Thus we will focus on the Newton iteration.

In our test example (taken from [25]), the interaction of CO_2 with minerals is considered. In these days, we are facing the global warming of the earth which is at least partly due to the CO_2-concentration in the atmosphere. Therefore, techniques have been investigated to inject CO_2 into the subsurface. The long term storage of CO_2 beneath the surface of our planet is the desired goal. This might be more likely if the carbon precipitates would form minerals than the carbon being dissolved in the ground water.

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

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

It consists of 3 minerals (calcite and mineral B are carbonates, mineral A is a silicate) and 6 species which are dissolved in the ground water and one aqueous tracer. More details and insights for this example, especially its internal functionality, can be found in [25, Subsection 4.5.2].

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

In the following calculation, we set the spatial and the time step size to $h = \tau = 0.1$. With this setting, we get 6100 grid nodes for an equidistant quadratic grid. The discretization was done via a second-order finite difference method. With the MinSchur algorithm we calculate the resulting concentrations for the 10 species for 3600 time steps, i.e. a time span of 360 seconds. The results have been checked to match the results from [25].

In Table 4.1 we compare the linear systems which arise in these two algorithms from Newton's method. Both of these sparse systems are solved with the GMRES(50) method. The numbers in the last two columns show the total number of inner GMRES iterations which are needed in both algorithms. The fifth and sixth columns display the condition numbers of the linear systems of both algorithms. Finally, we present in the third and fourth columns the dimensions of these linear systems. Of course, the linear system of the MinFull algorithm has always the same size, since the arising Jacobians always stem from the same function. While the linear system of the MinSchur approach is not the Jacobian of G_M itself but only a reordered submatrix, whose size depends on the size of the active set.

In this table, we have only listed three time steps since the displayed tendencies always remain unchanged. The linear system in the MinSchur algorithm is almost always four times smaller than the linear system in the MinFull algorithm (in the number of rows and in the number of columns). Furthermore, its condition number is usually smaller than 3,

time step	iteration	size MinSchur	size MinFull	cond. MinSchur	cond. MinFull	MinSchur GMRes iterations	MinFull GMRes iterations
1	0	9760	42700	2.3899	3981.27	14	271
	1	10175	42700	2.3813	3981.27	15	134
	2	10213	42700	2.3906	3981.27	15	90
	3	10219	42700	2.3906	3981.27	15	75
	4	10219	42700	2.3906	3981.27	15	
	5	10219		2.3906			
2	0	9817	42700	2.4212	3981.27	14	217
	1	10223	42700	2.3906	3981.27	15	139
	2	10226	42700	2.3906	3981.27	13	64
	3	10226	42700	2.3906	3981.27	15	109
	4	10226	42700	2.3906	3981.27		
3	0	9840	42700	2.3906	3981.27	14	185
	1	10231	42700	2.3902	3981.27	15	142
	2	10233	42700	2.3902	3981.27	14	89
	3	10233	42700	2.3902	3981.26	15	138
	4	10233	42700	2.3902	3981.27		

Table 4.1: comparison of the arising linear systems

while the condition number of the full Jacobian in the MinFull implementation is typically more than 1000 times greater. This is a numerical confirmation for the boundedness of the condition number of the Schur-complement matrix in the MinSchur algorithm predicted in Theorem 4.5.6. The last two columns show that the MinFull algorithm needs much more total GMRES iterations than the MinSchur algorithm especially in the starting iterations for the first linear system in each time step. Overall, it is therefore not surprising that the running time for the MinSchur algorithm is much faster (by a factor of about 8 for the current discretization) than for the MinFull algorithm.

Table 4.2 shows the quadratic convergence for both algorithms of our Newton-type methods as predicted in the previous theory. Deviations in the last Newton step are probably due to the machine accuracy of about $2.2 \cdot 10^{-16}$. The third column contains the errors of the MinSchur algorithm, whereas the fourth column gives the errors of the MinFull algorithm. The good consistency of these errors shows that these two algorithms realize the same Newton method where only the linear systems are solved differently. Usually these two algorithms need the same number of Newton iterations to get below the termination condition of 10^{-11} with reference to the maximum norm. With time step size $\tau = 0.1$, they both need almost always only three Newton iterations after about 10 time iterations.

Figures 4.1–4.3 visualize the numerical results (the graphics are compressed by a factor

time step	iteration	method Schur:$\|\|G(z)\|\|_\infty$	method full:$\|\|G(z)\|\|_\infty$
1	0	$4.323041 \cdot 10^{-1}$	$4.323041 \cdot 10^{-1}$
	1	$4.907447 \cdot 10^{-1}$	$2.509660 \cdot 10^{-2}$
	2	$3.061458 \cdot 10^{-3}$	$1.122566 \cdot 10^{-3}$
	3	$5.328815 \cdot 10^{-4}$	$4.223114 \cdot 10^{-7}$
	4	$9.517796 \cdot 10^{-8}$	$6.039613 \cdot 10^{-14}$
	5	$3.108624 \cdot 10^{-15}$	
2	0	$3.004508 \cdot 10^{-1}$	$3.004508 \cdot 10^{-1}$
	1	$5.582860 \cdot 10^{-3}$	$5.582860 \cdot 10^{-3}$
	2	$1.247748 \cdot 10^{-4}$	$1.247748 \cdot 10^{-4}$
	3	$5.335036 \cdot 10^{-9}$	$5.335037 \cdot 10^{-9}$
	4	$2.664535 \cdot 10^{-15}$	$2.664535 \cdot 10^{-15}$
3	0	$2.215680 \cdot 10^{-1}$	$2.215680 \cdot 10^{-1}$
	1	$3.743785 \cdot 10^{-3}$	$3.743785 \cdot 10^{-3}$
	2	$1.759327 \cdot 10^{-5}$	$1.759327 \cdot 10^{-5}$
	3	$1.069300 \cdot 10^{-10}$	$1.069295 \cdot 10^{-10}$
	4	$1.776357 \cdot 10^{-15}$	$2.664535 \cdot 10^{-15}$
8	0	$9.540300 \cdot 10^{-2}$	$9.540299 \cdot 10^{-2}$
	1	$7.372864 \cdot 10^{-4}$	$7.372864 \cdot 10^{-4}$
	2	$1.771312 \cdot 10^{-7}$	$1.771312 \cdot 10^{-7}$
	3	$1.065814 \cdot 10^{-14}$	$1.065814 \cdot 10^{-14}$
18	0	$4.920261 \cdot 10^{-2}$	$4.920261 \cdot 10^{-2}$
	1	$2.126615 \cdot 10^{-4}$	$2.126614 \cdot 10^{-4}$
	2	$1.462897 \cdot 10^{-8}$	$1.462897 \cdot 10^{-8}$
	3	$2.220446 \cdot 10^{-15}$	$2.664535 \cdot 10^{-15}$

Table 4.2: Comparison of errors

1.5 in vertical direction). Note that the differences to the results given in [25] are only due to a different color scaling. There is a slow water flow in horizontal direction from the left to the right. With it enters dissolved CO_2 into the computational domain. This decreases the pH value (the negative common logarithm of the concentration of H^+ ions in the water). The water stream of low pH value dissolves Mineral A and Calcite, when it reaches those areas. Moreover, the dissolution of Mineral A leads to an immediate precipitation of Mineral B.

We also made some tests with refined discretizations and, therefore, different dimensions of the discretized problem. The numerical behavior of our (two) method(s) remains almost unchanged; the number of Newton steps is essentially fixed. This is not surprising since a mesh independence result is known for the min-function approach, cf. [15].

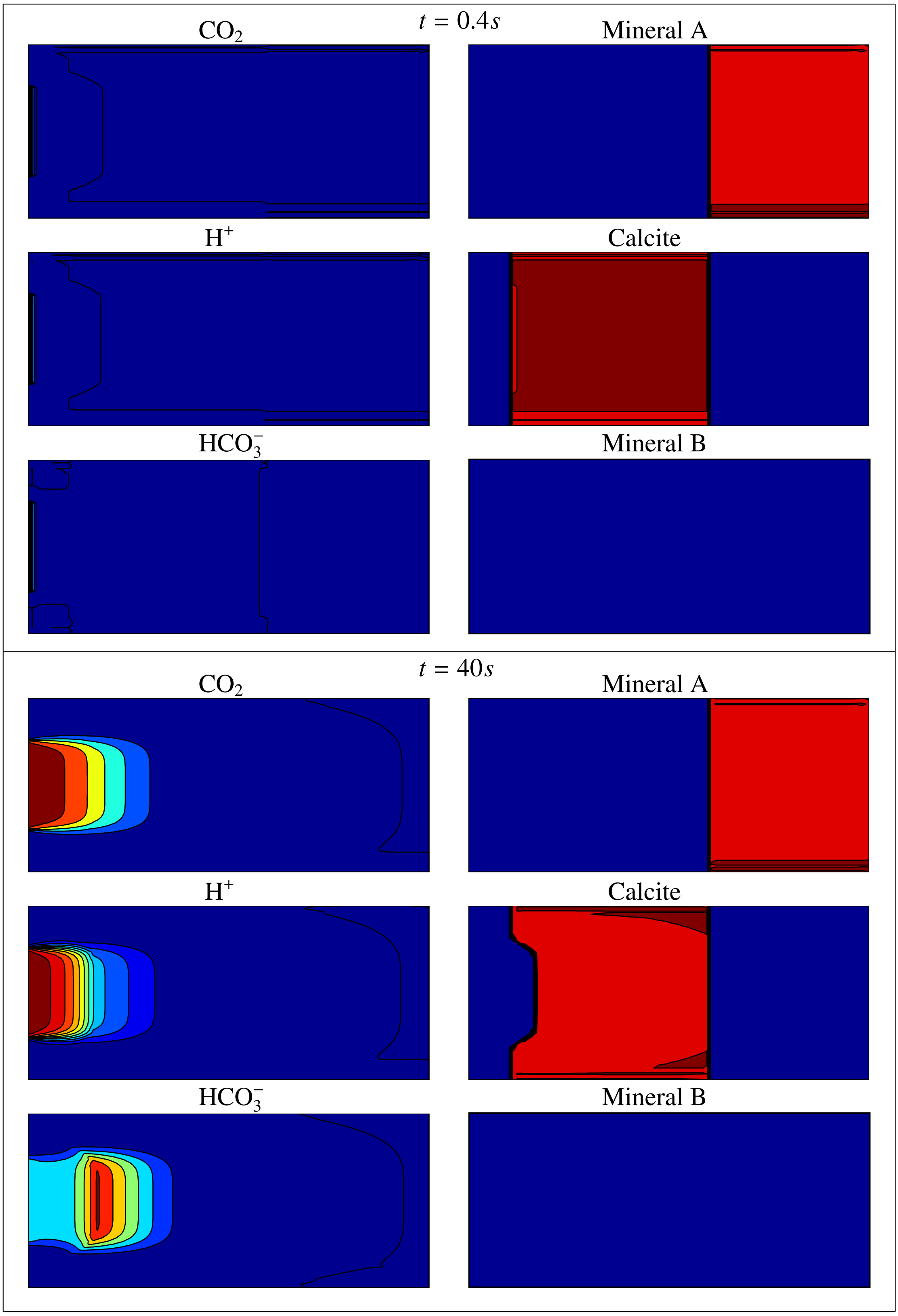

Figure 4.1: Results obtained after $t = 0.4$ seconds.

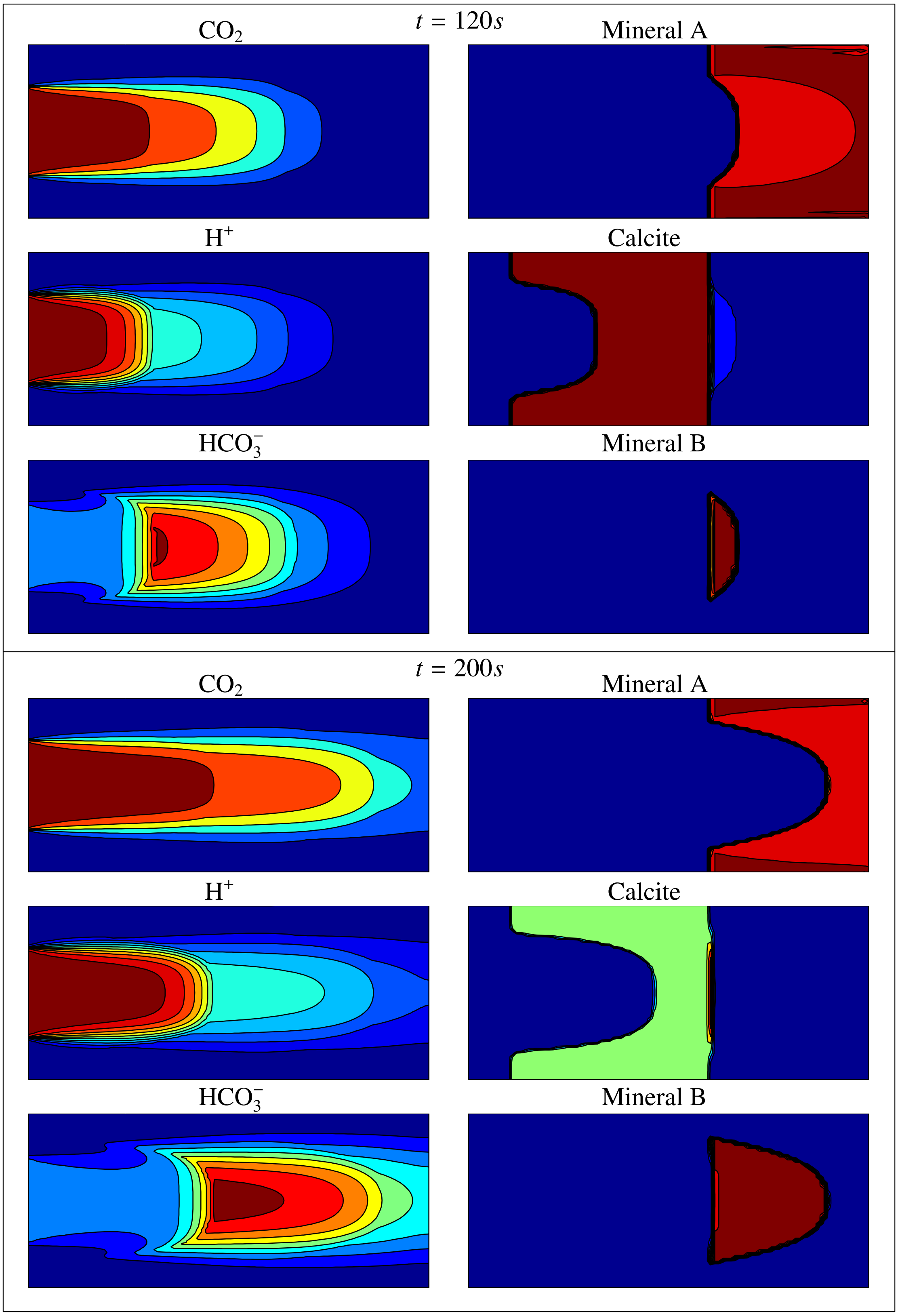

Figure 4.2: Results obtained after t = 120 seconds.

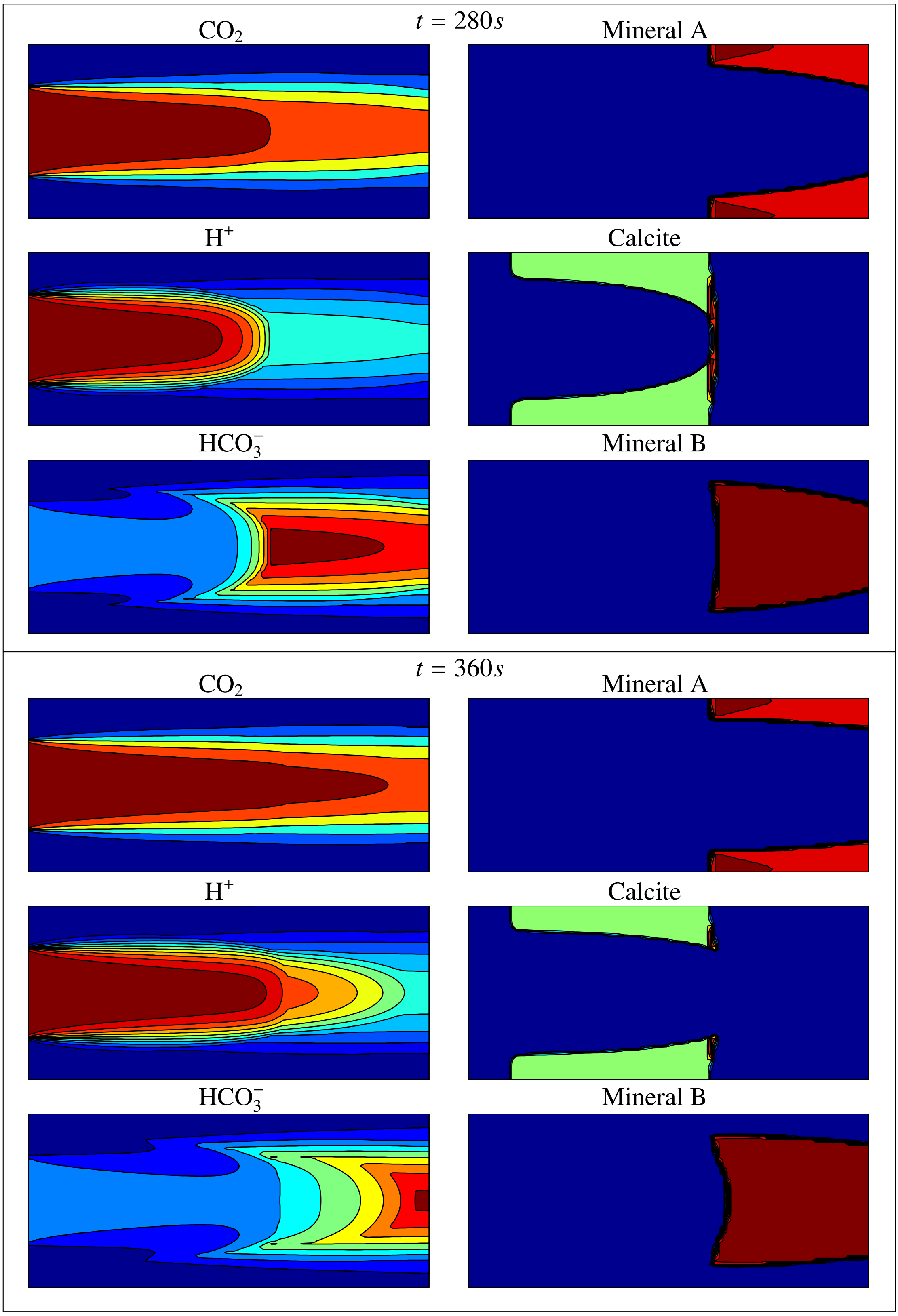

Figure 4.3: Results obtained after t = 280 seconds.

5 The Fischer-Burmeister Function Approach

In this chapter we apply the Fischer-Burmeister function as NCP-function to the system (3.31)–(3.33), i.e. we set

$$G_2\left(\xi_{min}, \xi_{mob}, \bar{c}\right) := -\varphi_F\left(\tilde{E}\left(\xi_{min}, \xi_{mob}\right), \bar{c}\right) . \tag{5.1}$$

This definition shall be valid throughout this chapter unless mentioned otherwise. The resulting total function is then denoted as

$$G_F : \mathcal{D} \longrightarrow \mathbb{R}^{(2J_{min}+J_{mob})p} .$$

Like in the previous chapter, we continue to use the notation introduced in Chapter 3, in particular $\mathcal{D}, \Omega_h, p = |\Omega_h|, \varphi_F, \varphi_M, J_{min}, J_{mob}, J$ and I. Our aim is to prove local convergence of the semismooth Newton method for solving

$$G_F(w) = 0\,,\; w \in \mathcal{D}\,, \tag{5.2}$$

and convergence for a globalized version of this algorithm. The Fischer-Burmeister function is a strongly semismooth NCP-function just like the minimum function. So many results and proofs in this chapter are similar to the results and proofs in Chapter 4. We can profit here very much from the (additional) work we did in that chapter. The reason why we consider this slightly different formulation with the Fischer-Burmeister function is that it is possible to globalize the resulting semismooth Newton method for G_F. This is not possible with the minimum function formulation.

The local existence and uniqueness results from Section 4.7 are also valid for the Fischer-Burmeister formulation G_F. This holds, because these two formulations are equivalent, i.e. for $w \in \mathcal{D}$ it holds

$$G_M(w) = 0 \qquad \Longleftrightarrow \qquad G_F(w) = 0\,.$$

So we don't state a similar result in this chapter.

The structuring in this chapter is as follows: In Section 5.1 we study the structure of the subdifferentials of G_F and show that it is strongly semismooth. In Section 5.2 we formulate the semismooth Newton method for this function, analyze the elements of the generalized Jacobian of G_F for nonsingularity and prove local convergence of this algorithm. In Section 5.3 we introduce a globalization of the semismooth Newton method for G_F and study related topics. And in the last Section 5.4 we present a numerical example.

5.1 Study of Subdifferentials of G_F

In this section we show that G_F is strongly semismooth and study its subdifferentials.

We have already seen in Section 4.1 that G_M is strongly semismooth. The proof that G_F has the same property can be copied, just replace 'minimum function' with 'Fischer-Burmeister function' and φ_M with φ_F.

Lemma 5.1.1. *The function $G_F : \mathcal{D} \longrightarrow \mathbb{R}^{(2J_{min}+J_{mob})p}$ is strongly semismooth.*

This Lemma justifies the use of the terms B-subdifferential, generalized Jacobian and C-subdifferential for our function G_F as we have defined them in Section 2.2, because they are only defined for locally Lipschitz continuous functions. And semismooth functions are locally Lipschitz continuous.

The following Lemma gives information about the structure of $\partial_B G_F$.

Lemma 5.1.2. *Let $\Omega_h = \left\{x_1, x_2, \dots, x_p\right\}$. Furthermore let $w = [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$ be arbitrary. Then the following statements hold:*

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

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

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

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

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

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

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

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

(4) Let x_i, a and b be $j \in \left\{1, \dots, \bar{I}\right\}$. Then with

$$c = \frac{\tilde{E}_j(a)}{\sqrt{\tilde{E}_j(a)^2 + b_j^2}} - 1,\ d = \frac{b_j}{\sqrt{\tilde{E}_j(a)^2 + b_j^2}} - 1$$

we have

$$\begin{aligned}
&\partial_B\varphi_F\left(\tilde{E}_j(a), b_j\right)\\
&= \begin{cases}
\left\{\left[c \cdot \frac{\partial\tilde{E}_j(a)}{\partial\xi_{min}} \mid c \cdot \frac{\partial\tilde{E}_j(a)}{\partial\xi_{mob}} \mid \ d \cdot e_l^T \ \right]\right\}, & if (\tilde{E}_j(a),\, b_j) \neq (0,0),\\
\left\{\left[\alpha \cdot \frac{\partial\tilde{E}_j(a)}{\partial\xi_{min}} \mid \alpha \cdot \frac{\partial\tilde{E}_j(a)}{\partial\xi_{mob}} \mid \ \beta \cdot e_l^T \ \right] \right. & \\
\qquad\qquad \left. :\ (\alpha,\beta) \in \partial B_1(-1,-1)\right\}, & if (\tilde{E}_j(a),\, b_j) = (0,0),
\end{cases}
\end{aligned}$$

where $e_l^T = \frac{\partial\bar{c}_j(x_i)}{\partial\bar{c}}$ is a unit vector with all components vanishing except for the component $l = i \cdot J_{min} + j$, which is one.

Proof. This proof goes like the proof of Lemma 4.1.2 except for (4). The B-subdifferential in (4) can be calculated using Example 2.2.4 and Proposition 2.2.5. □

Please note that the elements of $\partial_B\varphi_F\left(\tilde{E}_j(a), b_j\right)$ can always be written as linear combinations

$$\lambda_1 \cdot \left[\frac{\partial \tilde{E}_j(a)}{\partial \xi_{min}} \mid \frac{\partial \tilde{E}_j(a)}{\partial \xi_{mob}} \mid 0\right] + \lambda_2 \cdot \left[0 \mid 0 \mid e_l^T\right]$$

in the differentiable and the nondifferentiable case, where the scalar factors λ_1 and λ_2 are nonpositive with $[\lambda_1, \lambda_2] \neq [0, 0]$. Note also the similarity to the elements of $\partial_E G$ that were also defined with linear combinations in the corresponding rows, see (4.14) and the following equation.

Due to the fact that G_1 and G_3 are continuously differentiable we can conclude the following result.

Corollary 5.1.3. *The B-subdifferential of G_F is a cross product of the B-subdifferentials of its scalar components, i.e. with $w = [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$ we have*

$$\partial_B G_F(w) = \partial_B G^1(w) \times \partial_B G^2(w) \times \ldots \times \partial_B G^n(w)$$

where $n = (2J_{min} + J_{mob}) \cdot p$ and G^i is a scalar component function of G_F spanning over all functions G_1, G_2 and G_3.

This result simplifies the calculation of $\partial_B G_F$ very much. The generalized Jacobian of G_F is defined as convex hull of $\partial_B G_F$. And the convex hull of a cross product equals the cross product of the convex hulls of its components, see Lemma 4.1.4. As an immediate consequence of these facts we get the following result.

Corollary 5.1.4. *Let $w = [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$ be arbitrary. Then it holds*

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

Finally we would like to see how an element of $\partial_B G_F(w)$ looks like. To this end let $\omega : \{1, \ldots, J_{min} \cdot p\} \longrightarrow (\{1, \ldots, J_{min}\} \times \Omega_h)$ be the bijective enumeration function from Section 4.3. With this function we construct the matrix Q by defining its k-th row $Q(k)$. Let $(i, x) := \omega(k)$ then

$$\begin{aligned} Q(k) \;:=\; & \left[\; Q_1(k) \;\middle|\; Q_2(k) \;\middle|\; Q_3(k) \;\right] \\ & \left[\; -\alpha\frac{\partial \tilde{E}_i(x)}{\partial \xi_{min}} \;\middle|\; -\alpha\frac{\partial \tilde{E}_i(x)}{\partial \xi_{mob}} \;\middle|\; -\beta\frac{\partial \bar{c}_i(x)}{\partial \bar{c}} \;\right] \end{aligned}$$

where for differentiable points $[\tilde{E}_i(x), \bar{c}_i(x)] \neq [0, 0]$ we set

$$\alpha := \frac{\tilde{E}_i(x)}{\sqrt{\tilde{E}_i(x)^2 + \bar{c}_i(x)^2}} - 1,\; \beta := \frac{\bar{c}_i(x)}{\sqrt{\tilde{E}_i(x)^2 + \bar{c}_i(x)^2}} - 1$$

and for nondifferentiable points $[\tilde{E}_i(x), \bar{c}_i(x)] = [0, 0]$ we have

$$[\alpha,\beta] \in \partial B_1(-1, -1) = \partial_B\varphi_F(0, 0)\,. \tag{5.3}$$

Then an element of $H \in \partial_B G(w)$ looks like

$$H := \begin{bmatrix} \theta I + \tau L_h & 0 & I \\ Q_1 & Q_2 & Q_3 \\ \frac{\partial Q_{mob}}{\partial \xi_{min}} & \frac{\partial Q_{mob}}{\partial \xi_{mob}} & 0 \end{bmatrix} . \tag{5.4}$$

An element $J \in \partial G(w)$ is structured in exactly the same way only that (5.3) must be replaced with

$$[\alpha,\beta] \in \overline{B_1(-1,-1)} = \partial \varphi_F(0,0) .$$

5.2 Local Convergence and Nonsingularity

In this section we formulate the semismooth Newton method for G_F. Afterwards we study local convergence of this algorithm and nonsingularity of the elements of $\partial G_F(w)$.

