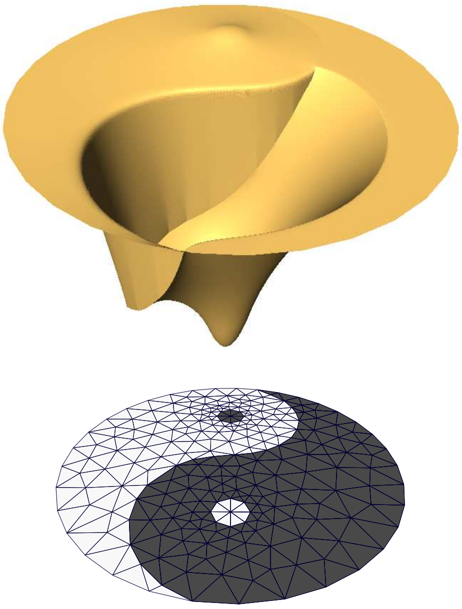
Inaugural-Dissertation
zur Erlangung der Doktorwürde
des Fachbereichs Mathematik und Informatik
der Freien Universität Berlin

Domain Decomposition Methods for Elliptic Problems with Jumping Nonlinearities and Application to the Richards Equation



vorgelegt von
Heiko Berninger
am 19. 10. 2007

Gutachter:
Prof. Dr. Ralf Kornhuber (Berlin)
Prof. Dr. Alfio Quarteroni (Lausanne)

Tag der Disputation: 12. 3. 2008

Betreuer: Prof. Dr. Ralf Kornhuber

Contents

Introduction

"Divide and conquer." — This saying has become one of the basic principles in mathematics and, in particular, in numerical analysis and scientific computing. And it will be a central feature of our approach to the problems which we study in this work. The *domain decomposition methods* we are going to apply to our problems — both on an analytical and on a computational level — are certainly the most prominent characteristic of this principle in our approach. Nonetheless, they are not the only one.

The starting and end point and the basis for the questions raised and investigated in this work is the *Richards equation*. It arises from the physical principle of mass conservation and serves as a mathematical model describing the saturated-unsaturated fluid flow through a porous medium in one single partial differential equation. As a consequence, the Richards equation assumes different types in different regions of the computational domain. In the saturated regime it is of elliptic type whereas it is parabolic in the unsaturated regime.

The equation contains two nonlinearities given by parameter functions, one of which is related to the time derivative while the other one is a factor in the spatial derivative. Moreover, it leads to heterogeneous problems since the parameter functions depend on the soil types given in the domain. The Richards equation is degenerate in the sense that the main part of the spatial derivative does not provide a uniformly elliptic operator because the relative permeability as a factor in it can become arbitrarily small for small pressure values in the fluid. Finally, the parameter functions may contain big slopes and can even degenerate into step functions for extreme soil parameters.

In view of these *heterogeneities and degeneracies* we propose a solution method for the Richards equation which does not rely on any linearization whatsoever. In order to realize this, however, we need to "divide" the problem into several partial problems which we can "conquer".

Therefore, in a first step we restrict ourselves to the Richards equation in a homogeneous setting, i.e. to a single soil type resulting in two fixed parameter functions. If, in addition, we neglect gravity in the equation, an implicit time discretization of the resulting problem provides quasilinear spatial problems which can be transformed into semilinear problems by an application of the *Kirchhoff transformation*. This simplification is not possible in case of spatially dependent parameter functions, i.e. in a heterogeneous setting.

Now, the semilinear problem, which we obtain for each time step after the time discretization and the Kirchhoff transformation, turns out to be equivalent to a *convex minimization problem*. This is even true if we treat the gravitational term explicitly in the time discretization, which we will do in this work. Whether gravity is included in this way or not, an existence and uniqueness result can be applied to the convex minimization problem and a convergence result for an appropriate finite element discretization can be derived. Moreover, with the *monotone multigrid method* a fast solver is at hand for the solution of the arising discrete problems. In addition, the performance of the monotone multigrid is *robust* with respect to varying soil parameters, even in the most extreme cases.

We point out that our solution method for the spatial problems in the homogeneous case is based on nonlinear minimization rather than linearization. The robustness of the method refers to the solution of the transformed problem, which only contains one nonlinearity, rather than to the solution of the original spatial problem, which is doubly nonlinear and degenerate for small pressure values. Here, our "divide and conquer" approach pays out on the practical level, since by virtue of the Kirchhoff transformation the degeneracy of the spatial problem is completely separated from the solution procedure. The degenerate character of the original problem reoccurs of course if we want to determine the solution in the physical pressure variables. Here, the inverse Kirchhoff transformation comes into play which is ill-conditioned for small pressure variables. But this inverse transformation is applied only once — after the solution has been calculated in the transformed variables.

Having understood the spatial problems arising from the Richards equation in the homogeneous case, we use this knowledge in the second step to address the situation in heterogeneous soil. Here, we consider the case of constant soil parameters on non-overlapping subdomains Ω_i, $i = 1, \ldots, n$, of a domain $\Omega \subset \mathbb{R}^d$ which change discontinuously across the interfaces between the subdomains. As a consequence, we obtain a problem with *jumping nonlinearities*, i.e. with nonlinear parameter functions which are fixed for each subdomain but different on each side of an interface separating the subdomains.

The global problem is formulated as a domain decomposition problem in which Richards equations in homogeneous soil on each subdomain are coupled by the continuity of the physical pressure p and the continuity of the normal fluid flux across the interfaces. Now, in order to apply the results from the homogeneous case, this domain decomposition problem is reformulated by the application of a Kirchhoff transformation on each subdomain. Since these transformations differ on different subdomains, the continuity condition on p now turns into nonlinear jump conditions for the transformed variables u_i on Ω_i, $i = 1, \ldots, n$, across the interfaces. As a consequence, we obtain a coupling of convex minimization problems on the subdomains with nonlinear transmission conditions given on the interfaces.

Now, we propose to solve this domain decomposition problem by iterative substructuring methods based on the transmission conditions which are well known for the solution of linear problems. More concretely, a *nonlinear Dirichlet–Neumann method* and a *nonlinear Robin method* are developed as generalizations of their linear counterparts. We point out that we apply these domain decomposition methods directly to the nonlinear substructuring problem *without any further linearization.* In particular, no linearization of the transmission conditions is involved.

To our knowledge, nonlinear domain decomposition methods of this kind have not yet been investigated in the literature. Here, we present an analysis for these methods and the underlying domain decomposition problems by generalizing existing linear theory on *Steklov–Poincaré operators* to our nonlinear case.

The main analytical results in this work concern the *convergence* of these nonlinear domain decomposition methods *in one space dimension* on two subdomains in case of relative permeabilities which are bounded from below by a positive constant. With these assumptions, the convergence of the damped Dirichlet–Neumann method and of the Robin method and *well-posedness* of the domain decomposition problem can be proved for the stationary Richards equation without gravity. Moreover, for the time-discretized Richards equation we obtain a convergence result for the Robin method as well as a theorem stating the well-posedness of the substructuring problem.

We emphasize that an existence and uniqueness result for the Richards equation in the heterogeneous case does not seem to be at hand so far. To our knowledge, the analysis on the Richards equation, which can be found in the literature, has only been developed for the Kirchhoff–transformed version, i.e. in the homogeneous case. The well-posedness of the domain decomposition problem for the time-discretized Richards equation, which we obtain in 1D for nondegenerate relative permeabilities, might serve as a starting point to establish an existence theorem for the Richards equation in heterogeneous soil at least in one space dimension.

Our one-dimensional analytical results are accompanied by *numerical results* for corresponding test cases *in two space dimensions.* In these examples we obtain convergence with reasonable convergence rates if the damping parameter in the Dirichlet–Neumann method or the acceleration parameter in the Robin method, respectively, is chosen appropriately. Unfortunately, we observe deteriorating convergence rates for the Robin method on higher levels, i.e. on fine grids. For the Dirichlet–Neumann method, however, we measure convergence rates and optimal damping parameters which are stable on higher levels. Therefore, even though our analytical results are more general for the Robin method than for the Dirichlet–Neumann method, the latter might be a promising tool for the solution of the Richards equation, too.

The last partial problem we have to deal with in view of a stable numerical solution method for the Richards equation is to find an appropriate space discretization of the gravitational term which is explicitly discretized in time. For this purpose we develop an artificial viscosity term in order to obtain an *upwind finite element discretization* which accounts for the constant direction of gravity. As a consequence, we obtain a numerically stable solution method with tolerable time step restrictions. Finally, we have a numerical example in 2D, wherein the Robin method is successfully applied to the Richards equation in four different soils and coupled with a surface water reservoir.

In the following we give an outline of this work which consists of four chapters. In Chapter 1 we introduce the Richards equation in homogeneous soil and study the Kirchhoff transformation and the parameter functions according to Brooks–Corey which we want to use in this work. We carry out a scaling of the Richards equation and take a look at hydrologically realistic, nondegenerate and limit cases. Furthermore, we present and investigate Signorini-type boundary value problems for the Richards equation and its Kirchhoff–transformed version in strong and weak formulations. Here, we also discuss the Kirchhoff transformation as a superposition operator. Finally, an overview of the analysis which can be found in the literature on initial boundary value problems for the Richards equation is given.

In Chapter 2 we discuss the numerical treatment of the Richards equation without gravity in homogeneous soil. We start with an overview of the numerics for the Richards equation to be found in the literature. Then we present our time discretization for the Richards equation with the secondary role attributed to the gravitational term. A uniquely solvable convex minimization problem is derived from the arising boundary value problem for which related or equivalent variational inequalities and variational inclusions are given. Furthermore, we present a finite element discretization of the convex minimization problem and prove a convergence result together with generalizations and consequences. Finally, in view of appropriate solvers for the discrete problems, monotone multigrid methods with the Gauss–Seidel relaxation as an essential ingredient are discussed and an asymptotic convergence theorem is given. The chapter ends with numerical results confirming the multigrid theory.

Chapter 3 is devoted to the Steklov–Poincaré theory for domain decomposition problems with jumping nonlinearities which are related to and motivated by the Richards equation. We start with a theorem on the substructuring of a Signorini-type problem for the Richards equation in homogeneous soil which serves as a definition for the heterogeneous case. Here, we also derive a Dirichlet–Neumann scheme for the time-discretized Richards equation in heterogeneous soil. Then we investigate a nonlinear Dirichlet–Neumann method applied to an elliptic transmission problem related to the nondegenerate stationary Richards equation without gravity. We give formulations via linear Steklov–Poincaré operators and prove convergence and well-posedness in one space dimension. Although we have counterexamples for our method of proof in 2D, a numerical example shows that the method can also be applied in two space dimensions. Moreover, we study a nonlinear Robin method for elliptic transmission problems related to the time-discretized nondegenerate

Richards equation. Here, we introduce nonlinear Steklov–Poincaré operators in which we formulate and analyse both the problem and the Robin method. The latter turns out to have an equivalent formulation as a nonlinear ADI method applied to the Steklov–Poincaré equation related to the problem. As before, we prove convergence and well-posedness in 1D. Then we discuss the Robin method applied to the time- and space-discretized Richards equation together with its convergence and its numerical treatment. Finally, we give numerical results obtained for the Robin method applied to the Richards equation without gravity and compare its performance to the performance of the Dirichlet–Neumann method in the stationary case.

In Chapter 4 we complete our numerical solution method for the Richards equation by appropriately addressing the explicitly time-discretized gravitational term with the help of an upwind finite element discretization. A numerical test in homogeneous soil demonstrates the stability and practicability of the resulting method. Finally, we present a numerical example in which we solve the Richards equation with the Robin method as the domain decomposition method. In this test case, which marks the end of this work, we include four different soil types and a surface water reservoir with realistic hydrological data.

I am deeply indebted to my supervisor Prof. Dr. Ralf Kornhuber for his invaluable support during the preparation of this work. Moreover, I am much obliged that Prof. Dr. Alfio Quarteroni agreed to render his expert opinion on my work. Also, I wish to thank Dr. Marco Discacciati for his rich expertise on Steklov–Poincaré operators by which he inspired me a lot. In addition, I am very grateful to my colleague Oliver Sander for his excellent work on the implementation of the algorithms. Furthermore, I would like to thank my colleague Carsten Gräser for many clever ideas on convex analysis and beyond. Finally, I think in gratitude of all the people around me these last years who appreciated both the mathematical and the non-mathematical contents of this work.

Berlin, October 2007 Heiko Berninger

Chapter 1

Richards equation in homogeneous soil

1.1 Introduction

"Very great rivers flow underground." — This citation by Leonardo da Vinci (see for example MacCurdy [67, III. 961.]) shall serve as a starting point of this work whose basis is the Richards equation which describes saturated-unsaturated groundwater flow and was first published in Richards [79].

At the beginning it seems to be in order to introduce the Richards equation by its hydrological derivation, which we give in Section 1.2, and then to look at it from different mathematical points of view, which we pursue in the subsequent sections. As a fist step we introduce the Kirchhoff transformation in Section 1.3 and apply it to the Richards equation. This transformation will be very helpful for the analytical treatment of the equation. Furthermore, it enables the numerical approach we want to apply in Chapter 2.

The next step, also carried out in Section 1.3, is the scaling of the Richards equation where we take a look at the equation and special parameter functions in a version in which the units have been eliminated. Together with realistic hydrological data given in Section 1.4 this will make clear the exact shape of the parameter functions which we choose according to Brooks and Corey in this work. Section 1.4 also contains a discussion of hydrologically relevant limit and nondegenerate cases arising from our parameter functions.

In Section 1.5 we derive different strong and weak variational inequalities for the Richards equation with Signorini-type boundary conditions which, for instance, occur in the case of the coupling of groundwater flow with a surface water reservoir. In this context we gain some important insights into the Kirchhoff transformation as a superposition operator which will also be helpful for the treatment of the heterogeneous Richards equation in Chapter 3. Finally, in Section 1.6 we give an overview of what is known about the analysis of initial boundary value problems for the Richards equation in different settings. The basic hydrological facts mentioned in this chapter can be found in Bear [13], Chavent and Jaffré [24] and van Genuchten [91].

1.2 Elements of hydrology

In this section we introduce some basic hydrological concepts and terms which occur throughout this work. In particular, we shall derive the Richards equation from fundamental physical laws. Finally, we discuss the Brooks–Corey functions which we use predominantly as the parameter functions in the Richards equation.

It is well known that the mathematical description of fluid flow is based on the principle of *mass conservation*. This principle states that for any control volume $\Omega' \subset \Omega$, where $\Omega \subset \mathbb{R}^3$ is the computational domain, the increase or decrease of mass of the fluid in Ω' within a certain time Δt is equalized by the flow of mass across $\partial\Omega'$ during that time, possibly influenced by an increase or decrease of mass given by a source or a sink in Ω'. In concrete terms, with $\Delta t \to 0$, we arrive at the mass conservation in the differential form

$$\frac{\partial}{\partial t} \int_{\Omega'} \theta n \varrho \, dx + \int_{\partial\Omega'} \theta n \varrho \, \bar{\mathbf{v}} \cdot \mathbf{n} \, d\sigma = \int_{\Omega'} \bar{f} dx \tag{1.2.1}$$

if we consider flow of water in a porous medium. In this formula $\theta n \varrho$ represents the rate of fluid mass per volume, $\bar{\mathbf{v}}$ is the microscopic velocity of the water and $\bar{f}$ is a source term. With the outer normal $\mathbf{n}$ of $\partial\Omega'$ the term $\bar{\mathbf{v}} \cdot \mathbf{n}$ gives the water flux across $\partial\Omega'$. To obtain the rate of mass per volume the water density ϱ is multiplied with the weighting factor θn in which n, the *porosity*, is a function on Ω giving the rate of void space per volume in the porous medium and θ, the *saturation*, gives the rate of water in this void space on Ω. We remark that in the hydrological literature θn is often called (volumetric) *water* or *moisture content* which, in contrast to our notation, is then denoted by θ. Taking into account the incompressibility of the water, i.e. $\varrho = \text{const}$, and assuming that the macroscopic velocity $\mathbf{v}$ is given by $\mathbf{v} = \theta n \bar{\mathbf{v}}$ while denoting $f = \bar{f}/\varrho$, an application of the divergence theorem A.1.2 to (1.2.1) gives the continuity equation

$$n\theta_t + \operatorname{div} \mathbf{v} = f\,. \tag{1.2.2}$$

Now, the following equation of motion, which is well known as the experimental *law of Darcy*, relates the water flow to the pressure of the water for any time t and states

$$\mathbf{v} = -K_c \nabla h\,. \tag{1.2.3}$$

In this formula the coefficient of proportionality K_c is called *hydraulic conductivity*. It is a scalar function if the flow takes place in an isotropic medium. In general it is a symmetric positive definite $3 \times 3-$matrix for any $x \in \Omega$. h is the so-called *piezometric head* which can be regarded as the groundwater level at the point $x = (x_1, x_2, z) \in \Omega$, see for example Bear [13, p. 64] or Richards [78]. It is related to the pressure of the water at x by the formula

$$h = \frac{p}{\varrho g} - z\,. \tag{1.2.4}$$

Here we assume that the z-axis of the coordinate system points downwards in the direction of gravity. Again ϱ is the density of the water, g is the gravitational constant and $p = p_w - p_a$ is the difference between the pressure of water and the constant pressure of air. The value $p/(\varrho g)$ is called *pressure head* ψ and comes from a hydrostatic pressure if $p > 0$ and from a *capillary pressure* or a suction if $p < 0$. In the first case the groundwater level is above the point x and in the second case it is below the point x.

Although our choice of the direction of the z-axis is contrary to how the piezometric head is measured, this choice is often made in the literature, such as in Chavent and Jaffré [24], Fuhrmann [40] or Eymard et al. [37]. As a consequence of (1.2.3) and (1.2.4) we obtain a positive component of water flux in the direction of the z-axis due to gravity. Note that we are free in the choice of the zero-level in z-direction because only spatial changes of h are relevant in (1.2.3). However, if we choose the zero-level in z-direction as the surface of the ground (in case it is horizontal), it is easy to see that $-h$ in (1.2.4) is just the (positive) distance between the groundwater level and the surface. As a consequence, in an isotropic medium Darcy's law (1.2.3) just states that at at any point in the ground the water flows to where the groundwater level above that point has its biggest decline.

If the saturation of water in the medium is maximal everywhere, $K_c = K_h(\cdot)$ is a function on Ω which is given by

$$K_h(x) = K(x)\,\mu^{-1}\varrho g \quad \forall x \in \Omega\,. \tag{1.2.5}$$

Here μ is the viscosity of the water. The remaining function $K(\cdot)$ on Ω is no longer dependent on the fluid and is called *permeability* of the soil. $K_h(x)$ can be regarded as a maximal hydraulic conductivity. In fact the law of Darcy (1.2.3) holds true in the more general setting of unsaturated soil. However, K_c is also dependent on the saturation θ of the water in that case. More concretely, we have

$$K_c(x, \theta) = K_h(x)\,kr(\theta) \tag{1.2.6}$$

in which $kr(\cdot)$ is the so-called *relative permeability* and provides a weighting factor in the interval $[0, 1]$. It is a monotonically increasing function of the saturation θ. In this general case $K_h(x)\,kr(\theta)$ is sometimes called *effective hydraulic conductivity* while $K(x)\,kr(\theta)$ is called *effective permeability.* It has to be pointed out that the concrete shape of the function kr is dependent on the soil so that the space dependency in (1.2.6) does not reflect spatial heterogeneity in full generality. Therefore, we still call the situation of the medium in (1.2.6) homogeneous soil. Figure 1.2 displays a typical shape of a relative permeability function $\theta \mapsto kr(\theta)$.

Finally, and again dependent on the soil, there is another equation of state which relates the saturation θ of the water to the pressure p. The function $p \mapsto \theta(p)$ is also monotonically increasing between a minimal saturation θ_m and a maximal saturation θ_M. In general we have $\theta_m > 0$ due to residual water and $\theta_M < 1$ due to residual air in the soil. For hydrostatic pressures $p \geq 0$ the saturation is maximal. Figure 1.1 shows an example of a function $p \mapsto \theta(p)$.

Altogether, if we put Darcy's law (1.2.3) and the equations of state into (1.2.2) with $f = 0$ we obtain the *Richards equation*

$$n(x)\theta(p)_t - \operatorname{div}\Big(K(x)\, \mu^{-1} kr(\theta(p))\nabla\big(p - \varrho g z\big)\Big) = 0 \tag{1.2.7}$$

for the unknown function p on $\Omega \times (0,T)$ with $T > 0$ in which

$$\mathbf{v} = -K(x)\, \mu^{-1} kr(\theta(p))\nabla\big(p - \varrho g z\big) \tag{1.2.8}$$

is the water flux. Obviously the Richards equation is a quasilinear elliptic-parabolic equation. More concretely, it is of elliptic type where the soil is fully saturated and parabolic in the unsaturated regime. In this case it might even be of hyperbolic type where $kr(\theta(p)) = 0$. In what is to come we will restrict ourselves to situations in which $kr(\theta(p))$ is always positive.

There are different methods how to obtain concrete analytical versions of the parameter functions $p \mapsto \theta(p)$ and $\theta \mapsto kr(\theta)$ for which van Genuchten [91] is the classical reference. Widely-used examples are the ones due to van Genuchten, which are applied in Fuhrmann [40] for instance, and the ones due to Brooks–Corey which shall be applied here (see also Section 1.3). Since the shapes of these fitting functions are very similar we restrict ourselves to presenting the Brooks–Corey model.

The main structural difference between the van Genuchten model and the Brooks–Corey model is that the van Genuchten functions are smooth and $\theta(p) < \theta_M$ holds for $p < 0$, whereas we have $\theta(p) = \theta_M$ for $p \geq p_b$ according to Brooks–Corey with a so-called *bubbling pressure* $p_b < 0$ in which the Brooks–Corey functions are non-differentiable. The latter model takes into account that the saturation remains maximal until the suction is large enough to suddenly allow air bubbles to enter the soil. However, the non-differentiability does not seem hydrologically essential. More refined models incorporate hysteresis effects which shall not be considered here. With a soil dependent parameter $\lambda > 0$ which is called the *pore size distribution factor* the so-called *soil-water retention curve* due to Brooks–Corey is given by

$$\Theta(p) := \frac{\theta(p) - \theta_m}{\theta_M - \theta_m} = \left[\frac{p}{p_b}\right]^{-\lambda} := \begin{cases} \left(\frac{p}{p_b}\right)^{-\lambda} & \text{for } p \leq p_b \\ 1 & \text{for } p \geq p_b\,. \end{cases} \tag{1.2.9}$$

In addition, using further theories, the following equation of state for the relative permeability is established for $\Theta \in (0,1]$ or $\theta \in (\theta_m, \theta_M]$, respectively:

$$kr(\theta) = \hat{kr}(\Theta) := \Theta^{e(\lambda)} \quad \text{with} \quad e(\lambda) := \begin{cases} 3 + \frac{2}{\lambda} & \text{due to Burdine} \\ \frac{5}{2} + \frac{2}{\lambda} & \text{due to Mualem.} \end{cases} \tag{1.2.10}$$

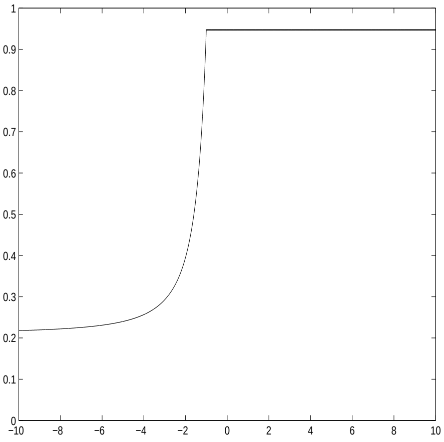


Figure 1.1: $p \mapsto \theta(p)$

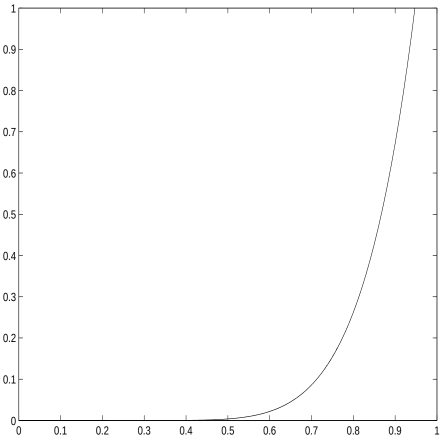


Figure 1.2: $\theta \mapsto kr(\theta)$

Thus, we obtain

$$kr(\theta(p)) = \hat{kr}(\Theta(p)) = \left[\frac{p}{p_b}\right]^{-\lambda e(\lambda)} . \tag{1.2.11}$$

According to van Genuchten [91], it is not clear whether the theory of Burdine, which was used by Brooks and Corey, or the theory of Mualem, which was found later, fits better to realistic data. Typical graphs of the Brooks–Corey functions are shown in Figures 1.1 and 1.2.

Considering the shape of the function $p \mapsto \theta(p)$ it becomes clear that the parameter λ is indeed related to how the pore sizes are distributed in the porous medium. If the pressure in the medium drops slightly below the bubbling pressure, then the capillaries with the biggest diameters are filled with air first. If the diameters of the pores in the medium vary a lot, i.e. if we have a big “pore size distribution”, then the pressure has to decrease considerably in order to drain the capillaries with small diameter, too. In this case the slope of the function $p \mapsto \theta(p)$ will be relatively small around the bubbling pressure p_b which means λ has to be relatively small. Consequently, $1/\lambda$ can be regarded as a measure for the pore size distribution. In the following section we will investigate the shape of the above functions in more detail and give concrete realistic values for the hydrological and the soil parameters involved in the Richards equation.

1.3 Kirchhoff transformation of the Richards equation and scaling with the Brooks–Corey parameter functions

In this section we introduce the Kirchhoff transformation which turns out to be a crucial tool both for the analysis and for our numerical treatment of the Richards equation in Chapters 2 and 4. We apply this transformation to the Richards equation with the Brooks–Corey parameter functions. Furthermore, regarding the concrete shape of the transformed functions, we carry out a scaling of the Richards equation in order to obtain the equation as well as the Brooks–Corey functions and the transformed functions in an *adimensional* form, i.e. in a form of a real-valued equation in which all units have been eliminated.

The Kirchhoff transformation is a crucial tool for simplifying a class of partial differential equations by eliminating certain nonlinearies (see, in particular, Alt and Luckhaus [4]). It is not only applied to problems dealing with saturated-unsaturated groundwater flow, where it is used in order to eliminate the relative permeability $kr(\cdot)$ in front of the gradient in (1.2.7). One also makes use of the Kirchhoff transformation in problems in which temperature-dependent material properties play the same role as $kr(\cdot)$ in our case, for example in the analysis of semiconductor devices (Bonani and Ghione [18]), thermoelasticity (Chen et al. [25]) or electrical eddy current fields (Breuer [20]).

For the Richards equation the Kirchhoff transformation $\kappa : \mathbb{R} \to \mathbb{R}$ is defined as follows:

$$\kappa : p \mapsto u := \int_0^p kr(\theta(q))\, dq\,. \tag{1.3.1}$$

The new variable u shall be called *generalized pressure.* The saturation as a function of u is denoted by

$$M(u) := \theta(\kappa^{-1}(u))\,. \tag{1.3.2}$$

Taking the chain rule into account which gives

$$\nabla u = kr(\theta(p))\nabla p \tag{1.3.3}$$

the transformed Richards equation (1.2.7) reads

$$n(x)M(u)_t - \operatorname{div}\Big(K(x)\,\mu^{-1}\big(\nabla u - kr(M(u))\varrho g \nabla z\big)\Big) = 0\,. \tag{1.3.4}$$

Thus, the transformed equation is a semilinear equation in which the nonlinearity in front of the spatial derivative has been eliminated.

Remark 1.3.1. We point out that the Kirchhoff transformation is of no use if the relative permeability kr is not only dependent on p but also explicitly on $x \in \Omega$, i.e. if we have $kr(x, \theta(p(x)))$. We can still carry out the Kirchhoff transformation (1.3.1) in this case, thus obtaining $u(x) = \kappa(x, p(x))$. Then, however, the chain rule provides

$$\nabla u(x) = kr(x, \theta(p(x)))\nabla p(x) + \nabla_x \kappa(x, p(x)) \quad \forall x \in \Omega$$

in which $\nabla_x \kappa(x, p(x))$ is to be understood as the vector of partial derivatives of $\kappa(\cdot, p)$ in its first d components corresponding to the entry $x \in \Omega \subset \mathbb{R}^d$. Therefore, the transformation does not simplify the Richards equation if kr is explicitly space-dependent. Nevertheless, such a case can be regarded as the full heterogeneous case. We address this problem in Chapter 3.

We will discuss the application and the validity of the chain rule in a weak formulation of the problem in Subsection 1.5.4 in more detail. At this stage we note that, although (1.3.3) can be understood in the classical sense here, usually we cannot assume $u \mapsto kr(M(u))$ to be spatially differentiable in case of the Brooks–Corey equations of state (1.2.9) and (1.2.11) which are both non-differentiable. However, as remarked in the previous section, this non-differentiability is not essential to the problem of groundwater flow. So at this stage one can either assume kr and M to be smooth (or smoothed) enough for the differential formulation or refer to our weak formulation of a boundary value problem for the Richards equation in Subsection 1.5.4.

The advantage of our choice of the parameter functions in (1.2.9) and (1.2.10) according to Brooks and Corey is that the Kirchhoff transformation and its inverse and the transformed functions involved in (1.3.4) can be given explicitly in a closed form. More concretely, from (1.2.11) we have

$$\begin{aligned} u = \kappa(p) &= \int_0^p \left[\frac{q}{p_b}\right]^{-\lambda e(\lambda)} dq \\ &= \begin{cases} \frac{p_b}{-\lambda e(\lambda)+1}\left(\frac{p}{p_b}\right)^{-\lambda e(\lambda)+1} + \frac{-\lambda e(\lambda) p_b}{-\lambda e(\lambda)+1} & \text{for } p \le p_b \\ p & \text{for } p \ge p_b\,. \end{cases} \end{aligned} \tag{1.3.5}$$

Obviously the generalized pressure is equal to the physical pressure in case of full saturation. In the unsaturated case, however, the interval $(-\infty, p_b)$ is mapped onto the bounded interval (u_c, p_b) in which we call

$$u_c := \frac{\lambda e(\lambda)}{\lambda e(\lambda) - 1}\, p_b < p_b \tag{1.3.6}$$

the *critical generalized pressure*. Consequently, the inverse transformation reads

$$p = \kappa^{-1}(u) = \begin{cases} p_b \left(\frac{u(-\lambda e(\lambda)+1)}{p_b} + \lambda e(\lambda)\right)^{\frac{1}{-\lambda e(\lambda)+1}} & \text{for } u_c < u \le p_b \\ u & \text{for } u \ge p_b\,. \end{cases} \tag{1.3.7}$$

Furthermore, the saturation as a function of the generalized pressure is given by

$$\begin{aligned} M(u) &= \theta(\kappa^{-1}(u)) = \theta_m + (\theta_M - \theta_m)\left[\frac{\kappa^{-1}(u)}{p_b}\right]^{-\lambda} \qquad (1.3.8) \\ &= \begin{cases} \theta_m + (\theta_M - \theta_m)\left(\frac{u(-\lambda e(\lambda)+1)}{p_b} + \lambda e(\lambda)\right)^{\frac{\lambda}{\lambda e(\lambda)-1}} & \text{for } u_c < u \le p_b \\ \theta_M & \text{for } u \ge p_b \end{cases} \end{aligned}$$

in which $M(u) \to \theta_m$ for $u \downarrow u_c$. Finally, the relative permeability as a function of u has the form

$$\begin{aligned} kr(M(u)) &= \left[\frac{\kappa^{-1}(u)}{p_b}\right]^{-\lambda e(\lambda)} \\ &= \begin{cases} \left(\frac{u(-\lambda e(\lambda)+1)}{p_b} + \lambda e(\lambda)\right)^{\frac{\lambda e(\lambda)}{\lambda e(\lambda)-1}} & \text{for } u_c < u \le p_b \\ 1 & \text{for } u \ge p_b\,. \end{cases} \end{aligned} \tag{1.3.9}$$

Now we want to investigate the shape of the Brooks-Corey parameter functions and the transformed functions obtained from the Kirchhoff transformation in more detail. Concretely, we are interested in how big the slope of these functions are in a situation with realistic hydrological data (see Section 1.4.1). To this end, we want to get rid of the physical units in the Richards equation and obtain only real-valued parameter functions and unknown functions in the problem. We scale the equation using characteristic unit values of the problem for the spatial coordinates, the time and the pressure.

For the sake of presentation in the rest of this section — *and exclusively in the rest of this section* — we want to alter our notation introduced in Section 1.2, where we had $x = (x_1, x_2, z) \in \Omega$, in the following way. Let $\mathbf{x} = (x, y, z) \in \Omega$ and x_0, y_0, z_0 be unit values for the corresponding coordinates. In addition, let t_0, p_0 and u_0 be unit values for the time, the pressure and the generalized pressure.

We introduce the transformation

$$\hat{x} := \frac{x}{x_0}\,, \;\; \hat{y} := \frac{y}{y_0}\,, \;\; \hat{z} := \frac{z}{z_0}\,, \;\; \hat{t} := \frac{t}{t_0}\,, \;\; \hat{\mathbf{x}} := \begin{pmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{pmatrix}, \;\; \hat{\nabla} := \begin{pmatrix} \frac{\partial}{\partial \hat{x}} \\ \frac{\partial}{\partial \hat{y}} \\ \frac{\partial}{\partial \hat{z}} \end{pmatrix},$$

obtaining $\hat{x}, \hat{y}, \hat{z} \in \mathbb{R}$ and $\hat{\mathbf{x}} \in \mathbb{R}^3$. Furthermore, we define the real functions

$$\hat{p}(\hat{\mathbf{x}}) := \frac{p(\mathbf{x})}{p_0} \quad \text{and} \quad \hat{u}(\hat{\mathbf{x}}) := \frac{u(\mathbf{x})}{u_0} \tag{1.3.10}$$

and

$$\hat{n}(\hat{\mathbf{x}}) := n(\mathbf{x})\,, \;\; \hat{\theta}(\hat{p}) := \theta(p)\,, \;\; \hat{K}(\hat{\mathbf{x}}) := K(\mathbf{x})\,, \;\; \hat{M}(\hat{u}) := M(u)\,, \tag{1.3.11}$$

where only $\hat{K}$ is not a real function, and keeping in mind that p and u also depend on t.

We have

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial \hat{x}} \frac{d\hat{x}}{dx} = x_0^{-1} \frac{\partial}{\partial \hat{x}} \quad \text{etc.}$$

leading to

$$\nabla = \operatorname{diag}(x_0^{-1}, y_0^{-1}, z_0^{-1}) \hat{\nabla} \quad \text{and} \quad \nabla \cdot F = \hat{\nabla} \cdot \big(\operatorname{diag}(x_0^{-1}, y_0^{-1}, z_0^{-1})\, F\big)$$

where $\operatorname{diag}(x_0^{-1}, y_0^{-1}, z_0^{-1})$ is the diagonal matrix with the entries x_0^{-1}, y_0^{-1} and z_0^{-1} and $F : \Omega \to \mathbb{R}^3$ is a differentiable vector field.

Finally, we define $e_{\hat{z}} := \hat{\nabla}\hat{z} = \nabla z =: e_z$ which is the unit vector in $\hat{z}$- or in z-direction, i.e. in the direction of gravity.

Now, from (1.2.7) we obtain the scaled Richards equation

$$\hat{n}(\hat{\mathbf{x}}) t_0^{-1} \hat{\theta}(\hat{p})_{\hat{t}} - \hat{\nabla} \cdot \Big(\Big[\hat{K}(\hat{\mathbf{x}}) \mu^{-1} kr(\hat{\theta}(\hat{p}))\, \operatorname{diag}(x_0^{-2}, y_0^{-2}, z_0^{-2}) \Big] \Big(\hat{\nabla}\, (p_0 \hat{p} - \varrho g z_0 \hat{z}) \Big) \Big) = 0 \tag{1.3.12}$$

and, analogously, from (1.3.4) we get the scaled Kirchhoff–transformed Richards equation

$$\hat{n}(\hat{\mathbf{x}})t_0^{-1}\hat{M}(\hat{u})_{\hat{t}} - \hat{\nabla}\cdot\Big(\Big[\hat{K}(\hat{\mathbf{x}})\mu^{-1}\,\mathrm{diag}\,(x_0^{-2},y_0^{-2},z_0^{-2})\Big] \Big(\hat{\nabla}(u_0\hat{u}) - kr(\hat{M}(\hat{u}))\varrho g\hat{\nabla}(z_0\hat{z})\Big)\Big) = 0 \quad (1.3.13)$$

in which we have diagonal matrices within the brackets $[\dots]$, respectively.

Now, in order to get rid of the units in these equations we define

$$\hat{K}_h(\hat{\mathbf{x}}) := K_h(\mathbf{x}) = \hat{K}(\hat{\mathbf{x}})\mu^{-1}\varrho g \quad (1.3.14)$$

with the hydraulic conductivity $K_h(\mathbf{x})$ from (1.2.5) which has the dimension of a velocity (assumed to be given in m/s). Furthermore, we define the following real-valued quotients of pressures

$$u_r := u_0(\varrho g z_0)^{-1}, \qquad p_r := p_0(\varrho g z_0)^{-1} \quad (1.3.15)$$

and the spatial scaling matrix

$$A_s := z_0\,\mathrm{diag}\,(x_0^{-2},y_0^{-2},z_0^{-2}) \quad (1.3.16)$$

which provides the unit m^{-1}. Then, altogether, if we consider t_0 as a real number given in the unit s, the scaled versions (1.3.12) and (1.3.13) provide the adimensional (i.e. unit-free) Richards equation

$$\hat{n}(\hat{\mathbf{x}})t_0^{-1}\hat{\theta}(\hat{p})_{\hat{t}} - \hat{\nabla}\cdot\Big(\Big[\hat{K}_h(\hat{\mathbf{x}})\,A_s\,kr(\hat{\theta}(\hat{p}))\Big]\,\hat{\nabla}\,(p_r\hat{p}-\hat{z})\Big) = 0 \quad (1.3.17)$$

and the adimensional Kirchhoff–transformed Richards equation

$$\hat{n}(\hat{\mathbf{x}})t_0^{-1}\hat{M}(\hat{u})_{\hat{t}} - \hat{\nabla}\cdot\Big(\hat{K}_h(\hat{\mathbf{x}})\,A_s\Big(\hat{\nabla}(u_r\hat{u}) - kr(\hat{M}(\hat{u}))\hat{\nabla}\hat{z}\Big)\Big) = 0\,. \quad (1.3.18)$$

Note that with our definitions in this section and with (1.2.2), (1.2.7) and (1.3.4) we obtain the physical water flux in the form

$$\begin{aligned}\mathbf{v} &= -\hat{K}_h(\hat{\mathbf{x}})\,kr(\hat{\theta}(\hat{p}))\,z_0\,\mathrm{diag}\,(x_0^{-1},y_0^{-1},z_0^{-1})\Big(\hat{\nabla}(p_r\hat{p}) - e_{\hat{z}}\Big)\\ &= -\hat{K}_h(\hat{\mathbf{x}})\,z_0\,\mathrm{diag}\,(x_0^{-1},y_0^{-1},z_0^{-1})\Big(\hat{\nabla}(u_r\hat{u}) - kr(\hat{M}(\hat{u}))e_{\hat{z}}\Big)\end{aligned}$$

and, with regard to (1.3.17) and (1.3.18), the formal flux as

$$\begin{aligned}\hat{\mathbf{v}} &= -\hat{K}_h(\hat{\mathbf{x}})\,kr(\hat{\theta}(\hat{p}))\Big(A_s\hat{\nabla}(p_r\hat{p}) - z_0^{-1}e_{\hat{z}}\Big)\\ &= -\hat{K}_h(\hat{\mathbf{x}})\Big(A_s\hat{\nabla}(u_r\hat{u}) - kr(\hat{M}(\hat{u}))z_0^{-1}e_{\hat{z}}\Big)\end{aligned}$$

which is clear from the transformation $\hat{\nabla}\cdot\hat{\mathbf{v}} = \nabla\cdot\mathbf{v} = \hat{\nabla}\cdot\big(\mathrm{diag}\,(x_0^{-1},y_0^{-1},z_0^{-1})\mathbf{v}\big)$, i.e.

$$\hat{\mathbf{v}} = \mathrm{diag}\,(x_0^{-1},y_0^{-1},z_0^{-1})\mathbf{v}\,.$$

Now, again, we take a look at our concrete choice of state equations, the Brooks–Corey functions (1.2.9) and (1.2.10) as well as their transformed versions and the Kirchhoff transformation given in (1.3.5) in an adimensional form. To this end, it is appropriate to set

$$u_0 := p_0 := -p_b \tag{1.3.19}$$

as the (positive) pressure unit. Then, with the notation in (1.3.10) and (1.3.11), the definitions (1.2.9) and (1.2.10) provide

$$\hat{\theta}(\hat{p}) = \begin{cases} (\theta_M - \theta_m)\,\hat{p}^{-\lambda} + \theta_m & \text{for } \hat{p} \le -1 \\ \theta_M & \text{for } \hat{p} \ge -1 \end{cases} \tag{1.3.20}$$

and

$$kr(\hat{\theta}(\hat{p})) = \left(\frac{\hat{\theta}(\hat{p}) - \theta_m}{\theta_M - \theta_m}\right)^{e(\lambda)} = \begin{cases} \hat{p}^{-\lambda e(\lambda)} & \text{for } \hat{p} \le -1 \\ 1 & \text{for } \hat{p} \ge -1\,. \end{cases} \tag{1.3.21}$$

By a straightforward definition of the adimensional Kirchhoff transformation, we obtain from (1.3.5)

$$\hat{\kappa}(\hat{p}) := \hat{u} = \frac{u}{-p_b} = \begin{cases} \frac{1}{\lambda e(\lambda)-1}(-\hat{p})^{-\lambda e(\lambda)+1} - \frac{\lambda e(\lambda)}{\lambda e(\lambda)-1} & \text{for } \hat{p} \le -1 \\ \hat{p} & \text{for } \hat{p} \ge -1\,. \end{cases} \tag{1.3.22}$$

With the adimensional critical generalized pressure

$$\hat{u}_c := -\frac{\lambda e(\lambda)}{\lambda e(\lambda) - 1} < -1 \tag{1.3.23}$$

due to (1.3.6) and (1.3.7) the adimensional inverse Kirchhoff transformation reads

$$\hat{p} = \hat{\kappa}^{-1}(\hat{u}) = \begin{cases} -((\lambda e(\lambda) - 1)\hat{u} + \lambda e(\lambda))^{-\frac{1}{\lambda e(\lambda)-1}} & \text{for } \hat{u}_c < \hat{u} \le -1 \\ \hat{u} & \text{for } \hat{u} \ge -1\,. \end{cases} \tag{1.3.24}$$

Furthermore, from (1.3.8) we obtain

$$\begin{aligned} \hat{M}(\hat{u}) &= M(u) \qquad (1.3.25) \\ &= \begin{cases} \theta_m + (\theta_M - \theta_m)((\lambda e(\lambda) - 1)\hat{u} + \lambda e(\lambda))^{\frac{\lambda}{\lambda e(\lambda)-1}} & \text{for } \hat{u}_c < \hat{u} \le -1 \\ \theta_M & \text{for } \hat{u} \ge -1 \end{cases} \end{aligned}$$

where $\hat{M}(\hat{u}) \to \theta_m$ for $\hat{u} \downarrow \hat{u}_c$. Finally, (1.3.9) gives

$$\begin{aligned} kr(\hat{M}(\hat{u})) &= kr(M(u)) \\ &= \begin{cases} ((\lambda e(\lambda) - 1)\hat{u} + \lambda e(\lambda))^{\frac{\lambda e(\lambda)}{\lambda e(\lambda)-1}} & \text{for } \hat{u}_c < \hat{u} \le -1 \\ 1 & \text{for } \hat{u} \ge -1\,. \end{cases} \end{aligned} \tag{1.3.26}$$

We will take a look at the realistic shape of these functions in the following section and come back to them again with concrete parameters when we deal with the numerical treatment of the Richards equation in Sections 2.8, 3.3.5 and 3.4.6 as well as in Chapter 4. However, since we have only introduced these adimensional functions with a ˆ for technical reasons, we will refer to these functions and also to the scaled, adimensional Richards equations (1.3.17) and (1.3.18) in the form without the ˆ from now on. Furthermore, as mentioned at the beginning of this section, we will consider the points $x \in \Omega$ whose last component we denote by z.

In view of a special time discretization of the Richards equation which we introduce in Section 2.3, a primitive Φ of M will be essential in order to treat the arising spatial problems. Note that M is monotonically increasing on (u_c, ∞) such that Φ turns out to be convex. We define this primitive as

$$\Phi(u) = \int_0^u M(s)\,ds \quad \forall u \in [u_c, \infty) \tag{1.3.27}$$

and obtain $\Phi(u) = \theta_M\, u$ for $u \geq -1$ and

$$\Phi(u) = \theta_m u + \frac{\theta_M - \theta_m}{\lambda(1+e(\lambda))-1}\left(\Big((\lambda e(\lambda)-1)u + \lambda e(\lambda)\Big)^{\frac{\lambda(1+e(\lambda))-1}{\lambda e(\lambda)-1}} - \lambda(1+e(\lambda))\right)$$

for $u_c < u \leq -1$ with the exponent

$$\frac{\lambda(1+e(\lambda))-1}{\lambda e(\lambda)-1} = \frac{(a+1)\lambda+1}{a\lambda+1} > 1$$

(where $a \in \{3, 5/2\}$ due to (1.2.10)) and the limit

$$\begin{aligned}\Phi(u_c) := \lim_{u \downarrow u_c} \Phi(u) &= \theta_m u_c - (\theta_M - \theta_m)\,\frac{\lambda(1+e(\lambda))}{\lambda(1+e(\lambda))-1} \\ &= -\frac{\lambda}{\lambda(1+e(\lambda))-1}\left((1+e(\lambda))\,\theta_M + \frac{\lambda}{\lambda e(\lambda)-1}\,\theta_m\right).\end{aligned}$$

1.4 Realistic situations and limit cases

The purpose of this section is to investigate the concrete shape of the functions given in (1.3.20)–(1.3.26) by considering a realistic hydrological situation. This will give rise to some limit cases for deteriorating soil parameters p_b and λ. With regard to realistic physical situations but also to mathematical conditions of nondegeneracy, we also take a look at the uniformly elliptic case $kr(\cdot) \geq c$ for some $c > 0$.

1.4.1 A hydrological example

Our realistic hydrological example was provided by hydrologists from the working group of Prof. Bronstert at the University of Potsdam. They have a research catchment of around $2000\,m \times 3000\,m \times 5\,m$ and measure bubbling pressures as small as $p_b = -0.1\,m$ water column. With an overflow of up to $2\,m$ they obtain maximal pressures of up to $7\,m$ water column in the ground. On the other hand, they observe capillary pressures as little as $-2\,m$ water column. With (1.3.10) and our convention (1.3.19), this gives the interval $[-20, 70]$ as the range for the variable p. We remark that extreme cases for the bubbling pressure which can be found in the literature Rawls et al. [77, Table 5.3.2] range from $p_b = -0.0136\,m$ for sand to $p_b = -1.872\,m$ for clay. The usual values are between $-0.1\,m$ (geometric mean for sand) and $-0.4\,m$ (geometric mean for clay).

The volumetric water content $n\theta$ attained its residual value, i.e. its minimum, at $n\theta_m = 0.08$ and its maximum at $n\theta_M = 0.36$. With a porosity in the soil of $n = 0.38$ we obtain approximately $\theta_m = 0.21$ and $\theta_M = 0.95$ (< 1 due to possible residual air in the soil). For completeness, we note the value for the hydraulic conductivity in case of full saturation $K_h = 0.002\,m/s$ which can easily vary by one order of magnitude even for largely homogeneous soil.

As far as the scaling factor for the pressure in (1.3.15) is concerned setting $u_0 = p_0 = -p_b = 0.1\,m$ we obtain $u_r = 0.02$ here since $\varrho g z_0$ represents a pressure of $5\,m$ water column. Furthermore, we point out that with the size of the research catchment above, we obtain an anisotropy in the scaled equation reflected by the scaling matrix (1.3.16)

$$A_s = 5\,m\ \mathrm{diag}((4{\cdot}10^6 m^2)^{-1}, (9{\cdot}10^6 m^2)^{-1}, (25\,m^2)^{-1}) \approx \mathrm{diag}(10^{-6}, 10^{-6}, 1)\,m^{-1}.$$

This anisotropy requires a special treatment in the numerical solution of the Richards equation which shall not be considered here. For a linear analogue we refer to Wittum [100].

Apart from the range of pressure and saturation, the crucial value for the shape of the adimensional parameter functions (1.3.20) and (1.3.21) and the functions in (1.3.22)–(1.3.26) is certainly the pore size distribution factor λ. For the loamy sand-type soils of the research catchment this value was between 0.48 and 0.65. It turns out that with the example $\lambda = 2/3$ already, the parameter functions have quite big slopes and partly look like step functions. This does not change much for other choices, even in extreme cases of $\lambda = 1.090$ for sand or $\lambda = 0.037$ for clay (the arithmetic mean values in sand and clay are $\lambda = 0.7$ and $\lambda = 0.1$, respectively) which can be found in Rawls et al. [77, Table 5.3.2].

For our choice of λ we apply Burdine's theory in (1.2.10) to obtain the most relevant terms:

$$\lambda = \frac{2}{3} \implies \begin{cases} e(\lambda) = 6 \\ \lambda e(\lambda) = 4 \\ \frac{\lambda}{\lambda e(\lambda)-1} = \frac{2}{9} \quad \text{(-th root in } M) \\ u_c = -\frac{\lambda e(\lambda)}{\lambda e(\lambda)-1} = -\frac{4}{3} \end{cases}$$

With these values, the soil water retention curve is given by

$$\theta(p) = \theta_m + (\theta_M - \theta_m)[-p]^{-\frac{2}{3}}$$

and the permeability as a function of the saturation reads

$$kr(\theta) = \left(\frac{\theta - \theta_m}{\theta_M - \theta_m}\right)^6 .$$

The graphs of these two functions and also of the functions below are shown on pages 20 and 21. Since the following functions of the physical pressure p or the generalized pressure u contain big slopes, the corresponding graphs are given both on the full, expected range of p or of u, respectively, and on a smaller neighbourhood around the adimensional bubbling pressure -1.

The permeability as a function of the pressure reads

$$kr(\theta(p)) = [-p]^{-4} .$$

The Kirchhoff transformation gives the relationship between the pressure and the generalized pressure by

$$u = \kappa(p) = \begin{cases} \frac{1}{3}(-p)^{-3} - \frac{4}{3} & \text{for } \ p \leq -1 \\ p & \text{for } \ p \geq -1 \,. \end{cases}$$

The inverse Kirchhoff transformation reads

$$p = \kappa^{-1}(u) = \begin{cases} -(3u+4)^{-\frac{1}{3}} & \text{for } \ -\frac{4}{3} < u \leq -1 \\ u & \text{for } \ u \geq -1 \,. \end{cases}$$

The saturation as a function of the generalized pressure has the form

$$M(u) = \begin{cases} 0.21 + 0.74\,(3u+4)^{\frac{2}{9}} & \text{for } \ -\frac{4}{3} < u \leq -1 \\ 0.95 & \text{for } \ u \geq -1 \,. \end{cases} \tag{1.4.1}$$

Finally, the permeability as a function of the generalized pressure is given by

$$kr(M(u)) = \begin{cases} (3u+4)^{\frac{4}{3}} & \text{for } \ -\frac{4}{3} < u \leq -1 \\ 1 & \text{for } \ u \geq -1 \,. \end{cases}$$

Note that by considering the full, expected range for p or u in the graphs of these functions we account for the influence of the bubbling pressure p_b on the shape of the functions. The bubbling pressure scales this range as in (1.3.5)–(1.3.9) at the beginning of this section and has been set as the negative pressure unit (1.3.19) later on. As already mentioned above, in the most extreme case to be found in Rawls et al. [77, Table 5.3.2] we have $p_b = 1.36 \cdot 10^{-2}\,[m]$ for very coarse sand.

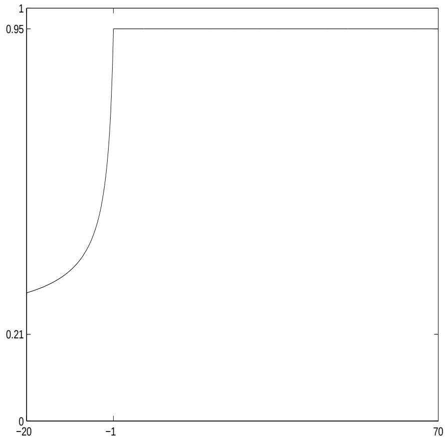


Figure 1.3: $p \mapsto \theta(p)$: full range

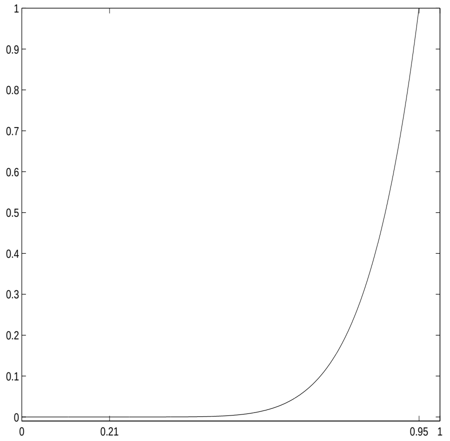


Figure 1.4: $\theta \mapsto kr(\theta)$

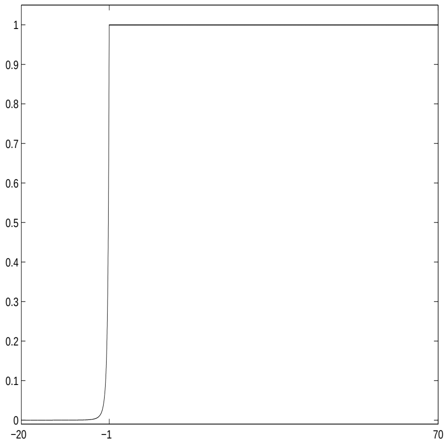


Figure 1.5: $p \mapsto kr(\theta(p))$: full range

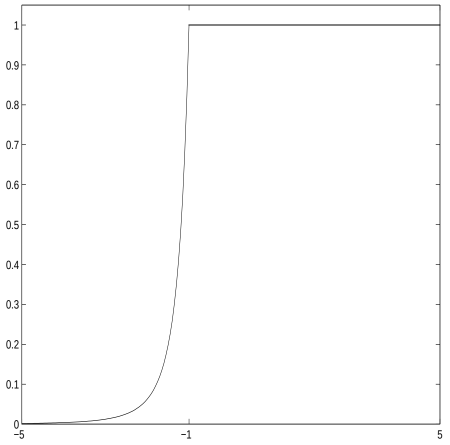


Figure 1.6: $p \mapsto kr(\theta(p))$: zoomed

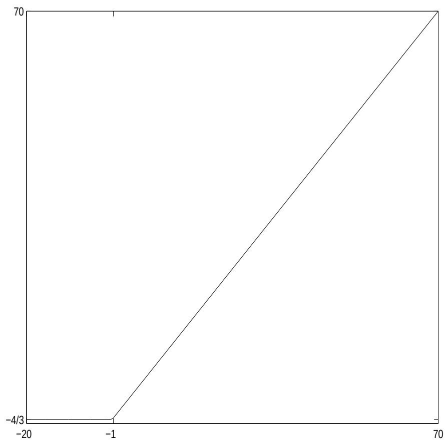


Figure 1.7: $p \mapsto \kappa(p)$: full range

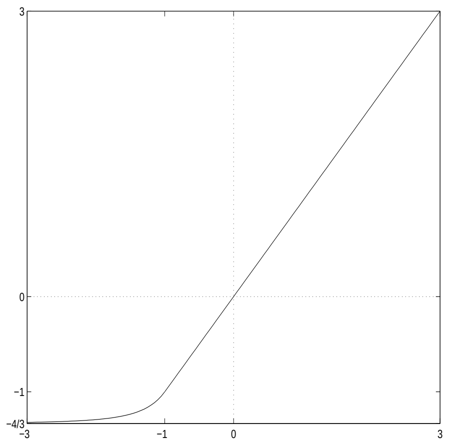


Figure 1.8: $p \mapsto \kappa(p)$: zoomed

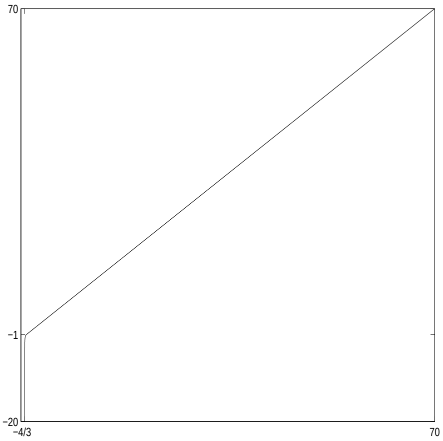


Figure 1.9: $u \mapsto \kappa^{-1}(u)$: full range

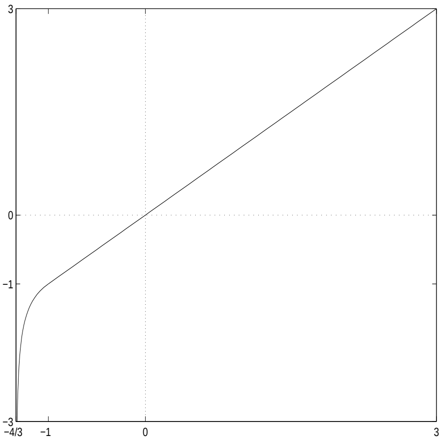


Figure 1.10: $u \mapsto \kappa^{-1}(u)$: zoomed

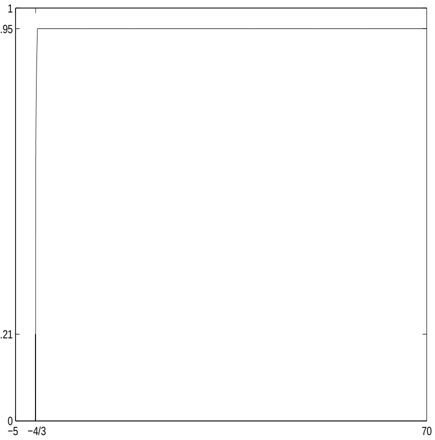


Figure 1.11: $u \mapsto M(u)$: full range

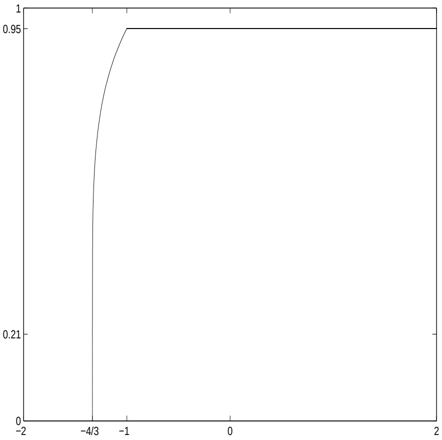


Figure 1.12: $u \mapsto M(u)$: zoomed

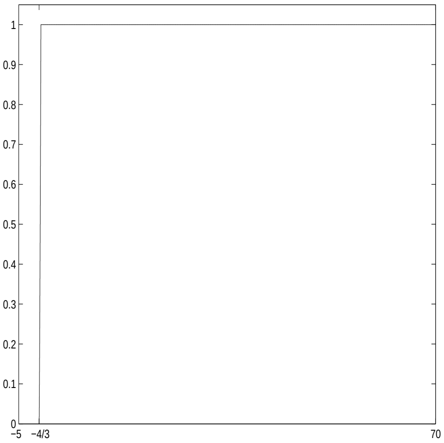


Figure 1.13: $u \mapsto kr(M(u))$: full range

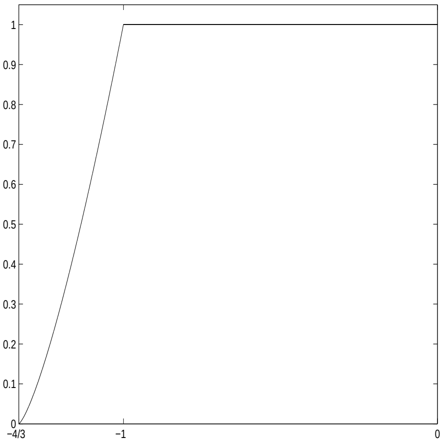


Figure 1.14: $u \mapsto kr(M(u))$: zoomed

1.4.2 Limit cases for the Brooks–Corey functions

Since the graphs of the functions shown on pages 20 and 21 already look quite extreme (i.e. quite steep) for our realistic hydrological data, it is natural from a mathematical point of view to ask how the functions in (1.3.20)–(1.3.26) behave for deteriorating soil parameters p_b and λ. Of course, such considerations are particularly important with respect to the robustness of numerical solution methods (see Section 2.8). We point out that especially the shape of M is interesting for the solution method, and also the inverse Kirchhoff transformation κ^{-1} which is, however, only applied to calculate the physical solution after one has already obtained the generalized solution. For the following discussion, we refer to the hydrological meaning of the soil parameters p_b and λ on page 10 and recall (1.2.10) as well as (1.3.23), i.e.

$$\lambda e(\lambda) = a\lambda + 2 \quad \text{and} \quad u_c = -\frac{a\lambda + 2}{a\lambda + 1} \tag{1.4.2}$$

with $a = 3$ according to Burdine and $a = 5/2$ according to Mualem. Both cases behave the same in the limits that we consider.

First, for $\lambda \to 0$ the slopes of the parameter function $p \mapsto \theta(p)$ in (1.3.20) decrease and the function becomes flatter, tending pointwise to the constant function

$$\theta_0 : p \mapsto \theta_M \quad \forall p \in \mathbb{R} \,.$$

In contrast, the slopes of $\theta \mapsto kr(\theta)$ in (1.2.10) increase and we get the step function kr_0 with

$$kr_0(\theta) = \begin{cases} 0 & \text{for } \theta_m \le \theta < \theta_M \\ 1 & \text{for } \theta = \theta_M \end{cases} \tag{1.4.3}$$

as the pointwise limit. The limit case seems hydrologically useless since it ignores the unsaturated case unless one defines something like

$$\theta_0(-\infty) := \theta_m \,. \tag{1.4.4}$$

Strangely enough, the function $p \mapsto kr(\theta(p))$ in (1.3.21) which turns into

$$k_0 : p \mapsto \begin{cases} p^{-2} & \text{for } p \le -1 \\ 1 & \text{for } p \ge -1 \end{cases} \tag{1.4.5}$$

and the Kirchhoff transformation do not reflect this singular situation. In the limit $\lambda \to 0$ we obtain

$$\lambda e(\lambda) \to 2 \quad \text{and} \quad u_c \downarrow -2$$

by (1.4.2) such that the interval $(-2, -1)$ accounts for the unsaturated regime. The inverse Kirchhoff transformation κ^{-1} in (1.3.24) tends pointwise to the function

$$\kappa_0^{-1} : u \mapsto \begin{cases} -(u+2)^{-1} & \text{for } -2 < u \le -1 \\ u & \text{for } u \ge -1 \end{cases}$$

On the one hand, this makes clear that for big λ the function κ^{-1} remains ill-conditioned due to big slopes for small pressures u. On the other hand, this is irrelevant in the limit case $\lambda \to 0$, in which the saturation M of the generalized pressure u in (1.3.25) tends pointwise to

$$M_0 : u \mapsto \theta_M \quad \forall u \in (-2, \infty) \tag{1.4.6}$$

while $kr(M(u))$ in (1.3.26) tends to 1 on $(-2, \infty)$. However, since one can extend M continuously by setting

$$M(u_c) := \theta_m$$

with $kr(M(u_c)) = 0$, one can extend the limit M_0 (discontinuously) in -2 with

$$M_0(-2) = M_0\left(\lim_{\lambda \to 0} u_c\right) := \lim_{\lambda \to 0} M(u_c) = \theta_m \tag{1.4.7}$$

reflecting the definition (1.4.4) since one could define

$$\kappa(-\infty) := u_c\,. \tag{1.4.8}$$

Another view on this case is the observation that the graph of M (which is a degenerating root function as the exponent in (1.3.25) vanishes) turns into the monotone graph

$$u \mapsto \left\{ \begin{array}{cl} [\theta_m, \theta_M] & \text{for } \ u = -2 \\ \theta_M & \text{for } \ u > -2\,. \end{array} \right. \tag{1.4.9}$$

It seems that the limit case for $\lambda \to 0$ can be regarded as hydrologically reasonable if one accepts the definitions in (1.4.4) and in (1.4.6)–(1.4.8). Obviously, this limit case produces a jump in the saturation (from θ_m to θ_M) and in the pressure (from $-\infty$ to $p \in \mathbb{R}$), respectively, across the wetting front in a soil (cf. [13, p. 303]), thus modelling the unsaturated regime in a degenerate way. We point out that such models are considered in the literature, e.g. for so-called *dam problems*, see page 52. Furthermore, we note that with our solution method, see Remark 2.4.5, Remark 2.7.5 and Section 2.8, we can also treat a version of this limit case generated by the maximal extension of the monotone graph in (1.4.9).

For $\lambda \to \infty$, while $kr(\cdot)$ (in (1.2.10) with (1.4.2)) becomes

$$kr_\infty : \theta \mapsto \left(\frac{\theta - \theta_m}{\theta_M - \theta_m}\right)^a \qquad \forall \theta \in [\theta_m, \theta_M]\,, \tag{1.4.10}$$

the parameter function $p \mapsto \theta(p)$ degenerates into a step function θ_∞ with

$$\theta_\infty(p) = \left\{ \begin{array}{ll} \theta_m & \text{for } \ p < -1 \\ \theta_M & \text{for } \ p \geq -1\,. \end{array} \right. \tag{1.4.11}$$

Analogously, so does $kr(\theta(\cdot))$ with the limit k_∞ satisfying

$$k_\infty(p) = kr_\infty(\theta_\infty(p)) = \left\{ \begin{array}{ll} 0 & \text{for } \ p < -1 \\ 1 & \text{for } \ p \geq -1\,. \end{array} \right. \tag{1.4.12}$$

Consequently, in the limit case the Kirchhoff transformation is no longer invertible for the unsaturated regime.