The semismooth Newton algorithm for G_F formulated with the Fischer-Burmeister function differs in nothing from the algorithm for G_M formulated with the minimum function. The linear equation system in this algorithm is

$$H \begin{pmatrix} \Delta\xi_{min} \\ \Delta\xi_{mob} \\ \Delta\bar{c} \end{pmatrix} = - \begin{pmatrix} G_1(w) \\ G_2(w) \\ G_3(w) \end{pmatrix} , \tag{5.5}$$

with $H \in \partial_B G_F(w)$ and $w = [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$. This semismooth Newton method is only well defined if all elements from $\partial_B G_F(w)$ are nonsingular, especially around a solution point. We proceed by stating the whole algorithm. Hereby we ignore the restriction $w = [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$, which is equivalent with $c(\xi_{min}, \xi_{mob}) > 0$. If we start sufficiently close to a solution in $\mathcal{D}$, the iterates will stay in $\mathcal{D}$.

Algorithm 5.2.1 (Semismooth Newton Method)**.**

(S.0) (Initialization)
Choose $w^0 = [\xi^0_{min}, \xi^0_{mob}, \bar{c}^0] \in \mathcal{D}$, $\epsilon \geq 0$ *and set* $k := 0$.

(S.1) (Termination Criterion)
If $\left\| G_F\left(w^k\right) \right\|_\infty \leq \epsilon$ *or* $w^k \notin \mathcal{D}$*, stop.*

(S.2) (Newton Direction Calculation)
Choose $H \in \partial_B G_F\left(w^k\right)$*. Find a solution* d^k *of the linear system*

$$Hd = -G_F(w^k) .$$

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

Now we deal with the nonsingularity of the elements of ∂G_F. For Algorithm 5.2.1 we need only the nonsingularity of the elements of $\partial_B G_F$. But it doesn't require more effort to study the generalized Jacobian instead. Here we can benefit from results in Chapter 4, which were too general or seemingly unnecessary at that time. Particularly the result about the nonsingularity of the elements of $\partial_E G(w)$ will be useful here. We remind the reader that $\tau_{max} > 0$ was defined in (4.25) as

$$\tau_{max} = \frac{\theta}{\|L_h\|_{sp} \cdot \sqrt{1+s^2}}$$

where $s > 0$ is a constant that depends only on the stoichiometric matrices S^1_{min} and S^1_{mob}.

Theorem 5.2.2. *Let $w = [\xi_{min}, \xi_{mob}, \bar{c}] \in \mathcal{D}$ and $J \in \partial G_F(w)$ be arbitrary. Furthermore let $0 \le \tau < \tau_{max}$. Then J is nonsingular, more precisely*

$$(-1)^{(J_{min}\cdot p)(J_{mob}\cdot p+2)} \det J > 0\,.$$

Proof. Let $J \in \partial G_F(w)$ be arbitrary. According to (5.4) we have

$$J := \begin{bmatrix} \theta I + \tau L_h & 0 & I \\ Q_1 & Q_2 & Q_3 \\ \frac{\partial Q_{mob}}{\partial \xi_{min}} & \frac{\partial Q_{mob}}{\partial \xi_{mob}} & 0 \end{bmatrix}$$

where the k-th row $Q(k)$ of $Q = [Q_1 \mid Q_2 \mid Q_3]$ looks like

$$Q(k) \;:=\; \left[\; -\alpha\frac{\partial \tilde{E}_i(x)}{\partial \xi_{min}} \;\middle|\; -\alpha\frac{\partial \tilde{E}_i(x)}{\partial \xi_{mob}} \;\middle|\; -\beta\frac{\partial \bar{c}_i(x)}{\partial \bar{c}} \;\right]$$

with coefficients $\alpha \le 0$, $\beta \le 0$ that can't vanish at the same time. In Section 4.3 we defined the elements of $\partial_E G(w)$ in a very similar way. Only the coefficients a_k, b_k were nonnegative there. The matrix Q has $J_{min}p$ rows. By multiplying every row of Q in J with (-1) we get an element $\tilde{J}$ from $\partial_E G(w)$ and it holds

$$\det \tilde{J} = (-1)^{J_{min}p} \det J\,.$$

Thanks to Theorem 4.5.3 we know that

$$(-1)^{(J_{min}\cdot p)(J_{mob}\cdot p+1)} \det \tilde{J} > 0$$

holds and therefore

$$(-1)^{(J_{min}\cdot p)(J_{mob}\cdot p+1)}(-1)^{J_{min}p} \det J > 0\,.$$

□

If we choose $\tau \in [0, \tau_{max})$ then Algorithm 5.2.1 is well-defined and we have no problems with nonsingularity even far away from a solution point. This guarantees that every point in $\mathcal{D}$ is BD-regular.

Now we state the main local convergence result for this Newton-type method.

Theorem 5.2.3. *Let $w^* := [\xi^*_{min}, \xi^*_{mob}, \bar{c}^*] \in \mathcal{D}$ be a BD-regular solution of the nonlinear system $G_F(w) = 0$. Then there exists an $\epsilon > 0$ such that for every starting point $w^0 \in B_\epsilon(w^*)$ the following assertions hold:*

(1) The Newton-type iteration defined in Algorithm 5.2.1 is well-defined and produces a sequence $\left\{w^k\right\}$ that converges to w^*.

(2) The rate of convergence is quadratic.

Proof. We have to choose $\epsilon > 0$ such that $B_\epsilon(w^*) \subset \mathcal{D}$ holds, because G_F is only defined on $\mathcal{D}$. This is possible because $\mathcal{D}$ is a convex and open set. Thanks to Lemma 5.1.1 we know that G_F is strongly semismooth. Then the assertion follows from [38, Theorem 3.1]. □

5.3 Globalization

It is well known that Newton's method is only locally convergent. In this section we discuss a minimizing approach to make it globally convergent. The idea is to minimize

$$F : \mathcal{D} \to \mathbb{R},\; F(w) := \frac{1}{2}\|G_F(w)\|_2^2$$

with $w = [\xi_{min}, \xi_{mob}, \bar{c}]$. If $w^* \in \mathcal{D}$ is a global minimum of F with $F(w^*) = 0$ then $G_F(w^*) = 0$ holds, too. Conversely if $w^* \in \mathcal{D}$ solves $G_F(w) = 0$ it is also a global minimizer of F with $F(w^*) = 0$. Any method of unrestricted optimization could be used to minimize F, where the fact that $\mathcal{D}$ is only a subset of $\mathbb{R}^{(2J_{min}+J_{mob})p}$ is simply ignored, i.e. the algorithm terminates for an unfeasible point. We will use a line search method with the Newton direction as descent direction and the Armijo rule as step size control.

First we examine F more closely. Explicitly we have

$$\begin{aligned} F(w) \;=\; & \frac{1}{2}\left\|\theta\xi_{min} + \bar{c} + \tau L_h \xi_{min} - \theta\xi_{min}^{old} - \bar{c}^{old}\right\|_2^2 \\ & + \sum_{(i,x)\in\{1,\dots,J_{min}\}\times\Omega_h} \psi\left(\tilde{E}_i\left(\xi_{min}(x), \xi_{mob}(x)\right), \bar{c}_i(x)\right) \\ & + \frac{1}{2} \sum_{(i,x)\in\{1,\dots,J_{mob}\}\times\Omega_h} \tilde{Q}_{mob,i}\left(\xi_{min}(x), \xi_{mob}(x)\right)^2 \end{aligned}$$

where ψ is just the squared Fischer-Burmeister function, i.e.

$$\psi : \mathbb{R}^2 \longrightarrow \mathbb{R},\; \psi(a,b) := \frac{1}{2}\varphi_F(a,b)^2\,. \tag{5.6}$$

For our globalization approach it is very important that ψ is continuously differentiable, because then F is also continuously differentiable on $\mathcal{D}$. This was first mentioned by Kanzow [22]. We bring the more sophisticated proof proposed by Facchinei and Soares [13].

Theorem 5.3.1. *The squared Fischer-Burmeister function ψ defined in (5.6) is continuously differentiable on the whole space $\mathbb{R}^2$. It's differential in the origin is*

$$\nabla\psi(0,0)^T = [0,0]\ .$$

Proof. With Proposition 2.2.5 it holds that $\partial\psi(0,0) = \varphi_F(0,0)\partial\varphi_F(0,0) = \{[0,0]\}$. The Fischer-Burmeister function φ_F is continuously differentiable for $[a,b] \neq [0,0]$ and so is the squared Fischer-Burmeister function ψ. With Proposition 2.2.2(b) we have that $\partial\psi(a,b) = \{\psi'(a,b)\}$ in $[a,b] \neq [0,0]$. Then the generalized gradient $\partial\psi(a,b)$ is everywhere single valued. Again thanks to Proposition 2.2.2(b) is ψ continuously differentiable on $\mathbb{R}^2$ and the only element of $\partial\psi(0,0)$ is the transposed gradient of ψ. □

The function $g(x) := \frac{1}{2}\|x\|_2^2$ is continuously differentiable and its differential is $g'(x) = x^T$. Therefore we can apply Proposition 2.2.5 to $F = g \circ G$ and we get as generalized gradient

$$\partial F(w) = G_F(w)^T \partial G_F(w), \quad w \in \mathcal{D}.$$

Since F is continuously differentiable we can conclude that

$$\partial F(w) = \{F'(w)\} = G_F(w)^T \partial G_F(w) \tag{5.7}$$

holds for $w \in \mathcal{D}$. Now we are ready for our next result.

Theorem 5.3.2. *Let $w \in \mathcal{D}$ and $\tau \in [0, \tau_{max})$. Every stationary point $w \in \mathcal{D}$ of F is a global minimizer with*

$$F(w) = 0 \quad \textit{and} \quad G_F(w) = 0\,.$$

Proof. In Theorem 5.2.2 we have seen that all elements of $\partial G_F(w)$ are nonsingular if $\tau \in [0, \tau_{max})$. With (5.7) we can conclude for a stationary point $w \in \mathcal{D}$ that

$$0 = \nabla F(w) = H^T G_F(w)$$

holds for all $H \in \partial G_F(w)$. Since H is nonsingular if follows that $G_F(w) = 0$. With the definition of F we can deduce $F(w) = 0$. □

Usually optimization methods only give local minima or stationary points. With this theorem we know that every such point is already a global minimizer and a solution of our primary problem. So we don't have to worry about local minima.

Now we formulate the globalized Newton's method.

Algorithm 5.3.3 (Globalized Semismooth Newton Method)**.**

(S.0) *(Initialisation)*
Choose $w^0 = \left(\xi^0_{min}, \xi^0_{mob}, \bar{c}^0\right) \in \mathcal{D}, \rho > 0, p > 2, \beta \in (0,1), \sigma \in (0, \frac{1}{2}), \epsilon \geq 0$ *and set* $k := 0$.

(S.1) *(Termination Criterion)*
If $\left\|F\left(w^k\right)\right\|_\infty \leq \epsilon$ *or* $w^k \notin \mathcal{D}$*, stop.*

(S.2) (Search Direction Calculation)
Choose $J_k \in \partial_B G_F\left(w^k\right)$ without restriction. Find a solution d^k of the linear system

$$J_k d = -G_F(w^k)\,. \tag{5.8}$$

If this linear system is not solvable or if the descent condition

$$\nabla F(w^k)^T d^k \leq -\rho\left\|d^k\right\|^p \tag{5.9}$$

is not satisfied, set $d^k := -\nabla F(w^k)$.

(S.3) (Line Search with Armijo rule)
Compute $t_k := \max\left\{\beta^l \mid l = 0, 1, 2, \ldots\right\}$ such that

$$F(w^k + t_k d^k) \leq F(w^k) + \sigma t_k \nabla F(w^k)^T d^k\,.$$

(S.4) (Update)
Set $w^{k+1} := w^k + t_k d^k$, $k \leftarrow k+1$, and go to (S.1).

For numerical calculations one would set ϵ as a small positive number, e.g. a multiple of the machine constant. For our theoretical analysis we assume $\epsilon = 0$. If this algorithm terminates with w^k then it is either a solution or an infeasible point. If it does not terminate, it produces an infinite sequence $\left\{w^k\right\}$. The following theorem covers this case. It is based on a paper by De Luca, Facchinei and Kanzow [10].

Theorem 5.3.4. *Let $\tau \in [0, \tau_{max})$ and let $\left\{w^k\right\}$ be a infinite sequence generated by Algorithm 5.3.3 with $\epsilon = 0$.*
Then a feasible accumulation point $w^* \in \mathcal{D}$ of this sequence is a global minimizer of F with $F(w^*) = 0$ and a solution of $G_F(w) = 0$. And it is the only limit of $\left\{w^k\right\}$, i.e. $\lim_{k\to\infty} w^k = w^*$. Furthermore

(a) Eventually d^k is always given by the solution of (5.8), i.e. there is an index k_1 such that for $k > k_1$ always $d^k \neq -\nabla F(w^k)$ holds.

(b) Eventually the step size of one is always accepted in the line search, i.e. there is an index k_2 such that for $k > k_2$ always $t_k = 1$ and consequently $w^{k+1} = w^k + d^k$ holds.

(c) The convergence rate is quadratic.

Proof. From Theorem 5.3.2 we know that every stationary point is a global minimizer of F. Thanks to Theorem 5.2.2 and the assumption $\tau \in [0, \tau_{max})$ we can conclude that the accumulation point $w^* \in \mathcal{D}$ is a BD-regular point. The inner function $\tilde{E}$ in G_2 is a C^∞-function on $\mathcal{D}$. Now all assertions follow with [10, Theorem 11]. The fact that our function G_F constitutes a mixed complementary problem and not a complementary problem as required in the paper is not important for the theorem. This can be seen in the proof of that theorem. Anyhow the block component functions G_1 and G_3 which do not relate to a complementary problem are C^∞-functions on $\mathcal{D}$. □

Remark 5.3.5. Without the condition $\tau \in [0, \tau_{max})$ the statements of this theorem still hold if w^* is a BD-regular point.

5.4 Numerical Example

The reactive transport problem from Chapter 3 formulated with the Fischer-Burmeister function was implemented in two versions using MATLAB®. Both Versions tackle the arising linear equation systems directly without any reformulation or transformation. The first version applies a line seach along the Newton direction while the second version just takes the Newton direction itself to calculate the next iterate. We will refer to the first version as FBglob algorithm and to the second version as FBloc algorithm. Both versions use the same discretization of the PDE-operator as the numerical algorithms from Section 4.8. Also they solve the decoupled linear system (for the η variables) in exactly the same way as the algorithms from Chapter 4. In the following we ignore this decoupled linear systems and focus on the Newton iteration for the remaining nonlinear system (5.2) formulated with the Fischer-Burmeister function as NCP-function.

In order to compare the results of the algorithms MinSchur and MinFull from Chapter 4 with these algorithms we consider the same example with the same constants and with the same step sizes $h = \tau = 0.1$. The algorithms FBglob and FBloc produce the same solution as the algorithms MinSchur and MinFull. The maximal difference of the solution vectors of these new algorithms and the previous algorithms is 10^{-8} per time step measured in the maximum norm. Visualizing the results from the new algorithms would produce the same pictures as in Section 4.8.

In Table 5.1 we list the Newton iterates of some typical time steps of the FBloc and FBglob algorithms. As termination criterion we chose $\|G(z)\|_\infty < 10^{-11}$. The second and third columns show the errors of both algorithms. The last column shows the step size that is accepted with the Armijo rule along the Newton direction in the FBglob algorithm. First of all we see that the Armijo step is almost always 1. After about 50 time steps deviations from the line search step 1 are very rare. This is probably due to good start values for the Newton iteration, which are listed as iteration 0 in the table. As start values we take the solution from the previous time step. This behavior is consistent with the theory, cf. Theorem 5.3.4. Since the FBglob algorithm differs from the FBloc algorithm only in the additional line search it is therefore no surprise that the iterates of both algorithms coincide almost always. Differences occur only in the first 50 time steps. This is displayed in the good consistency of the errors of both algorithms. The convergence rate for both methods is superlinear and sometimes quadratic. They need about the same number of iterations, which is usually 5 or 6 and rarely 4 or 8. It is quite clear that they are not as efficient as the algorithms MinSchur and MinFull, which clearly display quadratic convergence. Why the FBloc and FBglob algorithms only seldom show the theoretical predicted quadratic convergence is not clear. We suspect that the evaluation of the derivative of the Fischer-Burmeister function near $[0, 0]$ introduces very small errors.

Table 5.2 shows information about the linear systems that arise during the Newton iteration as subproblems in the FBloc algorithm. These numbers are of course valid for the FBglob algorithm, too. These linear systems are solved with the GMRES(50) method. The last column in this table shows the total number of inner iterations and the third column shows the condition numbers of the corresponding Jacobians. We see that the condition numbers have the same order of magnitude as for the MinFull algorithm. For the start-

time step	iteration	algo. FBloc:$\|\|G(z)\|\|_\infty$	algo. FBglob:$\|\|G(z)\|\|_\infty$	Armijo step
1	0	$6.264927 \cdot 10^{-1}$	$6.264927 \cdot 10^{-1}$	1
	1	$2.223732 \cdot 10^{-2}$	$2.223732 \cdot 10^{-2}$	1
	2	$6.133391 \cdot 10^{-4}$	$6.133391 \cdot 10^{-4}$	1
	3	$6.688679 \cdot 10^{-5}$	$6.688679 \cdot 10^{-5}$	1
	4	$4.190574 \cdot 10^{-8}$	$4.190574 \cdot 10^{-8}$	1
	5	$1.654763 \cdot 10^{-8}$	$1.654763 \cdot 10^{-8}$	1
	6	$6.615356 \cdot 10^{-10}$	$6.615356 \cdot 10^{-10}$	1
	7	$1.398322 \cdot 10^{-12}$	$1.398322 \cdot 10^{-12}$	
18	0	$4.920261 \cdot 10^{-2}$	$4.920261 \cdot 10^{-2}$	1
	1	$2.126614 \cdot 10^{-4}$	$2.126615 \cdot 10^{-4}$	1
	2	$1.071023 \cdot 10^{-6}$	$1.071021 \cdot 10^{-6}$	0.3
	3	$8.610043 \cdot 10^{-7}$	$9.508448 \cdot 10^{-7}$	1
	4	$1.024707 \cdot 10^{-7}$	$4.519288 \cdot 10^{-7}$	1
	5	$2.515772 \cdot 10^{-11}$	$1.410458 \cdot 10^{-9}$	1
	6	$1.096902 \cdot 10^{-13}$	$1.021403 \cdot 10^{-13}$	
102	0	$1.375459 \cdot 10^{-2}$	$1.375459 \cdot 10^{-2}$	1
	1	$9.300983 \cdot 10^{-4}$	$9.300983 \cdot 10^{-4}$	1
	2	$4.465934 \cdot 10^{-6}$	$4.465934 \cdot 10^{-6}$	1
	3	$3.136712 \cdot 10^{-10}$	$3.136712 \cdot 10^{-10}$	1
	4	$1.861855 \cdot 10^{-12}$	$1.861855 \cdot 10^{-12}$	
133	0	$1.271447 \cdot 10^{-2}$	$1.271447 \cdot 10^{-2}$	1
	1	$6.183039 \cdot 10^{-3}$	$6.183039 \cdot 10^{-3}$	1
	2	$8.608736 \cdot 10^{-4}$	$8.608736 \cdot 10^{-4}$	1
	3	$3.980050 \cdot 10^{-5}$	$3.980050 \cdot 10^{-5}$	1
	4	$1.016641 \cdot 10^{-7}$	$1.016641 \cdot 10^{-7}$	1
	5	$6.812328 \cdot 10^{-13}$	$6.812328 \cdot 10^{-13}$	
1230	0	$1.371732 \cdot 10^{-2}$	$1.371732 \cdot 10^{-2}$	1
	1	$9.092115 \cdot 10^{-3}$	$9.092115 \cdot 10^{-3}$	1
	2	$1.096420 \cdot 10^{-3}$	$1.096420 \cdot 10^{-3}$	1
	3	$1.950766 \cdot 10^{-4}$	$1.950766 \cdot 10^{-4}$	1
	4	$3.207753 \cdot 10^{-6}$	$3.207753 \cdot 10^{-6}$	1
	5	$1.325427 \cdot 10^{-9}$	$1.325427 \cdot 10^{-9}$	1
	6	$1.296740 \cdot 10^{-13}$	$1.296740 \cdot 10^{-13}$	

Table 5.1: Comparison of errors

ing iterations they are even worse. Therefore it is not surprising that the number of inner GMRES iterations is even greater than for the MinFull algorithm. The size of these linear systems (both columns and rows) is of course constant since the matrix from the linear systems is always the Jacobian of G_F.

time step	iteration	Condition FBloc	size full	FBloc GMRes iterations
1	0	7003.7459	42700	382
	1	5332.8800	42700	438
	2	4201.0067	42700	296
	3	3988.4142	42700	192
	4	3981.2698	42700	240
	5	3996.4149	42700	194
	6	3981.2660	42700	175
2	0	7003.7213	42700	343
	1	5328.7233	42700	484
	2	4199.4568	42700	321
	3	3988.3245	42700	180
	4	3981.2726	42700	190
	5	3981.2731	42700	235
	6	3981.2696	42700	199
3	0	7003.6752	42700	322
	1	5324.8248	42700	512
	2	4198.2968	42700	323
	3	3988.0174	42700	183
	4	3981.2704	42700	175
	5	3981.2733	42700	175
	6	3981.2703	42700	169
	7	3981.2703	42700	188

Table 5.2: Linear systems from the FBloc algorithm

Finally we used the FBloc algorithm to calculate a long term simulation for this example with some altered constants. In contrast to the previous calculations and Section 4.8 we chose $q = 0.03$, $\theta = 0.45$, $\beta_l = 0.3$ and $\beta_t = 0.4$. The step sizes remained unchanged, namely $\tau = h = 0.1$. The numerical results are displayed in Figures 5.1–5.3. The water flow from the left is twice times as fast as the water flow in the simulation from Section 4.8. But there are no substantial differences. The whole process just develops faster. In Figure 4.3 we see that Mineral B is dissolving again. The numerical results from the previous chapter only indicate this.

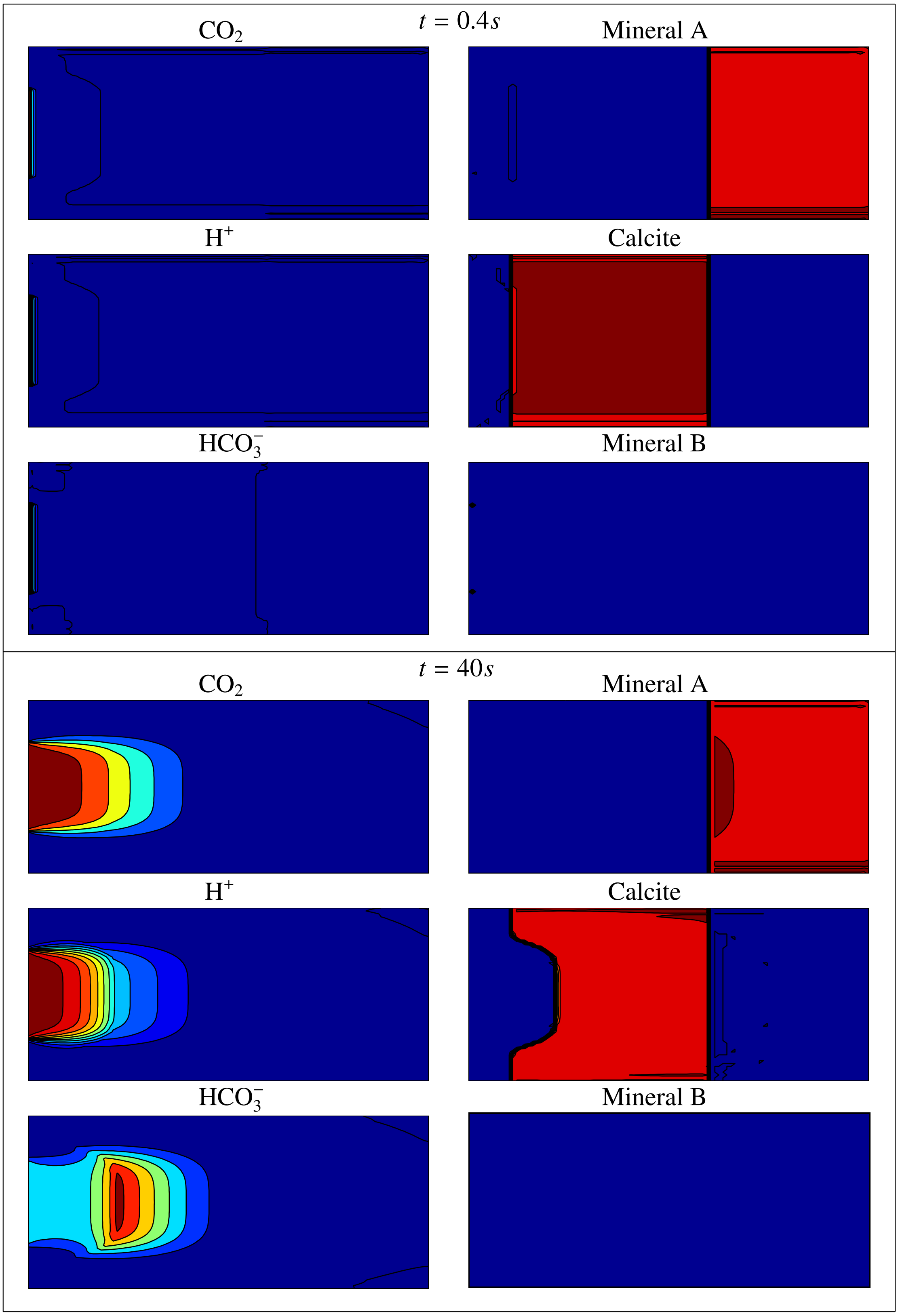

Figure 5.1: Results obtained after $t = 0.4$ seconds.

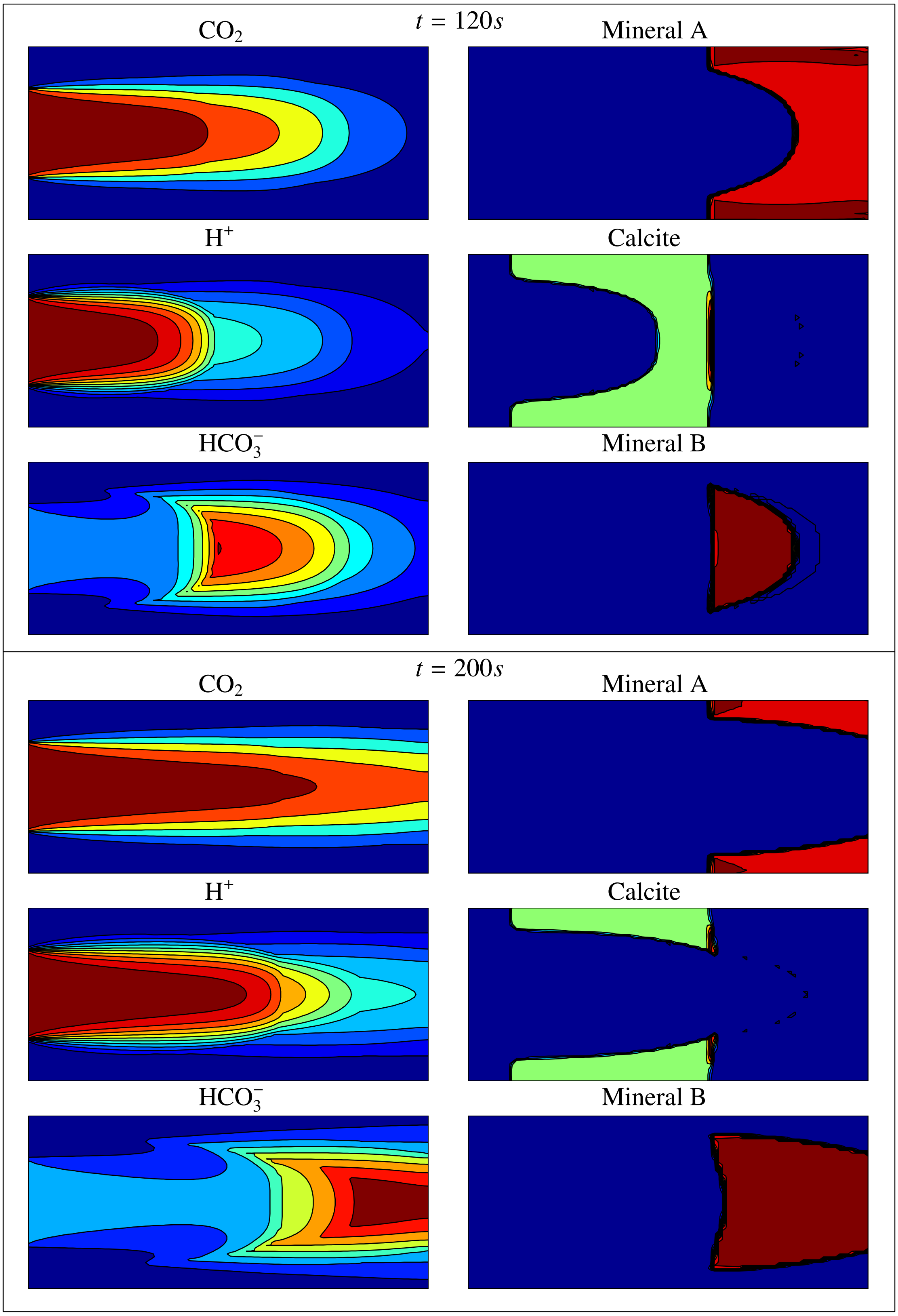

Figure 5.2: Results obtained after t = 120 seconds.

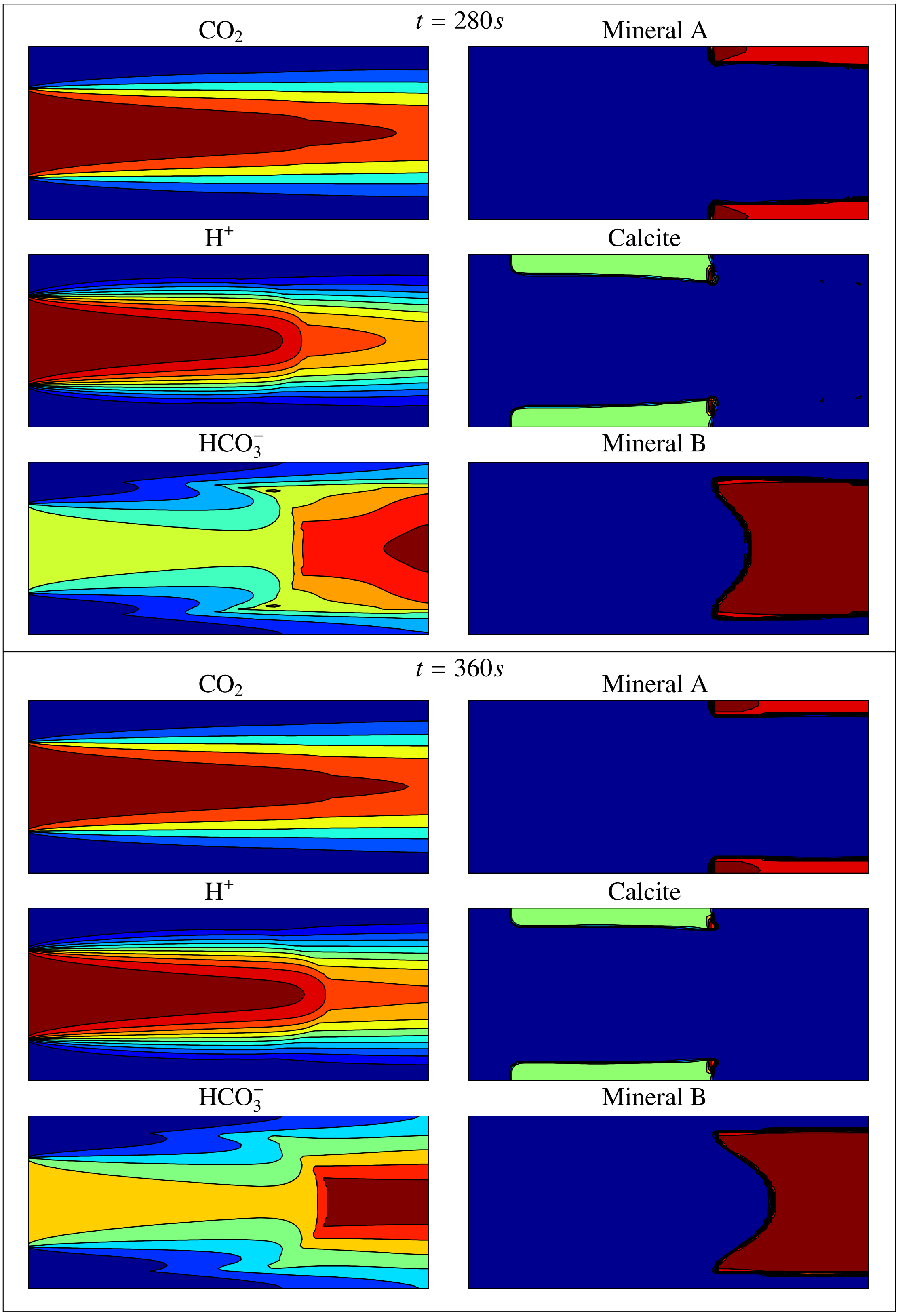

Figure 5.3: Results obtained after t = 280 seconds.

6 The Determinant Theory

In this chapter we will develop a small theory which is applied in Section 4.5. The theory is more general than we need for our problem. So it might be significant for other problems, too. First we introduce the problem of interest.

Let $n, m \in \mathbb{N}$ with $m < n$. Furthermore let $S \in \mathbb{R}^{n \times m}$, $S = (s_{i,j})$ be a matrix with full column rank, i.e. $\operatorname{rank} S = m$. Let $V \in \mathbb{R}^{n \times d}$ be an arbitrary matrix with $d \leq n$ and $\operatorname{rank} V = d$. Finally let $y \in \mathbb{R}^n$ be an arbitrary vector and $D(y) := \operatorname{diag}(y_1, y_2, \ldots, y_n)$. With this we define the matrix valued function

$$F : \mathbb{R}^n_+ \longrightarrow \mathbb{R}^{m \times d}, \quad y \mapsto \left(S^T D(y) S\right)^{-1} \cdot \left(S^T D(y) V\right) \tag{6.1}$$

where $\mathbb{R}^n_+$ denotes the set of real vectors with only positive entries. Our goal is to show that the spectral norm of $F(y)$ is bound for all $y \in \mathbb{R}^n_+$. Since all norms in a finite dimensional space are equivalent this holds for all other matrix norms, too. For our analysis of $F(y)$ during this chapter we will make intense use of the determinant and its properties. Thus the name of this chapter.

In the first section we will introduce some special notation for this chapter. And we cite some basic theorems upon which the rest of this chapter is build. In the second section we will develop a formula for the entries of $F(y)$ in Lemma-size steps. In Section 6.3 upper bounds for the entries of $F(y)$ and for some matrix norms of $F(y)$ are developed. And finally in the fourth section we will see how this theory applies to our main problem.

6.1 Preliminaries

This chapter is technically difficult. To make it as simple as possible, we use in this chapter special notation.

Let M be an arbitrary $p \times q$ matrix with $p, q \in \mathbb{N}$. With $\bar{p}$ we denote the set $\{1, 2, 3, \ldots, p\}$. Furthermore for $\mathcal{I} \subset \bar{p}$ the matrix $M_{\mathcal{I}}$ is the submatrix of M that contains only the rows in the index set $\mathcal{I}$. The ordering of these rows is preserved. For $j \in \bar{q}$ let $M^{\text{ӿ}j}$ be the submatrix of M that emerges from M by canceling its j-th column. For $i \in \bar{p}$ we denote with $M_{\text{ӿ}i}$ the matrix M without its i-th row. This notation can be combined e.g. $M^{\text{ӿ}j}_{\text{ӿ}i}$. Generally superscripts apply to columns and subscripts apply to rows. As a short form for $\mathcal{I} \setminus \{j\}$ we often write $\mathcal{I} \setminus j$, but only if we leave out just one element. Let us consider a short example. Let $\mathcal{I} \subset \bar{p}$, $i \in \mathcal{I}$ and $j \in \bar{q}$. Then $M^{\text{ӿ}j}_{\mathcal{I} \setminus i}$ denotes the submatrix of M, where the j-th column is left out and the row indices are in $\mathcal{I} \setminus \{i\}$. With $|\mathcal{I}|$ we denote the cardinal number of $\mathcal{I}$.

We state some results from linear algebra which will be important for our further investigation. The following theorem will play a crucial role. It is given here in a simplified form.

Theorem 6.1.1 (Binet-Cauchy)**.** *Let $p > q$ be positive integers. Let $A, B \in \mathbb{R}^{p \times q}$ be arbitrary matrices. Then*

$$\det\left(A^T \cdot B\right) = \sum_{\substack{\mathcal{I} \subset \bar{p} \\ |\mathcal{I}|=q}} \det\left(A_{\mathcal{I}}\right) \cdot \det\left(B_{\mathcal{I}}\right) . \tag{6.2}$$

Proof. see [17, p. 22]. □

The next result is a well known result about calculating the inverse of a square matrix.

Theorem 6.1.2. *Let A be a nonsingular $p \times p$ matrix. Then*

$$A^{-1} = \frac{1}{\det A} \cdot \tilde{A} ,$$

where $\tilde{A} = \left(\tilde{a}_{i,j}\right)_{i=1,\dots,p}^{j=1,\dots,p}$ is the adjoint of A with $\tilde{a}_{i,j} = (-1)^{i+j} \cdot \det\left(A_{\varkappa j}^{\varkappa i}\right)$.

Proof. see [17, p.20-21]. □

Finally we state the Laplace determinant expansion theorem.

Theorem 6.1.3 (Laplace expansion of the determinant)**.** *Let $A = \left(a_{i,j}\right)$ be a real $p \times p$ matrix. For $i \in \bar{p}$ it holds*

$$\det(A) = (-1)^i \sum_{k=1}^{p} (-1)^k \cdot a_{i,k} \cdot \det\left(A_{\varkappa i}^{\varkappa k}\right) \quad (\textit{expansion for row } i)$$

and for $j \in \bar{p}$ it holds

$$\det(A) = (-1)^j \sum_{k=1}^{p} (-1)^k \cdot a_{k,j} \cdot \det\left(A_{\varkappa k}^{\varkappa j}\right) \quad (\textit{expansion for column } j) .$$

Proof. see [17, p. 7]. □

6.2 Formulas for the Entries of $F(y)$

The function $y \mapsto F(y)$ is a matrix valued function. In this section we will develop direct formulas for the individual entries of $F(y)$. We do this in several Lemma-size steps.

For preparation we start by applying the Theorem of Binet-Cauchy from Section 6.1 to a matrix product. This yields a couple of very simple but useful results that will help to understand the technical proofs to come. In this section we do not need $y \in \mathbb{R}^n$ to have only positive components. But later on we have to restrict the domain of y somewhat.

Lemma 6.2.1. *Let T and R be real $p \times q$ matrices with $p > q$. Furthermore let $D(y) = diag(y_1, y_2, \dots, y_p)$ be a diagonal $p \times p$ matrix with $y \in \mathbb{R}^p$. Then*

$$\det\left(R^T D(y) T\right) = \sum_{\substack{\mathcal{I} \subset \bar{p} \\ |\mathcal{I}|=q}} \left(\left(\prod_{i \in \mathcal{I}} y_i \right) \cdot \det\left(R_{\mathcal{I}}\right) \cdot \det\left(T_{\mathcal{I}}\right) \right) .$$

Proof. Let

$$T = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_p \end{bmatrix}$$

with $t_i = \left(t_{i,1}, t_{i,2}, \ldots, t_{i,q}\right)$ the i-th row of $T = \left(t_{i,j}\right)_{i=1,\ldots,p}^{j=1,\ldots,q}$. With Theorem 6.1.1 we obtain

$$\det\left(R^T\,(D(y)T)\right) = \sum_{\mathcal{I}\subset\bar{p},\,|\mathcal{I}|=q} \det\left(R_{\mathcal{I}}\right)\cdot\det\left((D(y)\cdot T)_{\mathcal{I}}\right)\,.$$

Since

$$D(y)\cdot T = \begin{bmatrix} y_1\cdot t_1 \\ y_2\cdot t_2 \\ \vdots \\ y_p\cdot t_p \end{bmatrix},$$

it follows from the linearity of the determinant in every row that

$$\det\left((D(y)\cdot T)_{\mathcal{I}}\right) = \Big(\prod_{i\in\mathcal{I}} y_i\Big)\cdot\det\left(T_{\mathcal{I}}\right)$$

holds. Together with the second equation in this proof the assertion follows. □

From this Lemma we can easily deduce two Corollaries that we really need later.

Corollary 6.2.2. *Let S and $D(y)$ be as in (6.1). Then*

$$\det\left(S^TD(y)S\right) = \sum_{\substack{\mathcal{I}\subset\bar{n}\\|\mathcal{I}|=m}} \Big(\big(\prod_{i\in\mathcal{I}} y_i\big)\cdot\det\left(S_{\mathcal{I}}\right)^2\Big).$$

Proof. We just apply Lemma 6.2.1 with $R := S$, $T := S$, $q := m$ and $p := n$. □

Corollary 6.2.3. *Let S and $D(y)$ be as in (6.1). Then for all $i, j = 1,\ldots,m$*

$$\det\Big(\left(S^TD(y)S\right)_{\varkappa i}^{\varkappa j}\Big) = \sum_{\substack{\mathcal{I}\subset\bar{n}\\|\mathcal{I}|=m-1}} \Big(\big(\prod_{k\in\mathcal{I}} y_k\big)\cdot\det\left(S_{\mathcal{I}}^{\varkappa i}\right)\cdot\det\left(S_{\mathcal{I}}^{\varkappa j}\right)\Big).$$

Proof. It is not difficult to see that $\left(S^TD(y)S\right)_{\varkappa i}^{\varkappa j} = \left(S^{\varkappa i}\right)^T D(y)S^{\varkappa j}$ holds. Now we can apply Lemma 6.2.1 with $R := S^{\varkappa i}$, $T := S^{\varkappa j}$, $q := m-1$ and $p := n$. □

The first Corollary shows that $S^TD(y)S$ is singular if $y = 0$ holds. And the y-values where $S^TD(y)S$ is singular are roots of a polynomial. We define the complement set as

$$\mathcal{M} := \left\{y\in\mathbb{R}^n \mid S^TD(y)S \text{ nonsingular}\right\}\,. \tag{6.3}$$

If $y > 0$ or $y < 0$ then y is an element of $\mathcal{M}$.

In the next Lemma we calculate the inverse matrix of $S^TD(y)S$ using its adjoint matrix, cf. Theorem 6.1.2.

Lemma 6.2.4. *Let S and $D(y)$ be as in (6.1) and let $y \in \mathcal{M}$. Then*

$$\left(S^T D(y) S\right)^{-1} = \frac{1}{\det\left(S^T D(y) S\right)} \cdot H \,,$$

with $H = \left(h_{i,j}\right)_{i=1,\dots,m}^{j=1,\dots,m}$ the adjoint of $S^T D(y) S$ i.e.

$$h_{i,j} \;=\; (-1)^{i+j} \cdot \det\left(\left(S^T D(y) S\right)_{\not i}^{\not j}\right) = (-1)^{i+j} \cdot \det\left(\left(S^{\not i}\right)^T D(y) \left(S^{\not j}\right)\right) . \tag{6.4}$$

Proof. Since $S^T D(y) S$ is symmetric it holds $\left(S^T D(y) S\right)_{\not i}^{\not j} = \left(\left(S^T D(y) S\right)^T\right)_{\not i}^{\not j} = \left(S^T D(y) S\right)_{\not j}^{\not i}$. It is easy to see that $\left(S^T D(y) S\right)_{\not i}^{\not j} = \left(S^{\not i}\right)^T D(y) \left(S^{\not j}\right)$. Now we can apply Theorem 6.1.2 and the assertion follows immediately. □

In the following we need another special notation. Let T be a $p \times q$ matrix and $\mathcal{I} \subset \bar{q}$ a set of row indices. For $i \in \mathcal{I}$ the i-th row in T is still present in $T_{\mathcal{I}}$ but its index changes. We denote the new index with $\sigma(i)$. If there are two rows $i, j \in \mathcal{I}$ with equal entries and $i < j$ then $\sigma(i) < \sigma(j)$ shall also hold, i.e. the order of these rows in $T_{\mathcal{I}}$ is preserved. With this we have defined an obviously bijective function

$$\sigma : \mathcal{I} \longrightarrow \{1, \dots, |\mathcal{I}|\} \,,\; i \mapsto \sigma(i) \tag{6.5}$$

in a non-ambiguous way.

In the following Lemma we expand the determinant of $T_{\mathcal{I}}$ for its $\sigma(i)$-th row according to Theorem 6.1.3. Though this result is trivial, it is an important component for the further investigation.

Lemma 6.2.5. *Let $T = (t_{i,j})$ be a real $p \times q$ matrix with $p > q$. Furthermore let $\mathcal{I} \subset \bar{p}$, $|\mathcal{I}| = q$ and $i \in \mathcal{I}$. Then it holds*

$$(-1)^{\sigma(i)} \sum_{j=1}^{q} (-1)^j \det\left(T_{\mathcal{I}\setminus i}^{\not j}\right) \cdot t_{i,j} = \det T_{\mathcal{I}} \,,$$

where σ is the function defined above.

Proof. We apply Theorem 6.1.3 to expand $T_{\mathcal{I}}$ for the $\sigma(i)$-th row and obtain

$$\det T_{\mathcal{I}} = (-1)^{\sigma(i)} \sum_{j=1}^{p} (-1)^j \cdot \tilde{t}_{\sigma(i),j} \cdot \det\left(T_{\mathcal{I}}\right)_{\not\sigma(i)}^{\not j} \,.$$

Here $\tilde{t}$ is just the variable denoting the entries of $T_{\mathcal{I}}$, i.e. $T_{\mathcal{I}} = \left(\tilde{t}_{i.j}\right)$. Since the $\sigma(i)$-th row in $T_{\mathcal{I}}$ is the same as the i-th row in T, canceling the $\sigma(i)$-th row from $T_{\mathcal{I}}$ is the same as canceling the i-th row from T. Thus it holds $\left(T_{\mathcal{I}}\right)_{\not\sigma(i)} = T_{\mathcal{I}\setminus i}$. For the same reason we have $\tilde{t}_{\sigma(i),j} = t_{i,j}$. This is the assertion. □

The next Lemma is very similar. The difference is that this time the determinant of a matrix is expanded which has two identical rows. It is the square matrix

$$\begin{bmatrix} \tilde{r} \\ R_{\mathcal{J}} \end{bmatrix}$$

where $\tilde{r}$ is a row in $R_{\mathcal{J}}$. It is expanded for the first row $\tilde{r}$.

Lemma 6.2.6. *Let $R = (r_{i,j})$ be a real $p \times q$ matrix with $p > q$. Furthermore let $\mathcal{J} \subset \bar{p}$, $|\mathcal{J}| = q - 1$ and $i \in \mathcal{J}$. Then*

$$\sum_{j=1}^{q} (-1)^j \det\left(R_{\mathcal{J}}^{\not j}\right) \cdot r_{i,j} = 0\,.$$

Proof. Let $\tilde{r} := \left(r_{i,1}, r_{i,2}, \ldots, r_{i,q}\right)$ be the i-th row of R and $T := \begin{bmatrix} \tilde{r} \\ R_{\mathcal{J}} \end{bmatrix}$. Clearly T is a $q \times q$ matrix and $\det T = 0$. Now we expand T for the first row with Theorem 6.1.3 and get

$$0 = \det T = (-1)^1 \sum_{j=1}^{q} (-1)^j \cdot t_{1,j} \cdot \det\left(T_{\not 1}^{\not j}\right) = (-1) \sum_{j=1}^{q} (-1)^j \cdot r_{i,j} \cdot \det\left(R_{\mathcal{J}}^{\not j}\right).$$

Multiplying this equation by (-1) yields the result. □

For convenience we introduce an abbreviation. Let $H = \left(h_{i,j}\right)_{i=1,\ldots,m}^{j=1,\ldots,m}$ be the matrix from Lemma 6.2.4. For $y \in \mathcal{M}$ we define

$$B := H \cdot S^T \cdot D(y)\ , \ B = \left(b_{i,j}\right)_{i=1,\ldots,m}^{j=1,\ldots,n}\,. \tag{6.6}$$

According to the definition of H, it holds

$$B = \det\left(S^T D(y) S\right) \cdot \left(S^T D(y) S\right)^{-1} \cdot S^T D(y)\,.$$

By finding formulas for the entries $b_{i,j}$ we take a big step toward finding formulas for the entries of $F(y)$. This is done in the next Lemma.

Lemma 6.2.7. *Let $i \in \bar{m}$ and $j \in \bar{n}$ be arbitrary indices. Furthermore let $y \in \mathcal{M}$. Then it holds*

$$b_{i,j} \;=\; (-1)^i \cdot \sum_{\substack{\mathcal{I} \subset \bar{n} \\ |\mathcal{I}| = m}, \, j \in \mathcal{I}} (-1)^{\sigma(j)} \cdot \left(\prod_{l \in \mathcal{I}} y_l\right) \cdot \det\left(S_{\mathcal{I} \setminus j}^{\not i}\right) \cdot \det\left(S_{\mathcal{I}}\right),$$

where σ is the function defined in (6.5).

Proof. Entry (k, j) of $S^T D(y)$ is

$$\left(S^T D(y)\right)_k^j = y_j \cdot s_{j,k}\,.$$

We start by formulating the matrix multiplication $H \cdot \left(S^T D(y)\right)$ element-wise

$$
\begin{aligned}
b_{i,j} \quad &= \quad \sum_{k=1}^{m} h_{i,k} \cdot y_j \cdot s_{j,k} \\
&\overset{\text{Def. of } h_{i,k}}{=} \quad \sum_{k=1}^{m} (-1)^{i+k} \det\left(\left(S^T D(y) S\right)^{\not\ast k}_{\not\ast i}\right) \cdot y_j \cdot s_{j,k} \\
&\overset{\text{Cor. 6.2.3}}{=} \quad \sum_{k=1}^{m} (-1)^{i+k} \cdot y_j \cdot s_{j,k} \cdot \sum_{\mathcal{I}\subset\bar{n},\, |\mathcal{I}|=m-1} \left(\left(\prod_{l\in\mathcal{I}} y_l\right) \cdot \det\left(S^{\not\ast i}_{\mathcal{I}}\right) \cdot \det\left(S^{\not\ast k}_{\mathcal{I}}\right)\right) \\
&= \quad \sum_{\substack{\mathcal{I}\subset\bar{n} \\ |\mathcal{I}|=m-1}} (-1)^{i} \cdot \left(\prod_{l\in\mathcal{I}} y_l\right) \cdot y_j \cdot \det\left(S^{\not\ast i}_{\mathcal{I}}\right) \cdot \left[\sum_{k=1}^{m} s_{j,k} \cdot (-1)^k \cdot \det\left(S^{\not\ast k}_{\mathcal{I}}\right)\right],
\end{aligned}
$$

where we just reordered the finite sums in the last equation. According to Lemma 6.2.6, the last term in square brackets vanishes for $j \in \mathcal{I}$. We consider this and continue the equation chain

$$
\begin{aligned}
&= \quad \sum_{\substack{\mathcal{I}\subset\bar{n} \\ |\mathcal{I}|=m-1}, \, j\notin\mathcal{I}} (-1)^{i} \cdot \left(\prod_{l\in\mathcal{I}} y_l\right) \cdot y_j \cdot \det\left(S^{\not\ast i}_{\mathcal{I}}\right) \cdot \left(\sum_{k=1}^{m} s_{j,k} \cdot (-1)^k \cdot \det\left(S^{\not\ast k}_{\mathcal{I}}\right)\right) \\
&= \quad \sum_{\substack{\mathcal{I}\subset\bar{n} \\ |\mathcal{I}|=m}, \, j\in\mathcal{I}} (-1)^{i} \cdot \left(\prod_{l\in\mathcal{I}} y_l\right) \cdot \det\left(S^{\not\ast i}_{\mathcal{I}\setminus j}\right) \cdot \left(\sum_{k=1}^{m} s_{j,k} \cdot (-1)^k \cdot \det\left(S^{\not\ast k}_{\mathcal{I}\setminus j}\right)\right) \\
&\overset{\text{Lemma 6.2.5}}{=} \quad \sum_{\substack{\mathcal{I}\subset\bar{n} \\ |\mathcal{I}|=m}, \, j\in\mathcal{I}} (-1)^{i} \cdot \left(\prod_{l\in\mathcal{I}} y_l\right) \cdot \det\left(S^{\not\ast i}_{\mathcal{I}\setminus j}\right) \cdot \left((-1)^{\sigma(j)} \cdot \det\left(S_{\mathcal{I}}\right)\right),
\end{aligned}
$$

where σ is the function defined in (6.5) and applied to S. Now the assertion follows immediately. □

In the next lemma we go one step further.

Lemma 6.2.8. *Let $v \in \mathbb{R}^n$ and $y \in \mathcal{M}$. Then for $i \in \bar{n}$ we have*

$$
\begin{aligned}
[B \cdot v]_i \quad &= \quad \sum_{\substack{\mathcal{I}\subset\bar{n} \\ |\mathcal{I}|=m}} \left(\prod_{l\in\mathcal{I}} y_l\right) \cdot \det\left(S_{\mathcal{I}}\right) \cdot \det\left(S^{*i}_{\mathcal{I}}\right) \\
&= \quad \det\left(S^T D(y) S^{*i}\right),
\end{aligned}
$$

where S^{*i} emerges from S by replacing the i-th column of S with v.