Observe that for $\lambda \to \infty$ we have

$$\lambda e(\lambda) \to \infty \quad \text{and} \quad u_c \uparrow -1$$

due to (1.4.2) such that the slopes of the functions κ^{-1}, M and $kr(M(\cdot))$ increase while the intervals $(u_c, -1)$ corresponding to the unsaturated regime become smaller and smaller. So as above, the inverse Kirchhoff transformation is also ill-conditioned for big λ, i.e. it *always* has big slopes for small generalized pressures u. Unfortunately, in contrast to the case $\lambda \to 0$, we obtain the constant function

$$M_\infty : u \mapsto \theta_M \quad \forall u \in [-1, \infty) \tag{1.4.13}$$

as the pointwise limit of M, in which nothing accounts for the unsaturated regime anymore. So this limit case does not seem to make sense hydrologically. Nevertheless, it can be given a sense if we observe that for $\lambda \to \infty$ the graph of M approaches the monotone graph

$$u \mapsto \begin{cases} [\theta_m, \theta_M] & \text{for } u = -1 \\ \theta_M & \text{for } u > -1 \end{cases} \tag{1.4.14}$$

which looks essentially like (1.4.9). A plausible remedy would now be to redefine

$$\theta_\infty(-1) := \theta_m \quad \text{and correspondingly} \quad M_\infty(-1) = \theta_m \tag{1.4.15}$$

with $kr_\infty(M_\infty(-1)) = 0$, thus assigning the unsaturated regime to the (normed) bubbling pressure -1. As a consequence, the value $p = -1$ (corresponding to $u = -1$) would play the same role as $p = -\infty$ (corresponding to $u = -2$) in the case $\lambda \to 0$ above, and the two limit cases would result in the same model. Interestingly, our analytical and numerical approach to this limit case described in Remark 2.4.5, Remark 2.7.5 and Section 2.8, is the same whether we carry out the redefinition (1.4.15) or not. As mentioned above for $\lambda \to 0$, the crucial aspect of this case is the argument u_c and not the value $M_\infty(u_c)$ as already suggested by (1.4.14).

Remark 1.4.1. With regard to variations of p_b, our functions in (1.3.20)–(1.3.26) do not seem to alter. However, in (1.3.19) we have defined $p_0 := -p_b$ as the (positive pressure) unit for these functions which has to be taken into account when p_b is variable. Note that p_b is a negative (capillary) pressure (corresponding to a suction). Decreasing the unit $-p_b$ results in "compressing" the functions (1.3.20) with respect to the p- and the u-axis while increasing $-p_b$ means "expanding" the functions (or their graphs). These transformations become clear if we set $p = \hat{p}\, p_0$ and $u = \hat{u}\, p_0$ in (1.2.9), (1.2.11) and (1.3.5)–(1.3.9) with $p_0 = u_0 \neq -p_b$ instead of (1.3.19) and vary p_b/p_0 for the resulting functions in $\hat{p}$ or $\hat{u}$, respectively. Note that for the Kirchhoff transformation and its inverse this scaling (i.e. compression or expansion of the axes) takes place on both axes while for the other functions in (1.3.20)–(1.3.26) it just applies to one (the $\hat{p}$- or $\hat{u}$-) axis.

As far as the Richards equation is concerned, especially if one is only interested in the saturation rather than in the pressure, the influence of the bubbling pressure is restricted to the scaling factor $u_r = p_r = -p_b(\varrho g z_0)^{-1}$ which is a measure for the size of the spatial derivative (or the elliptic term) in (1.3.17) and (1.3.18). Of course, p_b as a pressure unit has to be taken into account for the scale of the solution p when it comes to posing boundary conditions (see Section 1.5) and possibly initial conditions (if they are not given in the saturation, see Section 1.6) and, in particular, in situations where different bubbling pressures occur (see Section 2.8 and Chapter 4). In the latter case, one could also choose $p_0 = u_0 = (\varrho g z_0)^{-1}$ as a canonical pressure unit, leading to $p_r = u_r = 1$ in (1.3.15), and alter the parameter functions as just described.

Now, with a fixed p_0, it becomes clear from the observation in Remark 1.4.1 that $p_b/p_0 \to -\infty$ results in functions (1.3.7)–(1.3.9) with increasing support and in which both endpoints of the interesting interval accounting for the unsaturated regime go to $-\infty$. The limit case with the pointwise limits

$$\theta_{-\infty} = M_{-\infty} : u \mapsto \theta_M \quad \forall u \in \mathbb{R}$$

of θ in (1.2.9) and M in (1.3.8) no longer "sees" the unsaturated case unless one defines

$$\theta_{-\infty}(-\infty) = M_{-\infty}(-\infty) := \theta_m$$

which, in contrast to (1.4.9) and (1.4.14), does not result in a reasonable nontrivial numerical problem, see also Section 2.4.

The limit case $p_b/p_0 \uparrow 0$ is more interesting since here we obtain essentially the same situation as for $\lambda \to \infty$. We only need to replace -1 by 0 in the definition of the relevant functions for that case above. Observe that the redefinition of θ_∞ and correspondingly M_∞ in (1.4.15), which might seem somewhat artificial in these limit cases, is not necessary if we apply an altered Kirchhoff transformation in the form

$$\tilde{\kappa} : p \mapsto u = \int_{-\infty}^{p} kr(\theta(q))\, dq \tag{1.4.16}$$

which is possible here since the improper integral exists for the Brooks–Corey functions. This entails a translation of the functions (1.3.5) and (1.3.7)–(1.3.9) to the right by $|u_c|$ (given in (1.3.6)) such that the corresponding functions obtained by the altered transformation are defined on $(0, \infty)$. Then, except for $p_b/p_0 \to -\infty$, the limit cases look similar as above, now with $\tilde{u}_c = 0$ representing the unsaturated regime (without redefinition). For $p_b/p_0 \to \infty$ we would get $M(u) \to 0$ for all $u \in (0, \infty)$, but this is just due to the fact that the interesting range around the physical (atmospheric) pressure 0 is now mapped on $|u_c|$ by $\tilde{\kappa}$ with $|u_c| \to \infty$.

1.4.3 Altered Brooks–Corey functions for the nondegenerate Richards equation

One difficulty in the analysis and the numerics of the Richards equation is that the factor $kr(\theta(p))$ in the spatial derivative can become arbitrarily small (if p tends to $-\infty$), such that an implicit time discretization of (1.3.17) does not lead to uniformly elliptic spatial problems. On the other hand, physically reasonable solutions $p : \Omega \times (0,T) \to \mathbb{R}$ should be bounded. Therefore, and since we consider this situation in Chapter 3, we take a look at the *nondegenerate* Richards equation in which $kr(\cdot)$ is replaced by an altered relative permeability function $kr_\alpha(\cdot)$ satisfying the nondegeneracy condition

$$kr_\alpha(\theta) \geq \alpha \quad \forall \theta \in [\theta_m, \theta_M] \tag{1.4.17}$$

for a (small) $\alpha \in (0,1)$. Consequently, if $K(\cdot) \geq \alpha$ holds, too, the main part $\operatorname{div}(K(x)\mu^{-1}kr_\alpha(\theta(p))\nabla p)$ of the spatial derivative in (1.2.7) is a quasilinear *uniformly elliptic* operator (cf. [44, p. 203]) in the sense that (with (A.2.13))

$$c\,\|p\|_1^2 \leq \int_\Omega K(x)\mu^{-1}kr_\alpha(\theta(p))|\nabla p|^2\,dx \leq C\|p\|_1^2 \quad \forall p \in H_0^1(\Omega)$$

holds for certain $c, C > 0$. Whenever we speak of the uniformly elliptic case in this work, we have these inequalities in mind with a focus on the left estimate.

The obvious way to achieve (1.4.17) is to define

$$kr_\alpha(\theta) := \max(kr(\theta), \alpha) \quad \forall \theta \in [\theta_m, \theta_M] \tag{1.4.18}$$

with $kr(\cdot)$ in (1.2.10). The function $\theta(\cdot)$ in (1.3.20) can remain untouched (see, however, (1.4.23)). With this definition and $p_\alpha < -1$ given by

$$(-p_\alpha)^{-\lambda e(\lambda)} = \alpha \quad \Longleftrightarrow \quad p_\alpha = -\alpha^{-\frac{1}{\lambda e(\lambda)}} \tag{1.4.19}$$

the altered relative permeability function with respect to p reads

$$kr_\alpha(\theta(p)) = \begin{cases} \alpha & \text{for } p \leq p_\alpha \\ kr(\theta(p)) & \text{for } p \geq p_\alpha \end{cases}$$

with $kr(\theta(\cdot))$ given in (1.3.21). Clearly, if a function $p : \Omega \times (0,T) \to \mathbb{R}$ is bounded by $|p_\alpha|$, it is a solution of (1.3.17) if and only if it solves this equation with kr_α instead of kr. For this altered nondegenerate Richards equation, however, the altered Kirchhoff transformation $\kappa_\alpha : \mathbb{R} \to \mathbb{R}$ is surjective and the corresponding improper integral in (1.4.16) no longer exists.

More concretely, with κ given in (1.3.22) and

$$u_\alpha := \kappa(p_\alpha) = \frac{1}{\lambda e(\lambda) - 1}\,\alpha^{1-\frac{1}{\lambda e(\lambda)}} - \frac{\lambda e(\lambda)}{\lambda e(\lambda) - 1} \tag{1.4.20}$$

satisfying $u_c < u_\alpha < -1$ we obtain

$$\kappa_\alpha(p) = \begin{cases} \alpha(p - p_\alpha) + u_\alpha & \text{for } p \leq p_\alpha \\ \kappa(p) & \text{for } p \geq p_\alpha \end{cases}$$

such that $\kappa_\alpha(p) \to -\infty$ holds for $p \to -\infty$. Consequently, the altered inverse Kirchhoff transformation $\kappa_\alpha^{-1} : \mathbb{R} \to \mathbb{R}$ no longer has a singularity u_c and reads

$$\kappa_\alpha^{-1}(u) = \begin{cases} \alpha^{-1}(u - u_\alpha) + p_\alpha & \text{for } u \leq u_\alpha \\ \kappa^{-1}(u) & \text{for } u \geq u_\alpha \end{cases}$$

with κ^{-1} given in (1.3.24), i.e. κ_α is an affine function on $(-\infty, u_\alpha]$ with the (big) slope α^{-1}.

Now, in contrast to the function $p \mapsto \theta(p)$, the saturation of the generalized pressure $u \mapsto M_\alpha(u) = \theta(\kappa_\alpha^{-1}(u))$ differs from $u \mapsto M(u)$. One can think of the function M on $(u_c, u_\alpha]$ being "stretched" onto the interval $(-\infty, u_\alpha]$ in order to obtain M_α given by

$$M_\alpha(u) = \begin{cases} \theta_m + (\theta_M - \theta_m)(-\alpha^{-1}(u - u_\alpha) - p_\alpha)^{-\lambda} & \text{for } u \leq u_\alpha \\ M(u) & \text{for } u \geq u_\alpha \,. \end{cases} \tag{1.4.21}$$

M_α is monotonically increasing with $M_\alpha(u) \to \theta_m$ for $u \to -\infty$ and

$$\theta_s := M_\alpha(u_\alpha) = \theta(p_\alpha) = \theta_m + (\theta_M - \theta_m)\, \alpha^{\frac{1}{e(\lambda)}} \,, \tag{1.4.22}$$

i.e. with a (small) range $M((-\infty, u_\alpha)) = (\theta_m, \theta_s)$. We point out that one could also treat $\theta(\cdot)$ similarly as or instead of $kr(\cdot)$ in (1.4.18) and obtain the same results for the altered functions that we just discussed. Even though in doing so we alter the Richards equation (1.3.17) in the saturation term, too, let us define the "cut" saturation

$$\theta_\alpha(p) := \max(\theta(p), \theta_s) \quad \forall p \in \mathbb{R} \,, \tag{1.4.23}$$

which provides the same results as above, for further use. For completeness, we note that

$$kr_\alpha(M_\alpha(u)) = \begin{cases} \alpha & \text{for } u \leq u_\alpha \\ kr(M(u)) & \text{for } u \geq u_\alpha \,. \end{cases}$$

We remark here that, instead of cutting $kr(\cdot)$ as in (1.4.18), one could of course think of parameter functions for which $kr(\cdot) \geq 0$ does not satisfy $kr(\cdot) \geq \alpha$ for an $\alpha > 0$, but which nevertheless generate surjective Kirchhoff transformations $\kappa : \mathbb{R} \to \mathbb{R}$ leading to similar results as just discussed for the nondegenerate case. As an example one could choose $\lambda = 1$ in (1.2.9) but replace $e(\lambda)$ by 1 in (1.2.10). This would lead to a logarithmic expression for κ with respect to $p \leq -1$ and to exponential terms in κ^{-1}, M and $kr(M(\cdot))$ with respect to $u \leq -1$, i.e. to big slopes of these functions which is characteristic in hydrologically realistic situations.

1.4.4 Limit cases for the altered Brooks–Corey functions

Finally, we take a look at the limit cases for our altered functions with a fixed $\alpha \in (0, 1)$, using the results obtained above for the original parameter functions. In our considerations in Section 2 we come back to these limit cases, too.

First, for $\lambda \to 0$ we get

$$p_\alpha \to -\alpha^{-\frac{1}{2}} =: p_{\alpha,0} < -1 \quad \text{and} \quad u_\alpha \to -2 + \alpha^{\frac{1}{2}} =: u_{\alpha,0} > -2$$

from (1.4.19) and (1.4.20). According to (1.4.3) and (1.4.5) the pointwise limit of the relative permeability functions $kr_\alpha(\cdot)$ and $kr_\alpha(\theta(\cdot))$ is given by

$$kr_{\alpha,0}(\theta) = \max(kr_0(\theta), \alpha) \quad \forall \theta \in [\theta_m, \theta_M]$$

and

$$k_{\alpha,0}(p) = \max(k_0(p), \alpha) \quad \forall p \in \mathbb{R}\,,$$

respectively. The limit of the inverse Kirchhoff transformation κ_α^{-1} reads

$$\kappa_{\alpha,0}^{-1} : u \mapsto \begin{cases} \alpha^{-1}(u - u_{\alpha,0}) + p_{\alpha,0} & \text{for } u \le u_{\alpha,0} \\ \kappa_0^{-1}(u) & \text{for } u \ge u_{\alpha,0}\,. \end{cases}$$

Unfortunately, the limit of M_α is

$$M_{\alpha,0} : u \mapsto \theta_M \quad \forall u \in \mathbb{R} \tag{1.4.24}$$

which is clear for $u > u_{\alpha,0}$ due to the behaviour of M_0. On the other hand, by definition of M_α for $u \le u_{\alpha,0}$ we obtain

$$-\alpha^{-1}(u - u_\alpha) - p_\alpha \to -\alpha^{-1}(u - u_{\alpha,0}) - p_{\alpha,0} > 1$$

and consequently

$$(-\alpha^{-1}(u - u_\alpha) - p_\alpha)^{-\lambda} \to 1$$

for $u \le u_{\alpha,0}$ and $\lambda \to 0$. The limit (1.4.24) does not account for the unsaturated case and is therefore useless as a hydrological model. This does not change if we use (1.4.23) with a fixed $\theta_s \in (\theta_m, \theta_M)$ instead of (1.4.18) with a fixed α at the beginning. Even though we get an altered saturation

$$M_{\theta_s}(u) = \begin{cases} \theta_s & \text{for } u \le u_\alpha \\ M(u) & \text{for } u \ge u_\alpha \end{cases}$$

in this case which is independent of λ on $(-\infty, u_\alpha)$, we still have the dependency (1.4.22) of θ_s and α which is equivalent to

$$\log \alpha = \log\left(\frac{\theta_s - \theta_m}{\theta_M - \theta_m}\right) e(\lambda)\,.$$

But this relationship forces $\alpha \to 0$ (and $u_\alpha \to -2$) for $\lambda \to 0$ because of $e(\lambda) \to \infty$ for $\lambda \to 0$. This, however, spoils our initial intention (1.4.17) to consider a nondegenerate Richards equation in which the factor in front of the spatial derivatives is bounded from below by a positive constant. Instead we basically regain the limit case $\lambda \to 0$ for our original parameter functions (with θ_m replaced by θ_s in (1.4.9)).

The situation is more promising for $\lambda \to \infty$. Here we have

$$p_\alpha \to -1 \quad \text{and} \quad u_\alpha \to -1$$

due to (1.4.19) and (1.4.20). The relative permeabilities converge pointwise to

$$kr_{\alpha,\infty}(\theta) = \max(kr_\infty(\theta), \alpha) \quad \forall \theta \in [\theta_m, \theta_M]$$

and

$$k_{\alpha,\infty}(p) = \max(k_\infty(p), \alpha) \quad \forall p \in \mathbb{R}\,,$$

respectively, with the functions in (1.4.10) and (1.4.12). In the limit the inverse of the altered Kirchhoff transformation reads

$$\kappa_{\alpha,\infty}^{-1} : u \mapsto \begin{cases} -\alpha^{-1}(u+1) - 1 & \text{for } u \le -1 \\ u & \text{for } u \ge -1\,. \end{cases}$$

The saturation M_α tends pointwise to the step function

$$M_{\alpha,\infty}(u) = \begin{cases} \theta_m & \text{for } u < -1 \\ \theta_M & \text{for } u \ge -1 \end{cases} \tag{1.4.25}$$

with $kr_\alpha(M_\alpha(\cdot))$ converging to

$$kr_{\alpha,\infty}(M_{\alpha,\infty}(u)) = \begin{cases} \alpha & \text{for } u < -1 \\ 1 & \text{for } u \ge -1\,. \end{cases}$$

This limit case makes sense hydrologically because it models both the saturated case to which all physical pressures values $p \ge -1$ refer and the unsaturated case for all $p < -1$. In addition, the full range of possible values $p \in \mathbb{R}$ is still contained in the model and the inverse Kirchhoff transformation is not ill-conditioned since its slopes are bounded by α^{-1}. If we choose (1.4.23) with a fixed $\theta_s \in (\theta_m, \theta_M)$ instead of (1.4.18), the limit looks the same with θ_m in (1.4.11) and (1.4.25) replaced by θ_s.

For realistic situations as discussed at the beginning of this section, one would of course choose α small enough, such that θ_s is close to θ_m and the resulting M_α almost “looks like” a step function (see Figure 1.11), thus resembling the situation in this limit case. Finally, we mention that our analytical and numerical approach to the Richards equation (compare (2.4.7), Remark 2.7.5 and Section 2.8) allows an efficient and robust treatment of this limit case, too.

As far as variations of p_b are concerned for the altered parameter functions, we obviously obtain the same case as above for $p_b/p_0 \to -\infty$ which is hydrologically senseless. For $p_b/p_0 \uparrow 0$ we obtain $p_\alpha \uparrow 0$ as well as $u_\alpha \uparrow 0$. Therefore, $kr_\alpha(\theta(\cdot))$ turns into the step function

$$p \mapsto \begin{cases} \alpha & \text{for } p < 0 \\ 1 & \text{for } p \ge 0 \end{cases}$$

and (keeping in mind that both axes are scaled for the transformation) κ_α converges to a piecewise linear function on $\mathbb{R}$ with the inverse

$$u \mapsto \begin{cases} \alpha^{-1}u & \text{for } u < 0 \\ u & \text{for } u \geq 0\,. \end{cases}$$

The saturation M_α tends to

$$u \mapsto \begin{cases} \theta_m & \text{for } u < 0 \\ \theta_M & \text{for } u \geq 0 \end{cases} \tag{1.4.26}$$

and $kr_\alpha(M_\alpha(\cdot))$ goes to

$$u \mapsto \begin{cases} \alpha & \text{for } u < 0 \\ 1 & \text{for } u \geq 0 \end{cases}$$

with θ_m replaced by θ_s if we consider the "cut" saturation (1.4.23). We conclude that, just as for the original parameter functions, this limit case $p_b/p_0 \uparrow 0$ is basically the same as the one obtained above for $\lambda \to \infty$ with the altered parameter functions.

1.5 Boundary value problems for the Richards equation: strong and weak formulations

The purpose of this section is to give strong and weak formulations of a standard boundary value problem we would like to consider for the homogeneous Richards equation (1.3.17) and its Kirchhoff–transformed version (1.3.18). In Subsection 1.5.1 we focus on a differential form of a Signorini-type boundary value problem for the Richards equation with surface water which we obtain from hydrological considerations on a reservoir model. Then, on the basis of this problem, we derive an equivalent variational formulation for the generalized physical pressure in Subsection 1.5.2 giving rise to an interpretation in a weak sense which we introduce in Subsection 1.5.3. Finally, in Subsection 1.5.4 we derive a weak variational inequality of the Signorini-type problem in the physical pressure variable and investigate the connection between this variational inequality and the one obtained in the generalized variables. To this end, we shall discuss intensively the Kirchhoff transformation as a superposition operator on different Sobolev spaces, which will be needed for the analysis of the heterogeneous Richards equation in Chapter 3, too.

For simplicity and without loss of generality we set $t_0 = 1\,s$, $A_s = m^{-1}$, $n \equiv 1$, $p_r = u_r = 1$ and $K_h \equiv 1$ in the equations (1.3.17) and (1.3.18). The $\hat{}$ in these equations has only been introduced to carry out the scaling in Section 1.3 and will be skipped from now on. So in the following we mostly deal with the Richards equation in the form

$$\theta(p)_t - \operatorname{div}\Big(kr(\theta(p))\nabla\,(p - z)\Big) = 0 \tag{1.5.1}$$

and its transformed version which reads

$$M(u)_t - \operatorname{div}\Big(\nabla u - kr(M(u))e_z\Big) = 0\,. \tag{1.5.2}$$

We will indicate how to deal with space dependent $n(\cdot)$ and $K_h(\cdot)$ where it is necessary and appropriate. In general, these functions are at least required to be positive and bounded and $K_h(\cdot) \geq c$ should be satisfied for a $c > 0$.

1.5.1 A problem of Signorini's type for the Richards equation with surface water

Let $\Omega \subset \mathbb{R}^d$ (in our concrete cases $d \in \{1, 2, 3\}$) be an open, bounded, connected and nonempty set with a Lipschitz boundary $\partial\Omega$ (see Definition A.2.1). This condition guarantees that the normal $\mathbf{n}$, which we assume to be directed outwards, exists almost everywhere on $\partial\Omega$ (cf. Ciarlet [28, pp. 32–37]). In practical cases we mostly consider Ω to be a polyhedron or at least having a boundary which is piecewise C^1 (see Definition A.1.1). Figure 1.15 shows an example of such a domain $\Omega \subset \mathbb{R}^2$ which could be regarded as a vertical cut through the ground in three space dimensions with γ_u representing the surface of the ground.

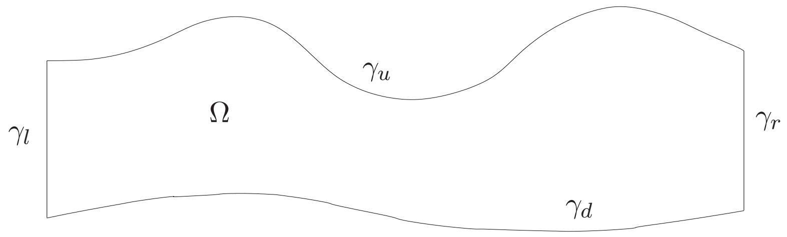


Figure 1.15: 2D-domain Ω (vertical cut through the ground)

For a given time $t \in [0, T]$ we assume $\partial\Omega$ to be decomposed into finitely many non-overlapping connected subsets each of which is contained in exactly one of the three subsets $\gamma_D(t)$, $\gamma_N(t)$ and $\gamma_S(t)$ of $\partial\Omega$. With given functions $u_D(t)$ on $\gamma_D(t)$ and $f_N(t)$ on $\gamma_N(t)$ we assume that the unknown function u and the unknown flux

$$\mathbf{v} = -(\nabla u - kr(M(u))e_z) = -kr(\theta(p))\nabla\,(p - z) \tag{1.5.3}$$

(compare (1.5.1), (1.5.2) with (1.2.2)) satisfy the following boundary conditions:

a) Dirichlet boundary conditions:

$$u = u_D(t) \quad \text{on} \quad \gamma_D(t)$$

b) Neumann boundary conditions:

$$\mathbf{v} \cdot \mathbf{n} = f_N(t) \quad \text{on} \quad \gamma_N(t)$$

c) Signorini-type boundary conditions:

$$u \leq 0\,, \quad \mathbf{v} \cdot \mathbf{n} \geq 0\,, \quad u \cdot (\mathbf{v} \cdot \mathbf{n}) = 0 \quad \text{on} \quad \gamma_S(t)$$

Dirichlet and Neumann boundary conditions are well known for all kinds of boundary value problems. Observe that in our case the Dirichlet boundary conditions could be equivalently expressed in terms of the physical water pressure $p = p_D(t) := \kappa^{-1}(u_D(t))$ using the Kirchhoff transformation κ (see (1.3.1)). The Neumann boundary conditions refer to the physical water flux $\mathbf{v}$, whether expressed by the physical variable p or the generalized pressure u. With regard to the Signorini-type boundary conditions observe that due to the definition of the Kirchhoff transformation (1.3.1) and $kr(\theta(p)) = 1$ for $p \geq 0$ we have $p = u$ for $p \geq 0$ or $u \geq 0$, respectively, and $p < 0 \Leftrightarrow u < 0$. Therefore, c) can be equivalently formulated if we replace u by p.

Dirichlet boundary conditions usually appear as hydrostatic pressures given by surface water (e.g. lakes or rivers) on γ_u or water (e.g. rivers or the sea) on a side γ_l or γ_r of $\partial\Omega$. Neumann boundary conditions specify water flow into or out of Ω due to rain or water movements around $\partial\Omega$ and often occur as homogeneous Neumann boundary conditions on dry parts of γ_u or on the border of an impermeable soil, e.g. on γ_d.

Apart from the Dirichlet and Neumann boundary conditions which are called boundary conditions of the first and second kind, respectively, one also encounters boundary conditions of the third kind (on semipermeable boundaries, see Bear [13, p. 265]), known as *Robin boundary conditions* (consult Gustafson and Abe [46] and [47]). They are conditions on linear combinations of Dirichlet and Neumann boundary values and will play an crucial role for our treatment of the heterogeneous Richards equation in Section 3.4. For simplicity we do not treat them in the first two chapters of this work. Further (nonlinear) generalizations of Robin boundary conditions are known as leaky boundary conditions (see e.g. Carrillo and Chipot [23] and Chipot and Lyaghfouri [26]), which we do not consider here.

Signorini boundary conditions are well known for contact problems in mechanics (see Signorini [86] and Krause [61]). Moreover, “Signorini’s problem” *is* the name for a problem in linear elasticity. Nevertheless, *Signorini-type boundary conditions* or boundary conditions *of Signorini’s type* occur in various fields, for instance in electrochemistry (cf. Gerbi et al. [42]), in connection with Stefan problems (cf. Calvo et al. [22]) and in hydrology (cf. Bagagiolo and Visintin [8] and Zheng et al. [103]). Therefore, we also attribute Signorini’s name to the boundary γ_S and the complementary conditions given above in c).

In hydrological settings, boundary conditions of Signorini’s type usually appear around surface water reservoirs on γ_u or in case of a so-called dam problem (see e.g. Alt [2]) above the part of the boundary where the surface water or the sea, respectively, is in contact with $\partial\Omega$ (in the latter case e.g. on a part of γ_l or γ_r). Such a situation is depicted in Figure 1.16 which shows a zoomed part of γ_u where a water reservoir occurs above γ_3. Therefore, we consider γ_3 to be

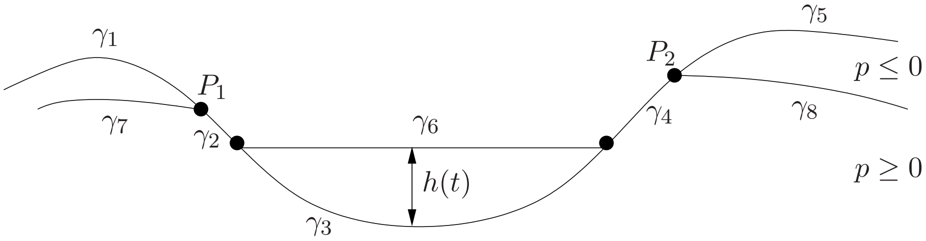


Figure 1.16: Boundary of Signorini's type around a water reservoir

a part of the Dirichlet boundary $\gamma_D(t)$ with the boundary values given as the hydrostatic pressure imposed by the surface water.

Although the surface of the reservoir is γ_6, the domain Ω can be fully saturated up to a curve lying above γ_6, here up to the curve $\gamma_7 \cup \gamma_2 \cup \gamma_4 \cup \gamma_8$, on which the pressure is vanishing. The part $\gamma_2 \cup \gamma_4$ of $\partial\Omega$ is called the *seepage face.* Here water can flow from the interior of Ω across $\partial\Omega$ tickling into the reservoir, thus $\mathbf{v} \cdot \mathbf{n} \geq 0$, but the pressure of the water is $p = u = 0$. On $\gamma_1 \cup \gamma_5$, however, we have no flow $\mathbf{v} \cdot \mathbf{n} = 0$, but then the water pressure cannot be positive, i.e. $p \leq 0 \Leftrightarrow u \leq 0$. This is the hydrological reason for the complementary conditions for u and $\mathbf{v} \cdot \mathbf{n}$ on the Signorini-type boundary $\gamma_S(t)$ given in c) that result in a decomposition of $\gamma_S(t)$ in seepage faces like $\gamma_2 \cup \gamma_4$ and adjacent parts of $\partial\Omega$ like $\gamma_1 \cup \gamma_5$. The points P_1 and P_2, which determine the boundary of the seepage face within the Signorini-type boundary $\gamma_S(t)$, are usually unknown a priori and arise as part of the solution, which satisfies the conditions given in c). Just as in mechanics, the boundary value problems with conditions of Signorini's type, which we encounter in the following, can therefore be regarded as free boundary problems. Finally, note the analogy of our conditions in c) and Signorini boundary conditions known from mechanics. For example, $u \leq 0$ on γ_S corresponds to the non-penetration condition in contact problems.

There is a hydrological necessity for the existence of a nontrivial seepage face if the so-called *phreatic surface*, which is given by $p = 0$, i.e. $\gamma_7 \cup \gamma_8$, is above the water table γ_6 around the reservoir. Otherwise, if P_1 and P_2 were the endpoints of γ_3, the water flow $\mathbf{v}$ in these points would have to be parallel to the phreatic surface on the one hand, but also perpendicular to γ_3 on the other hand, which is only possible if the phreatic surface is beneath γ_6 around the water reservoir. See Bear [13, pp. 260/261] for a detailed discussion of the seepage face.

Furthermore, we note that the area beneath $\gamma_7 \cup \gamma_2 \cup \gamma_3 \cup \gamma_4 \cup \gamma_8$ in Figure 1.16, where $p \geq 0$ holds, does not necessarily coincide with the region in which the ground is fully saturated. In general there is a so-called *capillary fringe* above $\gamma_7 \cup \gamma_8$ in which full saturation still occurs although $p < 0$ holds. This is due to the so-called bubbling pressure $p_b < 0$ that is discussed in Section 1.2 and that is also reflected by our special choice of Brooks–Corey parameter functions (1.2.9) and (1.2.10). Consequently, the *groundwater table*, i.e. the border between the saturated and the unsaturated region in the interior of Ω, is given by $p = p_b$ and generally lies above the phreatic surface.

Now, for any $t \in (0,T]$, $T > 0$, we consider the boundary value problem

$$\begin{aligned} M(u)_t - \operatorname{div}\Big(\nabla u - kr(M(u))e_z\Big) &= 0 \qquad \text{on } \Omega && (1.5.4)\\ u &= u_D(t) \quad \text{on } \gamma_D(t) && (1.5.5)\\ \mathbf{v}\cdot\mathbf{n} &= f_N(t) \quad \text{on } \gamma_N(t) && (1.5.6)\\ u \le 0, \quad \mathbf{v}\cdot\mathbf{n} \ge 0, \quad u\cdot(\mathbf{v}\cdot\mathbf{n}) &= 0 \qquad \text{on } \gamma_S(t) && (1.5.7) \end{aligned}$$

which we call *Signorini-type problem* or *problem of Signorini's type* for the Kirchhoff–transformed Richards equation. Of course, it can be easily reformulated for the Richards equation with the physical pressure p as the unknown. Due to ellipticity and monotonicity, however, most of the analysis is only carried out for the Kirchhoff–transformed version (compare Subsection 1.6), and so is our numerical treatment of the Richards equation in Chapter 2. This is why we choose the formulation above. An additional requirement needed in (1.5.4)–(1.5.7) is certainly $u_D > u_c$ and $u > u_c$ if $M : (u_c, \infty) \to \mathbb{R}$ with a $u_c < 0$, which is the case for the Brooks–Corey parameter functions. Finally, we point out that if subsets of $\gamma_D(t)$ and $\gamma_S(t)$ are adjacent, $u_D(t)$ needs to be compatible with the condition $u \le 0$ on γ_S if functions u from certain solution spaces are to satisfy both boundary conditions.

Remark 1.5.1. Before we derive a weak formulation for this free boundary problem we shortly consider a straightforward model for a dynamic coupling of ground and surface water. In the situation depicted in Figure 1.16 the Dirichlet data on γ_3 is given by the hydrostatic pressure $p_D(t) = u_D(t) = \varrho g h(t)$ (possibly modulo scaling factors, e.g. $|p_b|^{-1}$ with the bubbling pressure p_b as done in Section 1.3). The surface water level $h(t)$ over the ground is a function on the subset γ_3 of $\gamma_D(t)$ which is in general time-dependent, too, i.e. we have $\gamma_3(t)$. In fact, by considering the geometry of $\partial\Omega$, knowing $\gamma_3(t)$ is equivalent to knowing $h(t)$ if the water in the reservoir is assumed to remain static. In this case we speak of a *reservoir model* for the surface water. If the surface water is a big lake or the sea, $h(t)$ as a function of t can be regarded as practically not influenced by the groundwater in Ω, i.e. as a given boundary condition on $\gamma_3(t)$ for the Richards equation. This is assumed in (1.5.4)–(1.5.7).

If the reservoir is small enough such that the flow of groundwater from Ω into the reservoir or back cannot be neglected, $h(t)$ or the Dirichlet boundary $\gamma_3(t)$ are not known a priori. Then we need to consider a coupling of the surface water behaviour and the groundwater modelling which is given by the Richards equation. The easiest way to do this for the reservoir model is to consider the increase or decrease of volume $\frac{d}{dt}V(t)$ of the surface water given by the water flow $\mathbf{v}\cdot\mathbf{n}$ across $\gamma(t) := \gamma_2(t) \cup \gamma_3(t) \cup \gamma_4(t)$ (see Figure 1.16) and assuming mass conservation

$$\frac{d}{dt}m(t) = \varrho\frac{d}{dt}V(t) = \varrho\int_{\gamma(t)} \mathbf{v}(x,t)\cdot\mathbf{n}\,d\sigma(x) \qquad (1.5.8)$$

for the total mass $m(t)$ of the water in the reservoir. By the geometry of $\partial\Omega$, knowing $V(t)$ is both equivalent to knowing $h(t)$ and equivalent to knowing $\gamma_3(t)$. Consequently, considering these geometric relationships, equations (1.5.4)–(1.5.8) form a simple model for a coupling of ground and surface water.

1.5.2 A variational inequality in a classical sense

As a special type of boundary value problem the Signorini-type problem (1.5.4)–(1.5.7) has a weak formulation in terms of a variational inequality. This follows from an application of a generalization of Green's formula or partial integration, which we note below. See Theorem A.1.3 in the appendix for a derivation of it from Gauss's theorem. We also refer to Definition A.1.1 in the appendix for an exact definition of a C^1-polyhedron $\Omega \subset \mathbb{R}^d$ which is sometimes called a domain with a smooth boundary $\partial\Omega$ except for a $(d-1)$-nullset of singularities. Furthermore, see [55, pp. 362–369] for measurability of hyperfaces and Definition A.1.4 to recall the well-known spaces $C^k(\overline{\Omega})$ for $k \in \mathbb{N}_0$.

Theorem 1.5.2. *Let $\Omega \subset \mathbb{R}^d$ be a bounded C^1-polyhedron with Hausdorff measurable $\partial\Omega$, $G \in (C^1(\overline{\Omega}))^d$ and $v \in C^1(\overline{\Omega})$. Then the following identity holds:*

$$\int_\Omega \operatorname{div}(G(x))\, v(x)\, dx = -\int_\Omega G(x)\nabla v(x)\, dx + \int_{\partial\Omega} (G(x)\cdot \mathbf{n}(x))\, v(x)\, d\sigma(x)\,. \tag{1.5.9}$$

Furthermore, for

$$u_D(t) \in \{v = w_{|\gamma_D(t)} : w \in C^2(\overline{\Omega}) \,\wedge\, w_{|\gamma_S(t)} \le 0\} \tag{1.5.10}$$

we define the convex set $\mathcal{K}_c(t) \subset C^2(\overline{\Omega})$ as

$$\mathcal{K}_c(t) := \{w \in C^2(\overline{\Omega}) : w_{|\gamma_D(t)} = u_D(t) \,\wedge\, w_{|\gamma_S(t)} \le 0\}\,. \tag{1.5.11}$$

It is clear that $\mathcal{K}_c(t)$ is nonempty since $u_D(t)$ is chosen to be compatible with the Signorini-type boundary condition $u_{|\gamma_S(t)} \le 0$.

Note that due to the definition (1.3.1) of the Kirchhoff transformation, the function $\kappa : \mathbb{R} \to \mathbb{R}$ is continuously differentiable if kr is continuous and, therefore, $M = \theta \circ \kappa^{-1}$ is continuously differentiable if θ is. In what is to come we apply the well-known definitions of (differentiable) manifolds in [55, pp. 115/116] and Hausdorff measures on C^1-polyhedra in [3, p. 13], see also Definition A.1.1 in the appendix. Now we can prove the following equivalence.

Proposition 1.5.3. *Let $M, kr : \mathbb{R} \to \mathbb{R}$ be continuously differentiable real functions, and for $t \in (0,T]$ let $u(t) \in C^2(\overline{\Omega})$ on a C^1-polyhedron $\Omega \subset \mathbb{R}^d$ with Hausdorff measurable $\partial\Omega$. Furthermore, let $f_N(t)$ be continuous on $\gamma_N(t)$, and let $\gamma_N(t)$ and $\gamma_S(t)$ be piecewise $(d-1)$-dimensional manifolds. Then u satisfies the boundary value problem (1.5.4)–(1.5.7) for t if and only if it satisfies the* *variational inequality*

$$\int_\Omega M(u)_t\,(v-u)\,dx + \int_\Omega \nabla u \nabla (v-u)\,dx \geq \\ \int_\Omega kr(M(u))e_z \nabla(v-u)\,dx - \int_{\gamma_N(t)} f_N(t)\,(v-u)\,d\sigma \quad \forall v \in \mathcal{K}_c(t) \qquad (1.5.12)$$

in the convex set $\mathcal{K}_c(t)$. In addition, (1.5.12) becomes a variational equality if there is no boundary of Signorini's type, i.e. if $\gamma_S(t) = \emptyset$. In this case the set of test functions $v-u$ is the subspace $\{w \in C^2(\overline{\Omega}) : w_{|\gamma_D(t)} = 0\}$ of $C^2(\overline{\Omega})$.

Proof. First, let u satisfy the boundary value problem (1.5.4)–(1.5.7) for a $t \in (0,T]$. Then, of course, we have $u \in \mathcal{K}_c(t)$ due to (1.5.7). With a $v \in \mathcal{K}_c(t)$ we multiply (1.5.4) by $v-u$ on both sides and integrate over Ω obtaining

$$\int_\Omega M(u)_t\,(v-u)\,dx + \int_\Omega -\operatorname{div}(\nabla u - kr(M(u))e_z)\,(v-u)\,dx = 0\,.$$

Due to the conditions imposed on the involved functions we have $v-u \in C^1(\overline{\Omega})$ and $\mathbf{v} = -(\nabla u - kr(M(u))e_z) \in (C^1(\overline{\Omega}))^d$. So applying Theorem 1.5.2 for $G = \mathbf{v}$ and $v-u$ instead of v we arrive at

$$\int_\Omega M(u)_t\,(v-u)\,dx + \int_\Omega \nabla u \nabla (v-u)\,dx \\ - \int_\Omega kr(M(u))e_z \nabla(v-u)\,dx + \int_{\partial\Omega} \mathbf{v}\cdot\mathbf{n}\,(v-u)\,d\sigma = 0\,. \qquad (1.5.13)$$

Note that $\partial\Omega$ is a disjoint union of the sets $\gamma_D(t)$, $\gamma_N(t)$ and $\gamma_S(t)$. Therefore, taking $v-u=0$ on $\gamma_D(t)$ and $\mathbf{v}\cdot\mathbf{n} = f_N(t)$ on $\gamma_N(t)$ into account, we only need to prove

$$\int_{\gamma_S(t)} \mathbf{v}\cdot\mathbf{n}\,(v-u)\,d\sigma \leq 0$$

in order to obtain (1.5.12). But this follows from $(\mathbf{v}\cdot\mathbf{n})\cdot u = 0$ and $\mathbf{v}\cdot\mathbf{n} \geq 0$ and $v_{|\gamma_S(t)} \leq 0$ on $\gamma_S(t)$.

Conversely, let $u \in \mathcal{K}_c(t)$ satisfy (1.5.12) for a $t \in (0,T]$. An application of Theorem 1.5.2 as above for $G = \mathbf{v}$ and $v-u$ instead of v to (1.5.12) provides

$$\int_\Omega \Big(M(u)_t\,(v-u) - \operatorname{div}(\nabla u - kr(M(u))e_z)\Big)(v-u)\,dx \\ + \int_{\gamma_N(t)} f_N(t)\,(v-u)\,d\sigma - \int_{\partial\Omega} \mathbf{v}\cdot\mathbf{n}\,(v-u)\,d\sigma \geq 0 \quad \forall v \in \mathcal{K}_c(t)\,. \qquad (1.5.14)$$

We first prove that u satisfies (1.5.4) for t by assuming that this is not the case, i.e. there is a $x \in \Omega$ such that w.l.o.g. we have

$$M(u(x,t))_t - \operatorname{div}(\nabla u(x,t) - kr(M(u(x,t)))e_z) > 0\,,$$

which by continuity of the function on the left hand side we assume to be true on a ball $B_\varepsilon(x) \subset \Omega$ of radius $\varepsilon > 0$ around x. Now we choose a $v \in \mathcal{K}_c(t)$

such that $v = u$ on $\overline{\Omega}\backslash B_\varepsilon(x)$ and $v - u < 0$ on $B_\varepsilon(x)$. But with this choice of v the integrals over $\gamma_N(t)$ and $\partial\Omega$ in (1.5.14) vanish while the integral over Ω is negative due to the continuity of the integrands. This is a contradiction to (1.5.14), so u satisfies (1.5.4).

The construction of a smooth function v as used above is well known and relies on the existence of a nonnegative $\varphi \in C^\infty(\mathbb{R}^d)$ with $\operatorname{supp}\varphi = \overline{B_\varepsilon(x)}$. For the basic ingredient of such a construction consult e.g. [56, p. 277]. In the following we will use this idea to get test functions with certain properties on parts of the boundary $\partial\Omega$.

Since u satisfies (1.5.4) and $u = v$ on $\gamma_D(t)$, the variational inequality (1.5.14) reduces to

$$\int_{\gamma_N(t)} (f_N(t) - \mathbf{v}\cdot\mathbf{n})(v-u)\, d\sigma - \int_{\gamma_S(t)} \mathbf{v}\cdot\mathbf{n}\,(v-u)\, d\sigma \geq 0 \quad \forall v \in \mathcal{K}_c(t)\,. \quad (1.5.15)$$

Now, since $\gamma_N(t)$ is a piecewise $(n-1)$-dimensional manifold, we can conclude $f_N(t) = \mathbf{v}\cdot\mathbf{n}$ on $\gamma_N(t)$ as required. Otherwise, using the continuity of the involved functions, we find a subset $\tilde{\gamma} \subset \gamma_N(t)$ with a positive Hausdorff measure and a suitable $v \in \mathcal{K}_c(t)$ such that we have $(f_N(t) - \mathbf{v}\cdot\mathbf{n})(v-u) < 0$ on $\tilde{\gamma}$ while $v = u$ holds on $\gamma_S(t) \cup (\gamma_N(t)\backslash\tilde{\gamma})$, leading to a contradiction to (1.5.15). Using this result, (1.5.15) reduces to

$$\int_{\gamma_S} \mathbf{v}\cdot\mathbf{n}\,(v-u)\, d\sigma \leq 0 \quad \forall v \in K_c(t)\,, \quad (1.5.16)$$

from which we prove (1.5.7) by applying the same technique: Let $u(x,t) < 0$ for a point x in a subset $\tilde{\gamma}_\varepsilon = B_\varepsilon(x) \cap \gamma_S(t)$ with a positive Hausdorff measure. We assume $\mathbf{v}(x,t)\cdot\mathbf{n} \neq 0$ and ε to be small enough such that u and $\mathbf{v}\cdot\mathbf{n}$ do not change their sign on $\tilde{\gamma}_\varepsilon$. Then, we first construct an admissible test function $v \in \mathcal{K}_c(t)$ such that $u < v \leq 0$ on $\tilde{\gamma}_\varepsilon$, i.e. $v - u > 0$ on $\tilde{\gamma}_\varepsilon$, and $v = u$ elsewhere on $\gamma_S(t)$. (1.5.16) now gives $\mathbf{v}(x,t)\cdot\mathbf{n} < 0$. On the other hand, if we choose $v \in \mathcal{K}_c(t)$ with $v < u$ on $\tilde{\gamma}_\varepsilon$, i.e. $v - u < 0$ on $\tilde{\gamma}_\varepsilon$, and $v = u$ elsewhere on $\gamma_S(t)$, we obtain $\mathbf{v}(x,t)\cdot\mathbf{n} > 0$ from (1.5.16), which is a contradiction. So altogether, we conclude $\mathbf{v}\cdot\mathbf{n} = 0$ if $u < 0$. Since $u \in \mathcal{K}_c(t)$ we get $u\cdot(\mathbf{v}\cdot\mathbf{n}) = 0$ on $\gamma_S(t)$. Finally, if $\mathbf{v}\cdot\mathbf{n} < 0$ on a $\tilde{\gamma}_\varepsilon \subset \gamma_S(t)$ as above, we choose a $v < 0$ on $\tilde{\gamma}_\varepsilon$ with $v = 0$ on $\gamma_S(t)\backslash\tilde{\gamma}_\varepsilon$ which leads to a positive value of the integral in (1.5.16) and, thus, to a contradiction. Therefore, we also have $\mathbf{v}\cdot\mathbf{n} \geq 0$ on $\gamma_S(t)$.

Now if $\gamma_S(t) = \emptyset$, it is clear that the set of test functions $v - u$ in (1.5.12) is the linear space $\{w \in C^2(\overline{\Omega}) : w_{|\gamma_D(t)} = 0\}$. Therefore, we can also test with $u - v$ instead of $v - u$ in (1.5.12) and conclude that (1.5.12) is indeed a variational equality in this case. □

Remark 1.5.4. It is easy to see that a reformulation of (1.5.4)–(1.5.7) as a Signorini-type problem for the Richards equation with the physical pressure $p(t)$ (for a $t \in (0,T]$) is equivalent to an analogous reformulation of the variational inequality (1.5.12) in the corresponding convex set if $\theta, kr : \mathbb{R} \to \mathbb{R}$ are continuously differentiable.

Observe that, due to our special choice of parameter functions according to Brooks–Corey, the saturation M as a function of the generalized pressure u is not defined on the whole real line but only on (u_c, ∞) with a $u_c < 0$. In cases like these the range of u needs to be a subset of (u_c, ∞) in order to satisfy (1.5.4)–(1.5.7). Of course, the Dirichlet boundary condition $u_D(t)$ needs to be chosen in that way in the first place. Given this, Proposition 1.5.3 also applies if $M : (u_c, \infty) \to \mathbb{R}$ is continuously differentiable.

Alternatively, one can choose the convex set

$$\tilde{\mathcal{K}}_c(t) := \{v \in \mathcal{K}_c(t) : v(x) > u_c \ \ \forall x \in \Omega\}\,. \tag{1.5.17}$$

Then, the function $u \in \tilde{\mathcal{K}}_c(t)$ satisfies (1.5.4)–(1.5.7) if and only if it satisfies (1.5.12) in $\tilde{\mathcal{K}}_c(t)$ instead of in $\mathcal{K}_c(t)$. Provided that $\gamma_S(t) = \emptyset$, the assertion about the variational equality is also true in this case because of the strict inequality in (1.5.17) since with the compactness of $\overline{\Omega}$ and the norm in $C^2(\overline{\Omega})$, one can see that the set of test functions constitutes a neighbourhood of 0 in the subspace given in Proposition 1.5.3.

We adopt this approach from now on because of our special interest in the Brooks–Corey model and its analytical treatment. However, for a convex minimization result that we want to apply in Section 2.3, we will need the somewhat unphysical condition $u \geq u_c$ instead of $u > u_c$ which means that $p(x) = -\infty$ is assumed to be possible for the physical pressure $p(\cdot)$. The formulas (1.3.5) and (1.3.6) provide this straightforward extension of M in (1.3.25) by an improper integral, see also (1.4.8). This is needed to obtain a closed convex subset in the space considered (compare with the results in Section 1.6). Otherwise, unique solvability of the variational inequality cannot be guaranteed. Then, however, it has to be emphasized that the assertion in Proposition 1.5.3 about the variational equality is false in general because the test functions do no longer form a neighbourhood of 0 in the subspace considered there. So as soon as an equality in $u(x) = u_c$ holds (or can hold) for an $x \in \Omega$, we need to deal with the variational inequality instead of a variational equality. We will come back to this topic in Section 2.3 (see Remarks 2.3.12 and 2.3.17).

1.5.3 A weak variational inequality for the Kirchhoff–transformed Richards equation

Observe that in case of the Brooks–Corey parameter functions M is not even differentiable but only piecewise differentiable. Of course, one could consider a smooth approximation of M since non-differentiability of M for the (normalized) bubbling pressure -1 is hydrologically not essential. However, it seems to be in order at this point to generalize the notion of a solution to the Signorini-type problem (1.5.4)–(1.5.7) by extending the corresponding weak formulation (1.5.12) to a variational inequality in a closed convex subset of a Sobolev space. In the most simple case of $\gamma_D(t) = \partial\Omega$ and $u_D(t) = 0$ (with $u_c = -\infty$), this would be a variational equality for $u \in H_0^1(\Omega)$. In our general case, we need some more ingredients from the theory of Sobolev spaces which we have noted in Appendix A.2 and which we use in the following.

With regard to a weak formulation of the variational inequality (1.5.12) for models of Brooks–Corey type, we require the functions $M : [u_c, \infty) \to \mathbb{R}$ and $kr : M([u_c, \infty)) \to \mathbb{R}$ to be continuous, monotonically increasing and bounded with a $u_c < 0$. Furthermore, let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain and $\gamma_D(t)$, $\gamma_N(t)$ and $\gamma_S(t)$ be pairwise disjoint Hausdorff measurable submanifolds of $\partial\Omega$ for all $t \in [0, T]$ with $\partial\Omega = \gamma_D(t) \cup \gamma_N(t) \cup \gamma_S(t)$.

Concerning the weak Dirichlet boundary condition for $t \in [0, T]$, we choose

$$u_D(t) \in \{v = tr_{\gamma_D(t)} w : w \in H^1(\Omega) \wedge w \geq u_c \text{ a.e.} \wedge tr_{\gamma_S(t)} w \leq 0 \text{ a.e. on } \gamma_S(t)\} \tag{1.5.18}$$

as an element of $H^{1/2}(\gamma_D(t))$ analogously to (1.5.10) with the trace operators $tr_\Sigma : H^1(\Omega) \to H^{1/2}(\Sigma)$ for $\Sigma \in \{\gamma_D(t), \gamma_S(t)\}$. As in Remark 1.5.4 we require the range of $u_D(t)$ to be contained in $[u_c, \infty)$, now almost everywhere on $\gamma_D(t)$ and even for an extension of $u_D(t)$ in $H^1(\Omega)$ almost everywhere on Ω (for the reason see the proof of Proposition 1.5.5). In order to model realistic physical situations one would choose $u_D(t)$ with a range even contained in the open interval (u_c, ∞). The set in (1.5.18) is nonempty since the function $u_D(t) = 0$ with its trivial extension on Ω is contained in it. This is obvious but necessary to make sense of the following.

The convex set $\mathcal{K}(t) \subset H^1(\Omega)$ as a weak counterpart of $\tilde{\mathcal{K}}_c(t)$ given in (1.5.17) is now defined as

$$\mathcal{K}(t) := \{v \in H^1(\Omega) : v \geq u_c \;\wedge\; tr_{\gamma_D(t)} v = u_D(t) \;\wedge\; tr_{\gamma_S(t)} v \leq 0 \text{ a.e. on } \gamma_S(t)\}, \tag{1.5.19}$$

in which $v \geq u_c$ is again to be understood as $v(x) \geq u_c$ almost everywhere on Ω. Observe that this latter condition ensures that $\mathcal{K}(t)$ is a closed subset of $H^1(\Omega)$ which is not the case of $\tilde{\mathcal{K}}_c(t)$ in $C^2(\overline{\Omega})$. Since we will need the properties of $\mathcal{K}(t)$ in Section 2.3, we note them here.

Proposition 1.5.5. *$\mathcal{K}(t)$ is a nonempty, closed and convex subset of $H^1(\Omega)$.*

Proof. Using the linearity of the trace operators $tr_{\gamma_D(t)}$ and $tr_{\gamma_S(t)}$, it is easy to see that $\mathcal{K}(t)$ is a convex set. It is nonempty since the Dirichlet condition $u_D(t)$ in (1.5.18) is chosen to be compatible with the Signorini-type boundary condition $tr_{\gamma_S(t)} u \leq 0$ and there is an extension w of $u_D(t)$ with $w \geq u_c$ almost everywhere on Ω, i.e. $w \in \mathcal{K}(t)$. (If we only require $u_D(t) \geq u_c$ almost everywhere on $\gamma_D(t)$, we might not be able to guarantee the corresponding property for an extension of $u_D(t)$ in $H^1(\Omega)$. For the converse see the next proposition.)

Regarding the closedness, observe that for any $w \in H^1(\Omega)$ with $w(x) < u_c$ almost everywhere on a subset $\Omega' \subset \Omega$ with a positive Lebesgue measure, we also obtain a subset $\Omega'' \subset \Omega'$ with a positive Lebesgue measure such that $w(x) < u_c - \varepsilon$ for an $\varepsilon > 0$. This follows from

$$\bigcup_{n \in \mathbb{N}} \{x \in \Omega' : w(x) < u_c - 1/n \text{ a.e.}\} = \{x \in \Omega' : w(x) < u_c \text{ a.e.}\}$$

and the σ-additivity of the Lebesgue measure. As a consequence, the norm $\|w - v\|_{L^2(\Omega)}$ and therefore the $H^1(\Omega)$-norm $\|w - v\|_1$ is bounded from below by a positive constant, uniformly for all $v \in \mathcal{K}(t)$.

With the same arguments and with suitable constants $c, C > 0$ for a $w \in H^1(\Omega)$ for which $tr_{\gamma_D(t)} w = u_D(t)$ or $tr_{\gamma_S(t)} \leq 0$ is false, we obtain

$$0 < c \leq \|tr_\Sigma w - tr_\Sigma v\|_{L^2(\Sigma)} \leq \|tr_\Sigma w - tr_\Sigma v\|_{H^{1/2}(\Sigma)} \leq C\|w - v\|_1 \quad \forall v \in \mathcal{K}(t)$$

for $\Sigma \in \{\gamma_D(t), \gamma_S(t)\}$ by definition of $H^{1/2}(\Sigma)$ (see pages 248/249 in the appendix) and due to the trace theorem A.2.3.

Altogether, $H^1(\Omega)\backslash\mathcal{K}(t)$ is open and therefore, $\mathcal{K}(t) \subset H^1(\Omega)$ is closed. □

We find the following proposition instructive since it guarantees that nothing unnatural can happen to the elements in the set $\mathcal{K}(t)$. Nevertheless, one might be surprised that its proof is not straightforward.

Proposition 1.5.6. *Any $v \in \mathcal{K}(t)$ satisfies $tr_{\partial\Omega}\, v \geq u_c$ almost everywhere in the Hausdorff measure on $\partial\Omega$.*

Proof. In the first step we use the fact that

$$C^\infty(\overline{\Omega}) \cap \mathcal{K}_0 \text{ is dense in } \mathcal{K}_0 := \{v \in H^1(\Omega) : v \geq 0 \text{ a.e. on } \Omega\}$$

in the $H^1(\Omega)$-topology. This is proved in Glowinski [45, p. 61]. Therefore, if $v \in \mathcal{K}(t)$ we have $v_0 := v - u_c \in \mathcal{K}_0$ and the existence of a sequence $(v_n)_{n\in\mathbb{N}}$ in $C^\infty(\overline{\Omega}) \cap \mathcal{K}_0$ with $v_n \to v_0$ in $H^1(\Omega)$ for $n \to \infty$. Obviously, $v_n(x) \geq 0$ holds for all $x \in \Omega$ and all $n \in \mathbb{N}$.

In the second step we note that the embedding theorem A.2.2 and the trace theorem A.2.3 provide

$$v_{n|\partial\Omega} \to tr_{\partial\Omega}\, v_0 \quad \text{in } L^2(\partial\Omega) \text{ for } n \to \infty\,. \tag{1.5.20}$$

Due to a result from measure theory, see e.g. [82, p. 74, ex. 18], (1.5.20) entails the convergence $v_{n_k|\partial\Omega} \to tr_{\partial\Omega}\, v_0$ of a subsequence $(v_{n_k})_{k\in\mathbb{N}}$ almost everywhere on $\partial\Omega$ for $k \to \infty$. Therefore, we have $tr_{\partial\Omega}\, v_0 \geq 0$ and

$$tr_{\partial\Omega}\, v = tr_{\partial\Omega}\,(v_0 + u_c) = (tr_{\partial\Omega}\, v_0) + u_c \geq u_c$$

almost everywhere on $\partial\Omega$. □

Note that with the same approximating sequence as in this proof we can conclude the corresponding result with $tr_\Sigma\, v = (tr_{\partial\Omega}\, v)_{|\Sigma}$ for any $\Sigma \subset \partial\Omega$ as on page 249, e.g. $\Sigma \in \{\gamma_D(t), \gamma_S(t), \gamma_N(t)\}$.

For completeness, we point out that the weak Neumann boundary data $f_N(t)$ can be chosen as a distribution from $H^{-1/2}(\gamma_N(t))$ or even from $H_{00}^{1/2}(\gamma_N(t))'$. However, we restrict ourselves to functions $f_N(t) \in L^2(\gamma_N(t))$.

The canonical solution space for parabolic problems on the open time cylinder $Q := \Omega \times (0,T)$ usually is the function space $L^2(0,T;H^1(\Omega))$ (see pages 252–254 in the appendix for more details on such spaces). But then the partial time derivative u_t is in general an element of the space $H^{-1}(0,T;H^1(\Omega))$, i.e. it does in general no longer have an interpretation as a function. In order to make sense of a generalization of (1.5.12), we require a solution $u \in L^2(0,T;H^1(\Omega))$ to have a regularity such that $M(u)_t \in L^2(\Omega)$ almost everywhere on $(0,T]$. In this case, we say that u is a weak solution of the variational inequality (1.5.12) at the time $t \in (0,T]$ if $u(t) \in \mathcal{K}(t)$ and

$$\begin{aligned}\int_\Omega M(u)_t\,(v-u)\,dx + \int_\Omega \nabla u\,\nabla(v-u)\,dx \geq \\ \int_\Omega kr(M(u))e_z\nabla(v-u)\,dx - \int_{\gamma_N(t)} f_N(t)\,(v-u)\,d\sigma \quad \forall v \in \mathcal{K}(t)\,. \qquad (1.5.21)\end{aligned}$$

Note that $kr(M(u(\cdot)))$ is bounded on Ω since both M and kr are bounded. Moreover, since M and kr are both monotonically increasing, so is $kr \circ M$. Now, it is straightforward to prove that the composition of a monotonic and a measurable function is measurable again, hence $kr(M(u(\cdot))) \in L^\infty(\Omega)$ (without using the continuity of $kr \circ M$). So all of the terms in (1.5.21) make sense.

1.5.4 Kirchhoff transformation as a superposition operator and a weak variational inequality for the Richards equation in physical variables

Finally, as for the classical case (1.5.12), one can consider a weak variational inequality analogous to (1.5.21) for the untransformed Richards equation (1.5.1) rather than for the Kirchhoff–transformed version (1.5.2). However, in contrast to the classical case in which the formulations are equivalent, the equivalence of (1.5.21) with a corresponding weak formulation for the physical pressure is not clear. And it will turn out that we need quite a lot of preparatory work in order to be able to answer this question, see the Concluding Remarks 1.5.20 of this subsection.

The first problem is that u_c can be in the range of $u(t)$ in (1.5.21) which would correspond to $-\infty$ being in the range of $p(t)$. And even if this is not the case $p(t) = \kappa^{-1}(u(t))$ does not need to be in $H^1(\Omega)$ in case of the Brooks–Corey functions if $u(t) \in H^1(\Omega)$. The second problem is the question whether a chain rule

$$\nabla u = \nabla(\kappa(p)) = \kappa'(p)\nabla p = kr(\theta(p))\nabla p \qquad (1.5.22)$$

such as (1.3.3) also holds in a weak sense for Sobolev functions p on Ω with the Kirchhoff transformation κ in (1.3.1) applied pointwise almost everywhere on Ω with $kr\circ\theta \in L^\infty(\mathbb{R})$. Concerning this question we refer to Leoni and Morini [63] where (1.5.22) is proved for functions with values in a finite-dimensional space based on a known result for real-valued functions. The latter is needed here and states that (1.5.22) holds almost everywhere on Ω if $p \in W^{1,1}_{\mathrm{loc}}(\Omega)$ and

$\kappa : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous. Since $\nabla p(x) = 0$ holds for all x for which $\kappa'(p(x))$ is undefined due to Rademacher's theorem (see [97, pp. 341–349] or [104, pp. 50/51]) and a result from Serrin and Varberg [85], we have to interpret $\kappa'(p(x))\nabla p(x)$ as zero in (1.5.22) for such x. For $kr \circ \theta \in L^\infty(\mathbb{R})$ the function $\kappa : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with $\kappa' = kr \circ \theta$ (see for example [15, p. 25]) and the last equation in (1.5.22) holds. In the following we formulate some results based on these observations. We start with properties of the Kirchhoff transformation.

Lemma 1.5.7. *If $kr \circ \theta \in L^\infty(\mathbb{R})$ (is nonnegative almost everywhere), then $\kappa : \mathbb{R} \to \mathbb{R}$ as defined by (1.3.1) is (strictly monotonically increasing and) Lipschitz continuous with the Lipschitz constant*

$$L(\kappa) = \|kr \circ \theta\|_\infty \quad \textit{and} \quad \kappa' = kr \circ \theta \quad \textit{a.e. on } \mathbb{R}\,.$$

If, in addition, there is a $c > 0$ such that $kr(s) \geq c$ holds for almost all $s \in \mathbb{R}$, the Kirchhoff transformation has a strictly monotonically increasing and Lipschitz continuous inverse $\kappa^{-1} : \mathbb{R} \to \mathbb{R}$ with the Lipschitz constant

$$L(\kappa^{-1}) = \|(kr \circ M)^{-1}\|_\infty \leq c^{-1} \quad \textit{and} \quad (\kappa^{-1})' = (kr \circ M)^{-1} \quad \textit{a.e. on } \mathbb{R}$$

with M as defined in (1.3.2).

Proof. The assertions on κ are well known and follow from the fundamental theorem of calculus for Lebesgue integrals or the theory of Lebesgue points (see [97, pp. 341–349] and [15, p. 25] or [82, pp. 138–147] for more details). The assertions on κ^{-1} follow from the properties of κ. In particular, note that

$$(\kappa^{-1})'(u) = \frac{1}{\kappa'(\kappa^{-1}(u))} = \frac{1}{kr(M(u))} \tag{1.5.23}$$

is satisfied in the classical sense for all $u \in \mathbb{R}$ for which κ is differentiable in $p = \kappa^{-1}(u)$, that is for almost all $u \in \mathbb{R}$ because Lebesgue nullsets are invariant under Lipschitz mappings. □

Remark 1.5.8. Note that in case of $kr \geq c > 0$ and also in case of the Brooks–Corey functions we can write

$$\kappa^{-1}(u) = \int_0^u \frac{1}{kr(M(s))}\,ds$$

for all $u \in \mathbb{R}$ or $u \in (u_c, \infty)$, respectively, analogously to (1.3.1). Of course, the function kr only needs to be given on the range $\theta(\mathbb{R})$ in Lemma 1.5.7. For simplicity, however, we consider it to be extended to $\mathbb{R}$.

Furthermore, it should be clear from the proof of Lemma 1.5.7 that, conversely, any strictly monotonically increasing Lipschitz continuous function $\kappa : \mathbb{R} \to \mathbb{R}$ can be regarded as a Kirchhoff transformation via the fundamental theorem

$$\kappa(p) = \int_0^p \kappa'(q)\,dq + \kappa(0) \quad \forall p \in \mathbb{R} \tag{1.5.24}$$

induced by the nonnegative function $\kappa' \in L^\infty(\mathbb{R})$ with $\|\kappa'\|_\infty = L(\kappa)$. The same holds with κ^{-1} instead of κ if κ^{-1} exists and is Lipschitz continuous or equivalently if $\|(\kappa^{-1})'\|_\infty = \|(\kappa')^{-1}\|_\infty < \infty$. Although this result is straightforward, we note it here since we use the framework of these general Kirchhoff transformations in Chapter 3. Moreover, we remark that most of the following results hold in an analogous way if $\kappa(0) \neq 0$ in (1.5.24). In view of the Richards equation, however, we mostly stick to the integrand $kr \circ \theta \in L^\infty(\mathbb{R})$ as the notation in what is to come, even where we do not require its nonnegativity and the corresponding primitive does not provide a transformation.

So far we have not yet explicitly distinguished the Kirchhoff transformation $\kappa : \mathbb{R} \to \mathbb{R}$ acting as a function on real numbers p from the transformation which it provides by pointwise (almost everywhere) application on a function p defined (almost everywhere) on Ω. At this point it is appropriate to do this.

Definition 1.5.9. Let p be a real-valued function defined on a subset $S \subset \mathbb{R}^d$, possibly almost everywhere with respect to an appropriate measure. Furthermore, let $\kappa : \mathbb{R} \to \mathbb{R}$ be a real function. By the pointwise application

$$(\kappa_S(p))(x) := \kappa(p(x))$$

of κ to p (for x almost everywhere) on S the *superposition operator*

$$\kappa_S : p \mapsto \kappa(p)$$

is defined. Let X be a normed space consisting of a subset of all measurable functions on the open set S. If the superposition operator satisfies $\kappa_S(p) \in X$ for all $p \in X$, we say that it *acts on* the space X. In this case we write

$$\kappa_X : X \to X$$

for the restriction of κ_S on the space X and call κ_X superposition operator on X. Analogously one defines superposition operators *acting between* two spaces X_1 and X_2.