Proof. With the preceding Lemma (and its notation) we have

$$\begin{aligned}
[B \cdot v]_i &= \sum_{j=1}^{n} b_{i,j} \cdot v_j \\
&= (-1)^i \cdot \sum_{j=1}^{n} \sum_{\substack{\mathcal{I} \subset \bar{n} \\ |\mathcal{I}|=m}, j \in \mathcal{I}} (-1)^{\sigma(j)} \cdot \Big(\prod_{l \in \mathcal{I}} y_l \Big) \cdot \det \big(S^{\not{x} i}_{\mathcal{I} \setminus j} \big) \cdot \det \left(S_{\mathcal{I}} \right) \cdot v_j \\
&\overset{(*)}{=} \sum_{\substack{\mathcal{I} \subset \bar{n} \\ |\mathcal{I}|=m}} \Big(\prod_{l \in \mathcal{I}} y_l \Big) \cdot \det \left(S_{\mathcal{I}} \right) \Big(\sum_{j \in \mathcal{I}} (-1)^i \cdot (-1)^{\sigma(j)} \cdot \det \big((S_{\mathcal{I}})^{\not{x} i}_{\not{x} \sigma(j)} \big) \cdot v_j \Big) \\
&\overset{(**)}{=} \sum_{\substack{\mathcal{I} \subset \bar{n} \\ |\mathcal{I}|=m}} \Big(\prod_{l \in \mathcal{I}} y_l \Big) \cdot \det \left(S_{\mathcal{I}} \right) \Big(\sum_{k=1}^{m} (-1)^i \cdot (-1)^k \cdot \det \big((S_{\mathcal{I}})^{\not{x} i}_{\not{x} k} \big) \cdot (v_{\mathcal{I}})_k \Big).
\end{aligned}$$

In $(*)$ we change the order of summation. We cancel the summation condition $j \in \mathcal{I}$ in the first sum and compensated for that by adding up in the second sum only indices which are in $\mathcal{I}$. The σ-index function used here applies to S. In $(**)$ we use the fact that $\sigma(j)$ denotes the index of row j of S in $S_{\mathcal{I}}$. Then we replace $\sigma(j)$ with k using the fact that σ is a bijective map. Note that for $j \in \mathcal{I}$ we have $v_j = (v_{\mathcal{I}})_{\sigma(j)}$.

Considering the definition of S^{*i} it is not difficult to see that

$$\Big(\sum_{k=1}^{m} (-1)^i \cdot (-1)^k \cdot \det \big((S_{\mathcal{I}})^{\not{x} i}_{\not{x} k} \big) \cdot (v_{\mathcal{I}})_k \Big)$$

is the expansion of $S^{*i}_{\mathcal{I}}$ for the i-th column (cf. Theorem 6.1.3). So we can resume our equation chain by

$$\begin{aligned}
[B \cdot v]_i &= \sum_{\mathcal{I} \subset \bar{n}, |\mathcal{I}|=m} \Big(\prod_{l \in \mathcal{I}} y_l \Big) \cdot \det \left(S_{\mathcal{I}} \right) \cdot \Big(\sum_{k=1}^{m} (-1)^i \cdot (-1)^k \cdot \det \Big((S_{\mathcal{I}})^{\not{x} i}_{\not{x} k} \Big) \cdot (v_{\mathcal{I}})_k \Big) \\
&= \sum_{\mathcal{I} \subset \bar{n}, |\mathcal{I}|=m} \Big(\prod_{l \in \mathcal{I}} y_l \Big) \cdot \det \left(S_{\mathcal{I}} \right) \cdot \det \Big(S^{*i}_{\mathcal{I}} \Big) \\
&= \det \Big(S^T D(y) S^{*i} \Big),
\end{aligned}$$

where the last equation follows from Lemma 6.2.1. □

Building on the previous Lemma we can formulate the main result of this section. Remember that

$$F(y) = \left(S^T D(y) S \right)^{-1} \left(S^T D(y) V \right) .$$

We simplify the notation of the next Theorem by replacing V with one of its columns v. Then

$$\left(S^T D(y) S \right)^{-1} \left(S^T D(y) v \right)$$

is a vector. By giving formulas for each component of this vector, we have formulas for each entry of $F(y)$.

Theorem 6.2.9. *Let $v \in \mathbb{R}^n$ and $y \in \mathcal{M}$. For $i \in \bar{n}$ the matrix S^{*i} is defined as in Lemma 6.2.8. Then*

$$\left[\left(S^T D(y) S\right)^{-1} \left(S^T D(y) v\right)\right]_i = \frac{\sum_{\mathcal{I} \subset \bar{n}, |\mathcal{I}|=m} \left(\prod_{l \in \mathcal{I}} y_l\right) \cdot \det\left(S_{\mathcal{I}}\right) \cdot \det\left(S^{*i}_{\mathcal{I}}\right)}{\sum_{\mathcal{I} \subset \bar{n}, |\mathcal{I}|=m} \left(\prod_{l \in \mathcal{I}} y_l\right) \cdot \det\left(S_{\mathcal{I}}\right)^2} .$$

If additionally v is in the image of S then there exists a vector $a \in \mathbb{R}^m$, which is independent of y, so that

$$\left(S^T D(y) S\right)^{-1} \left(S^T D(y) v\right) = a .$$

Proof. The first statement is just Corollary 6.2.2, Lemma 6.2.8 along with the definition of matrix B in (6.6). The second statement should be quite obvious. If $v \in \Im\mathfrak{m}(S)$ (image of S) then there exists a vector $a \in \mathbb{R}^m$ with $v = S \cdot a$. The rest is trivial. □

Remark 6.2.10. Each component of $\left(S^T D(y) S\right)^{-1} \left(S^T D(y) v\right)$ is a rational function in the variables $y_1, \ldots, y_n$ and in each product $\prod_{l \in \mathcal{I}} y_l$ the involved variables are of first power. Note that the products $\prod_{l \in \mathcal{I}} y_l$ which show up in the enumerator surely show up in the denominator as well. This is due to the common factor $\det(S_{\mathcal{I}})$. But the denominator might have more products since the factor $\det\left(S^{*i}_{\mathcal{I}}\right)$ in the enumerator can vanish, too. The coefficients of the polynomial in the denominator are always non-negative. The denominator can't be zero, since $y \in \mathcal{M}$ and $\operatorname{rank}(S) = m$ have been our prerequisites.

6.3 Bounds for $F(y)$

In this section we want to present estimates for the magnitude and the absolute magnitude of the components of $F(y)$ and for several norms of $F(y)$. In this section we really need all components of y to be positive (although some results may hold even if this is not the case), i.e. $y \in \mathbb{R}^n_+$. This implies that $y \in \mathcal{M}$ holds, where $\mathcal{M}$ is the set defined in (6.3). Our first estimate is quite straightforward.

Theorem 6.3.1. *Let $v \in \mathbb{R}^n$ and $y \in \mathbb{R}^n_+$ be arbitrary. For $i \in \bar{m}$ it holds*

$$\left|\left[\left(S^T D(y) S\right)^{-1} \left(S^T D(y) v\right)\right]_i\right| \leq \sum_{\substack{\mathcal{I} \subset \bar{n} \\ |\mathcal{I}|=m}, \det(S_{\mathcal{I}}) \neq 0} \left|\frac{\det\left(S^{*i}_{\mathcal{I}}\right)}{\det\left(S_{\mathcal{I}}\right)}\right| ,$$

where the matrix S^{*i} emerges from S by replacing the i-th column of S with v.

Proof. We use the formula from Theorem 6.2.9. Since $y_i > 0$ for all i, we get using the triangle inequality

$$\left|\sum_{\mathcal{I} \subset \bar{n}, |\mathcal{I}|=m} \left(\prod_{l \in \mathcal{I}} y_l\right) \cdot \det\left(S_{\mathcal{I}}\right) \cdot \det\left(S^{*i}_{\mathcal{I}}\right)\right| \leq \sum_{\mathcal{I} \subset \bar{n}, |\mathcal{I}|=m} \left(\prod_{l \in \mathcal{I}} y_l\right) \cdot \left|\det\left(S_{\mathcal{I}}\right)\right| \cdot \left|\det\left(S^{*i}_{\mathcal{I}}\right)\right| .$$

For the denominator we even have the equation

$$\left|\sum_{\mathcal{I}\subset\bar{n},\,|\mathcal{I}|=m}\Big(\prod_{l\in\mathcal{I}}y_l\Big)\cdot\det(S_{\mathcal{I}})^2\right| = \sum_{\mathcal{I}\subset\bar{n},\,|\mathcal{I}|=m}\Big(\prod_{l\in\mathcal{I}}y_l\Big)\cdot\det(S_{\mathcal{I}})^2 \ ,$$

because $\det(S_{\mathcal{I}})^2 \geq 0$. Then we have

$$\begin{aligned}
&\frac{\left|\sum_{\mathcal{I}\subset\bar{n},\,|\mathcal{I}|=m}(\prod_{l\in\mathcal{I}}y_l)\cdot\det(S_{\mathcal{I}})\cdot\det\left(S^{*i}_{\mathcal{I}}\right)\right|}{\left|\sum_{\mathcal{I}\subset\bar{n},\,|\mathcal{I}|=m}(\prod_{l\in\mathcal{I}}y_l)\cdot\det(S_{\mathcal{I}})^2\right|}\\
\leq\ & \sum_{\mathcal{I}\subset\bar{n},\,|\mathcal{I}|=m}\frac{(\prod_{l\in\mathcal{I}}y_l)\cdot|\det(S_{\mathcal{I}})|\cdot\left|\det\left(S^{*i}_{\mathcal{I}}\right)\right|}{\sum_{\mathcal{I}\subset\bar{n},\,|\mathcal{I}|=m}(\prod_{l\in\mathcal{I}}y_l)\cdot\det(S_{\mathcal{I}})^2}\\
\leq\ & \sum_{\substack{\mathcal{I}\subset\bar{n}\\|\mathcal{I}|=m},\det(S_{\mathcal{I}})\neq 0}\frac{(\prod_{l\in\mathcal{I}}y_l)\cdot|\det(S_{\mathcal{I}})|\cdot\left|\det\left(S^{*i}_{\mathcal{I}}\right)\right|}{(\prod_{l\in\mathcal{I}}y_l)\cdot\det(S_{\mathcal{I}})^2}\\
=\ & \sum_{\substack{\mathcal{I}\subset\bar{n}\\|\mathcal{I}|=m},\det(S_{\mathcal{I}})\neq 0}\left|\frac{\det\left(S^{*i}_{\mathcal{I}}\right)}{\det(S_{\mathcal{I}})}\right| \ .
\end{aligned}$$

□

With this result we know that every entry of $F(y)$ is bounded for all $y > 0$. Therefore every norm of $F(y)$ must be bounded, too, for $y > 0$. From the proof of this theorem it is also clear that this result cannot hold for all $y \in \mathbb{R}^n$. Because there can be y values where the denominator vanishes but the numerator does not.

Note that Theorem 6.3.1 holds even if some of the y_i equal zero as long as

$$\sum_{\mathcal{I}\subset\bar{n},\,|\mathcal{I}|=m}\Big(\prod_{l\in\mathcal{I}}y_l\Big)\cdot\det(S_{\mathcal{I}})^2$$

is positive. The bound in this theorem might be quite crude. In the next result we are refining it.

Theorem 6.3.2. *Let $v \in \mathbb{R}^n$ and $y \in \mathbb{R}^n_+$ be arbitrary. For $i \in \bar{m}$ it holds*

$$\left[\left(S^TD(y)S\right)^{-1}\left(S^TD(y)v\right)\right]_i \in [l_i, u_i] \ ,$$

with

$$\begin{aligned}
l_i &= \sum_{\substack{\mathcal{I}\subset\bar{n}\\|\mathcal{I}|=m},\det(S_{\mathcal{I}})\cdot\det\left(S^{*i}_{\mathcal{I}}\right)<0} -\left|\frac{\det\left(S^{*i}_{\mathcal{I}}\right)}{\det(S_{\mathcal{I}})}\right|\\
u_i &= \sum_{\substack{\mathcal{I}\subset\bar{n}\\|\mathcal{I}|=m},\det(S_{\mathcal{I}})\cdot\det\left(S^{*i}_{\mathcal{I}}\right)>0} \left|\frac{\det\left(S^{*i}_{\mathcal{I}}\right)}{\det(S_{\mathcal{I}})}\right|
\end{aligned}$$

and S^{*i} is as in Theorem 6.3.1.

Proof. This proof is very similar to the previous one. We just use different estimates for $\sum_{\mathcal{I}\subset\bar{n},|\mathcal{I}|=m}\left(\prod_{l\in\mathcal{I}} y_l\right)\cdot\det\left(S_{\mathcal{I}}\right)\cdot\det\left(S^{*i}_{\mathcal{I}}\right)$. By omitting the negative terms we get the simple estimate

$$\begin{aligned}
&\sum_{\mathcal{I}\subset\bar{n},|\mathcal{I}|=m}\Big(\prod_{l\in\mathcal{I}} y_l\Big)\cdot\det\left(S_{\mathcal{I}}\right)\cdot\det\left(S^{*i}_{\mathcal{I}}\right)\\
\le\ &\sum_{\substack{\mathcal{I}\subset\bar{n}\\|\mathcal{I}|=m},\det(S_{\mathcal{I}})\cdot\det\left(S^{*i}_{\mathcal{I}}\right)>0}\Big(\prod_{l\in\mathcal{I}} y_l\Big)\cdot\det\left(S_{\mathcal{I}}\right)\cdot\det\left(S^{*i}_{\mathcal{I}}\right)\\
=\ &\sum_{\substack{\mathcal{I}\subset\bar{n}\\|\mathcal{I}|=m},\det(S_{\mathcal{I}})\cdot\det\left(S^{*i}_{\mathcal{I}}\right)>0}\Big(\prod_{l\in\mathcal{I}} y_l\Big)\cdot\left|\det\left(S_{\mathcal{I}}\right)\cdot\det\left(S^{*i}_{\mathcal{I}}\right)\right|.
\end{aligned}$$

With the same technique we get the lower estimate

$$\begin{aligned}
&\sum_{\mathcal{I}\subset\bar{n},|\mathcal{I}|=m}\Big(\prod_{l\in\mathcal{I}} y_l\Big)\cdot\det\left(S_{\mathcal{I}}\right)\cdot\det\left(S^{*i}_{\mathcal{I}}\right)\\
\ge\ &\sum_{\substack{\mathcal{I}\subset\bar{n}\\|\mathcal{I}|=m},\det(S_{\mathcal{I}})\cdot\det\left(S^{*i}_{\mathcal{I}}\right)<0}\Big(\prod_{l\in\mathcal{I}} y_l\Big)\cdot\det\left(S_{\mathcal{I}}\right)\cdot\det\left(S^{*i}_{\mathcal{I}}\right)\\
=\ &\sum_{\substack{\mathcal{I}\subset\bar{n}\\|\mathcal{I}|=m},\det(S_{\mathcal{I}})\cdot\det\left(S^{*i}_{\mathcal{I}}\right)<0}\Big(\prod_{l\in\mathcal{I}} y_l\Big)\cdot(-1)\cdot\left|\det\left(S_{\mathcal{I}}\right)\cdot\det\left(S^{*i}_{\mathcal{I}}\right)\right|.
\end{aligned}$$

Since the denominator $\sum_{\mathcal{I}\subset\bar{n},|\mathcal{I}|=m}\left(\prod_{l\in\mathcal{I}} y_l\right)\cdot\det\left(S_{\mathcal{I}}\right)^2$ is always positive we obtain our assertion by using the triangle inequality. □

For the rest of this section we will develop bounds for the column-sum norm and the row-sum norm of $F(y)$ for all $y>0$. To this end we need additional notation. Remember that S is a $n\times m$ and V a $n\times d$ matrix, cf. (6.1). We write these matrices with their columns as $S=\left[s^1\mid s^2\mid\ldots\mid s^m\right]$ and $V=\left[v^1\mid\ldots\mid v^d\right]$. For $i\in\bar{m}$ and $j\in\bar{d}$ we define S^{*i*j} as the matrix that emerges from S by replacing its i-th column with the j-th column of V, i.e.

$$S^{*i*j}=[s^1\mid\ldots\mid s^{i-1}\mid v^j\mid s^{i+1}\mid\ldots\mid s^m].$$

With this notation we can formulate the next theorem.

Theorem 6.3.3. *Let $y\in\mathbb{R}^n_+$ be arbitrary. Then for the column-sum norm of $F(y)$ it holds*

$$\|F(y)\|_C\le\max_{j=1,\ldots,d}\left(\sum_{i=1}^{m}\sum_{\substack{\mathcal{I}\subset\bar{n}\\|\mathcal{I}|=m},\det(S_{\mathcal{I}})\neq0}\left|\frac{\det\left(S^{*i*j}_{\mathcal{I}}\right)}{\det\left(S_{\mathcal{I}}\right)}\right|\right)$$

and for its row-sum norm it holds

$$\|F(y)\|_R\le\max_{i=1,\ldots,m}\left(\sum_{j=1}^{d}\sum_{\substack{\mathcal{I}\subset\bar{n}\\|\mathcal{I}|=m},\det(S_{\mathcal{I}})\neq0}\left|\frac{\det\left(S^{*i*j}_{\mathcal{I}}\right)}{\det\left(S_{\mathcal{I}}\right)}\right|\right).$$

Proof. First we remember that $F(y)$ is a $m \times d$ matrix. From Theorem 6.3.1 we know that for entry (i, j) of $F(y)$ it holds

$$\left|F(y)_i^j\right| \leq \sum_{\substack{\mathcal{I}\subset\bar{n} \\ |\mathcal{I}|=m}, \det(S_\mathcal{I})\neq 0} \left|\frac{\det\left(S_\mathcal{I}^{*i*j}\right)}{\det\left(S_\mathcal{I}\right)}\right| .$$

With this we can immediately estimate the absolute sum of the i-th row of $F(y)$ by

$$\sum_{j=1}^{d}\left|F(y)_i^j\right| \leq \sum_{j=1}^{d} \sum_{\substack{\mathcal{I}\subset\bar{n} \\ |\mathcal{I}|=m}, \det(S_\mathcal{I})\neq 0} \left|\frac{\det\left(S_\mathcal{I}^{*i*j}\right)}{\det\left(S_\mathcal{I}\right)}\right|$$

and the absolute sum of the j-th column is bounded by

$$\sum_{i=1}^{m}\left|F(y)_i^j\right| \leq \sum_{i=1}^{m} \sum_{\substack{\mathcal{I}\subset\bar{n} \\ |\mathcal{I}|=m}, \det(S_\mathcal{I})\neq 0} \left|\frac{\det\left(S_\mathcal{I}^{*i*j}\right)}{\det\left(S_\mathcal{I}\right)}\right| .$$

Applying the maximum over all rows or columns, respectively, yields the assertion. □

It is not possible to deduce a formula for the upper bound of the spectral norm of $F(y)$ directly from the upper bounds of its entries. But one can estimate the spectral norm of every $p \times q$ matrix A by

$$||A||_{sp} \leq \sqrt{||A||_C \cdot ||A||_R}\,,$$

cf. Theorem A.3.1 in the appendix. Together with the previous theorem we have shown the following corollary.

Corollary 6.3.4. *Let $y \in \mathbb{R}_+^n$ be arbitrary. Then the spectral norm of $F(y)$ can be estimated by*

$$||F(y)||_{sp} \leq \sqrt{c \cdot r}$$

with

$$c = \max_{j=1,\dots,d}\left(\sum_{i=1}^{m} \sum_{\substack{\mathcal{I}\subset\bar{n} \\ |\mathcal{I}|=m}, \det(S_\mathcal{I})\neq 0} \left|\frac{\det\left(S_\mathcal{I}^{*i*j}\right)}{\det\left(S_\mathcal{I}\right)}\right|\right)$$

and

$$r = \max_{i=1,\dots,m}\left(\sum_{j=1}^{d} \sum_{\substack{\mathcal{I}\subset\bar{n} \\ |\mathcal{I}|=m}, \det(S_\mathcal{I})\neq 0} \left|\frac{\det\left(S_\mathcal{I}^{*i*j}\right)}{\det\left(S_\mathcal{I}\right)}\right|\right).$$

We stress that all these bounds are independant of the particular value $y > 0$. For arbitrary $y \in \mathbb{R}^n$ this is not possible.

Remark 6.3.5. The bounds for several matrix norms of $F(y)$ we have given here were obtained in the same way. First we replaced each entry of $F(y)$ by an upper bound of its absolute value. Then we gave upper bounds for several norms of the resulting matrix. Sharper bounds could be achieved, if one would optimize these matrix norms of $F(y)$ directly using the representation of its entries from Theorem 6.2.9. This formidable restricted optimization problem could probably be solved only numerically.

6.4 Application to the Main Problem

In this section we want to see how all this theory applies to our main problem from chapter 3. To this end we use notation from Chapters 3 and 4. This means to apply this theory to matrix D_1, which was introduced in Section 4.3. We want to show that the spectral norm of this matrix is bounded independently of the concentration vector $c > 0$. The defining equation for D_1 and D_2 is

$$\begin{bmatrix} -\frac{\partial \tilde{E}_{\mathcal{J}}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{J}}} & -\frac{\partial \tilde{E}_{\mathcal{J}}(\xi_{min},\xi_{mob})}{\partial \xi_{mob}} \\ \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{J}}} & \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{mob}} \end{bmatrix} \cdot \begin{bmatrix} D_1 \\ D_2 \end{bmatrix} = -\begin{bmatrix} -\frac{\partial \tilde{E}_{\mathcal{J}}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{B}}} \\ \frac{\partial \tilde{Q}_{mob}(\xi_{min},\xi_{mob})}{\partial \xi_{min}^{\mathcal{B}}} \end{bmatrix} \tag{6.7}$$

or with the notation from Section 4.6

$$\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \cdot \begin{bmatrix} D_1 \\ D_2 \end{bmatrix} = -\begin{bmatrix} C_1 \\ C_2 \end{bmatrix}. \tag{6.8}$$

We denote the whole matrices as B, D and C respectively. Like already mentioned in Section 4.3 there is an orthogonal permutation matrix O such that $\tilde{B} := OBO^T$ is a block diagonal matrix. Since the blocks in B_{11}, B_{12}, C_1 and in B_{21}, B_{22}, C_2 always have the same number of rows, multiplying C with O from the left hand side yields the matrix

$$\tilde{C} := O \cdot C\,,$$

with the block diagonal structure

$$\tilde{C} = \operatorname{diag}\left(\tilde{C}(x_1), \tilde{C}(x_2), \ldots, \tilde{C}(x_p)\right), \quad x_i \in \Omega_h$$

and the blocks

$$\tilde{C}(x_i) = [S^1_{min,\mathcal{J}(x_i)} \mid S^1_{mob}]^T \cdot \Lambda_{c(x_i)} \cdot S^1_{min,\mathcal{B}(x_i)}\,.$$

Remember that the blocks of $\tilde{B}$ look like

$$\tilde{B}(x_i) = [S^1_{min,\mathcal{J}(x_i)} \mid S^1_{mob}]^T \Lambda_{c(x_i)} [S^1_{min,\mathcal{J}(x_i)} \mid S^1_{mob}]\,.$$

Now we multiply (6.8) from the left hand side with O and we insert $O^T O = I$ between B and D and get

$$\tilde{B} \cdot OD = \tilde{C}\,. \tag{6.9}$$

The spectral norm of D and OD coincide since O is orthogonal. Therefore it is sufficient to estimate the spectral norm of $\tilde{D} := OD$. We already know (cf. Section 4.3) that $\tilde{B}$ is nonsingular. Therefore we can transform (6.9) equivalently into

$$\tilde{D} = \tilde{B}^{-1} \cdot \tilde{C}\,.$$

Now $\tilde{D}$ inherits the block structure of $\tilde{B}$ and $\tilde{C}$ and is therefore a block diagonal matrix $\tilde{D} = \operatorname{diag}\left(\tilde{D}(x_1), \ldots, \tilde{D}(x_p)\right)$. Each block $\tilde{D}(x_i)$ of this matrix looks like

$$\begin{aligned} \tilde{D}(x_i) &= \tilde{B}(x_i)^{-1} \cdot \tilde{C}(x_i) \\ &= \left([S^1_{min,\mathcal{J}(x_i)} \mid S^1_{mob}]^T \Lambda_{c(x_i)} [S^1_{min,\mathcal{J}(x_i)} \mid S^1_{mob}]\right)^{-1} \\ &\qquad \cdot \left([S^1_{min,\mathcal{J}(x_i)} \mid S^1_{mob}]^T \cdot \Lambda_{c(x_i)} \cdot S^1_{min,\mathcal{B}(x_i)}\right)\,. \end{aligned}$$

Utilizing the block structure of $\tilde{D}$ we can write its spectral norm with Theorem A.3.2 as

$$\left\|\tilde{D}\right\|_{sp} = \max_{i=1,\dots,p} \left\|\tilde{D}(x_i)\right\|_{sp} .$$

In summary the spectral norm of D_1 can be estimated with

$$\|D_1\|_{sp} \leq \|D\|_{sp} = \|\tilde{D}\|_{sp} = \max_{i=1,\dots,p} \|\tilde{D}(x_i)\|_{sp} .$$

For $\hat{\mathcal{J}} \subset \{1, \dots, J_{min}\}$ and $\hat{\mathcal{B}} := \{1, \dots, J_{min}\} \setminus \hat{\mathcal{J}}$ we define the matrix valued function

$$F_{\hat{\mathcal{J}}} : \mathbb{R}^I_+ \longrightarrow \mathbb{R}^{(|\hat{\mathcal{J}}|+J_{mob})\times|\hat{\mathcal{B}}|}, \; F_{\hat{\mathcal{J}}}(y) := (S^T\hat{D}(y)S)^{-1}(S^T\hat{D}(y)V)$$

with $S := [S^1_{min,\widehat{\mathcal{J}}} \mid S^1_{mob}]$, $V := S^1_{min,\widehat{\mathcal{B}}}$ and $\hat{D}(y) := \operatorname{diag}(1/y_1, 1/y_2, \dots, 1/y_I)$. Then $F_{\hat{\mathcal{J}}}(y)$ is well-defined for $y > 0$ and has the same structure as $F(y)$ defined in (6.1) although the diagonal matrix $\hat{D}(y)$ is defined differently. For $\hat{\mathcal{J}} = \mathcal{J}(x_i)$ it holds

$$F_{\hat{\mathcal{J}}}(c(x_i)) = \tilde{D}(x_i) .$$

Remember $\mathcal{J}(x_i)$ is the projection of $\mathcal{J} \subset \{1, \dots, J_{min}\} \times \Omega_h$ on x_i, i.e.

$$\mathcal{J}(x_i) = \{k \in \{1, \dots, J_{min}\} \mid (k, x_i) \in \mathcal{J}\} .$$

Applying Corollary 6.3.4 to $F_{\hat{\mathcal{J}}}(y)$ for $\widehat{\mathcal{J}} \subset \{1, \dots, J_{min}\}$ yields an upper bound $s_{\widehat{\mathcal{J}}} > 0$ such that

$$\left\|F_{\hat{\mathcal{J}}}(y)\right\|_{sp} \leq s_{\widehat{\mathcal{J}}}$$

holds for all $y > 0$. Since every block matrix $\tilde{D}(x_i)$ can be written as $F_{\mathcal{J}(x_i)}(c(x_i))$ with $\mathcal{J}(x_i) \subset \{1, \dots, J_{min}\}$ it holds

$$\max_{i=1,\dots,p} \|\tilde{D}(x_i)\|_{sp} \leq \max_{\hat{\mathcal{J}}\subset\{1,\dots,J_{min}\}} s_{\hat{\mathcal{J}}} .$$

We have proven the following Lemma.

Lemma 6.4.1. *Let $c > 0$. Then the spectral norm of D_1 is bound by*

$$\|D_1\|_{sp} \leq \max_{\widehat{\mathcal{J}}\subset\{1,\dots,J_{min}\}} s_{\widehat{\mathcal{J}}} =: s$$

where $s_{\widehat{\mathcal{J}}}$ is the upper bound for $F_{\hat{\mathcal{J}}}(y)$ from Corollary 6.3.4.

Example 6.4.2. We demonstrate the calculation of the bound for the spectral norm of $F_{\hat{\mathcal{J}}}(y)$ for only one subset $\hat{\mathcal{J}} \subset \{1, \dots, J_{min}\}$. For this we use the data from the numerical example in Section 4.8. Then we get $I = 7$, $J_{min} = 3$, $J_{mob} = 1$. We choose the subset $\hat{\mathcal{J}} = \{1, 2\}$. Remember that the rows of the stoichiometric matrix correspond to species and the columns correspond to chemical reactions, cf. Subsection 3.1.4. For the assembling of S^1_{min} and S^1_{mob}

we use the following order of the species involved: $CO_2^{(aq)}$, HCO_3^-, H^+, Ca^{2+}, Me^{3+}, SiO_2, Tracer. The order of the chemical reactions shall be as in Section 4.8. Then we get

$$S^1_{min} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ -1 & -3 & -2 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; S^1_{mob} = \begin{bmatrix} -1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

From the definition of $F_{\{1,2\}}(y)$, which is a 3×1 matrix, we get

$$S = \begin{bmatrix} 0 & 0 & -1 \\ 1 & 0 & 1 \\ -1 & -3 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; V = \begin{bmatrix} 0 \\ 1 \\ -2 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

Since V has only one column there are 3 matrices S^{*i*j}, namely S^{*1*1}, S^{*2*1}, S^{*3*1}, where S^{*3*1} coincides with S^1_{min}. For this three matrices and for S we have to calculate the determinants of all 3×3 submatrices. For each of these 4 matrices there are $\binom{7}{3} = 35$ submatrices and the same number of determinants. Since the Tracer is not involved in chemical reactions we can reduce this to $\binom{6}{3} = 20$ submatrices (the last row of S^1_{min} and S^1_{mob} vanishes). Summarizing all calculations we get

$$\begin{aligned} \sum_{\substack{\mathcal{I} \subset \bar{7} \\ |\mathcal{I}|=3}, \det(S_{\mathcal{I}}) \neq 0} \left| \frac{\det(S^{*1*1}_{\mathcal{I}})}{\det(S_{\mathcal{I}})} \right| &= 7.5 \\ \sum_{\substack{\mathcal{I} \subset \bar{7} \\ |\mathcal{I}|=3}, \det(S_{\mathcal{I}}) \neq 0} \left| \frac{\det(S^{*2*1}_{\mathcal{I}})}{\det(S_{\mathcal{I}})} \right| &= 8 \\ \sum_{\substack{\mathcal{I} \subset \bar{7} \\ |\mathcal{I}|=3}, \det(S_{\mathcal{I}}) \neq 0} \left| \frac{\det(S^{*3*1}_{\mathcal{I}})}{\det(S_{\mathcal{I}})} \right| &= 7.5\,. \end{aligned}$$

Then we get in Corollary 6.3.4

$$c = 23, \; r = 8$$

and finally

$$\left\| F_{\{1,2\}}(y) \right\|_{sp} \leq \sqrt{c \cdot r} = s_{\{1,2\}} = 13.56\,.$$

The same work has to be done for all other subsets $\hat{\mathcal{J}} \subset \{1, \ldots, J_{min}\}$ and the corresponding functions $F_{\hat{\mathcal{J}}}(y)$.

From this example we see that it is in principle easy and straightforward to calculate the bound s from Lemma 6.4.1. But we also see that it is quite some work to do this task by hand, although none of the steps involved is difficult. In Section A.4 in the appendix we have listed two MATLAB functions, which perform this task much faster.

7 The Singular Value Theory

In this chapter we introduce a theory that gives very sharp bounds to estimate the smallest and greatest eigenvalues of certain symmetric tridiagonal matrices. With this theory one can give sharp bounds for the smallest and greatest singular value of $\theta I + \tau L_h$ from the decoupled linear equation system (3.30) if the discretization was done with finite differences. This matrix also appears in the nonlinear equation system (3.36) and in the elements of the general Jacobians $\partial G_M(w)$ and $\partial G_F(w)$. With this theory one can also tell whether $\theta I + \tau L_h$ is positive definite, which depends on the choice of the step sizes h and τ. This information is important for the application of GMRES to the equation system (3.30).

In Section 7.1 we introduce the already mentioned class of symmetric tridiagonal matrices, which we study for the rest of the chapter, and some additional notation. To obtain our results, we take a closer look at these matrices and exploit heavily the particular structure using a careful analysis of the corresponding characteristic polynomial. To this end, we begin with some preliminary results in Section 7.2. The main results are contained in Section 7.3, where we give suitable bounds for the two extremal eigenvalues. We apply our results to the discretization of a partial differential equation in Section 7.4 where matrices arise that can be decomposed as a Kronecker product of tridiagonal matrices of the matrix class under investigation. The PDE we consider in this section is the same that we introduced in Section 3.2 but without a coupled ODE and without algebraic and complementary conditions. This application was, in fact, the original motivation for the theory of this chapter.

7.1 Introduction

Consider a tridiagonal matrix of the form

$$J = \begin{bmatrix} \alpha & \beta & & & & \\ \beta & \alpha & \gamma & & & \\ & \gamma & \alpha & \ddots & & \\ & & \ddots & \ddots & \gamma & \\ & & & \gamma & \alpha & \delta \\ & & & & \delta & \alpha \end{bmatrix} \in \mathbb{R}^{m \times m} \tag{7.1}$$

with given entries $\alpha, \beta, \gamma, \delta \in \mathbb{R}$. Matrices of this form arise quite frequently in many contexts, and the eigenvalues of such matrices can often be used to compute eigenvalues of more complicated matrices which arise, e.g. from the discretization of partial differential equations.

Our aim is to give sharp bounds for the smallest and largest eigenvalues of such a matrix. There is no previous treatment of this problem in literature. There exist many results for more general matrices like Gershgorin's, Ostrowski's or Brauer's Theorem (see, e.g., [14, 17, 47]) that estimate the area to which the eigenvalues belong to, however, the bounds one obtains from these results for the particular class of matrices considered here are by far too weak. The lower bound on the smallest eigenvalue for the matrix J may also be used to obtain a lower bound for the smallest singular value of a possibly nonsymmetric matrix. This lower bound seems to be much stronger than existing ones, see, e.g.,[19, 20, 37].

Notation: Given an arbitrary matrix $A \in \mathbb{R}^{m \times m}$, we denote by $A^s := \frac{1}{2}(A + A^T)$ the symmetric part of A. The singular values of A are denoted by $\sigma_i(A)$ $(i = 1, \dots, m)$ and ordered in such a way that $\sigma_1(A) \leq \sigma_2(A) \leq \ldots \leq \sigma_m(A)$, in particular, $\sigma_1(A)$ and $\sigma_m(A)$ denote the smallest and the largest singular value of A, respectively. Similarly, given a symmetric matrix $A \in \mathbb{R}^{m \times m}$, the corresponding (real) eigenvalues are denoted by $\lambda_i(A)$ $(i = 1, \dots, m)$ and ordered in such a way that $\lambda_1(A) \leq \lambda_2(A) \leq \ldots \leq \lambda_m(A)$ so that the symbol $\lambda_1(A)$ $(\lambda_m(A))$ always stands for the smallest (largest) eigenvalue of A. We sometimes also write $\lambda_{\min}(A)$ $(\lambda_{\max}(A))$ for the smallest (largest) eigenvalue of A.

7.2 Preliminaries

Let us begin by recalling some known facts about symmetric tridiagonal matrices of the form

$$T := \begin{bmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \beta_3 & \alpha_3 & \ddots & \\ & & \ddots & \ddots & \beta_m \\ & & & \beta_m & \alpha_m \end{bmatrix} \in \mathbb{R}^{m \times m}$$

satisfying (without loss of generality) $\beta_k \neq 0$ for all $k = 2, \dots, m$. Furthermore, let

$$T_k := \begin{bmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \beta_3 & \alpha_3 & \ddots & \\ & & \ddots & \ddots & \beta_k \\ & & & \beta_k & \alpha_k \end{bmatrix} \in \mathbb{R}^{k \times k}$$

be the leading $k \times k$ principal submatrix of T, and let

$$q_k(x) := \det(T_k - xI) \qquad \forall k = 1, \dots, m$$

be the corresponding characteristic polynomial. Then the following recursion holds, cf. [14, p. 437]:

$$\begin{aligned} q_0(x) &:= 1, \\ q_1(x) &= \alpha_1 - x, \\ q_{k+1}(x) &= (\alpha_{k+1} - x) q_k(x) - \beta_{k+1}^2 q_{k-1}(x) \quad \forall k = 1, 2, \dots, m-1. \end{aligned} \tag{7.2}$$

Furthermore, the next result is also well-known, see [14, Thm. 8.4.1] or [47, Section 5.6].

Theorem 7.2.1 (Sturm Sequence Property).
Assume that $\beta_k \neq 0$ for all $k = 2, \dots, m$. Then the following statements hold:

(a) The eigenvalues of all principal submatrices T_k are real and simple.

(b) The eigenvalues of T_{k-1} strictly separate the eigenvalues of T_k in the sense that

$$\lambda_1(T_k) < \lambda_1(T_{k-1}) < \lambda_2(T_k) < \ldots < \lambda_{k-1}(T_k) < \lambda_{k-1}(T_{k-1}) < \lambda_k(T_k).$$

(c) In the Sturm sequence $\{q_0(\lambda), q_1(\lambda), \dots, q_m(\lambda)\}$ *let $w(\lambda)$ denote the number of sign changes (where we use the convention that vanishing entries $q_k(x) = 0$ are removed from this sequence before counting the sign changes). Then $w(\lambda)$ equals the number of eigenvalues of the matrix T that are strictly less than λ.*

An immediate consequence of the previous result is the following one which can be used to develop the well-known bisection method to compute single eigenvalues of symmetric tridiagonal matrices.

Corollary 7.2.2. *Let $a, b \in \mathbb{R}$ be given with $a < b$. Then $w(b) - w(a)$ is the number of eigenvalues of the symmetric tridiagonal matrix T lying in the interval $[a, b)$.*

We next want to give a lower bound for the smallest singular value of a given positive (semi-) definite (but asymmetric) matrix A in terms of the smallest eigenvalue of the corresponding symmtric part A^s. This result is a special case of [18, Cor. 3.1.5] but we give a proof anyway.

Lemma 7.2.3. *Let $A \in \mathbb{R}^{m \times m}$ be positive semidefinite (not necessarily symmetric). Then $\sigma_1(A) \geq \lambda_1(A^s) \geq 0$.*

Proof. For an arbitrary (not necessarily symmetric or positive definite) matrix A, we have

$$\min_{\|x\|=1} x^T A x = \min_{\|x\|=1} x^T A^s x = \lambda_1(A^s).$$

In particular, the assumed positive semidefiniteness of A implies the inequality $\lambda_1(A^s) \geq 0$.

In order to verify the first inequality, let us define the matrix $B := A - \lambda_1(A^s)I$. This definition implies

$$B^T + B = A^T + A - 2\lambda_1(A^s)I = 2 \cdot (A^s - \lambda_1(A^s)I).$$

Since $\lambda_1(A^s) \geq 0$ is the smallest eigenvalue of A^s, it follows that 0 is the smallest eigenvalue of B^s. The symmetry of $B^T + B$ therefore gives

$$\min_{\|x\|=1} x^T (B^T + B) x = 0.$$

Using the fact that the smallest eigenvalue of the symmetric matrix A^TA is given by $(\sigma_1(A))^2$, cf. [9, Thm. 3.3], and taking into account the definition of the matrix B, we therefore obtain

$$\begin{aligned}
(\sigma_1(A))^2 &= \min_{\|x\|=1} x^T A^T A x \\
&= \min_{\|x\|=1} \left[\lambda_1(A^s)^2 x^T x + \lambda_1(A^s) \cdot x^T \left(B^T + B\right) x + x^T B^T B x\right] \\
&= \lambda_1(A^s)^2 + \min_{\|x\|=1} \left[\lambda_1(A^s) \cdot x^T \left(B^T + B\right) x + x^T B^T B x\right] \\
&\geq \lambda_1(A^s)^2 + \lambda_1(A^s) \cdot \min_{\|x\|=1} \left[x^T \left(B^T + B\right) x\right] + \min_{\|x\|=1} \left[x^T B^T B x\right] \\
&= \lambda_1(A^s)^2 + \min_{\|x\|=1} \left[x^T B^T B x\right] \\
&= \lambda_1(A^s)^2 + (\sigma_1(B))^2 \\
&\geq \lambda_1(A^s)^2.
\end{aligned}$$

Taking the square root and using the fact that $\sigma_1(A) \geq 0$ and (as already noted) $\lambda_1(A^s) \geq 0$, we obtain the desired statement. □

Applying Gershgorin's Theorem to $\lambda_1(A^s)$ and using Lemma 7.2.3 gives the lower bound

$$\sigma_1(A) \geq \min_{i=1,\dots,n} \left\{ a_{ii} - \frac{1}{2} \sum_{\substack{j=1 \\ j \neq i}}^{m} (|a_{ij}| + |a_{ji}|) \right\}$$

for the smallest singular value of a possibly nonsymmetric matrix A which is precisely the bound given in [19, Theorem 1].

Assume, for the moment that $A \in \mathbb{R}^{m \times m}$ is symmetric positive definite. Then $\sigma_1(A) = \lambda_1(A) = \lambda_1(A^s)$, so that the inequality from Lemma 7.2.3 is actually an equality. Now, since both the singular values and the eigenvalues of A and A^s, respectively, depend continuously on the entries of the corresponding matrices (c.f. [48]), it follows that we still have $\sigma_1(A) \approx \lambda_1(A^s)$ for matrices A which are close to being symmetric, hence the estimates from Lemma 7.2.3 are likely to provide very sharp bounds in this case. Of course, this is not true for highly asymmetric matrices. However, later, in our applications, we have to deal with matrices which are close to being symmetric.

We next investigate some properties of the one-dimensional mapping

$$f : (0, \infty) \longrightarrow \mathbb{R}, \; y \mapsto (\alpha - x) - \frac{\gamma^2}{y} \tag{7.3}$$

that will play an essential role in Section 7.3. Here α, γ, and x are given, whereas y is the variable. We are particularly interested in the properties of the corresponding fixed point iteration $y_{k+1} := f(y_k)$ for $k \in \mathbb{N}$. The following result gives all the necessary information.

Lemma 7.2.4. *Let $z := \alpha - x$. Choose an initial element $y_1 > 0$ and define $y_{k+1} := f(y_k)$ recursively for $k \in \mathbb{N}$. Then the following statements hold:*

Case $z \geq 2|\gamma|$*: Here* f *has a repelling fixed point* $f_1 := \frac{z-\sqrt{z^2-4\gamma^2}}{2}$ *and an attracting fixed point* $f_2 := \frac{z+\sqrt{z^2-4\gamma^2}}{2}$ *which coincide for* $z = \pm 2|\gamma|$.

- *For* $y_1 \in (f_1, f_2)$ *we have*

  $$f_1 < y_1 < y_2 < y_3 < \ldots < y_k < y_{k+1} < \cdots < f_2$$

  for all $k \in \mathbb{N}$*. Furthermore, it holds that* $\lim_{k\to\infty} y_k = f_2$.
- *For* $y_1 > f_2$ *we have*

  $$f_2 < \ldots < y_{k+1} < y_k < \ldots < y_3 < y_2 < y_1$$

  for all $k \in \mathbb{N}$*. Furthermore, it holds that* $\lim_{k\to\infty} y_k = f_2$.
- *For* $y_1 = f_2$ *we have* $y_k = f_2$ *for all* $k \in \mathbb{N}$.
- *For* $y_1 = f_1$ *we have* $y_k = f_1$ *for all* $k \in \mathbb{N}$.
- *For* $y_1 \in (0, f_1)$ *we have*

  $$f_1 > y_1 > y_2 > y_3 > \ldots$$

  and there exists a smallest $k_0 \in \mathbb{N}$ *with* $y_{k_0} \leq 0$*. From that on, the sequence is no longer well-defined.*

Case $z < 2|\gamma|$*: Here* f *has no fixed points. We have* $y > f(y)$ *for all* $y > 0$*, and for every starting point* $y_1 > 0$*, we obtain*

$$y_1 > y_2 > y_3 > \ldots,$$

and there is a smallest $k_0 \in \mathbb{N}$ *with* $y_{k_0} \leq 0$*. From that on, the sequence is no longer well-defined.*

Instead of giving the simple proof, we illustrate this result in Figure 7.1. The left picture shows the first case where we have two (possibly identical) fixed points f_1 and f_2. When $f_1 < f_2$ (so the two fixed points do not coincide), then the derivative f' at the first fixed point is larger than one, hence this fixed point is repelling, whereas the derivative at the second fixed point is smaller than one, hence this fixed point is attracting. The right picture, on the other hand, illustrates the second case where $y > f(y)$ holds for all $y > 0$, so that no fixed points exist.

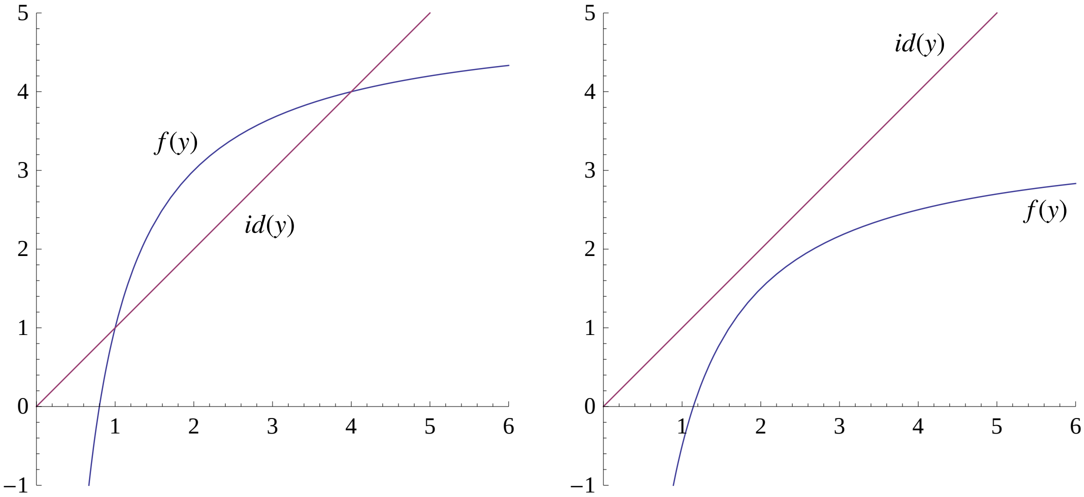

Figure 7.1: Illustration of Lemma 7.2.4, left: case 1, right: case 2

7.3 Estimates for the Extremal Eigenvalues

Here we investigate the symmetric tridiagonal matrix J from (7.1) and assume, without loss of generality that $\beta \cdot \gamma \cdot \delta \neq 0$ and that $m \geq 4$, since otherwise J is not defined properly. Our aim in this section is to develop accurate estimates for the smallest and largest eigenvalue of J. For the special case where $\beta = \gamma = \delta$, the matrix J becomes a tridiagonal Toeplitz matrix whose eigenvalues are known explicitly and given by

$$\lambda_j = \alpha + 2\,|\gamma| \cos\left(\frac{j}{m+1}\pi\right) \qquad \forall j = 1, \ldots, m, \tag{7.4}$$

cf. [5, Thm 2.4].

Remark 7.3.1. Consider, for the moment, once again the special case $\beta = \gamma = \delta$ and let us denote the corresponding Toeplitz matrix by T. Then it follows from (7.4) that $\lambda_{\min}(T) = \alpha + 2|\gamma| \cos\left(\frac{m}{m+1}\pi\right)$ and $\lambda_{\max}(T) = \alpha + 2|\gamma| \cos\left(\frac{1}{m+1}\pi\right)$. In particular, for increasing dimension $m \to \infty$, we therefore get $\lambda_{\min}(T) \to \alpha - 2|\gamma|$ and $\lambda_{\max}(T) \to \alpha + 2|\gamma|$. For the general case, where β, γ, and δ are not necessarily equal, this still implies that for any bound of the form $\lambda_{\min}(J) \geq \alpha - K$ and $\lambda_{\max}(J) \leq \alpha + K$ for some suitable constant $K > 0$, we must have $K \geq 2|\lambda|$ if this bound should hold for all (sufficiently large) dimensions $m \in \mathbb{N}$. This observation follows from the previous fact by noting that we can reorder the entries of J by a symmetric permutation such the first $m-2$ principal submatrices of J are Toeplitz matrices T of different dimensions, hence the claim follows from the interlacing property from Theorem 7.2.1 (b).

For the matrix J, which may be viewed as a (small) perturbation of the Toeplitz case, an analytic representation of the eigenvalues is not known. Our aim is therefore to obtain suitable lower and upper bounds for the extremal eigenvalues of J. Simple estimates can be obtained using Gershgorin's Theorem, see [14, Thm. 7.2.1], which implies that

$$\begin{aligned} \lambda_{\min}(J) &\geq \alpha - \max\{2|\gamma|, |\beta| + |\gamma|, |\delta| + |\gamma|\} \qquad \text{and} \\ \lambda_{\max}(J) &\leq \alpha + \max\{2|\gamma|, |\beta| + |\gamma|, |\delta| + |\gamma|\}. \end{aligned}$$

These estimates can be improved using suitable scalings of J, but it seems that the corresponding estimates are still worse than those that we develop in our subsequent theory.

To this end, let J_k be the principal $k \times k$ submatrix of J, and let

$$p_k(x) := \det(J_k - xI) \qquad \forall k = 1, 2, \ldots, m$$

be the corresponding characteristic polynomial. From (7.2), we obtain that these polynomials satisfy the following recursion:

$$\begin{aligned} p_0(x) &:= 1, \\ p_1(x) &= \alpha - x, \\ p_2(x) &= (\alpha - x) \cdot p_1(x) - \beta^2 \cdot p_0(x), \\ p_k(x) &= (\alpha - x) \cdot p_{k-1}(x) - \gamma^2 \cdot p_{k-2}(x) \quad \forall k = 3, \ldots, m-1, \\ p_m(x) &= (\alpha - x) \cdot p_{m-1}(x) - \delta^2 \cdot p_{m-2}(x). \end{aligned} \tag{7.5}$$

Here, $p_m(x)$ is the characteristic polynomial of J.

In view of Theorem 7.2.1, the characteristic polynomials $p_k(x)$ have real and single roots. Furthermore, given an arbitrary $\alpha \in \mathbb{R}$, the number $w(\alpha)$, denoting the number of sign changes in the Sturm sequence $p_0(\alpha), p_1(\alpha), \ldots, p_m(\alpha)$, is equal to the number of roots of $p_m(x)$ which are smaller than α.

Based on the above recursion and a simple induction argument, we can easily deduce that the polynomials p_k are symmetric in the following sense:

$$p_k(\alpha - y) = \begin{cases} p_k(\alpha + y), & \text{if } k \text{ is even}, \\ -p_k(\alpha + y), & \text{if } k \text{ is odd}. \end{cases} \tag{7.6}$$

In particular, for $k = m$, this implies that $\alpha - y$ is an eigenvalue of J if and only if $\alpha + y$ is an eigenvalue of J, hence the eigenvalues of J are distributed symmetrically around the point α. Consequently, $\alpha - K$ is a lower bound for the smallest eigenvalue (for some $K > 0$) if and only if $\alpha + K$ is an upper bound for the largest eigenvalue. Hence we only have to find suitable lower bounds for the smallest eigenvalue of J.

The basic idea to find suitable estimates of K is the following: We will find conditions (on K and, sometimes, also on the dimension m) which guarantee that all the numbers $p_k(\alpha - K)$ have the same sign which is equivalent to saying that all these numbers are positive since $K > 0$ is equivalent to $p_1(\alpha - K) > 0$. Then it follows from our previous considerations that $w(\alpha - K) = 0$, hence all zeros of p_m must be greater or equal to $\alpha - K$. However, since $p_m(\alpha - K) > 0$, all zeros must actually be greater than $\alpha - K$.

Instead of studying the Sturm sequence $\{p_1(x), p_2(x), \ldots, p_m(x)\}$ directly, we consider the quotients

$$r_k(x) := \frac{p_{k+1}(x)}{p_k(x)} \qquad \forall k = 1, 2, \ldots, m-1. \tag{7.7}$$

Using the recursion of the polynomials $p_k(x)$ in (7.5), we obtain the corresponding recursion

$$\begin{aligned} r_1(x) &= \frac{(\alpha - x)^2 - \beta^2}{(\alpha - x)}, && && (7.8)\\ r_{k+1}(x) &= (\alpha - x) - \gamma^2 \frac{p_k(x)}{p_{k-1}(x)} = (\alpha - x) - \frac{\gamma^2}{r_k(x)} && \text{for } k = 1, 2, \ldots, m-3, && (7.9)\\ r_{m-1}(x) &= (\alpha - x) - \frac{\delta^2}{r_{m-2}(x)} && \text{for } k = m-2. \end{aligned}$$

Based on these quotients, we have the following criterion.

Proposition 7.3.2. *Let $x < \alpha$. Then every member of the Sturm sequence $p_1(x), \ldots, p_m(x)$ is positive if and only if $r_k(x)$ is positive for all $k = 1, \ldots, m-2$ and $r_{m-2}(x) > h(x)$ holds, where*

$$h(x) := \frac{\delta^2}{\alpha - x}. \tag{7.10}$$

Proof. First suppose that all numbers $p_1(x), \ldots, p_m(x)$ are positive. Then (7.7) immediately implies $r_k(x) > 0$ for all $k = 1, \ldots, m-2$ (and also for $k = m-1$, but this part is not needed for our assertion). Furthermore, since $p_{m-2}(x) > 0$, we have the following equivalences that will also be used in order to verify the converse direction:

$$\begin{aligned} p_m(x) > 0 &\iff (\alpha - x)\,p_{m-1}(x) - \delta^2 p_{m-2}(x) > 0\\ &\iff (\alpha - x)\,p_{m-1}(x) > \delta^2 p_{m-2}(x)\\ &\overset{p_{m-2}(x)>0}{\iff} (\alpha - x)\tfrac{p_{m-1}(x)}{p_{m-2}(x)} > \delta^2 \qquad (7.11)\\ &\iff \tfrac{p_{m-1}(x)}{p_{m-2}(x)} > \tfrac{\delta^2}{\alpha - x}\\ &\iff r_{m-2}(x) > h(x). \end{aligned}$$

Since, in the proof of this direction, we have $p_m(x) > 0$, the above chain of equivalences therefore gives $r_{m-2}(x) > h(x)$.

Conversely, assume that $r_1(x), \ldots, r_{m-2}(x)$ are all positive, and that, in addition, we have $r_{m-2}(x) > h(x)$. Since the recursion (7.7) implies

$$p_{k+1}(x) = r_k(x) p_k(x) \quad \forall k = 1, \ldots, m-1,$$

and since we have $p_1(x) = \alpha - x > 0$ by assumption, we immediately obtain $p_{k+1}(x) > 0$ for all $k = 1, \ldots, m-2$. In particular, we therefore have $p_{m-2}(x) > 0$. The chain of equivalences (7.11) then shows that we also have $p_m(x) > 0$. □

Note that, in the statement of Proposition 7.3.2, we could alternatively require the positivity of $r_k(x)$ only for all $k = 1, \ldots, m-3$ since $r_{m-2}(x) > 0$ follows directly from the additional condition $r_{m-2}(x) > h(x)$ due to the fact that $h(x)$ is positive in view of the assumption that $x < \alpha$. We further note that it is indeed enough to consider the positivity of the Sturm sequence $\{p_1(x), \ldots, p_m(x)\}$ instead of $\{p_0(x), p_1(x), \ldots, p_m(x)\}$ since $p_0(x) \equiv 1$ is positive by definition and therefore does not imply additional sign changes.