There is a vast theory on superposition operators, also known as *Nemytskij operators*, acting on function spaces of all kinds. For an introduction into this theory we refer to the monograph of Appell and Zabrejko [6] which contains a large reference list on the topic. Note that we have restricted ourselves to the *autonomous case* $p \mapsto \kappa(p)$ on $\mathbb{R}$ instead of the general one given by a function $(x, p) \mapsto \kappa(x, p(x))$ on $\Omega \times \mathbb{R}$. In the following we investigate the Kirchhoff transformation as a superposition operator acting on different spaces which are relevant for us. As usual in this work, the "appropriate measures" mentioned in the definition are the Lebesgue measure on a Lipschitz domain $S = \Omega \subset \mathbb{R}^d$ or else the Hausdorff measure on a submanifold $S = \Sigma \subset \partial\Omega$ of its boundary.

Lemma 1.5.10. *If $kr \circ \theta \in L^\infty(\mathbb{R})$ then $\kappa : \mathbb{R} \to \mathbb{R}$ given in (1.3.1) induces a Lipschitz continuous superposition operator*

$$\kappa_{L^2(\Omega)} : L^2(\Omega) \to L^2(\Omega)\,.$$

The corresponding Lipschitz constants satisfy

$$L(\kappa_{L^2(\Omega)}) = L(\kappa) = \|kr \circ \theta\|_\infty .$$

If, in addition, $kr \geq c > 0$ holds almost everywhere on $\mathbb{R}$, then $\kappa_{L^2(\Omega)}$ has a Lipschitz continuous inverse given by the superposition operator

$$\kappa^{-1}_{L^2(\Omega)} = (\kappa^{-1})_{L^2(\Omega)} : L^2(\Omega) \to L^2(\Omega) .$$

The Lipschitz constants satisfy

$$L(\kappa^{-1}_{L^2(\Omega)}) = L(\kappa^{-1}) = \|(kr \circ M)^{-1}\|_\infty \leq c^{-1}$$

with M as defined in (1.3.2).

With a glance at Lemma 1.5.7 the proof is straightforward, and so is the proof of the next result. We just remark that Lebesgue measurability of composites of Lebesgue measurable functions can be proved by going back to Borel measurable representatives.

Lemma 1.5.11. *Let $\Omega \subset \mathbb{R}^d$ be bounded and open and $\Sigma \subset \partial\Omega$ a Lipschitz submanifold. If $kr \circ \theta \in L^\infty(\mathbb{R})$ then $\kappa : \mathbb{R} \to \mathbb{R}$ given in (1.3.1) induces a Lipschitz continuous superposition operator*

$$\kappa_{L^2(\Sigma)} : L^2(\Sigma) \to L^2(\Sigma) .$$

The corresponding Lipschitz constants satisfy

$$L(\kappa_{L^2(\Sigma)}) = L(\kappa) = \|kr \circ \theta\|_\infty .$$

If, in addition, $kr \geq c > 0$ holds almost everywhere on $\mathbb{R}$, then $\kappa_{L^2(\Sigma)}$ has a Lipschitz continuous inverse given by the superposition operator

$$\kappa^{-1}_{L^2(\Sigma)} = (\kappa^{-1})_{L^2(\Sigma)} : L^2(\Sigma) \to L^2(\Sigma) .$$

The Lipschitz constants satisfy

$$L(\kappa^{-1}_{L^2(\Sigma)}) = L(\kappa^{-1}) = \|(kr \circ M)^{-1}\|_\infty \leq c^{-1}$$

with M as defined in (1.3.2).

Now, with these preliminaries and based on the weak chain rule (1.5.22), we obtain the following

Proposition 1.5.12. *Let $\Omega \subset \mathbb{R}^d$ be an open set, $kr \circ \theta \in L^\infty(\mathbb{R})$ and $\kappa : \mathbb{R} \to \mathbb{R}$ as defined in (1.3.1). Then for $p \in H^1(\Omega)$ we obtain $u = \kappa_\Omega(p) \in H^1(\Omega)$, and the weak chain rule (1.5.22) holds in $(L^2(\Omega))^d$. Moreover, we have*

$$\|u\|_1 \leq \|kr \circ \theta\|_\infty \|p\|_1 . \tag{1.5.25}$$

Conversely, if in addition to $kr \circ \theta \in L^\infty(\mathbb{R})$ there is a $c > 0$ such that $kr(s) \geq c$ holds for almost all $s \in \mathbb{R}$, then we also have

$$c\,\|p\|_1 \leq \|u\|_1 . \tag{1.5.26}$$

Finally, with this latter condition $u \in H^1(\Omega)$ implies $p = (\kappa^{-1})_\Omega(u) \in H^1(\Omega)$.

Proof. For $p \in H^1(\Omega)$ the chain rule (1.5.22) holds almost everywhere on Ω. But since $p \in H^1(\Omega)$ we also have

$$\int_\Omega |\nabla u(x)|^2\,dx = \int_\Omega |kr(\theta(p(x)))|^2 |\nabla p(x)|^2\,dx \le \|kr\circ\theta\|_\infty^2 \int_\Omega |\nabla p(x)|^2\,dx \tag{1.5.27}$$

and furthermore

$$\int_\Omega |u(x)|^2\,dx \le \|kr\circ\theta\|_\infty^2 \int_\Omega |p(x)|^2\,dx \tag{1.5.28}$$

because of

$$|u(x)| = \left|\int_0^{p(x)} kr(\theta(q))\,dq\right| \le \|kr\circ\theta\|_\infty |p(x)| \tag{1.5.29}$$

almost everywhere on Ω. This proves (1.5.25). The converse (1.5.26) follows in the same way from (1.5.22) and $kr \ge c > 0$. However, (1.5.22) is not immediately clear because we do not know anything about p. Therefore, we first apply (1.5.22) in the form

$$\nabla p = \nabla(\kappa^{-1}(u)) = (\kappa^{-1})'(u)\nabla u \tag{1.5.30}$$

almost everywhere on Ω using the regularity of u. With the nondegeneracy assumption $kr \ge c > 0$ we can apply the formula (1.5.23) and obtain

$$(\kappa^{-1})'(u(x)) = \frac{1}{\kappa'(p(x))} = \frac{1}{kr(\theta(p(x)))} > 0 \tag{1.5.31}$$

wherever κ' is (pointwise classically) differentiable which is almost everywhere. Now, (1.5.22) follows from (1.5.30) and (1.5.31). □

Remark 1.5.13. Note that in case of $\kappa(0) \neq 0$, (1.5.29) produces an additional additive constant in the integrand on the right hand side of (1.5.28) such that the assertions of Proposition 1.5.12 cannot be established in this case if Ω is unbounded. Furthermore, we remark that $\kappa : \mathbb{R} \to \mathbb{R}$ is linear (and induces linear maps on $L^2(\Omega)$ and $H^1(\Omega)$) if and only if $kr \circ \theta : \mathbb{R} \to \mathbb{R}$ is constant. Otherwise observe that although we have (1.5.27), we cannot expect κ to induce a Lipschitz continuous map on $H^1(\Omega)$ (except for affine κ) because the Kirchhoff transformation of a function is defined pointwise and does not “see” derivatives of the function. However, under reasonable conditions we can still prove the continuity by rather elementary means. The next proposition covers the Brooks–Corey functions but, of course, not the limit cases in Section 1.4.

Proposition 1.5.14. *Let $\Omega \subset \mathbb{R}^d$ be an open set, $kr \circ \theta : \mathbb{R} \to \mathbb{R}$ uniformly continuous and bounded and the primitive $\kappa : \mathbb{R} \to \mathbb{R}$ as defined in (1.3.1). Then the superposition operator*

$$\kappa_{H^1(\Omega)} : H^1(\Omega) \to H^1(\Omega)$$

obtained in Proposition 1.5.12 is continuous. If $d = 1$ and $kr \circ \theta : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous and bounded, the superposition operator is Lipschitz continuous on any bounded subset of $H^1(\Omega)$.

Proof. We choose a fixed $p_0 \in H^1(\Omega)$ and a sequence $(p_n)_{n\in\mathbb{N}} \subset H^1(\Omega)$ converging to p_0 in $H^1(\Omega)$. Since κ_Ω is Lipschitz continuous on $L^2(\Omega)$ and the chain rule (1.3.3) holds, it suffices to show that

$$\int_\Omega |kr(\theta(p_n))\nabla p_n - kr(\theta(p_0))\nabla p_0|^2 \, dx$$

becomes small for small $\|p_n - p_0\|_1$. This term can be estimated by

$$2\|kr\circ\theta\|_\infty^2 \int_\Omega |\nabla(p_n - p_0)|^2 \, dx + 2\int_\Omega |kr(\theta(p_n)) - kr(\theta(p_0))|^2 |\nabla p_0|^2 \, dx \quad (1.5.32)$$

which means that it is enough to show that the second integral I in (1.5.32) becomes small for small $\|p_n - p_0\|_1$. But this follows already from the convergence of $p_n \to p_0$ in $L^2(\Omega)$ for $n \to \infty$ which forces the Lebesgue measure of the set

$$M_n := \{x \in \Omega : |p_n(x) - p_0(x)|^2 > \delta\}$$

to go to 0 for $n \to \infty$ and any fixed $\delta > 0$. With the ε-δ-criterion for the uniform continuity of $kr \circ \theta$ we can estimate I by

$$\varepsilon^2 \int_{\Omega\setminus M_n} |\nabla p_0|^2 \, dx + 2\|kr\circ\theta\|_\infty^2 \int_{M_n} |\nabla p_0|^2 \, dx$$

in which the second integral goes to 0 for $n \to \infty$ since the first integral goes to $\int_\Omega |\nabla p_0|^2 \, dx$ due to the theorem of Lebesgue (or, alternatively, of Beppo Levi).

If $d = 1$ and $kr \circ \theta : \mathbb{R} \to \mathbb{R}$ is even Lipschitz continuous, we can estimate $|kr(\theta(p_n)) - kr(\theta(p_0))|$ in the integral I by

$$L(kr\circ\theta)\|p_n - p_0\|_\infty \leq C\|p_n - p_0\|_1$$

for a $C > 0$ with the help of Sobolev's embedding theorem (2.5.36) in one space dimension. □

To put it mildly, the next result is astonishing. Not only does it state that Proposition 1.5.14 can be generalized, but in fact, that its assertion can virtually never be wrong.

Theorem 1.5.15. *Let $\Omega \subset \mathbb{R}^d$ be a bounded open set and $\kappa : \mathbb{R} \to \mathbb{R}$ a Borel function. The superposition operator κ_Ω acts on $H^1(\Omega)$, i.e. it induces a map*

$$\kappa_{H^1(\Omega)} : H^1(\Omega) \to H^1(\Omega),$$

if and only if it is continuous on $H^1(\Omega)$ or, equivalently, if and only if κ is Lipschitz continuous for $d > 1$ or locally Lipschitz in the case $d = 1$, respectively.

This result was proved in Marcus and Mizel [70, pp. 218–220] for (locally) Lipschitz continuous functions $\kappa : \mathbb{R} \to \mathbb{R}$ and also for unbounded sets $\Omega \subset \mathbb{R}^d$ if $\kappa(0) = 0$, see Remark 1.5.13. For bounded $\Omega \subset \mathbb{R}^d$ the (local) Lipschitz continuity of $\kappa : \mathbb{R} \to \mathbb{R}$ was deduced in Marcus and Mizel [69] as an *acting* *condition* (see [6, pp. 239–243]) from the sheer fact that the Borel function κ induces a superposition operator from $H^1(\Omega)$ into itself (see also [83, p. 267]). We already know from Proposition 1.5.12 (and we can easily see it for $d = 1$) that the converse assertion in Theorem 1.5.15 is true. In addition to the continuity of $\kappa_{H^1(\Omega)}$, one also obtains its boundedness in the sense that

$$\|\kappa_{H^1(\Omega)} v\|_1 \leq C(1 + \|v\|_1) \quad \forall v \in H^1(\Omega)$$

holds for a $C > 0$ independent of v for $d > 1$ analogously to (1.5.25), and for $C(M) > 0$ with $\|v\|_1 \leq M$ for $d = 1$. There are analogous statements about superposition operators mapping $W^{1,p}(\Omega)$ into $W^{1,r}(\Omega)$ for $p, r \geq 1$.

We note a nice consequence of Theorem 1.5.15. If this fact was not written down here, one would probably take it for granted as something the trace theorem should certainly provide, and one would be shocked if it did not hold for reasonable, i.e. continuous $\kappa : \mathbb{R} \to \mathbb{R}$. Fortunately (in this case), our functions defined almost everywhere behave as naturally as they "should". In fact, the condition that κ should be continuous turns out to be redundant.

Proposition 1.5.16. *Let $\Omega \subset \mathbb{R}^d$ be bounded and open and $\Sigma \subset \partial\Omega$ a Lipschitz submanifold. If $\kappa : \mathbb{R} \to \mathbb{R}$ is a Borel function and the superposition operator κ_Ω acts on $H^1(\Omega)$, then we have the commutativity*

$$\kappa_\Sigma(tr_\Sigma v) = tr_\Sigma(\kappa_\Omega v) \quad \forall v \in H^1(\Omega)\,. \tag{1.5.33}$$

Proof. We prove that for any $v \in H^1(\Omega)$

$$\|tr_\Sigma(\kappa_\Omega v) - \kappa_\Sigma(tr_\Sigma v)\|_{L^2(\Omega)} \tag{1.5.34}$$

is arbitrarily small by considering a sequence $(v_n)_{n\in\mathbb{N}} \subset C^\infty(\overline{\Omega})$ converging to v in $H^1(\Omega)$. In fact, since Theorem 1.5.15 provides the continuity of κ, the norm in (1.5.34) can be estimated by

$$\|tr_\Sigma(\kappa_\Omega v) - (\kappa_\Omega v_n)_{|\Sigma}\|_{L^2(\Omega)} + \|\kappa_\Sigma(v_{n|\Sigma}) - \kappa_\Sigma(tr_\Sigma v)\|_{L^2(\Omega)}\,. \tag{1.5.35}$$

The first term in (1.5.35) is at most

$$\|tr_\Sigma\|\,\|\kappa_\Omega v - \kappa_\Omega v_n\|_1$$

due to the trace theorem A.2.3, and this estimate goes to 0 for $n \to \infty$ by the continuity of $\kappa_{H^1(\Omega)}$. The second term in (1.5.35) can be estimated by

$$L(\kappa_{L^2(\Sigma)})\,\|v_{n|\Sigma} - tr_\Sigma v\|_{L^2(\Sigma)} \leq L(\kappa_{L^2(\Sigma)})\,\|tr_\Sigma\|\,\|v_n - v\|_1$$

with Lemma 1.5.11 (for $d > 1$ where $\kappa : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous) and the trace theorem A.2.3 and, therefore, tends to 0 for $n \to \infty$, too. Note that for $d = 1$ the real function κ is locally Lipschitz continuous on $\mathbb{R}$ due to Theorem 1.5.15 and, consequently, it is also locally Lipschitz continuous as a superposition operator on $H^{1/2}(\Sigma)$ which is isomorphic to $\mathbb{R}$ or to $\mathbb{R} \times \mathbb{R}$ with any norm. Moreover, in one space dimension (1.5.33) is trivial anyway due to the Sobolev embedding (2.5.36). □

The message of this proposition is that (1.5.33) is true as soon as the right hand side of it makes sense for all $v \in H^1(\Omega)$, and this cannot be the case for discontinuous $\kappa : \mathbb{R} \to \mathbb{R}$. However, it seems that Proposition 1.5.16 cannot be proved without using the continuity of the superposition operator $\kappa_{H^1(\Omega)}$, i.e. the very strong general result from Theorem 1.5.15. Strangely enough, the commutativity (1.5.33) appears to be a very natural property for continuous $\kappa : \mathbb{R} \to \mathbb{R}$ (one may think of C^∞-functions converging in $H^1(\Omega)$ and $H^{1/2}(\Sigma)$ and w.l.o.g. at the same time almost everywhere on Ω to v and almost everywhere on Σ to $tr_\Sigma v$.) Still, it seems to depend on something which, even in this case, can be regarded as quite surprising.

The commutativity (1.5.33) will be important in the proof of Theorem 1.5.18 which relates the physical and the generalized solution of the weak Signorini-type problem for the Richards equation. Moreover, it will already be essential for an adequate formulation and treatment of Dirichlet boundary conditions in the domain decomposition in Chapter 3. With regard to our theory presented in that chapter we note another important consequence of (1.5.33) here.

Proposition 1.5.17. *Let $\Omega \subset \mathbb{R}^d$ be bounded and open and $\Sigma \subset \partial\Omega$ a Lipschitz submanifold. If $\kappa : \mathbb{R} \to \mathbb{R}$ is a Borel function and the corresponding superposition operator κ_Ω acts on $H^1(\Omega)$, then the superposition operator κ_Σ acts on $H^{1/2}(\Sigma)$ and is continuous. If $\kappa(0) = 0$, then κ_Σ also acts on $H_0^{1/2}(\Sigma)$ and $H_{00}^{1/2}(\Sigma)$ and is continuous on these trace spaces, too.*

Proof. With the continuous extension operator $R_\Sigma : H^{1/2}(\Sigma) \to H^1(\Omega)$ given by the trace theorem A.2.3 and using Proposition 1.5.16 we can write

$$\kappa_\Sigma = \kappa_\Sigma \circ tr_\Sigma \circ R_\Sigma = tr_\Sigma \circ \kappa_{H^1(\Omega)} \circ R_\Sigma$$

and the operator on the right hand side is a composition of continuous operators which obviously acts on $H^{1/2}(\Sigma)$. This also shows the continuity of the superposition operator on $H_0^{1/2}(\Sigma)$ once it acts on this space. The latter can be seen by the definition of $H_0^{1/2}(\Sigma)$ on page 248.

Assume that $\eta \in H_0^{1/2}(\Sigma)$ and $(v_n)_{n\in\mathbb{N}}$ is a sequence of functions in $C^\infty(\overline{\Omega})$ such that each v_n vanishes on a neighbourhood of $\partial\Omega\backslash\Sigma$ and $v_{n|\Sigma} \to \eta$ for $n \to \infty$ in $H^{1/2}(\Sigma)$. Then the sequence $(\kappa_\Omega(v_n))_{n\in\mathbb{N}}$ lies in $C(\overline{\Omega}) \subset H^1(\Omega)$ and the support of each $\kappa_\Omega(v_n)$ is contained in the support of v_n since $\kappa(0) = 0$. In addition, we have

$$\kappa_\Sigma(v_{n|\Sigma}) \to \kappa_\Sigma(\eta) \quad \text{for } n \to \infty \text{ in } H^{1/2}(\Sigma)$$

due to the continuity of $\kappa_{H^{1/2}(\Sigma)}$. Therefore, since each $\kappa_\Sigma(v_{n|\Sigma}) \in H_0^{1/2}(\Sigma)$ can be approximated in $H^{1/2}(\Sigma)$ by a sequence $(w_{n,m})_{m\in\mathbb{N}} \subset C^\infty(\overline{\Omega})$ such that each $w_{n,m}$ vanishes on a neighbourhood of $\partial\Omega\backslash\Sigma$, we also obtain such an approximating sequence for $\kappa_\Sigma(\eta)$, i.e. by definition $\kappa_\Sigma(\eta) \in H_0^{1/2}(\Sigma)$.

In order to see that κ_Σ acts on $H_{00}^{1/2}(\Sigma)$ and is continuous, we refer to the definition (A.2.4). Let $\eta \in H_{00}^{1/2}(\Sigma)$ and $\tilde{\eta}$ be a trivial extension of η in $H^{1/2}(\partial\Omega)$.

Then, since $\kappa(0) = 0$ and $\kappa_{\partial\Omega}$ acts on the space $H^{1/2}(\partial\Omega)$, we can conclude $\kappa_{\partial\Omega}(\tilde{\eta}) \in H^{1/2}(\partial\Omega)$ and $\kappa_{\partial\Omega}(\tilde{\eta})_{|\Sigma}$ is a trivial extension of $\kappa_\Sigma(\eta) \in H^{1/2}(\Sigma)$, i.e. by definition $\kappa_\Sigma(\eta) \in H_{00}^{1/2}(\Sigma)$. Moreover, if $\mu \in H_{00}^{1/2}(\Sigma)$ is treated as η, then $\kappa_{\partial\Omega}(\tilde{\eta}) - \kappa_{\partial\Omega}(\tilde{\mu}) \in H^{1/2}(\partial\Omega)$ is a trivial extension of $\kappa_\Sigma(\eta) - \kappa_\Sigma(\mu) \in H_{00}^{1/2}(\Sigma)$. Now, (A.2.5) and the continuity of $\kappa_{\partial\Omega}$ provide that for any $\varepsilon > 0$ we have

$$\|\kappa_\Sigma(\eta) - \kappa_\Sigma(\mu)\|_{H_{00}^{1/2}(\Sigma)} = \|\kappa_{\partial\Omega}(\tilde{\eta}) - \kappa_{\partial\Omega}(\tilde{\mu})\|_{H^{1/2}(\partial\Omega)} \leq \varepsilon$$

if $\|\tilde{\eta} - \tilde{\mu}\|_{H^{1/2}(\partial\Omega)} = \|\eta - \mu\|_{H_{00}^{1/2}(\Sigma)} \leq \delta$ holds with a suitable $\delta > 0$. □

With the results collected so far in this subsection, we have given an overview of how the pointwise application of κ on p almost everywhere on Ω or on $\Sigma \subset \partial\Omega$ and in all spaces, which are relevant for us, can or should be understood in terms of superposition operators. With these results and knowing what we are talking about, we can again — and will from now on — just talk simultaneously of $\kappa(p)$ as a real number or as a function on Ω or on Σ. Note that the latter would be impossible if (1.5.33) did not hold.

We are now in a position to relate (1.5.21) to a corresponding weak variational inequality for the Richards equation in the physical pressure p. Analogously to (1.5.18) and (1.5.19) let

$$p_D(t) \in \{v = tr_{\gamma_D(t)}w : w \in H^1(\Omega) \ \wedge \ tr_{\gamma_S(t)}w \leq 0 \ \text{ a.e. on } \ \gamma_S(t)\} \qquad (1.5.36)$$

and the nonempty closed and convex set

$$\mathcal{K}_0(t) := \{v \in H^1(\Omega) : tr_{\gamma_D(t)}v = p_D(t) \wedge tr_{\gamma_S(t)}v \leq 0 \ \text{ a.e. on } \ \gamma_S(t)\}\,. \quad (1.5.37)$$

Then we say $p : \Omega \times (0, T] \to \mathbb{R}$ is a weak solution of the Signorini-type problem for the Richards equation (1.5.1) corresponding to (1.5.4)–(1.5.7) at the time $t \in (0, T]$ if $p(t) \in \mathcal{K}_0(t)$ and

$$\begin{aligned}\int_\Omega \theta(p)_t\,(v-p)\,dx + \int_\Omega kr(\theta(p))\nabla p\,\nabla(v-p)\,dx \geq \\ \int_\Omega kr(\theta(p))e_z\nabla(v-p)\,dx - \int_{\gamma_N(t)} f_N(t)\,(v-p)\,d\sigma \quad \forall v \in \mathcal{K}_0(t)\,. \quad (1.5.38)\end{aligned}$$

As for (1.5.21), in order to make sense of (1.5.38) we assume that a solution $p \in L^2(0, T; H^1(\Omega))$ is regular enough such that $\theta(p)_t \in L^2(\Omega)$ holds almost everywhere on $(0, T]$. Then we can state the following

Theorem 1.5.18. *Let $\theta : \mathbb{R} \to \mathbb{R}$ and $kr : \theta(\mathbb{R}) \to (0, 1]$ be bounded and monotonically increasing while $\kappa : \mathbb{R} \to \mathbb{R}$ is defined by (1.3.1). In addition, let $u_D(t) := \kappa(p_D(t))$. Then $u(t) = \kappa(p(t))$ solves (1.5.21) if $p(t)$ solves (1.5.38). If, in addition, $kr \geq c$ holds for a $c > 0$ and $\gamma_S(t) = \emptyset$, then (1.5.21) and (1.5.38) are equivalent in the sense that $u(t)$ satisfies (1.5.21) if and only if $p(t) = \kappa^{-1}(u(t))$ satisfies (1.5.38).*

Proof. First, if $p(t)$ solves (1.5.38) and $p \in L^2(0,T;H^1(\Omega))$, then by Proposition 1.5.12 we have $u(t) = \kappa(p(t)) \in H^1(\Omega)$ for almost all $t \in (0,T]$ and $\|u(t)\|_1 \leq \|kr \circ \theta\|_\infty \|p(t)\|_1$ for these t, which also gives $u \in L^2(0,T;H^1(\Omega))$. Furthermore, we have $\theta(p(t)) = \theta(\kappa^{-1}(\kappa(p(t)))) = M(u(t)) \in L^\infty(\Omega)$ as well as $kr(\theta(p(t))) = kr(M(u(t))) \in L^\infty(\Omega)$ for almost all $t \in (0,T]$, in particular these composite functions are all measurable due to the conditions on θ and kr. This and the chain rule (Proposition 1.5.12) give an equivalent formulation of (1.5.38) in terms of u in which only the test functions $v - p$ with $v \in \mathcal{K}_0(t)$ differ from those in (1.5.21). However, the set of test functions $\mathcal{K}_0(t) - p(t)$ contains the set of test functions $\mathcal{K}(t) - u(t)$ considered in (1.5.21). In order to see that

$$\mathcal{K}(t) - u(t) = \{v \in H^1(\Omega) : v \geq u_c - u \wedge tr_{\gamma_D(t)}v = 0 \wedge tr_{\gamma_S(t)}v \leq -tr_{\gamma_S(t)}\, u(t)\}$$

is a subset of

$$\mathcal{K}_0(t) - p(t) = \{v \in H^1(\Omega) : tr_{\gamma_D(t)}v = 0 \wedge tr_{\gamma_S(t)}v \leq -tr_{\gamma_S(t)}\, p(t)\}$$

note that $tr_{\gamma_S(t)}\, p(t) \leq 0$ and therefore

$$tr_{\gamma_S(t)}\, p(t) \leq tr_{\gamma_S(t)}\, \kappa(p(t)) = tr_{\gamma_S(t)}\, u(t) \leq 0$$

holds almost everywhere on $\gamma_S(t)$. Here the two inequalities are due to the (pointwise) definition (1.3.1) of κ and the fact that the range of kr is contained in $(0,1]$. The first equality is (1.5.33) and the second is the definition of $u(t)$.

Finally, observe that $p(t) \in \mathcal{K}_0(t)$ entails $u(t) \in \mathcal{K}(t)$. First, with $p_D(t)$ as in (1.5.36), the function $u_D(t) := \kappa(p_D(t))$ almost everywhere on $\gamma_D(t)$ is an admissible Dirichlet condition for $\mathcal{K}(t)$ contained in the set given in (1.5.18). To see this we choose $\tilde{w} := \kappa(w)$ as an admissible extension of $u_D(t)$ in $H^1(\Omega)$ if w is an extension of $p(t)$ in (1.5.36) and do not forget to apply (1.5.33). (Knowing that $p(t)$ solves (1.5.38) we could of course choose $w = p(t)$ and $\tilde{w} = u(t)$.) Secondly, from $tr_{\gamma_D(t)}\, p(t) = p_D(t)$ we can conclude $tr_{\gamma_D(t)}\, u(t) = u_D(t)$, once again with (1.5.33). This proves the first statement of the proposition.

For the equivalence result one can also argue backwards from $u(t)$ to $p(t)$ using Proposition 1.5.12 and the observation that the set of test functions $\mathcal{K}(t) - u(t)$ and $\mathcal{K}_0(t) - p(t)$ both coincide with $H^1_{\gamma_D}(\Omega)$ if $\gamma_S(t) = \emptyset$. □

We note that even though parameter functions θ and kr in the Richards equation are always monotonically increasing, the above proposition holds for general $\theta, kr \in L^\infty(\mathbb{R})$ with $kr > 0$. As indicated earlier, the Lebesgue measurability of composites $kr(\theta(p(\cdot)))$ for Lebesgue measurable $p(\cdot)$ on Ω can then be deduced by considering Borel measurable representatives.

Remark 1.5.19. In Section 2.3 we will show that a time-discretized version of (1.5.21) is uniquely solvable in $\mathcal{K}(t)$. Theorem 1.5.18 carries over to the corresponding time discretizations of the variational inequalities (1.5.21) and (1.5.38). In hydrologically interesting situations such as in the case of Brooks–Corey functions or similar parameter functions, the first set of conditions in Theorem 1.5.18 is satisfied. Consequently, if (1.5.38) has a physical meaning, i.e. a solution $p(t)$, then $u(t) = \kappa(p(t))$ satisfies (1.5.21) and the uniqueness of $u(t)$ implies the uniqueness of $p(t)$ since $\kappa : \mathbb{R} \to \mathbb{R}$ is invertible due to $kr > 0$. In order to have this chance at all in case of $\kappa^{-1} : (u_c, \infty) \to \mathbb{R}$ with the singularity in $u_c < 0$, one certainly needs to consider the transformed set

$$\kappa(\{w \in H^1(\Omega) : tr_{\gamma_S(t)} w \leq 0\}) \supset \{w \in H^1(\Omega) : w \geq u_c \,\wedge\, tr_{\gamma_S(t)} w \leq 0\}$$

of (1.5.36) as the admissible set for the generalized Dirichlet values in (1.5.18). Observe that this transformed set reads

$$\{w \in H^1(\Omega) : \kappa^{-1}(w) \in H^1(\Omega) \,\wedge\, tr_{\gamma_S(t)} w \leq 0\}$$

and that $\kappa^{-1}(w) \in H^1(\Omega)$ entails $\kappa^{-1}(w) > u_c$, but certainly not conversely. And even if $u_D(t) = \kappa(p_D(t))$ is "physically compatible" in this sense, we do not have an equality in

$$\kappa(\mathcal{K}_0(t)) \subset \mathcal{K}(t) = \{v \in H^1(\Omega) : v \geq u_c \,\wedge\, tr_{\gamma_D(t)} v = u_D(t) \,\wedge\, tr_{\gamma_S(t)} v \leq 0\}\,.$$

This does not change if we sharpen the condition $v \geq u_c$ to $v > u_c$ in (1.5.19). And even in this case it is easy to see that $\kappa^{-1}(u)$ does not need to be in $L^1(\Omega)$ for $u \in \mathcal{K}(t)$. On the other hand, with respect to our solution theory in Chapter 2, it is not possible to consider the smaller transformed set

$$\kappa(\mathcal{K}_0(t)) = \{v \in H^1(\Omega) : \kappa^{-1}(v) \in H^1(\Omega) \,\wedge\, tr_{\gamma_D(t)} v = u_D(t) \,\wedge\, tr_{\gamma_S(t)} v \leq 0\}$$

instead of $\mathcal{K}(t)$ because in general this set is not convex. Recall that for the Brooks–Corey functions, $\kappa : \mathbb{R} \to \mathbb{R}$ is convex and, consequently, the inverse $\kappa^{-1} : (u_c, \infty) \to \mathbb{R}$ is (strictly) concave (on $(u_c, -1]$).

In case of $kr \geq c > 0$, i.e. if the main part of the spatial derivative in the Richards equation (1.5.1) is uniformly elliptic (see the nondegenerate case in Subsection 1.4.3), we can deduce the unique solvability of (1.5.38) from the unique solvability of (1.5.21) if $\gamma_S(t) = \emptyset$. In the next chapter we prove the unique solvability of a time discretized version of (1.5.21) for general boundary conditions. In this context, we will come back to the variational inequality for the physical pressure in Remark 2.5.14. Furthermore, this inequality serves as a crucial starting point for the treatment of the Richards equation in heterogeneous soil in Section 3.2

Concluding Remarks 1.5.20. One can consider the connection between the variational inequalities (1.5.38) and (1.5.21) in Theorem 1.5.18 as the main result of this subsection. However, for the proof of Theorem 1.5.18 one needs to apply results which are of interest in themselves, such as the weak chain rule (1.5.22) in $H^1(\Omega)$, Proposition 1.5.12 and, finally, the commutativity (1.5.33) in order to deal with the situation on the Dirichlet and the Signorini-type boundary. Although at first glance, the property (1.5.33) of the Kirchhoff transformation on trace functions seems to be quite natural, a strong result (Theorem 1.5.15) from the theory of superposition operators was required to prove it. From this perspective one might be surprised about how long and deep the proof of Theorem 1.5.18 turned out to be. On the other hand, the solution theory in Chapter 2 focusses on the (time-discretized) variational inequality (1.5.21) rather than (1.5.38) such that this proposition and also a version of it for the time-discretized variational inequalities can be regarded as a basis for this approach (see Subsection 2.5.4 and in particular Remark 2.5.14).

Moreover, the theory on the Kirchhoff transformation as a superposition operator in Sobolev spaces on domains and their boundaries, in particular the commutativity (1.5.33) and the continuity (Theorem 1.5.15 and Proposition 1.5.17) will be an essential ingredient in Chapter 3 where we deal with the (time-discretized) Richards equation in heterogeneous soil and related problems. Since in that chapter, too, we will mainly examine the treatment of Kirchhoff–transformed problems rather than the original ones, the theory in this subsection will serve both as a starting point (see Remark 3.3.2 and Proposition 3.4.1) and as an endpoint (see Proposition 3.3.8 and Theorem 3.4.23) in order to obtain the desired results for the original problems.

If (1.5.38) does not have a physical meaning, one might think of variational formulations in other (possibly bigger) solution spaces or sets for $p(t)$ which we do not further investigate here. However, one often ignores this question and only deals with the generalized pressure rather than the physical one. This can be seen in the following section, where we consider time-integrated versions of (1.5.21), i.e. weak initial boundary value problems for the Kirchhoff–transformed Richards equation (with and without Signorini-type conditions) on the time cylinder $Q = \Omega \times (0, T)$, for which unique solvability is known.

1.6 Overview of analytical results for the Richards equation

In the following, we give some insight into results that have been obtained so far in the analysis of initial boundary value problems (Cauchy problems) for the Richards equation on a time cylinder $Q = \Omega \times (0, T)$. Our overview covers a one-dimensional result by van Duyn and Peletier in Subsection 1.6.1 and the elaborate theory due to Alt, Luckhaus, Visintin and Otto in Subsection 1.6.2. As far as the technicalities of the solution spaces for the considered problems are concerned, we refer to the appendix, pages 252–254. In the literature, (initial) boundary value problems for saturated-unsaturated groundwater flow are often presented as *(evolution) dam problems*, and quite a lot can be found on their analysis, see e.g. Gilardi [43], Alt [2], Carrillo and Chipot [23] or Chipot and Lyaghfouri [26]. However, these problems often assume the physical water pressure $p \geq 0$ and the saturation θ to be a step function with $\theta(0) = 0$ and $\theta(p) = 1$ for $p > 0$. We do not consider these degenerate cases here.

Obviously, whether boundary conditions of Signorini's type are included or not, the analysis of (initial) boundary value problems for the Richards equation has to deal with considerable difficulties. This is certainly due to the nonlinearities involved, but also due to the changing type of the equation, which is elliptic in the saturated and parabolic in the unsaturated regime with the water table constituting a free boundary between the two regimes.

1.6.1 One-dimensional theory of van Duyn and Peletier

One of the first papers dealing intensively with the free boundary problem from an analytical point of view seems to be of van Duyn and Peletier [90]. But here already, a simplified problem is considered only. More concretely, in the 1D-setting $\Omega = (0,1)$ and ignoring gravity, the Cauchy–Dirichlet problem for the Kirchhoff–transformed Richards equation

$$\begin{cases} M(u)_t = u_{xx} & \text{in } \Omega \times (0,T) \\ u(0,t) = -1, \;\; u(1,t) = 1 & \text{for } 0 < t \leq T \\ M(u(x,0)) = M(u_0(x)) & \text{for } x \in \Omega \end{cases} \tag{1.6.1}$$

is investigated in a straightforward weak formulation which we note below. Here, $M : [u_c, \infty) \to [0,1]$, with $u_c \leq -1$, is assumed to be Lipschitz continuous and strictly increasing on $[u_c, 0]$ and $M(s) = 1$ for $s \in [0, \infty)$. An initial value $u_0 : [0,1] \to \mathbb{R}$ is assumed to exist as a Lipschitz continuous function respecting the boundary values $u_0(0) = -1$ and $u_0(1) = 1$ as well as $u_0(x) \geq u_c$ on $[0,1]$.

We define the function $\bar{u} : [0,1] \to \mathbb{R}$ as

$$\bar{u}(x) = 2x - 1 \quad \forall x \in [0,1] \tag{1.6.2}$$

which is obviously a stationary solution of (1.6.1).

A measurable function $u : \overline{Q} \to \mathbb{R}$ is called a weak solution of (1.6.1) if we have $u \in \bar{u} + L^2(0,T; H_0^1(\Omega))$ and $M(u) \in C(\overline{Q})$ (with u possibly altered on a Lebesgue nullset) such that

$$\int_0^T \int_0^1 u_x v_x - M(u) v_t \, dx \, dt = \int_0^1 M(u_0) v(x,0) \, dx$$

holds for all test functions $v \in C^1(\overline{Q})$ vanishing on $\partial Q \backslash (\Omega \times \{0\})$.

The authors prove the existence and the uniqueness of such a weak solution. Moreover, a maximum principle for $M(u)$ is obtained, i.e. for weak solutions u_1 and u_2 corresponding to initial conditions u_{01} and u_{02} with $u_{01} \geq u_{02}$, we have $M(u_1) \geq M(u_2)$. As far as the regularity is concerned, $u \in L^2(0,T; H^2(\Omega))$ can be proved without assuming further conditions, and u is a classical solution on the unsaturated part of Q if $M_{|[u_c,0]} \in C^2([u_c, 0])$. Finally, results on the continuity of the function $g : [0,T] \to (0,1)$ determining the free boundary between the saturated and the unsaturated regime are obtained as well as the convergence of $c(u(x,t)) \to c(\bar{u})$ as $t \to \infty$ with the stationary solution $\bar{u}$ in (1.6.2).

We point out that these results do not apply to the parameter functions according to Brooks–Corey since in this case, M is not Lipschitz continuous but rather behaving like a root function, see (1.3.25) or (1.4.1).

1.6.2 Theory of Alt, Luckhaus, Visintin and Otto

A more general approach is pursued in the pioneering paper by Alt and Luckhaus [4] for a large class of quasilinear elliptic-parabolic equations. The Richards equation does not fit into this framework except in case of uniform ellipticity, i.e. if $kr(\theta) \geq c \ \forall \theta \in [\theta_m, \theta_M]$ holds for a $c > 0$. However, the results apply to the Kirchhoff–transformed Richards equation if M is a function on the whole real line. More concretely, starting with (1.5.4)–(1.5.6) and $\gamma_S(t) = \emptyset \ \forall t \in [0, T]$, we consider the initial boundary value problem

$$\begin{aligned} M(u)_t - \operatorname{div}\Big(\nabla u - kr(M(u))e_z\Big) &= 0 && \text{on } \Omega \times (0,T) && (1.6.3) \\ u &= u_D(t) && \text{on } \gamma_D \times (0,T) && (1.6.4) \\ -(\nabla u - kr(M(u))e_z) \cdot \mathbf{n} &= 0 && \text{on } \gamma_N \times (0,T) && (1.6.5) \\ M(u) &= M_0 && \text{on } \Omega \times \{0\} && (1.6.6) \end{aligned}$$

on an open bounded and connected Lipschitz domain Ω with fixed $\gamma_D \subset \partial\Omega$ of positive Hausdorff measure and $\gamma_N = \partial\Omega \backslash \gamma_D$. $M : \mathbb{R} \to \mathbb{R}$ is assumed to be monotonically increasing and continuous and $kr : M(\mathbb{R}) \to \mathbb{R}$ to be bounded and continuous. With regard to the data, let $u_D(t) \in H^{1/2}(\gamma_D)$ be the trace of a function $u^D \in L^2(0,T; H^1(\Omega)) \cap L^\infty(\Omega \times (0,T))$ for $t \in (0,T)$ (almost everywhere). Furthermore, let $M_0 \in L^1(\Omega)$ with M_0 mapping into the range of M. Then the existence of a measurable function u_0 on Ω with $M(u_0) = M_0$ can be proved.

Now, $u \in u^D + L^2(0,T; H^1_{\gamma_D}(\Omega))$ is called a weak solution of (1.6.3)–(1.6.6) if the following two conditions hold.

a) $M(u) \in L^\infty(0,T; L^1(\Omega))$ and $M(u)_t \in L^2(0,T; H^1_{\gamma_D}(\Omega)')$ and the initial values M_0 are attained in the sense

$$\int_0^T \langle M(u)_t, v\rangle \, dt + \int_0^T \int_\Omega (M(u) - M_0) v_t \, dx \, dt = 0 \qquad (1.6.7)$$

tested with all functions $v \in L^2(0,T; H^1_{\gamma_D}(\Omega)) \cap W^{1,1}(0,T; L^\infty(\Omega))$ satisfying $v(T) = 0$.

b) $\nabla u - kr(M(u))e_z \in L^2(\Omega \times (0,T))$ and u satisfies

$$\int_0^T \langle M(u)_t, v\rangle \, dt + \int_0^T \int_\Omega (\nabla u - kr(M(u))e_z) \nabla v \, dx \, dt = 0 \qquad (1.6.8)$$

for all test functions $v \in L^2(0,T; H^1_{\gamma_D}(\Omega))$.

In (1.6.7) and (1.6.8) the expression $\langle\,\cdot\,,\,\cdot\,\rangle$ stands for the duality pairing of $H^1_{\gamma_D}(\Omega)$ and $H^1_{\gamma_D}(\Omega)'$. We can replace the corresponding term in these variational equalities by $\int_\Omega M(u)_t\, v\, dx$ if $M(u)_t \in L^2(\Omega)$. Then integration by parts applied to (1.6.8) gives (1.6.3)–(1.6.5) if the regularity of the terms involved allows for the application of Green's formula (1.5.9) or its weak counterpart (A.2.12). With regard to how the initial condition is required to be respected, observe that partial time-integration applied to (1.6.7), if allowed, provides

$$\int_\Omega (M(u(x,0)) - M_0)v(x,0)\, dx = 0 \quad \forall v \in L^\infty(\Omega)\,,$$

i.e. $M(u(x,0)) = M_0$ in $L^1(\Omega)$.

Amongst others, the following results are obtained in Alt and Luckhaus [4]. Via a priori estimates with respect to certain energy integrals involving the Legendre transform of the primitive of M (see [34, pp. 16–20] and [72]), the existence of a weak solution in the sense given above is proved. Moreover, using backward Euler time-discretization, Galerkin approximations converge strongly to a weak solution in the topology given by these energy integrals. In general, $M(u)_t$ is not a function. However, if M and kr are both Lipschitz continuous and the data are sufficiently regular, i.e. if $u^D \in H^1(0,T;H^1(\Omega))$ ($\hookrightarrow C([0,T];H^1(\Omega))$!) and $M_0 = M(u_0)$ for a $u_0 \in u^D(\,\cdot\,,0) + H^1_{\gamma_D}(\Omega)$, then there is a weak solution u with $u \in L^2(\Omega \times (0,T))$. In fact, with these assumptions the solution constructed by the Galerkin approximations has this property. Finally, using a maximum principle, the authors prove the uniqueness of such a solution assuming these regularity conditions.

Observe here, again, that the latter assumptions are not satisfied for the parameter functions according to Brooks–Corey where M is not Lipschitz continuous. Furthermore, it is unclear whether the existence result holds in this case since $M : (u_c, \infty) \to \mathbb{R}$ is not a function on the whole real line and we have the additional obstacle condition $u > u_c$.

As regards these objections, we refer to the paper of Alt, Luckhaus and Visintin [5] in which the results from Alt and Luckhaus [4] have been further generalized, first to the situation of Brooks–Corey parameter functions and secondly to boundary conditions of Signorini's type. More concretely, an initial boundary value problem for the Kirchhoff–transformed Richards equation (1.6.3)–(1.6.6) with nonnegative $u_D(t)$ and additional Signorini-type boundary conditions

$$u \;\leq\; 0\,, \quad \mathbf{v}\cdot\mathbf{n} \;\geq\; 0\,, \quad u\cdot(\mathbf{v}\cdot\mathbf{n}) = 0 \qquad \text{on } \gamma_S \times (0,T)$$

with $\mathbf{v} = -(\nabla u - kr(M(u))e_z)$ is considered. The functions $M : [u_c, \infty) \to \mathbb{R}$ as well as $kr : M([u_c,\infty)) \to \mathbb{R}$ are supposed to be continuous and monotonically increasing. A weak formulation (P) of this problem is given in terms of a variational inequality that (in contrast to (1.6.7) and (1.6.8)) involves dual convex functions (see [34, pp. 16–20]). Here, the data need to satisfy $u^D \in H^1(Q) \cap C(0,T;H^1(\Omega))$ and $M_0 \in L^\infty(\Omega)$. The solution u is required to be in the convex set $K = \{v \in L^2(0,T;H^1(\Omega)) : u = u_D(t) \text{ on } \gamma_D \cap (0,T)\}$ with $M(u) \in L^\infty(Q) \cap H^1(0,T;H^1_{\gamma_D}(\Omega)')$.

The main result in Alt, Luckhaus and Visintin [5] is: There is at least one solution to the problem (P). Furthermore, (P) is considered in the limit case where M degenerates into a step function on the whole real line or, more precisely, to a maximal monotone multifunction. Here, an existence result for a very weak notion of a solution is established which is proved to have a physical meaning in one space dimension.

Finally, we point out that in the papers of Otto [72] and [73] L^1-contraction and uniqueness of the solutions in the setting of [4] and [5], respectively, is proved without assuming further regularity of the parameter functions or assuming that $M(u)_t$ is a function.

The papers [5] and [73] seem to provide the most general analytical results obtained so far on weak solutions to initial boundary value problems for the Richards equation with Signorini-type boundary conditions that also cover the case of Brooks–Corey parameter functions.

However, it should be emphasized that in order to obtain a semilinear, uniformly elliptic operator in the main part of the spatial derivative, the whole analysis in [5] and [73] is carried out *after* having applied the Kirchhoff transformation to the Richards equation. That includes the Galerkin approximations to the solution of the transformed equation. We point that out since we will pursue the same approach for our space discretization of the Richards equation in Section 2.5, which is applied to the Kirchhoff–transformed version only.

Unfortunately, it is not possible to treat heterogeneous versions of the Richards equation with this approach. In general, the hydraulic conductivity $K_c(x,\theta)$ can not be decomposed multiplicatively in a space-dependent function $K_h(x)$ and a factor depending only on the saturation as in (1.2.6). More concretely, in the Brooks–Corey permeability function

$$p \mapsto kr(\theta(p)) = \left[\frac{p}{p_b}\right]^{-\lambda e(\lambda)}$$

given in (1.2.11), the soil parameters p_b and λ can be space-dependent if the soil is not homogeneous. Now, as noted in Remark 1.3.1, the Kirchhoff transformation (1.3.1) applied to space-dependent relative permeabilities kr does not provide semilinear transformed equations in general.

In Chapter 3 we will investigate analytically heterogeneous boundary value problems for the nondegenerate time-discretized Richards equation involving discontinuous soil parameters. To our knowledge, so far no analytical results on (initial) boundary value problems for the Richards equation in a heterogeneous case have been presented in the literature.

Chapter 2

Numerical treatment of the Richards equation without gravity in homogeneous soil

2.1 Introduction

In this chapter we introduce our approach for the numerical treatment of the Richards equation in homogeneous soil which was introduced in Chapter 1. Our ansatz aims at separating the difficulties contained in the structure of the Richards equation and treating them independently in different steps. First, the nonlinearity in the spatial derivative is addressed by the Kirchhoff transformation, which has been done in Section 1.3. Secondly, by a suitable time discretization in which the gravitational term in the equation is treated explicitly, the influence of this term onto the arising spatial problems is minor and easily dealt with. Thirdly, the difficulty coming from the convective part due to gravity can be addressed by an upwind technique, independently from solving the spatial problems. This will be done in Chapter 4. The spatial problems arising after the time discretization and their solution are the main topic of this section. They will be treated in full generality for the Richards equation with gravity. Nevertheless, the performance of the solution method for the spatial problems can already be demonstrated for the Richards equation without gravity.

We start the presentation in Section 2.2 by giving some overview of the numerics that has been done so far for the Richards equation. This will contrast with our approach which is presented in the following sections beginning with our implicit–explicit time discretization of the Kirchhoff–transformed Richards equation in Section 2.3. With this discretization each of the arising spatial problems is a variational inequality in a convex subset of a Sobolev space. Equivalently, such a variational inequality can be regarded as a convex minimization problem for which we give analytical results including existence and uniqueness of a solution in that same Section 2.3. A generalization and reformulations of this continuous problem are discussed in terms of variational inclusions and further variational inequalities in Section 2.4. Equipped with these results, we introduce our finite element discretization of the continuous problem in Section 2.5, in which we also prove convergence of the finite dimensional solution to the continuous solution. The finite dimensional spatial problem is treated, without using further regularization, by convex minimization, the basis of which is a nonlinear Gauss–Seidel method presented in Section 2.6. This method serves as a smoother for the monotone multigrid that provides a nonlinear solver for the discrete problem and that will be described in Section 2.7. Finally, in Section 2.8 we give numerical results for the solution of the Richards equation without gravity which demonstrate the efficiency and the robustness of our solver for the spatial problem.

2.2 Previous numerical approaches to the Richards equation

A lot of work has been done in recent years on the numerical analysis and the numerical solution of the Richards equation. In the literature dealing with this topic the time discretization is mostly based on the full backward Euler method. This avoids time step restrictions regarding the stability of the numerical scheme which arise due to the convective (gravitational) term in the equation.

As far as the space discretization is concerned, intensive research has been done on the finite volume method in Fuhrmann [40], Fuhrmann and Langmach [41], Eymard et al. [37], [36] and on the mixed finite element method in Soucie [88], Schneid et al. [84], Radu et al. [76]. In Bastian et al. [10] discontinuous Galerkin schemes are used before the time discretization is carried out such that ordinary differential equations are obtained. A finite element method applied to the Richards equation in the physical form is investigated in Forsyth and Kropinski [38] concerning monotonicity of the discretization. Earlier works by Hornung [48], [49] contain a longitudinal line method and a finite element method, respectively, for the Richards equation without gravity, which can be regarded as a degenerate Fokker–Planck equation. A mixed finite element method for an equation similar to the Richards equation, in which the physical pressure can be written as a function of the saturation (in contrast to (1.2.9)), is presented in Arbogast et al. [7].

The numerical results in Fuhrmann [39] and [40] are obtained by the application of an algebraic Newton multigrid applied to a nonlinear finite element and finite volume scheme, respectively, for the Richards equation in its physical form (1.5.1). Stability and existence of solutions for the latter is proved in Fuhrmann and Langmach [41] under smoothness conditions on the parameter functions $p \mapsto \theta(p)$ and $\theta \mapsto kr(\theta)$. Of course, smoothness together with Lipschitz continuity of the parameter functions is also required for a successful application of Newton's method. With Regard to these regularity conditions, we note that in Fuhrmann [39] and [40] as well as in Wagner et al. [96] and in Knabner and Schneid [53], the parameter functions are chosen according to van Genuchten [91].

An advantage of the finite volume scheme for the physical Richards equation is that it provides discrete mass conservation which reflects the mass conservation in the continuous case. In addition, this approach is flexible with respect to variations of the conductivity and the soil parameters in the relative permeability function kr and can therefore be successfully applied in case of heterogeneous soil, too. This is done in Fuhrmann [39] and Fuhrmann and Langmach [41]. On the other hand, the performance of the Newton method is not robust in the case of deteriorating slopes of the parameter functions.

High flexibility and lack of robustness also apply to the finite volume approach persued in Wagner et al. [96] for a regularized and a transformed version of the physical Richards equation. Here, a Schur complement multigrid is used as a solver for the linear systems obtained by the Newton method. This was particularly designed for problems with strong variations of the hydraulic conductivity $K_h(x)$.

Convergence of finite volume schemes is proved in Eymard et al. [37] and [36]. In [36] the Kirchhoff–transformed Richards equation (1.5.2) is treated, however, assuming M to be Lipschitz continuous. In [37] the physical Richards equation (1.5.1) with the piezometric head $p/(\varrho g) - z$ as the unknown is considered, but here, the nondegeneracy condition $kr \geq c > 0$ is required.

Also assuming nondegeneracy $kr \geq c > 0$, convergence of a linearization scheme for the Richards equation is proved in Slodička [87]. The linearization is applied to the implicitly time-discretized Richards equation in physical form without a discretization in space.

A priori error estimates for mixed finite element discretizations of the Richards equation are proved in Soucie [88], Schneid et al. [84] and Radu et al. [76]. It seems that error estimates for the Richards equation with the physical pressure as the unknown can only be found in Soucie [88]. Here, however, gravity is ignored and a model with slight compressibility of the water is used. In addition, except for the derivation of estimates for the H^{-1}-norm, boundedness of $\theta(p)_t$ or ellipticity $kr \geq c > 0$ is required.

In Schneid et al. [84] and Radu et al. [76] the Kirchhoff–transformed Richards equation is considered in a time-integrated form. Error estimates are proved for the semidiscrete and the discrete scheme. But to obtain this, M needs to be continuously differentiable and Lipschitz continuous. As in Fuhrmann [40] and Wagner et al. [96], the resulting equations are solved with Newton's method (see also Knabner and Schneid [53]).

Despite the many results in the literature presented so far on the numerical treatment of the Richards equation, no robust solver of the spatial problems occurring after time discretization seems to be at hand. In the next section, we present a time discretization of the Kirchhoff–transformed Richards equation that — in case of a homogeneous soil — can be treated by convex minimization rather than regularization. The resulting spatial problems can then be solved robustly by a monotone multigrid method.

2.3 Time discretization and convex minimization

In this section, we present our special time discretization of the Richards equation and our approach to solve the corresponding spatial problems using convex analysis. The basis for our numerical solution of the Richards equation is the variational inequality (1.5.21) which for any $t \in (0,T]$ is a weak formulation of the Signorini-type problem (1.5.4)–(1.5.7) for the Kirchhoff–transformed version of the equation with the Brooks–Corey model. For reasons of notation, we set $f_N(t) = 0 \ \forall t \in (0,T]$ without loss of generality and we obtain the variational formulation to find $u(t) \in \mathcal{K}(t)$ with

$$\int_\Omega M(u(t))_t\,(v-u(t))\,dx + \int_\Omega \nabla u(t)\nabla(v-u(t))\,dx \geq \int_\Omega kr(M(u(t)))e_z\nabla(v-u(t))\,dx \quad \forall v \in \mathcal{K}(t) \tag{2.3.1}$$

in the convex set $\mathcal{K}(t)$ introduced in (1.5.19).

Remark 2.3.1. We point out at this stage that our further treatment of the variational inequality (2.3.1) does not depend on the special form of the functions M and kr in the Brooks–Corey model but on their basic properties. In what is to come in this section and in the following sections, we will make clear where these properties are required and often keep the presentation as general as possible. Therefore, the results can also be applied to the Richards equation with other parameter functions like in the van Genuchten model [91]. (Recall that an advantage of the Brooks–Corey model is that the Kirchhoff transformation and the transformed parameter functions can be given explicitly, see Section 1.3.) Moreover, our results are open to generalizations which will be essential in the treatment of the Robin method for the Richards equation that we address in Section 3.4.

2.3.1 Implicit–explicit time discretization

With regard to our aim to apply convex minimization rather than regularization to the spatial problems arising from the discretization, we choose our time discretization to be implicit in the main part of the equation and explicit in the convective part coming from the gravitation. This is already indicated in how the variational inequality (2.3.1) is formulated. For a partition $0 = t_0 < t_1 < \ldots < t_N = T$ of $[0,T]$ and $\tau_n := t_n - t_{n-1}$ for $n \in \{1,\ldots,N\}$, we discretize $M(u(t_n))_t$ in (2.3.1) by the backward Euler differential quotient

$$\frac{M(u(t_n)) - M(u(t_{n-1}))}{\tau_n},$$

setting $\nabla u(t)$ on the left hand side implicitly $\nabla u(t_n)$ and $kr(M(u(t)))$ on the right hand side explicitly $kr(M(u(t_{n-1})))$. Thus, with a given $u(0) = u^0$ and denoting the corresponding approximation of $u(t_n)$ by u^n for $n \in \{1, \dots, N\}$, the discrete version of (2.3.1) reads: Find $u^n \in \mathcal{K}(t_n)$ with

$$\int_\Omega M(u^n)\,(v - u^n)\,dx + \tau_n \int_\Omega \nabla u^n \nabla (v - u^n)\,dx \geq \tag{2.3.2}$$
$$\int_\Omega M(u^{n-1})\,(v - u^n)\,dx + \tau_n \int_\Omega kr(M(u^{n-1}))e_z \nabla (v - u^n)\,dx \quad \forall v \in \mathcal{K}(t_n)\,.$$

In order to proceed from here, we need to take a close look at the structure of this variational inequality. Since in the following we only consider spatial problems, we introduce some abbreviations. Given some $n \in \{1, \dots, N\}$, we set $\mathcal{K} := \mathcal{K}(t_n) \subset H^1(\Omega)$ and also $\gamma_D := \gamma_D(t_n)$, $\gamma_S := \gamma_S(t_n)$ and $\gamma_N := \gamma_N(t_n)$ as well as $u_D := u_D(t_n)$, i.e. we can write

$$\mathcal{K} = \{v \in H^1(\Omega) : v \geq u_c \,\wedge\, tr_{\gamma_D} v = u_D \,\wedge\, tr_{\gamma_S} v \leq 0\}\,. \tag{2.3.3}$$

Recall from Proposition 1.5.5 that, with an appropriate choice of $u_D \in H^{1/2}(\gamma_D)$ compatible to the Signorini-type boundary conditions and the constraint u_c, the set $\mathcal{K}$ is a nonempty, closed and convex subset of $H^1(\Omega)$.

We define the symmetric bilinear form $a(\cdot,\cdot)$ on $H^1(\Omega)$ by

$$a(v, w) := \tau_n \int_\Omega \nabla v \nabla w\,dx \quad \forall v, w \in H^1(\Omega)\,. \tag{2.3.4}$$

Of course, $a(\cdot,\cdot)$ is continuous on $H^1(\Omega)$, i.e. there is a $C > 0$ such that

$$a(v, w) \leq C\|v\|_1\|w\|_1 \quad \forall v, w \in H^1(\Omega)\,. \tag{2.3.5}$$

It is well known that $a(\cdot,\cdot)$ is coercive on the subspace $H^1_{\gamma_D}(\Omega)$ for any $\gamma_D \subset \partial\Omega$ with a positive Hausdorff measure (see Theorem A.2.5), which we assume here, i.e. there is a $c > 0$ such that

$$a(v, v) \geq c\|v\|_1^2 \quad \forall v \in H^1_{\gamma_D}(\Omega)\,. \tag{2.3.6}$$

This inequality leads to the following property of $a(\cdot,\cdot)$ on an affine space P_D in $H^1(\Omega)$ containing $\mathcal{K}$ which can be regarded as a more general notion of coercivity, see Proposition 2.3.15.

Lemma 2.3.2. *Let $w \in H^1(\Omega)$ with $tr_{\gamma_D} w = u_D$ and $P_D := w + H^1_{\gamma_D}(\Omega)$. Then we have $\mathcal{K} \subset P_D$, and with the constant c from (2.3.6) and certain positive c_1 and c_2 we obtain*

$$a(v, v) \geq c\|v\|_1^2 - c_1\|v\|_1 - c_2 \quad \forall v \in P_D\,.$$

Proof. $\mathcal{K} \subset P_D$ is obvious from $\mathcal{K} - w \subset H^1_{\gamma_D}(\Omega)$. Now, setting $v = w + \tilde{v} \in P_D$ with $\tilde{v} \in H^1_{\gamma_D}(\Omega)$ and using (2.3.5) and (2.3.6), we obtain

$$a(v, v) = a(\tilde{v}, \tilde{v}) + 2a(w, \tilde{v}) + a(w, w) \geq c\|\tilde{v}\|_1^2 - 2C\|w\|_1\|\tilde{v}\|_1 - C\|w\|_1^2\,,$$

which can be further estimated from below by

$$c\|v\|_1^2 - 2(C + c)\|w\|_1\|v\|_1 - (3C + c)\|w\|_1^2\,. \qquad \square$$

If $M : [u_c, \infty) \to \mathbb{R}$ and $kr : M([u_c, \infty)) \to \mathbb{R}$ are monotonically increasing and bounded functions, the linear form ℓ on $\mathcal{K} \subset H^1(\Omega)$ defined by

$$\ell(v) := \int_\Omega M(u^{n-1})\, v\, dx + \tau_n \int_\Omega kr(M(u^{n-1}))e_z \nabla v\, dx \quad \forall v \in H^1(\Omega) \qquad (2.3.7)$$

is continuous, i.e. an element of $H^1(\Omega)'$. Now, replacing u^n by u, we can write (2.3.2) more compactly as the variational inequality

$$u \in \mathcal{K} : \quad \int_\Omega M(u)(v-u)\, dx + a(u, v-u) - \ell(v-u) \geq 0 \quad \forall v \in \mathcal{K}\,. \qquad (2.3.8)$$

2.3.2 The convex function Φ and its properties

With regard to the first integral on the left hand side of (2.3.8), we define the function $\Phi : [u_c, \infty) \to \mathbb{R}$ as the integral

$$\Phi(z) := \int_0^z M(s)\, ds \quad \forall z \in [u_c, \infty) \qquad (2.3.9)$$

which turns out to be a convex function if M is monotonically increasing. Concerning the Richards equation with parameter functions according to Brooks and Corey, we have calculated Φ in (1.3.27). In general we need to assume $u_c < 0$ in the definition (2.3.9).

For what is to come we recall the definition of convex functionals and some of their properties (see [56, pp. 144–156] and [80, pp. 45–47] for details).

Definition 2.3.3. Let V be a real vector space and $K \subset V$ a convex set, i.e. for $y, z \in K$ and $\lambda \in (0,1)$ we have $(1-\lambda)y + \lambda z \in K$. $F : K \to \mathbb{R}$ is called a *convex functional* if

$$F((1-\lambda)y + \lambda z) \leq (1-\lambda)F(y) + \lambda F(z) \quad \forall y, z \in K\,.$$

If the inequality is strict in all cases where $y \neq z$, F is called *strictly convex*.

The next lemma points to the one-dimensional nature of convex functionals and is easy to prove.

Lemma 2.3.4. *Let V be a real vector space and $K \subset V$ a convex set. If the functional $F : K \to \mathbb{R}$ is convex and $u, v \in K$, then the real function $g : [0,1] \to \mathbb{R}$ defined by*

$$g(\lambda) = F(u + \lambda(v-u)) \quad \forall \lambda \in [0,1]$$

is also convex.

In the following we note a criterion for the convexity of real functions and a fundamental property of real convex functions. Both facts will be very helpful in the sequel.

Lemma 2.3.5. *Let $I \subset \mathbb{R}$ be an interval. Then for $f : I \to \mathbb{R}$ the following holds.*

a) f is convex if and only if the inequality

$$\frac{f(z) - f(z_1)}{z - z_1} \leq \frac{f(z_2) - f(z)}{z_2 - z}$$

holds for any $z_1, z, z_2 \in I$ with $z_1 < z < z_2$.

b) If f is convex, then for any $z \in I$ the difference quotient

$$\Delta_z(y) := \frac{f(y) - f(z)}{y - z}$$

is a monotonically increasing function of $y \in I \backslash \{z\}$.

These monotonicity properties of slopes of convex functions are well known and easy to derive from the definition.

Now we can relate basic properties of Φ in (2.3.9) to properties of M.

Lemma 2.3.6. *Let $M : [u_c, \infty) \to \mathbb{R}$ be monotonically increasing. Then Φ in (2.3.9) is convex. Φ is differentiable (from the right) in u_c, and, in addition, we have $\Phi'(u_c) = M(u_c)$ if M is continuous in u_c. Furthermore, Φ is differentiable in $z \in (u_c, \infty)$ if and only if M is continuous in z, which is true for all but countably many points, and in this case $\Phi'(z) = M(z)$ holds. If M is bounded, then Φ is Lipschitz continuous with Lipschitz constant $\|M\|_\infty$.*

Proof. Since M is monotonically increasing we can estimate

$$M(z_1) \leq \frac{\int_{z_1}^{z} M(s)\,ds}{z - z_1} \leq \lim_{y \uparrow z} M(y) \leq M(z) \leq \lim_{y \downarrow z} M(y) \leq \frac{\int_{z}^{z_2} M(s)\,ds}{z_2 - z} \leq M(z_2)$$

for $z_1, z, z_2 \in [u_c, \infty)$ with $z_1 < z < z_2$. Then, with

$$\frac{\Phi(z) - \Phi(z_1)}{z - z_1} = \frac{\int_{z_1}^{z} M(s)\,ds}{z - z_1} \quad \text{and} \quad \frac{\int_{z}^{z_2} M(s)\,ds}{z_2 - z} = \frac{\Phi(z_2) - \Phi(z)}{z_2 - z}$$

we immediately obtain the convexity of Φ from Lemma 2.3.5 a).

The same estimates show the one-sided differentiability of Φ on $[u_c, \infty)$. Furthermore, we obtain $\Phi'(z) = M(z)$ if and only if M is continuous in z, which is the case in all but countably many points $z \in [u_c, \infty)$ due to the monotonicity of M (see [56, p. 103]).

The assertion about the Lipschitz continuity of Φ follows in the same way. □

We remark that, except for an additive constant, every convex function on an open interval has a representation as in (2.3.9) (see [56, p. 156] and [98, p. 488]).

Observe that Lemma 2.3.6 then states that differentiable convex functions are immediately continuously differentiable. We emphasize that the special situation in the endpoint u_c is crucial for our further analysis and the main result of this section, Theorem 2.3.16. We refer to the Sections 1.4 and 2.4 for the situation of a discontinuity of M in u_c with regard to the Richards equation.

2.3.3 The convex functional ϕ and its properties

Now the function Φ gives rise to a functional $\phi : \mathcal{K} \to \mathbb{R}$ by

$$\phi(v) := \int_\Omega \Phi(v(x))\, dx \quad \forall v \in \mathcal{K} \tag{2.3.10}$$

which is well-defined if Φ is Lipschitz continuous. Convex functionals arising from this definition are investigated in a more general setting in Kornhuber [59]. Here we note their basic properties in our case.