	$r_1(x) < f_1(x)$	$r_1(x) \in (f_1(x), f_2(x))$	$r_1(x) > f_2(x)$
$h(x) < f_1(x)$	$\forall m \leq m_0$	$\forall m \in \mathbb{N}$	$\forall m \in \mathbb{N}$
$h(x) \in (f_1(x), f_2(x))$	never	$\forall m \geq m_0$	$\forall m \in \mathbb{N}$
$h(x) > f_2(x)$	never	never	$\forall m \leq m_0$

Table 7.1: Lower bounds x for $\lambda_{\min}(J)$ depending on the sizes of $r_1(x)$ and $h(x)$

The interesting part of Proposition 7.3.2 is the fact that we can characterize the positivity of all members from the Sturm sequence $p_1(x), \dots, p_m(x)$ in terms of the quotients $r_1(x), \dots, r_{m-2}(x)$ (together with the function $h(x)$ from (7.10)). Hence the quotient $r_{m-1}(x)$ is not needed in this characterization which is important since the recursion for r_{m-1} is different from the recursion of all the other quotients $r_k(x)$.

This observation is also useful from the following point of view: We will sometimes consider the dimension m of the given matrix J to be variable, i.e., we consider matrices of the form J with different dimensions. Now, the polynomials p_k and, therefore, also the quotients r_k obviously depend on the dimension of J. However, taking into account the particular structure of J, it follows immediately that, for two different dimensions m and $\tilde{m}$ with $m < \tilde{m}$, the quotients $r_k(x)$ $(k = 1, 2, \dots, m-2)$ are the same for both matrices.

We now take a closer look at the recursion (7.9). The initial element $r_1(x)$ is given by (7.8), whereas the recursion itself can be written as

$$r_{k+1}(x) = f(r_k(x)) \qquad \forall k = 1, 2, \dots, m-3$$

by using the function f from (7.3). The (fixed point) properties of the mapping f were already discussed in Lemma 7.2.4. In particular, it follows from this result that, in the only interesting case $x \leq \alpha - 2|\gamma|$, there are two fixed points f_1 and f_2, with f_1 being a repelling fixed point and f_2 being an attracting fixed point. Since these fixed points depend on the given x, we denote them by $f_1(x)$ and $f_2(x)$ from now on. In view of Proposition 7.3.2 we want the sequence $r_1(x), \dots, r_{m-3}(x)$ (note that $r_k(x)$ plays the role of y_k in Lemma 7.2.4) to be positive and, in addition, $r_{m-2}(x) > h(x)$. Obviously, whether these relations hold depends on how the starting point $y_1 = r_1(x)$ and the number $h(x) > 0$ are related to the fixed points of f.

In fact, using Proposition 7.3.2 and Lemma 7.2.4, we have the situation from Table 7.3 whose entries will be explained immediately.

This table assumes (implicitly) that $x \leq \alpha - 2|\gamma|$ and shows in which situation and under which conditions the given x provides a lower bound for the smallest eigenvalue of the matrix J.

More precisely, the table has the following meaning: There are nine cases depending on whether $h(x)$ is smaller than the fixed point $f_1(x)$ or strictly between the two fixed points $f_1(x)$ and $f_2(x)$ or larger than $f_2(x)$, and whether the quotient $r_1(x)$ is smaller than $f_1(x)$, between $f_1(x)$ and $f_2(x)$ or larger than $f_2(x)$. For simplicity of presentation, we do not consider the (often trivial) cases where $h(x)$ or $r_1(x)$ are equal to one of the two fixed points. Then the entry "never" indicates that the given x does not provide a lower bound on the smallest eigenvalue of J regardless of the dimension m of J. The entry "for all $m \in \mathbb{N}$"

indicates that the given x is a lower bound of the smallest eigenvalue of J for all dimensions $m \in \mathbb{N}$, whereas the entries "$\forall m \leq m_0$" and "$\forall m \geq m_0$" indicate that the given x provides a lower bound on the smallest eigenvalue of J for all sufficiently small and all sufficiently large m, respectively.

We still have to explain how these entries were obtained. We do not consider all nine cases since the argument is often the same, but let us take a closer view at some of these cases. First, consider the case $h(x) < f_1(x)$ and $r_1(x) < f_1(x)$. Lemma 7.2.4 then implies that the sequence $r_k(x)$ is monotonically decreasing and eventually becomes negative. Hence, only the first few elements of this sequence are positive, and the additional requirement $r_{m-2}(x) > h(x)$ can therefore also hold only for sufficiently small (possibly no) dimensions m. In view of Proposition 7.3.2, it therefore follows that the given x is a lower bound for the smallest eigenvalue of J only for all sufficiently small m, i.e., for all $m \leq m_0$ with some $m_0 \in \mathbb{N}$. This explains the corresponding entry in the upper left corner of Table 7.3.

Next, consider the case $h(x) < f_1(x)$ and $r_1(x) \in (f_1(x), f_2(x))$. Then Lemma 7.2.4 shows that the sequence $r_k(x)$ is monotonically increasing and converges to the fixed point $f_2(x)$. In particular, $r_k(x)$ is positive for all k, and $r_k(x) > h(x)$ holds for all $k \in \mathbb{N}$, especially, this holds for $k = m-2$ for any given dimension $m \in \mathbb{N}$. Hence it follows from Proposition 7.3.2 that the given x provides a lower bound on the smallest eigenvalue of J for all dimensions $m \in \mathbb{N}$ which again explains the corresponding entry in this case.

Now consider the case $h(x) \in (f_1(x), f_2(x))$ and $r_1(x) < f_1(x)$. Then Lemma 7.2.4 shows, in particular that all quotients $r_k(x)$ stay less than $f_1(x)$, so that the condition $r_{m-2}(x) > h(x)$ never holds in this case regardless of the dimension $m \in \mathbb{N}$. Hence Proposition 7.3.2 implies that the given x does not provide a lower bound for the smallest eigenvalue for any dimension $m \in \mathbb{N}$.

Finally, consider the case $h(x) \in (f_1(x), f_2(x))$ and $r_1(x) \in (f_1(x), f_2(x))$. Lemma 7.2.4 then implies that the sequence $r_1(x), r_2(x), r_3(x), \ldots$ is monotonically increasing and converges to the fixed point $f_2(x)$. Hence, all these quotients are positive, and eventually they are larger than the number $h(x)$. In particular, for sufficiently large dimensions $m \in \mathbb{N}$, we have $r_{m-2}(x) > h(x)$. Hence Proposition 7.3.2 shows that x is a lower bound for J's smallest eigenvalue for all sufficiently large dimensions m. In addition, the following note holds for this case (which is of particular interest in our further development).

Remark 7.3.3. Consider once again the case $h(x) \in (f_1(x), f_2(x))$ and $r_1(x) \in (f_1(x), f_2(x))$. Then it is possible that we already have $r_1(x) > h(x)$. Using a similar reasoning as before, this implies $r_{m-2}(x) > h(x)$ for *all* dimensions $m \in \mathbb{N}$. Consequently, and in addition to the corresponding entry in Table 7.3, it follows from Proposition 7.3.2 that x is a lower bound for the smallest eigenvalue of J for *all* dimensions $m \in \mathbb{N}$. — We further note that the condition $r_1(x) > h(x)$ is equivalent to $x < \alpha - \sqrt{\beta^2 + \delta^2}$ (provided that $x < \alpha$).

All the other entries in the Table 7.3 follow by a similar reasoning. Now, it is clear how to proceed. The previous table gives clear statements on how to get lower bounds for the smallest eigenvalue of J in terms of $r_1(x)$ and $h(x)$ compared to the two fixed points $f_1(x)$ and $f_2(x)$. Our aim is therefore to re-interpret these conditions in terms of the original data of the matrix J. The following technical result investigates these data and shows how $r_1(x)$ is related to the fixed points $f_1(x)$ and $f_2(x)$ depending on the relation between the data β

and γ of the matrix J whose dimension m is fixed in this lemma.

Lemma 7.3.4. *Let $x \leq \alpha - 2\,|\gamma|$, $f_1\,(x)$ and $f_2\,(x)$ be the fixed points of the function f, and let $r_1\,(x)$ be the first quotient from* (7.8)*. Then the following statements hold:*

Case $|\beta| > \sqrt{2}\,|\gamma|$: Then $r_1\,(x) > f_1\,(x) \iff x < \alpha - \frac{\beta^2}{\sqrt{\beta^2-\gamma^2}}$, whereas the inequality $r_1\,(x) < f_2\,(x)$ always holds (i.e. this inequality holds for all $x \leq \alpha - 2|\gamma|$).

Case $|\beta| = \sqrt{2}\,|\gamma|$: Then $f_1\,(x) < r_1\,(x) < f_2\,(x)$ holds for all $x < \alpha - 2\,|\gamma|$ (whereas at the boundary point $x = \alpha - 2\,|\gamma|$, we have $r_1\,(x) = f_1\,(x) = f_2\,(x)$.)

Case $|\gamma| < |\beta| < \sqrt{2}\,|\gamma|$: Then $r_1\,(x) > f_1\,(x)$ always holds (i.e. for all $x \leq \alpha - 2|\gamma|$), and $r_1\,(x) < f_2\,(x) \iff x < \alpha - \frac{\beta^2}{\sqrt{\beta^2-\gamma^2}}$.

Case $|\beta| \leq |\gamma|$: Then we always have $r_1\,(x) > f_2\,(x) \geq f_1\,(x)$.

Proof. To simplify the notation, we set $z := \alpha - x$, so that $x \leq \alpha - 2\,|\gamma|$ is equivalent to $z \geq 2\,|\gamma|$. We divide the (technical, but completely elementary) proof into three steps: Part (A) contains some facts that will be used in the subsequent parts. In part (B), we study how $r_1\,(x)$ and $f_1\,(x)$ relate to each other, and in part (C) we study the relation between $r_1\,(x)$ and $f_2\,(x)$.

(A) We start by finding candidates $x \leq \alpha - 2\,|\gamma|$ for which $r_1\,(x) = f_1\,(x)$ or $r_1\,(x) = f_2\,(x)$ is possible. We do this simultaneously for both fixed points. To this end, note that

$$\begin{aligned} r_1\,(x) = f_{1/2}\,(x) &\iff \frac{z^2-\beta^2}{z} = \frac{1}{2}\left(z \mp \sqrt{z^2-4\gamma^2}\right) \\ &\overset{z>0}{\iff} 2z^2 - 2\beta^2 = z^2 \mp z\sqrt{z^2-4\gamma^2} \\ &\iff z^2 - 2\beta^2 = \mp z\sqrt{z^2-4\gamma^2} \\ &\implies z^4 - 4\beta^2 z^2 + 4\beta^4 = z^2\left(z^2-4\gamma^2\right) \\ &\iff \beta^4 = z^2\left(\beta^2-\gamma^2\right). \end{aligned} \tag{7.12}$$

Hence, for both equations, we have the same necessary condition. It is satisfied for $z_1 = -\frac{\beta^2}{\sqrt{\beta^2-\gamma^2}}$ and $z_2 = +\frac{\beta^2}{\sqrt{\beta^2-\gamma^2}}$. Therefore, the possible candidates are $x_1 = \alpha + \frac{\beta^2}{\sqrt{\beta^2-\gamma^2}}$ and $x_2 = \alpha - \frac{\beta^2}{\sqrt{\beta^2-\gamma^2}}$. We call them the roots of the above equation. Using (7.12), it is also clear that there are no (real) roots if $|\beta| = |\gamma|$ or $|\beta| < |\gamma|$.

(B) Here we discuss the relation between $r_1(x)$ and the fixed point $f_1(x)$ in terms of the orginal data of the matrix J. To this end, first note that, since $\alpha - x \geq 2|\gamma| > 0$, we have

$$\begin{aligned} r_1(x) > f_1(x) &\iff \frac{(\alpha-x)^2-\beta^2}{(\alpha-x)} > \frac{1}{2}\left((\alpha-x) - \sqrt{(\alpha-x)^2-4\gamma^2}\right) \\ &\iff 2\beta^2 < (\alpha-x)^2 + (\alpha-x)\sqrt{(\alpha-x)^2-4\gamma^2}, \end{aligned}$$

and the last inequality obviously holds for all sufficiently small x. We call this observation (O1).

At this point, we have to discuss several cases:

Let $|\beta| > \sqrt{2}\,|\gamma|$. In this case, it turns out, by inserting the two candidate points x_1 and x_2 that x_2 is the only root in the interval $(-\infty, \alpha - 2|\gamma|)$. Together with observation (O1), it therefore follows that $r_1(x) > f_1(x)$ if $x < x_2$, whereas we have $r_1(x) < f_1(x)$ if $x > x_2$.

Let $|\beta| = \sqrt{2}\,|\gamma|$. Here we can write $x_{1/2} = \alpha \pm 2\,|\gamma|$. A simple calculation shows that both candidates are indeed roots. Since $x_2 < x_1$, we obtain from observation (O1) that $r_1(x) > f_1(x)$ for $x < x_2$ (note that $x \leq \alpha - 2\,|\gamma| = x_2$ was the prerequisite of this lemma, so the case $x > x_2$ does not occur), whereas we have $r_1(x) = f_1(x)$ at $x = x_2$ since x_2 is a root of our equation.

Let $|\gamma| < |\beta| < \sqrt{2}\,|\gamma|$. Here it is easy to see that x_1 is the only root among the two candidates. Together with observation (O1), we therefore get $r_1(x) > f_1(x)$ for $x < x_1$. However, in this case, we have $x_1 > \alpha - 2\,|\gamma|$. Hence, we obtain $r_1(x) > f_1(x)$ for all $x \leq \alpha - 2\,|\gamma|$.

Let $|\beta| \leq |\gamma|$. In this case, there are no roots in view of part (A). It therefore follows from observation (O1) that $r_1(x) > f_1(x)$ holds for all $x \leq \alpha - 2|\gamma|$.

(C) Here we discuss the relation between $r_1(x)$ and the fixed point $f_2(x)$, again in terms of the orginal data of the matrix J. The considerations are similar to those from part (B). To this end, we first note that $r_1(x) > f_2(x)$ is equivalent to $2\beta^2 < z^2\left(1 - \sqrt{1 - 4\gamma^2/z^2}\right) =: g(z)$. Using l'Hospital's rule, we obtain

$$\lim_{z\to\infty} g(z) = \lim_{z\to\infty} \frac{\left(1 - \sqrt{1 - 4\gamma^2/z^2}\right)}{z^{-2}} = \lim_{z\to\infty} \frac{2\gamma^2}{\sqrt{1 - 4\gamma^2/z^2}} = 2\gamma^2 .$$

Taking into account that $z = \alpha - x$, it follows that $r_1(x) > f_2(x)$ for all x sufficiently small if $|\beta| < |\gamma|$. Similarly, one can show that $r_1(x) < f_2(x)$ for all x sufficiently small if $|\beta| > |\gamma|$ (whereas the case $|\beta| = |\gamma|$ has to be treated separately). We call these statements observation (O2).

Like before, we proceed by considering several cases:

Let $|\beta| > \sqrt{2}\,|\gamma|$. Through simple calculation, we get that x_1 is the only root. Observation (O2) therefore implies $r_1(x) < f_2(x)$ for all $x < x_1$. But since $x_1 > \alpha - 2\,|\gamma|$ we even have $r_1(x) < f_2(x)$ for all $x \leq \alpha - 2\,|\gamma|$.

Let $|\beta| = \sqrt{2}\,|\gamma|$. Here $x_{1/2} = \alpha \pm 2\,|\gamma|$ are the two roots, but x_1 is greater than $\alpha - 2\,|\gamma|$ and hence irrelevant for our case. Using observation (O2) once again, we get $r_1(x) < f_2(x)$ for all $x < \alpha - 2\,|\gamma|$ as in the previous case, whereas we have $r_1(x) = f_2(x)$ at the boundary point $x = x_2 = \alpha - 2|\gamma|$.

Let $|\gamma| < |\beta| < \sqrt{2}\,|\gamma|$. Then x_2 is the only root, and observation (O2) gives $r_1(x) < f_2(x)$ if and only if $x < x_2$.

Let $|\beta| = |\gamma|$. We know from part (A) that there are no roots in this case, hence either $r_1(x) < f_2(x)$ or $r_1(x) > f_2(x)$ holds for all $x \leq \alpha - 2|\gamma|$. To decide which of these two inequalities holds, observation (O2) cannot be applied directly. However, direct calculation shows that $2\beta^2 = 2\gamma^2 < 4\gamma^2 = g(2\,|\gamma|)$, so that observation (O2) now gives $r_1(\alpha - 2\,|\gamma|) > f_2(\alpha - 2\,|\gamma|)$. Hence $r_1(x) > f_2(x)$ holds for all $x \leq \alpha - 2\,|\gamma|$.

Let $|\beta| < |\gamma|$. According to part (A), there are no roots in this case. Together with observation (O2), it follows that $r_1(x) > f_2(x)$ holds for all $x \leq \alpha - 2\,|\gamma|$.

The statement now follows by summarizing all subcases considered in parts (B) and (C). □

The following result is similar to the previous one (so we skip its proof) and shows how the number $h(x)$ is related to the two fixed points $f_1(x)$ and $f_2(x)$ depending on the original data δ and γ of our matrix J whose dimension m is again assumed to be fixed.

Lemma 7.3.5. *Let $x \leq \alpha - 2\,|\gamma|$, $f_1(x)$ and $f_2(x)$ be the fixed points of the function f, and let $h(x)$ be defined by (7.10). Then the following statements hold:*

Case $|\delta| > \sqrt{2}\,|\gamma|$: Then $h(x) < f_2(x) \iff x < \alpha - \frac{\delta^2}{\sqrt{\delta^2-\gamma^2}}$, whereas the inequality $h(x) > f_1(x)$ always holds (i.e. for all $x \leq \alpha - 2|\gamma|$).

Case $|\delta| = \sqrt{2}\,|\gamma|$: Then $f_1(x) < h(x) < f_2(x)$ for all $x < \alpha - 2\,|\gamma|$ (and $h(x) = f_1(x) = f_2(x)$ for the boundary point $x = \alpha - 2\,|\gamma|$).

Case $|\gamma| < |\delta| < \sqrt{2}\,|\gamma|$: Then $h(x) > f_1(x) \iff x < \alpha - \frac{\delta^2}{\sqrt{\delta^2-\gamma^2}}$, whereas the inequality $h(x) < f_2(x)$ always holds (i.e. for all $x \leq \alpha - 2|\gamma|$).

Case $|\delta| \leq |\gamma|$: Then we always have $h(x) < f_1(x) \leq f_2(x)$.

Now we are going to combine the previous results in order to get estimates for the extremal eigenvalues of the matrix J. We stress, however that it cannot be avoided that these bounds (in addition to the data of the matrix) sometimes also depend on the dimension m of this matrix, cf. Table 7.3 and the discussion to derive the entries of this table.

Unfortunately, we have to distinguish several cases in the presentation of our main result. In view of Lemmas 7.3.4 and 7.3.5, there are actually 16 different cases to consider, namely those that occur by combining the four possibilities

$$|\beta| > \sqrt{2}|\gamma|, \quad |\beta| = \sqrt{2}|\gamma|, \quad |\beta| \in (|\gamma|, \sqrt{2}|\gamma|), \text{ and } |\beta| \leq |\gamma|$$

from Lemma 7.3.4 with the corresponding four possibilities

$$|\delta| > \sqrt{2}|\gamma|, \quad |\delta| = \sqrt{2}|\gamma|, \quad |\delta| \in (|\gamma|, \sqrt{2}|\gamma|), \text{ and } |\delta| \leq |\gamma|$$

from Lemma 7.3.5.

Theorem 7.3.6. *Define $\bar{\beta} := \frac{\beta^2}{\sqrt{\beta^2-\gamma^2}}$ and $\bar{\delta} := \frac{\delta^2}{\sqrt{\delta^2-\gamma^2}}$. Then the inequalities*

$$\lambda_{\min}(J) \geq \alpha - K \qquad \textit{and} \qquad \lambda_{\max}(J) \leq \alpha + K$$

holds

(a) *for* all *dimensions $m \in \mathbb{N}$ with K being the constant from the following table:*

	$\lvert\delta\rvert > \sqrt{2}\lvert\gamma\rvert$	$\lvert\delta\rvert = \sqrt{2}\lvert\gamma\rvert$	$\lvert\delta\rvert \in (\lvert\gamma\rvert, \sqrt{2}\lvert\gamma\rvert)$	$\lvert\delta\rvert \leq \lvert\gamma\rvert$
$\lvert\beta\rvert > \sqrt{2}\lvert\gamma\rvert$	$\sqrt{\beta^2+\delta^2}$	$\sqrt{\beta^2+\delta^2}$	$\max\{\bar{\beta}, \sqrt{\beta^2+\delta^2}\}$	$\bar{\beta}$
$\lvert\beta\rvert = \sqrt{2}\lvert\gamma\rvert$	$\sqrt{\beta^2+\delta^2}$	$2\lvert\gamma\rvert$	$2\lvert\gamma\rvert$	$2\lvert\gamma\rvert$
$\lvert\beta\rvert \in (\lvert\gamma\rvert, \sqrt{2}\lvert\gamma\rvert)$	$\max\{\bar{\delta}, \sqrt{\beta^2+\delta^2}\}$	$2\lvert\gamma\rvert$	$2\lvert\gamma\rvert$	$2\lvert\gamma\rvert$
$\lvert\beta\rvert \leq \lvert\gamma\rvert$	$\bar{\delta}$	$2\lvert\gamma\rvert$	$2\lvert\gamma\rvert$	$2\lvert\gamma\rvert$

(b) *for* all sufficiently large *dimensions* $m \in \mathbb{N}$ *with the (usually sharper) constant* K *from the following table:*

	$\lvert\delta\rvert > \sqrt{2}\lvert\gamma\rvert$	$\lvert\delta\rvert = \sqrt{2}\lvert\gamma\rvert$	$\lvert\delta\rvert \in (\lvert\gamma\rvert, \sqrt{2}\lvert\gamma\rvert)$	$\lvert\delta\rvert \leq \lvert\gamma\rvert$
$\lvert\beta\rvert > \sqrt{2}\lvert\gamma\rvert$	$\max\{\bar{\beta}, \bar{\delta}\}$	$\bar{\beta}$	$\bar{\beta}$	$\bar{\beta}$
$\lvert\beta\rvert = \sqrt{2}\lvert\gamma\rvert$	$\bar{\delta}$	$2\lvert\gamma\rvert$	$2\lvert\gamma\rvert$	$2\lvert\gamma\rvert$
$\lvert\beta\rvert \in (\lvert\gamma\rvert, \sqrt{2}\lvert\gamma\rvert)$	$\bar{\delta}$	$2\lvert\gamma\rvert$	$2\lvert\gamma\rvert$	$2\lvert\gamma\rvert$
$\lvert\beta\rvert \leq \lvert\gamma\rvert$	$\bar{\delta}$	$2\lvert\gamma\rvert$	$2\lvert\gamma\rvert$	$2\lvert\gamma\rvert$

Proof. In view of our previous observation, $\alpha - K$ is a lower bound for $\lambda_{\min}(J)$ if and only if $\alpha + K$ is an upper bound for $\lambda_{\max}(J)$ for some $K > 0$. Hence it is enough to verify the lower bounds for the minimum eigenvalue of J. We further note that, in view of Remark 7.3.1, we (have to) assume throughout this proof that $x \leq \alpha - 2\,\lvert\gamma\rvert$ since there cannot be a lower bound greater than $\alpha - 2\,\lvert\gamma\rvert$ that fits for all (sufficiently large) matrix sizes m.

We begin by stating some elementary inequalities (without proof) that are useful for the subsequent considerations:

(I) If $\lvert\beta\rvert > \lvert\gamma\rvert$, then $\bar{\beta} \geq 2\lvert\gamma\rvert$ and $\bar{\beta} = 2\lvert\gamma\rvert$ holds if and only if $\lvert\beta\rvert = \sqrt{2}\lvert\gamma\rvert$.

(II) If $\lvert\delta\rvert > \lvert\gamma\rvert$, then $\bar{\delta} \geq 2\lvert\gamma\rvert$ and $\bar{\delta} = 2\lvert\gamma\rvert$ holds if and only if $\lvert\delta\rvert = \sqrt{2}\lvert\gamma\rvert$.

(II) If $\lvert\beta\rvert \geq \sqrt{2}\lvert\gamma\rvert$ and $\lvert\delta\rvert \geq \sqrt{2}\lvert\gamma\rvert$, then $\sqrt{\beta^2 + \delta^2} \geq \max\{2\lvert\gamma\rvert, \bar{\beta}, \bar{\delta}\}$.

(IV) If $\lvert\beta\rvert = \sqrt{2}\lvert\gamma\rvert$ and $\lvert\delta\rvert \in (\lvert\gamma\rvert, \sqrt{2}\lvert\gamma\rvert)$, then $\bar{\delta} \geq \sqrt{\beta^2 + \delta^2}$.

(V) If $\lvert\beta\rvert \in (\lvert\gamma\rvert, \sqrt{2}\lvert\gamma\rvert)$ and $\lvert\delta\rvert = \sqrt{2}\lvert\gamma\rvert$, then $\bar{\beta} \geq \sqrt{\beta^2 + \delta^2}$.

We now verify statements (a) and (b) simultaneously. In principle, we have to consider each of the possible 16 cases separately. However, it will be enough to consider only one of these cases (in fact, one of the more interesting ones), since the remaining cases can be treated in essentially the same way by referring to the corresponding cases from Lemmas 7.3.4 and 7.3.5 as well as to the corresponding entries of Table 7.3.

The case that we consider in more detail is the one where $\lvert\beta\rvert > \sqrt{2}\lvert\gamma\rvert$ and $\lvert\delta\rvert > \sqrt{2}\lvert\gamma\rvert$ holds. Then Lemma 7.3.4 shows that $r_1(x) < f_2(x)$ holds for all $x \leq \alpha - 2\lvert\gamma\rvert$, whereas $r_1(x) > f_1(x)$ is equivalent to $x < \alpha - \bar{\beta}$. Moreover, Lemma 7.3.5 shows that $h(x) > f_1(x)$ holds for all $x \leq \alpha - 2\lvert\gamma\rvert$, whereas we have $h(x) < f_2(x)$ if and only if $x < \alpha - \bar{\delta}$. Table 7.3 therefore shows that, for all $x < \min\{\alpha - \bar{\beta}, \alpha - \bar{\delta}\}$ and all $x \leq \alpha - 2\lvert\gamma\rvert$, this x provides a lower bound for $\lambda_{\min}(J)$ provided that the dimension m is sufficiently large. By continuity, we therefore get the lower bound

$$\lambda_{\min}(J) \geq \min\{\alpha - 2\lvert\gamma\rvert, \alpha - \bar{\beta}, \alpha - \bar{\delta}\} = \alpha - \max\{2\lvert\gamma\rvert, \bar{\beta}, \bar{\delta}\}$$

for all $m \in \mathbb{N}$ sufficiently large. Using observations (I) and (II), this lower bound reduces to

$$\lambda_{\min}(J) \geq \alpha - \max\{\bar{\beta}, \bar{\delta}\}.$$

This is precisely the lower bound given for the case considered here in statement (b).

However, in this particular case, we can also apply Remark 7.3.3 and obtain a lower bound for $\lambda_{\min}(J)$ for *all* dimensions $m \in \mathbb{N}$ if, in addition, x is chosen in such a way that $r_1(x) > h(x)$. Since this condition is equivalent to $x < \alpha - \sqrt{\beta^2 + \delta^2}$ according to Remark 7.3.3, it follows, together with our previous considerations that the lower bound

$$\lambda_{\min}(J) \geq \alpha - \max\{2|\gamma|, \bar{\beta}, \bar{\delta}, \sqrt{\beta^2 + \delta^2}\}$$

holds for all dimensions $m \in \mathbb{N}$. In view of observation (III), this lower bound boils down to

$$\lambda_{\min}(J) \geq \alpha - \sqrt{\beta^2 + \delta^2}$$

and therefore justifies the corresponding bound given in statement (a). □

We close this section with some remarks about the previous result.

Remark 7.3.7. (a) Except for the trivial case $|\beta|, |\delta| \leq |\gamma|$, our bounds on the extremal eigenvalues of the matrix J are better than those that come from Gershgorin's Theorem.

(b) The case $|\beta| = |\delta| = \sqrt{2}|\gamma|$ gives $\alpha - 2|\gamma|$ and $\alpha + 2|\gamma|$ as lower and upper bounds for $\lambda_{\min}(J)$ and $\lambda_{\max}(J)$, respectively. However, in this case these bounds are exact, i.e. $\lambda_{\min}(J) = \alpha - 2|\gamma|$ and $\lambda_{\max}(J) = \alpha + 2|\gamma|$. This follows from the recursion (7.5) which, in this case, gives $p_0(x) = 1, p_1(x) = 2|\gamma|, p_2(x) = 2|\gamma|^2, p_k(x) = 2|\gamma|^k$ for all $k = 3, \ldots, m-1$ and $p_m(x) = 0$ for $x := \alpha - 2|\gamma|$.

(c) Consider a matrix of the form

$$A = \begin{bmatrix} \alpha & \bar{\beta} & & & & & \\ \hat{\beta} & \alpha & \bar{\gamma}_3 & & & & \\ & \hat{\gamma}_3 & \alpha & \ddots & & & \\ & & \ddots & \ddots & \bar{\gamma}_{m-1} & & \\ & & & \hat{\gamma}_{m-1} & \alpha & \bar{\delta} \\ & & & & \hat{\delta} & \alpha \end{bmatrix}.$$

Define $\beta, \gamma \in \mathbb{R}$ in such a way that $\hat{\beta} \cdot \bar{\beta} = \beta^2$ and $\hat{\delta} \cdot \bar{\delta} = \delta^2$. Suppose that there is an element $\gamma \in \mathbb{R}$ with $\hat{\gamma}_i \cdot \bar{\gamma}_i = \gamma^2$ for all $i = 3, \ldots, m-1$. Then the characteristic polynomials of all principal submatrices of A coincide with the polynomials $p_k(x)$ from (7.5). Consequently, all the previous considerations for the matrix J also hold for the nonsymmetric matrix A. In particular, the same bounds for the extremal eigenvalues are valid for A.

(d) Statement (b) of Theorem 7.3.6 holds only for all sufficiently large dimensions m, say, for all $m \geq m_0$. Here, the smallest dimension m_0 can be computed in the following way: We are in the situation where $r_1(x) > f_1(x)$ and $h(x) < f_2(x)$ for $x = \alpha - K$, K the bound given in the tables of Theorem 7.3.6. Then the sequence $r_1(x), r_2(x), r_3(x), \ldots$ is monotonically increasing and converges to $f_2(x)$. So there exists a smallest integer s such that $r_s(x) > h(x)$.

Then $m_0 = s + 2$ is the required dimension since $r_s(x)$ determines the behaviour of $p_{s+2}(x)$. Hence we need to compute $r_1(x)$ and $h(x)$ as well as (if still necessary) the other quotients $r_k(x)$ for $k \geq 2$ via the corresponding recursion (7.9) until, for the first time, $r_s(x)$ is greater than $h(x)$.

(e) Theorem 7.3.6 (b) shows that $\lambda_{\min}(J) \geq \alpha - K$ holds for all dimensions $m \geq m_0$ with the constant K given in the corresponding table and a sufficiently large dimension $m_0 \in \mathbb{N}$ that can be computed via the previous remark. However, in some cases it might be enough to satisfy a weaker bound of the form $\lambda_{\min}(J) \geq \alpha - \tilde{K}$ for some $\tilde{K} \geq K$. This bound is certainly satisfied for all dimensions $m \geq m_0$, but it might already be satisfied for smaller dimensions, say, for all $m \geq \tilde{m}_0$ with some $\tilde{m}_0 \leq m_0$. The practical computation of $\tilde{m}_0$ can be done as in (d) with $x = \alpha - K$ replaced by $x = \alpha - \tilde{K}$.

7.4 Application

In this section we want to discuss how the previous theory can be applied to an example. As example we choose the PDE from our mathematical model without source or sink term on the right hand side, cf. Chapter 3. Hereby we assume the water flow field q to be constant in time and space, i.e. $q(t, x) = [q_0, 0]^T$. The result is the homogeneous PDE of second order (3.5) that was mentioned in Subsection 3.5 with the right hand side set to zero. This equation has the typical form of a convection-diffusion equation. It looks like

$$\theta \cdot \frac{\partial c\,(t,x,y)}{\partial t} - \beta_l \cdot q_0 \cdot \frac{\partial^2 c\,(t,x,y)}{\partial x^2} - \beta_t \cdot q_0 \cdot \frac{\partial^2 c\,(t,x,y)}{\partial y^2} + q_0 \cdot \frac{\partial c\,(t,x,y)}{\partial x} = 0$$

defined on $[0, T] \times \Omega$, where $[0, T]$ for some $T > 0$ denotes the time interval and $\Omega = [0, \omega_x] \times \left[0, \omega_y\right] \subseteq \mathbb{R}^2$ for some constants ω_x, ω_y denotes the spatial domain. In addition, we assume that we have boundary conditions described by a Dirichlet condition on the left border and by Neumann conditions on the other boundaries of the domain. The scalar constants $\theta, q_0, \beta_l, \beta_t > 0$ are used to specify some further properties of the given problem; their meaning is described in Subsection 3.1.3. For some additional background material regarding this particular application, we refer the interested reader to [26]. The following strategy is valid for many convection-diffusion PDEs.

Since we have a rectangular domain, the simplest discretization is by finite differences. To this end, we denote by h the step size in the spatial directions x and y, and by τ the step size for the discretization in time. Then we have $n = \frac{\omega_x}{h}$ unknown points in each grid row (for $x = 0$ the values are known by the Dirichlet boundary condition) and $m + 1$ unknown points in each grid column, with $m := \frac{\omega_y}{h}$. With $c_{i,j} := c\,(t_l, i \cdot h, j \cdot h)$ and $c_{i,j}^{old} := c\,(t_{l-1}, i \cdot h, j \cdot h)$ we denote the concentrations of the species at the discretized point (ih, jh) in the current time step $t_l = l \cdot \tau$ and the previous time step, respectively.

To obtain a suitable finite difference approximation of the original PDE, we use forward differences for the first term $\frac{\partial c(t,x,y)}{\partial t}$ (which is the only part that includes a derivate with respect to time), resulting in the first-order Euler approximation

$$\frac{c_{i,j} - c^{old}_{i,j}}{\tau}$$

in every grid point $(x_i, y_i) := (ih,\ jh)$. On the other hand, for the second-order derivative $-\beta_l q_0 \frac{\partial^2 c(t,x,y)}{\partial x^2}$ we use the standard central difference approximation. We also apply a second-order central difference approximation to the first-order derivative $q_0 \frac{\partial c(t,x,y)}{\partial x}$. The resulting approximation in each grid row $j = 0, \ldots, m$ for the inner grid points $i = 2, \ldots, n-1$ is

$$\left(-\frac{\beta_l q_0}{h^2} - \frac{q_0}{2h}\right) c_{i-1,j} + \frac{2\beta_l q_0}{h^2} c_{i,j} + \left(-\frac{\beta_l q_0}{h^2} + \frac{q_0}{2h}\right) c_{i+1,j},$$

while for $i = 1$ the value of $c_{0,j}$ is known from the Dirichlet boundary condition, so we obtain the approximation

$$\frac{2\beta_l q_0}{h^2} c_{1,j} + \left(-\frac{\beta_l q_0}{h^2} + \frac{q_0}{2h}\right) c_{2,j},$$

whereas for $i = n$ we get, taking into account the Neumann boundary condition on the right side of the domain, the discretization

$$\left(-\frac{2\beta_l q_0}{h^2}\right) c_{n-1,j} + \frac{2\beta_l q_0}{h^2} c_{n,j} = 0\,.$$

To write these expressions in matrix notation, we define the vectors

$$c^j := \left(c_{1,j}, c_{2,j}, \ldots, c_{n,j}\right)^T \qquad \forall j = 1, \ldots, m-1$$

and the $n \times n$ matrix

$$L_x := a_x \cdot M_x \quad \text{with} \quad M_x := \begin{bmatrix} 2 & -1+b & & & & \\ -1-b & 2 & -1+b & & & \\ & -1-b & \ddots & \ddots & & \\ & & \ddots & 2 & -1+b & \\ & & & -1-b & 2 & -1+b \\ & & & & -2 & 2 \end{bmatrix}$$

and

$$a_x := a_x(h) := \frac{\beta_l \cdot q_0}{h^2} \qquad \text{and} \qquad b := b(h) := \frac{h}{2\beta_l}.$$

Now the resulting equations for each grid row $j = 0, \ldots, m$ read

$$L_x \cdot c^j\,.$$

Similarly, applying the standard central difference approximation to the second-order derivative $-\beta_t q_0 \frac{\partial^2 c(t,x,y)}{\partial y^2}$, we obtain in each grid column $i = 1, \ldots, n$ for the inner points $j =$

$1, \dots, m-1$ (each boundary point in y direction has a Neumann condition) the discretized equation

$$-\frac{\beta_t q_0}{h^2} c_{i,j-1} + \frac{2\beta_t q_0}{h^2} c_{i,j} - \frac{\beta_t q_0}{h^2} c_{i,j+1} \,,$$

whereas on the lower bound $j = 0$ and the upper bound $j = m$, we have

$$\frac{2\beta_t q_0}{h^2} c_{i,0} - \frac{2\beta_t q_0}{h^2} c_{i,1} \quad \text{and} \quad -\frac{2\beta_t q_0}{h^2} c_{i,m-1} + \frac{2\beta_t q_0}{h^2} c_{i,m},$$

respectively. Like before we define for every grid column $i = 1, \dots, n$ the vectors

$$c_i := \left(c_{i,0}, c_{i,1}, \dots, c_{i,m}\right)^T$$

and the $(m+1) \times (m+1)$ matrix

$$L_y := a_y \cdot M_y \quad \text{with} \quad M_y := \begin{bmatrix} 2 & -2 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -2 & 2 \end{bmatrix}, \quad a_y := a_y(h) := \frac{\beta_t \cdot q_0}{h^2}.$$

Using this notation, we can write these expressions in every grid column $i = 1, \dots, n$ in matrix-vector form as

$$L_y \cdot c_i \,.$$

We now want to write all these $n \cdot (m+1)$ equations in one linear system. To this end, we order our unknowns $c_{i,j}$ in time step l lexicographically into one big vector by stacking together the grid rows c^j into

$$c_h = \left(c_{1,0}, c_{2,0}, \dots, c_{n,0}, c_{1,1}, \dots, c_{n,1}, \dots c_{1,m-1}, c_{2,m-1}, \dots, c_{n,m-1}\right)^T \,.$$

We do the same for the previous time step and call the resulting vector c_h^{old}, whose entries are no longer unknowns in the current time step t_l. We now formulate the matrix of the full linear system with the help of L_x and L_y. Writing $l^y_{i,j}$ for the elements of L_y and defining the $n(m+1) \times n(m+1)$ matrix

$$L_h = \begin{bmatrix} L_x + l^y_{1,1} I_n & l^y_{1,2} I_n & & & \\ l^y_{2,1} I_n & L_x + l^y_{2,2} I_n & l^y_{2,3} I_n & & \\ & l^y_{3,2} I_n & L_x + l^y_{3,3} I_n & \ddots & \\ & & \ddots & \ddots & l^y_{m,m+1} I_n \\ & & & l^y_{m+1,m} I_n & L_x + l^y_{(m+1),(m+1)} I_n \end{bmatrix},$$

the resulting linear system in one time step becomes

$$\left(\theta I_{n(m+1)} + \tau L_h\right) \cdot c_h = \theta \cdot c_h^{old} \,. \tag{7.13}$$

In order to solve this sparse linear system efficiently by an iterative solver, we study the properties of the system matrix $(\theta I_{n(m+1)} + \tau L_h)$. In particular, we are interested in the smallest singular value to know whether the matrix is nonsingular and to compute the condition number. Moreover, we would like to show that this matrix is positive definite (although nonsymmetric), because this guarantees that, e.g., the restarted GMRES solver used in our numerical test is known to converge in this case, cf. [44].

To this end we shortly review the notions of the Kronecker product and the Kronecker sum together with some basic properties. The following results can be found in [18, Section 4.2 and 4.4]. Let $A = \left(a_{i,j}\right)$ be a real $k \times k$ matrix and let $B = \left(b_{i,j}\right)$ be a real $l \times l$ matrix. Then the *Kronecker product* is defined as the $kl \times kl$ matrix

$$A \otimes B = \begin{bmatrix} a_{1,1}B & a_{1,2}B & \cdots & a_{1,k}B \\ a_{2,1}B & a_{2,2}B & \cdots & a_{2,k}B \\ \vdots & & & \vdots \\ a_{k,1}B & a_{k,2}B & \cdots & a_{k,k}B \end{bmatrix}.$$

Given another $l \times l$ matrix C, it holds that

$$A \otimes B + A \otimes C = A \otimes (B + C).$$

Similarly, we also have

$$(A + B) \otimes C = A \otimes C + B \times C$$

for all matrices A, B, C of appropriate dimension. For any real scalar r, we obviously have

$$r \cdot (A \otimes B) = (rA) \otimes B = A \otimes (rB) \ .$$

In addition, it is known that

$$(A \otimes B)^T = A^T \otimes B^T$$

holds for all suitable matrices A, B. Finally, the *Kronecker sum* of $A \in \mathbb{R}^{k \times k}$ and $B \in \mathbb{R}^{l \times l}$ is defined as the $kl \times kl$ matrix

$$I_l \otimes A + B \otimes I_k \, .$$

The eigenvalues $\mu_{i,j}$ of the Kronecker sum are given by $\lambda_i(A) + \lambda_j(B)$ for all $i = 1, \ldots, k$ and all $j = 1, \ldots, l$.

Now let us go back to our example. Using the notion of the Kronecker sum, the matrix L_h can be written as $L_h = I_{m+1} \otimes L_x + L_y \otimes I_n$, so that the matrix of the linear system (7.13) becomes

$$L(\tau, h) := \theta I_{(m+1)n} + \tau \left(I_{m+1} \otimes L_x + L_y \otimes I_n\right).$$

We want to compute a lower bound for the smallest singular value of this matrix. To this end, we first give a lower bound for the smallest eigenvalue of the corresponding symmetric part which is given by

$$L^s(\tau, h) = \theta I_{(m+1)n} + \tau \left(I_{m+1} \otimes L_x^s + L_y^s \otimes I_n\right) \, .$$

The previous considerations show that the smallest eigenvalue of this symmetric part is given by

$$\lambda_1\left(L^s(\tau,h)\right) = \theta + \tau\lambda_1\left(L_x^s\right) + \tau\lambda_1\left(L_y^s\right). \tag{7.14}$$

Hence we obtain a lower bound for the smallest eigenvalue $\lambda_1\left(L^s(\tau,h)\right)$ by calculating lower bounds for $\lambda_1\left(L_x^s\right)$ and $\lambda_1\left(L_y^s\right)$. Since both matrices L_x^s and L_y^s have the structure of the matrix J from (7.1), we can apply the theory from the previous section. Note, however that these bounds depend on our step size h. We will show that, for suitable choices of these step sizes, the matrix $L^s(\tau,h)$ has only positive eigenvalues. This implies that the (nonsymmetric) system matrix $L(\tau,h)$ itself is positive definite (recall that a nonsymmetric matrix A is positive definite if and only if its symmetric part A^s is positive definite). Furthermore, Lemma 7.2.3 then also gives a lower bound for the smallest singular value of $L(\tau,h)$.

Before we proceed, we note that it would alternatively be possible to consider the nonsymmetric matrix $L(\tau,h)$ directly since Remark 7.3.7 (c) can be applied in our particular application. The subsequent analysis, however, deals with the symmetric part $L^s(\tau,h)$ and calculates a lower bound for the smallest eigenvalue using the representation from (7.14).

In this example it is possible according to Remark 7.3.7(c) to find two symmetric tridiagonal matrices which have the same eigenvalues as L_x and L_y. With the theory in the previous section it is then possible to give accurate bounds for the extremal eigenvalues of these two matrices.

Let us first consider the matrix $L_x^s = a_x(h)\cdot M_x^s$. We now give a lower bound for the smallest eigenvalue of the $n\times n$ matrix

$$M_x^s = \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & \ddots & \ddots & & \\ & & \ddots & 2 & -1 & \\ & & & -1 & 2 & -1.5+0.5\cdot b \\ & & & & -1.5+0.5\cdot b & 2 \end{bmatrix}.$$

We adapt the results from the previous section and get the following corollary.

Corollary 7.4.1. *If $b \in \left[3-2\sqrt{2},\, 3+2\sqrt{2}\right]$ then $\lambda_1\left(M_x^s\right) \geq 0$ holds for all $n \geq 4$. If $b < 3-2\sqrt{2}$ or $b > 3+2\sqrt{2}$ then $\lambda_1\left(M_x^s\right) \geq 2 - \frac{d^2}{\sqrt{d^2-1}}$ holds for all $n \geq 4$, where $d = -1.5+0.5b$ is the perturbed entry of M_x^s.*

Proof. We first consider the simple case $b \in \left[3-2\sqrt{2},\, 3+2\sqrt{2}\right]$ which is equivalent with $|-1.5+0.5b| \leq \sqrt{2}$. With Theorem 3.6 (a) applied in the case "$|\beta| \leq |\gamma|$ and $|\delta| \leq \sqrt{2}|\gamma|$", we get the first estimate. The case $b < 3-2\sqrt{2}$ or $b > 3+2\sqrt{2}$ is equivalent to $|-1.5+0.5b| > \sqrt{2}$. Using Theorem 7.3.6 (a) once again, but applied in the case "$|\beta| \leq |\gamma|$ and $|\delta| > \sqrt{2}|\gamma|$", we obtain $2 - \frac{d^2}{\sqrt{d^2-1}}$ as a lower bound. The restriction regarding the dimension is simply due to the fact that all considerations in the previous section implicitly assumed that the matrices are at least 4×4-dimensional. □

Note that $2 - \frac{d^2}{\sqrt{d^2-1}}$ in the previous Corollary is always negative in the case where it is applied. Hence the corresponding matrix M_x^s is not necessarily positive definite in this case.

Similarly, we now study the $(m+1) \times (m+1)$ matrix $L_y^s = a_y(h) \cdot M_y^s$. We therefore give a lower bound for the smallest eigenvalue of

$$M_y^s = \begin{bmatrix} 2 & -1.5 & & & \\ -1.5 & 2 & \ddots & & \\ & -1 & \ddots & -1 & \\ & & \ddots & 2 & -1.5 \\ & & & -1.5 & 2 \end{bmatrix}.$$

To achieve the most accurate bounds, we distinguish different matrix sizes.

Corollary 7.4.2. *If $m \geq 65$ then $\lambda_1\left(M_y^s\right) \geq -0.0125$.*
If $m \geq 26$ then $\lambda_1\left(M_y^s\right) \geq -0.015$.
If $m \geq 16$ then $\lambda_1\left(M_y^s\right) \geq -0.02$.
If $m \geq 12$ then $\lambda_1\left(M_y^s\right) \geq -0.025$.

Proof. Theorem 7.3.6 (b) applied in the case "$|\beta| > \sqrt{2}|\gamma|$ and $|\delta| > \sqrt{2}|\gamma|$" shows that $\lambda_1\left(M_y^s\right) \geq 2 - \frac{2.25}{\sqrt{1.25}} \approx -0.01246$ holds for all sufficiently large m. Replacing this lower bound by the less restrictive numbers -0.0125, -0.015, -0.02 and -0.025, respectively, we obtain the desired statements in a way described in Remark 7.3.7 (e). □

Using (7.14), we therefore obtain

$$\begin{aligned} \lambda_{min}\left(L^s(\tau,h)\right) &= \theta + \tau a_x(h)\,\lambda_{min}\left(M_x^s\right) + \tau a_y(h)\,\lambda_{min}\left(M_y^s\right) \\ &= \theta + \tau\frac{\beta_l \cdot q_0}{h^2}\lambda_{min}\left(M_x^s\right) + \tau\frac{\beta_t \cdot q_0}{h^2}\lambda_{min}\left(M_y^s\right). \end{aligned}$$

From Lemma 7.2.3 we know that $\sigma_{min}(L(\tau,h)) \geq \lambda_{min}(L^s(\tau,h))$ if $L^s(\tau,h)$ is positive definite, which is equivalent to $\lambda_{min}(L^s(\tau,h)) > 0$. Recall that $b = b(h) = \frac{h}{2\beta_l}$ and therefore $b(h) > 0$ for all $h > 0$. Taking into account the two different cases considered in Corollary 7.4.1, we obtain the lower bound

$$\lambda_{min}\left(L^s(\tau,h)\right) \geq \theta + \tau\frac{\beta_t \cdot q_0}{h^2}\lambda_{min}\left(M_y^s\right) \quad \text{for } h \in \left[(1.5-\sqrt{2})4\beta_l, (1.5+\sqrt{2})4\beta_l\right],$$

whereas we have

$$\lambda_{min}\left(L^s(\tau,h)\right) \geq \theta + \tau\frac{\beta_t \cdot q_0}{h^2}\lambda_{min}\left(M_y^s\right) + \tau\frac{\beta_l \cdot q_0}{h^2}\cdot\left(2 - \frac{d^2}{\sqrt{d^2-1}}\right)$$

for $h \notin [(1.5-\sqrt{2})4\beta_l, (1.5+\sqrt{2})4\beta_l]$, where $d = -1.5 + 0.5b$. The possibly negative eigenvalues $\lambda_{min}\left(M_x^s\right)$ and $\lambda_{min}\left(M_y^s\right)$ get amplified by the numbers $\frac{\beta_l \cdot q_0}{h^2} > 0$ and $\frac{\beta_t \cdot q_0}{h^2} > 0$, respectively. These factors increase for $h \to 0$. Suppose a time step size $\tau > 0$ is given. Then we need to calculate a minimal step size h_0 such that

$$\theta + \tau \frac{\beta_l \cdot q_0}{h_0^2} \lambda_{min}\left(M_x^s\right) + \tau \frac{\beta_t \cdot q_0}{h_0^2} \cdot \lambda_{min}\left(M_y^s\right) > 0$$

holds and therefore our matrix $L(\tau, h)$ is positive definite and nonsingular. Then we can solve our linear system with all step sizes $h \geq h_0$. Here it is important to have an accurate lower bound for $\lambda_{min}(M_x^s)$ and $\lambda_{min}(M_y^s)$ so that we can use step sizes h as small as possible.

Example 7.4.3. We set $\omega_x = 10$ and $\omega_y = 6$ and therefore use the domain $\Omega = [0, 10] \times [0, 6]$. We further use the scalars $\tau = 0.1, \beta_l = 0.3, \beta_t = 0.03, q_0 = 0.18$ and $\theta = 0.3$. Depending on the choice of h, we now get different matrix sizes and eigenvalues. In the following table we compare the lower bound of $\lambda_{min}\left(L^s(\tau, h)\right)$ according to our theory (column 'λ_{min} lower bound') with the exact eigenvalue calculated from the corresponding system matrix with the MATLAB function eigs (column 'λ_{min} exact').

h	n	m	size	λ_{min} exact	λ_{min} lower bound
0.5	20	12	260	0.300412963667855	0.2999550000
0.2	50	30	1550	0.300213215599023	0.2997975000
0.1	100	60	6100	0.299416896200867	0.2991835345
0.05	200	120	24200	0.289617314238473	0.2896089457
0.02	500	300	150500	0.170625390123707	0.1705729827
0.01	1000	600	600600	−0.324076096559832	−0.3242857259

We see that the lower bounds obtained from our theory are very sharp. In fact, a rounding process after the first three digits gives identical values for all different matrix sizes.

From Lemma 7.2.3 we know that our estimate for $\lambda_{min}\left(L^s(\tau, h)\right)$ is also a lower bound for $\sigma_{min}\left(L(\tau, h)\right)$ as long as $L^s(\tau, h)$ is positive semidefinite, i.e., for all step sizes except $h = 0.01$. However, it is clear from Lemma 7.2.3 that this lower bound will be much less accurate, especially when the matrix $L(\tau, h)$ is far away from being symmetric (this will be the case for smaller values of h). Nevertheless, we will give a comparison of our lower bound for σ_{min} with prior results in this area. To this end, let us define the values $r_k(A) := \sum_{j=1, j\neq k}^{n} |a_{kj}|$ and $c_l(A) := \sum_{i=1, i\neq l}^{n} |a_{il}|$ for an arbitrary matrix $A = [a_{ij}] \in \mathbb{R}^{n \times n}$. Then, Johnson [19, Theorem 3] showed that

$$\sigma_{min}(A) \geq \min_{i=1,\dots,n} \left\{ |a_{ii}| - \frac{r_i(A) + c_i(A)}{2} \right\}. \tag{7.15}$$

whereas Johnson and Szulc [20, Theorem 2] proved the lower bound

$$\sigma_{min}(A) \geq \min_{i=1,\dots,n} \left\{ \sqrt{|a_{ii}|^2 + \left(\frac{r_k(A) - c_k(A)}{2} \right)^2} - \frac{r_k(A) + c_k(A)}{2} \right\}. \tag{7.16}$$

Another interesting lower bound was given by Qi [37, Theorem 3]:

$$\begin{aligned}&\sigma_{min}(A) \geq \max\{0, \min\{l_1, \ldots, l_n\}\} \qquad \text{with}\\ &l_i := \min\left\{\sqrt{a_{ii}^2 + a_{ii}r_i(A) + \frac{c_i(A)^2}{4}} - \frac{c_i(A)}{2}, \sqrt{a_{ii}^2 + a_{ii}c_i(A) + \frac{r_i(A)^2}{4}} - \frac{r_i(A)}{2}\right\}.\end{aligned} \tag{7.17}$$

Finally Li [31, Theorem 2] introduced the following lower bound for a matrix A which has no isolated vertex:

$$\sigma_{min}(A) \geq \min_{(i,j)\in E(A)} \{g_{ij}\}, \tag{7.18}$$

where

$$g_{ij} = \frac{|a_{ii}| + |a_{jj}|}{2} - \frac{1}{2} \cdot \left[(|a_{ii}| - |a_{jj}|)^2 + (r_i(A) + c_i(A))(r_j(A) + c_j(A))\right]^{1/2}$$

and $(i, j) \in E(A)$ if and only if $a_{ij} \neq 0$ or $a_{ji} \neq 0$. This bound is an improvement of a corresponding result in [20]. Further lower bounds for the smallest singular value may be found in [42], but they are based on the determinant of A which is expensive to compute in our case.