Proposition 2.3.7. *If Φ is a convex function, then $\phi : \mathcal{K} \to \mathbb{R}$ is a convex functional. In addition, if Φ is Lipschitz continuous, then ϕ is Lipschitz continuous, and, with a $C > 0$, satisfies*

$$|\phi(v)| \leq C\|v\|_1 \quad \forall v \in \mathcal{K}\,. \tag{2.3.11}$$

Proof. The convexity of ϕ follows directly from the convexity of the function Φ. Also, Lipschitz continuity of ϕ follows directly from the Lipschitz continuity of Φ since the Lipschitz constant L of Φ and the Cauchy–Schwarz inequality in $L^2(\Omega)$ provide $C = L\,\|1\|_{L^2(\Omega)}$ with

$$|\phi(v) - \phi(w)| \leq \int_\Omega |\Phi(v(x)) - \Phi(w(x))|\, dx \leq C\|v - w\|_{L^2(\Omega)} \leq C\|v - w\|_1$$

for all $v, w \in \mathcal{K}$. In particular, this leads to (2.3.11) since $w = 0 \in \mathcal{K}$ and $\phi(0) = 0$. □

In order to see how ϕ is related to the first integral on the left hand side of (2.3.8) we recall definitions of different notions of derivatives (consult e.g. [34, p. 23] and [98, p. 113]). We use the duality brackets $\langle \cdot, \cdot \rangle$ for the duality (V', V).

Definition 2.3.8. Let $F : S \to \mathbb{R}$ on a subset $S \subset V$ of a normed space V, $u \in S$ and $v \in V$.

a) If there is an $\varepsilon > 0$ such that $u + \lambda v \in S$ for all $\lambda \in [0, \varepsilon]$, we call

$$\partial_v F(u) := \lim_{\lambda \downarrow 0} \frac{F(u + \lambda v) - F(u)}{\lambda} \tag{2.3.12}$$

the *directional derivative* of F at u in the direction of v if the one-sided limit in (2.3.12) exists.

b) If, in addition to a), there is a $u' \in V'$ such that

$$\partial_v F(u) = \langle u', v \rangle \quad \forall v \in V\,,$$

then F is called *Gâteaux–differentiable* at the point u with the *Gâteaux–derivative* $F'(u) := u'$.

c) If u is an interior point of S and, in addition to b), the convergence in (2.3.12) is uniform with respect to the elements in the unit ball of V, then F is said to be *Fréchet–differentiable* and $F'(u)$ is called the *Fréchet–derivative* of F at u.

Proposition 2.3.9. *Let $\Phi : [u_c, \infty) \to \mathbb{R}$ be convex and differentiable. Then, for any $u, v \in \mathcal{K}$ the directional derivative $\partial_{v-u}\phi(u)$ exists and can be written as*

$$\partial_{v-u}\phi(u) = \int_\Omega \Phi'(u(x))(v(x) - u(x))\,dx = \int_\Omega M(u(x))(v(x) - u(x))\,dx\,. \tag{2.3.13}$$

Proof. If $u, v \in \mathcal{K}$, we obviously have $u + \lambda(v - u) \in \mathcal{K}$ for $\lambda \in [0,1]$ since $\mathcal{K}$ is convex. We set $w := v - u$ and consider the difference quotient

$$\frac{\phi(u + \lambda w) - \phi(u)}{\lambda} = \int_\Omega \frac{\Phi(u(x) + \lambda w(x)) - \Phi(u(x))}{\lambda}\,dx \tag{2.3.14}$$

as $\lambda \downarrow 0$. In light of Lemma 2.3.5 b) we obtain

$$\frac{\Phi(u(x) + \lambda w(x)) - \Phi(u(x))}{\lambda} \le \frac{\Phi(u(x) + w(x)) - \Phi(u(x))}{1} =: G(x)$$

for $w(x) \ge 0$ and $\lambda \in (0,1]$ and

$$H(x) := \frac{\Phi(u(x) - w(x)) - \Phi(u(x))}{1} \le \frac{\Phi(u(x) + \lambda w(x)) - \Phi(u(x))}{\lambda}$$

for $w(x) \ge 0$ and $\lambda \in (0,1]$. Altogether, the integrands in (2.3.14) are bounded by the integrable function $\max(|H(\cdot)|, |G(\cdot)|)$ on Ω independently of $\lambda \in (0,1]$.

Due to the differentiability of Φ, the pointwise values of the integrands

$$\frac{\Phi(u(x) + \lambda w(x)) - \Phi(u(x))}{\lambda}$$

converge to $\Phi'(u(x))w(x)$ almost everywhere in Ω as $\lambda \downarrow 0$, even as a monotonically increasing sequence for $w(x) < 0$ and as a monotonically decreasing sequence for $w(x) \ge 0$ due to Lemma 2.3.5 b).

Consequently, by the theorems of Lebesgue or of Beppo Levi (see [98, p. 492]), we get the convergence of the integral in (2.3.14) and the assertion (2.3.13) with $\Phi' = M$ from Lemma 2.3.6. □

Remark 2.3.10. Note that the proof only uses the convexity of $\Phi : [u_c, \infty) \to \mathbb{R}$ besides its differentiability. Of course, applying the mean value theorem, boundedness of $\Phi' = M$ also leads to (2.3.13) without assuming convexity.

Observe that if $M : [u_c, \infty) \to \mathbb{R}$ is monotonically increasing and bounded and $u \in \mathcal{K}$, then the map

$$w \mapsto \int_\Omega M(u(x))w(x)\,dx \quad \forall w \in H^1(\Omega)$$

defines a bounded linear functional on $H^1(\Omega)$, i.e. we obtain Gâteaux–differentiability of ϕ in $H^1(\Omega)$.

We remark that one obtains analogous results for ϕ defined on the whole space $H^1(\Omega)$ in (2.3.10) if M is defined on the whole real line or if $M : [u_c, \infty) \to \mathbb{R}$ is extended by $M(u_c)$ on $(-\infty, u_c)$. In this case, if M is uniformly continuous and bounded, one can show that the Gâteaux–derivative mapping $\phi' : H^1(\Omega) \to H^1(\Omega)'$ is continuous (with respect to $u \in H^1(\Omega)$). It is even Hölder continuous with a Hölder exponent $\alpha \in (0, 1]$ if M is Hölder continuous with the same exponent. For these two results consult Propositions 2.5.11 and 2.5.12. Note that in the Brooks–Corey case (1.3.8), M is Hölder continuous. Now, continuity of $\phi' : H^1(\Omega) \to H^1(\Omega)'$ guarantees the Fréchet–differentiability of ϕ on $H^1(\Omega)$, see [98, p. 120].

2.3.4 From the variational inequality to a convex minimization problem with a unique solution

With Proposition 2.3.9 and our notation in (2.3.4) and (2.3.7), assuming continuity of M, the variational inequality (2.3.8) for the solution $u \in \mathcal{K}$ reads

$$\partial_{v-u}\phi(u) + a(u, v - u) - \ell(v - u) \geq 0 \quad \forall v \in \mathcal{K}\,. \tag{2.3.15}$$

Now, it is well known that the quadratic functional $\mathcal{J} : H^1(\Omega) \to \mathbb{R}$ defined by

$$\mathcal{J}(v) := \frac{1}{2}a(v, v) - \ell(v) \quad \forall v \in H^1(\Omega) \tag{2.3.16}$$

is strictly convex (see [34, p. 36]) and continuous if kr and M are monotonically increasing and bounded (see (2.3.5) and (2.3.7)), with the bilinear form in (2.3.4) which is coercive on $H^1_{\gamma_D}(\Omega)$ and satisfies Lemma 2.3.2. Furthermore, its Fréchet–differentiability in $u \in H^1(\Omega)$ with the derivative

$$\mathcal{J}'(u)(v) = \partial_v \mathcal{J}(u) = a(u, v) - \ell(v) \quad \forall v \in H^1(\Omega) \tag{2.3.17}$$

is easy to see, too (cf. [60, p. 4]).

Consequently, the functional $F : \mathcal{K} \to \mathbb{R}$ defined by

$$F(v) := \phi(v) + \mathcal{J}(v) \quad \forall v \in \mathcal{K} \tag{2.3.18}$$

is strictly convex with existing $\partial_{v-u}F(u)$ for any $u, v \in \mathcal{K}$, and (2.3.8) has the following form: Find $u \in \mathcal{K}$ such that

$$\partial_{v-u}F(u) \geq 0 \quad \forall v \in \mathcal{K} . \tag{2.3.19}$$

Now, this variational inequality can be regarded as a convex minimization problem. For this equivalence we refer to [34, p. 37], however, we state it in a more general form using only directional derivatives.

Proposition 2.3.11. *Let V be a real vector space, $K \subset V$ a convex set and $F : K \to \mathbb{R}$ a convex functional whose directional derivative $\partial_{v-u}F(u)$ exists for all $u, v \in K$. Then*

$$u \in K : \quad \partial_{v-u}F(u) \geq 0 \quad \forall v \in K \tag{2.3.20}$$

is equivalent to

$$u \in K : \quad F(u) \leq F(v) \quad \forall v \in K . \tag{2.3.21}$$

Proof. Assuming (2.3.20), for any $v \in K$ and $\lambda \in (0, 1]$ we can estimate

$$F(v) - F(u) = \frac{F(u + (v-u)) - F(u)}{1} \geq \frac{F(u + \lambda(v-u)) - F(u)}{\lambda}$$

if we consider Lemma 2.3.4 and Lemma 2.3.5 b). Passing to the limit $\lambda \downarrow 0$, we obtain

$$F(v) - F(u) \geq \partial_{v-u}F(u) \geq 0 .$$

Conversely, if (2.3.21) is satisfied, for any $\lambda \in (0, 1]$ we have

$$F(u + \lambda(v-u)) \geq F(u) \quad \forall v \in K$$

and consequently

$$\frac{F(u + \lambda(v-u)) - F(u)}{\lambda} \geq 0 \quad \forall v \in K$$

which leads to (2.3.20) for $\lambda \downarrow 0$. □

Remark 2.3.12. In general the inequality (2.3.20) does not only account for the fact that u could be an element of the boundary ∂K. Strict inequalities can occur in such a case, but they can also occur for an inner point u of K, e.g. in case of the absolute value function $F : x \mapsto |x|$ on $K = [-1, 1]$. If, however, u is an inner point of K (which is always true if $K = V$ for example) and F is Gâteaux–differentiable in u, then (2.3.20) is always an equality and is in fact equivalent to

$$\partial_v F(u) = 0 \quad \forall v \in V \quad \text{or} \quad F'(u) = 0 . \tag{2.3.22}$$

Let $B_\varepsilon(u) = \{v \in V : \|v - u\| < \varepsilon\} \subset \operatorname{int} K$ with a certain $\varepsilon > 0$. Then for any $\tilde{v} \in B_\varepsilon(u)$ we can insert $v = u + \tilde{v} \in K$ in (2.3.20) to obtain $\partial_w F(u) \geq 0$ for $w = \pm\tilde{v}$ and conclude (2.3.22).

In the following, we cite the well-known general existence and uniqueness result for convex minimization problems (2.3.21) in reflexive Banach spaces, the proof of which can be found in Ekeland and Temam [34, p. 35]. We recall two important terms (cf. [34, pp. 8–10, 34]).

Definition 2.3.13. Let V be a topological space and $K \subset V$ nonempty, closed and convex. A functional $F : K \to \mathbb{R} \cup \{+\infty\}$ is called *lower semicontinuous* if $\liminf_{w \to v} F(w) \geq F(v)$ holds for all $v \in K$ (with $w \in K$). A convex functional $F : K \to \mathbb{R} \cup \{+\infty\}$ is called *proper* if it is not identically $+\infty$. In this case $\operatorname{dom} F := \{v \in K : \ F(v) < \infty\}$ is called the *domain* of F.

Remark 2.3.14. As in [34] one often prefers to work with the canonical extension $\tilde{F} : V \to \mathbb{R} \cup \{+\infty\}$ of $F : K \to \mathbb{R} \cup \{+\infty\}$ which is given by $\tilde{F}(z) = +\infty$ for all $z \in V \backslash K$. Then, since K is nonempty, closed and convex, $\tilde{F}$ is lower semicontinuous and proper if and only if F is, and the minimization problem (2.3.21) for F on K is equivalent to the corresponding minimization problem for $\tilde{F}$ on V.

Proposition 2.3.15. *Let V be a reflexive Banach space, $K \subset V$ a nonempty, closed and convex subset of V and $F : K \to \mathbb{R} \cup \{+\infty\}$ a convex, lower semicontinuous and proper functional. Furthermore, let F be coercive, i.e. for any sequence $(u_n) \subset K$ with $\|u_n\| \to \infty$ we have $F(u_n) \to \infty$. Then the minimization problem (2.3.21) has a solution. It is unique if F is strictly convex.*

Now we can state the main result of this section.

Theorem 2.3.16. *Let $\mathcal{K} \subset H^1(\Omega)$, $a(\cdot,\cdot)$ and $\ell(\cdot)$ in (2.3.8) be defined as at the beginning of this section. If $M : [u_c, \infty) \to \mathbb{R}$ is monotonically increasing, bounded and continuous and $kr : M(\mathbb{R}) \to \mathbb{R}$ is monotonically increasing and bounded, then the variational inequality (2.3.8) has a unique solution. More specifically, it is equivalent to the minimization problem*

$$u \in \mathcal{K} : \quad \mathcal{J}(u) + \phi(u) \leq \mathcal{J}(v) + \phi(v) \quad \forall v \in \mathcal{K} \tag{2.3.23}$$

with $\mathcal{J}$ and ϕ as defined in (2.3.16) and in (2.3.10), respectively.

Proof. We only need to check that $F = \mathcal{J} + \phi$ on $\mathcal{K}$ satisfies the conditions required in Proposition 2.3.15. The Hilbert space $H^1(\Omega)$ is reflexive and $\mathcal{K} \subset H^1(\Omega)$ is nonempty, closed and convex. $F : \mathcal{K} \to \mathbb{R}$ is strictly convex since $\mathcal{J}$ is strictly convex and ϕ is convex, and F is proper. Furthermore, F is lower semicontinuous since it is continuous as $\mathcal{J}$ and ϕ are, see (2.3.16) and Proposition 2.3.7. This latter proposition and Proposition 2.3.2 provide the coerciveness of F by

$$|\mathcal{J}(v)+\phi(v)| \geq \frac{1}{2}a(v,v) - |\ell(v)| - |\phi(v)| \geq \frac{1}{2}c\|v\|_1^2 - (c_1 + \|\ell\| + C)\|v\|_1 - c_2 \to \infty$$

for $\|v\|_1 \to \infty$, $v \in \mathcal{K}$. □

Remark 2.3.17. Note that, to guarantee coerciveness of $a(\cdot,\cdot)$, we always assume γ_D to be a subset of $\partial\Omega$ with a positive Hausdorff measure. Also note that we can admit $f_N \neq 0$ and still obtain the result in Theorem 2.3.16 since f_N contributes to the linear functional $\ell(\cdot)$ only. By the trace theorem A.2.3, continuity of the corresponding expression given by the integral in (1.5.21) with respect to $v - u \in H^1_{\gamma_D}(\Omega)$ is given for any $f_N \in L^2(\gamma_N)$.

As a further generalization, we can easily extend our theory to space-dependent functions $n(\cdot)$ for the porosity and $K_h(\cdot)$ for the hydraulic conductivity as given in the original Richards equation (1.3.17) or (1.3.18). As long as $n(\cdot)$ is a nonnegative and bounded function, our results above can be carried over without further change if we replace $\Phi(v(x))$ by $n(x)\Phi(v(x))$ in (2.3.10) and $M(u(x))$ by $n(x)M(u(x))$ in Propositions 2.3.7 and 2.3.9 and in Remark 2.3.10. With regard to the hydraulic conductivity, it is clear that we need to impose boundedness of $K_h(\cdot)$ and $K_h(\cdot) \geq c$ for a $c > 0$ in order to preserve the continuity and the coerciveness of $a(\cdot,\cdot)$.

We point out that Theorem 2.3.16 also holds for monotonically increasing, bounded and continuous functions $M : \mathbb{R} \to \mathbb{R}$ defined on the whole real line. In fact, this leads to an easier problem (2.3.8), in which ϕ is defined and Gâteaux–differentiable on the whole space $H^1(\Omega)$. To reduce it further, assume that $\gamma_S = \emptyset$ and $u_D = 0$ on γ_D. Then (2.3.8) is equivalent to the variational equality

$$u \in H^1_{\gamma_D}(\Omega) : \quad \int_\Omega M(u)\, v \, dx + a(u,v) - \ell(v) = 0 \quad \forall v \in H^1_{\gamma_D}(\Omega) \tag{2.3.24}$$

according to Remark 2.3.22. For $u_D \neq 0$, too, we obtain a variational equality in $H^1_{\gamma_D}(\Omega)$ from (2.3.8) by replacing $u = w + \tilde{u}$ with $w \in H^1(\Omega)$ and $tr_{\gamma_D} w = u_D$ as well as $\tilde{u} \in H^1_{\gamma_D}(\Omega)$ in (2.3.24) (compare what follows (2.4.11) in Section 2.4). Recall from Section 1.4 that $M : \mathbb{R} \to \mathbb{R}$ occurs for the Richards equation if kr is replaced by a kr_α which is uniformly bounded away from 0 and, as a discontinuous function, in certain hydrologically reasonable limit cases.

Finally, we remark that boundedness of M is not necessary for Theorem 2.3.16 to hold. For example, one could replace boundedness of M by

$$\textit{Hölder continuity of } M \textit{ outside of an interval } [-R, R] \tag{2.3.25}$$

for an $R > 0$, leading to an affine estimate on the right hand side of (2.3.11) which is not hard to see (consult e.g. Kornhuber [59, pp. 22–26]). In the next section we give a generalization of Theorem 2.3.16 for $\mathcal{K} = H^1_{\gamma_D}(\Omega)$ in which we can even omit the requirement for M to be continuous, thus addressing limit cases for the Richards equation as discussed in Section 1.4. Note that continuity of M is only needed (in Proposition 2.3.9) for the equivalence of (2.3.8) and (2.3.23), and that monotonicity and boundedness (or (2.3.25) instead) of M is enough to ensure the coercivity of $\mathcal{J} + \phi$ and therefore guarantee unique solvability of (2.3.23).

2.4 Variational inclusions and further variational inequalities

The purpose of this section is first to generalize and secondly to reformulate (2.3.8) or equivalently (2.3.23) in terms of further variational inequalities and also in terms of variational inclusions. Some of these reformulations are of interest in themselves (e.g. leading to a straightforward proof that (2.4.11) is well-posed at the end of this section), and some of them will be helpful in the analysis of our space discretization which will be carried out in the next section. Besides, they will serve as a starting point for the numerical solution method that we apply to the spatially discretized problem in Section 2.6 and Section 2.7.

2.4.1 Subdifferentials and a variational inclusion for the limit cases

We proceed by giving a generalization of Theorem 2.3.16 to monotonically increasing but possibly unbounded and, moreover, discontinuous $M : [u_c, \infty) \to \mathbb{R}$ or $M : \mathbb{R} \to \mathbb{R}$ in case of $\mathcal{K} = H^1_{\gamma_D}(\Omega)$. Even though such a situation does not occur for the Richards equation with continuous parameter functions as introduced in Section 1.2, one might think of it as a limit case of the Brooks–Corey model (1.2.9) and (1.2.10) for deteriorating soil parameters p_b and λ where the parameter functions degenerate into step functions. A detailed discussion of the limit cases for the Brooks–Corey model and also of the nondegenerate case can be found in Section 1.4. Recall that in hydrological applications M might be close to a step function, see Figure 1.11. In the limit cases for the Brooks–Corey model (see (1.4.6) and (1.4.7) or (1.4.13) and (1.4.15), respectively), M *is* discontinuous in u_c (and constant otherwise). However, the exact value of M in u_c is "less important" than the argument u_c itself whose role is to just represent the unsaturated regime in these cases. This phenomenon is already reflected by (1.4.9) or (1.4.14), respectively, and a wider mathematical basis for it is developed further in the following.

Observe in the proof of Lemma 2.3.6 that for any point $z_0 \in (u_c, \infty)$ we still have

$$\Phi(z) - \Phi(z_0) \geq m_{z_0}(z - z_0) \quad \forall z \in \mathbb{R} \tag{2.4.1}$$

for all

$$m_{z_0} \in [\lim_{y \uparrow z_0} M(y), \lim_{y \downarrow z_0} M(y)] =: I_{z_0} \tag{2.4.2}$$

even if M is not continuous in z_0, and certainly with $I_{z_0} = \{\Phi'(z_0)\} = \{M(z_0)\}$ if M is continuous. More generally, the proof of Proposition 2.3.11 shows that for convex $F : \mathcal{K} \to \mathbb{R}$ we have

$$F(v) - F(v_0) \geq \partial_{v-v_0} F(v_0) \quad \forall v \in \mathcal{K} \tag{2.4.3}$$

wherever $\partial_{v-v_0} F(v_0)$ exists. This remains true for all $v \in H^1(\Omega)$ if we extend F according to Remark 2.3.14. The following definition contains the well-known generalization of this phenomenon.

Definition 2.4.1. Let V be a normed space, $F : V \to \mathbb{R} \cup \{+\infty\}$ a convex functional and $v_0 \in \operatorname{dom} F$. Any element $g \in V'$ with

$$F(v) - F(v_0) \geq \langle g, v - v_0 \rangle \quad \forall v \in V$$

is called a *subgradient* of F at v_0. The set $\partial F(v_0)$ of all subgradients is called the *subdifferential* of F at v_0 and $\operatorname{dom} \partial F := \{v \in \operatorname{dom} F : \partial F(v) \neq \emptyset\}$.

Obviously, the subdifferential ∂F is a multivalued function $\partial F : \operatorname{dom} \partial F \to 2^{V'}$. Now, let the scalar convex function Φ defined in (2.3.9) be canonically extended by $+\infty$ on $(-\infty, u_c)$ as in Remark 2.3.14. Furthermore, for $M : [u_c, \infty) \to \mathbb{R}$ as above and the definition in (2.4.2), we define the corresponding multivalued function $\tilde{M} : [u_c, \infty) \to 2^{\mathbb{R}}$ by

$$\tilde{M}(z_0) := I_{z_0} \;\; \forall z_0 \in (u_c, \infty) \quad \text{and} \quad \tilde{M}(u_c) := (-\infty, \lim_{u \downarrow u_c} M(u)] \,. \tag{2.4.4}$$

Then, in light of Lemma 2.3.6, we obtain $\partial \Phi = \tilde{M}$ and $\operatorname{dom} \partial \Phi = [u_c, \infty)$. This scalar multifunction is *maximal monotone* which means that if

$$(m_z - m_{z_0})(z - z_0) \geq 0 \quad \forall z \in \operatorname{dom} \partial \Phi \;\; \forall m_z \in \partial \Phi(z)$$

holds, then we have $z_0 \in \operatorname{dom} \partial \Phi$ and $m_{z_0} \in \partial \Phi(z_0)$ (see Kornhuber [59, p. 33]).

It is clear that for the definition of $\tilde{M}$ we only need M to be monotonically increasing. In addition, we point out that the above considerations can analogously be carried out for any function $M : \mathbb{R} \to \mathbb{R}$ defined on the whole real line instead of on $[u_c, \infty)$ as long as it is monotonically increasing.

In either case, however, we now want to restrict ourselves on monotonically increasing functions M for which

$$\Phi(v) \in L^2(\Omega) \quad \forall v \in H^1(\Omega)$$

holds and (2.3.23) is uniquely solvable for the corresponding convex function Φ in (2.3.9) with $\partial \Phi = \tilde{M}$. This is of course satisfied for any hydrologically interesting M (see Section 1.4), which is bounded, but also for certain unbounded M as in (2.3.25). With the notation

$$\partial \Phi(v_0) := \{w \in L^2(\Omega) : w(x) \in \partial \Phi(v_0(x)) \;\text{ a.e. on }\; \Omega\} \tag{2.4.5}$$

for a $v_0 \in H^1(\Omega)$ we write $(\partial \Phi(v_0), \cdot)_{L^2(\Omega)}$ to indicate the corresponding subset of all bounded linear functionals on $H^1(\Omega)$ arising, by application of the L^2-scalar product, from elements of the subdifferential $\tilde{M} = \partial \Phi$ of the scalar convex function Φ. Observe that for any $w \in \partial \Phi(v_0)$ we have

$$\int_\Omega w(x)(v(x) - v_0(x))\, dx \leq \int_\Omega \Phi(v(x)) - \Phi(v_0(x))\, dx \tag{2.4.6}$$

which provides the inclusion $(\partial \Phi(v_0), \cdot)_{L^2(\Omega)} \subset \partial \phi(v_0)$. In fact, $(\partial \Phi(v_0), \cdot)_{L^2(\Omega)}$ is the subset of all elements in $\partial \phi(v_0) \subset H^1(\Omega)'$ which are also functionals on $L^2(\Omega)$, see Barbu [9, pp. 61/62].

For the following very general existence result we assume $\gamma_S = \emptyset$ and homogeneous Dirichlet conditions on γ_D. Neumann conditions on γ_N can be chosen according to Remark 2.3.17. In order to motivate what is to come, we first assume that we have a monotonically increasing and continuous function $M : \mathbb{R} \to \mathbb{R}$ which is bounded or else satisfies property (2.3.25). Then we obtain $\mathcal{K} = H^1_{\gamma_D}(\Omega)$ and the variational equality (2.3.24) as an equivalent formulation of (2.3.8). Now, in the general case of monotonically increasing and possibly discontinuous $M : \mathbb{R} \to \mathbb{R}$ as in the limit (1.4.25) of the nondegenerate case (1.4.18) for the Richards equation, we replace M by the multifunction $\tilde{M} = \partial\Phi$. Then, while taking (2.4.5) into account, we can consider the *variational inclusion*

$$u \in H^1_{\gamma_D}(\Omega) : \quad 0 \in (\tilde{M}(u), v)_{L^2(\Omega)} + a(u, v) - \ell(v) \quad \forall v \in H^1_{\gamma_D}(\Omega) \qquad (2.4.7)$$

as a generalization of (2.3.24). More precisely, (2.4.7) has to be understood as the inclusion of the 0-functional on $H^1_{\gamma_D}(\Omega)$ in the subset

$$(\partial\Phi(u), \cdot)_{L^2(\Omega)} + a(u, \cdot) - \ell(\cdot)$$

of $H^1_{\gamma_D}(\Omega)'$. A solution u of (2.4.7) is therefore accompanied with an L^2-function $w_u \in \partial\Phi(u)$ such that

$$(w_u, v)_{L^2(\Omega)} + a(u, v) - \ell(v) = 0 \quad \forall v \in H^1_{\gamma_D}(\Omega) \qquad (2.4.8)$$

holds, and since $H^1_{\gamma_D}(\Omega)$ is dense in $L^2(\Omega)$, it is clear that w_u is uniquely defined for any solution u of (2.4.7). The existence of a solution is guaranteed in a very general setting, and the following theorem can be found in a more general form in Jerome [51, pp. 91–94].

Theorem 2.4.2. *For monotonically increasing $M : \mathbb{R} \to \mathbb{R}$ or $M : [u_c, \infty) \to \mathbb{R}$ for $u_c < 0$ with a coercive and continuous $a(\cdot, \cdot)$ and $\ell \in H^1_{\gamma_D}(\Omega)'$, the variational inclusion (2.4.7) has a solution.*

Remark 2.4.3. Note that the theorem is not restricted to functions M defined on the whole real line, in fact one can apply it to $M : I \to \mathbb{R}$ on any interval $I \subset \mathbb{R}$ with 0 in its interior. In particular, the range of the solution u is contained in this interval. Moreover, the translated $\tilde{M}(\cdot + u_0)$ is also an admissible maximal monotone multifunction if $u_0 > u_c$. Therefore, the theorem covers at least constant Dirichlet boundary conditions u_0 (see Remark 2.3.17). The question how far this result can be extended to more general boundary conditions shall not be further discussed here.

Since we are particularly interested in hydrologically relevant M and we want to have a unique solution of (2.3.23), we assume M to be monotonically increasing and bounded from now on and still assume Φ as given in (2.3.9). For discontinuous M we cannot relate the variational inequality (2.3.8) to the convex minimization problem (2.3.23) as done in Propositions 2.3.9 and 2.3.11. However, it will turn out later (see Proposition 2.4.8) that a variational inclusion (2.4.7) is related to the corresponding convex minimization problem (2.3.23) in the sense that every solution of (2.4.7) solves (2.3.23), too. But since (2.3.23) is uniquely solvable, we obtain the following

Proposition 2.4.4. *For monotonically increasing and bounded $M : \mathbb{R} \to \mathbb{R}$ or $M : [u_c, \infty) \to \mathbb{R}$ the solution of (2.4.7) given in Theorem 2.4.2 is unique.*

Observe that w_u in (2.4.8) is defined almost everywhere in the set

$$\Omega_c := \{x \in \Omega : u(x) = \tilde{u}_c \text{ a.e.}\} \tag{2.4.9}$$

for any discontinuity $\tilde{u}_c$ of M. Theorem 2.4.2 does not give any further information on the values of w_u in $\tilde{u}_c$, but since u is unique, the variational equality (2.3.24) cannot be solvable for a discontinuous M if Ω_c has a positive Lebesgue measure and the value $M(\tilde{u}_c)$ is incompatible to the values of w_u on Ω_c.

Remark 2.4.5. With regard to the hydrological limit case given by the step function $M_0 : [u_c, \infty) \to \mathbb{R}$ in (1.4.6) and (1.4.7) and the corresponding monotone multifunction in (1.4.9), we consider the constant function $M : u \mapsto \theta_M$ on $[u_c, \infty)$ as in the limit case (1.4.13) with some $u_c < 0$. In order to motivate with the help of Theorem 2.4.2 that both cases could be interpreted as a similar problem, we still assume $\gamma_S = \emptyset$ and $u_D = 0$. Note that for the discontinuous function M_0 we can neither show the equivalence of the corresponding variational inequality (2.3.8) and a convex minimization problem (2.3.23), nor do we obtain a variational equality in this case for these relaxed boundary conditions (see Remark 2.3.22). The latter is also true for the constant function M. But we can still consider the minimization problem (2.3.23) for M_0 which is, however, the same as the one for M and in this sense not reflecting the different variational inequalities (2.3.8) generated by M_0 and M. Since M is continuous, the solution

$$u \in \mathcal{K} = \{v \in H^1_{\gamma_D}(\Omega) : v \geq u_c \text{ a.e.}\}$$

of (2.3.23) solves (2.3.8) according to Theorem 2.3.16. In the variational inclusion (2.4.7), which u solves due to Theorem 2.4.2 and Proposition 2.4.8, the (constant) function $M(u)$ in (2.3.8) is replaced by a $w_u \in \tilde{M}(u)$, i.e. with $w_u \leq M(u)$ almost everywhere. If u is an inner point of $\mathcal{K}$, the variational inequality (2.3.8) is equivalent to the variational equality (2.3.24) due to Remark 2.3.22. Then, we conclude $M(u) = M_0(u) = w_u = \theta_M$ almost everywhere. However, we can have strict inequality in (2.3.8) if Ω_c in (2.4.9), with $u_c = \tilde{u}_c$, has a positive Lebesgue measure and we have a test function $v \in \mathcal{K}$ with $v(x) > u(x) = u_c$ almost everywhere on Ω_c. In such a case we obtain

$$(M(u_c) - w_u, v - u)_{L^2(\Omega_c)} > 0$$

if we subtract (2.4.8) from (2.3.8). Therefore, we necessarily have

$$w_u(x) < M(u_c) = \theta_M \quad \text{on } \Omega_c' \subset \Omega_c$$

with an Ω_c' of positive Lebesgue measure. Even though we do not know the values of w_u on Ω, we can interpret $w_u \in \tilde{M}(u) = \tilde{M}_0(u)$ as a generalized saturation which is the same for both limit cases M_0 and M and which constitutes an unsaturated region Ω_c'. (Note, however, that the same considerations apply in the regular case where $M : [u_c, \infty) \to \mathbb{R}$ is continuous and u_c is regarded as a singular generalized pressure associated to physical pressure $p = -\infty$.) We come back to this topic at the end of Section 2.5.

2.4.2 Reformulations of the convex minimization problem

The rest of this section is devoted to further reformulations of the minimization problem (2.3.23). For example, there is an equivalent variational inclusion formulation of (2.3.23) that will lead to Proposition 2.4.8 which was already addressed above. In order to establish these results we assume the conditions imposed on M and kr in Theorem 2.3.16 except for the continuity of M. Problems like (2.3.23) or variational versions like (2.3.8) are often considered in the form of minimization problems, variational inclusions or variational inequalities on a (reflexive) Banach space rather than on a closed and convex subset of a (reflexive) Banach space, see e.g. Ekeland and Temam [34, p. 34] or Kornhuber [59]. As indicated in Remark 2.3.14, this could be achieved by adding the characteristic functional $\chi_{\mathcal{K}}$ of the convex set $\mathcal{K} \subset H^1(\Omega)$ defined by

$$\chi_{\mathcal{K}} : x \mapsto \begin{cases} \;\;0 & \text{for } \; x \in \mathcal{K} \\ +\infty & \text{for } \; x \in H^1(\Omega) \backslash \mathcal{K} \end{cases}$$

to the functional ϕ on $H^1(\Omega)$. However, for transparency and further analysis (see e.g. Section 2.5), we treat the three constraints in $\mathcal{K}$ separately. Moreover, we deal with the Dirichlet boundary conditions differently but in a well-known way by introducing a translated problem.

As above and as done in Kornhuber [59], we consider the scalar convex function $\Phi : [u_c, \infty) \to \mathbb{R}$ to be extended by $+\infty$ on $(-\infty, u_c)$. This corresponds to extending $M : [u_c, \infty) \to \mathbb{R}$ to a maximal monotone multifunction $\tilde{M}$ by (2.4.4) and (2.4.2), such that $\partial\Phi = \tilde{M}$. Consequently, $\phi : \mathcal{K} \to \mathbb{R}$ naturally extends to a lower semicontinuous and proper convex functional $\phi : H^1(\Omega) \to \mathbb{R} \cup \{+\infty\}$ by definition (2.3.10). For the extended Φ we still use the same notation. As a consequence, with the definition

$$\mathcal{K}_b := \{v \in H^1(\Omega) : tr_{\gamma_D} v = u_D \;\wedge\; tr_{\gamma_S} v \leq 0\} \,,$$

the minimization problem (2.3.23) is equivalent to the minimization problem

$$u \in \mathcal{K}_b : \quad \mathcal{J}(u) + \phi(u) \leq \mathcal{J}(v) + \phi(v) \quad \forall v \in \mathcal{K}_b \tag{2.4.10}$$

in which the constraints constituting the nonempty, closed and convex set $\mathcal{K}_b \subset H^1(\Omega)$ only concern boundary conditions.

In order to obtain a variational inclusion formulation of (2.4.10), we also need to encode the boundary conditions in the convex functional and construct an equivalent minimization problem on some suitable Hilbert space rather than on a closed and convex subset of a Hilbert space. This is done by translation of the Dirichlet values and by the introduction of an additional functional referring to the boundary conditions of Signorini's type. Therefore, analogously to Lemma 2.3.2, we first choose a fixed

$$w \in H^1(\Omega) \quad \text{with} \quad tr_{\gamma_D} w = u_D \,, \tag{2.4.11}$$

set $u = w + \tilde{u}$ and $v = w + \tilde{v}$ in (2.4.10) and find (2.4.10) to be equivalent to

$$\tilde{u} \in \mathcal{K}_b - w : \;\; \mathcal{J}(w + \tilde{u}) + \phi(w + \tilde{u}) \leq \mathcal{J}(w + \tilde{v}) + \phi(w + \tilde{v}) \quad \forall \tilde{v} \in \mathcal{K}_b - w$$

with

$$\mathcal{K}_b - w = \{v \in H^1(\Omega) : tr_{\gamma_D} v = 0 \wedge tr_{\gamma_S} v \leq -tr_{\gamma_S} w\} \subset H^1_{\gamma_D}(\Omega)\,.$$

Furthermore, the characteristic function

$$\chi_{\mathbb{R}_0^-} : x \mapsto \begin{cases} 0 & \text{for } x \leq 0 \\ +\infty & \text{for } x > 0 \end{cases}$$

of the nonempty, closed and convex subset $\mathbb{R}_0^-$ of $\mathbb{R}$ induces a convex, lower semicontinuous and proper functional $\psi^S : H^1(\Omega) \to \mathbb{R} \cup \{+\infty\}$ defined by

$$\psi^S(v) = \int_{\gamma_S} \chi_{\mathbb{R}_0^-}(v(x))\, d\sigma(x) \quad \forall v \in H^1(\Omega)\,. \tag{2.4.12}$$

Obviously, ψ^S is just the characteristic functional $\chi_{\mathcal{C}}$ of the closed and convex subset $\mathcal{C} := \{v \in H^1(\Omega) : tr_{\gamma_S} v \leq 0\}$ of $H^1(\Omega)$. With the definition

$$F_w(\cdot) := F(w + \cdot) \tag{2.4.13}$$

for translated mappings $F : V \to W$ between vector spaces V and W, the functional $\mathcal{J}_w + \phi_w + \psi^S_w$ is still convex, lower semicontinuous and proper on the Hilbert space $H^1_{\gamma_D}(\Omega)$, and ψ^S_w is the characteristic functional of the subset $\mathcal{K}_b - w$ of $H^1_{\gamma_D}(\Omega)$. In light of our considerations so far and using these notations, the following proposition is straightforward.

Proposition 2.4.6. *The minimization problem (2.3.23) is equivalent to*

$$\tilde{u} \in H^1_{\gamma_D}(\Omega) : \;\; \mathcal{J}_w(\tilde{u}) + \phi_w(\tilde{u}) + \psi^S_w(\tilde{u}) \leq \mathcal{J}_w(v) + \phi_w(v) + \psi^S_w(v) \quad \forall v \in H^1_{\gamma_D}(\Omega) \tag{2.4.14}$$

in the sense that the solution u of (2.3.23) equals $w + \tilde{u}$.

In the following, a variational inclusion is found to be a reformulation of (2.4.14) in terms of the subdifferentials of $\phi(w+\tilde{u})$ and $\psi^S(w+\tilde{u})$ as subsets of $H^1_{\gamma_D}(\Omega)'$.

Proposition 2.4.7. *The minimization problem (2.4.14) is equivalent to the variational inclusion*

$$\tilde{u} \in H^1_{\gamma_D}(\Omega) : \;\; 0 \in a(w + \tilde{u}, \cdot) - \ell(\cdot) + \partial\phi(w + \tilde{u}) + \partial\psi^S(w + \tilde{u}) \tag{2.4.15}$$

in $H^1_{\gamma_D}(\Omega)'$.

Proof. Considering (2.3.17), it is easy to see that first,

$$\partial(\mathcal{J}_w)(v_0)(v) = (\mathcal{J}_w)'(v_0)(v) = \mathcal{J}'(w + v_0)(v) = a(w + v_0, v) - \ell(v)$$

and secondly,

$$\partial(\mathcal{J}_w + \phi_w + \psi_w^S)(v_0)(v) = a(w + v_0, v) - \ell(v) + (\partial\phi)_w(v_0)(v) + (\partial\psi^S)_w(v_0)(v)$$

holds for all $v \in H^1_{\gamma_D}(\Omega)$ and $v_0 \in \mathrm{dom}(\phi_w + \psi_w^S) \subset H^1_{\gamma_D}(\Omega)$. Now, if $\tilde{u} \in H^1_{\gamma_D}(\Omega)$ solves the minimization problem (2.4.14), i.e.

$$(\mathcal{J}_w(v)+\phi_w(v)+\psi_w^S(v))-(\mathcal{J}_w(\tilde{u})+\phi_w(\tilde{u})+\psi_w^S(\tilde{u})) \geq 0 \quad \forall v \in H^1_{\gamma_D}(\Omega)\,, \tag{2.4.16}$$

we obviously have

$$0 \in a(w + \tilde{u}, \cdot) - \ell(\cdot) + (\partial\phi)_w(\tilde{u}) + (\partial\psi^S)_w(\tilde{u})$$

in $H^1_{\gamma_D}(\Omega)'$ by Definition 2.4.1 of the subdifferential. Conversely, if (2.4.15) holds then, with the same reasoning, we obtain (2.4.16). □

Now we are in a position to establish the connection between variational inclusions (2.4.7) and convex minimization problems (2.3.23) which we announced above. Since we have $(\partial\Phi(u), \cdot)_{L^2(\Omega)} \subset \partial\phi(u)$ as already seen in (2.4.6), the following consequence of Proposition 2.4.7 is straightforward.

Proposition 2.4.8. *Every solution of the variational inclusion (2.4.7) is a solution of the corresponding convex minimization problem (2.3.23).*

Observe that the converse of the assertion in Proposition 2.4.8 is not true and we cannot replace $\partial\phi(\tilde{u} + w)(\cdot)$ in Proposition (2.4.7) by $(\partial\Phi(\tilde{u} + w), \cdot)_{L^2(\Omega)}$ since, for $w = 0$ already, we have $\partial\phi(u) \neq (\partial\Phi(u), \cdot)_{L^2(\Omega)}$ as mentioned in (2.4.6). However, an analogous reformulation of $\partial\phi(u)$ in terms of $\partial\Phi(u(x))$ can be obtained in the space-discretized version of (2.3.23) or (2.4.14) which we consider in the following section. We remark that a contribution to the convex functional coming from the boundary will also play an important role in the treatment of Robin boundary conditions which we discuss in Section 3.4.

Finally, instead of considering the subdifferentials of the nondifferentiable parts ϕ_w and ψ_w^S of the convex functional considered in (2.4.14), one can also restrict oneself to differentiating $\mathcal{J}_w$ as done in (2.4.15) while maintaining the convex contributions of ϕ_w and ψ_w^S as they are. In this way, one arrives at another variational inequality that follows from a generalization of Proposition 2.3.11 which we note here.

Proposition 2.4.9. *In addition to the assumptions in Proposition 2.3.11 let* $G : K \to \mathbb{R} \cup \{+\infty\}$ *be convex. Then*

$$u \in K : \quad \partial_{v-u}F(u) + G(v) - G(u) \geq 0 \quad \forall v \in K \tag{2.4.17}$$

is equivalent to

$$u \in K : \quad (F + G)(u) \leq (F + G)(v) \quad \forall v \in K\,.$$

The proof can be easily obtained by adapting the proof of the earlier result and exploiting the convexity of G, see also [34, p. 38]. It is a convention in the literature (see [34, p. 7]) to treat the inequality in (2.4.17) as a valid assertion only if the left hand side is defined, i.e. in this case only if $G(v)$ and $G(u)$ are not both equal to $+\infty$.

For the second statement in the following proposition, we define the translation of our convex set $\mathcal{K}$ (see (2.3.3)) by w onto a subset of $H^1_{\gamma_D}(\Omega)$ as

$$\mathcal{K}_{\gamma_D} := \mathcal{K} - w = \{v \in H^1_{\gamma_D}(\Omega) : v \geq u_c - w \,\wedge\, tr_{\gamma_S} v \leq -tr_{\gamma_S} w\} . \qquad (2.4.18)$$

With regard to the first statement of the proposition, observe that the sum of the functionals ϕ_w and ψ^S_w is just the characteristic functional of $\mathcal{K}_{\gamma_D}$.

Proposition 2.4.10. *The minimization problem (2.4.14) is equivalent to the variational inequality*

$$\begin{aligned}\tilde{u} \in H^1_{\gamma_D}(\Omega) : \quad & a(w+\tilde{u}, v-\tilde{u}) - \ell(v-\tilde{u}) \\ & + \phi_w(v) - \phi_w(\tilde{u}) + \psi^S_w(v) - \psi^S_w(\tilde{u}) \geq 0 \quad \forall v \in H^1_{\gamma_D}(\Omega) \quad (2.4.19)\end{aligned}$$

and to the variational inequality

$$\begin{aligned}\tilde{u} \in \mathcal{K}_{\gamma_D} : \quad & a(w+\tilde{u}, v-\tilde{u}) - \ell(v-\tilde{u}) \\ & + \phi(w+v) - \phi(w+\tilde{u}) \geq 0 \quad \forall v \in \mathcal{K}_{\gamma_D} . \quad (2.4.20)\end{aligned}$$

Proof. First, setting $K = V = H^1_{\gamma_D}(\Omega)$ and $F = \mathcal{J}_w$ with (2.3.17) as well as $G = \phi_w + \psi^S_w$, we get the variational inequality (2.4.19) from Proposition 2.4.9. Setting instead $K = \mathcal{K}_{\gamma_D} \subset H^1_{\gamma_D}(\Omega)$ and $G = \phi_w$, we obtain (2.4.20) as a variational formulation of the translated minimization problem

$$\tilde{u} \in \mathcal{K}_{\gamma_D} : \quad \mathcal{J}(w+\tilde{u}) + \phi(w+\tilde{u}) \leq \mathcal{J}(w+v) + \phi(w+v) \quad \forall v \in \mathcal{K}_{\gamma_D} \quad (2.4.21)$$

which is certainly equivalent to the untranslated problem (2.3.23). □

Although $\phi(w+\cdot)$ is differentiable on $\mathcal{K}_{\gamma_D}$, it is useful not to compute the corresponding directional derivative for the variational formulation (2.4.20). This formulation will be crucial in the analysis of the finite element discretization which we present in the next section. Another advantage of the variational inequality (2.4.20) is that it allows an easy proof for the well-posedness of the convex minimization problem (2.3.23) whose solution even depends Lipschitz continuously on the linear functional ℓ.

Proposition 2.4.11. *Assume that the conditions in Theorem 2.3.16 are satisfied (and possibly the boundedness of M replaced by (2.3.25)). Furthermore, for $i = 1, 2$ let $\ell_i \in H^1(\Omega)'$ and let u_i be the unique solutions of*

$$u_i \in \mathcal{K} : \quad \frac{1}{2}a(u_i, u_i) - \ell_i(u_i) + \phi(u_i) \leq \frac{1}{2}a(v,v) - \ell_i(v) + \phi(v) \quad \forall v \in \mathcal{K} . \quad (2.4.22)$$

Then we have

$$\|u_1 - u_2\|_1 \le c^{-1}\|\ell_1 - \ell_2\| \qquad (2.4.23)$$

where c is the coercivity constant of $a(\cdot,\cdot)$ in (2.3.6).

Proof. With (2.4.18), (2.4.20) and Proposition 2.4.6 we can write (2.4.22) in the form

$$u_i \in \mathcal{K}: \quad a(u_i, v - u_i) + \phi(v) - \phi(u_i) \ge \ell_i(v - u_i) \quad \forall v \in \mathcal{K}\,, \;\; i = 1, 2\,.$$

Now, setting $v = u_2$ for $i = 1$ and $v = u_1$ for $i = 2$ we obtain

$$a(-u_1, u_2 - u_1) - \phi(u_2) + \phi(u_1) \le \ell_1(u_1 - u_2) \qquad (2.4.24)$$

and

$$a(u_2, u_2 - u_1) - \phi(u_1) + \phi(u_2) \le -\ell_2(u_1 - u_2)\,. \qquad (2.4.25)$$

Adding (2.4.24) and (2.4.25) gives

$$a(u_2 - u_1, u_2 - u_1) \le (\ell_1 - \ell_2)(u_1 - u_2) \qquad (2.4.26)$$

which provides (2.4.23) with the coercivity constant c in (2.3.6). □

Using the differentiability of ϕ, one can establish the well-posedness (2.4.23) of the minimization problem (2.3.23) with the variational inequality (2.3.15) in the same way as done in the proof with the variational inequality (2.4.20). Then the contributions of $\phi(u_1)$ and $\phi(u_2)$ do not cancel each other out but appear as an additional term

$$\partial_{u_1 - u_2}\phi(u_1) - \partial_{u_1 - u_2}\phi(u_2)$$

on the left hand side (2.4.26). But this term is nonnegative due to the convexity of ϕ which can easily be derived by (2.4.3). In fact, (2.4.3) shows that if the directional derivative of a convex functional on $\mathcal{K}$ exists, it is a monotone operator on $\mathcal{K}$ (compare also (2.3.13)). We will turn to further monotonicity considerations in Chapter 3 where we apply Proposition 2.4.11 in the proof of Theorem 3.4.23.

2.5 Finite element discretization

In this section we present our finite element discretization of (2.3.23) or (2.4.14), respectively, following ideas and the notation in Kornhuber [59, pp. 36–43] (see also Glowinski [45, pp. 12–15]). As stated in Remark 2.3.1, we keep our assumptions on the problem as general as possible (see e.g. the conditions on the unique solvability of (2.3.23)) in order to make clear what properties of the problem are really needed. We bear in mind that these properties are satisfied in case of the Richards equation with the Brooks–Corey parameter functions but also for the limit cases as given in Section 1.4 where M is monotonically increasing and bounded but not continuous.

2.5.1 The discrete problem: properties and reformulations

For the sake of presentation we consider the two-dimensional case of a polygonal domain $\Omega \subset \mathbb{R}^2$. At the same time we emphasize that the convergence results for our discretization can be obtained analogously for polyhedral domains in higher and lower dimensions. Let $\mathcal{T}_j$, $j \in \mathbb{N}_0$, be a partition of Ω into triangles $t \in \mathcal{T}_j$ with minimal diameter of order $\mathcal{O}(2^{-j})$. We assume the triangulation $\mathcal{T}_j$ to be regular in the sense that the intersection of two different triangles in $\mathcal{T}_j$ is either empty or consists of a common edge or a common vertex. The set of all vertices of the triangles in $\mathcal{T}_j$ is denoted by $\mathcal{N}_j$.

For a consistent discretization, the set $\mathcal{N}_j \cap \partial\Omega$ should resolve the parts of the boundary corresponding to different boundary conditions properly. Therefore, we require that each intersection point of two closures (in $\mathbb{R}^2$) of the subsets γ_D, γ_N and γ_S of the boundary $\partial\Omega$ is contained in $\mathcal{N}_j$. Furthermore, we assume γ_D and $\gamma_S \cup \gamma_D$ to be closed and we define $\mathcal{N}_j^D := \mathcal{N}_j \cap \gamma_D$ as well as $\mathcal{N}_j^S := \mathcal{N}_j \cap \gamma_S$.

We choose the finite element space $\mathcal{S}_j \subset H^1(\Omega)$ as the subspace of all continuous functions in $H^1(\Omega)$ which are linear on each triangle $t \in \mathcal{T}_j$. Analogously, we define $\mathcal{S}_j^D \subset H^1_{\gamma_D}(\Omega)$. $\mathcal{S}_j$ and $\mathcal{S}_j^D$ are spanned by the nodal bases

$$\Lambda_j := \{\lambda_p^{(j)} : p \in \mathcal{N}_j\} \quad \text{and} \quad \Lambda_j^D := \{\lambda_p^{(j)} : p \in \mathcal{N}_j \backslash \mathcal{N}_j^D\},$$

respectively, where the latter is only guaranteed because of our special choice of $\mathcal{N}_j^D$ containing all intersection points of parts of $\partial\Omega$ adjacent to γ_D.

Note that $\mathcal{N}_j$ and $\mathcal{S}_j$, $j \geq 0$, should not be confused with the corresponding notation in Kornhuber [59] for the homogeneous case where $\mathcal{N}_j$ is the set of all vertices in $\mathcal{T}_j$ which are interior points of Ω and $\mathcal{S}_j \subset H_0^1(\Omega)$ is defined accordingly. Of course this also applies to $\mathcal{K}$ and $\mathcal{K}_j$ which we define now.

For the definition of the finite dimensional analogue of $\mathcal{K}$ we assume that the Dirichlet boundary condition u_D is continuous in each node $p \in \mathcal{N}_j^D$ such that writing $u_D(p)$ makes sense in these nodes. Then it is natural to define this convex set $\mathcal{K}_j \subset \mathcal{S}_j$ by

$$\mathcal{K}_j := \{v \in \mathcal{S}_j : v(p) \geq u_c \, \forall p \in \mathcal{N}_j \wedge v(p) = u_D(p) \, \forall p \in \mathcal{N}_j^D \wedge v(p) \leq 0 \, \forall p \in \mathcal{N}_j^S\} \tag{2.5.1}$$

which, as a subset of the finite dimensional space $\mathcal{S}_j$, is clearly nonempty and closed.

Remark 2.5.1. Obviously, $\mathcal{K}_j$ is the set of all piecewise linear interpolations of functions in $\mathcal{K}$ on the triangulation $\mathcal{T}_j$. However, $\mathcal{K}_j \subset \mathcal{K}$ is false in general because the Dirichlet boundary values in $\mathcal{K}_j$ differ from those in $\mathcal{K}$ in general. Observe that, as a consequence of the piecewise linearity of $v \in \mathcal{S}_j$, the two properties $v(x) \geq u_c$ $\forall x \in \Omega$ and $v(x) \leq 0$ $\forall x \in \gamma_S$ if $v \in \mathcal{K}_j$, valid in the continuous case, are preserved in the discretization. The second property is due to our choice of $\mathcal{N}_j^S$ which contains all intersection points of the closures of γ_S and γ_N while the intersection points of the closure of γ_S and γ_D are contained in $\mathcal{N}_j^D$. But in such points p the Dirichlet boundary condition u_D is supposed to be continuous, which entails $u_D(p) \le 0$ because u_D is chosen to be compatible with the Signorini-type boundary condition, see (1.5.18).

Furthermore, we approximate the integral in the definition (2.3.10) of ϕ by a quadrature formula arising from $\mathcal{S}_j$-interpolation of the integrand $\Phi(v)$. In this way, we arrive at the discrete functional $\phi_j : \mathcal{S}_j \to \mathbb{R} \cup \{+\infty\}$ defined by

$$\phi_j(v) := \sum_{p\in\mathcal{N}_j} \Phi(v(p))\, h_p \quad \forall v \in \mathcal{S}_j \tag{2.5.2}$$

with the positive weights

$$h_p := \int_\Omega \lambda_p^{(j)}(x)\, dx\,.$$

Of course, the properties of functionals in the continuous case (see Proposition 2.3.7) should be preserved by the discretized functionals, preferably in a uniform way. With regard to the convergence result for our discretization in Theorem 2.5.9, the following properties of the discrete functionals ϕ_j and, in particular, their relation to the continuous counterpart ϕ in (2.5.5) will be crucial.

Lemma 2.5.2. *Provided M in (2.3.9) is monotonically increasing and bounded, the functional ϕ_j is convex and Lipschitz continuous on its domain*

$$\operatorname{dom}\phi_j = \{v \in \mathcal{S}_j : v(p) \ge u_c \;\; \forall p \in \mathcal{N}_j\} \tag{2.5.3}$$

with a Lipschitz constant independent of $j \ge 0$. Furthermore, ϕ_j is lower semicontinuous and proper and it admits an estimate

$$\phi_j(v) \ge C\|v\|_1 \quad \forall v \in \mathcal{S}_j \tag{2.5.4}$$

with a constant $C \in \mathbb{R}$ independent of $j \ge 0$. Moreover, for $v_j \in \mathcal{S}_j$, $j \ge 0$, and $v \in H^1(\Omega)$ we have

$$v_j \rightharpoonup v,\; j \to \infty \implies \liminf_{j\to\infty} \phi_j(v_j) \ge \phi(v) \tag{2.5.5}$$

where $v_j \rightharpoonup v$ denotes the weak convergence of v_j to v in $H^1(\Omega)$.

Proof. Since the function Φ is convex on $[u_c, \infty)$ and the weights h_p are positive, the functional ϕ_j is convex on its domain

$$\operatorname{dom}\phi_j = \{v \in \mathcal{S}_j : v(p) \ge u_c \;\; \forall p \in \mathcal{N}_j\}$$

which is closed and compact in $\mathcal{S}_j$. Next, let $L = \|M\|_\infty$ be the Lipschitz constant of Φ according to Lemma 2.3.6 and $v, \bar{v} \in \operatorname{dom}\phi_j$. Then, due to the nonnegativity of the nodal basis functions λ_p, $p \in \mathcal{N}_j$, we have

$$|\phi_j(v) - \phi_j(\bar{v})| \le L \sum_{p\in\mathcal{N}_j} |v(p) - \bar{v}(p)|\, h_p \le L \sum_{p\in\mathcal{N}_j} \int_{\operatorname{supp}\lambda_p^{(j)}} |v(x) - \bar{v}(x)|\, dx \tag{2.5.6}$$

where the last estimate is a consequence of the linearity of v and $\bar{v}$ on each triangle $t \in \mathcal{T}_j$. Since each triangle is contained in the support of the three nodal basis functions corresponding to its vertices, we can go on estimating

$$|\phi_j(v) - \phi_j(\bar{v})| \leq 3L \int_\Omega |v(x) - \bar{v}(x)|\, dx \leq 3L\, \|1\|_{L^2(\Omega)} \|v - \bar{v}\|_1$$

where the last estimate follows from the Cauchy–Schwarz inequality in $L^2(\Omega)$. This shows the Lipschitz continuity of ϕ_j, with a Lipschitz constant independent of $j \geq 0$. Of course, $\phi_j : \mathcal{S}_j \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous and proper.

With regard to (2.5.4) and (2.5.5), observe that

$$\phi_j(v) = \int_\Omega \sum_{p \in \mathcal{N}_j} \Phi(v(p))\, \lambda_p^{(j)}(x)\, dx \geq \int_\Omega \Phi\Big(\sum_{p \in \mathcal{N}_j} v(p) \lambda_p^{(j)}(x) \Big)\, dx = \phi(v) \;\; \forall v \in \mathcal{S}_j \tag{2.5.7}$$

because $\sum_{p \in \mathcal{N}_j} \lambda_p^{(j)}(x) = 1$ and $\lambda_p^{(j)}(x) \geq 0$ holds for all $x \in \Omega$ and the scalar function Φ is convex. Now, (2.5.4) follows from (2.3.11). Furthermore, since the convex functional $\phi : H^1(\Omega) \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous, it is weakly lower semicontinuous (see [34, p. 11]) such that (2.5.7) provides

$$\liminf_{j \to \infty} \phi_j(v_j) \geq \liminf_{j \to \infty} \phi(v_j) \geq \phi(v) \tag{2.5.8}$$

for $v_j \rightharpoonup v$, $j \to \infty$, as given above. □

Remark 2.5.3. Observe that the above proof depends heavily on our choice of linear finite elements. Although we only need the positivity of the weights h_p to show the convexity of ϕ_j, we also take into account the special shape of the nodal basis functions and their nonnegativity in order to obtain (2.5.6). For (2.5.7), too, we need the nonnegativity of $\lambda_p^{(j)}$ in each point $x \in \Omega$. It is therefore unclear if one can derive some variant of Lemma (2.5.2) for elements of higher order (whose nodal basis functions are in general not nonnegative on their domain) which can be used to obtain analogous results as we do below in our convergence analysis.

Now, our discrete version of the minimization problem (2.3.23) reads

$$u_j \in \mathcal{K}_j : \;\; \mathcal{J}(u_j) + \phi_j(u_j) \leq \mathcal{J}(v) + \phi_j(v) \quad \forall v \in \mathcal{K}_j\, . \tag{2.5.9}$$

Since $\mathcal{K}_j$, $\mathcal{J}$ and ϕ_j satisfy the required properties of Theorem 2.3.16, now in the subspace $\mathcal{S}_j$ of the Hilbert space $H^1(\Omega)$, we obtain

Theorem 2.5.4. *The discrete minimization problem (2.5.9) has a unique solution.*

As in the continuous case above, we can reformulate the discrete minimization problem in the convex set as a discrete minimization problem in a finite dimensional Hilbert space and rewrite the latter as a discrete variational inclusion. To this end, we choose

$$w_j \in \mathcal{S}_j \quad \text{with} \quad w_j(p) = u_D(p) \;\; \forall p \in \mathcal{N}_j^D \tag{2.5.10}$$

and consider the characteristic functional $\psi^S = \chi_{\mathcal{C}}$ in (2.4.12) restricted on $\mathcal{S}_j^D$. Then, again with the notation (2.4.13) and carrying over the considerations, which led to Propositions 2.4.6 and 2.4.7, on the discrete level, the following can be proved analogously.

Proposition 2.5.5. *The minimization problem (2.5.9) is equivalent to*

$$\tilde{u}_j \in \mathcal{S}_j^D: \ \mathcal{J}_{w_j}(\tilde{u}_j) + (\phi_j)_{w_j}(\tilde{u}_j) + \psi^S_{w_j}(\tilde{u}_j) \leq \mathcal{J}_{w_j}(v) + (\phi_j)_{w_j}(v) + \psi^S_{w_j}(v) \quad \forall v \in \mathcal{S}_j^D \tag{2.5.11}$$

in the sense that the solution u_j of (2.5.9) equals $w_j + \tilde{u}_j$. Furthermore, it is equivalent to the variational inclusion

$$\tilde{u}_j \in \mathcal{S}_j^D: \quad 0 \in a(w_j + \tilde{u}_j, \cdot) - \ell(\cdot) + \partial\phi_j(w_j + \tilde{u}_j) + \partial\psi^S(w_j + \tilde{u}_j) \tag{2.5.12}$$

in $(\mathcal{S}_j^D)'$.

Remark 2.5.6. Observe that the characteristic functional $\psi^S = \chi_{\mathcal{C}}$ restricted to $\mathcal{S}_j$ is the characteristic functional $\chi_{\mathcal{C}_j}$ of the subset

$$\mathcal{C}_j := \{v \in \mathcal{S}_j^D : v(p) \leq 0 \ \ \forall p \in \mathcal{N}_j^S\}$$

of $\mathcal{S}_j^D$. This is due to the piecewise linearity of the functions in $\mathcal{S}_j^D$ and the definition of $\mathcal{N}_j^S$ as a subset of γ_S including all intersection points with adjacent parts of the boundary except those contained in γ_D. Therefore, instead of considering the integral in (2.4.12), we can equivalently compute

$$\psi^S(v) = \sum_{p \in \mathcal{N}_j^S} \chi_{\mathbb{R}_0^-}(v(p)) \quad \forall v \in \mathcal{S}_j^D .$$

While in the continuous case in the previous section it was not possible to interchange integration with taking the subdifferential, in the discrete case we have $\operatorname{dom}\partial\phi_j = \operatorname{dom}\phi_j$ with

$$\partial\phi_j(v_0)(v) = \sum_{p \in \mathcal{N}_j} \partial\Phi(v_0(p))v(p)\, h_p \quad \forall v \in \mathcal{S}_j^D \tag{2.5.13}$$

for all $v_0 \in \operatorname{dom}\partial\phi_j$ and $\operatorname{dom}\partial\psi^S = \operatorname{dom}\psi^S \subset \mathcal{S}_j^D$ with

$$\partial\psi^S(v_0)(v) = \sum_{p \in \mathcal{N}_j^S} \partial\chi_{\mathbb{R}_0^-}(v_0(p))v(p) \quad \forall v \in \mathcal{S}_j^D \tag{2.5.14}$$

for all $v_0 \in \operatorname{dom}\psi^S \subset \mathcal{S}_j^D$. This is due to a general result stating a summation rule for subdifferentials of convex, lower semincontinuous and proper functionals which are continuous on their domains, see [34, p. 26] or [59, pp. 35, 38].

Finally, we turn to the discrete analogue of Proposition 2.4.10. Therefore, we define the translation of our discrete convex set $\mathcal{K}_j$ (see (2.5.1)) by w_j onto a subset of $\mathcal{S}_j^D$ as

$$\begin{aligned} \mathcal{K}_j^D &:= \mathcal{K}_j - w_j \qquad\qquad (2.5.15) \\ &= \{v \in \mathcal{S}_j^D : v(p) \geq u_c - w_j(p) \ \ \forall p \in \mathcal{N}_j \ \wedge \ v(p) \leq -w_j(p) \ \ \forall p \in \mathcal{N}_j^S\} \end{aligned}$$

analogously as in the continuous case $\mathcal{K}_{\gamma_D}$ in (2.4.18). As for $\mathcal{K}_j$ and $\mathcal{K}$, observe that $\mathcal{K}_j^D$ is the set of all piecewise linear interpolations of the functions in $\mathcal{K}_{\gamma_D}$ on the triangulation $\mathcal{T}_j$ if we choose w in (2.4.11) to be continuous and w_j in (2.5.10) as its piecewise linear interpolation in $\mathcal{S}_j$. However, the constraints in the continuous and the discrete case may differ such that we have $\mathcal{K}_j^D \not\subseteq \mathcal{K}_{\gamma_D}$ in general.

Proposition 2.5.7. *The minimization problem (2.5.9) is equivalent to the variational inequality*

$$\begin{aligned}\tilde{u}_j \in \mathcal{S}_j^D :\ & a(w_j + \tilde{u}_j, v - \tilde{u}_j) - \ell(v - \tilde{u}_j) \\ & + (\phi_j)_{w_j}(v) - (\phi_j)_{w_j}(\tilde{u}_j) + \psi^S_{w_j}(v) - \psi^S_{w_j}(\tilde{u}_j) \geq 0 \quad \forall v \in \mathcal{S}_j^D \end{aligned} \tag{2.5.16}$$

and to the variational inequality

$$\begin{aligned}\tilde{u}_j \in \mathcal{K}_j^D :\ & a(w_j + \tilde{u}_j, v - \tilde{u}_j) - \ell(v - \tilde{u}_j) \\ & + \phi_j(w_j + v) - \phi_j(w_j + \tilde{u}_j) \geq 0 \quad \forall v \in \mathcal{K}_j^D \end{aligned} \tag{2.5.17}$$

in the sense that the solution u_j of (2.5.9) equals $w_j + \tilde{u}_j$.

The proof is the same as for Proposition 2.4.10 if one replaces $H^1_{\gamma_D}(\Omega)$ by $\mathcal{S}_j^D$, $\mathcal{K}_{\gamma_D}$ by $\mathcal{K}_j^D$ and ϕ by ϕ_j for the application of Proposition 2.4.9. Again, observe that (2.5.17) can be regarded as a variational formulation of the translated minimization problem

$$\tilde{u}_j \in \mathcal{K}_j^D :\ \mathcal{J}(w_j + \tilde{u}_j) + \phi_j(w_j + \tilde{u}_j) \leq \mathcal{J}(w_j + v) + \phi_j(w_j + v) \quad \forall v \in \mathcal{K}_j^D \tag{2.5.18}$$

which is equivalent to the untranslated problem (2.5.9). Finally, note that an analogous well-posedness result as in Proposition 3.3.8 in the continuous case also holds for the discrete problem (2.5.9) with the same proof.

2.5.2 A classical convergence result

Now, we deal with the convergence of our finite element solutions from (2.5.9) to the solution of the continuous problem (2.3.23) for which the variational inequalities (2.4.20) and (2.5.17) will play a central role. The derivation of our results is largely based on the arguments given in Kornhuber [59, pp. 38–42] for the case of homogeneous Dirichlet boundary conditions on all of $\partial\Omega$. We need to take special care of the inhomogeneous Dirichlet values and the Signorini-type boundary conditions defined only on parts of $\partial\Omega$.

As in Kornhuber [59, pp. 38–42], the convergence results depend on the assumption that the corresponding sequence of triangulations has a decreasing mesh size

$$h_j = \max_{t \in \mathcal{T}_j} \operatorname{diam} t \to 0 \quad \text{for} \quad j \to \infty\,. \tag{2.5.19}$$

In addition, we assume that the sequence of triangulations

$$(\mathcal{T}_j)_{j\geq 0} \text{ is } \textit{shape regular} \tag{2.5.20}$$

which denotes the well-known property that the minimal interior angle of all triangles contained in $\cup_{j\geq 0}\mathcal{T}_j$ is bounded from below by a positive constant.