In the following table, we compare these estimates with our estimate for $L(\tau, h)$ for different step sizes h.

h	σ_1 exact	σ_1 l.b.	(7.15)	(7.16)	(7.17)	(7.18)
0.5	0.30046	0.29996	0.29712	0.29713	0.29689	0.28831
0.2	0.30031	0.29980	0.24825	0.25049	0.22481	0.05821
0.1	0.30005	0.29918	0.048	0.06919	0.00000	− 0.89218
0.05	0.29978	0.28960	−0.798	−0.68005	0.08769	− 4.91874
0.02	0.30006	0.17057	−6.9	−6.04810	0.20315	−35.69999

The column entitled “σ_1 l.b.” is our lower bound. We see that the estimates from (7.15)–(7.18) all become zero or negative at a certain stage (and, hence, are useless as a lower bound for the smallest singular value). Furthermore, our lower bound is (much) better in almost all situations.

8 Final Remarks

In the previous chapters we have presented a special application from the field of hydrogeology and its modeling as a PDE-ODE-AE-CC system. After some transformations and discretization we finally obtained a nonlinear nonsmooth equation system formulated either with the minimum function or the Fischer-Burmeister function. Both nonlinear nonsmooth equation systems were solved with the semismooth Newton method. The minimum function approach enabled the full exploitation of the structure of the linear equation systems arising from Newton's method. This resulted in a very specialized and efficient algorithm and in strong theoretical results. The Fischer-Burmeister approach did not allow the exploitation of the arising linear equation systems. But this formulation enabled the globalization of the semismooth Newton method. Also good theoretical results could be shown for this globalized Newton method. The numerical behavior of the MinSchur algorithm corresponding to the Schur complement approach for the minimum function formulation proved to be stable and much more efficient then the globalized algorithm FBglob stemming from the Fischer-Burmeister formulation. In our numerical test runs there was never a need for the use of a globalization technique, because the starting vectors from the previous time step were always good enough. But this algorithm could very well be useful for problems with bad starting vectors.

In [25] the author studied a more general model then we did. There he also considered kinetic reactions and so-called sorption reactions. One interesting question would be: can the strategy presented in this thesis be adapted to that more general case? For the answer one must distinguish if only kinetic reactions or only sorption reactions or both are added to our model. The strategy from Section 4.4 for τ could always be adapted. Much of what we did in Chapter 4 and 5 was subject to the special structure of our problem especially to the structure of the elements of $\partial_B G_M, \partial_B G_F$. Adding both kinetic and sorption reactions to our model would completely alter the structure of these matrices. In that case I reckon that it is hardly possible to give considerably stronger results then the ones that Kräutle brought in [25]. If we add only sorption reactions (and species) to our model, still the structure of the problem would be altered so much, that the strategies in this thesis could not be applied. This problem could be a subject of a different research project, where the results of this thesis might be helpful. The third and last scenario would be to add only kinetic reactions to our model. Hereby the structure of our problem is alerted in such a way that the strategies presented in the previous chapters could be adapted. Some results would still be valid unchanged and some results would be weaker. For example Theorem 4.5.2, saying that all $H^{\mathcal{B}} \in \partial_D G(w)$ are nonsingular for $0 \leq \tau < \tau_{max}$ in every time step (note τ_{max} can be calculated), would not be valid any more. This statement would have to be replaced by: In every time step there is a constant τ_{max} such that all $H^{\mathcal{B}} \in \partial_D G(w)$ are nonsingular for $0 \leq \tau < \tau_{max}$ (and τ_{max} could not be calculated). And of course any result that depends on this theorem would have do be adapted as well.

Finally we want to discuss briefly two open questions. Much effort was put in the solution of the first open question, yet without result. Assume that we have found a solution $[\eta, \xi_{min}, \xi_{mob}, \bar{c}]$ so that $G(\xi_{min}, \xi_{mob}, \bar{c}) = 0$ holds and η solves the decoupled equation system $(\theta I + \tau L_h) \cdot \eta = \theta \eta^{old}$. Then it is important to know: Are all components of

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

positive? The opposite would make no physical sense. Our numerical test runs would suggest $c > 0$ always holds. But can it be proved mathematically? Also it would be desirable to have a more powerful local existence result then that provided in Section 5.2. In Theorem 4.7.2 it was shown that there is an $\tau_s > 0$ and a function $g : [0, \tau_s) \longrightarrow \mathcal{P} \times \mathbb{R}^{J_{min} p}$ such that

$$F\,(\tau, g(\tau)) = 0\,,\ \ \forall \tau \in [0, \tau_s)$$

and

$$g(0) = \left[\eta^{old}, \xi^{old}_{min}, \xi^{old}_{mob}, \bar{c}^{old}\right] \in \mathcal{P} \times \mathbb{R}^{J_{min} p}$$

hold (with the notation of that section). Remember that $\left[\eta^{old}, \xi^{old}_{min}, \xi^{old}_{mob}, \bar{c}^{old}\right]$ is the solution of the previous time step and $\mathcal{P}$ is the set of all vectors $\eta, \xi_{min}, \xi_{mob}$ such that $c\,(\xi_{min}, \xi_{mob}, \eta) > 0$ holds. But this theorem says nothing about the magnitude of τ_s. It would be desirable to have a minimal constant $\tau_{min} > 0$ such that $\tau_s \geq \tau_{min}$ always holds so that τ_s can't tend to zero. This matter is connected to the previous question because $\tau_s > 0$ must be chosen in such a way, that $g(\tau)$ stays in $\mathcal{P} \times \mathbb{R}^{J_{min} p}$ for all $\tau \in [0, \tau_s)$.

A Appendix

A.1 Determinant Sum Expansion Formula

The following result was used in Section 4.4. The result itself can be found in [8, p. 60] but without proof. Since we are not aware of an explicit reference containing the proof, we give the details here.

Theorem A.1.1. *Let $B, D \in \mathbb{R}^{n\times n}$ with D being a diagonal matrix, and let $M = D + B$. Then*

$$\det M = \sum_{\alpha\subset I} \det D^{\alpha,\alpha} \cdot \det B^{\bar{\alpha},\bar{\alpha}} ,$$

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

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

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

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

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

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

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

Now we iterate this and eventually get

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

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

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

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

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

since $\alpha \cap \{i\} \neq \emptyset$ guarantees, on the one hand that d_i is always on the diagonal of $D^{\alpha,\alpha}$, and, on the other hand that the index i does not belong to $\bar{\alpha}$ so that we can replace B_i by B. Now we can insert this result in (A.1) and get

$$\det M = \sum_{i=1}^{n+1} \left[\sum_{\substack{\alpha \cap \{i\} \neq \emptyset \\ \alpha \subset \{i, i+1, \ldots, n+1\}}} \det D^{\alpha,\alpha} \cdot \det B^{\bar{\alpha},\bar{\alpha}} \right] + \det B \, .$$

Now it holds that $\cup_{i=1}^{n+1} \{\alpha \mid \alpha \subset \{i, i+1, \ldots, n+1\} \, , \, \alpha \cap \{i\} \neq \emptyset\}$ equals the power set of $\{1, 2, \ldots, n+1\}$ off the empty set. Furthermore, for different i, two sets

$$\{\alpha \mid \alpha \subset \{i, i+1, \ldots, n+1\} \, , \, \alpha \cap \{i\} \neq \emptyset\}$$

do not have an intersection. Therefore α runs through every subset of $\{1, 2, \ldots, n+1\}$ once except for the empty set. But for the empty set, we have

$$\det D^{\emptyset,\emptyset} \cdot \det B^{\bar{\emptyset},\bar{\emptyset}} = \det B \, .$$

Hence we obtain

$$\det M = \sum_{\alpha \subset \{1,\ldots,n+1\}} \det D^{\alpha,\alpha} \cdot \det B^{\bar{\alpha},\bar{\alpha}} \ ,$$

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

A.2 The Determinant and Block Permutations

In this section we will show some simple results about the effects on the determinant of a matrix under special permutations of its rows or columns. The aim is to get an easy formula for swapping block rows or block columns of a matrix.

Let $v_1, v_2, \ldots, v_n \in \mathbb{R}^n$ and let $\sigma \in \mathfrak{S}_n$ be an arbitrary permutation. Then we define the real $n \times n$ matrix

$$V_\sigma := \left[v_{\sigma(1)},\ v_{\sigma(2)}, \ldots,\ v_{\sigma(n)}\right] .$$

As abbreviation we set $V := V_{id}$. With sign we denote the signum function, which assigns each permutation $\sigma \in \mathfrak{S}_n$ one of the numbers -1, $+1$. Our first Lemma is based on the fact that each permutation in $\mathfrak{S}_n$ can be written as a product of transpositions (swapping of two numbers). We leave the proof to the reader.

Lemma A.2.1. *For all permutations $\sigma \in \mathfrak{S}_n$ it holds*

$$\det(V_\sigma) = sign(\sigma) \cdot \det(V) .$$

Proof. trivial. □

Let $w_1, \ldots, w_n \in \mathbb{R}^n$ and $\sigma \in \mathfrak{S}_n$. Then we define the $n \times n$ matrix

$$W^\sigma := \begin{bmatrix} w_{\sigma(1)}^T \\ w_{\sigma(2)}^T \\ \vdots \\ w_{\sigma(n)}^T \end{bmatrix}$$

and with $W := W^{id}$ we again denote the unpermuted matrix. Since the determinant of a matrix and and its transposed version is the same we can directly deduce

Corollary A.2.2. $\det(W^\sigma) = sign(\sigma) \cdot \det(W)$.

Now let $m_i \in \mathbb{N}_0$ so that $\sum_{i=1}^4 m_i = n$. And let $V_i \in \mathbb{R}^{n \times m_i}$ and $W_i \in \mathbb{R}^{m_i \times n}$ for $i = 1, \ldots, 4$. We consider the matrices composed out of this matrix blocks. In the next result we study what happens if two neighboring blocks are switched.

Theorem A.2.3. *It holds*

$$\begin{aligned} \det([V_1 \mid V_2 \mid V_3 \mid V_4]) &= (-1)^{m_2 \cdot m_3} \cdot \det([V_1 \mid V_3 \mid V_2 \mid V_4]) \\ \det\left(\begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ W_4 \end{bmatrix}\right) &= (-1)^{m_2 \cdot m_3} \cdot \det\left(\begin{bmatrix} W_1 \\ W_3 \\ W_2 \\ W_4 \end{bmatrix}\right) \end{aligned}$$

Proof. It is sufficient to prove the first equation, since the second one is just the transposed version of the first one.

This proof is in two steps. First we find the permutation which causes this switch of blocks. Then we have to calculate the signum value of this permutation. We proceed with step one.

We write $V := [V_1 \mid V_2 \mid V_3 \mid V_4]$ with column vectors

$$V = \left[v_1, \ldots, v_{m_1} \mid v_{m_1+1}, \ldots, v_{m_1+m_2} \mid v_{m_1+m_2+1}, \ldots, v_{m_1+m_2+m_3} \mid v_{m_s+1}, \ldots, v_{m_s+m_4}\right]$$

with $m_s = m_1 + m_2 + m_3$.

Consider the permutations in cycle notation $\sigma_i = (m_1 + i, m_1 + i + 1, \ldots, m_1 + m_2 + i)$. Then σ_1 applied to V switches the column $v_{m_1+m_2+1}$ with the block V_2

$$V_{\sigma_1} = [V_1 \mid v_{m_1+m_2+1} \mid V_2 \mid v_{m_1+m_2+2}, \ldots, v_{m_1+m_2+m_3} \mid V_4] \; .$$

Applying σ_2 to V_{σ_1} switches V_2 with the column $v_{m_1+m_2+2}$ and results in

$$V_{\sigma_2 \circ \sigma_1} = [V_1 \mid v_{m_1+m_2+1}, v_{m_1+m_2+2} \mid V_2 \mid v_{m_1+m_2+3}, \ldots, v_{m_1+m_2+m_3} \mid V_4] \; .$$

We repeat this until we get

$$V_{\sigma_{m_3} \circ \sigma_{(m_3-1)} \circ \ldots \circ \sigma_1} = [V_1 \mid V_3 \mid V_2 \mid V_4] \; .$$

Now σ_i can be written with m_2 transpositions

$$\begin{aligned} \sigma_i \;&=\; (m_1 + i,\, m_1 + i + 1)\,(m_1 + i + 1,\, m_1 + i + 2) \\ &\quad\; (m_1 + i + 2,\, m_1 + i + 3) \ldots (m_1 + m_2 - 1 + i,\, m_1 + m_2 + i) \end{aligned}$$

this means that $sign(\sigma_i) = (-1)^{m_2}$. Using the homomorphism property of sign-function we can conclude

$$\begin{aligned} \operatorname{sign}(\sigma_{m_3} \circ \sigma_{m_3-1} \circ \ldots \circ \sigma_1) \;&=\; \prod_{i=1}^{m_3} \operatorname{sign}(\sigma_i) \\ &=\; \prod_{i=1}^{m_3} (-1)^{m_2} = ((-1)^{m_2})^{m_3} \\ &=\; (-1)^{m_2 \cdot m_3} \end{aligned}$$

and we have proved our formula. □

Remark A.2.4. We want to consider shortly the special case $m_2 = m_3$, i.e. the block matrices V_2 and V_3 have the exact same size. Now this two blocks can be switched just by exchanging column after column. That is we swap the i-th column of V_2 with the i-th column of V_3. This procedure yields the same result as the construction in the proof above. Here we need exactly m_2 column changes and therefore we get

$$\det([V_1 \mid V_2 \mid V_3 \mid V_4]) = (-1)^{m_2} \cdot \det([V_1 \mid V_3 \mid V_2 \mid V_4]) \; .$$

But since $m_2 \equiv (m_2)^2 \;\, mod\, 2$ this result does not contradict the Theorem above.

In Theorem A.2.3 we were swapping neighboring blocks of a matrix. But this result may not hold if the blocks are not neighboring.

A.3 Results for the Spectral Norm

Here we bring the proof of a fairly known result. It can be found in [17, Sec. 5.6] but without proof. We have used it in Section 6.3.

Theorem A.3.1. *Let $A \in \mathbb{R}^{m \times n}$. Then*

$$\|A\|_{sp} \leq \sqrt{\|A\|_C \cdot \|A\|_R}\,.$$

Proof. Let $M \in \mathbb{R}^{s \times r}$ be arbitrary. Let $\lambda \in \mathbb{C}$ be an eigenvalue of M such that $|\lambda| = \rho(M)$, where $\rho(M)$ is the spectral radius of M. Let $x \in \mathbb{R}^r$ be an eigenvector corresponding to λ with $\|x\|_\infty = 1$. Then it holds

$$|\lambda| = |\lambda| \cdot \|x\|_\infty = \|\lambda x\|_\infty = \|Mx\|_\infty \leq \|M\|_R \, \|x\|_\infty = \|M\|_R \ .$$

That means, that

$$\rho(M) \leq \|M\|_R$$

holds. Now we can conclude

$$\|A\|_{sp}^2 = \rho\left(A^T A\right) \leq \left\|A^T A\right\|_R \leq \left\|A^T\right\|_R \cdot \|A\|_R = \|A\|_C \cdot \|A\|_R$$

where we used the fact that $\|\cdot\|_R$ is sub-multiplicative. Applying the square root on both sides of this equation yields the assertion. □

The next result helps to calculate the spectral norm of a block diagonal matrix. To this end we need some definitions.

Let $m, n, r \in \mathbb{N}$ be arbitrary numbers. Furthermore let $n_i, m_i \in \mathbb{N} (i = 1, \ldots, r)$ be numbers such that $\sum_{i=1}^r n_i = n$ and $\sum_{i=1}^r m_i = m$ holds. For $i = 1, \ldots, r$ let $M_i \in \mathbb{R}^{m_i \times n_i}$. Then we define the block diagonal matrix M as

$$M = \operatorname{diag}\left(M_1, M_2, \ldots, M_r\right) \ .$$

Theorem A.3.2. *For the spectral norm of M is holds*

$$\|M\|_{sp} = \max_{i=1,\ldots,r} \|M_i\|_{sp} \ .$$

Proof. Let $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$. We partition the column vector x as

$$x = \left[x^1, x^2, \ldots, x^r\right]$$

such that $x^i \in \mathbb{R}^{n_i}$. With this notation we mean that the column vectors x^i are parts of x. Then it holds

$$\|x\|_2^2 = \sum_{i=1}^r \|x_i\|_2^2$$

and

$$\|Mx\|_2^2 = \sum_{i=1}^r \|M_i x_i\|_2^2 \ .$$

Since the functions $x \mapsto x^2$ and $x \mapsto \sqrt{x}$ are strictly increasing for positive arguments, we can swap the composition of 'max' with these functions. In the following equation chain x shall always be partitioned this way. Then

$$\begin{aligned}
\max_{||x||_2^2=1} ||Mx||_2^2 &= \max_{||x_1||_2^2+\ldots+||x_r||_2^2=1} \sum_{i=1}^{r} ||M_i x_i||_2^2 \\
&= \max_{\substack{\alpha_1+\ldots+\alpha_r=1 \\ \alpha_i\geq 0}} \sum_{i=1}^{r} \max_{||x_i||_2^2=\alpha_i} ||M_i x_i||_2^2 \\
&= \max_{\substack{\alpha_1+\ldots+\alpha_r=1 \\ \alpha_i\geq 0}} \sum_{i=1}^{r} \Big(\max_{||x_i||_2=\sqrt{\alpha_i}} ||M_i x_i||_2 \Big)^2 \\
&= \max_{\substack{\alpha_1+\ldots+\alpha_r=1 \\ \alpha_i\geq 0}} \sum_{i=1}^{r} \Big(\sqrt{\alpha_i} \max_{||x_i||_2=1} ||M_i x_i||_2 \Big)^2 \\
&= \max_{\substack{\alpha_1+\ldots+\alpha_r=1 \\ \alpha_i\geq 0}} \sum_{i=1}^{r} \Big(\sqrt{\alpha_i}\, ||M_i||_{sp} \Big)^2 \\
&= \max_{\substack{\alpha_1+\ldots+\alpha_r=1 \\ \alpha_i\geq 0}} \sum_{i=1}^{r} \alpha_i\, ||M_i||_{sp}^2 \\
&= \max_{i=1,\ldots,r} ||M_i||_{sp}^2 \\
&= \Big(\max_{i=1,\ldots,r} ||M_i||_{sp} \Big)^2 .
\end{aligned}$$

Then it holds

$$\max_{||x||_2^2=1} ||Mx||_2^2 = \Big(\max_{||x||_2=1} ||Mx||_2 \Big)^2 = ||M||_{sp}^2 .$$

And in summary we have

$$||M||_{sp}^2 = \Big(\max_{i=1,\ldots,r} ||M_i||_{sp} \Big)^2 .$$

Applying the square root on both sides of the equation yields the assertion. □

A.4 Calculating the Bound for $F(y)$ with MATLAB®

In this section we briefly list two MATLAB functions. Both functions implements the theory from Chapter 6 in a straightforward way. The first function calculates norm bounds of the matrix-valued function $F(y)$ for $y > 0$ according to Theorem 6.3.3 and the following corollary. The input arguments are the matrices S, V that define the function $F(y)$, cf. (6.1). The three return values are bounds for $F(y)$ corresponding to the spectral, column sum and row sum norm. Here is the listing of this function:

sbound.m

```
function [sp,C,R] = sbound(S,V)
```

```
%% This function calculates the bound of the function f(y),
%%  y>0 from chapter 6 "The Determinant Theory". The bound
%% is calculated according to the spectral, row-sum and
%% column-sum norm.
%% S,V must have the same number of rows and they both must
%% have at least so many rows as columns



[n1,m1] = size(S);
[n2,m2] = size(V);

if( n1 ~= n2 )
  error('S␣and␣V␣do␣not␣have␣the␣same␣number␣of␣rows!');
else
    n = n1;
end

if( n < m1 || n < m2 )
  error([ 'The␣matrices␣S␣and␣V␣must␣have␣at␣least' ...
                              'as␣many␣rows␣as␣columns!' ]);
end


nsub = nchoosek(n,m1);
Ssub = nchoosek(1:n, m1);

Serg = zeros(1, nsub);

for( i=1:nsub )
    Serg(i) = det(S(Ssub(i,:),:));
end

Werg = zeros(m1,m2);

for( i=1:m1 )
    for( j=1:m2 )
        T = S;
        T(:,i) = V(:,j);
        for( k=1:nsub )
            if( Serg(k) ~= 0 )
                Werg(i,j) = Werg(i,j) ...
                         + abs(det(T(Ssub(k,:),:))/Serg(k));
            end
        end
    end
end

C  = norm(Werg, 1);
R  = norm(Werg, inf);
sp = sqrt(C*R);
```

The second function computes the bound s from Lemma 6.4.1 for matrix D_1 corresponding to not only the spectral norm but also to the row-sum and column-sum norms. This function is specially fitted for the main problem of this thesis, cf. Chapter 3. The input arguments are the stoichiometric matrices S_{min} and S_{mob}. This is done by calling the previous function with different arguments. Here is the listing of this function.

exabound.m

```
function  [spmax,Cmax,Rmax] = exabound(Smin, Smob)
%% This function calculates for several norms the bounds for
%% D_1 Smin should be a I X J_min matrix and Smob should be
%% a I X J_mob matrix


%% Here n1 should be equal to n2. Then n1 = I
%% Furthermore m1=J_min  and m2=J_mob
[n1,m1] = size(Smin);
[n2,m2] = size(Smob);

%% some sanity checks
if( n1 ~= n2 )
  error('Smin␣and␣Smob␣do␣not␣have␣the␣same␣number␣of␣rows');
else
    n = n1;
end

if( n < m1 || n < m2 )
  error(['The␣matrices␣Smin␣and␣Smob␣must␣have' ...
                         '␣at␣least␣so␣many␣rows␣as␣columns']);
end



%% \hat{J} is the emptyset
[spmax,Cmax,Rmax] = sbound(Smob, Smin);



%% the next loop calculates all possible sets \hat{J} and
%% \hat{B}. The case \hat{J}={1,...,J_min} is excluded
%% because this results in \hat{B}=emptyset and therefore
%% V would be a Ix0 matrix.

for(k=1:m1-1)
    %% now we treat subsets \hat{J} with k element(s)
    %% then \hat{B} has m1-k element(s)

    nsub = nchoosek(m1,k);
    Ssub = nchoosek(1:m1, k);

    for i=1:nsub
        hat_J = Ssub(i,:);
        hat_B = setdiff(1:m1, hat_J);
        S = [Smin(:,hat_J), Smob];
        V =  Smin(:,hat_B);
        [sp,C,R] = sbound(S, V);
        spmax    = max( spmax, sp );
        Cmax     = max( Cmax, C );
        Rmax     = max( Rmax, R );
    end
end
```

Bibliography

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

[2] J. Bear: *Dynamics of fluids in porous media.* American Elsevier, New York, 1972.

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

[4] H. Buchholzer, C. Kanzow, P. Knabner, S. Kräutle: *The semismooth Newton method for the solution of reactive transport problems including mineral precipitation-dissolution reactions.* Technical Report, University of Würzburg, Würzburg, Germany, January 2010, submitted for publication.

[5] A. Böttcher and S. Grudsky: *Spectral Properties of Banded Toeplitz Matrices.* SIAM, Philadelphia, PA, 2005.

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

[7] F. H. Clarke: *Optimization and Nonsmooth Analysis*. John Wiley and Sons, New York, 1983.

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

[9] J.W. Demmel: *Applied Numerical Linear Algebra.* SIAM, Philadelphia, PA, 1997.

[10] T. De Luca, F. Facchinei, C. Kanzow: *A semismooth equation approach to the solution of nonlinear complementary problems*. Mathematical Programming 75, 1996, pp. 407–439.

[11] A. Fischer: *Solution of monotone complementary problems with locally Lipschitzian functions*. Mathematical Programming 76, 1997, pp. 513–532.

[12a] F. Facchinei and J.S. Pang: *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume I.* Springer, New York, NY, 2003.

[12b] F. Facchinei and J.S. Pang: *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume II.* Springer, New York, NY, 2003.

[13] F. Facchinei and J. Soares: *A new merit function for nonlinear complementarity problems and a related algorithm*. SIAM Journal on Optimization 7, 1997, pp. 225–247.

[14] G.H. Golub and C.F. van Loan: *Matrix Computations.* The Johns Hopkins University Press, Baltimore, MD, second edition 1989.

[15] M. Hintermüller and M. Ulbrich: *A mesh-independence result for semismooth Newton methods*. Mathematical Programming 101, 2004, pp.~151–184.

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

[17] R. Horn, C. Johnson: *Matrix Analysis*, Cambridge University Press, reprinted 1991.

[18] R. Horn, C. Johnson: *Topics in Matrix Analysis*, Cambridge University Press, 1991.

[19] C. Johnson: *A Gersgorin-type lower bound for the smallest singular value*. Linear Algebra and its Applications 112, 1989, pp. 1–7.

[20] C. Johnson and T. Szulc: *Further lower bounds for the smallest singular value*. Linear Algebra and its Applications 272, 1998, pp. 169–179.

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

[22] C. Kanzow: *An unconstrained optimization technique for large-scale linearly constrained convex minimization problems*. Computing 53, 1994, pp. 101–117.

[23] P. Knabner and L. Angermann: *Numerik partieller Differentialgleichungen.* Springer, Berlin, 2000.

[24] K. Königsberger: *Analysis 2, 3. überarbeitete Auflage.* Springer, Berlin, 2000.

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

[26] S. Kräutle: *The semismooth Newton method for multicomponent reactive transport with minerals.* Advances in Water Resources 34, 2011, pp. 137–151.

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

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

[29] B. Kummer: *Newton's method for non-differentiable functions.* In J. Guddat et al. (eds.): *Advances in Mathematical Optimization.* Akademie-Verlag, Berlin, 1988, pp. 114–125.

[30] B. Kummer: *Newton's method based on generalized derivatives for nonsmooth functions: Convergence analysis.* Lecture Notes in Economics and Mathematical Systems 382, Springer-Verlag,1991, pp.~171–194.

[31] L. Li: *Lower bounds for the smallest singular value.* Computers and Mathematics with Applications 41, 2001, pp. 483–487.

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

[33] F. Morel, J. Hering: *Principles and applications of aquatic chemistry*. Wiley, New York, 1993.

[34] T.S. Munson, F. Facchinei, M.C. Ferris, A. Fischer and C. Kanzow: *The semismooth algorithm for large scale complementarity problems*. INFORMS Journal on Computing 13, 2001, pp. 294-311.

[35] R. Mifflin: *Semismooth and semiconvex functions in constrained optimization*. SIAM Journal on Control and Optimization 15, 1977, pp. 957–972.

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

[37] L. Qi: *Some simple estimates for singular values of a matrix.* Linear Algebra and its Applications 56, 1984, pp. 105–119.

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

[39] L. Qi and J. Sun: *A nonsmooth version of Newton's method*. Mathematical Programming 58, 1993, pp. 353–368.

[40] H. Rademacher: *über partielle und totale Differenzierbarkeit von Funktionen mehrerer Variablen und über die Transformation der Doppelintegrale*. Mathematische Annalen 79, 1919, pp. 340–359.

[41] R. T. Rockafellar: *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[42] O. Rojo: *Further bounds for the smallest singular value and the spectral condition number.* Computers and Mathematics with Applications 38, 1999, pp. 215–228.

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

[44] Y. Saad: *Iterative Methods for Sparse Linear Systems.* SIAM, Philadelphia, PA, second edition 2003.

[45] C.I. Steefel and 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.

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

[47] J. Stoer and R. Bulirsch: *Introduction to Numerical Analysis.* Springer, New York, NY, third edition 2002.

[48] H. Weyl: *über beschränkte quadratische Formen, deren Differenz vollstetig ist.* Rend. Circ. Mat. Palermo 27, 1909, pp. 373–392.

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