For further discussion, we introduce the piecewise linear interpolation operator

$$I_{\mathcal{S}_j} : H^1(\Omega) \cap C(\overline{\Omega}) \to \mathcal{S}_j$$

defined by $(I_{\mathcal{S}_j} v)(p) = v(p) \;\; \forall p \in \mathcal{N}_j$ for $v \in H^1(\Omega) \cap C(\overline{\Omega})$.

It will turn out that we can only guarantee convergence if the Dirichlet boundary condition u_D on γ_D is the trace of a uniformly continuous function $w \in H^1(\Omega)$, i.e. if we have

$$u_D = tr_{\gamma_D} w \quad \text{for a} \quad w \in H^1(\Omega) \cap C(\overline{\Omega})\,. \tag{2.5.21}$$

It is well known that if u_D is continuous on γ_D (which we assumed to be closed), then it can be extended to a continuous function on the closure of Ω (see [98, p. 498]). We require that there exists such an extension in $H^1(\Omega)$. Moreover, for the proof of convergence we will assume w_j to be the piecewise linear interpolations of w in $\mathcal{S}_j$ approximating w in $H^1(\Omega)$, i.e. we require

$$w_j = I_{\mathcal{S}_j} w \quad \text{with} \quad w_j \to w \text{ for } j \to \infty \text{ in } H^1(\Omega). \tag{2.5.22}$$

In general, according to the interpolation theory in Ciarlet [27, pp. 122–124], the latter can only be guaranteed if w is regular enough. To check the assumptions stated there, we recall that the Sobolev embedding theorem (see [21, p. 1.52]) provides the compact embedding

$$H^k(t) \hookrightarrow C(\overline{t}) \quad \Longleftrightarrow \quad k > \frac{d}{2}$$

for polyhedra $t \subset \mathbb{R}^d$ and $k \in \mathbb{N}$. Now, with $t \in \mathcal{T}_j$ and $d = 2$ in our case we obtain (2.5.22) (with the order $\mathcal{O}(h_j)$) for $w \in H^2(\Omega)$ from the results in [27, pp. 122–124], provided (2.5.19) and also (2.5.20) hold. Consequently, we could also replace (2.5.21) and (2.5.22) by $u_D = tr_{\gamma_D} w$ with the condition $w \in H^2(\Omega)$ or a corresponding condition for $d > 2$.

In Kornhuber [59, pp. 38/39] it is proved that for $\tilde{\mathcal{K}} = \{v \in H_0^1(\Omega) : v \geq u_c\}$ the subset $C_0^\infty(\Omega) \cap \tilde{\mathcal{K}}$ is dense in $\tilde{\mathcal{K}}$. Since $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, for given $v \in \tilde{\mathcal{K}}$ there is always a sequence $(v_k)_{k\geq 0} \subset C_0^\infty(\Omega)$ with $v_k \to v$ for $k \to \infty$. In order to ensure $v_k \in \tilde{\mathcal{K}}$, however, regularizations of v with mollifiers are considered. It is by far a nontrivial task to extend this result to more general settings like our convex set $\mathcal{K}_{\gamma_D}$ in $H^1_{\gamma_D}(\Omega)$. The technique can be refined (see Glowinski [45, pp. 36–38]) to generalize the result to continuous obstacles on $\overline{\Omega}$ which are nonnegative in a neighbourhood of $\gamma_D = \partial\Omega$ as $u_c - w$ is in our case (2.4.18) with the property (2.5.21). Furthermore, an exercise in [45, pp. 38/39] suggests that the latter result can be extended to $H^1_{\gamma_D}(\Omega)$ for sufficiently smooth $\gamma_D \subset \partial\Omega$ if one applies the density of

$$C^\infty_{\gamma_D}(\overline{\Omega}) := \{v \in C^\infty(\overline{\Omega}) : v = 0 \text{ in a neighbourhood of } \gamma_D\} \tag{2.5.23}$$

in $H^1_{\gamma_D}(\Omega)$. As to the boundary conditions of Signorini's type one finds a proof for the density of $C^\infty(\overline{\Omega}) \cap \bar{\mathcal{K}}$ in the convex set $\bar{\mathcal{K}} = \{v \in H^1(\Omega) : tr_{\partial\Omega}\, v \geq 0\}$ in [45, p. 61]. Since we do not want to go into more details here, it seems to be in order to require

$$C^\infty_{\gamma_D}(\overline{\Omega}) \cap \mathcal{K}_{\gamma_D} \text{ is dense in } \mathcal{K}_{\gamma_D} \tag{2.5.24}$$

as an additional condition for our translated convex set $\mathcal{K}_{\gamma_D}$ defined in (2.4.18). Property (2.5.24) will provide an essential density argument in the proof of the convergence result in Theorem 2.5.9.

As a necessary ingredient for convergence, the following lemma provides the consistency of the discrete functionals ϕ_j. The proof is essentially the same as in Kornhuber [59, pp. 38–40] for the homogeneous case. However, we state it here in order to make clear where we need the assumptions on the extension w of u_D and the interpolating w_j.

Lemma 2.5.8. *We assume (2.5.19), (2.5.20) and M in (2.3.9) to be bounded and monotonically increasing. If $v \in C^\infty(\overline{\Omega})$ and $v_j = I_{\mathcal{S}_j} v \in \mathcal{S}_j$ for $j \geq 0$, then we have the convergence*

$$v_j \to v \;\; in \;\; H^1(\Omega) \quad and \quad \phi_j(v_j) \to \phi(v) \quad for \;\; j \to \infty\,. \tag{2.5.25}$$

Assuming in addition (2.5.21) and (2.5.22), the assertion (2.5.25) also holds for $v = w + \tilde{v} \in w + C^\infty(\overline{\Omega})$ and $v_j = I_{\mathcal{S}_j} v = I_{\mathcal{S}_j} w + I_{\mathcal{S}_j} \tilde{v} = w_j + \tilde{v}_j$, $j \geq 0$.

Proof. Since $C^\infty(\overline{\Omega})$ is dense in all $H^k(\Omega)$, $k \geq 0$, (see Ciarlet [27, p. 114]) we obtain $v_j \to v$ by applying the results in [27, pp. 122–124] under the assumptions (2.5.19) and (2.5.20). The same applies for $v = w + \tilde{v} \in w + C^\infty(\overline{\Omega})$ and $v_j = I_{\mathcal{S}_j} v$ because of (2.5.22).

Now, we show $\phi_j(v_j) \to \phi(v)$ if $v_j \to v$ and if v is uniformly continuous which is true for $v \in C^\infty(\overline{\Omega})$ and for $v \in w + C^\infty(\overline{\Omega})$ due to (2.5.21). If $\phi(v) = \infty$, then there is an open subset $\Omega' \subset \Omega$ with $v(x) < u_c \;\forall x \in \Omega'$. Due to (2.5.19) we obtain $\mathcal{N}_j \cap \Omega' \neq \emptyset$ for all $j \geq j_0$ with a suitable $j_0 \geq 0$. This provides $\phi_j(v_j) \to \infty$ for $j \to \infty$.

If $\phi(v) < 0$, i.e. $v(x) \geq u_c \;\forall x \in \overline{\Omega}$, then the function $\Phi(v(\cdot))$ is uniformly continuous on $\overline{\Omega}$ since $v : \overline{\Omega} \to [u_c, \infty)$ is uniformly continuous and $\Phi : [u_c, \infty) \to \mathbb{R}$ is Lipschitz continuous (see Lemma 2.3.6). Let $p_1, p_2, p_3 \in \mathcal{N}_j$ denote the vertices of $t \in \mathcal{T}_j$. Then we have $\sum_{i=1}^3 \lambda^{(j)}_{p_i} = 1$ on t and therefore altogether

$$\begin{aligned} |\phi(v) - \phi_j(v_j)| &\leq \sum_{t \in \mathcal{T}_j} \int_t \left(\sum_{i=1}^3 \lambda^{(j)}_{p_i}(x) |\Phi(v(x)) - \Phi(v(p_i))| \right) dx \\ &\leq |\Omega| \max_{\{x,\xi \in \Omega,\, |x-\xi| \leq h_j\}} |\Phi(v(x)) - \Phi(v(\xi))| \to 0 \end{aligned} \tag{2.5.26}$$

for $j \to \infty$. □

Now, with the consistency result in Lemma 2.5.8 and the further properties of ϕ_j in Lemma 2.5.2 we can prove convergence, adapting the ideas in Kornhuber [59, pp. 41/42] to our inhomogeneous case.

Theorem 2.5.9. *We assume (2.5.19)–(2.5.24) and the conditions imposed in Theorem 2.3.16 except for the continuity of M. Then the solutions u_j of the discrete minimization problem (2.5.9) converge to the solution u of (2.3.23) in the sense that*

$$u_j \to u \;\; in \;\; H^1(\Omega) \quad and \quad \phi_j(u_j) \to \phi(u) \quad for \;\; j \to \infty\,. \tag{2.5.27}$$

Proof. As in [59, pp. 41/42], the proof is carried out in three steps. For the argumentation we consider the translated minimization problems (2.4.21) and (2.5.18) and the corresponding variational inequalities in the form (2.4.20) and (2.5.17), respectively. Since we assume $w_j \to w$, we can mainly concentrate on the functions $\tilde{u}_j \in \mathcal{K}_j^D \subset \mathcal{S}_j^D \subset H^1_{\gamma_D}(\Omega)$ instead of $u_j = w_j + \tilde{u}_j \in w_j + \mathcal{S}_j^D$ and $\tilde{u} \in \mathcal{K}_{\gamma_D} \subset H^1_{\gamma_D}(\Omega)$ instead of $u = w + \tilde{u} \in w + H^1_{\gamma_D}(\Omega)$. Indeed, this is necessary in the first and in the third step where we need to exploit the fact that due to (2.3.5) and (2.3.6), $a(\cdot,\cdot)$ induces an equivalent norm on $H^1_{\gamma_D}(\Omega)$ defined by

$$|v|_a := a(v,v)^{1/2} \quad \forall v \in H^1_{\gamma_D}(\Omega)\,.$$

First, we prove that $(\tilde{u}_j)_{j\geq 0}$ is bounded in $H^1_{\gamma_D}(\Omega)$. Using (2.5.24) we choose a $v \in C^\infty(\overline{\Omega}) \cap \mathcal{K}_{\gamma_D}$ and define $v_j = I_{\mathcal{S}_j} v \in \mathcal{K}_j^D$ for $j \geq 0$. This is well-defined since $I_{\mathcal{S}_j}(\mathcal{K}_{\gamma_D}) = \mathcal{K}_j^D$ because of $w_j = I_{\mathcal{S}_j} w$, see (2.5.15). Since $\tilde{u}_j$ satisfies the variational inequality (2.5.17), we have

$$|\tilde{u}_j|_a^2 = a(\tilde{u}_j,\tilde{u}_j) \leq a(w_j+\tilde{u}_j, v_j) - a(w_j,\tilde{u}_j) + \phi_j(w_j+v_j) - \phi_j(w_j+\tilde{u}_j) - \ell(v_j-\tilde{u}_j)\,.$$

Lemma 2.5.8 provides uniform upper bounds for $\|v_j\|_1$ and $|\phi_j(w_j+v_j)|$, and we have the uniform lower estimate (2.5.4). Using this, the continuity (2.3.5) of $a(\cdot,\cdot)$ and $\ell(\cdot)$ as well as the equivalence (2.3.6) of $|\cdot|_a$ and $\|\cdot\|_1$, then, with some $C > 0$, we can go on estimating

$$\begin{aligned} c\|\tilde{u}\|_1^2 &\leq C(\|w_j\|_1 + \|\tilde{u}_j\|_1 + \|w_j\|_1\|\tilde{u}_j\|_1) + C(\|w_j\|_1 + \|\tilde{u}_j\|_1) + C\|\tilde{u}_j\|_1 + C \\ &\leq \tilde{C}\|\tilde{u}_j\|_1 + \tilde{C} \end{aligned}$$

where $\tilde{C} > 0$ is sufficiently large, taking into account that $(\tilde{w}_j)_{j\geq 0}$ is bounded due to (2.5.22).

Therefore, $(\tilde{u}_j)_{j\geq 0}$ must be bounded in $H^1_{\gamma_D}(\Omega)$.

In the second step, we prove the weak convergence of $(\tilde{u}_j)_{j\geq 0}$ to $\tilde{u}$ in $H^1_{\gamma_D}(\Omega)$. Since $H^1_{\gamma_D}(\Omega)$ is a Hilbert space, i.e. reflexive, there is a weakly convergent subsequence $(\tilde{u}_{j_k})_{k\geq 0}$ with a limit $\tilde{u}^* \in H^1_{\gamma_D}(\Omega)$ due to the boundedness of $(\tilde{u}_j)_{j\geq 0}$ (see e.g. [98, p. 107]). In order to prove $\tilde{u} = \tilde{u}^*$, it is enough to show that $\tilde{u}^*$ is a solution of (2.4.20), since (for fixed w) (2.4.20) is uniquely solvable due to Proposition 2.4.10 and Theorem 2.3.16.

We need to make clear that $\tilde{u}^* \in \mathcal{K}_{\gamma_D}$ holds in the first place. As we have

$$w_{j_k} + \tilde{u}_{j_k} \in \mathcal{K}_a := \{v \in H^1(\Omega) : v \geq u_c \wedge tr_{\gamma_S} v \leq 0\} \quad \forall k \geq 0$$

in the closed and convex set $\mathcal{K}_a \subset H^1(\Omega)$ and $w_{j_k} \to w$, i.e. $w_{j_k} + \tilde{u}_{j_k} \rightharpoonup w + \tilde{u}^*$, this leads to $w + \tilde{u}^* \in \mathcal{K}_a$ since $\mathcal{K}_a$ is also weakly closed (see e.g. [98, p. 108]). Due to $tr_{\gamma_D} w = u_D$ and $\tilde{u}^* \in H^1_{\gamma_D}(\Omega)$ we get $w + \tilde{u}^* \in \mathcal{K}$ and therefore $\tilde{u}^* \in \mathcal{K}_{\gamma_D}$.

To show that $\tilde{u}^* \in \mathcal{K}_{\gamma_D}$ satisfies the variational inequality (2.4.20), assume $v \in C^\infty(\overline{\Omega}) \cap \mathcal{K}_{\gamma_D}$ and let $v_j = I_{\mathcal{S}_j} v \in \mathcal{K}_j^D$ for $j \geq 0$ as in the first step. With these discrete test functions in the discrete variational inequality (2.5.17) we obtain

$$\begin{aligned} a(\tilde{u}_{j_k}, \tilde{u}_{j_k}) + \phi_{j_k}(w_{j_k} + \tilde{u}_{j_k}) \leq a(w_{j_k} + \tilde{u}_{j_k}, v_{j_k}) - a(w_{j_k}, \tilde{u}_{j_k}) \\ + \phi_{j_k}(w_{j_k} + v_{j_k}) - \ell(v_{j_k} - \tilde{u}_{j_k}) \end{aligned}$$

as above. Since we have $w_{j_k} \to w$, $\tilde{u}_{j_k} \rightharpoonup \tilde{u}^*$ and $v_{j_k} \to v$ (by Lemma 2.5.8), passing to the limit while using the consistency (2.5.25) of ϕ_j gives

$$\begin{aligned} \liminf_{k\to\infty}(a(\tilde{u}_{j_k}, \tilde{u}_{j_k}) + \phi_{j_k}(w_{j_k} + \tilde{u}_{j_k})) \leq a(w + \tilde{u}^*, v) - a(w, \tilde{u}^*) \\ + \phi(w + v) - \ell(v - \tilde{u}^*) . \end{aligned} \tag{2.5.28}$$

Now, using the elementary estimate

$$0 \leq a(\tilde{u}^* - \tilde{u}_{j_k}, \tilde{u}^* - \tilde{u}_{j_k}) = a(\tilde{u}^*, \tilde{u}^*) - 2a(\tilde{u}^*, \tilde{u}_{j_k}) + a(\tilde{u}_{j_k}, \tilde{u}_{j_k})$$

and the weak convergence of $\tilde{u}_{j_k}$, we get

$$a(\tilde{u}^*, \tilde{u}^*) \leq \liminf_{k\to\infty}(a(\tilde{u}_{j_k}, \tilde{u}_{j_k})) .$$

In connection with (2.5.5) in Lemma 2.5.2, this provides

$$a(\tilde{u}^*, \tilde{u}^*) + \phi(w + \tilde{u}^*) \leq \liminf_{k\to\infty}(a(\tilde{u}_{j_k}, \tilde{u}_{j_k}) + \phi_{j_k}(w_{j_k} + \tilde{u}_{j_k})) .$$

Combining this estimate with (2.5.28) we conclude

$$a(w + \tilde{u}^*, v - \tilde{u}^*) - \ell(v - \tilde{u}^*) + \phi(w + v) - \phi(w + \tilde{u}^*) \geq 0 \quad \forall v \in C^\infty(\overline{\Omega}) \cap \mathcal{K}_{\gamma_D} . \tag{2.5.29}$$

Now, the density assumption (2.5.24) comes into play to extend (2.5.29) to all $v \in \mathcal{K}_{\gamma_D}$. For any $v \in \mathcal{K}_{\gamma_D}$ assumption (2.5.24) provides a sequence $(v_k)_{k\geq 0}$ in $C^\infty(\overline{\Omega}) \cap \mathcal{K}_{\gamma_D}$ converging to v. Since (2.5.29) is satisfied for all v_k, $k \geq 0$, and the functionals on the left hand side of (2.5.24) acting on v_k are continuous in this argument on $\mathcal{K}_{\gamma_D}$ (in particular ϕ, see Proposition 2.3.7), (2.5.24) is satisfied for v.

We have proved that $\tilde{u}^* = \tilde{u}$ is the solution of (2.4.20). The uniqueness of the solution entails $\tilde{u}_j \rightharpoonup \tilde{u}$ for $j \to \infty$.

In the final step, we show the strong convergence of $(\tilde{u}_j)_{j\geq 0}$. Again, we consider a $v \in C^\infty(\overline{\Omega}) \cap \mathcal{K}_{\gamma_D}$ with $v_j = I_{\mathcal{S}_j} v \in \mathcal{K}_j^D$ for $j \geq 0$. And again, using the discrete variational inequality (2.5.17) we compute

$$\begin{aligned}
&|\tilde{u} - \tilde{u}_j|_a^2 + \phi_j(w_j + \tilde{u}_j) \\
&= a(\tilde{u}, \tilde{u}) - 2a(\tilde{u}, \tilde{u}_j) + a(\tilde{u}_j, \tilde{u}_j) + \phi_j(w_j + \tilde{u}_j) \\
&\leq a(\tilde{u}, \tilde{u}) - 2a(\tilde{u}, \tilde{u}_j) + a(w_j + \tilde{u}_j, v_j) - a(w_j, \tilde{u}_j) + \phi_j(w_j + v_j) - \ell(v_j - \tilde{u}_j) \,.
\end{aligned} \tag{2.5.30}$$

With the same arguments as used above, the right-hand side in the estimate (2.5.30) converges to $a(w + \tilde{u}, v - \tilde{u}) + \phi(w + v) - \ell(v - \tilde{u})$ as $j \to \infty$. Hence, we can estimate

$$\begin{aligned}
&\liminf_{j\to\infty} \phi_j(w_j + \tilde{u}_j) \\
&\qquad \leq \liminf_{j\to\infty} (|\tilde{u} - \tilde{u}_j|_a^2 + \phi_j(w_j + \tilde{u}_j)) \\
&\qquad \leq \limsup_{j\to\infty} (|\tilde{u} - \tilde{u}_j|_a^2 + \phi_j(w_j + \tilde{u}_j)) \\
&\qquad \leq a(w + \tilde{u}, v - \tilde{u}) + \phi(w + v) - \ell(v - \tilde{u}) \quad \forall v \in C^\infty(\overline{\Omega}) \cap \mathcal{K}_{\gamma_D} \,.
\end{aligned} \tag{2.5.31}$$

Again, applying (2.5.24) and using the same density argument as above, we obtain (2.5.31) for all $v \in \mathcal{K}_{\gamma_D}$. Then, inserting $v = \tilde{u}$ in (2.5.31) and using (2.5.5) we deduce

$$\phi(w + \tilde{u}) \leq \liminf_{j\to\infty} \phi_j(w_j + \tilde{u}_j) \leq \limsup_{j\to\infty} (|\tilde{u} - \tilde{u}_j|_a^2 + \phi_j(w_j + \tilde{u}_j)) \leq \phi(w + \tilde{u}) \,.$$

From this we conclude $\phi_j(w_j + \tilde{u}_j) \to \phi(w + \tilde{u})$ as well as $|\tilde{u}_j - \tilde{u}|_a \to 0$, i.e. $\|\tilde{u}_j - \tilde{u}\|_1 \to 0$ due to the equivalence of these norms in $H^1_{\gamma_D}(\Omega)$ and therefore $u_j = w_j + \tilde{u_j} \to u = w + \tilde{u}$ in $H^1(\Omega)$. This completes the proof. □

Remark 2.5.10. With regard to the structure of the proof and in comparison to the homogeneous case given in Kornhuber [59, pp. 41/42], several notes seem to be in order.

Concerning the first step of the proof, observe the following: Since the Dirichlet value u_D is taken out of the set given in (1.5.18), it obviously has an extension w in the convex set $\mathcal{K}$ in (1.5.19) or (2.3.3). Now, if there is such an extension w also satisfying (2.5.21) and (2.5.22), which is certainly true for homogeneous Dirichlet boundary conditions for instance, we can choose it for our argumentation in here. Consequently, we can choose $v = 0 \in \mathcal{K}_{\gamma_D} = \mathcal{K} - w$ and $v_j = I_{\mathcal{S}_j} v = 0 \in \mathcal{K}_j^D = \mathcal{K}_j - w_j$ as the test functions above, which simplifies the first step considerably.

Furthermore, we note that in the second step we do not get $\tilde{u}^* \in \mathcal{K}_{\gamma_D}$ immediately from the weak convergence $\tilde{u}_{j_k} \rightharpoonup \tilde{u}^*$, since although

$$\tilde{u}_{j_k} \in \mathcal{K}_{j_k}^D = I_{\mathcal{S}_{j_k}}(\mathcal{K}_{\gamma_D}) \quad \forall k \geq 0 \,,$$

we have $\mathcal{K}^D_{j_k} \not\subseteq \mathcal{K}_{\gamma_D}$ in general (see (2.5.15)). We rather need to deal with the solutions $u_{j_k} = w_{j_k} + \tilde{u}_{j_k}$ of (2.5.9) which we consider as elements of $\mathcal{K}_a$ since the discrete and the continuous Dirichlet values differ in general. Observe, however, that $w + \tilde{u}^* \geq 0$ also follows from (2.5.29) which gives $\phi(w + \tilde{u}^*) < \infty$.

It does not seem possible to relax our density assumption (2.5.24) (e.g. to the well-known density (2.5.23) of $C^\infty_{\gamma_D}(\overline{\Omega}) \cap H^1_{\gamma_D}(\Omega)$ in $H^1_{\gamma_D}(\Omega)$) by applying the variational inequality (2.4.19) and its discrete counterpart (2.5.16) in the proof. Although the latter is feasible by establishing a straightforward analogue of Lemma 2.5.8 and (2.5.5) for ψ^S, the density argument in the second and in the third step cannot be carried out if for a $v \in \mathcal{K}_{\gamma_D}$ there is only a sequence $(v_k)_{k\geq 0} \subset C^\infty_{\gamma_D}(\overline{\Omega}) \cap (H^1_{\gamma_D}(\Omega) \backslash \mathcal{K}_{\gamma_D})$ converging to v. Since $\mathrm{dom}(\phi + \psi^S) = \mathcal{K}_{\gamma_D}$, we would have $(\phi + \psi^S)(v_k) = \infty$ for all $k \geq 0$ although $(\phi + \psi^S)(v) < 0$ in such a case (see also [59, p. 42]).

In order not to confuse our notation with the one chosen in (2.4.20) and (2.5.17), $w_j + \tilde{u}_j$ and $w + \tilde{u}$ are not abbreviated by u_j and u, respectively, in this proof. However, the assertions about $(\tilde{u}_j)_{j\geq 0}$ and its limit correspond to the analogous ones for $(u_j)_{j\geq 0}$ and its limit, and the proof, as well as (2.4.20) and (2.5.17), could be reformulated accordingly.

It is clear that the conditions (2.5.21) and (2.5.22) on w and w_j only need to be satisfied for a single $w \in H^1(\Omega)$ with $tr_{\gamma_D} w = u_D$ in order to obtain the convergence result. For the assertion of Theorem 2.5.9 to hold, it is certainly irrelevant how the function w_j satisfying (2.5.10) is chosen to compute the solution u_j of (2.5.9) since u_j does not depend on a special choice of w_j. However, the iterates $\tilde{u}_j$, $j \geq 0$, do not converge unless the interpolants w_j, $j \geq 0$, do.

2.5.3 Generalizations of the convergence result

With regard to the continuous setting discussed in Remark 2.3.17 it is natural to ask to what extent the convergence results for our finite element discretization of (2.3.23) can be generalized. The hydraulic conductivity $K_h(\cdot)$ can be chosen to be space-dependent because it is included in the bilinear form $a(\cdot,\cdot)$ which is just restricted to $\mathcal{S}_j \times \mathcal{S}_j$ in the discretization and of which only the continuity and the coerciveness are used. Therefore, $K_h(\cdot)$ can be chosen as in the continuous case.

The question how a space-dependent porosity $n(\cdot)$ can be treated in the discretization seems to be more interesting, since ϕ is not evaluated exactly on $\mathcal{S}_j$. Keeping in mind how $n(\cdot)$ is included in the continuous case (see Remark 2.3.17), we replace the weights h_p in (2.5.2) by

$$\tilde{h}_p := \int_\Omega n(x)\lambda_p(x)\,dx \tag{2.5.32}$$

which provides a natural discretization $\tilde{\phi}_j$ of the generalized convex functional

$$\phi : v \mapsto \int_\Omega n(x)\Phi(v(x))\,dx \quad \forall v \in \mathcal{K} \tag{2.5.33}$$

instead of (2.3.10). We choose $n(\cdot)$ to be positive in order to obtain positive weights and the same domain $\operatorname{dom}\tilde{\phi}_j$ as in (2.5.3). Furthermore, $n(\cdot)$ should be bounded as in the continuous setting. Then the proofs of the crucial Lemmas 2.5.2 and 2.5.8 can be easily carried over to the general case.

Possible generalizations of the convergence results above to approximations of $a(\cdot,\cdot)$ by $a_j(\cdot,\cdot)$ and $\ell(\cdot)$ by $\ell_j(\cdot)$, respectively, e.g. obtained by quadrature, shall not be considered here. Nevertheless, it is tempting to replace the generalized ϕ in (2.5.33) by a "full" $\mathcal{S}_j$-interpolation of the integrand which results in a discretization $\overline{\phi}_j$ with the weights

$$\overline{h}_p := n(p) \int_\Omega \lambda_p(x)\, dx = n(p)\, h_p$$

instead of (2.5.32). Then $\Phi(v(\cdot))$ in (2.5.26) has to be replaced by $n(\cdot)\Phi(v(\cdot))$ such that $n(\cdot)$ needs to be uniformly continuous on Ω (or equivalently on $\overline{\Omega}$) if we want to preserve the argument in (2.5.26) for the consistency result in Lemma 2.5.8. Unfortunately, although the rest of the proof of Lemma 2.5.2 can be carried out analogously, the crucial property (2.5.5) seems to be unclear or at least more difficult to derive since the estimate in (2.5.7) might no longer be true for space-dependent $n(\cdot)$. The integrands in (2.5.7) can be multiplied with $n(x)$, which provides

$$\tilde{\phi}_j(v) \geq \phi(v) \quad \forall v \in \mathcal{S}_j\,,$$

but $n(x)$ cannot be replaced by $n(p)$ in general in the sum on the left hand side of the inequality in (2.5.7). (We remark that one could instead replace $n(x)$ by $\max\{n(x) : x \in \operatorname{supp}\lambda_p\}$ in order to obtain this and all the other assertions to be checked here.) However, one could hope that if $n(\cdot)$ is regular enough, this replacement "is allowed in the limit" $j \to \infty$ such that the essential assertion (2.5.8) is still satisfied. This would require to prove

$$|\tilde{\phi}_j(v_j) - \overline{\phi}_j(v_j)| \to 0 \quad \text{for } v_j \rightharpoonup v,\ j \to \infty\,, \tag{2.5.34}$$

which does not seem to be straightforward. With a glance at (2.5.26), we can estimate

$$|\tilde{\phi}_j(v_j) - \overline{\phi}_j(v_j)| \leq |\Omega| \max_{\{x,\xi\in\Omega,\, |x-\xi|\leq h_j\}} |n(x)-n(\xi)| \sup_{j\geq 0,\, p\in\mathcal{N}_j} |\Phi(v_j(p))|$$

in which

$$\sup_{j\geq 0,\, p\in\mathcal{N}_j} |\Phi(v_j(p))| < \infty \tag{2.5.35}$$

does not seem to be satisfied in general. Due to the weak convergence of $(v_j)_{j\geq 0}$, we only have the uniform bound $\|v_j\|_1 \leq C \ \ \forall j \geq 0$ with a $C > 0$. Therefore, we can guarantee (2.5.35) at least in one space dimension because of the well-known continuous Sobolev embedding

$$H^1(\Omega) \hookrightarrow C(\overline{\Omega}) \tag{2.5.36}$$

for bounded intervals $\Omega \subset \mathbb{R}$ (see for example [1, p. 98] for a version of it including also unbounded Ω). In this case at least, the uniform continuity of

$n(\cdot)$ and (2.5.35) provide (2.5.34) and therefore (2.5.8) in Lemma 2.5.2. We close these considerations by noting that our numerical method for the solution of (2.5.9), which we present in the following two sections, is not affected by the special choice of the weights h_p as long as they are nonnegative.

With regard to the properties of M, it is clear that the same convergence results as above can be obtained for $M : \mathbb{R} \to \mathbb{R}$ defined on the whole real line, thus accounting for the nondegenerate case and corresponding limit cases concerning the Richards equation (see Section 1.4). In Kornhuber [59] one finds analogue versions of Lemmas 2.5.2 and 2.5.8 for an even more general situation including property (2.3.25) (note for example that Lipschitz continuity of Φ is not necessary in the proof of Lemma 2.5.8).

2.5.4 Convergence of the discrete saturation and the physical pressure for the Brooks–Corey functions

One can consider the convergence result $\phi_j(u_j) \to \phi(u)$ in Theorem 2.5.9 as unsatisfactory from the point of view of our original problem, the variational inequality (2.3.8). With the hydrological background from the first chapter one might be more interested in the convergence of the saturation $M(u_j) \to M(u)$. To address this issue we note the following general convergence results in $L^2(\Omega)$ which can be formulated in the framework of superposition operators introduced in Definition 1.5.9.

Proposition 2.5.11. *Let $\Omega \subset \mathbb{R}^d$ be bounded and $M : \mathbb{R} \to \mathbb{R}$ uniformly continuous and bounded. Then the superposition operator M_Ω acts on $L^2(\Omega)$ and is continuous.*

Sketch of the proof. M_Ω acts on $L^2(\Omega)$ because Ω and M are bounded. For the proof of continuity let $(u_n)_{n\geq 0} \subset L^2(\Omega)$ with $u_n \to u$ for $n \to \infty$ in $L^2(\Omega)$. Now, we can argue similarly as in the proof of Proposition 1.5.14. One can split Ω in

$$\Omega^n_{>\varepsilon} = \{x \in \Omega : |u(x) - u_n(x)|^2 > \varepsilon\}$$

and $\Omega^n_{\leq\varepsilon} = \Omega \backslash \Omega^n_{>\varepsilon}$ for an $\varepsilon > 0$ and derive $|\Omega^n_{>\varepsilon}| \to 0$ for $n \to \infty$ with the Lebesgue measure $|\cdot|$ from the convergence $u_n \to u$ in $L^2(\Omega)$. Using the uniform continuity of M on $\Omega^n_{\leq\varepsilon}$ and its boundedness on $\Omega^n_{>\varepsilon}$, one can show $M(u_n) \to M(u)$ in $L^2(\Omega)$. □

If M is more regular we obtain more.

Proposition 2.5.12. *Let $\Omega \subset \mathbb{R}^d$ be bounded. If $M : \mathbb{R} \to \mathbb{R}$ is Hölder continuous with respect to the exponent $\alpha \in (0, 1]$, then M induces a superposition operator*

$$M_\alpha : L^2(\Omega) \to L^{2/\alpha}(\Omega)$$

which is also Hölder continuous with respect to α.

Sketch of the proof. One can show that $|M(u(\cdot))|^{2/\alpha}$ is integrable by considering $|M(u(\cdot))| \leq |M(u(\cdot)) - M(u(x_0))| + |M(u(x_0))|$ for an $x_0 \in \Omega$, using the inequality

$$(a+b)^q \leq 2^{q-1}(a^q + b^q) \tag{2.5.37}$$

for $a, b \geq 0$ and $q \geq 1$ (consult e.g. [54, p. 161]) with $q = 2/\alpha$ and then the Hölder continuity of M. The claimed Hölder continuity of M_α is straightforward. □

With the continuous embedding $i : L^{2/\alpha}(\Omega) \hookrightarrow L^2(\Omega)$ for bounded $\Omega \subset \mathbb{R}^d$, it follows easily that $i \circ M_\alpha : L^2(\Omega) \to L^2(\Omega)$ is Hölder continuous with respect to α which improves the convergence result in Proposition 2.5.11 for Hölder continuous M. Recall that the generalized saturation M in (1.3.25) from the Brooks–Corey parameter functions for the Richards equation is Hölder continuous. Of course, we can apply the results to $u_j \to u$ from Theorem 2.5.9. Note that in order to obtain

$$M(u_j) \to M(u) \quad \text{in } L^2(\Omega) \text{ for } j \to \infty\,, \tag{2.5.38}$$

it is enough to assume the properties of M (in Propositions 2.5.11 or 2.5.12) only on the union of the ranges of the functions u_j, $j \geq 0$, and u. Otherwise, one can think of $M : [u_c, \infty) \to \mathbb{R}$ to be extended on $\mathbb{R}$ by the value $M(u_c)$. Furthermore, it is clear that Propositions 2.5.11 or 2.5.12 hold for any $L^p(\Omega)$ with $p \geq 1$ instead of $p = 2$.

For practical purposes one might be even more interested in the $\mathcal{S}_j$-interpolation of $M(u_j)$ rather than in the exact function $M(u_j)$, see (2.7.36) and (2.7.37). With regard to the convergence of this inexact evaluation of $M(u_j)$, we state the following result.

Theorem 2.5.13. *Let $M : \mathbb{R} \to \mathbb{R}$ be Hölder continuous with respect to the exponent $\alpha \in (0, 1]$. Then, for linear finite element functions $u_j \in \mathcal{S}_j$, $j \geq 0$, satisfying $u_j \to u$ in $H^1(\Omega)$ for $j \to \infty$, we have*

$$M(u_j) - I_{\mathcal{S}_j} M(u_j) \to 0 \quad \textit{in } L^2(\Omega) \textit{ for } j \to \infty\,.$$

Proof. For any point x contained in a triangle $t \in \mathcal{T}_j$ with the vertices p_1, p_2, p_3 there are $\lambda_i \in [0, 1]$, $i = 1, 2, 3$, with $\sum_{i=1}^3 \lambda_i = 1$ such that

$$I_{\mathcal{S}_j} M(u_j)(x) = \sum_{i=1}^{3} \lambda_i\, M(u(p_i))\,.$$

Therefore, using binomial formulas (as (2.5.37)) and the Hölder continuity of M with the Hölder constant C_α, we can estimate

$$\begin{aligned} |M(u_j(x)) - I_{\mathcal{S}_j} M(u_j)(x)|^2 &\leq \left(\sum_{i=1}^{3} \lambda_i\, |M(u_j(x)) - M(u_j(p_i))| \right)^2 \\ &\leq 3 \sum_{i=1}^{3} |M(u_j(x)) - M(u_j(p_i))|^2 \leq 3\, C_\alpha^2 \sum_{i=1}^{3} |u_j(x) - u_j(p_i)|^{2\alpha}\,. \end{aligned} \tag{2.5.39}$$

Using the mean value theorem

$$|u_j(x) - u_j(p_i)| \le |\nabla u_j||x - p_i|$$

on the triangle t (with the Euclidean norm $|\cdot|$ on $\mathbb{R}^d$) while considering that $|\nabla u_j|$ is constant on t, we can go on estimating the last term in (2.5.39) by

$$9\, C_\alpha^2\, |\nabla u_j|^{2\alpha}\, h_j^{2\alpha}$$

with h_j as in (2.5.19). Now, integration over Ω provides

$$\begin{aligned}\int_\Omega |M(u_j(x)) - I_{\mathcal{S}_j} M(u_j)(x)|^2\, dx &\le \sum_{t\in\mathcal{T}_j} \int_t |M(u_j(x)) - I_{\mathcal{S}_j} M(u_j)(x)|^2\, dx \\ &\le 9\, C_\alpha^2\, h_j^{2\alpha} \int_\Omega (|\nabla u_j|^2 + 1)\, dx\,.\end{aligned}$$

Since $(u_j)_{j\ge 0}$ converges in $H^1(\Omega)$, the last integral is uniformly bounded and therefore this whole last term tends to 0 as $j \to \infty$ due to (2.5.19). □

Note that due to the Sobolev embedding (2.5.36), Propositions 2.5.11 and 2.5.12 as well as Theorem 2.5.13 also hold in one space dimension if $L^2(\Omega)$ and $L^{2/\alpha}(\Omega)$ are replaced by $(C(\overline{\Omega}), \|\cdot\|_\infty)$. In this case the assertions of Proposition 2.5.11 and Theorem 2.5.13 are already satisfied for any uniformly continuous $M : \mathbb{R} \to \mathbb{R}$. Again, Theorem 2.5.13 can also be applied to $M : [u_c, \infty) \to \mathbb{R}$ for our $u_j \in \mathcal{K}_j$, $j \ge 0$, and $u \in \mathcal{K}$. Note that with the continuous embedding $L^2(\Omega) \hookrightarrow L^p(\Omega)$ for all $p \in [1, 2]$ and (2.5.37), the proof can be generalized to these $L^p(\Omega)$, in particular to $p = 1$. Furthermore, with the Cauchy–Schwarz inequality we obtain

$$M(u_j)v_j \to M(u)v \quad \text{in } L^1(\Omega) \text{ for } j \to \infty \tag{2.5.40}$$

from (2.5.38) for

$$v_j \in \mathcal{S}_j,\ j \ge 0\,, \text{ with } v_j \to v \text{ in } H^1(\Omega) \text{ for } j \to \infty\,. \tag{2.5.41}$$

Now, it is not hard to adapt the proof of Theorem 2.5.13 in order to guarantee

$$M(u_j)v_j - I_{\mathcal{S}_j}(M(u_j)v_j) \to 0 \quad \text{in } L^1(\Omega) \text{ for } j \to \infty \tag{2.5.42}$$

for Hölder continuous M and these functions u_j, v_j and u, v. This is interesting because (2.5.40)–(2.5.42) entail

$$\sum_{p\in\mathcal{N}_j} M(u_j(p))\,(v_j(p) - u_j(p))\, h_p \longrightarrow \int_\Omega M(u)\,(v - u)\, dx \quad \text{for } j \to \infty\,. \tag{2.5.43}$$

With a glance at (2.5.12) and (2.5.13), an application of Proposition 2.3.11 shows that the discretization

$$u_j \in \mathcal{K}_j : \sum_{p\in\mathcal{N}_j} M(u_j(p))\,(v(p) - u_j(p))\, h_p + a(u_j, v - u_j) - \ell(v - u_j) \ge 0 \quad \forall v \in \mathcal{K}_j \tag{2.5.44}$$

of the original variational inequality (2.3.8) is equivalent to the discretization (2.5.9) on the level of our convex functional if $M : [u_c, \infty) \to \mathbb{R}$ is continuous. We can rewrite (2.5.43) as the convergence

$$\partial_{v_j - u_j} \phi_j(u_j) \to \partial_{v-u} \phi(u) \quad \text{for } j \to \infty$$

on the level of the variational inequality (2.3.8) and its discretization for our solutions u_j of (2.5.44) and u of (2.3.8) with v_j, v as in (2.5.41). Note that for any $v \in \mathcal{K}$ there is a sequence $(v_j)_{j \geq 0}$ satisfying (2.5.41) because of condition (2.5.24) and Lemma 2.5.8.

It turns out that via the definition (2.3.9) and the analysis on the convex minimization problem (2.3.23), we obtain all the desirable results concerning the variational inequality (2.3.8) and the saturation M which are closer to our original hydrological problem. The same strategy will be further pursued in the following Sections 2.6 and 2.7 in which we deal with the numerical treatment of (2.5.9) or (2.5.44), respectively. However, at the end (see e.g. Figure 2.1), we also come back again to the derivative M of Φ. In particular, we will never need to evaluate Φ (as given in (2.3.9)) in the whole numerical process.

Remark 2.5.14. The above discussion was mainly motivated by the fact that a hydrological situation is much better reflected by the saturation $M(u)$ than the value of the functional $\phi(u)$. But the same can be said in such a case for the physical pressure $p = \kappa^{-1}(u)$ as compared to the generalized pressure $u \in \mathcal{K}$ with the Kirchhoff transformation (1.3.1). It was already indicated in Remark 1.5.19 that we are limited in answering this question for the Brooks–Corey parametrization. For example, we do not know the regularity of the function p which is given almost everywhere on Ω by pointwise application $\kappa^{-1}(u(x))$ of the inverse transformation on the values of u, say, in the Lebesgue points x of u in Ω (see Rudin [82, p. 140]). Depending on the range of u in $[u_c, \infty)$ and the singularity of $\kappa^{-1} : (u_c, \infty) \to \mathbb{R}$ (see (1.3.24) or Figure 1.9) one might not even have $p \in L^1(\Omega)$.

However, if there is a physically meaningful interpretation of (2.3.8) which is given by the corresponding variational inequality in the physical pressure p analogously to (1.5.38), then as derived in Remark 1.5.19 from Theorem 1.5.18, this variational inequality is uniquely solvable by $p = \kappa^{-1}(u) \in H^1(\Omega)$. But even then it is unclear if the iterates $p_j := \kappa^{-1}(u_j)$, $j \geq 0$, converge to p because of the singularity of κ^{-1} in u_c.

It seems that a nontrivial result can only be obtained if the ranges of u and u_j, $j \geq 0$, are uniformly bounded from below by a constant $a > u_c$. In this case κ^{-1} can be regarded as a Lipschitz continuous function on these ranges as in the nondegenerate case $kr \geq c$ for a $c > 0$ in Lemma 1.5.7. Still, these situations can be considered as somewhat natural in a realistic physical setting where one does not expect deteriorating physical pressure values. In these cases $\kappa^{-1} : (a, \infty) \to \mathbb{R}$ can be extended to a Lipschitz continuous function on the real line and the strong result in Theorem 1.5.15 provides the convergence

$$p_j = \kappa^{-1}(u_j) \to p = \kappa^{-1}(u) \quad \text{in } H^1(\Omega) \text{ for } j \to \infty \, .$$

For the interpolants Theorem 2.5.13 guarantees

$$I_{\mathcal{S}_j} p_j \to p \quad \text{in } L^2(\Omega) \text{ for } j \to \infty .$$

Again, in one space dimension the Sobolev embedding theorem (2.5.36) allows us to replace the spaces in these convergence results by $(C(\overline{\Omega}), \|\cdot\|_\infty)$. Note that the Lipschitz continuity of κ^{-1} on compact subsets of (u_c, ∞) provides the same convergence speed of $p_j \to p$ as $u_j \to u$ in $L^2(\Omega)$ or $(C(\overline{\Omega}), \|\cdot\|_\infty)$ due to Lemma 1.5.10.

We cannot close these considerations without mentioning that of course we also obtain

$$M(u_j) \to M(u) \quad \text{in } H^1(\Omega) \text{ for } j \to \infty$$

for the saturation iterates in these nondegenerate cases by Theorem 1.5.15 since $M : (a, \infty) \to \mathbb{R}$ as in (1.3.25) can be extended to a Lipschitz continuous function on the real line, too.

2.5.5 Weak* convergence of the discrete generalized saturation in the limit cases

We finally want to address the situation of discontinuous saturation M, in particular the limit cases in (1.4.25) and in Remark 2.4.5, where M is interpreted in terms of a maximal monotone multifunction $\tilde{M} = \partial\Phi$. As in Remark 2.4.5, but now in the discretized case for general boundary conditions, we can naturally associate $u_j \in \mathcal{K}_j$, $j \geq 0$, to a generalized discrete saturation $m_{u_j} \in \mathcal{S}_j$ which satisfies

$$m_{u_j}(p) = M(u_j(p)) \quad \text{on} \quad \{p \in \mathcal{N}_j : u_j(p) \neq u_c\} \tag{2.5.45}$$

where $\tilde{M}$ is single-valued. The definition is motivated by the equivalent formulation (2.5.12) with (2.5.13) of our discretization and reads

$$m_{u_j}(p) := h_p^{-1}(\ell(\lambda_p) - a(u_j, \lambda_p)) \quad \forall p \in \mathcal{N}_j \backslash \mathcal{N}_j^D , \tag{2.5.46}$$

thus extending (2.5.45) to the singular case $u_j(p) = u_c$ for $p \in \mathcal{N}_j \backslash \mathcal{N}_j^D$. Of course, m_{u_j} can be prescribed on the Dirichlet nodes by choosing values $m_{u_j}(p) \in \tilde{M}(u_D(p))$ for $p \in \mathcal{N}_j^D$. Analogously as in Remark 2.4.5, and again for general boundary conditions, definition (2.5.46) leads to $M(u_j(p)) < \theta_M$ for the limit case discussed in Remark 2.4.5 if $u_j(p) = u_c$ and the variational inequality (2.5.44) is strict for the test function $v = u_j + \lambda_p$. Thus, the set of nodes with this property gives rise to a discrete generalized unsaturated region for the discretization of this limit case. (Note that as in Remark 2.4.5 these considerations are not restricted to the limit case, but that u_c is a singular value in the case of continuous $M : [u_c, \infty) \to \mathbb{R}$ where the unsaturated regime can be represented by hydrologically sensible generalized pressure values in $(u_c, -1)$.)

On the other hand, as in Theorem 2.4.2, i.e. assuming $\gamma_S = \emptyset$ with homogeneous Dirichlet boundary conditions now, the general result in Jerome [51, p. 93] also guarantees the existence of

$$\bar{u}_j \in \mathcal{S}_j^D \quad \text{and} \quad w_{\bar{u}_j} \in \partial\Phi(\bar{u}_j) \tag{2.5.47}$$

(see definition (2.4.5)) such that

$$(w_{\bar{u}_j}, v)_{L^2(\Omega)} + a(\bar{u}_j, v) - \ell(v) = 0 \quad \forall v \in \mathcal{S}_j^D \tag{2.5.48}$$

is satisfied. Of course, $w_{\bar{u}_j} \in L^2(\Omega)$ does not have to be unique or accessible in discrete terms. Instead, one might consider $I_{\mathcal{S}_j} \tilde{w}_{\bar{u}_j}$ for some pointwise evaluable L^2-approximation $\tilde{w}_{\bar{u}_j}$ of $w_{\bar{u}_j}$ (see [82, p. 140]) or the discrete variational equality (2.5.48) in which $w_{\bar{u}_j}$ is replaced by its L^2-projection $\bar{w}_{\bar{u}_j}$ on $\mathcal{S}_j^D$. $\bar{w}_{\bar{u}_j}$ is uniquely determined but does not need to fulfill $\bar{w}_{\bar{u}_j} \in \partial\Phi(\bar{u}_j)$ or the pointwise discrete variant $\bar{w}_{\bar{u}_j}(p) \in \partial\Phi(\bar{u}_j(p))$ for $p \in \mathcal{N}_j \backslash \mathcal{N}_j^D$. The latter, however, holds for m_{u_j} which satisfies the analogue

$$u_j \in \mathcal{S}_j^D \quad \text{and} \quad m_{u_j}(p) \in \partial\Phi(u_j(p)) \;\text{ for }\; p \in \mathcal{N}_j \backslash \mathcal{N}_j^D \tag{2.5.49}$$

with the property

$$(m_{u_j}, v)_j + a(u_j, v) - \ell(v) = 0 \quad \forall v \in \mathcal{S}_j^D \tag{2.5.50}$$

of (2.5.47) and (2.5.48). Here, the *lumped L^2-scalar product* $(\cdot,\cdot)_j$ instead of $(\cdot,\cdot)_{L^2(\Omega)}$ on S_j^D occurs which is given according to (2.5.12) and (2.5.13) by the definition

$$(u, v)_j := \int_\Omega I_{\mathcal{S}_j}(u \cdot v)\, dx = \sum_{p \in \mathcal{N}_j} u(p)v(p)\, h_p \quad \forall u, v \in \mathcal{S}_j \,. \tag{2.5.51}$$

There is a crucial observation with regard to the variational equality (2.5.48) which can be derived with analogous arguments as Proposition 2.4.8.

Proposition 2.5.15. *The solution $\bar{u}_j$ of (2.5.48) is the unique solution of the convex minimization problem*

$$\bar{u}_j \in \mathcal{S}_j^D : \;\; \mathcal{J}(\bar{u}_j) + \phi(\bar{u}_j) \le \mathcal{J}(v) + \phi(v) \quad \forall v \in \mathcal{S}_j^D \,. \tag{2.5.52}$$

Proof. It is easy to see that an analogous version of Proposition 2.4.7 holds for the minimization problem (2.5.52) in terms of a variational inclusion with the subdifferential $\partial\phi$ of ϕ on $\mathcal{S}_j^D$. Now, (2.4.6) shows that the restrictions of the elements in $(\partial\Phi(\bar{u}_j), \cdot)_{L^2(\Omega)}$ onto $\mathcal{S}_j^D$ form a subset of $\partial\phi(\bar{u}_j)$. Therefore, $\bar{u}_j$ in (2.5.48) solves (2.5.52). Finally, (2.5.52) is uniquely solvable due to Theorem 2.3.16. □

It seems obvious that (2.5.52) provides a better approximation to the solution of (2.3.23) than (2.5.9) in which ϕ is not exactly evaluated on $\mathcal{S}_j^D$. Therefore, it is not surprising that we have the following result.

Proposition 2.5.16. *If the conditions of Theorem 2.5.9 hold, then the solutions $\bar{u}_j$, $j \ge 0$, of (2.5.48) converge in $H^1_{\gamma_D}(\Omega)$ to the solution u of (2.3.23).*

The proof can be carried out analogously as above for the convergence result in Theorem 2.5.9. It is just easier.

Equipped with Proposition 2.5.16 we can also ask if convergence of $w_{\bar{u}_j} \to w_u$ or $\bar{w}_{\bar{u}_j} \to w_u$ for $j \to \infty$ in some sense can be established. It turns out that the variational equalities (2.4.8) and (2.5.48) together with the inclusion conditions on the functions w_u and $w_{\bar{u}_j}$ do not seem to provide more than poor results if one does not want to impose unacceptable regularity or stability conditions on these functions. For example, one obtains $\bar{w}_{\bar{u}_j} \to w_u$ in $L^2(\Omega)$ for $j \to \infty$ if $\bar{w}_{\bar{u}_j}$, $j \geq 0$, and w_u are uniformly bounded in $H^1_{\gamma_D}(\Omega)$. If $\|w_{\bar{u}_j}\|_{L^\infty(\Omega)} \leq C$ is satisfied for $j \geq 0$ and a $C > 0$, which can be regarded as the only situation in which a physically interpretable generalized saturation is given for $\bar{u}_j$, $j \geq 0$, and which is certainly fulfilled in the case (1.4.25), one can show that a subsequence of $(w_{\bar{u}_j})_{j\geq 0}$ converges in the weak* sense in $L^\infty(\Omega)$ to w_u and a subsequence of $(\bar{w}_{\bar{u}_j})_{j\geq 0}$ converges weakly in $L^2(\Omega)$ to w_u.

A more interesting result, which requires no further assumptions on $w_{\bar{u}_j}$, $j \geq 0$, and w_u, can be obtained in the dual space $H^1_{\gamma_D}(\Omega)'$. Still, for the obstacle problem in the limit case of Remark (2.4.5) we can only prove it in one space dimension where the Sobolev embedding (2.5.36) is valid.

Proposition 2.5.17. *Let Ω be a bounded interval in case of $M : [u_c, \infty) \to \mathbb{R}$ with a $u_c < 0$ and $\Omega \subset \mathbb{R}^d$ as at the beginning of this section in case of $M : \mathbb{R} \to \mathbb{R}$. In both cases M is assumed to be monotonically increasing and bounded. Furthermore, let u, w_u as well as $\bar{u}_j$, $w_{\bar{u}_j}$, $j \geq 0$, be given according to (2.4.8) as well as (2.5.48), respectively. Then the sequence $((w_{\bar{u}_j}, \cdot)_{L^2(\Omega)})_{j\geq 0}$ of functionals on $H^1_{\gamma_D}(\Omega)$ converges in the weak* sense in $H^1_{\gamma_D}(\Omega)'$ to the functional $(w_u, \cdot)_{L^2(\Omega)}$.*

Proof. First, the uniform boundedness

$$\|w_{\bar{u}_j}\|_{L^\infty(\Omega)} \leq C_0' \quad \forall j \geq 0$$

with a $C_0' > 0$ and therefore

$$|(w_{\bar{u}_j}, v)_{L^2(\Omega)}| \leq \|w_{\bar{u}_j}\|_{L^\infty(\Omega)} \|v\|_{L^2(\Omega)} \leq C' \|v\|_1 \quad \forall v \in H^1_{\gamma_D}(\Omega) \tag{2.5.53}$$

with a $C' > 1$ is clear for $M : \mathbb{R} \to \mathbb{R}$.

The case $M : [u_c, \infty) \to \mathbb{R}$, where $\tilde{M}(u_c)$ is unbounded, is more interesting. Here, we have at least the uniform boundedness

$$\|w_{\bar{u}_j}\|_{L^1(\Omega)} \leq C_0 \quad \forall j \geq 0 \tag{2.5.54}$$

for a $C_0 > 0$, which can be shown in the following way.

Testing (2.5.48) with $v = \bar{u}_j$ and considering the convergence $\bar{u}_j \to u$ in $H^1_{\gamma_D}(\Omega)$ for $j \to \infty$ due to Theorem 2.5.9, one obtains the uniform bound

$$\left| \int_\Omega w_{\bar{u}_j} \bar{u}_j \, dx \right| \leq C_1 \quad \forall j \geq 0 \tag{2.5.55}$$

for a $C_1 > 0$. Furthermore, with Ω_c as in (2.4.9), but now dependent on $j \geq 0$, we have

$$\int_\Omega w_{\bar{u}_j} \bar{u}_j \, dx = u_c \int_{\Omega_c} w_{\bar{u}_j} \, dx + \int_{\Omega \setminus \Omega_c} w_{\bar{u}_j} \bar{u}_j \, dx \, . \tag{2.5.56}$$

Since $M : [u_c, \infty) \to \mathbb{R}$ is bounded, so is the corresponding maximal monotone multifunction $\widetilde{M}(\cdot)$ on (u_c, ∞). Together with (2.5.47), the convergence of $(\bar{u}_j)_{j \geq 0}$ and the boundedness of Ω this leads to

$$\left| \int_{\Omega \setminus \Omega_c} w_{\bar{u}_j} \bar{u}_j \, dx \right| \leq C_2 \quad \forall j \geq 0 \tag{2.5.57}$$

for a $C_2 > 0$. Since the range of $w_{\bar{u}_j}$, $j \geq 0$, is uniformly bounded from above due to the boundedness of M, we obtain (2.5.54) from (2.5.55)–(2.5.57).

Furthermore, the uniform bound (2.5.54) and the Sobolev embedding (2.5.36) provide a $C > 0$ such that we have

$$|(w_{\bar{u}_j}, v)_{L^2(\Omega)}| \leq \|w_{\bar{u}_j}\|_{L^1(\Omega)} \|v\|_{L^\infty(\Omega)} \leq C \|v\|_1 \quad \forall v \in H^1_{\gamma_D}(\Omega) \tag{2.5.58}$$

and also $C \geq C'$ in (2.5.53). Consequently, in both cases considered, the sequence $((w_{\bar{u}_j}, \cdot)_{L^2(\Omega)})_{j \geq 0}$ of linear functionals on $H^1_{\gamma_D}(\Omega)$ is uniformly bounded by C. The function $w_u \in L^2(\Omega)$ also induces a linear functional on $H^1_{\gamma_D}(\Omega)$ in the canonical way

$$|(w, v)_{L^2(\Omega)}| \leq \|w\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leq \|w_u\|_{L^2(\Omega)} \|v\|_1 \quad \forall v \in H^1_{\gamma_D}(\Omega)$$

with a norm bounded by $\|w_u\|_{L^2(\Omega)}$. Let $v \in H^1_{\gamma_D}(\Omega)$. Using (2.5.24) and Lemma 2.5.8 we have $v_j \in \mathcal{S}_j^D$, $j \geq 0$, with $v_j \to v$ in $H^1_{\gamma_D}(\Omega)$ for $j \to \infty$. With these test functions the variational equalities (2.4.8) and (2.5.48) lead to

$$(w_u - w_{\bar{u}_j}, v_j)_{L^2(\Omega)} = a(u - \bar{u}_j, v_j) \quad \forall j \geq 0$$

such that, with a $C_a > 0$, we can estimate

$$\begin{aligned} |(w_u - w_{\bar{u}_j}, v)_{L^2(\Omega)}| &\leq |a(u - \bar{u}_j, v_j)| + |(w_u - w_{\bar{u}_j}, v - v_j)_{L^2(\Omega)}| \\ &\leq C_a \|u - \bar{u}_j\|_1 \|v_j\|_1 + (\|w_u\|_{L^2(\Omega)} + C) \|v - v_j\|_1 \end{aligned}$$

where the last sum converges to 0 because of the convergence $\|u - \bar{u}_j\|_1 \to 0$ (Proposition 2.5.16) and $\|v - v_j\|_1 \to 0$ for $j \to \infty$. □

The last estimate in the proof shows that a strong convergence of $(w_{\bar{u}_j}, \cdot)_{L^2(\Omega)}$, $j \geq 0$, to $(w_u, \cdot)_{L^2(\Omega)}$ in $H^1_{\gamma_D}(\Omega)'$ seems to require a uniform approximation property of $(\mathcal{S}_j^D)_{j \geq 0}$ in $H^1_{\gamma_D}(\Omega)$.

The rest of this subsection is devoted to find a connection between w_{u_j} which satisfies the continuous inclusion in (2.5.47) with m_{u_j} which satisfies the corresponding discrete inclusion in (2.5.49) for $j \to \infty$. This will be achieved in

(2.5.62) in the proof of the next proposition. Observe that our discrete generalized saturation m_{u_j}, $j \ge 0$, in (2.5.46) gives rise to a linear functional at least on a dense subset of $H^1_{\gamma_D}(\Omega)$ via the lumped scalar product

$$v \mapsto (m_{u_j}, v)_j \quad \forall v \in C^\infty_{\gamma_D}(\overline{\Omega}) \tag{2.5.59}$$

given in (2.5.51) and according to (2.5.23). Due to (2.5.50) and Theorem 2.5.9 the norm of these functionals $(m_{u_j}, \cdot)_j$ on $\mathcal{S}_j^D$ is uniformly bounded for $j \ge 0$. The following result shows that they are uniformly bounded on $C^\infty_{\gamma_D}(\overline{\Omega})$ and therefore in $H^1_{\gamma_D}(\Omega)'$ in one space dimension and how they are connected with the generalized saturation w_u of the infinite dimensional problem in this case.

Theorem 2.5.18. *Let $\Omega \subset \mathbb{R}$ be a bounded interval and $M : [u_c, \infty) \to \mathbb{R}$ with a $u_c < 0$ or $M : \mathbb{R} \to \mathbb{R}$ monotonically increasing and bounded. Let u, w_u and u_j, m_{u_j}, $j \ge 0$, be given as in (2.4.8) and (2.5.46), respectively. Then both the sequence $((m_{u_j}, \cdot)_j)_{j\ge 0}$ and the sequence $((m_{u_j}, \cdot)_{L^2(\Omega)})_{j\ge 0}$ of functionals are contained in $H^1_{\gamma_D}(\Omega)$ and converge in the weak* sense in $H^1_{\gamma_D}(\Omega)'$ to the functional $(w_u, \cdot)_{L^2(\Omega)}$.*

Proof. With the same reasoning as in the continuous case (see (2.5.54)) in the proof of Proposition 2.5.17 one can show the existence of a $C_0 > 0$ such that

$$\sum_{p \in \mathcal{N}_j} |m_{u_j}(p)|\, h_p \le C_0 \quad \forall j \ge 0 \tag{2.5.60}$$

is satisfied. Again, this is clear for $M : \mathbb{R} \to \mathbb{R}$, where we even have an analogous $L^\infty(\Omega)$-bound, and it is more interesting for $M : [u_c, \infty) \to \mathbb{R}$. In this case (2.5.60) is proved just as in the derivation of (2.5.54) by testing (2.5.50) with the test function u_j for each $j \ge 0$ and using the convergence $u_j \to u$ in $H^1_{\gamma_D}(\Omega)$ for $j \to \infty$. Focussing on the set of nodes $\mathcal{N}_j^c(u_j)$ where $u_j(p) = u_c$ holds and using (2.5.49), one can prove (2.5.60) with $\mathcal{N}_j^c(u_j)$ replaced by $\mathcal{N}_j$ and thus (2.5.60) altogether.

Now, (2.5.60) shows that $(m_{u_j}, \cdot)_j$, $j \ge 0$, according to (2.5.59) even give rise to bounded linear functionals on $(C(\overline{\Omega}), \|\cdot\|_\infty)$. Moreover, their norms are uniformly bounded by C_0 in this case. In one space dimension, the Sobolev embedding (2.5.36) provides the corresponding statement for these functionals considered on $H^1_{\gamma_D}(\Omega)$, i.e. there is a $C_m > 0$ independent of $j \ge 0$ such that we have

$$(m_{u_j}, v)_j \le C_m \|v\|_1 \quad \forall v \in H^1_{\gamma_D}(\Omega) \ \ \forall j \ge 0\,. \tag{2.5.61}$$

In view of Proposition 2.5.17 we only need to show

$$|(w_{\bar{u}_j}, v)_{L^2(\Omega)} - (m_{u_j}, v)_j| \to 0 \quad \text{for } j \to \infty \tag{2.5.62}$$

for all $v \in H^1_{\gamma_D}(\Omega)$ in order to obtain the weak* convergence of $(m_{u_j}, \cdot)_j$ to $(w_u, \cdot)_{L^2(\Omega)}$ for $j \to \infty$. Now, for any $v \in H^1_{\gamma_D}(\Omega)$ we choose a sequence of functions $v_j \in \mathcal{S}_j^D$ with $v_j \to v$ in $H^1_{\gamma_D}(\Omega)$ according to (2.5.24) and Lemma 2.5.8. Testing the variational equalities (2.5.48) and (2.5.50) with v_j, we obtain

$$(w_{\bar{u}_j}, v_j)_{L^2(\Omega)} - (m_{u_j}, v_j)_j = a(\bar{u}_j - u_j, v_j) \quad \forall j \ge 0\,.$$

Then, with C as in (2.5.58) and C_m as in (2.5.61), we can estimate

$$\begin{aligned}
&|(w_{\bar{u}_j}, v)_{L^2(\Omega)} - (m_{u_j}, v)_j| \\
&\qquad \le |(w_{\bar{u}_j}, v_j)_{L^2(\Omega)} - (m_{u_j}, v_j)_j| + |(w_{\bar{u}_j}, v - v_j)_{L^2(\Omega)} - (m_{u_j}, v - v_j)_j| \\
&\qquad \le C_a \, \|\bar{u}_j - u_j\|_1 \, \|v_j\|_1 + (C + C_m)\|v - v_j\|_1 \, .
\end{aligned}$$

The last term goes to 0 for $j \to \infty$ because $\|\bar{u}_j - u_j\|_1 \to 0$ due to Theorem 2.5.9 and Proposition 2.5.16 and $\|v - v_j\|_1 \to 0$ holds by construction.

With the weak* convergence of $(m_{u_j}, \cdot)_j$ to $(w_u, \cdot)_{L^2(\Omega)}$ for $j \to \infty$ we can now derive the weak* convergence of $(m_{u_j}, \cdot)_{L^2(\Omega)}$ to $(w_u, \cdot)_{L^2(\Omega)}$ for $j \to \infty$ by proving

$$|(m_{u_j}, v)_j - (m_{u_j}, v)_{L^2(\Omega)}| \to 0 \quad \text{for } j \to \infty \tag{2.5.63}$$

for all $v \in H^1_{\gamma_D}(\Omega)$. To see this, we need to use the uniform continuity of $v \in H^1_{\gamma_D}(\Omega) \subset C(\overline{\Omega})$ given by (2.5.36). The left hand side in (2.5.63) can be written and estimated as

$$\left| \int_\Omega \sum_{p \in \mathcal{N}_j} m_{u_j}(p) v(p) \lambda_p(x) - \sum_{p \in \mathcal{N}_j} m_{u_j}(p) \lambda_p(x) v(x) \, dx \right| =$$

$$\left| \sum_{p \in \mathcal{N}_j} m_{u_j}(p) \int_\Omega (v(p) - v(x)) \lambda_p(x) \, dx \right| \le \sum_{p \in \mathcal{N}_j} |m_{u_j}(p)| \|v(p) - v(x)\|_{C(\overline{\operatorname{supp} \lambda_p})} \, h_p \, .$$

This last term tends to 0 for $j \to \infty$ because of (2.5.60) and the uniform continuity of v. The latter provides $\|v(p) - v(x)\|_{C(\overline{\operatorname{supp} \lambda_p})} \to 0$ for $j \to \infty$ uniformly for $p \in \mathcal{N}_j$ since $|p - x| \le h_j$ holds for all $x \in \operatorname{supp} \lambda_p$ and $p \in \mathcal{N}_j$ and we have $h_j \to 0$ for $j \to \infty$ due to (2.5.19). □

Recall that according to Remark 2.4.3, our Propositions 2.5.15–2.5.17 and Theorem 2.5.18 cover constant Dirichlet boundary conditions which is no restriction in one space dimension if γ_D contains one point.

In contrast to Proposition 2.5.17 the situation in Theorem 2.5.18 appears more complicated in higher dimensions for the case $M : \mathbb{R} \to \mathbb{R}$. Although (2.5.60) is always satisfied here with $C_0 = \|M\|_\infty |\Omega|$, we do not have a Hölder-like inequality

$$|(m_{u_j}, v)_j| \le \|m_{u_j}\|_{L^\infty(\Omega)} \|v\|_{L^1(\Omega)} \quad \forall v \in C^\infty_{\gamma_D}(\overline{\Omega})$$

for the lumped scalar product (with $C^\infty_{\gamma_D}(\overline{\Omega})$ defined in (2.5.23)). Of course, one would like to obtain something stronger like

$$|(m_{u_j}, v)_j| \le C \|m_{u_j}\|_{L^\infty(\Omega)} \|v\|_k \quad \forall v \in C^\infty_{\gamma_D}(\overline{\Omega}) \ \forall j \ge 0 \tag{2.5.64}$$

for a $C > 0$ and $k = 1$ with regard to (2.5.61) and in order to define $(m_{u_j}, \cdot)_j$ for $j \ge 0$ on $H^1_{\gamma_D}(\Omega)$ in the first place (with the usual technique as described in [98, p. 48]). Using the continuity of $I_{S_j} : H^2(\Omega) \to H^2(\Omega)$ for $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, see Ciarlet [27, p. 123], one can prove (2.5.64) for $k = 2$ in two or three space dimensions (and similarly for arbitrary $\mathbb{R}^d$ with some $k > 2$). One could derive (2.5.64) for $k \in \mathbb{N}$ by an interpolation error estimate of the form

$$\left| \int_\Omega m \cdot v - I_{\mathcal{S}_j}(m \cdot v)\, dx \right| \leq C h_j \|m\|_{L^2(\Omega)} \|v\|_k \quad \forall m \in \mathcal{S}_j \;\; \forall v \in C^\infty(\overline{\Omega}) \quad (2.5.65)$$

with a $C > 0$ independent of $j \geq 0$ and the mesh size h_j from (2.5.19). For $k = 1$ the estimate is true for all $m, v \in \mathcal{S}_j$ according to Blowey and Elliot [17, p. 149], and it is clear that a derivation of it for a triangle $t \in \mathcal{T}_j$ instead of Ω would be enough to prove it. Again, using some interpolation theory to be found in Ciarlet [27, pp. 122–124] for $d = 1, 2, 3$ (and with uniform refinement, see the beginning of Subsection 2.7.3, to get C independent of j), we can only prove a version of (2.5.65) where we have h_j^2 instead of h_j but with $L^2(\Omega)$ replaced by $W^{2,\infty}(\Omega)$ and $k = 2$ on the right hand side.

It is clear that (2.5.65) for $k = 1$ would prove (2.5.63), however, both (2.5.64) and (2.5.65) are false for $k = 1$ in more than one space dimension because in case of $\Omega \subset \mathbb{R}^2$ already there are well-known examples of unbounded $H^1(\Omega)$-functions, see for example [19, p. 30]. And one can consider a sequence of small translations of such functions with the singularity around a node p with $m_{u_j}(p) \neq 0$ or $m(p) \neq 0$ (for a fixed $j \geq 0$) in order to see that the left hand side of (2.5.64) or (2.5.65), respectively, can be arbitrarily large while the right hand side remains bounded.

Another interesting task in this context, which is not dealt with here, is an asymptotic convergence analysis for the limit cases discussed in Section 1.4. In concrete terms, one could investigate the behaviour of the solutions u_λ or $u_{\lambda,j}$ as well as u_0 or $u_{0,j}$ of the problems (2.3.23) and (2.5.9), respectively, for the continuous Brooks–Corey functions (or (1.4.21)) with the parameter $\lambda > 0$ and the corresponding discontinuous limit case "$\lambda = 0$". For example, $u_\lambda \to u_0$ in $H^1_{\gamma_D}(\Omega)$ for $\lambda \to 0$ and (2.4.8) would entail strong convergence of $(w_{u_\lambda}, \cdot)_{L^2(\Omega)}$ to $(w_u, \cdot)_{L^2(\Omega)}$ in $H^1_{\gamma_D}(\Omega)'$ for $\lambda \to 0$. On the discrete level $u_{\lambda,j} \to u_{0,j}$ in $\mathcal{S}_j^D$ for $\lambda \to 0$ would provide $m_{u_{\lambda,j}} \to m_{u_{0,j}}$ in $\mathcal{S}_j^D$ for $\lambda \to 0$ in view of (2.5.46) (compare also Remark 2.7.5).

2.6 Nonlinear Gauss–Seidel relaxation

The purpose of this and the next section is to present a numerical solution method for the finite dimensional problem (2.5.9). The presentation including the notation is again based on Kornhuber [59], but also on Kornhuber [58]. As indicated in the Introduction 2.1, our method is not based on regularization but on minimization. Therefore, the scalar function Φ only needs to be piecewise smooth on its domain, which it is in our case of the Brooks–Corey functions, on which we will focus now, with the saturation $M = \Phi'$ of the generalized pressure as given in (1.3.25). Moreover, our method will turn out to be robust with respect to the size of the slope of M. In fact, we can also use it to treat the degenerated situations as in the limit cases in Section 1.4 in which M turns into a maximal monotone multifunction.

2.6.1 General setting and convergence result

To start with, we apply the nonlinear Gauss–Seidel relaxation to solve (2.5.9). Although this well-known method (see Glowinski [45, 142–147] or Kornhuber [59, pp. 45–50]) lacks efficiency, it already has the properties just mentioned for solving (2.5.9). Therefore, it is used as a basic ingredient, a smoother, for the monotone multigrid method which, in addition to the properties of the Gauss–Seidel method, provides the desired efficiency of a solver for (2.5.9) and which we present in the next section.

The nonlinear Gauss–Seidel method results from successive minimization of the convex functional $\mathcal{J}+\phi_j$ in the direction of the nodal basis functions $\lambda_p^{(j)} \in \Lambda_j$, $p \in \mathcal{N}_j$, defined at the beginning of Section 2.5. Here, we chose some ordering of the nodes $p \in \mathcal{N}_j$, i.e. $\mathcal{N}_j = \{p_l\}_{l=1,\dots,n_j}$ with the cardinality $n_j = \#\mathcal{N}_j$ of $\mathcal{N}_j$. For a precise formulation, we introduce the splitting

$$\mathcal{S}_j = \sum_{l=1}^{n_j} V_l \quad \text{with} \quad V_l = \operatorname{span}\{\lambda_{p_l}^{(j)}\} \quad \forall l = 1,\dots,n_j$$

of $\mathcal{S}_j$ into one-dimensional subspaces $V_l \subset \mathcal{S}_j$. Analogously, observe that the special form of the convex set $\mathcal{K}_j$ defined in (2.5.1) allows for the splitting

$$\mathcal{K}_j = \sum_{l=1}^{n_j} \mathcal{K}_{jl} \tag{2.6.1}$$

in which we denote the one-dimensional "traces" as

$$\mathcal{K}_{jl} := \mathcal{K}_j \cap V_l \quad \forall l = 1,\dots,n_j \,.$$

Moreover, for an element $w \in \mathcal{K}_j$ with $w = \sum_{l=1}^{n_j} v_l$, $v_l \in \mathcal{K}_{jl}$ for $l = 1,\dots,n_j$, the definition (2.5.2) of ϕ_j provides the decoupling property

$$\phi_j(w) = \sum_{l=1}^{n_j} \Phi(v_l(p_l))\, h_{p_l} \,. \tag{2.6.2}$$

Now, with a given iterate $w_0^\nu = u_j^\nu \in \mathcal{K}_j$, $\nu \geq 0$, we compute a sequence of intermediate iterates $w_l^\nu = w_{l-1}^\nu + \bar{v}_l^\nu$, $l = 1,\dots,n_j$, by solving the one-dimensional convex minimization problems of finding corrections

$$\begin{aligned} \bar{v}_l^\nu \in V_l \text{ with } w_{l-1}^\nu + \bar{v}_l^\nu \in \mathcal{K}_j :\ & \mathcal{J}(w_{l-1}^\nu + \bar{v}_l^\nu) + \phi_j(w_{l-1}^\nu + \bar{v}_l^\nu) \\ & \leq \mathcal{J}(w_{l-1}^\nu + v) + \phi_j(w_{l-1}^\nu + v) \quad \forall v \in V_l \text{ with } w_{l-1}^\nu + v \in \mathcal{K}_j \,. \end{aligned} \tag{2.6.3}$$

Then we define the next iterate by

$$u_j^{\nu+1} =: \mathcal{M}_j u_j^\nu = w_{n_j}^\nu = u_j^\nu + \sum_{l=1}^{n_j} \bar{v}_l^\nu \,. \tag{2.6.4}$$

Much has been done so far on this kind of relaxation methods. So the following global convergence result is well known. Its proof is heavily based on the two properties (2.6.1) and (2.6.2).

Theorem 2.6.1. *We assume the conditions given in Theorem 2.3.16 apart from the continuity of M. Then for any initial iterate $u_j^0 \in \mathcal{K}_j$, the sequence of iterates $(u_j^\nu)_{\nu\geq 0}$ provided by the nonlinear Gauss–Seidel relaxation (2.6.4) converges to the solution u_j of the discrete problem (2.5.9).*

A proof of a more general theorem can be found in Glowinski [45, 142–147]. However, one might find the approach in Kornhuber [59, pp. 47–49], based on an idea in Mandel [68], more instructive. In order to understand the basic ingredients of this proof, observe first that the corrections given by (2.6.3) are unique and that we have

$$\mathcal{J}(w_l^\nu) + \phi_j(w_l^\nu) \leq \mathcal{J}(w_{l-1}^\nu) + \phi_j(w_{l-1}^\nu) \tag{2.6.5}$$

such that we get equality in (2.6.5) if and only if $w_l^\nu = w_{l-1}^\nu$. Therefore, we obtain the property

$$\mathcal{J}(\mathcal{M}_j w) + \phi_j(\mathcal{M}_j w) \leq \mathcal{J}(w) + \phi_j(w) \tag{2.6.6}$$

for any fixed point w of $\mathcal{M}_j : \mathcal{K}_j \to \mathcal{K}_j$. Furthermore, we show below (see Remark 2.6.3) that $\mathcal{M}_j$ is continuous on $\mathcal{K}_j$. Finally, in light of Proposition 2.5.7, the local problems (2.6.3) can be rewritten as the variational inequalities

$$\begin{aligned}\bar{v}_l^\nu \in V_l \ \text{ with } \ w_{l-1}^\nu + \bar{v}_l^\nu \in \mathcal{K}_j : \ \ & a(w_{l-1}^\nu + \bar{v}_l^\nu, v - \bar{v}_l^\nu) - \ell(v - \bar{v}_l^\nu) \\ & + \phi_j(w_{l-1}^\nu + v) - \phi_j(w_{l-1}^\nu + \bar{v}_l^\nu) \geq 0 \quad \forall v \in V_l \ \text{ with } \ w_{l-1}^\nu + v \in \mathcal{K}_j\end{aligned}$$

which by virtue of (2.6.2) are equivalent to

$$\begin{aligned}\bar{v}_l^\nu \in V_l \ \text{ with } \ w_{l-1}^\nu + \bar{v}_l^\nu \in \mathcal{K}_j : \ \ & a(w_{l-1}^\nu + \bar{v}_l^\nu, v - \bar{v}_l^\nu) - \ell(v - \bar{v}_l^\nu) \\ & + \Phi(w_{l-1}^\nu(p_l) + v(p_l))\, h_{p_l} - \Phi(w_{l-1}^\nu(p_l) + \bar{v}_l^\nu(p_l))\, h_{p_l} \geq 0 \\ & \qquad \forall v \in V_l \ \text{ with } \ w_{l-1}^\nu + v \in \mathcal{K}_j\,.\end{aligned} \tag{2.6.7}$$

Now, the essential steps of the proof given in [59, pp. 47–49] for the homogeneous case can also be applied to our case without further technical problems.

Sketch of the proof for Theorem 2.6.1. Using to the coercivity of $\mathcal{J} + \phi_j$ (see Theorem 2.3.16) as well as the monotonicity (2.6.5), one obtains the boundedness of the sequence $(u_j^\nu)_{\nu\geq 0}$.

Let $u^* \in \mathcal{K}_j$ be the limit of a subsequence $(u_j^{\nu_k})_{k\geq 0}$. The monotonicity (2.6.5) provides

$$\mathcal{J}(u_j^{\nu_{k+1}}) + \phi_j(u_j^{\nu_{k+1}}) \leq \mathcal{J}(\mathcal{M}_j u_j^{\nu_k}) + \phi_j(\mathcal{M}_j u_j^{\nu_k}) \leq \mathcal{J}(u_j^{\nu_k}) + \phi_j(u_j^{\nu_k})\,.$$

With this estimate and the continuity of $\mathcal{M}_j$ and $\mathcal{J} + \phi_j$ on $\mathcal{K}_j$, one concludes (2.6.6) for $w = u^*$, i.e. u^* is a fixed point of $\mathcal{M}_j$.

Now, for any $v_j \in \mathcal{K}_j$, one adds up the inequalities (2.6.7) for $w_0^\nu = w_{l-1}^\nu = u^*$, i.e. with $\bar{v}_l^\nu = 0$, for $l = 1, \dots, n_j$, tested with the one-dimensional interpolations $v = I_{V_l}(v_j - u^*)$. Thus the property (2.6.1) and the special structure (2.6.2) of the functional ϕ_j show that any fixed point of $\mathcal{M}_j$ satisfies the variational inequality (2.5.17) in the form

$$u_j \in \mathcal{K}_j : \quad a(u_j, v_j - u_j) - \ell(v_j - u_j) + \phi_j(v_j) - \phi_j(u_j) \geq 0 \quad \forall v_j \in \mathcal{K}_j \, .$$

This variational inequality is uniquely solvable, i.e. $u^* = u_j$.

Finally, since any convergent subsequence of $(u_j^\nu)_{\nu \geq 0}$ converges to u_j, so does the whole sequence $(u_j^\nu)_{\nu \geq 0}$. □

Remark 2.6.2. In order to keep the notation as simple as possible, nodes $p_l \in \mathcal{N}_j^D$ also contribute corrections in (2.6.3), which are always vanishing of course since we have $\mathcal{K}_{jl} = \{u_D(p_l)\}$ in these cases. In practical computations the values of any iterate at these nodes are always kept fixed as the prescribed Dirichlet values, such that any w_l^ν can be regarded as a suitable w_j in the variational inequality (2.5.17) cited in the proof. Note that these points also contribute to (a constant part) of ϕ_j in (2.5.2).

Observe that we do not obtain global convergence in the sense that we can start with any initial iterate $u_j^0 \in \mathcal{S}_j$ since in this case, the convex functional in the one-dimensional minimization problem (2.6.3) for $l = 1$ might be identically $+\infty$ (i.e. not proper), and so the problem is possibly not solvable (or not uniquely solvable if we replace $\mathcal{K}_j$ by $\mathcal{S}_j$). However, one can choose a canonical candidate for a solution by replacing considering the problems (2.6.7) instead of the equivalent ones given in (2.6.3) and replacing $\mathcal{K}_j$ by $\mathcal{S}_j$ in (2.6.7) (or just by altering (2.6.3) accordingly). These latter subproblems are proper and uniquely solvable for any initial iterate $u_j^0 \in \mathcal{S}_j$ and lead to $u_j^1 \in \mathcal{K}_j$. In this sense, we can define $\mathcal{M}_j : \mathcal{S}_j \to \mathcal{K}_j$ and obtain global convergence on the whole space $\mathcal{S}_j$. This is reflected by the practical treatment of (2.6.3) to which we turn now.

2.6.2 Practical realization of the method

Since we can assume $w_{l-1}^\nu = u_D(p) \; \forall p \in \mathcal{N}_j^D$ for $l = 1$ and $\nu = 0$ and therefore for all $l = 1, \dots, n_j$ and $\nu \geq 0$, in the following, we only consider points

$$p_l \in \mathcal{N}_j \backslash \mathcal{N}_j^D \, .$$

In order to make clear how the corrections $\bar{v}_l^\nu \in V_l$ in (2.6.3) can be computed practically, we observe that in light of Proposition 2.5.5, one Gauss–Seidel step (2.6.3) is equivalent to

$$\bar{v}_l^\nu \in V_l : \quad 0 \in a(w_{l-1}^\nu + \bar{v}_l^\nu, \cdot) - \ell(\cdot) + \partial\phi_j(w_{l-1}^\nu + \bar{v}_l^\nu)(\cdot) + \partial\psi^S(w_{l-1}^\nu + \bar{v}_l^\nu)(\cdot) \quad \text{on } V_l' . \tag{2.6.8}$$

Of course, the subdifferential given on the right-hand side of this variational inclusion is to be understood as a set of functionals on the subspace $V_l \subset \mathcal{S}_j$.

Equivalently (Hahn–Banach), it can be interpreted as the set of restrictions of the given subgradients in $\mathcal{S}'_j$ on V_l. Since V_l is one-dimensional, this set is uniquely determined as a set of numbers (which has to contain $0 \in \mathbb{R}$) provided by (2.6.8) if we insert $\lambda_{p_l}^{(j)}$ with $\lambda_{p_l}^{(j)}(p_l) = 1$.

In this way we can reformulate (2.6.8) as a variational inclusion on the real line, setting $\bar{v}_l^\nu \in V_l$ as

$$\bar{v}_l^\nu := z_l \lambda_{p_l}^{(j)}$$

with the unknown correction factor $z_l \in \mathbb{R}$ (while dropping the dependence on ν for notational reasons). Furthermore, we define the real numbers

$$a_{ll} := a(\lambda_{p_l}^{(j)}, \lambda_{p_l}^{(j)}) \quad \text{and} \quad r_l := \ell(\lambda_{p_l}^{(j)}) - a(w_{l-1}^\nu, \lambda_{p_l}^{(j)}) \tag{2.6.9}$$

as well as the real convex functionals

$$\Phi_l(z) := \phi_j(w_{l-1}^\nu + z\lambda_{p_l}^{(j)}) \quad \forall z \in \mathbb{R} \quad \text{and} \quad \Psi_l^S(z) := \psi^S(w_{l-1}^\nu + z\lambda_{p_l}^{(j)}) \quad \forall z \in \mathbb{R}\,.$$

Applying the "chain rule" for subdifferentials (consult e.g. Ekeland and Temam [34, pp. 27/28]) while considering (2.5.13) we obtain

$$\partial\Phi_l(z) = \partial\phi_j(w_{l-1}^\nu + z\lambda_{p_l}^{(j)})(\lambda_{p_l}^{(j)}) = \partial\Phi(w_{l-1}^\nu(p_l) + z)\, h_{p_l} \quad \forall z \geq u_c - w_{l-1}^\nu(p_l)$$

and (2.5.14) provides analogously

$$\begin{aligned}\partial\Psi_l^S(z) &= \partial\psi^S(w_{l-1}^\nu + z\lambda_{p_l}^{(j)})(\lambda_{p_l}^{(j)}) \\ &= \begin{cases} \partial\chi_{\mathbb{R}_0^-}(w_{l-1}^\nu(p_l) + z) & \forall z \geq -w_{l-1}^\nu(p_l) \quad \text{if } p_l \in \mathcal{N}_j^S \\ 0 & \forall z \in \mathbb{R} \quad \text{if } p_l \in \mathcal{N}_j \backslash (\mathcal{N}_j^D \cup \mathcal{N}_j^S)\,. \end{cases}\end{aligned}$$

We recall from (2.4.2) and (2.4.4) that the maximal monotone multifunction $\partial\Phi : \mathbb{R} \to 2^{\mathbb{R}}$ reads

$$\partial\Phi(y) = \begin{cases} \emptyset & \text{for } y < u_c \\ (-\infty, \lim_{u \downarrow u_c} M(u)] & \text{for } y = u_c \\ \{M(y)\} & \text{for } y \geq u_c \end{cases} \tag{2.6.10}$$

with $M(= \hat{M})$ calculated in (1.3.25). Furthermore, the subdifferential $\partial\chi_{\mathbb{R}_0^-}$ is obviously given by

$$\partial\chi_{\mathbb{R}_0^-}(y) = \begin{cases} 0 & \text{for } y \in \mathbb{R}^- \\ [0, +\infty) & \text{for } y = 0 \\ \emptyset & \text{for } y \in \mathbb{R}^+\,. \end{cases} \tag{2.6.11}$$

Altogether, inserting $\lambda_{p_l}^{(j)}$ in (2.6.8), we obtain the scalar inclusion

$$z_l \in \mathbb{R} : \quad 0 \in a_{ll}\, z_l - r_l + \partial\Phi_l(z_l) + \partial\Psi_l^S(z_l) \tag{2.6.12}$$

as a reformulation of (2.6.8). Of course, there is a version of (2.6.12) in terms of classical derivatives which is given by

$$z_l \in I_l : \quad 0 = a_{ll}\, z_l - r_l + \Phi'(w_{l-1}^\nu(p_l) + z_l)\, h_{p_l} \tag{2.6.13}$$

with $I_l := [u_c - w^\nu_{l-1}(p_l), \infty)$ for $p_l \in \mathcal{N}_j \backslash (\mathcal{N}_j^D \cup \mathcal{N}_j^S)$ and with the additional constraint $I_l := [u_c - w^\nu_{l-1}(p_l), -w^\nu_{l-1}(p_l)]$ for $p_l \in \mathcal{N}_j^S$. This formulation could have been derived with analogous arguments for classical derivatives using Theorem 2.3.11.

In the following, we abbreviate

$$\bar{w}_l := w^\nu_{l-1}(p_l) \in \mathbb{R}$$

for any $p_l \in \mathcal{N}_j \backslash \mathcal{N}_j^D$. Interpreting the right-hand side of (2.6.12) as a maximal monotone graph $\Gamma \subset \mathbb{R}^2$, solving (2.6.12) requires to find the x-coordinate of the intersection point of Γ with the x-axis in $\mathbb{R}^2$. Equivalently, we can determine the intersection point of the real linear function

$$G : x \mapsto -\frac{a_{ll}}{h_{p_l}} x + \frac{r_l}{h_{p_l}} \tag{2.6.14}$$

with the real multifunction

$$x \mapsto \begin{cases} \partial\Phi(\bar{w}_l + x) & \text{for } \; p_l \in \mathcal{N}_j \backslash (\mathcal{N}_j^D \cup \mathcal{N}_j^S) \\ \partial\Phi(\bar{w}_l + x) + \partial\chi_{\mathbb{R}_0^-}(\bar{w}_l + x) & \text{for } \; p_l \in \mathcal{N}_j^S \,. \end{cases}$$

Instead, for simplicity, we introduce the translation

$$y := \bar{w}_l + x \,,$$

now, with (2.6.10) and (2.6.11), considering the multifunction

$$\mathcal{H}_l : y \mapsto \begin{cases} \partial\Phi(y) & \text{for } \; p_l \in \mathcal{N}_j \backslash (\mathcal{N}_j^D \cup \mathcal{N}_j^S) \\ \partial\Phi(y) + \partial\chi_{\mathbb{R}_0^-}(y) & \text{for } \; p_l \in \mathcal{N}_j^S \end{cases} \tag{2.6.15}$$

and the translated linear function

$$G_{-\bar{w}_l} : y \mapsto G(y - \bar{w}_l) \,.$$

Consequently, (2.6.12) can be written as

$$y_l \in \mathbb{R} : \quad G_{-\bar{w}_l}(y_l) = \mathcal{H}_l(y_l) \tag{2.6.16}$$

with $z_l = y_l - \bar{w}_l$.

Observe that the linear functions $G_{-\bar{w}_l}$ are strictly decreasing since both a_{ll} and h_{p_l} are positive numbers. Now, Figure 2.1 shows the concrete solution of (2.6.16) for $p_l \in \mathcal{N}_j \backslash (\mathcal{N}_j^D \cup \mathcal{N}_j^S)$ if we choose M according to our choice of parameter functions due to Brooks and Corey, see (1.3.25) (and more concretely (1.4.1) or Figure 1.12). The additional (vertical) constraint in $y = 0$ which would occur for points $p_l \in \mathcal{N}_j^S$ on the Signorini-type boundary is clear.

We distinguish three cases:

1st case (bottom line g_b): $G_{-\bar{w}_l}(u_c) = G(u_c - \bar{w}_l) \leq \theta_m$

$$\Longrightarrow z_l = u_c - \bar{w}_l \,.$$

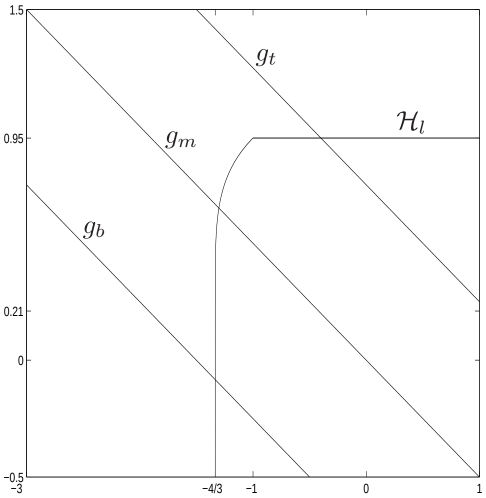


Figure 2.1: Possible intersections of $\mathcal{H}_l$ and $G_{-\bar{w}_l} \in \{g_b, g_m, g_t\}$

2nd case (top line g_t): $G_{-\bar{w}_l}(-1) = G(-1 - \bar{w}_l) \geq \theta_M$

$$\Longrightarrow z_l = \begin{cases} a_{ll}^{-1}(-h_{p_l}\theta_M + r_l) & \text{if } \; p_l \in \mathcal{N}_j \backslash (\mathcal{N}_j^D \cup \mathcal{N}_j^S) \\ \min\{a_{ll}^{-1}(-h_{p_l}\theta_M + r_l), -\bar{w}_l\} & \text{if } \; p_l \in \mathcal{N}_j^S \,. \end{cases}$$

3rd case (middle line g_m): “nontrivial” intersection of M and $G_{-\bar{w}_l}$

$$\Longrightarrow u_c < y_l = \bar{w}_l + z_l < -1 \,.$$

Remark 2.6.3. We solve the third case numerically (up to machine precision) using the bisection method. In order to increase the convergence speed, this can be replaced by Newton’s method once an iterate obtained from the bisection is smaller than y_l. Alternatively, one could solve the third case inexactly in such a way that a damped Gauss–Seidel method is obtained with $\omega_l^\nu \bar{v}_l^\nu$ for instead of $\bar{v}_l^\nu$ in (2.6.3) with $\omega_l^\nu \in [\omega_0, 1]$ for some $\omega_0 \in (0, 1]$. Convergence of such an inexact Gauss–Seidel relaxation is shown in Kornhuber [58, pp. 3/4].

With a glance at Figure 2.1, one can “see” that $z_l = y_l - \bar{w}_l$ depends continuously on $\bar{w}_l$, i.e. $\bar{v}_l^\nu$ depends continuously on w_{l-1}^ν which shows the continuity of $\mathcal{M}_j$ in (2.6.4). More precisely, one can interpret one Gauss–Seidel step as a one-dimensional convex minimization problem in which the functional depends linearly, via the residual, on the fixed coordinates of the unknown vector. Then one can apply iteratively a one-dimensional version of Proposition 2.4.11 in order to see the continuity of $\mathcal{M}_j$.

Furthermore, we note that for $\phi_j = 0$ and no constraints (apart from the Dirichlet values on $\mathcal{N}_j$) the above considerations reduce to the well-known Gauss–Seidel method for the solution of a linear system $Ax = b \in \mathbb{R}^n$ with a symmetrical positive definite matrix $A \in \mathbb{R}^{n^2}$ and $n = n_j - \#\mathcal{N}_j^D$, see for example Glowinski [45, pp. 147/148].

2.7 Monotone multigrid methods

In this section, we present monotone multigrid methods including discrete convergence results for the solution of (2.5.9) in our hydrological setting with Brooks–Corey functions and their limit cases (see Subsection 1.4.2). As already stated at the beginning of Section 2.6, nonlinear Gauss–Seidel relaxations alone do not provide an efficient numerical method for solving problems like (2.5.9).

2.7.1 Monotone coarse grid corrections

In the linear case already, i.e. for $\phi_j = 0$ and no constraints, convergence rates of the Gauss–Seidel method deteriorate (exponentially with j) if one passes to more and more refined (uniform) triangulations $\mathcal{T}_j$, see e.g. Kornhuber and Schütte [60, pp. 115–121]. The well-known reason for this fact is the small support or "high frequency" of the nodal basis functions used for the successive minimization on the (fine) grid. This restricts the information transport for the computation of the correction $\bar{v}_l^\nu = z_l \lambda_{p_l}^{(j)}$ in one Gauss–Seidel step (2.6.3) to the values of w_{l-1}^ν at the neighbouring points of p_l (see (2.6.12) and the definition of the terms used therein). Therefore, the application of $\mathcal{M}_j$ to an iterate u_j^ν only reduces the high frequency contribution of the error $u_j^\nu - u_j$. In order to account for lower frequencies in the error and to accelerate the information transport, one extends the set Λ_j by additional functions (or search directions) with larger support. Therefore, we introduce the ordered subsets

$$M^\nu = (\mu_1^\nu, \dots, \mu_{m_j^\nu}^\nu)$$

of $\mathcal{S}_j$ for any $\nu \geq 0$ in which we assume that

$$\mu_l^\nu = \lambda_{p_l}^{(j)}, \quad l = 1, \dots n_j\,,$$

are the *fine grid* functions and that μ_l^ν, $l = n_j + 1, \dots, m_j^\nu$, are suitably chosen functions with larger support in general.

Now, $(M^\nu)_{\nu \geq 0}$ induces an *extended relaxation method* resulting from the successive minimization of $\mathcal{J} + \phi_j$ in the direction of $\mu_l^\nu \in M^\nu$, $l = 1, \dots, m_j^\nu$, for $\nu = 0, 1, \dots$. Therefore, setting $V_l^\nu = \operatorname{span}\{\mu_l^\nu\}$, we can consider the problems

$$\begin{aligned} \bar{v}_l^\nu \in V_l^\nu \ \text{ with } \ w_l^\nu := w_{l-1}^\nu + \bar{v}_l^\nu \in \mathcal{K}_j :\ & \mathcal{J}(w_{l-1}^\nu + \bar{v}_l^\nu) + \phi_j(w_{l-1}^\nu + \bar{v}_l^\nu) \\ \leq \mathcal{J}(w_{l-1}^\nu + v) + \phi_j(w_{l-1}^\nu + v) \quad & \forall v \in V_l^\nu \ \text{ with } \ w_{l-1}^\nu + v \in \mathcal{K}_j \end{aligned} \tag{2.7.1}$$

and define the next iterate by

$$u_j^{\nu+1} = w_{m_j^\nu}^\nu = w_{n_j}^\nu + \sum_{l=n_j+1}^{m_j^\nu} \bar{v}_l^\nu =: \tilde{\mathcal{C}}_j^\nu \mathcal{M}_j u_j^\nu \tag{2.7.2}$$

with the so-called *smoothed iterate* from (2.6.4) which we denote by

$$\bar{u}_j^\nu := w_{n_j}^\nu = \mathcal{M}_j u_j^\nu\,.$$

In general, due to the form of (the nonlinear) ϕ_j and the constraints in (2.7.1), the exact evaluation of (2.7.1) for $l = n_j + 1, \ldots, m_j^\nu$, i.e. of the *coarse grid correction* $\tilde{\mathcal{C}}_j^\nu \bar{u}_j^\nu$, is too costly for practical calculations. Therefore, much work has been done on the construction of efficient approximations $\mathcal{C}_j^\nu$ of $\tilde{\mathcal{C}}_j^\nu$. It turns out that the only property such an approximation $\mathcal{C}_j^\nu$ has to satisfy in order to maintain the global convergence result (Theorem 2.6.1) for the extended relaxation method is the monotonicity condition

$$\mathcal{J}(\mathcal{C}_j^\nu w) + \phi_j(\mathcal{C}_j^\nu w) \leq \mathcal{J}(w) + \phi_j(w) \quad \forall w \in \mathcal{K}_j\,. \tag{2.7.3}$$

Theorem 2.7.1. *Assume that $\tilde{\mathcal{C}}_j^\nu$ in (2.7.2) is replaced by any coarse grid correction $\mathcal{C}_j^\nu$ satisfying (2.7.3). Then the iteration given by (2.7.2) is globally convergent.*

For the notion of global convergence see Theorem 2.6.1 and Remark 2.6.2. With our knowledge gathered so far the proof of Theorem 2.7.1 is surprisingly easy and unveils how powerful the monotonicity condition (2.7.3) is.

Proof. We only need to replace "monotonicity (2.6.5)" by "monotonicity (2.6.5) and (2.7.3)" in the proof of Theorem 2.6.1. The rest is literally the same. □

Often the monotonicity

$$\mathcal{J}(w_l^\nu) + \phi_j(w_l^\nu) \leq \mathcal{J}(w_{l-1}^\nu) + \phi_j(w_{l-1}^\nu) \quad \forall \nu \geq 0,\ l = 1, \ldots, m_j^\nu$$

is enforced for every one-dimensional correction step such that

$$\mathcal{J}(u_j^{\nu+1}) + \phi_j(u_j^{\nu+1}) \leq \mathcal{J}(w_l^\nu) + \phi_j(w_l^\nu) = \mathcal{J}(u_j^\nu) + \phi_j(u_j^\nu) \quad \forall \nu \geq 0,\ l = 1, \ldots, m_j^\nu$$

together with the continuity of $\mathcal{J} + \phi_j$ on $\mathcal{K}_j$ and the uniqueness of the limit u_j of u_j^ν provide the convergence

$$w_l^\nu \to u_j \quad \text{for} \quad \nu \to \infty,\ l = 1, \ldots, m_j^\nu \tag{2.7.4}$$

of the whole sequence $(w_l^\nu)_{\nu \geq 0,\, l=1,\ldots,m_j^\nu}$ of intermediate iterates.

2.7.2 Constrained Newton linearization with local damping in case of Brooks–Corey parametrization

It is well known that Newton's method does not converge globally, consider for instance finding the zero of the real (strictly increasing) function $\arctan : \mathbb{R} \to \mathbb{R}$ or equivalently the minimum of a (strictly convex) primitive of arctan. However, since the coarse grid correction $\mathcal{C}_j^\nu$ alone does not need to provide a convergent iteration and since our Φ is at least piecewise smooth, some linearization of $\mathcal{J} + \phi_j$ for an approximate solution of (2.7.1) seems feasible. Based on Kornhuber [59] this idea is carried out in Kornhuber [58].

We sketch this idea and the main results in [58] considering the example of the convex function $\Phi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ given by (1.3.27) due to our choice of Brooks–Corey parameter functions and $\Phi(u) = +\infty$ for $u < u_c$ according to (2.4.2) and (2.4.4). In order to incorporate the points in $\mathcal{N}_j$ located on the Signorini-type boundary (as in (2.6.15)) or the Dirichlet boundary, we define the point-dependent convex functions

$$\Phi_p : u \mapsto \begin{cases} \Phi(u) & \text{for } \ p \in \mathcal{N}_j \backslash (\mathcal{N}_j^D \cup \mathcal{N}_j^S) \\ \Phi(u) + \chi_{\mathbb{R}_0^-}(u) & \text{for } \ p \in \mathcal{N}_j^S \\ \chi_{\{u_D(p)\}} & \text{for } \ p \in \mathcal{N}_j^D \end{cases}$$

with the subdifferentials

$$\partial\Phi_p : u \mapsto \begin{cases} \partial\Phi(u) & \text{for } \ p \in \mathcal{N}_j \backslash (\mathcal{N}_j^D \cup \mathcal{N}_j^S) \\ \partial\Phi(u) + \partial\chi_{\mathbb{R}_0^-}(u) & \text{for } \ p \in \mathcal{N}_j^S \\ \partial\chi_{\{u_D(p)\}} & \text{for } \ p \in \mathcal{N}_j^D \end{cases}$$

where

$$\partial\chi_{\{u_D(p)\}} : u \mapsto \begin{cases} \mathbb{R} & \text{if } \ u = u_D(p) \\ \emptyset & \text{if } \ u \neq u_D(p)\,. \end{cases}$$

For $p \in \mathcal{N}_j \backslash \mathcal{N}_j^D$ the function Φ_p is infinitely times differentiable on

$$I_1 := (u_c, -1)$$

and on

$$I_2 := \begin{cases} (-1, \infty) & \text{for } \ p \in \mathcal{N}_j \backslash (\mathcal{N}_j^D \cup \mathcal{N}_j^S) \\ (-1, 0) & \text{for } \ p \in \mathcal{N}_j^S \end{cases} \tag{2.7.5}$$

where it is the identity. It is continuously differentiable on a neighbourhood of -1 but not twice differentable in this point.

We call $p \in \mathcal{N}_j$ a *critical node* of $v \in \mathcal{K}_j$ if $p \in \mathcal{N}_j^D$ or

$$v(p) \in C_p := \begin{cases} \{u_c, -1\} & \text{for } \ p \in \mathcal{N}_j \backslash (\mathcal{N}_j^D \cup \mathcal{N}_j^S) \\ \{u_c, -1, 0\} & \text{for } \ p \in \mathcal{N}_j^S \end{cases}$$

and define $\mathcal{N}_j^\bullet(v)$ as the set of critical nodes of v. We call the elements of C_p critical values and the elements of $\mathbb{R}\backslash C_p$ regular values for Φ_p. Accordingly, we define $\mathcal{N}_j^\circ(v) := \mathcal{N}_j \backslash \mathcal{N}_j^\bullet(v)$ as the set of *regular nodes* of v.

We point out that the results given in [58] can be generalized to our situation of point-dependent Φ_p and more than one critical value. In fact, the results in [59], where finitely many critical values are treated for piecewise quadratic Φ, can be merged with the ones in [58], where Φ is only required to be twice differentiable on the interior of its domain with locally Lipschitz continuous Φ''. According to [59, p. 59], further classification of the regular nodes of v with respect to connected subsets of regular values is helpful. Therefore, we define the *discrete phases* of v by

$$\mathcal{N}_j^i(v) := \{p \in \mathcal{N}_j : v(p) \in I_i\}\,, \quad i = 1, 2\,. \tag{2.7.6}$$

Now, for a given smoothed iterate $\bar{u}_j^\nu$ and each regular node $p \in \mathcal{N}_j^\circ(\bar{u}_j^\nu)$, we can find real numbers

$$\underline{\varphi}_{\bar{u}_j^\nu}(p) < \bar{u}_j^\nu(p) < \overline{\varphi}_{\bar{u}_j^\nu}(p)$$

such that on the neighbourhood $[\underline{\varphi}_{\bar{u}_j^\nu}(p), \overline{\varphi}_{\bar{u}_j^\nu}(p)]$ of $\bar{u}_j^\nu(p)$ the function Φ_p is twice differentiable with

$$|\Phi_p''(z_1) - \Phi_p''(z_2)| \le L_p^\nu |z_1 - z_2| \quad \forall z_1, z_2 \in [\underline{\varphi}_{\bar{u}_j^\nu}(p), \overline{\varphi}_{\bar{u}_j^\nu}(p)] \tag{2.7.7}$$

and a pointwise Lipschitz constant $L_p^\nu > 0$. More concretely, if $p \in \mathcal{N}_j^1(\bar{u}_j^\nu)$ we set

$$[\underline{\varphi}_{\bar{u}_j^\nu}(p), \overline{\varphi}_{\bar{u}_j^\nu}(p)] := [(u_c + \bar{u}_j^\nu(p))/2, (\bar{u}_j^\nu(p) - 1)/2] \tag{2.7.8}$$

and if $p \in \mathcal{N}_j^2(\bar{u}_j^\nu)$ we choose

$$[\underline{\varphi}_{\bar{u}_j^\nu}(p), \overline{\varphi}_{\bar{u}_j^\nu}(p)] := \begin{cases} [(-1 + \bar{u}_j^\nu(p))/2, 2|\bar{u}_j^\nu(p)| + 1] & \text{for } p \in \mathcal{N}_j \backslash (\mathcal{N}_j^D \cup \mathcal{N}_j^S) \\ [(-1 + \bar{u}_j^\nu(p))/2, \bar{u}_j^\nu(p)/2] & \text{for } p \in \mathcal{N}_j^S\,. \end{cases} \tag{2.7.9}$$

Setting in addition

$$\underline{\varphi}_{\bar{u}_j^\nu}(p) = \overline{\varphi}_{\bar{u}_j^\nu}(p) = \bar{u}_j^\nu(p)$$

for $p \in \mathcal{N}_j^\bullet(\bar{u}_j^\nu)$, we introduce the closed and convex set

$$\mathcal{K}_{\bar{u}_j^\nu} := \{w \in \mathcal{S}_j : \underline{\varphi}_{\bar{u}_j^\nu}(p) \le w(p) \le \overline{\varphi}_{\bar{u}_j^\nu}(p) \;\; \forall p \in \mathcal{N}_j\}\,. \tag{2.7.10}$$

Then, due to the special form of ϕ_j given by (2.5.2), we obtain the representation

$$\phi_j(w) = \phi_{\bar{u}_j^\nu}(w) + \text{const.} \quad \forall w \in \mathcal{K}_{\bar{u}_j^\nu} \tag{2.7.11}$$

with the smooth functional

$$\phi_{\bar{u}_j^\nu} : w \mapsto \sum_{p \in \mathcal{N}_j^\circ(\bar{u}_j^\nu)} \Phi(w(p))\, h_p \quad \forall w \in \mathcal{K}_{\bar{u}_j^\nu}\,. \tag{2.7.12}$$

Even though $\mathcal{K}_{\bar{u}_j^\nu}$ is not a neighbourhood of $\bar{u}_j^\nu$, we can regard $\phi_{\bar{u}_j^\nu}$ as given on an open neighbourhood of $\bar{u}_j^\nu$ since the critical nodes of $\bar{u}_j^\nu$ do not contribute to $\phi_{\bar{u}_j^\nu}$.

Now, we consider the *constrained minimization* of the *smooth energy* $\mathcal{J} + \phi_{\bar{u}_j^\nu}$:

$$u_{\bar{u}_j^\nu} \in \mathcal{K}_{\bar{u}_j^\nu} : \quad \mathcal{J}(u_{\bar{u}_j^\nu}) + \phi_{\bar{u}_j^\nu}(u_{\bar{u}_j^\nu}) \le \mathcal{J}(v) + \phi_{\bar{u}_j^\nu}(v) \quad \forall v \in \mathcal{K}_{\bar{u}_j^\nu}\,. \tag{2.7.13}$$

Replacing $\phi_{\bar{u}_j^\nu}$ by a Taylor expansion of second order around $\bar{u}_j^\nu$, one obtains a functional $\mathcal{J}_{\bar{u}_j^\nu}$ as the quadratic approximation of $\mathcal{J} + \phi_{\bar{u}_j^\nu}$ around $\bar{u}_j^\nu$. The quadratic obstacle problem

$$w_{\bar{u}_j^\nu} \in \mathcal{K}_{\bar{u}_j^\nu} : \quad \mathcal{J}_{u_{\bar{u}_j^\nu}}(w_{\bar{u}_j^\nu}) \le \mathcal{J}_{\bar{u}_j^\nu}(v) \quad \forall v \in \mathcal{K}_{\bar{u}_j^\nu} \tag{2.7.14}$$

can then be regarded as a *constrained Newton linearization* of (2.7.13).

Now, we choose search directions μ_l^ν, $l = n_j + 1, \ldots, m_j^\nu$, possibly depending on $\mathcal{K}_{\bar{u}_j^\nu}$, and replace the nonlinear relaxation steps (2.7.1) by the constrained one-dimensional linear problems

$$v_l^\nu \in \mathcal{D}_l^\nu : \quad \mathcal{J}_{\bar{u}_j^\nu}(w_{l-1}^\nu + v_l^\nu) \le \mathcal{J}_{\bar{u}_j^\nu}(w_{l-1}^\nu + v) \quad \forall v \in \mathcal{D}_l^\nu \tag{2.7.15}$$

with constraints $\mathcal{D}_l^\nu \subset V_l^\nu$ satisfying

$$0 \in \mathcal{D}_l^\nu \subset \{v \in V_l^\nu : w_{l-1}^\nu + v \in \mathcal{K}_{\bar{u}_j^\nu}\}\,. \tag{2.7.16}$$

Since $\mathcal{J}_{\bar{u}_j^\nu}$ is a quadratic approximation of $\mathcal{J} + \phi_{\bar{u}_j^\nu}$ around $\bar{u}_j^\nu$ and $\phi_{\bar{u}_j^\nu}$ is convex, the following is clear for $l = n_j + 1$, i.e. for $w_{l-1}^\nu = \bar{u}_j^\nu$. If $v_l^\nu \neq 0$ in (2.7.1), then v_l^ν is a direction in which $\mathcal{J} + \phi_{\bar{u}_j^\nu}$ is (also) decreasing (at least) locally around the iterate w_{l-1}^ν. Consequently, (2.7.1) holds if the next iterate

$$w_l^\nu = w_{l-1}^\nu + \omega_l^\nu v_l^\nu \tag{2.7.17}$$

is chosen with a suitable damping parameter $\omega_l^\nu \in (0, 1]$. Of course this does not have to be true for $l > n_j + 1$. (Considering quadratic expansions of $\phi_{\bar{u}_j^\nu}$ around all intermediate iterates w_l^ν, $l = n_j + 1, \ldots, m_j^\nu$, would in general lead to suboptimal numerical complexity, see also (2.7.26) and Kornhuber [58, p. 15].) However, in [58, pp. 8, 10] an upper bound is available for $w_l^\nu \in [0, 1]$ and any $\|\mu_l^\nu\|_\infty = 1$, $l = n_j + 1, \ldots, m_j^\nu$, $\nu \ge 0$, which guarantees the monotonicity (2.7.1) for the choice of w_l^ν according to (2.7.17). For the computation of a suitable ω_l^ν only (local) properties on $\operatorname{supp} \mu_l^\nu$ are required and we have

$$\omega_l^\nu \to 0 \qquad \text{for} \quad \max_{p \in \operatorname{int} \operatorname{supp} \mu_l^\nu} L_p \to \infty\,. \tag{2.7.18}$$

Choosing the *local damping parameters* ω_l^ν according to this bound, one obtains a *monotone coarse grid correction*

$$\mathcal{C}_j^\nu \bar{u}_j^\nu = \bar{u}_j^\nu + \sum_{l=n_j+1}^{m_j^\nu} \omega_l^\nu v_l^\nu \tag{2.7.19}$$

with local damping which preserves global convergence due to Theorem 2.7.1. Such a damped version of an extended relaxation is also called *extended underrelaxation.*

2.7.3 Standard and truncated multigrid: asymptotic convergence result

Monotone multigrid methods now provide realizations of coarse grid corrections $\mathcal{C}_j^\nu$, first with optimal numerical complexity, i.e. with $\mathcal{O}(n_j)$ point operations for one iteration step

$$u_j^{\nu+1} = \mathcal{C}_j^\nu \mathcal{M}_j u_j^\nu$$

and secondly, with a considerable acceleration of the convergence $u_j^\nu \to u_j$ for $\nu \to \infty$.

Again, we sketch the main definitions and results about monotone multigrid methods for our special setting and refer to Kornhuber [59] and [58] for further details. The only aspects of the method described above that still need to be specified is the concrete choice of search directions μ_l^ν and the corresponding constraints $\mathcal{D}_l^\nu$ for $l = n_j + 1, \dots, m_j^\nu$ and $\nu \geq 0$.

For convenience assume that a sequence of (nested) triangulations $\mathcal{T}_0$, $\mathcal{T}_1, \dots, \mathcal{T}_j$ of Ω resulting from uniform refinement is at hand, i.e. each triangle $t \in \mathcal{T}_k$ is subdivided into four congruent subtriangles constituting $\mathcal{T}_{k+1}$, $k = 0, \dots, j-1$. This also gives nested sets of nodes $\mathcal{N}_0 \subset \cdots \subset \mathcal{N}_j$ and a nested sequence $\mathcal{S}_0 \subset \cdots \subset \mathcal{S}_j$ of subspaces of $\mathcal{S}_j$ which correspond to the levels $k = 0, \cdots, j$. Let

$$\Lambda_S := (\lambda_{p_1}^{(j)}, \dots, \lambda_{p_{n_j}}^{(j)}, \lambda_{p_1}^{(j-1)}, \dots, \lambda_{p_{n_{j-1}}}^{(j-1)}, \dots, \lambda_{p_1}^{(0)} \dots, \lambda_{p_{n_0}}^{(0)})$$

be the multilevel nodal basis consisting of all $m_S = n_j + \cdots + n_0$ nodal basis functions from all refinement levels, ordered from fine to coarse, and set

$$\mu_l^\nu = \lambda_{p_l}^{(k_l)}, \quad l = n_j + 1, \dots, m_j^\nu = n_j + m_S\,,$$

also in an ordering from fine to coarse. Note that with this choice, the nodal basis functions on the finest grid also enter the coarse grid correction. This is reasonable due to the different nature of the problems (2.6.3) and (2.7.15) for these search directions.

In order to obtain optimal numerical complexity, the resulting algorithm should allow for an implementation as a multigrid V-cycle in which calculations of corrections on a level $k \in \{0, \dots, j\}$ should only require to access information on nodes $p \in \mathcal{N}_k$. Therefore, the fine grid obstacles

$$\underline{\varphi}_{\bar{u}_j^\nu} - w_{l-1}^\nu \leq v \leq \overline{\varphi}_{\bar{u}_j^\nu} - w_{l-1}^\nu$$

for an admissible correction v in (2.7.15) are enforced by approximated local coarse grid obstacles $\underline{\psi}_l^\nu, \overline{\psi}_l^\nu \in V_l^\nu$ which satisfy

$$\underline{\varphi}_{\bar{u}_j^\nu}(p) - w_{l-1}^\nu(p) \leq \underline{\psi}_l^\nu(p) \leq 0 \leq \overline{\psi}_l^\nu(p) \leq \overline{\varphi}_{\bar{u}_j^\nu}(p) - w_{l-1}^\nu(p) \quad \forall p \in \mathcal{N}_j\,.$$

We refer to [59, pp. 74–76] for inductive constructions of *quasioptimal restrictions* which provide such obstacles. Now, the local constraints $\mathcal{D}_l^\nu$ in the local problems (2.7.15) are given by

$$\mathcal{D}_l^\nu := \{v \in V_l^\nu : \underline{\psi}_l^\nu \leq v \leq \overline{\psi}_l^\nu\}\,, \quad l = n_j + 1, \dots, m_j^\nu\,.$$

Altogether, these selections of search directions μ_l^ν and constraints $\mathcal{D}_l^\nu$ lead to a ν-independent coarse grid correction $\mathcal{C}_j^{\mathrm{std}} = \mathcal{C}_j^\nu$ of the *standard monotone multigrid method*

$$u_j^{\nu+1} = \mathcal{C}_j^{\mathrm{std}} \mathcal{M}_j u_j^\nu\,, \quad \nu \geq 0\,. \tag{2.7.20}$$

A drawback of this method is that, due to the definition of $\mathcal{K}_{\bar{u}_j^\nu}$, any μ_l^ν with

$$\operatorname{int}\operatorname{supp}\mu_l^\nu \cap \mathcal{N}_j^\bullet(\bar{u}_j^\nu) \neq \emptyset \tag{2.7.21}$$

gives a trivial correction $v_l^\nu = 0$. This can exclude many coarse grid search directions and therefore affect the information transport provided by low-frequency functions, which was the reason for setting up extended relaxations in the first place. Therefore, *truncated multilevel nodal bases* $\tilde{\Lambda}_S^\nu$ have been defined. They depend on the set $\mathcal{N}_j^\bullet(\bar{u}_j^\nu)$, i.e. on $\nu \geq 0$, and are based on the idea of appropriately "cutting" any $p \in \mathcal{N}_j^\bullet(\bar{u}_j^\nu)$ "out" of $\operatorname{int}\operatorname{supp}\lambda_p^{(k)}$ for any nodal basis function $\lambda_p^{(k)} \in \Lambda_S$. More concretely, for $k = 0, \dots, j$ and $\nu \geq 0$, the modified basis functions

$$\tilde{\lambda}_p^{(k)} := T_{j,k}^\nu \lambda_p^{(k)}$$

on level k are defined for any $p \in \mathcal{N}_k$ with the truncation operators

$$T_{j,k}^\nu := I_{\mathcal{S}_j^\nu} \circ \cdots \circ I_{\mathcal{S}_k^\nu}$$

arising from the interpolations $I_{\mathcal{S}_k^\nu} : \mathcal{S}_j \to \mathcal{S}_k^\nu$ on the reduced subspaces

$$\mathcal{S}_k^\nu := \{v \in \mathcal{S}_k : v(p) = 0 \ \ \forall p \in \mathcal{N}_k^\nu\}$$

of $\mathcal{S}_k$ induced by the critical nodes $\mathcal{N}_k^\nu = \mathcal{N}_k \cap \mathcal{N}_j^\bullet(\bar{u}_j^\nu)$ on level k. The modified basis functions $\tilde{\lambda}_p^{(k)}$ now constitute the truncated multilevel nodal basis

$$\tilde{\Lambda}_{\mathcal{S}}^\nu := (\tilde{\lambda}_{p_1}^{(j)}, \dots, \tilde{\lambda}_{p_{n_j}}^{(j)}, \tilde{\lambda}_{p_1}^{(j-1)}, \dots, \tilde{\lambda}_{p_{n_{j-1}}}^{(j-1)}, \dots, \tilde{\lambda}_{p_1}^{(0)} \dots, \tilde{\lambda}_{p_{n_0}}^{(0)})$$

and vanish if and only if all nodes $p \in \operatorname{int}\operatorname{supp}\tilde{\lambda}_p^{(k)}$ are critical nodes of $\bar{u}_j^\nu$. We point out that this also gives additional search directions for $p \in \mathcal{N}_j^D$ if the Dirichlet boundary γ_D is not resolved on the coarse grid given by $\mathcal{N}_0$ (as it is the case in the numerical example in Section 4.3).

As above, we define the search directions

$$\mu_l^\nu = \tilde{\lambda}_{p_l}^{(k_l)}, \quad l = n_j + 1, \dots, m_j^\nu = n_j + m_S^\nu ,$$

ordered again from fine to coarse. As in (2.6.4) for the search directions λ_p with $p \in \mathcal{N}_j^D$, the elements of $\tilde{\Lambda}_S^\nu$ which are equal to 0 can be skipped.

The local coarse grid constraints D_l^ν, $l = n_j+1, \dots, m_j^\nu$, $\nu \geq 0$, can be obtained analogously as in the standard case by quasioptimal restrictions, see [59, p. 81].

As a result of these search directions μ_l^ν and constraints $\mathcal{D}_l^\nu$ we obtain ν-dependent coarse grid corrections $\mathcal{C}_j^{\mathrm{trc},\nu} = \mathcal{C}_j^\nu$ and the *truncated monotone multigrid method*

$$u_j^{\nu+1} = \mathcal{C}_j^{\mathrm{trc},\nu} \mathcal{M}_j u_j^\nu , \quad \nu \geq 0 . \tag{2.7.22}$$

It is observed that truncated monotone multigrid usually converges more rapidly than the standard one. Unfortunately, this is not yet reflected by the analytical results which cannot guarantee better asymptotic convergence rates for the truncated version, see Theorem 2.7.4 (or consult [58, pp. 17–19] and [59, pp. 87–92] for more details). The difficulty for the analysis lies in the fact that the truncated basis functions $\tilde{\lambda}_p^{(k)}$ can have a rather strange shape and are in general not contained in with $\mathcal{S}_k$. Despite this, it may surprise that truncated monotone multigrid methods can be implemented as multigrid V-cycles with only minor changes as compared to the standard ones, see [59, p. 82].

Asymptotic bounds of convergence rates demonstrating the fast convergence of the monotone multigrid methods described above can be proved for problems (2.5.9) whose solution u_j satisfies the *non-degeneracy condition*

$$\ell(\lambda_p^{(j)}) - a(u_j, \lambda_p^{(j)}) \in \operatorname{int} \partial\phi_j(u_j)(\lambda_p^{(j)}) + \partial\psi^S(u_j)(\lambda_p^{(j)}) \quad \forall p \in \mathcal{N}_j^{\bullet}(u_j) \backslash \mathcal{N}_j^D\,. \tag{2.7.23}$$

This condition guarantees that all the critical nodes of u_j and their (exact) values are detected by the nonlinear Gauss–Seidel method after a finite number of iterations. (The values on $\mathcal{N}_j^D$ are known anyway.)

Lemma 2.7.2. *Assume that the solution u_j of the discrete minimization problem (2.5.9) is non-degenerate in the sense of (2.7.23). Then there is a $\nu_0 \geq 0$ such that*

$$w_l^\nu(p) = u_j(p) \quad \forall p \in \mathcal{N}_j^{\bullet}(u_j) \tag{2.7.24}$$

and

$$\mathcal{N}_j^i(w_l^\nu) = \mathcal{N}_j^i(u_j) \tag{2.7.25}$$

holds for all $l = 1, \ldots, m_j^\nu$ and $\nu \geq \nu_0$.

Remark 2.7.3. The proof of this result is based on continuity arguments and the convergence (2.7.4) of the intermediate iterates. It is mainly a consequence of the nonlinear Gauss–Seidel smoother $\mathcal{M}_j$ since critical nodes remain untouched in the coarse grid correction. Observe that the essential assertion of Lemma 2.7.2 is (2.7.24) and that (2.7.25) follows from (2.7.4).

Figure 2.1 displays the abstract arguments of the proof and the relevance of (2.7.23) for the example of the critical value u_c: Small perturbations of the bottom line g_b do not change the x-coordinate u_c of its intersection with $\operatorname{int}\partial\Phi(u_c)$. With a glance at the coefficients constituting the line g_b, one can see that this demonstrates (2.7.23), on a scalar level, as a stability condition for critical nodes $\mathcal{N}_j^{\bullet}(u_j)$ with respect to small perturbations of u_j (consider u_j or small perturbations of it as coming from a Gauss–Seidel iteration). Observe that if (2.7.23) does not hold for the critical value $0 = u_j(p)$ at a critical node $p \in \mathcal{N}_j^S$, then $u_j(p)$ assumes this value whether the obstacle condition $u_j(p) \leq 0$ at this node is imposed or not. Finally, we remark that the condition in (2.7.23) can obviously be never fulfilled in case of any critical point p in which u_j assumes the value -1 where Φ is differentiable but not "smooth enough". We come back to this phenomenon in more detail in Subsection 2.7.4.

As a consequence of Lemma 2.7.2, of our choice of the obstacles defining $\mathcal{K}_{\bar{u}_j^\nu}$ in (2.7.10) and the convergence (2.7.4) we conclude that $u_j \in \mathcal{K}_{\bar{u}_j^\nu}$ holds for all $\nu \geq \nu_1$ with some $\nu_1 \geq 0$. Now, since u_j minimizes $\mathcal{J} + \phi_j$ on $\mathcal{K}_j$ and $\phi_{\bar{u}_j^\nu}$ from (2.7.11) only differs by a constant from ϕ_j on $\mathcal{K}_{\bar{u}_j^\nu}$, which is contained in $\mathcal{K}_j$ for all $\nu \geq 0$, the solution u_j of the non-smooth problem (2.5.9) solves the constrained smooth problem (2.7.13).

With this crucial observation and the local Lipschitz continuity (2.7.7) of Φ'' one can prove that there is a $\nu_2 \geq 0$ such that the constrained Newton linearization (2.7.14) is equivalent to classical Newton linearization of (2.7.13) at $\bar{u}_j^\nu$ for $\nu \geq \nu_2$. The latter becomes a linear problem for which the local obstacle problems (2.7.15) asymptotically (for $\nu \geq \nu_3$ and a $\nu_3 \geq 0$) provide an extended linear underrelaxation. This property can be guaranteed with the special choice of the coarse grid constraints $\mathcal{D}_l^\nu$ described in [59, pp. 71–76, 81]. Now, with the technical assumption that all non-zero corrections v_l^ν have the property

$$\|v_k^\nu\|_\infty^2 = o(\|v_l^\nu\|_\infty) \quad \text{for } \nu \to \infty\,, \quad k = n_j + 1, \dots, l-1\,, \tag{2.7.26}$$

there is a $\nu_4 \geq 0$ such that the damping parameters w_l^ν in (2.7.19) satisfy $\omega_l^\nu = 1$ for all $\nu \geq \nu_4$. (2.7.26) accounts for the fact that derivatives of $\phi_{\bar{u}_j^\nu}$ are not evaluated at the intermediate iterates w_l^ν, $\nu \geq 0$, $l = n_j + 1, \dots, m_j^\nu$.

Altogether, one concludes that our coarse grid corrections $\mathcal{C}_j^\nu$ asymptotically become iterations of linear (Newton) multigrid methods or more generally linear subspace correction methods for which convergence rates can be derived. These results can be combined with an asymptotic error estimate for the nonlinear Gauss–Seidel smoothing (2.6.4) for which (2.7.7) is exploited. The asymptotic convergence rates are based on the local energy norm $\|\cdot\|_{u_j}$ defined by

$$\|v\|_{u_j}^2 := a(v,v) + \phi''_{u_j}(u_j)(v,v) \quad \forall v \in \mathcal{S}_j^\circ \tag{2.7.27}$$

on the reduced space

$$\mathcal{S}_j^\circ := \{v \in \mathcal{S}_j : v(p) = 0 \ \ \forall p \in \mathcal{N}_j^\bullet(u_j)\}\,.$$

Observe that by Lemma 2.7.2 the errors $u_j - u_j^\nu$ are elements of $\mathcal{S}_j^\circ$ for $\nu \geq \nu_0$. We assume that

$$\|v\|_{u_j} \leq \gamma_j \|v\| \quad \forall v \in \mathcal{S}_j^\circ \tag{2.7.28}$$

is satisfied for the energy norm $\|\cdot\|$ given by

$$\|v\|^2 := a(v,v) \quad \forall v \in \mathcal{S}_j^\circ\,.$$

Theorem 2.7.4. *Assume that the non-degeneracy condition (2.7.23) as well as (2.7.26) and (2.7.28) is satisfied. Then there is a $\nu_j \geq 0$ such that the iterates produced by the standard (2.7.20) or the truncated (2.7.22) monotone multigrid method fulfill the error estimate*

$$\|u_j - u_j^{\nu+1}\|_{u_j} \leq (1 - c\gamma_j^{-1}(j+1)^{-4})\|u_j - u_j^\nu\|_{u_j} \quad \forall \nu \geq \nu_j$$

with a positive constant c depending only on the ellipticity of $a(\cdot,\cdot)$ and on the initial triangulation $\mathcal{T}_0$.

2.7.4 Practical treatment and nature of the critical values

We proceed with some further comments on the critical nodes which are also of interest in the practical computation. As a matter of fact, the three possible critical values u_c, -1 and 0 for a Φ_p are quite different in nature.

The critical value 0 is imposed as an obstacle condition for nodes on the Signorini-type boundary whereas the critical value u_c occurs as a singularity of Φ'' with $\Phi''(u) \to \infty$ and $\Phi'''(u) \to \infty$ for $u \downarrow u_c$. Therefore, the local Lipschitz condition (2.7.7) deteriorates for regular nodes p with values $\bar{u}_j^\nu(p)$ in a small neighbourhood of u_c in the sense that L_p^ν explodes. Now, due to (2.7.18), this produces small damping parameters, i.e. hardly any contribution of search directions μ_l^ν with $p \in \operatorname{int}\operatorname{supp}\mu_l^\nu$ to the coarse grid correction of $\bar{u}_j^\nu$. In this way, such nodes play a similar role as critical nodes in (2.7.21). Therefore, it seems reasonable to consider such nodes as critical, too, and leave their values $\bar{u}_j^\nu(p)$ untouched in the coarse grid correction. Skipping all μ_l^ν with $p \in \operatorname{int}\operatorname{supp}\mu_l^\nu$ as the corresponding search directions obviously does not violate the global convergence.

In concrete terms, given $v \in \mathcal{K}_j$ we define

$$\mathcal{N}_j^*(v) := \{p \in \mathcal{N}_j : p \in \mathcal{N}_j^\bullet(v) \wedge L_p^\nu > L\} \tag{2.7.29}$$

with a threshold $L > 0$ as an extended set of critical nodes of v and replace $\mathcal{N}_j^\bullet(\bar{u}_j^\nu)$ by $\mathcal{N}_j^*(\bar{u}_j^\nu)$ for the definition (2.7.10) of $\mathcal{K}_{\bar{u}_j^\nu}$. If L is large enough, then we have $\mathcal{N}_j^*(u_j) = \mathcal{N}_j^\bullet(u_j)$ and $\mathcal{N}_j^*(w_l^\nu) = \mathcal{N}_j^\bullet(w_l^\nu) = \mathcal{N}_j^\bullet(u_j)$ for $l = n_j, \dots, m_j^\nu$ and $\nu \geq \bar{\nu}_0$ with a $\bar{\nu}_0 \geq 0$ due to (2.7.4) and Lemma 2.7.2. Consequently, a suitable choice of L preserves the asymptotic convergence result in Theorem 2.7.4.

In our practical computations (as in Section 2.8), we determine the solution $\tilde{u}_c > u_c$ of $\Phi'''(u) - L = 0$ (see (1.3.27)) and choose

$$\underline{\varphi}_{\bar{u}_j^\nu}(p) := \tilde{u}_c$$

for any regular node $p \in \mathcal{N}_j^1(\bar{u}_j^\nu)$ instead of (2.7.8). As a consequence, (2.7.7) holds with the global Lipschitz constant $L := \Phi'''(\tilde{u}_c) = L_p^\nu$, uniformly for all regular nodes $p \in \mathcal{N}_j \backslash \mathcal{N}_j^*(\bar{u}_j^\nu)$. Note that some balance has to be kept in the choice of L versus $\tilde{u}_c$ in order to get reasonable corrections. On the one hand, the larger L (i.e. the smaller $\tilde{u}_c$) is, the smaller the damping parameters in (2.7.19) are due to (2.7.18). On the other hand, the bigger $\tilde{u}_c$ (i.e. the smaller L) is, bigger damping parameters might be useless since in this case the constraints $\mathcal{K}_{\bar{u}_j^\nu}$ and $\mathcal{D}_l^\nu$ become more restrictive.

With regard to the other critical values observe that even though Φ'' does not exist in -1, we still have existing one-sided limits $\Phi''_-(-1) := \lim_{u\uparrow -1}\Phi''(u) \in \mathbb{R}^+$ and $\Phi''_+(-1) := \lim_{u\downarrow -1}\Phi''(u)(=0)$ as well as $\Phi''_-(0) := \lim_{u\uparrow 0}\Phi''(u)(=0)$ in contrast to $\lim_{u\downarrow u_c}\Phi''(u) = \infty$ for the singularity u_c. Using these real one-sided limits in the endpoints of the intervals in (2.7.9) we can simplify the choice of these intervals altogether and

$$\textit{replace}\ \ [\underline{\varphi}_{\bar{u}_j^\nu}(p), \overline{\varphi}_{\bar{u}_j^\nu}(p)]\ \ \textit{by phase-dependent intervals} \tag{2.7.30}$$

as, for example, by $[\tilde{u}_c, -1]$ if $p \in \mathcal{N}_j^1(\bar{u}_j^\nu)$, by $[-1, 0]$ for $p \in \mathcal{N}_j^2(\bar{u}_j^\nu) \cap \mathcal{N}_j^S$ and by $[-1, \infty)$ for $p \in \mathcal{N}_j^2(\bar{u}_j^\nu) \backslash \mathcal{N}_j^S$. With the definitions of $\Phi''(u)$ as the suitable one-sided limits of Φ'' in the endpoints of these intervals, property (2.7.7) still holds and the analysis in Kornhuber [58] can be carried out analogously leading to the same results. We use these modifications in our practical calculations in Section 2.8.

Observe that there is no "naturally occurring" obstacle corresponding to the value -1, which is critical just because Φ is not smooth enough in -1 (for our analysis), even though Φ' exists and is continuous in -1. Again, we point out that this is a special feature of the Brooks–Corey parameter functions (1.2.9) and (1.2.10) in which a "sharp" bubbling pressure p_b is incorporated. Other parameter functions, e.g. according to van Genuchten [91], do not have this property and are hydrologically reasonable, too.

However, for the parametrization according to the Brooks–Corey functions, one could argue that nodes referring to the critical value -1 (which corresponds to the bubbling pressure p_b in (1.2.9), see Section 1.3) mark the free boundary of the waterfront separating the saturated from the unsaturated region. By Lemma 2.7.2 the nodes on the free boundary of Signorini's type are found after a finite number of nonlinear Gauss–Seidel steps (guaranteeing the asymptotic error estimates) if the non-degeneracy condition (2.7.23) is satisfied. This certainly applies to all critical nodes, but condition (2.7.23) cannot hold for nodes p referring to the critical value -1 because Φ is "too regular" at this point such that $u_j(p)$ is always unstable under small perturbations of u_j (see also Remark 2.7.3). So u_j can only be non-degenerate in the sense of (2.7.23) if

$$u_j(p) \neq -1 \quad \forall p \in \mathcal{N}_j \backslash \mathcal{N}_j^D \tag{2.7.31}$$

is satisfied, which is sensible from a numerical point of view. Nevertheless, together with our considerations on (2.7.29) above, this motivates the definition of even more extended sets of critical nodes in order to "get rid" of the non-degeneracy condition (and therefore (2.7.31)) by introducing an *ε-non-degeneracy.*

For $\varepsilon > 0$ we define

$$\mathcal{N}_j^{\varepsilon}(v) := \{p \in \mathcal{N}_j : |v(p) - c_p| \leq \varepsilon \text{ for a } c_p \in C_p\} \tag{2.7.32}$$

with the set C_p of all critical values of Φ_p. As C_p is finite, c_p is unique in (2.7.32) if ε is sufficiently small. We call the solution u_j of (2.5.9) *ε-non-degenerate* if

$$\mathcal{N}_j^{\varepsilon}(u_j) = \mathcal{N}_j^{\bullet}(u_j)\,. \tag{2.7.33}$$

Since u_j is an element of a finite-dimensional space, it is always ε-non-degenerate for some $\varepsilon > 0$ even if it is non-degenerate. Due to (2.7.4) there is also a $\nu_0 \geq 0$ such that $\mathcal{N}_j^{\varepsilon/2}(w_l^{\nu}) = \mathcal{N}_j^{\varepsilon}(u_j)$ is satisfied for all $l = 1, \ldots, m_j^{\nu}$ and $\nu \geq \nu_0$. Consequently, if we choose $\varepsilon > 0$ small enough and $\nu_0 \geq 0$ large enough while setting $w_l^{\nu}(p) = c_p$ for $p \in \mathcal{N}_j^{\varepsilon/2}(w_l^{\nu})$ according to (2.7.32) for all $l = 1, \ldots, m_j^{\nu}$ and $\nu \geq \nu_0$, the assertions (2.7.24) and (2.7.25) of Lemma 2.7.2 hold and the asymptotic convergence analysis can be applied to the corresponding algorithm with the same results.

2.7.5 Monotone multigrid for the limit cases

So far, our focus concerning the numerical treatment of (2.3.23) in this and in the previous section was on the special situation according to Brooks and Corey in (1.3.25). However, the Gauss–Seidel method as well as the monotone multigrid method in Kornhuber [59] and [58] can be applied to much more general cases. For the arguments in this section to hold, it is obviously enough to impose the following conditions on $M : I \to \mathbb{R}$, given on a nontrivial interval $I \subset \mathbb{R}$ with $0 \in I$. First, the function M should be monotonically increasing and bounded (or else satisfying (2.3.25)). Secondly, M should be continuously differentiable on the interior of finitely many subintervals I_k, $k = 1, \ldots, n$, of I, whose union is I, such that M' is still Lipschitz continuous on compact subsets of $\operatorname{int} I_k$ for each $k = 1, \ldots, n$ (compare [59, pp. 23/24] and [58, p. 1]). These conditions include the nondegenerate case and all the limit cases discussed in Section 1.4.

For the nondegenerate case in (1.4.21) and its limit cases, we do not have a lower obstacle $u_c < 0$, thus the convex function $\Phi : \mathbb{R} \to \mathbb{R}$ given by (2.3.9) is defined on the whole real line. In (1.4.21) the value u_α occurs as an additional critical value (due to lack of regularity of M) which plays a similar role as the critical value -1 above, but does not have a physical meaning other than perhaps as an upper bound for "unrealisticallly small" (generalized) pressure values. In the limit case (1.4.25) we obtain a piecewise linear Φ on $\mathbb{R}$ with only one critical value -1 (apart from 0 for the Signorini-type boundary), for which (in contrast to -1 in the Brooks–Corey case) $\partial\Phi = \tilde{M}$ is setvalued with $\tilde{M}(-1) = [\theta_m, \theta_M]$. Note that in the limit case $p_b/p_0 \to 0$ in (1.4.26) the critical value -1 is replaced by 0 such that we only have one critical value here.

The numerical treatment of the (hydrologically reasonable) limit case for the Brooks–Corey parameter functions treated in this section is not only possible, but it is in fact much easier than what has been done above. In Remark 2.4.5 it is discussed how the limit cases (1.4.6) with (1.4.7) and (1.4.13) can be regarded as the same situation, in which a linear Φ or, equivalently, a constant $\Phi' = M \equiv \theta_M$ on an interval $[u_c, \infty)$ with $u_c \leq 0$ resulting in

$$\partial\Phi(u) = \tilde{M}(u) = \begin{cases} \emptyset & \text{for } u < u_c \\ (-\infty, \theta_M] & \text{for } u = u_c \\ \theta_M & \text{for } u > u_c \end{cases} \tag{2.7.34}$$

constitutes a linear constrained problem. We obtained $u_c \in \{-2, -1, 0\}$ in Section 1.4, the value 0 coming from $p_b/p_0 \to 0$. The effect of this latter case is that the boundary conditions of Signorini's type now turn into homogeneous Dirichlet boundary conditions which makes the situation even more simple. Otherwise, the value of $u_c < 0$, which we discuss in the following, does not matter.

With (2.7.34) instead of (2.6.10) the construction in Figure 2.1 degenerates and becomes very easy. Furthermore, we only have the critical values u_c and 0 (for the Signorini-type boundary) which constitute the intervals I_2 as in (2.7.5) (with -1 replaced by u_c) for the discrete phases $\mathcal{N}_j^2(v)$ as in (2.7.6). Accordingly, the constraints in (2.7.30) can be replaced by the closures of these intervals, on which we have vanishing L_p^ν due to $\Phi'' \equiv 0$, such that we can choose the constraints in

$$\mathcal{K}_{\bar{u}_j^\nu} = \mathcal{K}_j$$

independently of $\bar{u}_j^\nu$ with the full convex set $\mathcal{K}_j$ from (2.5.1). Since ϕ_j is linear on $\mathcal{K}_j$, the constrained Newton linearization (2.7.14) is a constrained problem for the original quadratic functional

$$\mathcal{J}_{\bar{u}_j^\nu} = \mathcal{J} + \phi_{\bar{u}_j^\nu} = \mathcal{J} + \phi_j \qquad (2.7.35)$$

on $\mathcal{K}_j$, and so are the one-dimensional problems (2.7.15). The constraints $\mathcal{D}_l^\nu$ are chosen as above and damping is not necessary due to the identity (2.7.35). As a consequence, we end up with a usual multigrid method for a linear constrained problem. The truncated version can be used as above without the adjustments described in Section 2.7.4 for exploding Φ''.

2.7.6 Convergence of the iterates for the saturation, the physical pressure and the generalized saturation

Finally, as in Subsection 2.5.4, we remark that the convergence $u_j^\nu \to u_j$ in $\mathcal{K}_j$ for $\nu \to \infty$ entails

$$I_{\mathcal{S}_j} M(u_j^\nu) \to I_{\mathcal{S}_j} M(u_j) \quad \text{in } \mathcal{S}_j \text{ for } \nu \to \infty \qquad (2.7.36)$$

for the sequence of discrete saturations $(I_{\mathcal{S}_j} M(u_j^\nu))_{\nu \geq 0}$ if $M : \mathbb{R} \to \mathbb{R}$ or, alternatively, $M : [u_c, \infty) \to \mathbb{R}$ is uniformly continuous. This is obvious if one considers the equivalent $\|\cdot\|_\infty$-norm on $\mathcal{S}_j$ and it completes our convergence results given in Propositions 2.5.11 and 2.5.12 as well as in Theorem 2.5.13. Note that the map $M_{\mathcal{S}_j} : \mathcal{S}_j \to \mathcal{S}_j$ defined by

$$M_{\mathcal{S}_j}(v) := I_{\mathcal{S}_j} M(v) \quad \forall v \in \mathcal{S}_j \qquad (2.7.37)$$

is Hölder continuous with respect to $\alpha \in (0,1]$ if $M : \mathbb{R} \to \mathbb{R}$ or $M : [u_c, \infty) \to \mathbb{R}$ is. In this case we also obtain asymptotic geometric convergence

$$\|M_{\mathcal{S}_j}(u_j) - M_{\mathcal{S}_j}(u_j^\nu)\|_\infty \leq C \|u_j - u_j^{\nu_j}\|_{u_j} (\rho_j^\alpha)^{\nu - \nu_j} \quad \forall \nu \geq \nu_j \qquad (2.7.38)$$

on S_j for some $C > 0$ and with $\rho_j := (1 - c\gamma_j^{-1}(j+1)^{-4}) \in (0,1)$ from Theorem 2.7.4.

With regard to the iterates $p_j^\nu := I_{\mathcal{S}_j} \kappa^{-1}(u_j^\nu)$ of the discrete physical pressure $I_{\mathcal{S}_j} p_j = I_{\mathcal{S}_j} \kappa^{-1}(u_j)$ the same restrictions concerning the singularity of κ^{-1} in u_c apply here as already discussed in Remark 2.5.14. If, however, a physically realistic situation with $u_j(p) > u_c$ or equivalently $u_j(p) > u_c + \varepsilon$ with an $\varepsilon > 0$ for all $p \in \mathcal{N}_j$ is given, then the Lipschitz continuity of κ^{-1} on $[u_c + \varepsilon, \infty)$ provides the same convergence results for the sequence $(p_j^\nu)_{\nu \geq 0} = (I_{\mathcal{S}_j} \kappa^{-1}(u_j^\nu))_{\nu \geq 0}$ and its limit $I_{\mathcal{S}_j} p_j = I_{\mathcal{S}_j} \kappa^{-1}(u_j)$ as we have just obtained in (2.7.36) and (2.7.38) for the sequence $(I_{\mathcal{S}_j} M(u_j^\nu))_{\nu \geq 0}$ and its limit $I_{\mathcal{S}_j} M(u_j)$.

If $M : \mathbb{R} \to \mathbb{R}$ or $M : [u_c, \infty) \to \mathbb{R}$ is discontinuous as in the limit cases (1.4.25) and in Remark 2.4.5, we can associate the generalized saturation $m_{u_j^\nu} \in \mathcal{S}_j$ defined by

$$m_{u_j^\nu}(p) := h_p^{-1}(\ell(\lambda_p) - a(u_j^\nu, \lambda_p)) \quad \forall p \in \mathcal{N}_j \backslash \mathcal{N}_j^D \tag{2.7.39}$$

(and with fixed prescribed values for $p \in \mathcal{N}_j^D$) to the iterate $u_j^\nu \in \mathcal{K}_j$, analogously as for u_j in (2.5.46). Now, we can interpret (2.7.39) in concrete terms if we assume that u_j^ν is obtained from a Gauss–Seidel iteration (2.6.4). In this case the number $m_{u_j^\nu}(p)$ for the node $p = p_l$ corresponding to the last Gauss–Seidel step (2.6.3) occurs as the second coordinate of the intersection point of the line in $\mathbb{R}^2$, given by the right hand side of (2.7.39) (compare (2.6.9) and (2.6.14)), with the monotone graph $\mathrm{gr}(\tilde{M})$ as depicted in Figure 2.1. The convergence $u_j^\nu \to u_j$ in $\mathcal{K}_j$ for $\nu \to \infty$ provides

$$|m_{u_j}(p) - m_{u_j^\nu}(p)| \leq h_p^{-1} |a(u_j - u_j^\nu, \lambda_p)| \leq C \|u_j - u_j^\nu\|_1 \, h_p^{-1} \|\lambda_p\|_1 \to 0$$

for $\nu \to \infty$ and any $p \in \mathcal{N}_j \backslash \mathcal{N}_j^D$. Again, with Theorem 2.7.4 we can conclude the asymptotic convergence

$$\|m_{u_j} - m_{u_j^\nu}\|_\infty \leq C \|u_j - u_j^{\nu_j}\|_{u_j} \, \rho_j^{\nu - \nu_j} \quad \forall \nu \geq \nu_j$$

on S_j for a $C > 0$ and with ρ_j as in (2.7.38). Strangely enough, this general result guarantees better asymptotic convergence rates than (2.7.38) for Hölder continuous M. Note, however, that $m_{u_j^\nu}(p) < M(u_j^\nu(p))$ or $m_{u_j}(p) < M(u_j(p))$ is possible for $u_j^\nu(p) = u_c$ or $u_j(p) = u_c$ as discussed for (2.5.46) in Section 2.5.

2.8 Numerical results in 2D

The purpose of this section is to demonstrate the robustness and the efficiency of our spatial solver, the truncated monotone multigrid method from Section 2.7, when applied to the Richards equation. Since we have not yet dealt with the issue of how to treat the gravitational term, we consider the Richards equation in (1.2.7) with no gravity, i.e. with $g = 0$. Recall that neglecting this term does not restrict the difficulties in the spatial problems (2.3.2) or (2.3.8) arising after our implicit–explicit time discretization because the gravitational term "disappears" in the linear functional $\ell(\cdot)$ by the time discretization.

Concretely, based on the Brooks–Corey parameter functions, we consider the Kirchhoff–transformed Richards equation in two space dimensions in the form (1.3.18) in which the last product accounting for the gravitation is skipped. In the pressure unit $\varrho g z_0$ in (1.3.15) we still choose the density $\varrho = 10^3 \, [kg/m^3]$ of the water and the gravitational constant $g = 10 \, [m/s^2]$ in order to obtain the usual hydraulic conductivity K_h given in (1.3.14). Altogether, with $\Omega \subset \mathbb{R}^2$ and $T > 0$ given below, we consider the equation

$$n\, t_0^{-1} M(u)_t - K_h \Delta u = 0 \quad \text{on } \Omega \times [0,T] \tag{2.8.1}$$

in which the porosity $n = 0.4$ and $K_h = 2 \cdot 10^{-3}\,[m/s]$ are chosen in a realistic range, see Subsection 1.4.1. Consequently, the time unit is $t_0 = 200\,[s]$. The spatial units are $x_0 = z_0 = 1\,[m]$ and the scaling factor in (1.3.15) is chosen to be constant $u_r = 1$. With this choice, varying bubbling pressures p_b are dealt with according to Remark 1.4.1. We give the pressure values with the unit $[m]$ in terms of water column level under earth's gravitation conditions. Our standard bubbling pressure is $p_b = -0.1\,[m]$ and the pore size distribution factor is set as $\lambda = 1$ which is realistic according to Subsection 1.4.1.

The domain Ω is chosen to be the triangle in $\mathbb{R}^2$ given by the vertices $(0,0)$, $(2,0)$ and $(0,2)$. Recall from (1.6.3)–(1.6.6) that an initial saturation M_0 has to be given on Ω in order to obtain a well-posed problem. With the ball

$$B := \{x \in \mathbb{R}^2 : |x| \le 1.38\}$$

given by the Euclidean norm $|\cdot|$ on $\mathbb{R}^2$, we choose an extreme saturation

$$M_0(x) = \begin{cases} \theta_M := 1 & \text{for } x \in \Omega \cap B \\ \theta_m := 0 & \text{for } x \in \Omega \backslash B\,. \end{cases} \tag{2.8.2}$$

The initial condition is depicted in Figure 2.2. The set of points between $(0,0)$ and $(0,1.38)$ on the y-axis including $(0,0)$ and $(0,1.38)$ is chosen to be the Dirichlet boundary γ_D which is constant in time. On γ_D we impose Dirichlet boundary conditions equal to the pressure a water column between the x-axis and the parallel line through $(0,1.38)$ would generate in case of existing gravity. Consequently, with the atmospheric pressure $0\,[m]$, we have

$$u_D(x,y,t) = (1.38 - y) \quad \forall (x,y) \in \gamma_D\,.$$

We impose homogeneous Neumann boundary conditions on the edge between the endpoints $(0,0)$ and $(2,0)$ on the x-axis and also between the points $(0,1.38)$ and $(0,2)$ on the y-axis, again constant in time. The hypotenuse between $(2,0)$ and $(0,2)$ including these points is the Signorini-type boundary γ_S. Note that the Dirichlet condition and the initial condition are compatible according to (1.3.25), but that the initial condition on $\Omega \backslash B$ corresponds to $u = u_c$, i.e. to the unphysical pressure $p = -\infty$.

The initial triangulation $\mathcal{T}_0$ of Ω, which resolves the Dirichlet boundary, is given by the triangles t_1 and t_2 with the vertices $(0,0)$, $(1,1)$ and $(0,1.38)$ or $(1.38,0)$, respectively, and the two remaining triangles $\Omega \backslash (t_1 \cup t_2)$, see Figure 2.3. The refined triangulations $\mathcal{T}_j$, $j = 1, \dots, 7$, are obtained by uniform refinement, indicated on page 113, and result in 33024 unknowns on the finest level $j = 7$ with the mesh size $h_j = 1.38/128 \approx 0.01$.

For the iterative solution of the discretized Signorini-type problems (2.5.44) or, equivalently, (2.5.9) on each refinement level $j = 0, \dots, 7$, we use the truncated

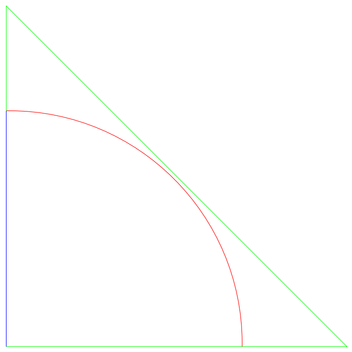


Figure 2.2: Initial condition

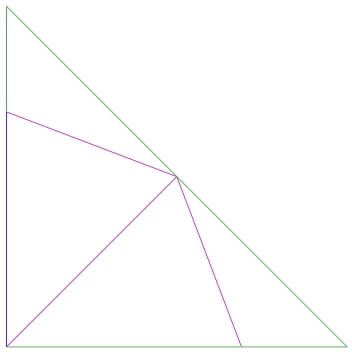


Figure 2.3: Initial triangulation

monotone multigrid V-cycle as described in Section 2.7, but with 3 presmoothing and 3 postsmoothing steps. The latter means that the relaxation procedures per level within the multigrid method as in Section 2.7 are repeated twice on each level and are then carried out backwards from the coarsest to the in the same manner. This whole process is repeated twice and gives the $V(3,3)$-cycle which is intended to increase the convergence speed.

We apply the modifications in Remark 2.7.4, in particular we take into account the threshold Lipschitz constant $L = 10^{12}$ in (2.7.29). Remarkably, $\tilde{u}_c$ with $\Phi'''(\tilde{u}_c) = L$ can be determined explicitly due to the special form of the Brooks–Corey parametrization, see (1.3.27). The size of L is not adapted for variations of p_b because the scaling factor $u_r = 1$ in (1.3.18) is always kept fixed (see Remark 1.4.1) such that the factor in front of Δu in (2.8.1) does not change either.

Since the solution from the previous time step usually serves as a good initial iterate for the spatial problem at the next time step, nested iteration as described in Kornhuber [57, p. 7] is not necessary in general. However, we carry out numerical studies by variations of the parameters. Therefore, we use nested iteration in order to obtain suitable initial iterates throughout. As a stopping criterion for the multigrid on level j we accept the approximate solution $\tilde{u}_j = u_j^{\nu^*}$ as soon as the relative accuracy condition

$$\|u_j^{\nu^*} - u_j^{\nu^*-1}\|_{u_j^{\nu^*}} \leq 10^{-12} \|u_j^{\nu^*}\|_{u_j^{\nu^*}} \tag{2.8.3}$$

is satisfied. Here, $\|\cdot\|_{u_j^{\nu^*}}$ is intended to approximate the local energy norm $\|\cdot\|_{u^{\nu^*}_j}$ and is defined analogously as in (2.7.27).

Note that the treatment of the constrained linear problems (2.7.15) on the algebraic level requires the stiffness matrix $(a(\lambda_p, \lambda_q))_{p,q\in\mathcal{N}_j}$ as well as the additional vector $b := (\phi'_{\bar{u}_j^\nu}(\bar{u}_j^\nu)(\lambda_p))_{p\in\mathcal{N}_j}$ added to the right hand side and the matrix $D := (\phi''_{\bar{u}_j^\nu}(\bar{u}_j^\nu)(\lambda_p, \lambda_q))_{p,q\in\mathcal{N}_j}$. Due to the special form of $\phi_{\bar{u}_j^\nu}$ in (2.7.12), the entries of b are $M(\bar{u}_j^\nu(p))\, h_p$ for $p \in \mathcal{N}_j^\circ(\bar{u}_j^\nu)$ and 0 otherwise, and D is a diagonal matrix with the nontrivial entries $M'(\bar{u}_j^\nu(p))\, h_p$ for $p = q \in \mathcal{N}_j^\circ(\bar{u}_j^\nu)$. Therefore, the norm $\|\cdot\|_{u_j^{\nu^*}}$ given according to (2.7.27) can be computed easily.

Restrictions and prolongations of the matrices and the residual are carried out as stated in Kornhuber [59, pp. 77, 82].

We also refer to the numerical example of a stationary porous medium equation treated in Kornhuber [58]. It contains a similar nonlinearity as $M : [u_c, \infty) \to \infty$ considered here, however, with a discontinuity for the critical value -1. (Nevertheless, recall from (2.7.31) that this situation can be regarded as "less degenerate" than the one considered here, where M is continuous but non-differentiable in -1.) As in [58] the implementation for our case was carried out in the framework of the finite element toolbox KASKADE [14].

With a glance at the initial condition (2.8.2) observe that the interior of $\Omega \cap B$ is still connected such that all of γ_S is completely unsaturated at the beginning. In the time development, for which we choose the constant time step size $\tau = 0.1 \, (\hat{=} 20 \, [s])$ and the end time $T = 1 \, (\hat{=} 200 \, [s])$, free boundaries are detected on γ_S and water flows out of the domain across parts of γ_S until Ω is fully saturated.

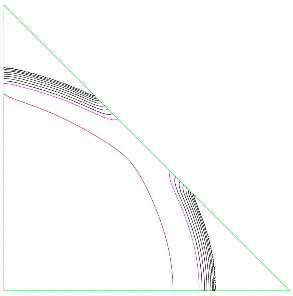


Figure 2.4: $t = 0.1$

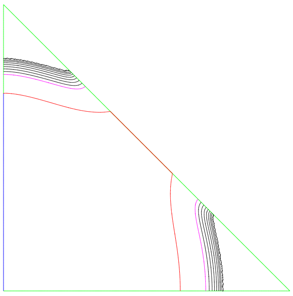


Figure 2.5: $t = 0.2$

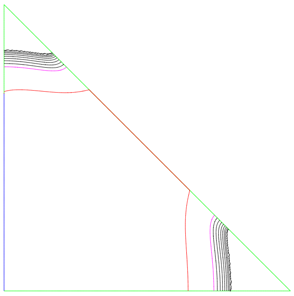


Figure 2.6: $t = 0.3$

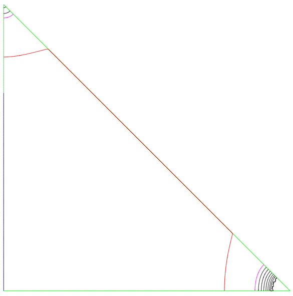


Figure 2.7: $t = 0.8$

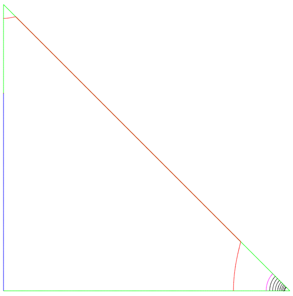


Figure 2.8: $t = 0.9$

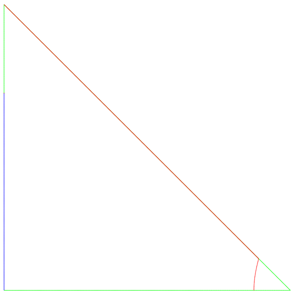


Figure 2.9: $t = 1.0$

Figures 2.8–2.8 show the development of the pressure p or the saturation $\theta(p)$ in time. In these graphics the innermost (red) line in each triangle is the isoline $p = 0$ and the succeeding (magenta) line is the isoline $p = p_b$. The latter marks the border of the fully saturated region outside of which both pressure and saturation decrease. All the following (black) lines are isolines on which constant equidistant saturation values $\theta(p) < \theta_M = 1$ occur, which correspond to constant pressure values $p < p_b$.

Our set of parameters is chosen in such a way that in the first time step a nontrivial part of γ_S, where $p \geq p_b$ occurs, is already fully saturated but not yet with $p \geq 0$. In the second time step, a free boundary is detected on γ_S separating the part where $p < 0$ prevails from the part where we have $p \geq 0$, and outflow is possible.

Apart from the fixed data given at the beginning of this section, the following set of model parameters (with the units as introduced above) is the basis of our numerical results.

1.	end time	$T = 0.1$
2.	time step size	$\tau = 0.1$
3.	maximal refinement level	$j = 7$
4.	pore size distribution factor	$\lambda = 1$
5.	bubbling pressure	$p_b = -0.1$

In the following, we present average convergence rates ρ of the truncated multigrid and the numbers of multigrid iterations $\#it \, (= \nu^*)$ until (2.8.3) is satisfied. The convergence rates are based on the relative errors with respect to the previous iterates. We determine the average convergence rate ρ by the geometric mean of the relative errors, i.e. we set

$$\rho = \left(\frac{\|u_j^{\nu^*} - u_j^{\nu^*-1}\|_{u_j^{\nu^*}}}{\|u_j^1 - u_j^0\|_{u_j^{\nu^*}}} \right)^{\frac{1}{\nu^*-1}} .$$

Our intention is to show that both ρ and $\#it$ do not degenerate if a parameter from the above set is varied. The following Tables 2.1–2.7 show the results we have obtained.

T	ρ	$\#it$
0.1	0.273	18
0.2	0.288	18
0.3	0.295	18
0.4	0.324	19
0.5	0.317	19
0.6	0.353	21
0.7	0.363	22
0.8	0.338	20
0.9	0.328	20
1.0	0.202	14

Table 2.1: Variations of T

τ	ρ	$\#it$
0.001	0.223	16
0.01	0.186	14
0.05	0.244	16
0.1	0.273	18
0.25	0.383	23
0.5	0.438	27
0.75	0.528	34
0.9	0.576	40
0.95	0.288	18
1.0	0.288	18

Table 2.2: Variations of $\tau = T$

Table 2.1 refers to the time development discussed above and depicted in Figures 2.8–2.8. Here, ρ and $\#it$ are given for each time step $T = 0.1, 0.2, \ldots, 1.0$. In Table 2.2 the time step size τ (which is set equal to the end time T in

j	ρ	$\#it$
1	0.780	5
2	0.496	10
3	0.104	12
4	0.196	16
5	0.251	18
6	0.192	14
7	0.273	18
8	0.392	24

Table 2.3: Variations of the maximal refinement level j

this case) is varied. With regard to variations of the spatial step size, i.e. the mesh size h_j, Table 2.3 shows the results for different maximal refinement levels $j = 1, 2, \ldots, 8$.

Tables 2.4–2.7 provide the most interesting results since they refer to variations of the Brooks–Corey parameter functions which are intensely discussed in Section 1.4, especially regarding their big slopes. First, Tables 2.4 and 2.5 display the effects which variations of the soil parameters λ and p_b (the latter given in the unit $[m]$) in their realistic ranges have on ρ and $\#it$. According to Rawls et al. [77, Table 5.3.2], we have $\lambda \in [0.1, 0.7]$ (and up to $\lambda \in [0.037, 1.090]$ including standard deviations) and $p_b \in [-0.4, -0.1]$ (with deviations up to $p_b \in [-1.872, -0.0136]$) in naturally occurring situations. Furthermore, Tables 2.6 and 2.7 show the asymptotic behaviour for decreasing or increasing soil parameters, i.e. referring to extreme shapes of the parameter functions, most importantly $M : [u_c, \infty) \to \mathbb{R}$, "near" the limit cases derived in Subsection 1.4.2.

The results demonstrate the robustness and the efficiency of our spatial solver. Yet, we have found some extreme cases documented in Tables 2.4 and 2.5, in which the convergence results are comparatively unsatisfactory. For $\lambda = 0.1$, for example, the isoline $p = p_b$ touches the boundary of Signorini's type. For $p_b = -1.8$ there is only a small neighbourhood around the vertex $(0, 1)$ in which Ω is not fully saturated, and for $p_b = -4.0$ the same situation occurs around the vertex $(1, 0)$ (this is also the case for $T = 0.9$ in Table 2.1). These situations have in common that small pertubations of the corresponding parameter change the (topological) shape of the unsaturated regime considerably. Therefore, the convergence results could be worse here than on the average because of the sensitivity of the solution rather than of the solver.

Finally, the convergence results for λ and p_b in an asymptotic range in Tables 2.6 and 2.7 make clear that the performance of the solver is not restricted by big slopes of $M : [u_c, \infty) \to \mathbb{R}$. Altogether, the initial iterates obtained from nested iteration usually seem to be sufficiently good so that the asymptotic linear convergence guaranteed by Theorem 2.7.4 governs most of the iteration history and provides convergence rates that are independent of Φ.

λ	ρ	$\#it$
0.01	0.384	23
0.05	0.457	28
0.09	0.511	34
0.1	0.584	41
0.105	0.526	34
0.2	0.401	25
0.3	0.328	21
0.4	0.265	17
0.5	0.332	21
0.6	0.248	17
0.7	0.294	19
0.8	0.267	17
0.9	0.264	17
1.0	0.273	18
1.25	0.260	17
1.5	0.252	16
1.75	0.249	16
2.0	0.248	16
2.5	0.232	16
3.0	0.237	16

Table 2.4: λ in a realistic range

p_b	ρ	$\#it$
-0.005	0.235	16
-0.01	0.248	16
-0.05	0.237	16
-0.1	0.273	18
-0.2	0.268	18
-0.3	0.299	19
-0.4	0.310	20
-0.5	0.342	22
-0.75	0.433	28
-1.0	0.523	37
-1.25	0.643	52
-1.5	0.683	61
-1.7	0.756	81
-1.8	0.810	112
-1.9	0.643	52
-2.0	0.470	30
-2.5	0.564	39
-3.0	0.619	47
-4.0	0.786	94
-5.0	0.274	17

Table 2.5: p_b in a realistic range

λ	ρ	$\#it$
10^{-10}	0.270	17
10^{-9}	0.270	17
10^{-8}	0.270	17
10^{-7}	0.270	17
10^{-6}	0.270	17
10^{-5}	0.258	17
10^{-4}	0.260	17
10^{-3}	0.282	18
10^{1}	0.253	16
10^{2}	0.376	22
10^{3}	0.400	23
10^{4}	0.469	28
10^{5}	0.292	18
10^{6}	0.282	18
10^{7}	0.278	17
10^{8}	0.278	17
10^{9}	0.278	17
10^{10}	0.282	18

Table 2.6: λ in an asymptotic range

p_b	ρ	$\#it$
-10^{-10}	0.169	13
-10^{-9}	0.170	13
-10^{-8}	0.170	13
-10^{-7}	0.169	13
-10^{-6}	0.168	13
-10^{-5}	0.321	19
-10^{-4}	0.294	18
-10^{-3}	0.299	18
-10^{1}	0.274	17
-10^{2}	0.278	15
-10^{3}	0.275	13
-10^{4}	0.270	11
-10^{5}	0.263	9
-10^{6}	0.234	7
-10^{7}	0.268	5
-10^{8}	0.300	6
-10^{9}	0.302	6
-10^{10}	0.302	6

Table 2.7: p_b in an asymptotic range

Chapter 3

Steklov–Poincaré theory for domain decomposition problems with jumping nonlinearities

3.1 Introduction and overview

In the first chapter of this work, mainly in Subsections 1.5.3 and 1.5.4, we provided weak formulations of a Signorini-type problem for the Richards equation and its Kirchhoff–transformed version in homogeneous soil. The second chapter was devoted to a numerical treatment of such a problem if we ignore gravity. However, we already pointed out there that our implicit–explicit time discretization of the Richards equation leads to spatial problems which were discussed generally enough in the last chapter to cover a treatment of the gravitational term by an upwind technique which we present in Chapter 4. The same applies to the next two chapters which are motivated by the question of how we can treat the Richards equation in heterogeneous soil. We only consider spatial problems in a heterogeneous setting which might have arisen from time-dependent problems after a suitable time discretization. Moreover, we do not restrict ourselves to the time-discretized Richards equation although the starting point (in Section 3.2) and the most important result of this chapter (Theorem 3.4.23) are concerned with it.

As already indicated in Section 1.6, there seems to be a lack of analysis for the Richards equation in a heterogeneous setting so far. Moreover, to our knowledge, a numerical treatment of the Richards equation in heterogeneous soil has only been carried out in Fuhrmann [39], Fuhrmann and Langmach [41] as well as in Bastian et al. [10], however, with a solver that is not robust with respect to deteriorating soil parameters (see Section 2.2). By contrast, our approach presented in Chapter 2 provides such a solver in the homogeneous setting (see Sections 2.6 and 2.7). Recall, however, that the Kirchhoff transformation, on which our approach is based, does not lead to semilinear problems in the heterogeneous case (see Remark 1.3.1).

It should be pointed out at this stage that we consider the setting given by the Richards equation (1.2.7) still as homogeneous, even though the permeability $K(\cdot)$ is allowed to depend on $x \in \Omega$. In this form the spatial problems arising from the Richards equation after the implicit–explicit time discretization can be treated just as described in the last chapter. The heterogeneous case comes into play if we also consider the parameter functions $\theta(\cdot)$ and $kr(\cdot)$ as explicitly dependent on space since the choice of these functions depends on the soil type that occurs in Ω. Therefore, what we call heterogeneous soil in the following will be a setting in which different subdomains Ω_i, $i = 1, \ldots, n$, of Ω contain different soil types which are themselves homogeneous in each Ω_i in the sense just described. But then the question arises how a partial differential equation with nonlinearities which have jumps across the interfaces between two subdomains can be interpreted or given sense at all.

In Section 3.2 we give a weak formulation of and therefore a meaning to a Signorini-type problem for the Richards equation in such a heterogeneous setting. It consists of a multi-domain formulation which is motivated by an equivalence result that holds in the homogeneous case. This formulation is a generalization of a result from the linear theory in Quarteroni and Valli [75] and gives rise to a Dirichlet–Neumann algorithm for the Richards equation which we also present.

In Section 3.3 we discuss a Dirichlet–Neumann method for a nonlinear transmission problem on two subdomains which includes the special case of a stationary Richards equation without gravity in heterogeneous soil with nondegenerating parameter functions (as in Subsection 1.4.3). Despite the heterogeneity, the main idea for our further treatment of this domain decomposition problem is the same as in the homogeneous case, namely the application of the Kirchhoff transformation, now independently in the subdomains. We point out that the equivalence of the reformulated problem with the original one depends strongly on the results we obtained in Subsection 1.5.4, in which the Kirchhoff transformation on a domain or on (parts of) a boundary was investigated in the framework of superposition operators. The reformulated substructuring problem can then be analysed on the basis of the linear Steklov–Poincaré theory, to be found in Quarteroni and Valli [75], which we extend to our nonlinear case. Doing so, we find sufficient conditions for the convergence of the nonlinear Dirichlet–Neumann method which turn out to be satisfied in one space dimension. Counterexamples show that these conditions need not hold in higher dimensions. However, numerical computations suggest that the algorithm can also be applied successfully to higher-dimensional problems.

Section 3.4 is devoted to an analysis of Robin's method applied to a larger class of nonlinear transmission problems than in Section 3.3, which also contains the implicit-explicitly time-discretized Richards equation in the nondegenerate case of Subsection 1.4.3. Nevertheless, we proceed similarly as in Section 3.3 by first applying Kirchhoff transformations on the subdomains, again using the results on superposition operators in Subsection 1.5.4. Then, a nonlinear Steklov–Poincaré theory is established for the transformed problem which extends the linear theory in Discacciati [33, Chapter 5]. As in Section 3.3, this leads to sufficient conditions for the convergence of the nonlinear Robin method which can be proved to be satisfied in one space dimension. However, these considerations are more complicated in Section 3.4 because here we have two nonlinearities in the partial differential equation of the original problem, whereas in Section 3.3 we only have one. In addition, and in contrast to Section 3.3, the subproblems in the Robin iteration procedure are always nonlinear, which is also discussed on the continuous and on the discrete level. Finally, although the same counterexamples as in Section 3.3 apply here, the Robin method shows a satisfying convergence behaviour in numerical tests carried out in two space dimensions.

3.2 Substructuring of a Signorini-type problem for the Richards equation in heterogeneous soil

The purpose of this section is to obtain a weak formulation of a Signorini-type problem for the Richards equation in heterogeneous soil. This is achieved via substructuring of a corresponding problem in homogeneous soil, which leads to an equivalence between the global problem and local problems that are coupled by suitable interface conditions. The latter set of problems is then taken as a definition of a Signorini-type problem for the Richards equation in heterogeneous soil. We start with the global homogeneous problem and some necessary notation in Subsection 3.2.1. Then, Subsection 3.2.2 contains the theoretical results (especially Theorem 3.2.4) for the substructuring of this problem. Finally, in Subsection 3.2.3 we address the Dirichlet–Neumann scheme arising from the substructuring for the Richards equation and note Definition 3.2.7, which gives sense to a Signorini-type problem for the Richards equation in a heterogeneous setting.

3.2.1 Global problem and notation

The setting that we want to consider is given by a decomposition of a bounded open Lipschitz domain $\Omega \subset \mathbb{R}^d$ into two *non-overlapping* open and nonempty subdomains Ω_1 and Ω_2 (i.e. $\overline{\Omega}_1 \cup \overline{\Omega}_2 = \Omega$ and $\Omega_1 \cap \Omega_2 = \emptyset$) with the *interface*

$$\Gamma := \overline{\Omega}_1 \cap \overline{\Omega}_2.$$

As in Quarteroni and Valli [75, p. 6] we assume that Γ is a $(d-1)$-dimensional Lipschitz manifold so that the results on trace spaces in the appendix (on pages 248–251) are applicable. In addition, both Ω_1 and Ω_2 are assumed to have Lipschitz boundaries. Figure 3.1 displays such a situation for $d = 2$ which we have already considered in Subsection 1.5.1 in the homogeneous case.

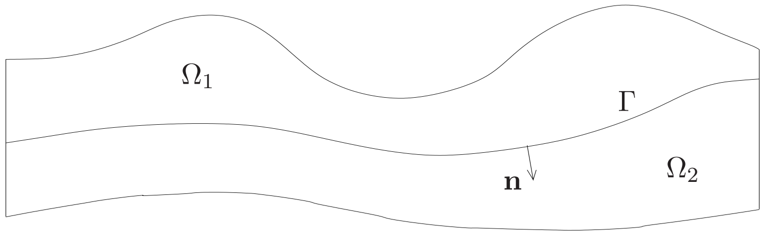


Figure 3.1: 2D-domain Ω decomposed into two subdomains

As already stated in the introduction to this chapter, many of our considerations here are motivated by and generalizations of the corresponding theory presented in Quarteroni and Valli [75] for the linear case. However, since we consider weak formulations of Signorini-type problems for the Richards equation, which involve nonlinearities and convex sets in Sobolev spaces rather than the full spaces, our notation needs to be different in this section. Nevertheless, we will use the notation in [75] wherever it is appropriate, for example, we will denote the "restriction" of a $p \in H^1(\Omega)$ to $\Sigma \subset \partial\Omega$ by $p_{|\Sigma}$ instead of $tr_\Sigma\, p$ with the trace operator $tr_\Sigma : H^1(\Omega) \to H^{1/2}(\Sigma)$. Moreover, for different domains and $i \in \mathbb{N}$ we will from now on distinguish H^1-norms as

$$\|v\|_{1,\Omega} := \|v\|_{H^1(\Omega)} \quad \forall v \in H^1(\Omega) \quad \text{and} \quad \|v_i\|_{1,\Omega_i} := \|v_i\|_{H^1(\Omega_i)} \quad \forall v_i \in H^1(\Omega_i)\,.$$

We start with the homogeneous case, i.e. we assume constant soil parameters (λ and p_b in case of the Brooks–Corey functions, see Section 1.2) on Ω_1 as on Ω_2 providing space-independent nonlinearities $\theta(\cdot)$ and $kr(\cdot)$ on the whole domain Ω. Instead of the weak formulation (1.5.38) of the Signorini-type problem for the Richards equation we consider the problem arising from an implicit–explicit time discretization of (1.5.38). We point out, however, that our equivalence result in Theorem 3.2.4 can equally be established for the problem (1.5.38), see also Subsection 3.2.3.

Setting $f_N(t) = 0$ without loss of generality as in Section 2.3 and applying our implicit–explicit time discretization to (1.5.38) explained in that section, we obtain the variational inequality

$$\begin{aligned} p \in \mathcal{K}_0 : \quad & \int_\Omega \theta(p)\,(v-p)\,dx + \tau \int_\Omega kr(\theta(p))\nabla p\,\nabla(v-p)\,dx \geq \\ & \int_\Omega \theta(\tilde{p})\,(v-p)\,dx + \tau \int_\Omega kr(\theta(\tilde{p}))e_z\nabla(v-p)\,dx \quad \forall v \in \mathcal{K}_0\,. \end{aligned} \tag{3.2.1}$$

Here, $\tau > 0$ is some time step size and we assume that $\tilde{p} \in H^1(\Omega)$ is a known physical pressure from the previous time step. We omit the time step t or t_n in the notation as done in Subsection 2.3.1. Recall from (1.5.37) that

$$\mathcal{K}_0 = \{v \in H^1(\Omega) : v_{|\gamma_D} = p_D \,\wedge\, v_{|\gamma_S} \leq 0\}$$

with an appropriate p_D as in (1.5.36). As already stated in Remark 1.5.19, Theorem 1.5.18 holds equally for the time-discretized versions, i.e. if p solves the variational inequality (3.2.1), then the Kirchhoff–transformed $u = \kappa(p)$ solves (2.3.8) with $u_D = \kappa(p_D)$. The converse is true provided $kr \geq c > 0$ (compare Subsection 1.4.3) and $\gamma_S = \emptyset$.

For the purpose of this section it is indicated to introduce the abbreviations

$$\tilde{a}(p, v-p) := \int_\Omega \theta(p)\,(v-p)\,dx + \tau \int_\Omega kr(\theta(p))\nabla p\,\nabla(v-p)\,dx \quad \forall v \in \mathcal{K}_0$$

and

$$\ell(v-p) := \int_\Omega \theta(\tilde{p})\,(v-p)\,dx + \tau \int_\Omega kr(\theta(\tilde{p}))e_z\nabla(v-p)\,dx \quad \forall v \in \mathcal{K}_0$$

in (3.2.1) with the solution $p \in \mathcal{K}_0$ if it exists. The form $\tilde{a}(\cdot,\cdot)$ on $(H^1(\Omega))^2$ is nonlinear in the first and linear in the second entry while $\ell(\cdot)$ is a linear form on $H^1(\Omega)$. Analogously, we introduce the convex sets

$$\mathcal{K}_i := \{v \in H^1(\Omega_i) : v_{|\gamma_{D_i}} = p_{D|\gamma_{D_i}} \wedge v_{|\gamma_{S_i}} \leq 0\}$$

as well as the forms

$$\tilde{a}_i(p_i, v_i{-}p_i) := \int_{\Omega_i} \theta(p_i)\,(v_i{-}p_i)\,dx{+}\tau \int_{\Omega_i} kr(\theta(p_i))\nabla p_i\,\nabla(v_i{-}p_i)\,dx \quad \forall v_i \in \mathcal{K}_i$$

and

$$\ell_i(v_i-p_i) := \int_{\Omega_i} \theta(\tilde{p}_i)\,(v_i-p_i)\,dx + \tau \int_\Omega kr(\theta(\tilde{p}_i))e_z\nabla(v_i-p_i)\,dx \quad \forall v_i \in \mathcal{K}_i$$

with $p_i \in \mathcal{K}_i$ and given $\tilde{p}_i := \tilde{p}_{|\Omega_i}$ for $i = 1, 2$, which correspond to the subdomains Ω_1 and Ω_2. Here we have used the definitions $\gamma_{D_i} := \partial\Omega_i \cap \gamma_D$ and $\gamma_{S_i} := \partial\Omega_i \cap \gamma_S$.

We also need to introduce convex sets with prescribed Dirichlet values on the interface which we define as

$$\mathcal{K}_i^{p_j} := \{v \in \mathcal{K}_i : v_{|\Gamma} = p_{j|\Gamma}\}$$

for $p_j \in \mathcal{K}_j$ and $i, j \in \{1, 2\}$. In addition, we introduce the convex set of traces

$$\Lambda_0 := \{\eta \in H^{1/2}(\Gamma) : \eta = v_{|\Gamma} \text{ for a } v \in \mathcal{K}_0\}$$

and its translated copy

$$\tilde{\Lambda} := \Lambda_0 - p_{|\Gamma} = \{\eta \in H^{1/2}(\Gamma) : \eta = v_{|\Gamma} \text{ for a } v \in \mathcal{K}_0 - p\}\,.$$

with a $p \in \mathcal{K}_0$ (which will later be the assumed solution of (3.2.1)). We refer to Lemma 3.2.3 to make sure that traces of $H^1(\Omega)$-functions in the interior of Ω are well defined.

With respect to the setting for the Poisson problem considered in Quarteroni and Valli [75, p. 6] we note that our convex sets degenerate and fit into that setting if we only have homogeneous Dirichlet values imposed on $\partial\Omega$. With the definitions

$$\begin{array}{rcl} V & := & H_0^1(\Omega) \\ V_i & := & \{v \in H^1(\Omega_i) : v_{i|\partial\Omega\cap\partial\Omega_i} = 0\} \\ V_i^0 & := & H_0^1(\Omega_i) \\ \Lambda & := & \{\eta \in H^{1/2}(\Gamma) : \eta = v_{|\Gamma} \text{ for a } v \in V\} \end{array} \tag{3.2.2}$$

for $i = 1, 2$ we obtain $V = \mathcal{K}_0$, $V_i = \mathcal{K}_i$ and $V_i^0 = \mathcal{K}_i^{p_i} - p_i$ as well as $\Lambda_0 = \tilde{\Lambda} = \Lambda$ in this case. We recall that $\Lambda = H^{1/2}(\Gamma)$ for $\Gamma \cap \partial\Omega = \emptyset$ and $\Lambda = H_{00}^{1/2}(\Gamma)$ if $\Gamma \cap \partial\Omega \neq \emptyset$ which is the case we mostly consider here (compare (A.2.4) and [75, pp. 6/7]). However, the structure of $\tilde{\Lambda}$ is more delicate in our general case where we can only guarantee that $\tilde{\Lambda}$ is a convex subset of $H^{1/2}(\Gamma)$. For our equivalence result (Theorem 3.2.4) we need the vector space structure of $\tilde{\Lambda}$ which we cannot expect if $\Gamma \cap \overline{\gamma}_S \neq \emptyset$.

3.2.2 Substructuring equivalence result in homogeneous soil

We start with a result which guarantees that $\tilde{\Lambda}$ is a vector space.

Proposition 3.2.1. *We assume $\Gamma \cap \overline{\gamma}_S = \emptyset$. Then $\tilde{\Lambda}$ is a subspace of $H^{1/2}(\Gamma)$ with the property*

$$\tilde{\Lambda} = \{\eta \in H^{1/2}(\Gamma) : \eta = v_{|\Gamma} \textit{ for a } v \in H^1_{\gamma_D\cup\gamma_S}(\Omega)\}\,, \tag{3.2.3}$$

i.e. containing $H_{00}^{1/2}(\Gamma)$. If, in addition, $\Gamma \cap \overline{\gamma}_N = \emptyset$ and $\Gamma \cap \partial\Omega \neq \emptyset$ or $\Gamma \cap \overline{\gamma}_D = \emptyset$, then we have $\tilde{\Lambda} = H_{00}^{1/2}(\Gamma)$ or $\tilde{\Lambda} = H^{1/2}(\Gamma)$, respectively. In the general case $\tilde{\Lambda}$ is a Hilbert space with the quotient norm

$$\|\eta\|_{\tilde{\Lambda}} = \inf\{\|v\|_{1,\Omega} : v \in H^1_{\gamma_D\cup\gamma_S}(\Omega) \wedge \eta = v_{|\Gamma}\}\,. \tag{3.2.4}$$

Moreover, with the subspace

$$\tilde{H}^1_{\gamma_{D_i}\cup\gamma_{S_i}}(\Omega_i) := \{v \in H^1_{\gamma_{D_i}\cup\gamma_{S_i}}(\Omega_i) : v_{i|\Gamma} \in \tilde{\Lambda}\}$$

of the Hilbert space $H^1_{\gamma_{D_i}\cup\gamma_{S_i}}(\Omega_i)$ the trace operator

$$tr_\Gamma : H^1_{\gamma_D\cup\gamma_S}(\Omega) \to \tilde{\Lambda} \tag{3.2.5}$$

induces continuous linear trace operators

$$tr_{\Gamma,i} : \tilde{H}^1_{\gamma_{D_i}\cup\gamma_{S_i}}(\Omega_i) \to \tilde{\Lambda}\,, \quad i = 1, 2\,,$$

for which, in addition, continuous linear extension operators

$$R_i : \tilde{\Lambda} \to H^1_{\gamma_{D_i}\cup\gamma_{S_i}}(\Omega_i)\,, \quad i = 1, 2\,, \tag{3.2.6}$$

with $tr_{\Gamma,i}R_i\eta = \eta$ for all $\eta \in \tilde{\Lambda}$ exist.

Proof. In order to see "$\supset$" in (3.2.3) observe that $v+p \in \mathcal{K}_0$ for $p \in \mathcal{K}_0$ and any $v \in H^1_{\gamma_D \cup \gamma_S}(\Omega)$. Conversely, since we have $\operatorname{dist}(\Gamma, \gamma_S) > 0$ or $\gamma_S = \emptyset$ there are open neighbourhoods O_Γ and O_{γ_S} of Γ and γ_S, respectively, with $O_\Gamma \cap O_{\gamma_S} = \emptyset$ and an open ball $B \subset \mathbb{R}^d$ with $\overline{O}_\Gamma \cup \overline{O}_{\gamma_S} \cup \overline{\Omega} \subset B$. It is well known that there is a $\varphi \in C_0^\infty(B)$ with a range in $[0,1]$ satisfying $\varphi_{|O_{\gamma_S}} = 0$ and $\varphi_{|O_\Gamma} = 1$, consult e.g. [56, p. 277]. Let $\eta \in \tilde{\Lambda}$ and $v \in \mathcal{K}_0$ with $v_{|\Gamma} = \eta$. Then one can check with the Leibniz rule (compare e.g. [1, p. 21]) and $\varphi_{|\Omega} \in W^{1,\infty}(\Omega)$ that $\varphi v \in H^1_{\gamma_D \cup \gamma_S}(\Omega)$ holds. Moreover, we have $(\varphi v)_{|\Gamma} = v_{|\Gamma} = \eta$. In particular, (3.2.3) entails $\tilde{\Lambda} \supset H^{1/2}_{00}(\Gamma)$. Note that the arguments can also be applied to the case $\gamma_S = \emptyset$.

If, in addition, $\Gamma \cap \overline{\gamma}_N = \emptyset$, then we can replace $H^1_{\gamma_D \cup \gamma_S}(\Omega)$ by $H^1_0(\Omega)$ in (3.2.3) and obtain $\tilde{\Lambda} = H^{1/2}_{00}(\Gamma)$ due to $\Gamma \cap \partial\Omega \neq \emptyset$. This can be seen in the same manner as above: If O_Γ and O_{γ_N} are chosen analogously as O_Γ and O_{γ_S} and a function $\varphi \in C_0^\infty(B)$ with $\varphi_{|O_{\gamma_N}} = 0$ and $\varphi_{|O_\Gamma} = 1$ is at hand, then for any $v \in H^1_{\gamma_D \cup \gamma_S}(\Omega)$ the function $\varphi v \in H^1_0(\Omega)$ satisfies $(\varphi v)_{|\Gamma} = v_{|\Gamma}$.

If, instead, $\Gamma \cap \overline{\gamma}_D = \emptyset$, then we can replace $H^1_{\gamma_D \cup \gamma_S}(\Omega)$ by $H^1(\Omega)$ in (3.2.3). Now we choose O_Γ and $O_{\gamma_D \cup \gamma_S}$ analogously as O_Γ and O_{γ_S} above and a $\varphi \in C_0^\infty(B)$ with $\varphi_{|O_{\gamma_D \cup \gamma_S}} = 0$ and $\varphi_{|O_\Gamma} = 1$. As a consequence, for any $v \in H^1(\Omega)$ we have $\varphi v \in H^1_{\gamma_D \cup \gamma_S}(\Omega)$ and $(\varphi v)_{|\Gamma} = v_{|\Gamma}$.

With regard to the general case it is easily checked that the quotient norm in (3.2.4) is indeed a norm (compare [98, p. 34]). With this norm, tr_Γ in (3.2.5) is a quotient map and therefore $\tilde{\Lambda}$ is isometrically isomorphic to the quotient $H^1_{\gamma_D \cup \gamma_S}(\Omega)/\ker(tr_\Gamma)$, see [98, pp. 54, 56]. Since $H^1_{\gamma_D \cup \gamma_S}(\Omega)$ is a Hilbert space we have the canonical representation $H^1_{\gamma_D \cup \gamma_S}(\Omega) = \ker(tr_\Gamma) \oplus \ker(tr_\Gamma)^\perp$ in which $\ker(tr_\Gamma)^\perp$ is the orthogonal complement of the (closed) kernel $\ker(tr_\Gamma)$, see [98, p. 221]. Therefore, we can conclude the isometric isomorphisms

$$\ker(tr_\Gamma)^\perp \cong H^1_{\gamma_D \cup \gamma_S}(\Omega)/\ker(tr_\Gamma) \cong \tilde{\Lambda}$$

in which tr_Γ induces the isomorphism $\ker(tr_\Gamma)^\perp \cong \tilde{\Lambda}$. In particular, $\tilde{\Lambda}$ is a Hilbert space. The inverse

$$R : \tilde{\Lambda} \to \ker(tr_\Gamma)^\perp \subset H^1_{\gamma_D \cup \gamma_S}(\Omega)$$

of tr_Γ restricted to $\ker(tr_\Gamma)^\perp$ is a continuous linear map with the property $tr_\Gamma R\eta = \eta$ for all $\eta \in \tilde{\Lambda}$. The definition $R_i\eta := (R\eta)_{|\Omega_i}$ for all $\eta \in \tilde{\Lambda}$ and $i = 1, 2$ provides continuous linear operators (3.2.6) with the properties

$$\|R_i\eta\|_{1,\Omega_i} \leq \|R\eta\|_{1,\Omega} = \|\eta\|_{\tilde{\Lambda}} \quad \forall \eta \in \tilde{\Lambda}$$

and (with a glance at Lemma 3.2.3) $tr_{\Gamma,i} R_i\eta = tr_\Gamma R\eta = \eta$ as required. □

Observe that for the existence of the extension operators R_i, $i = 1, 2$, we needed to use the Hilbert space structure of $H^1_{\gamma_{D_i} \cup \gamma_{S_i}}(\Omega_i)$, in particular the existence of an orthogonal complement of $\ker(tr_\Gamma)$. In contrast, a closed subspace in a general Sobolev space $W^{1,p}_{\gamma_{D_i}\cup\gamma_{S_i}}(\Omega)$ for $p \geq 1$ and $p \neq 2$ does not necessarily have a complemented subspace in $W^{1,p}_{\gamma_{D_i}\cup\gamma_{S_i}}(\Omega)$, see [98, pp. 162, 248]. In such a case one would have to define extension operators as in (3.2.6) or, equivalently, projections from $W^{1,p}_{\gamma_{D_i}\cup\gamma_{S_i}}(\Omega)$ on $\ker(tr_\Gamma)$ more explicitly. Concerning this question we refer to Lions and Magenes [64, pp. 19–23, 38–43].

Remark 3.2.2. Since $H^{1/2}_{00}(\Gamma)$ is intrinsically definable (see (A.2.4) but also Quarteroni and Valli [75, p. 7]) and thus only dependent on Γ, it seems that $\tilde{\Lambda} = H^{1/2}_{00}(\Gamma)$ holds whenever $\Gamma \cap \overline{\gamma}_D = \Gamma \cap \partial\Omega$ is satisfied (at least if we have $\Gamma \cup \overline{\gamma}_D \subset \partial\tilde{\Omega}$ with some Lipschitz domain $\tilde{\Omega}$ and except for possibly higher dimensional "degenerate cases"). With the same reasoning it seems that $\tilde{\Lambda}$ is also a vector space (satisfying (3.2.3)) if $\Gamma \cap \overline{\gamma}_S \neq \emptyset$ holds with $\Gamma \cap \overline{\gamma}_S = \Gamma \cap \overline{\gamma}_D$ (again possibly except for degeneracies). Note that if Γ intersects arbitrary parts of γ_D or γ_N (but not of $\overline{\gamma}_S$), then the trace space $\tilde{\Lambda}$ could be regarded as a "partially 00" $H^{1/2}(\Gamma)$-space. Depending on the geometry, especially in higher dimensions, one might obtain cases, which could result in different "00-degrees", for example, if $d = 3$ and $\Gamma \cap \overline{\gamma}_D$ contains one point. (Note that we always assume γ_D to have positive Hausdorff measure if $\gamma_D \neq \emptyset$.) At least in such degenerate cases $\tilde{H}^1_{\gamma_{D_i}\cup\gamma_{S_i}}(\Omega_i)$ might be a proper subspace of $H^1_{\gamma_{D_i}\cup\gamma_{S_i}}(\Omega_i)$.

The following basic result is crucial for any substructuring in $H^1(\Omega)$.

Lemma 3.2.3. *If $p \in H^1(\Omega)$, then we have $p_i := p_{|\Omega_i} \in H^1(\Omega_i)$ for $i = 1,2$ and $p_{1|\Gamma} = p_{2|\Gamma}$. Conversely, if $p_i \in H^1(\Omega_i)$ for $i = 1,2$ and $p_{1|\Gamma} = p_{2|\Gamma}$ holds, then*

$$p := \begin{cases} p_1 & \text{on } \Omega_1 \\ p_2 & \text{on } \Omega_2 \end{cases}$$

is contained in $H^1(\Omega)$.

Proof. The first assertion is easy to see by considering a sequence of functions $(\varphi_n)_{n\in\mathbb{N}} \subset C^\infty(\overline{\Omega})$ converging to p in $H^1(\Omega)$ and observing that their restrictions to Ω_i or to Γ converge to p_i or to $p_{i|\Gamma}$, respectively, for $i = 1,2$ in the corresponding norms. Conversely, to see that p is weakly differentiable we apply partial integration (A.2.11) in $H^1(\Omega_i)$ to the weak derivatives of p_i for $i = 1,2$ tested with test functions in $C^\infty_0(\overline{\Omega})$ and observe that the contributions on Γ cancel each other due to the gluing $p_{1|\Gamma} = p_{2|\Gamma}$. □

We can now prove the main result of this section which is a generalization of Lemma 1.2.1 in Quarteroni and Valli [75] to problems of Signorini's type for nonlinear equations. Recall that extension operators are defined as right inverses to corresponding trace maps.

Theorem 3.2.4. *Let $\Gamma \cap \overline{\gamma}_S = \emptyset$. Then the variational problem (3.2.1) which in short reads*

$$p \in \mathcal{K}_0 : \quad \tilde{a}(p, v-p) - \ell(v-p) \geq 0 \quad \forall v \in \mathcal{K}_0 \tag{3.2.7}$$

can be equivalently reformulated as: Find $p_1 \in \mathcal{K}_1$ *and* $p_2 \in \mathcal{K}_2$ *such that*

$$\tilde{a}_1(p_1, v_1 - p_1) - \ell_1(v_1 - p_1) \geq 0 \quad \forall v_1 \in \mathcal{K}_1^{p_1} \tag{3.2.8}$$

$$p_1 = p_2 \quad \textit{on } \Gamma \tag{3.2.9}$$

$$\tilde{a}_2(p_2, v_2 - p_2) - \ell_2(v_2 - p_2) \geq 0 \quad \forall v_2 \in \mathcal{K}_2^{p_2} \tag{3.2.10}$$

$$\tilde{a}_2(p_2, R_2\mu) = \ell_2(R_2\mu) + \ell_1(R_1\mu) - \tilde{a}_1(p_1, R_1\mu) \quad \forall \mu \in \tilde{\Lambda} \tag{3.2.11}$$

where R_i *denotes any possible extension operator from* $\tilde{\Lambda}$ *to* $H^1_{\gamma_{D_i} \cup \gamma_{S_i}}(\Omega_i)$ *for* $i = 1, 2$*.*

Note that R_i, $i = 1, 2$, exist and can be chosen as the continuous linear extension operators given by Proposition 3.2.1.

Proof. First let p be a solution of (3.2.7). Then we have $p_i := p_{|\Omega_i} \in \mathcal{K}_i$ for $i = 1, 2$ and (3.2.9) due to Lemma 3.2.3. Let $v_1 \in \mathcal{K}_1^{p_1}$. Since (3.2.9) holds, the function

$$v := \begin{cases} v_1 & \text{on } \Omega_1 \\ p_2 & \text{on } \Omega_2 \end{cases}$$

is contained in $\mathcal{K}_0$ (Lemma 3.2.3), and (3.2.8) follows from (3.2.7). Analogously, we obtain (3.2.10). Now, for each $\mu \in \tilde{\Lambda}$ the function $R\mu$ defined by

$$R\mu := \begin{cases} R_1\mu & \text{on } \Omega_1 \\ R_2\mu & \text{on } \Omega_2 \end{cases} \tag{3.2.12}$$

belongs to $H^1_{\gamma_D \cup \gamma_S}(\Omega)$ (Lemma 3.2.3) and we have $\pm R\mu + p \in \mathcal{K}_0$ ($\tilde{\Lambda}$ is a vector space!). The variational inequality (3.2.7) applied to both $v = R\mu + p \in \mathcal{K}_0$ and $v = -R\mu + p \in \mathcal{K}_0$ provides (equality in) (3.2.11).

Conversely, let $p_i \in \mathcal{K}_i^{p_i}$, $i = 1, 2$, be solutions of (3.2.8)–(3.2.11). Setting

$$p := \begin{cases} p_1 & \text{on } \Omega_1 \\ p_2 & \text{on } \Omega_2 \end{cases}$$

we obtain $p \in \mathcal{K}_0$ due to (3.2.9), the definitions of $\mathcal{K}_i^{p_i}$ and Lemma 3.2.3. Choosing a $v \in \mathcal{K}_0$ we set $\mu := v_{|\Gamma}$ and $\lambda := p_{|\Gamma}$ and obtain $\mu - \lambda \in \tilde{\Lambda}$ by definition of $\tilde{\Lambda}$. Defining $R(\mu - \lambda)$ according to (3.2.12) we see that

$$v_i := v_{|\Omega_i} - R_i(\mu - \lambda) \in \mathcal{K}_i^{p_i}\,, \quad i = 1, 2\,,$$

holds. Now, (3.2.8), (3.2.10) and (3.2.11) entail

$$\begin{aligned} \tilde{a}(p, v - p) - \ell(v - p) &= \sum_{i=1}^{2} \tilde{a}_i(p_i, v_{|\Omega_i} - p_i) - \ell_i(v_{|\Omega_i} - p_i) \\ &= \sum_{i=1}^{2} \Big(\tilde{a}_i(p_i, v_i - p_i) - \ell_i(v_i - p_i) \\ &\qquad + \tilde{a}_i(p_i, R_i(\mu - \lambda)) - \ell_i(R_i(\mu - \lambda)) \Big) \geq 0 \end{aligned}$$

as required. □

Remark 3.2.5. We point out that it seems unrealistic to generalize Theorem 3.2.4 in a satisfying way to situations in which Γ and $\overline{\gamma}_S$ have a nonempty intersection (and thus $\tilde{\Lambda}$ is in general no longer a vector space). Observe that for the second part of the proof we need extension operators

$$R_i : \tilde{\Lambda} \to \mathcal{K}_i - \mathcal{K}_i^{p_i} \subset H^1_{\gamma_{D_i}}(\Omega_i)\,, \quad i = 1, 2\,, \tag{3.2.13}$$

(not necessarily linear or continuous) with the property

$$(v - R(v_{|\Gamma} - p_{|\Gamma}))_{|\gamma_S} \le 0 \quad \forall v \in \mathcal{K}_0 \tag{3.2.14}$$

(with R as in (3.2.12)) and for which (3.2.11) is satisfied with "$\ge$" instead of "$=$". And, indeed, with this modified condition (3.2.11) we obtain the equivalence — if such extension operators exist. However, we have the following proposition, in which the second assertion presumably also holds in all cases where the intersection of Γ and $\overline{\gamma}_S$ leads to a $\tilde{\Lambda}$ without a vector space structure.

Proposition 3.2.6. *Let $p \in \mathcal{K}_0$ and $R : \tilde{\Lambda} \to H^1_{\gamma_D}(\Omega)$ be some (not necessarily linear or continuous) extension operator, i.e. satisfying $tr_\Gamma R\mu = \mu$ for all $\mu \in \tilde{\Lambda}$. In addition, assume with (3.2.12) that R satisfies (3.2.13) and (3.2.14). Then we have*

$$R : \tilde{\Lambda} \to \{v \in H^1_{\gamma_D}(\Omega) : v_{|\gamma_S} \ge 0\}\,. \tag{3.2.15}$$

In particular, if there is a neighbourhood O_Γ of Γ with $O_\Gamma \cap \partial\Omega = O_\Gamma \cap \overline{\gamma}_S \neq \emptyset$, then such a map R does not exist.

Proof. For simplicity but without loss of generality we assume $p = 0$. Then (3.2.14) reads

$$v_{|\gamma_S} \le (R(v_{|\Gamma}))_{|\gamma_S} \quad \forall v \in \mathcal{K}_0\,. \tag{3.2.16}$$

The following construction can be carried out as in the proof of Proposition 3.2.1. Assuming that γ_S is open in the relative topology of $\partial\Omega$, we consider neighbourhoods $O_n \subset \mathbb{R}^d$ of $\overline{\gamma}_S \backslash \gamma_S$ and neighbourhoods U_n of Γ for $n \in \mathbb{N}$ with Lebesgue measure $|O_n|$, $|U_n| \to 0$ for $n \to \infty$. Now, for any $\eta \in \tilde{\Lambda}$ it is possible to construct a sequence of functions $(v_n)_{n\in\mathbb{N}} \subset \mathcal{K}_0$ which satisfy $v_{n|\Gamma} = \eta$ and $v_{n|\gamma_S \backslash (O_n \cup U_n)} = 0$ for all $n \in \mathbb{N}$. $(v_n)_{n\in\mathbb{N}}$ can probably even be chosen uniformly bounded in $H^1(\Omega)$ if one uses the example function in Braess [19, p. 30]. Now, it follows from (3.2.16) that R has to provide $(R\eta)_{|\gamma_S} \ge 0$ for all $\eta \in \tilde{\Lambda}$. This proves the first assertion (3.2.15) of Proposition 3.2.6 (which can analogously be obtained for arbitrary $p \in \mathcal{K}_0$). Consequently, the elements of $\tilde{\Lambda}$ are both traces of functions v with $v_{|\gamma_S} \ge 0$ and of functions w with $w_{|\gamma_S} \le 0$.

In particular, if $O_\Gamma \cap \partial\Omega = O_\Gamma \cap \gamma_S \neq \emptyset$ for some neighbourhood O_Γ of Γ, then we obtain $\tilde{\Lambda} \subset H^{1/2}_{00}(\Gamma)$ (and therefore $\tilde{\Lambda} = H^{1/2}_{00}(\Gamma)$) with the help of arguments as in Proposition 3.2.1 for the case $\Gamma \cap \overline{\gamma}_N = \emptyset$ and $\Gamma \cap \partial\Omega \neq \emptyset$. However, there is an $\tilde{\eta} \in H^{1/2}(\Gamma) \backslash H^{1/2}_{00}(\Gamma)$ and a $v \in H^1(\Omega)$ with $v_{|\Gamma} = \tilde{\eta}$. In addition, we (generally) have

$$v^+ := \max(v, 0) \in H^1(\Omega) \quad \text{and} \quad v^- := \min(v, 0) \in H^1(\Omega)$$

due to Theorem 1.5.15. Again, by a localization technique as in Proposition 3.2.1 for the case $\Gamma \cap \overline{\gamma}_D = \emptyset$, using $O_\Gamma \cap \partial\Omega = O_\Gamma \cap \gamma_S \neq \emptyset$, we conclude

$$-v^+_{|\Gamma},\, v^-_{|\Gamma} \in \{w_{|\Gamma} : w \in H^1_{\gamma_D}(\Omega) \wedge w_{|\gamma_S} \leq 0\} \subset \tilde{\Lambda} \tag{3.2.17}$$

with the help of Proposition 1.5.6. But since we have $v_{|\Gamma} \notin H^{1/2}_{00}(\Gamma)$, at least one of the functions $v^+_{|\Gamma}$, $v^-_{|\Gamma}$ is not contained in $H^{1/2}_{00}(\Gamma)$ which, however, contradicts the inclusion $\tilde{\Lambda} \subset H^{1/2}_{00}(\Gamma)$. □

Using an $\tilde{\eta}$ in the gap between a "partially 00" $H^{1/2}(\Gamma)$-space ("00" on $\Gamma \cap \gamma_D$) and another suitable "partially 00" $H^{1/2}(\Gamma)$-space ("00" on $\Gamma \cap (\gamma_D \cup \gamma_S)$), in which $\tilde{\Lambda}$ is contained, one might obtain (3.2.17) by a possibly more refined localization technique. In this way, it seems feasible to extend the assertion about the non-existence of the operator R to much more general nontrivial intersections of Γ and γ_S, which lead to proper convex subsets $\tilde{\Lambda}$ of a "partially 00" trace space (see also Remark 3.2.2).

Although the considerations in Remark 3.2.5 lead to the poor non-existence result in Proposition 3.2.6 they still contain a positive message for the discrete setting. If we discretize the problems (3.2.7) and (3.2.8)–(3.2.11) analogously as in Subsection 2.5.1, then obviously Theorem 3.2.4 and Remark 3.2.5 can be established accordingly in the discrete setting. Now, however, the properties (3.2.13) and (3.2.14) are satisfied if we consider R to be the trivial extension, setting $R\mu$ as 0 on the nodes in $\Omega\backslash\Gamma$ while respecting the Dirichlet values. In this case we would also obtain an equivalence between the discretized version of (3.2.7) and the discretized versions of (3.2.8)–(3.2.11) where "=" is replaced by "≥" in the discretization of (3.2.11).

3.2.3 Dirichlet–Neumann scheme for the time-discretized Richards equation and the heterogeneous case

As far as the interpretation of the interface conditions is concerned, it is clear that (3.2.9) indicates the continuity of the pressure p across the interface Γ. Furthermore, with the help of Green's formula (1.5.9) one can easily verify that (3.2.11) is the weak formulation for the continuity of the implicit-explicitly time-discretized flux

$$(kr(\theta(p_1))\nabla p_1 - kr(\theta(\tilde{p}_1))e_z) \cdot \mathbf{n} = (kr(\theta(p_2))\nabla p_2 - kr(\theta(\tilde{p}_2))e_z) \cdot \mathbf{n} \quad \text{on } \Gamma \tag{3.2.18}$$

related to the implicit-explicitly time-discretized Richards equation (see (1.5.1) and (1.5.3)) which reads

$$\frac{\theta(p) - \theta(\tilde{p})}{\tau} - \operatorname{div}\Big(kr(\theta(p))\nabla p - kr(\theta(\tilde{p}))e_z\Big) = 0 \tag{3.2.19}$$

in strong form (see also Subsection 3.4.5). For simplicity, we dropped the minus sign on both sides of (3.2.18), so we actually deal with the negative discretized water flux here. Moreover, we adapt to the usual convention that $\mathbf{n}$ is the outward normal of the subdomain Ω_1 (see Figure 3.1 and compare [75, pp. 1/2]).

As already stated above, we can also establish Theorem 3.2.4 for the variational form (1.5.38) of the Richards equation before time discretization. Then, we have $p_i = \tilde{p}_i$ for $i = 1, 2$ in (3.2.18), and the strong form of (3.2.11) is just the continuity of the water flux at the time t.

As in the linear case in Quarteroni and Valli [75, p. 7], the subproblems (3.2.8) and (3.2.10) are underdetermined because of lacking boundary values for p_i, $i = 1, 2$, on Γ. If these problems are nontrivial problems of Signorini's type, even the convex sets of test functions $\mathcal{K}_i^{p_i}$ are unknown a priori. This is not the case if $\gamma_{S_i} = \emptyset$ since then $\mathcal{K}_i^{p_i} - p_i = V_i^0$ is a vector space. Nevertheless, both from an algorithmic and from an analytic point of view, it is quite important to know how one can "complete" these problems in order to make them well-posed.

As in the linear case this can be done within an iterative process with the help of the Dirichlet interface condition (3.2.9) and the Neumann interface condition (3.2.11). Given an iterate $p_2^k \in \mathcal{K}_2$, one can compute an iterate $p_1^{k+1} \in \mathcal{K}_1$ by solving the convex problem

$$p_1^{k+1} \in \mathcal{K}_1^{p_2^k} : \quad \tilde{a}_1(p_1^{k+1}, v_1 - p_1^{k+1}) - \ell_1(v_1 - p_1^{k+1}) \geq 0 \quad \forall v_1 \in \mathcal{K}_1^{p_2^k} \tag{3.2.20}$$

imposing the Dirichlet condition $p_1^{k+1} = p_2^k$ on Γ which refers to (3.2.9). Given $p_1^{k+1} \in \mathcal{K}_1$, one can obtain $p_2^{k+1} \in \mathcal{K}_2$ by solving the convex problem

$$\begin{aligned} &p_2^{k+1} \in \mathcal{K}_2 : \quad \tilde{a}_2(p_2^{k+1}, \tilde{v}_2 - p_2^{k+1}) - \ell_2(\tilde{v}_2 - p_2^{k+1}) \qquad (3.2.21)\\ &- \Big(\ell_1(R_1(\tilde{v}_2 - p_2^{k+1})_{|\Gamma}) - \tilde{a}_1(p_1^{k+1}, R_1(\tilde{v}_2 - p_2^{k+1})_{|\Gamma})\Big) \geq 0 \quad \forall \tilde{v}_2 \in \mathcal{K}_2 \end{aligned}$$

into which the weak form (3.2.11) of the Neumann condition (3.2.18) (with p_i replaced by p_i^{k+1}, i=1,2) is wired. To see this observe first that (3.2.21) is uniquely solvable due to Theorem 2.3.16 with the assumptions given there. Now, for any $\tilde{v}_2 \in \mathcal{K}_2$ consider the trace function

$$\mu := (\tilde{v}_2 - p_2^{k+1})_{|\Gamma} \in \Lambda_0 - p_{2|\Gamma}^{k+1} \quad \text{and} \quad v_2 := \tilde{v}_2 - R_2\mu \in \mathcal{K}_2^{p_2^{k+1}} .$$

Then, with these functions, adding (3.2.10) and (3.2.11) leads to (3.2.21).

Together with an initial guess p_2^0, the iterative procedure given by (3.2.20) and (3.2.21) for $k \geq 0$ is a *nonlinear Dirichlet–Neumann scheme* in a weak formulation applied to the Signorini-type problem (3.2.1) for the time-discretized Richards equation. However, as in the linear case one would apply an additional damping (compare (3.3.10)) in order to have a chance to get a convergent sequence.

It is also possible to combine the two interface conditions (3.2.9) and (3.2.11) in order to obtain convex problems with Robin boundary conditions on Γ. They also turn out to be uniquely solvable for the Richards equation in homogeneous soil under natural conditions. In Subsection 3.4.1 we will address this option, resulting in a Robin method for the Richards equation, in detail.

Finally, we turn to the case of heterogeneous soil, i.e. to the case of possibly different parameter functions $\theta_1(\cdot)$ and $kr_1(\cdot)$ in Ω_1 and $\theta_2(\cdot)$ and $kr_2(\cdot)$ in Ω_2. This is the case referred to as *jumping nonlinearities* in the title of this work. We assume that $\tilde{a}_i(\cdot,\cdot)$ and $\ell_i(\cdot)$, $i=1,2$, are constituted accordingly as in Subsection 3.2.1. With these ingredients one can also define $\tilde{a}(\cdot,\cdot)$ and $\ell(\cdot)$ on Ω and give sense to the corresponding variational inequality (3.2.7).

However, our solution theory from Chapter 2 cannot be applied in this heterogeneous setting since the Kirchhoff transformation cannot be carried out globally on Ω (compare Remark 1.3.1). Therefore, we turn to the corresponding substructuring problem (3.2.8)–(3.2.11) for which the arising Dirichlet and Neumann problems can be solved uniquely after (different) Kirchhoff transformations in the subdomains. The interface conditions which require continuity of the physical pressure (3.2.9) and continuity of the (time-discretized) water flux as discussed above (see (3.2.11) and (3.2.18)) also seem to be hydrologically very well justified. Therefore, we can finally note

Definition 3.2.7. Let $\Gamma \cap \overline{\gamma}_S = \emptyset$ and $\theta_i(\cdot)$, $kr_i(\cdot)$ given (possibly different) parameter functions on Ω_i for $i=1,2$. Let $\tilde{a}_i(\cdot,\cdot)$, $\ell_i(\cdot)$, $i=1,2$, be defined with these functions according to Subsection 3.2.1. We call a function p defined a.e. on Ω with $p_i := p_{|\Omega_i}$, $i=1,2$, a *weak solution of the corresponding Signorini-type problem for the Richards equation in heterogeneous soil* on Ω if we have $p_1 \in \mathcal{K}_1$ and $p_2 \in \mathcal{K}_2$ such that

$$\tilde{a}_1(p_1, v_1 - p_1) - \ell_1(v_1 - p_1) \geq 0 \quad \forall v_1 \in \mathcal{K}_1^{p_1} \tag{3.2.22}$$

$$p_1 = p_2 \quad \text{on } \Gamma \tag{3.2.23}$$

$$\tilde{a}_2(p_2, v_2 - p_2) - \ell_2(v_2 - p_2) \geq 0 \quad \forall v_2 \in \mathcal{K}_2^{p_2} \tag{3.2.24}$$

$$\tilde{a}_2(p_2, R_2\mu) = \ell_2(R_2\mu) + \ell_1(R_1\mu) - \tilde{a}_1(p_1, R_1\mu) \quad \forall \mu \in \tilde{\Lambda} \tag{3.2.25}$$

where R_i denotes any possible extension operator from $\tilde{\Lambda}$ to $H^1_{\gamma_{D_i} \cup \gamma_{S_i}}(\Omega_i)$ for $i=1,2$.

As usual, for more than two subdomains one would consider convex problems for each subdomain and impose continuity of the pressure and the (time-discretized) water flux on each interface.

3.3 Nonlinear Dirichlet–Neumann method

The topic of this section is a quasilinear elliptic transmission problem where the nonlinearity changes discontinuously across two subdomains. This problem is motivated by Definition 3.2.7 of a boundary value problem for the Richards equation on a domain with two different soils in the subdomains. In fact, it can be regarded as a nondegenerate stationary Richards equation without gravity in such a setting.

We treat this problem applying a nonlinear version of the linear Steklov–Poincaré theory introduced in Quarteroni and Valli [75, pp. 8–11]. Therefore, in Subsection 3.3.1, we start with a short introduction to the basic idea underlying the usage of Steklov–Poincaré operators for the analytical treatment of the Dirichlet–Neumann method applied to linear problems.

In order to extend the linear theory to cover our case, we reformulate our nonlinear transmission problem via Kirchhoff transformation, thus obtaining linear problems on the subdomains (cf. Bonani and Ghione [18]). However, this entails nonlinear transmission conditions. Still, this allows a reformulation of the problem as a nonlinear Steklov–Poincaré interface equation. Then, we introduce a Dirichlet–Neumann iteration for this problem which, in analogy to the linear case, can be regarded as a preconditioned Richardson iteration applied to the nonlinear Steklov–Poincaré equation. All this is done in Subsection 3.3.2.

Then, in Subsection 3.3.3, we present a convergence analysis based on Banach's fixed point theorem for our nonlinear Dirichlet–Neumann iteration, generalizing related results for the linear case in Quarteroni and Valli [75, pp. 117–120]. This leads to sufficient conditions for a convergence of the scheme to a unique solution which are satisfied in one space dimension. As a by-product, we obtain well-posedness of our transmission problem in this case.

In Subsection 3.3.4 we present counterexamples in 2D, one of them given and discussed analytically, showing that the sufficient conditions for the convergence of the nonlinear Dirichlet–Neumann algorithm, on which our proof via the contraction argument is based, are violated in higher dimensions. This still leaves the open question whether (at least local) convergence of the Dirichlet–Neumann method can be proved by other techniques.

Finally, in view of the convergence in 1D and the considerations on the counterexamples in 2D, numerical results in two space dimensions are given for our transmission problem in Subsection 3.3.5, suggesting that the algorithm can be applied successfully to higher-dimensional problems and in the general case of the Richards equation.

3.3.1 Basic idea of linear Steklov–Poincaré theory

In order to understand the basic problem that we encounter when we try to analyse the substructuring problem (3.2.22)–(3.2.25) with jumping nonlinearities across the interface Γ, it is helpful to recall the principles of linear Steklov–Poincaré theory. In the following, this is pursued in loose terms and also leads to a formulation of the Dirichlet–Neumann method in terms of Steklov–Poincaré operators which our nonlinear approach in Subsection 3.3.2 extends.

Consider a non-overlapping decomposition of $\Omega \subset \mathbb{R}^d$ into Ω_1 and Ω_2 as in Section 3.2, compare Figure 3.1. The starting point is a boundary value problem

$$L\,u = f \quad \text{on } \Omega \tag{3.3.1}$$

with boundary conditions φ on $\partial\Omega$ (linear in u) where L is some linear partial differential operator. As seen in the (even nonlinear) example in Section 3.2, such a problem can turn out to be equivalent to solving

$$L_i u_i = f \quad \text{on } \Omega_i \tag{3.3.2}$$

where L_i, are the restrictions of L on Ω_i for $i = 1, 2$, together with the restrictions due to the boundary conditions φ and with suitable transmission conditions on Γ. The regularity of u on Ω in some solution space often requires

$$u_{1|\Gamma} = u_{2|\Gamma} \tag{3.3.3}$$

and the fact that u also solves (3.3.1) "across Γ" generally leads to a continuity condition on some "flux" across Γ

$$\Psi_1 u_1 = -\Psi_2 u_2 \tag{3.3.4}$$

given by certain linear operators Ψ_1 and Ψ_2.

Now, usually the aim is to obtain an *interface equation* for the trace $\lambda := u_{|\Gamma}$ of the solution in the trace space Λ which is equivalent to the global problem (3.3.1) or to the substructuring problem (3.3.2)–(3.3.4). For simplicity, we assume that φ are homogeneous Dirichlet boundary conditions on $\partial\Omega$. Then the general approach in the linear case is the following.

We write $L_i^{-1}(f, \mu)$, $i = 1, 2$, for the solution of (3.3.2) where μ is either a Dirichlet boundary value or a "flux" (Neumann) boundary value as in (3.3.4). Both the unknown interface value λ on Γ and the known source term f can be regarded as inhomogeneities for the solution of the subproblems (3.3.2) on Ω_i for $i = 1, 2$. Using the linearity of L_i, $i = 1, 2$, one can decouple these inhomogeneities and separate the known one f from the unknown one λ. Then, obviously, the solutions for the

$$\begin{array}{ll} \text{homogeneous system + inhomogeneity } \lambda \text{ on } \Gamma: & u_i^0 = L_i^{-1}(0, \lambda) =: L_{i,hom}^{-1}(\lambda) \\ \text{inhomogeneous system + homogeneity on } \Gamma: & u_i^* = L_i^{-1}(f, 0) =: L_{i,inh}^{-1}(f) \\ \text{give the solution for the data } f,\, \lambda \text{ on } \Omega_i: & u_i = u_i^0 + u_i^* \,. \end{array} \tag{3.3.5}$$

Considering (3.3.5) for any interface value λ, the first continuity condition (3.3.3) is satisfied. Now, with the linearity of Ψ_i, $i = 1, 2$, the second condition (3.3.4) can be written as

$$\Psi_1 u_1^0 + \Psi_2 u_2^0 = -\Psi_1 u_1^* - \Psi_2 u_2^* \tag{3.3.6}$$

so that the influence of the two inhomogeneities occurs on different sides. The left hand side of this equation results from the action of the Steklov–Poincaré operators, defined by

$$S_i := \Psi_i L_{i,hom}^{-1}\,, \quad i = 1, 2\,, \quad S := S_1 + S_2\,, \tag{3.3.7}$$

on the interface value λ. Together with the right hand side

$$\chi := (-\Psi_1 L_{1,inh}^{-1} - \Psi_2 L_{2,inh}^{-1})(f) \tag{3.3.8}$$

of (3.3.6) one obtains the desired Steklov–Poincaré interface equation in the form

$$S\lambda = \chi\,. \tag{3.3.9}$$

Now, with a given initial iterate λ^0, and as already seen in (3.2.22)–(3.2.25) in a weak formulation for our nonlinear problem, the Dirichlet–Neumann iteration is obtained by the Dirichlet step and the Neumann step

$$\begin{aligned} u_1^{k+1} &= L_1^{-1}(f, \lambda^k) \\ u_2^{k+1} &= L_2^{-1}(f, -\Psi_1 u_1^{k+1}) \end{aligned}$$

for $k \geq 0$, respectively. It provides the next iterate

$$\lambda^{k+1} = \vartheta u_{2|\Gamma}^{k+1} + (1-\vartheta)\lambda^k \tag{3.3.10}$$

after a damping with the factor $0 < \vartheta < 1$ which is necessary in general to obtain convergence.

The natural decomposition $S = S_1 + S_2$ allows for a reformulation of (3.3.10) in terms of Steklov–Poincaré operators. As a result, the damped Dirichlet–Neumann method turns out to be a Richardson procedure

$$\lambda^{k+1} = \lambda^k + \vartheta S_2^{-1}(-S\lambda^k + \chi)$$

for the Steklov–Poincaré interface equation $S\lambda = \chi$ with S_2 as a preconditioner, see Quarteroni and Valli [75, pp. 13/14]. In what is to come, these basic results are generalized to certain nonlinear problems.

3.3.2 Steklov–Poincaré formulation for elliptic problems related to the nondegenerate stationary Richards equation without gravity

In this subsection we introduce a nonlinear elliptic problem in a heterogeneous setting as considered in Subsection 3.2.3 which can be regarded as a stationary Richards equation without gravity in heterogeneous soil with nondegenerate relative permeability (see Subsection 1.4.3). Furthermore, we generalize the ideas from the linear Steklov–Poincaré theory indicated in the last subsection to this nonlinear problem.

As in Subsection 3.2.1, let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain divided into two non-overlapping subdomains Ω_1, Ω_2 with the Lipschitz continuous interface $\Gamma = \overline{\Omega}_1 \cap \overline{\Omega}_2$, compare Figure 3.2.

Given $f \in L^2(\Omega)$, k_1, $k_2 \in L^\infty(\mathbb{R})$ with $k_i \geq \alpha > 0$ for $i = 1, 2$, we consider the following quasilinear elliptic transmission problem in strong formulation.

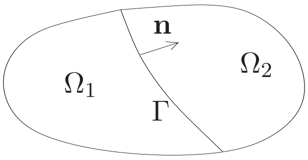


Figure 3.2: Non-overlapping partition of the domain Ω.

Find a function p on Ω, $p_{|\Omega_i} = p_i \in H^1(\Omega_i)$, $i = 1, 2$, $p_{|\partial\Omega} = 0$, such that

$$\begin{aligned} -\operatorname{div}(k_i(p_i)\nabla p_i) &= f && \text{on } \Omega_i\,,\ \ i = 1,2 && (3.3.11)\\ p_1 &= p_2 && \text{on } \Gamma && (3.3.12)\\ k_1(p_1)\nabla p_1 \cdot \mathbf{n} &= k_2(p_2)\nabla p_2 \cdot \mathbf{n} && \text{on } \Gamma\,. && (3.3.13) \end{aligned}$$

Compare this problem to (3.2.19) with (3.2.18) in order to see the claimed relationship with the stationary Richards equation without gravity. We remark that the gravitational term cannot be treated explicitly in the stationary case and would lead to an additional nonlinearity which our approach does not cover. In addition, note that we impose uniform ellipticity

$$k_i(\cdot) \geq \alpha > 0 \quad \text{for } \ i = 1,2 \tag{3.3.14}$$

here, so this case is comparable to the nondegenerate case induced by altered Brooks–Corey functions treated in Subsection 1.4.3. On the other hand, observe that the nonlinearities k_i need not be continuous here, so this setting covers much more and in particular the limit cases discussed in Subsection 1.4.4.

Due to lack of regularity of the nonlinearities, Newton-type linearization is ruled out in advance. However, by our approach based on Kirchhoff transformation, we can reformulate the two nonlinear partial differential equations (3.3.11) as linear Poisson equations in both subdomains.

Since large parts of what we present here can be regarded as generalizations of the linear theory given in Quarteroni and Valli [75], we use much of the notation accordingly. This includes the spaces defined in (3.2.2) with the trace space $\Lambda = H^{1/2}_{00}(\Gamma)$ in the case we consider here. Furthermore, for $i = 1, 2$, we use the abbreviation

$$(w_i, v_i)_{\Omega_i} := \int_{\Omega_i} w_i v_i \, dx \quad \forall w_i, v_i \in L^2(\Omega_i) \tag{3.3.15}$$

for the L^2-scalar product on Ω_i and define the forms

$$\begin{aligned} a_i(w_i, v_i) &:= (\nabla w_i, \nabla v_i)_{\Omega_i} && (3.3.16)\\ b_i(w_i, v_i) &:= (k(w_i)\nabla w_i, \nabla v_i)_{\Omega_i} && (3.3.17) \end{aligned}$$

for $w_i, v_i \in V_i$. As usual, the norm in $H^1(\Omega_i)$ will be denoted by $\|\cdot\|_{1,\Omega_i}$, the norm in Λ with $\|\cdot\|_\Lambda$.

Let R_i, $i = 1, 2$, be any linear continuous extension operator from Λ to V_i. Then, with Green's formula (1.5.9), the variational formulation of problem (3.3.11)–(3.3.13) reads as follows:

Find $p_i \in V_i$, $i = 1, 2$, such that

$$\begin{aligned} b_i(p_i, v_i) &= (f, v_i)_{\Omega_i} \qquad \forall v_i \in V_i^0,\ \ i = 1,2 && (3.3.18)\\ p_{1|\Gamma} &= p_{2|\Gamma} \qquad \text{in } \Lambda && (3.3.19)\\ b_1(p_1, R_1\mu) - (f, R_1\mu)_{\Omega_1} &= -b_2(p_2, R_2\mu) + (f, R_2\mu)_{\Omega_2} \quad \forall \mu \in \Lambda\,. && (3.3.20) \end{aligned}$$

We now introduce new variables u_i, $i = 1, 2$, by Kirchhoff transformations κ_i defined as

$$u_i(x) := \kappa_i(p_i(x)) = \int_0^{p_i(x)} k_i(q)\,dq \quad \text{a.e. in } \Omega_i \tag{3.3.21}$$

according to (1.3.1), which yields $k_i(p_i)\nabla p_i = \nabla u_i$ in a weak sense due to Proposition 1.5.12. For the details we refer to Subsection 1.5.4, where we discussed the topic of weak Kirchhoff transformations in terms of superposition operators in appropriate depth. There, we had $kr \circ \theta$ from the Richards equation instead of the function k_i but with equal generality. Nevertheless, we note the following basic properties of κ_i here.

Proposition 3.3.1. *The following holds for $i = 1, 2$. $\kappa_i : \mathbb{R} \to \mathbb{R}$ satisfies $k_i(0) = 0$ and is a.e. differentiable with $\kappa_i' = k_i$, strictly increasing and Lipschitz continuous with Lipschitz constant $\|k_i\|_\infty$. The inverse κ_i^{-1} is also a.e. differentiable, strictly increasing and Lipschitz continuous with Lipschitz constant $\|k_i^{-1}\|_\infty$.*
Furthermore, with $\alpha \geq 0$ from (3.3.14) we have

$$\alpha\,\|p_i\|_{1,\Omega_i} \leq \|\kappa_i(p_i)\|_{1,\Omega_i} \leq \|k_i\|_\infty\,\|p_i\|_{1,\Omega_i} \tag{3.3.22}$$

and there exist positive constants c, C with

$$c\,\|p_{i|\Gamma}\|_\Lambda \leq \|\kappa_i(p_i)_{|\Gamma}\|_\Lambda \leq C\,\|p_{i|\Gamma}\|_\Lambda\,. \tag{3.3.23}$$

Finally, interpreted as superposition operators, $\kappa_i : V_i \to V_i$ and $\kappa_i : \Lambda \to \Lambda$ are homeomorphisms.

Proof. The first statements have already been noted in Lemma 1.5.7. The estimates in (3.3.22) come from Proposition 1.5.12. Property (3.3.23) is entailed by (3.3.22) and the last statement on κ_i, interpreted as superposition operators, which is a direct consequence of Theorem 1.5.15 and Proposition 1.5.17. The bijectivity of these operators and the commutativity (1.5.33) applied to (3.3.22) provide

$$\alpha\,\|p_{i|\Gamma}\| \leq \|\kappa_i(p_i)_{|\Gamma}\| \leq \|k_i\|_\infty\,\|p_{i|\Gamma}\|$$

for the norm $\|\cdot\|$ in Λ defined by

$$\|\eta\| := \inf_{p \in V_i,\, p_{|\Gamma} = \eta} \|p\|_{1,\Omega_i} \quad \forall \eta \in \Lambda$$

which is known to be equivalent to $\|\cdot\|_\Lambda$, see (A.2.10). □

Now, with the transformations κ_i, problem (3.3.18)–(3.3.20) becomes:

Find $u_i \in V_i$, $i = 1, 2$, such that

$$\begin{aligned} a_i(u_i, v_i) &= (f, v_i)_{\Omega_i} \qquad \forall v_i \in V_i^0, \ \ i = 1,2 && (3.3.24) \\ \kappa_1^{-1}(u_{1|\Gamma}) &= \kappa_2^{-1}(u_{2|\Gamma}) \qquad \text{in } \Lambda && (3.3.25) \\ a_1(u_1, R_1\mu) - (f, R_1\mu)_{\Omega_1} &= -a_2(u_2, R_2\mu) + (f, R_2\mu)_{\Omega_2} \quad \forall \mu \in \Lambda\,. && (3.3.26) \end{aligned}$$

Remark 3.3.2. At first glance, the equivalence of problem (3.3.18)–(3.3.20) with $p_i \in V_i$ and its transformed version (3.3.24)–(3.3.26) with $u_i = \kappa_i(p_i) \in V_i$ seems to be an easy matter. Notice, however, that for (3.3.18) ⇔ (3.3.24) and (3.3.20) ⇔ (3.3.26) we need to apply the weak chain rule (1.5.22) which is not trivial. In addition, for (3.3.12) ⇔ (3.3.19) we need the commutativity of the trace map and the Kirchhoff transformation (see Proposition 1.5.16), which is not straightforward either and which requires a strong result from the theory of superposition operators in the general setting occurring here (compare Proposition 1.5.14 and Theorem 1.5.15).

Moreover, bearing these results from Subsection 1.5.4 in mind, we can interpret κ_i, $i = 1, 2$, equally as a pointwise evaluation almost everywhere on Ω_i or Γ or as a superposition operator in the relevant function spaces on these sets. In particular, the crucial commutativity (1.5.33) makes it possible to talk about the Kirchhoff transformation on Γ of a function in V_i. To simplify the notation, we usually leave out the brackets for operators as in $\kappa_i(\cdot)$ from now on.

Now, we turn to a Steklov–Poincaré formulation of problem (3.3.24)–(3.3.26). The latter is a weak formulation of a problem as discussed in Subsection 3.3.1 including linear subproblems (3.3.24), but now with nonlinear Dirichlet transmission conditions (3.3.25) for the transformed variables in the two subdomains. Since the linearity of the subproblems can be used now, the argumentation is quite similar as usual and outlined in Subsection 3.3.1.

For a given $\lambda \in \Lambda$ and $i = 1, 2$ we now consider the harmonic extensions $u_i^0 = H_i(\kappa_i\lambda) \in V_i$ of the Dirichlet boundary value $\kappa_i\lambda$ on Γ. Furthermore, let $u_i^* = \mathcal{G}_i f$ be the solutions of the subproblems (3.3.24) with homogeneous Dirichlet data $u^*_{i|\partial\Omega_i} = 0$. Due to the linearity of the local problems (3.3.24), the functions

$$u_i = H_i\kappa_i\lambda + \mathcal{G}_i f\,, \quad i = 1, 2\,, \tag{3.3.27}$$

satisfy (3.3.24)–(3.3.26) if and only if

$$\begin{aligned} &a_1(H_1\kappa_1\lambda, R_1\mu) + a_2(H_2\kappa_2\lambda, R_2\mu) = \\ &\qquad (f, R_1\mu)_{\Omega_1} - a_1(\mathcal{G}_1 f, R_1\mu) + (f, R_2\mu)_{\Omega_2} - a_2(\mathcal{G}_2 f, R_2\mu) \quad \forall \mu \in \Lambda\,. \end{aligned} \tag{3.3.28}$$

Since the extension operators R_i, $i = 1, 2$, can be chosen arbitrarily, we set $R_i = H_i$. Denoting by $\langle \cdot, \cdot \rangle$ the duality pairing between Λ' and Λ, we recall the definition of the Steklov–Poincaré operators $S_i : \Lambda \to \Lambda'$,

$$\langle S_i\eta, \mu\rangle = a_i(H_i\eta, H_i\mu) \quad \forall \eta, \mu \in \Lambda\,, \quad i = 1, 2\,, \tag{3.3.29}$$

and furthermore the functional $\chi = \chi_1 + \chi_2 \in \Lambda'$,

$$\langle \chi_i, \mu \rangle = (f, H_i\mu)_{\Omega_i} - a_i(\mathcal{G}_i f, H_i\mu) \quad \forall \mu \in \Lambda \,, \quad i = 1, 2 \,, \tag{3.3.30}$$

which can be found in Quarteroni and Valli [75, pp. 8/9].

Now, (3.3.28) can be written as the nonlinear Steklov–Poincaré interface equation

$$\text{find } \lambda \in \Lambda : \qquad (S_1\kappa_1 + S_2\kappa_2)\lambda = \chi \tag{3.3.31}$$

or, equivalently,

$$\text{find } \lambda_2 \in \Lambda : \qquad (S_1\kappa_1\kappa_2^{-1} + S_2)\lambda_2 = \chi \tag{3.3.32}$$

if we set $\lambda_2 = \kappa_2\lambda$. Note that since $\kappa_2 : \Lambda \to \Lambda$ is a homeomorphism due to Proposition 3.3.1, the convergence of a sequence of iterates λ_2^k to λ_2 implies the convergence of $\lambda^k = \kappa_2^{-1}\lambda_2^k$ to λ an vice versa. We state the main result of this subsection.

Proposition 3.3.3. *Solving problem (3.3.18)–(3.3.20) is equivalent to solving the nonlinear Steklov–Poincaré equations (3.3.31) or (3.3.32) in the sense of (3.3.21) and (3.3.27).*

3.3.3 Convergence result for a Dirichlet–Neumann method and well-posedness in 1D

In this subsection we present a nonlinear Dirichlet–Neumann algorithm for our problem (3.3.18)–(3.3.20) which is a straightforward generalization of the linear one introduced in Subsection 3.3.1. It requires the solution of two linear problems in each iteration step and a nonlinear transformation only on the interface, but it does not involve any further linearization. We analyse the algorithm along the lines of the linear theory in Quarteroni and Valli [75] leading us to sufficient conditions for convergence which are satisfied in one space dimension. As a consequence, we get an existence and uniqueness result for the original nonlinear heterogeneous problem (3.3.18)–(3.3.20).

Since it turns out that for a rigorous analysis the damping has to be carried out in the transformed variables, we state our Dirichlet–Neumann algorithm for the transformed version (3.3.24)–(3.3.26) as follows.

Given $\lambda_2^0 \in \Lambda$, find successively $u_1^{k+1} \in V_1$ and $u_2^{k+1} \in V_2$ for each $k \geq 0$ such that

$$\begin{aligned} a_1(u_1^{k+1}, v_1) &= (f, v_1)_{\Omega_1} \qquad \forall v_1 \in V_1^0 && (3.3.33) \\ u_{1|\Gamma}^{k+1} &= \kappa_1\kappa_2^{-1}(\lambda_2^k) \qquad \text{in } \Lambda && (3.3.34) \end{aligned}$$

and then

$$\begin{aligned} a_2(u_2^{k+1}, v_2) &= (f, v_2)_{\Omega_2} \qquad \forall v_2 \in V_2^0 && (3.3.35) \\ a_2(u_2^{k+1}, H_2\mu) - (f, H_2\mu)_{\Omega_2} &= -a_1(u_1^{k+1}, H_1\mu) + (f, H_1\mu)_{\Omega_1} \; \forall \mu \in \Lambda \,. && (3.3.36) \end{aligned}$$

Then, with some damping parameter $\vartheta \in (0,1)$, the new iterate is defined by

$$\lambda_2^{k+1} := \vartheta\, u_{2|\Gamma}^{k+1} + (1-\vartheta)\lambda_2^k\,. \tag{3.3.37}$$

As for the linear case in Subsection 3.3.1, one can reformulate the nonlinear Dirichlet–Neumann algorithm (3.3.37) in terms of Steklov–Poincaré operators (3.3.29), including the nonlinear Kirchhoff transformations now. Considering the harmonic extensions $H_i u_{i|\Gamma}^{k+1}$ and the solutions $\mathcal{G}_i f$ of the problems (3.3.24) with homogeneous boundary data for $i = 1, 2$, the intermediate iterates are obtained by

$$u_1^{k+1} = H_1 \kappa_1 \kappa_2^{-1} \lambda_2^k + \mathcal{G}_1 f \quad \text{and} \quad u_2^{k+1} = H_2 u_{2|\Gamma}^{k+1} + \mathcal{G}_2 f\,.$$

Thus, equation (3.3.36) provides

$$\begin{aligned} a_1(H_1\kappa_1\kappa_2^{-1}\lambda_2^k, H_1\mu) &+ a_2(H_2 u_{2|\Gamma}^{k+1}, H_2\mu) \\ &= \sum_{i=1}^{2} (f, H_i\mu)_{\Omega_i} - a_i(\mathcal{G}_i f, H_i\mu) \quad \forall \mu \in \Lambda \end{aligned}$$

which, due to (3.3.29) and (3.3.30), is the same as

$$\langle S_2 u_{2|\Gamma}^{k+1}, \mu\rangle = \langle -S_1\kappa_1\kappa_2^{-1}\lambda_2^k + \chi, \mu\rangle \quad \forall \mu \in \Lambda$$

and, regarding (3.3.37), altogether yields

$$S_2(\lambda_2^{k+1} - \lambda_2^k) = \vartheta(\chi - (S_1\kappa_1\kappa_2^{-1} + S_2)\lambda_2^k) \quad \text{in } \Lambda'\,. \tag{3.3.38}$$

Consequently, the damped Dirichlet–Neumann algorithm applied to (3.3.24)–(3.3.26) is a preconditioned Richardson procedure for the nonlinear Steklov–Poincaré formulation (3.3.32) with S_2 as a preconditioner.

Note that an analogous formulation for the interface equation (3.3.31) cannot be obtained due to the nonlinearity of $S_2\kappa_2$. However, (3.3.38) can be treated just as in the linear case if we apply the following generalization of an abstract convergence result in Quarteroni and Valli [75, pp. 118/119]. Let X be a Hilbert space, let $Q_1 : X \to X'$ be a not necessarily linear operator and let $Q_2 : X \to X'$ be linear and invertible. With the definition $Q := Q_1 + Q_2$ and for given $G \in X'$ we consider the problem

$$\text{find } \lambda \in X : \quad Q\lambda = G \tag{3.3.39}$$

together with the corresponding Richardson iteration

$$\lambda^{k+1} = \lambda^k + \vartheta\, Q_2^{-1}(G - Q\lambda^k) \tag{3.3.40}$$

with the linear operator Q_2 as a preconditioner.

Theorem 3.3.4. *Let Q_2 be continuous and coercive, i.e. there are positive constants β_2 and α_2 such that*

$$\langle Q_2\eta, \mu\rangle \leq \beta_2\|\eta\|_X\|\mu\|_X \ \ \forall \eta, \mu \in X\,, \quad \langle Q_2\eta, \eta\rangle \geq \alpha_2\|\eta\|_X^2 \ \ \forall \eta \in X\,. \quad (3.3.41)$$

Let Q_1 be Lipschitz continuous, i.e. there is a constant $\beta_1 > 0$ such that

$$\langle Q_1\eta - Q_1\mu, \lambda\rangle \leq \beta_1\|\eta - \mu\|_X\|\lambda\|_X \quad \forall \eta, \mu, \lambda \in X\,. \qquad (3.3.42)$$

Suppose there exists a constant $\kappa^* > 0$ such that

$$\langle Q_2(\eta - \mu), Q_2^{-1}(Q\eta - Q\mu)\rangle + \langle Q\eta - Q\mu, \eta - \mu\rangle \geq \kappa^*\|\eta - \mu\|_X^2 \quad \forall \eta, \mu \in X\,. \qquad (3.3.43)$$

Then (3.3.39) has a unique solution $\lambda \in X$. Furthermore, for any given $\lambda^0 \in X$ and any $\vartheta \in (0, \vartheta_{\max})$ with

$$\vartheta_{\max} := \frac{\kappa^*\alpha_2^2}{\beta_2(\beta_1 + \beta_2)^2}\,,$$

the sequence given by (3.3.40) converges in X to λ.

The proof is an application of Banach's fixed point theorem and is carried out here along the lines of the one given in Quarteroni and Valli [75, pp. 119].

Proof. First, the operator Q_2 is invertible as a consequence of (3.3.41) and of the Lax–Milgram lemma (see e.g. [98, pp. 240]). With this observation we introduce the Q_2-scalar product

$$(\eta, \mu)_{Q_2} := \frac{1}{2}(\langle Q_2\eta, \mu\rangle + \langle Q_2\mu, \eta\rangle) \quad \forall \eta, \mu \in X\,.$$

The corresponding Q_2-norm given by

$$\|\eta\|_{Q_2} := (\eta, \eta)_{Q_2}^{1/2} = \langle Q_2\eta, \eta\rangle^{1/2} \quad \forall \eta \in X$$

is equivalent to the norm $\|\cdot\|_X$; actually, it satisfies the two-sided inequality

$$\alpha_2\|\eta\|_X^2 \leq \|\eta\|_{Q_2}^2 \leq \beta_2\|\eta\|_X^2 \quad \forall \eta \in X\,. \qquad (3.3.44)$$

To prove the convergence of the sequence $\{\lambda^k\}_{k\geq 0}$ it is sufficient to show that the mapping

$$T_\vartheta : X \to X\,, \quad T_\vartheta\eta := \eta - \vartheta\, Q_2^{-1}Q\eta \quad \forall \eta \in X\,,$$

is a contraction with respect to the Q_2-norm. With this aim, assuming that $\vartheta \geq 0$, we have for $\eta, \mu \in X$:

$$\begin{aligned} \|T_\vartheta\eta - T_\vartheta\mu\|_{Q_2}^2 &= \|\eta - \mu\|_{Q_2}^2 + \vartheta^2\langle Q\eta - Q\mu, Q_2^{-1}(Q\eta - Q\mu)\rangle \\ &\quad -\vartheta(\langle Q_2(\eta - \mu), Q_2^{-1}(Q\eta - Q\mu)\rangle + \langle Q\eta - Q\mu, \eta - \mu\rangle) \\ &\leq \|\eta - \mu\|_{Q_2}^2 + \vartheta^2\frac{(\beta_1 + \beta_2)^2}{\alpha_2}\|\eta - \mu\|_X^2 - \vartheta\kappa^*\|\eta - \mu\|_X^2\,. \end{aligned}$$

To verify the second term of the estimate (the third one is (3.3.43)), observe that due to (3.3.41) and (3.3.42) we have

$$\langle Q\eta - Q\mu, Q_2^{-1}(Q\eta - Q\mu)\rangle \leq (\beta_1 + \beta_2)\|\eta - \mu\|_X \|Q_2^{-1}(Q\eta - Q\mu)\|_X$$

and that due to (3.3.44) we get

$$\|Q_2^{-1}(Q\eta - Q\mu)\|_X^2 \leq \frac{1}{\alpha_2}\|Q_2^{-1}(Q\eta - Q\mu)\|_{Q_2}^2 = \frac{1}{\alpha_2}\langle Q\eta - Q\mu, Q_2^{-1}(Q\eta - Q\mu)\rangle\,.$$

Now, setting

$$K_\vartheta = 1 + \vartheta^2 \frac{(\beta_1 + \beta_2)^2}{\alpha_2^2} - \vartheta\frac{\kappa^*}{\beta_2}$$

we obtain

$$\|T_\vartheta\eta - T_\vartheta\mu\|_{Q_2}^2 \leq K_\vartheta \|\eta - \mu\|_{Q_2}^2\,.$$

The bound $K_\vartheta < 1$ is guaranteed if $0 < \vartheta < \vartheta_{\max}$. □

Remark 3.3.5. Observe that the proof does not depend on the splitting $Q = Q_1 + Q_2$ if Q_1 in (3.3.42) is replaced by Q. Furthermore, note that condition (3.3.43) reduces to a much simpler expression if, in addition, Q_2 is symmetric. In the linear case (3.3.43) is just the coerciveness of Q or, equivalently, the coerciveness of Q_1. In our nonlinear case (3.3.43) states the *strong monotonicity* of Q (see e.g. [102, p. 501]) which reads

$$\langle Q\eta - Q\mu, \eta - \mu\rangle \geq \frac{\kappa^*}{2}\|\eta - \mu\|_X^2 \quad \forall \eta, \mu \in X\,. \tag{3.3.45}$$

Now, it is well known that in the particular situation of (3.3.32) and (3.3.38) both Steklov–Poincaré operators S_1 and S_2 are symmetric, continuous and coercive, see [75, pp. 8/9]. Thus in order to apply Theorem 3.3.4 to the case $X = \Lambda$, $G = \chi$, $Q_2 = S_2$ and $Q_1 = S_1\kappa_1\kappa_2^{-1}$, we have to make sure that the conditions (3.3.42) and (3.3.45) are satisfied for $Q_1 = S_1\kappa_1\kappa_2^{-1}$. So we arrive at the following

Theorem 3.3.6. *The nonlinear Steklov–Poincaré equation (3.3.32) admits a unique solution λ_2 in Λ to which the nonlinear Dirichlet–Neumann scheme (3.3.33)–(3.3.37) converges for sufficiently small ϑ and any $\lambda_2^0 \in \Lambda$ if the following two conditions are satisfied:*
$\kappa_1\kappa_2^{-1} : \Lambda \to \Lambda$ is Lipschitz continuous, i.e., there is a constant $L(\kappa_1\kappa_2^{-1}) > 0$ such that

$$\|\kappa_1\kappa_2^{-1}\eta - \kappa_1\kappa_2^{-1}\mu\|_\Lambda \leq L(\kappa_1\kappa_2^{-1})\|\eta - \mu\|_\Lambda \quad \forall \eta, \mu \in \Lambda\,, \tag{3.3.46}$$

and $S_1\kappa_1\kappa_2^{-1} : \Lambda \to \Lambda'$ is a strongly monotone operator, i.e. there is a constant $\alpha_1 > 0$ such that

$$\langle S_1(\kappa_1\kappa_2^{-1}\eta - \kappa_1\kappa_2^{-1}\mu), \eta - \mu\rangle \geq \alpha_1\|\eta - \mu\|_\Lambda^2 \quad \forall \eta, \mu \in \Lambda\,. \tag{3.3.47}$$

Note that the conditions (3.3.46) and (3.3.47) do not require $\kappa_i : \Lambda \to \Lambda$, $i = 1, 2$, to be superposition operators or Kirchhoff transformations defined by (3.3.21). In case of linear $\kappa_i : \Lambda \to \Lambda$, condition (3.3.46) reduces to the boundedness of κ_i, $i = 1, 2$, and (3.3.47) is the coercivity of the linear map $S_1\kappa_1\kappa_2^{-1} : \Lambda \to \Lambda'$. Note that (3.3.46) and (3.3.47) are trivially satisfied for linear Kirchhoff transformations corresponding to constant functions k_i as in (3.3.14). However, our abstract theorem also has a concrete relevance for our general setting in problem (3.3.11)–(3.3.13).

Proposition 3.3.7. *The conditions (3.3.46) and (3.3.47) are satisfied in one space dimension.*

Proof. Let $\Omega_1 = [a, b]$, $\Omega_2 = [b, c]$ with $\Gamma = \{b\}$ and $a < b < c$. Then we have $\Lambda = H_{00}^{1/2}(\Gamma) = H^{1/2}(\Gamma) \cong (\mathbb{R}, |\cdot|)$ and condition (3.3.46) follows from Proposition 3.3.1.

Let $L(\kappa_1^{-1})$ and $L(\kappa_2)$ be the Lipschitz constants of the real functions κ_1^{-1} and κ_2 according to Proposition 3.3.1. In order to prove (3.3.47), let $\eta, \mu, \lambda \in \mathbb{R}$. The harmonic extension $H_1(\lambda)$ is the affine function $x \mapsto \frac{\lambda}{b-a}\, x - \frac{\lambda}{b-a}\, a$. As κ_1^{-1} and κ_2 are monotonically increasing, then (3.3.29) provides

$$\begin{aligned}
&\langle S_1(\kappa_1\kappa_2^{-1}\eta - \kappa_1\kappa_2^{-1}\mu), \eta - \mu\rangle \\
&\qquad = \int_a^b \nabla H_1(\kappa_1\kappa_2^{-1}\eta - \kappa_1\kappa_2^{-1}\mu)\nabla H_1(\eta - \mu)\, dx \\
&\qquad = \int_a^b \frac{\kappa_1\kappa_2^{-1}\eta - \kappa_1\kappa_2^{-1}\mu}{b-a} \cdot \frac{\eta - \mu}{b-a}\, dx \\
&\qquad = \frac{(\kappa_1\kappa_2^{-1}\eta - \kappa_1\kappa_2^{-1}\mu)(\eta - \mu)}{b-a} \\
&\qquad \geq \frac{1}{(b-a)L(\kappa_1^{-1})L(\kappa_2)}\, |\eta - \mu|^2 .
\end{aligned}$$

□

Now, we are able to state an existence and uniqueness result for the nonlinear heterogeneous problem (3.3.18)–(3.3.20), which is satisfied at least in one space dimension.

Proposition 3.3.8. *If the conditions (3.3.46) and (3.3.47) are satisfied, then problem (3.3.18)–(3.3.20) is well-posed. Moreover, we have $p_i^k \to p_i$, $k \to \infty$, in V_i for the iterates*

$$p_i^k = \kappa_i^{-1}(H_i\kappa_i\lambda^k + \mathcal{G}_i f) \in V_i\,, \quad i = 1, 2\,, \tag{3.3.48}$$

on Ω_i which correspond via $\lambda^k = \kappa_2^{-1}\lambda_2^k$, $k \geq 0$, to the iterates $(\lambda_2^k)_{k\geq 0}$ of the Dirichlet–Neumann scheme (3.3.33)–(3.3.37).

Proof. The equivalence of problem (3.3.18)–(3.3.20) and (3.3.32) has been obtained in Proposition 3.3.3. We recall from (3.3.27) that the functions

$$p_i = \kappa_i^{-1}(H_i\kappa_i\lambda + \mathcal{G}_i f) \in V_i\,, \quad i = 1, 2\,, \tag{3.3.49}$$

solve (3.3.11)–(3.3.13) if $\lambda_2 = \kappa_2 \lambda$ solves (3.3.32). With Theorem 3.3.6 this gives existence and uniqueness of these solutions for (3.3.18)–(3.3.20).

In order to prove the continuous dependency of p_i, $i = 1, 2$, on the data $f \in L^2(\Omega)$, it is enough to prove the continuous dependency of λ on f since all operators in (3.3.49) are known to be continuous. For $\kappa_i : \Lambda \to \Lambda$ and $\kappa_i^{-1} : V_i \to V_i$ see Proposition 3.3.1, for $H_i : \Lambda \to V_i$ see e.g. [75, p. 9], and for $\mathcal{G}_i : L^2(\Omega_i) \to H_0^1(\Omega_i)$ we refer to [60, p. 82].

To see that $\lambda = \kappa_2^{-1}\lambda_2$ depends continuously on f in (3.3.32) observe first that $\chi \in \Lambda'$ depends continuously on f due to (3.3.30). Secondly, the strong monotonicity (3.3.47) provides the Lipschitz continuity of the inverse operator $(S_1\kappa_1\kappa_2^{-1})^{-1} : \Lambda' \to \Lambda$ and therefore of $(S_1\kappa_1\kappa_2^{-1} + S_2)^{-1}$ in (3.3.32). Consequently, λ_2 and finally λ depend continuously on f. This shows the well-posedness of problem (3.3.18)–(3.3.20).

Now, the claimed convergence of the iterates p_i^k in (3.3.48) to the solutions p_i, $i = 1, 2$, on the subdomains is a consequence of the convergence of $(\lambda_2^k)_{k \geq 0}$ to λ_2 in Λ, the continuity of the Kirchhoff transformations as superposition operators on Λ and V_i, $i = 1, 2$, (Theorem 1.5.15 and Proposition 1.5.17) as well as the continuity of H_i in (3.3.48) for $i = 1, 2$. □

Remark 3.3.9. In Subsection 3.3.5 we are going to carry out the Dirichlet–Neumann iteration (3.3.33)–(3.3.36) for the solution of (3.3.24)–(3.3.26) on a discrete level. In light of Subsection 2.5.1 it is quite clear how to do this. We discretize (3.3.24) as well as (3.3.33) and (3.3.35) by linear finite elements in suitable finite element spaces $\mathcal{S}_j^i$ on Ω_i for $i = 1, 2$, such that the restrictions of the finite element functions from both sides of Γ constitute a common interface space Λ^j of finite elements on Γ (which needs to be polygonal in this case). The Dirichlet transmission conditions (3.3.25) as well as (3.3.34) are imposed on the nodes on Γ only, involving corresponding discrete Kirchhoff transformations $\kappa_{i,j} : \Lambda^j \to \Lambda^j$. The Neumann conditions (3.3.26) as well as (3.3.36) are discretized for all $\mu \in \Lambda^j$ with discrete (linear continuous or, more specifically, harmonic) extension operators $R_{i,j}, H_{i,j} : \Lambda^j \to \mathcal{S}_j^i$ for $i = 1, 2$. Accordingly, discrete Green operators $\mathcal{G}_{i,j}$ and χ_j as in Quarteroni and Valli [75, pp. 46/47] come into play.

Then, due to the coercivity and the continuity of the discrete Steklov–Poincaré operators $S_{i,j} : \Lambda^j \to (\Lambda^j)'$ (with constants independent of $j \geq 0$, consult [75, p. 105/106]), the convergence theory with the corresponding results from Theorem 3.3.6 and Proposition 3.3.7 in the discrete setting is literally the same as in the continuous case. For the proof of (3.3.46) and (3.3.47) in one space dimension we even have $\Lambda^j = \Lambda$, $\kappa_{i,j} = \kappa_i$ and $H_{i,j} = H_i$ for $j \geq 0$ and $i = 1, 2$.

However, in contrast to the continuous setting, the discrete Kirchhoff transformations $\kappa_{i,j} : \Lambda^j \to \Lambda^j$ do not satisfy the weak chain rule (1.5.22). Therefore, the discretized transmission problem (3.3.24)–(3.3.26) is only equivalent to a corresponding retransformed transmission problem similar to (3.3.18)–(3.3.20) with solutions $p_{i,j} = \kappa_i(u_{i,j}) \notin \mathcal{S}_j^i$ and with equality $p_{1,j|\Gamma} = p_{2,j|\Gamma}$ just for the nodes on Γ. Accordingly, Proposition 3.3.8 has to be understood and thus has a satisfying discrete counterpart only for the discretized transformed problem (3.3.24)–(3.3.26).

Recall that we have $\kappa_i = \kappa_{i,j}$ for $i = 1, 2,\ j \geq 0$, in the trivial case of linear Kirchhoff transformations κ_i and that for $\kappa_1 = \kappa_2$ the discretization of (3.3.24)–(3.3.26) can be shown to be equivalent to the corresponding discretized global problem as done in the proof of Theorem 3.2.4. In the general nonlinear case, however, we do not know whether the solutions $u_{i,j}$ of the discretized problem (3.3.24)–(3.3.26) converge to u_i, $i = 1, 2$, for $j \to \infty$.

Remark 3.3.10. Observe that we do not know whether a Dirichlet–Neumann method applied to (3.3.18)–(3.3.20) in the "physical pressure" $\lambda = p_{i|\Gamma}$ converges if the damping parameter is chosen small enough. Due to the nonlinearities $S_i\kappa_i$ for $i = 1, 2$ in the symmetric equation (3.3.31) we do not even get a Steklov–Poincaré formulation for this method.

We note that there is another proof of the convergence for the Dirichlet–Neumann method to be found in Marini and Quarteroni [71] which does not involve Steklov–Poincaré operators directly. Nevertheless, one can also adapt and extend that proof to our situation and obtain Theorem 3.3.6 as well.

Remark 3.3.11. It is possible to formulate a nonlinear Neumann–Neumann method for problem (3.3.24)–(3.3.26) in terms of Steklov–Poincaré operators and involving the Kirchhoff transformations as done in Quarteroni and Valli [75, pp. 14–16] for the linear case. With a $\lambda_1^0 \in \Lambda$ and given positive constants σ_1, σ_2, one obtains the iteration

$$\lambda_1^{k+1} = \lambda_1^k + \vartheta(\sigma_1 S_1^{-1} + \sigma_2 \tilde{S}_2^{-1})(\chi - \tilde{S}\lambda_1^k), \quad k \geq 0, \tag{3.3.50}$$

with $\tilde{S}_2 = S_2\kappa_2\kappa_1^{-1}$ and $\tilde{S} = S_1 + \tilde{S}_2$ referring to the Steklov–Poincaré equation

$$(S_1 + \tilde{S}_2\kappa_2\kappa_1^{-1})\lambda_1 = \chi$$

for the transformed variable $\lambda_1 = \kappa_1\lambda$ on Γ from the first subdomain with λ as in (3.3.31). However, both for (3.3.50) and its symmetric counterpart

$$\lambda^{k+1} = \lambda^k + \vartheta(\sigma_1(S_1\kappa_1)^{-1} + \sigma_2(S_2\kappa_2)^{-1})(\chi - (S_1\kappa_1 + S_2\kappa_2)\lambda^k), \quad k \geq 0,$$

for the untransformed variable λ referring to (3.3.31) we do not have a linear preconditioner such that a direct application of Theorem 3.3.4 is not possible. It is unclear whether this theorem can be extended to nonlinear preconditioners which do not induce a norm in which convergence could be measured.

3.3.4 Counterexamples in 2D for the strong monotonicity of the nonlinear Steklov–Poincaré operator

Let us point out at the beginning of this subsection that we do not know whether the assertions of Theorem 3.3.6 or Proposition 3.3.8 or weaker ones, for example local convergence, hold in higher dimensions if some natural conditions on κ_1 and κ_2, possibly other than (3.3.46) and (3.3.47), are satisfied. Therefore, although this subsection is about counterexamples, we do not have a counterexample for which the Dirichlet–Neumann scheme diverges in higher dimensions. The counterexamples that we talk about here concern our approach to prove the assertions in Theorem 3.3.6 by a contraction argument and to use Theorem 3.3.4 to achieve that. This special approach seems to be ruled out for higher dimensions in which the trace space Λ, in contrast to 1D, is an abstract infinite dimensional function space with a norm that is not so easily treatable in concrete terms (see pages 248–251 in the appendix).

Concerning the question whether the Kirchhoff transformations κ_1 and κ_2 are Lipschitz continuous as superposition operators acting on the trace space Λ, recall that we deduced Proposition 1.5.17 from Lemma 1.5.11, Theorem 1.5.15 and Proposition 1.5.16, i.e. the continuity of κ_i on Λ is a consequence of the continuity of κ_i on $H^1(\Omega_i)$, $i = 1, 2$. But with a glance at the proof of the latter, even for restrictive conditions on k_i (see Proposition 1.5.14), and even more so in the general case treated in Theorem 1.5.15, one will find it extremely unlikely that $\kappa_i : H^1(\Omega_i) \to H^1(\Omega_i)$ and therefore $\kappa_i : \Lambda \to \Lambda$ are Lipschitz continuous.

Much more so, the strong monotonicity (3.3.47) is in general not satisfied in higher dimensions. This is what the two counterexamples in 2D, which we are going to present, are intended to show. Unfortunately, in higher dimensions condition (3.3.47) may be violated to the extent that

$$\langle S_1(\kappa_1\kappa_2^{-1}\eta - \kappa_1\kappa_2^{-1}\mu), \eta - \mu\rangle < 0 \tag{3.3.51}$$

can occur for certain $\eta, \mu \in \Lambda$ and $\kappa_i : \Lambda \to \Lambda$, $i = 1, 2$.

As a starting point for the considerations leading to possible counterexamples observe that in case of $\kappa_i = id : \Lambda \to \Lambda$, condition (3.3.47) is just the coercivity of the Steklov–Poincaré operator

$$\langle S_i\eta, \eta\rangle = \int_{\Omega_i} |\nabla H_i\eta|^2 \, dx \geq c_{\Omega_i} \|H_i\eta\|_{1,\Omega_i}$$

which is a consequence of the Poincaré inequality (see Theorem A.2.5) with a $C_{\Omega_i} > 0$, $i = 1, 2$. It is well known that if η and $\partial\Omega_i$ are smooth enough, then the harmonic extension $H_i\eta$ on Ω_i, $i = 1, 2$, is harmonic in the strong sense (see e.g. Werner [99, pp. 5, 11, 36, 62, 212]) and the classical normal derivatives of $u_i := H_i\eta$ exist on $\Gamma \subset \partial\Omega_i$, i.e. we have

$$\langle S_i\eta, \eta\rangle = \int_\Gamma \frac{\partial u_i}{\partial \mathbf{n}_i} \cdot u_i \, d\sigma > 0 \tag{3.3.52}$$

for $i = 1, 2$ due to Green's formula (1.5.9). Now, the crucial observation is that although (3.3.52) holds for any harmonic function u_i on Ω_i which is in $C^2(\overline{\Omega}_i)$, there may well be subsets $\Gamma_- \subset \Gamma$ of positive Hausdorff measure on which

$$\frac{\partial u_i}{\partial \mathbf{n}_i} \cdot u_i < 0 \tag{3.3.53}$$

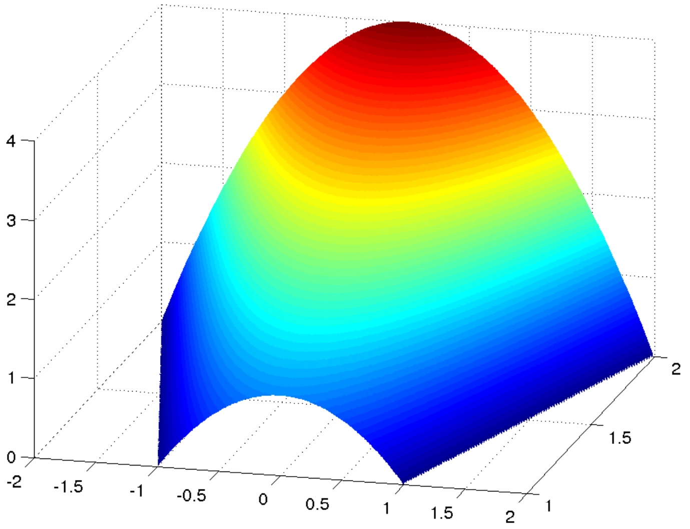


Figure 3.3: $u : (x, y) \mapsto -x^2 + y^2$ satisfying (3.3.53) between $(-1, 1)$ and $(1, 1)$

is possible. This fact is exploited in the construction of the counterexamples. Loosely speaking, with a certain choice of a Kirchhoff transformation one can provide a bigger weight to the product on the left hand side of (3.3.53) such that altogether one can obtain (3.3.51).

We start with an analytical counterexample based on the harmonic function

$$u : (x, y) \mapsto -x^2 + y^2$$

given on the trapezium $T \subset \mathbb{R}^2$ with the vertices $(-1, 1)$, $(1, 1)$, $(2, 2)$ and $(-2, 2)$, see Figure 3.3. For $i = 1, 2$ let Γ_i be the edge between the vertices $(-i, i)$ and (i, i), not containing the vertices in both cases. We compute

$$\frac{\partial u}{\partial \mathbf{n}} = -\frac{\partial u}{\partial y} = -2y = -2 \quad \text{on } \Gamma_1 , \qquad \frac{\partial u}{\partial \mathbf{n}} = \frac{\partial u}{\partial y} = 2y = 4 \quad \text{on } \Gamma_2$$

and, moreover,

$$\begin{aligned} \int_{\partial T} \frac{\partial u}{\partial \mathbf{n}} \cdot u \, d\sigma &= \int_{\Gamma_1} \frac{\partial u}{\partial \mathbf{n}} \cdot u \, d\sigma + \int_{\Gamma_2} \frac{\partial u}{\partial \mathbf{n}} \cdot u \, d\sigma \\ &= \int_{-1}^{1} (-2) \cdot (-x^2 + 1) \, dx + \int_{-2}^{2} 4 \cdot (-x^2 + 4) \, dx \\ &= -2 \, \frac{2}{3} + 29 \, \frac{1}{3} = 26 \, \frac{2}{3} , \end{aligned} \tag{3.3.54}$$

i.e. we have $\Gamma_- = \Gamma_1$ on which (3.3.53) holds in this example. Now, we choose a $\kappa : \mathbb{R} \to \mathbb{R}$ intended to replace $(\kappa_1 \kappa_2^{-1})^{-1}$ in (3.3.51) and some $\eta, \mu \in H^{1/2}(\partial T)$ with $u_{|\partial T} = \eta - \mu$ such that

$$\langle S_1(\eta - \mu), \kappa\eta - \kappa\mu \rangle = \int_{\partial T} \frac{\partial u}{\partial \mathbf{n}} \cdot (\kappa\eta - \kappa\mu) \, d\sigma < 0 .$$

Let $\mu_{|\Gamma_1} := 0$, $\mu_{|\Gamma_2} := 1$ and $\eta_{|\Gamma_1} := u_{|\Gamma_1}$, $\eta_{|\Gamma_2} := u_{|\Gamma_2} + 1$, and extend both μ and η linearly on the edges between $(-1,1)$ and $(-2,2)$ as well as between $(1,1)$ and $(2,2)$. Furthermore, we define $\kappa := id$ on $(-\infty, 1]$ and

$$\kappa : x \mapsto ax + 1 - a \quad \forall x \in [1, \infty) \tag{3.3.55}$$

for a constant $a > 0$. As a consequence, we have

$$\langle S_1(\eta - \mu), \kappa\eta - \kappa\mu \rangle = \int_{\partial T} \frac{\partial u}{\partial \mathbf{n}} \cdot (\kappa\eta - \kappa\mu)\, d\sigma = \int_{\Gamma_1} \ldots + \int_{\Gamma_2} \ldots + \int_{\partial T \backslash (\Gamma_1 \cup \Gamma_2)} \ldots$$

in which the last integrand vanishes since $\eta = \mu$ on $\partial T \backslash (\Gamma_1 \cup \Gamma_2)$. Moreover, we still have

$$\int_{\Gamma_1} \frac{\partial u}{\partial \mathbf{n}} \cdot (\kappa\eta - \kappa\mu)\, d\sigma = -2\frac{2}{3}$$

as in (3.3.54) and

$$\int_{\Gamma_2} \frac{\partial u}{\partial \mathbf{n}} \cdot (\kappa\eta - \kappa\mu)\, d\sigma = 4 \int_{\Gamma_2} \kappa\eta - \kappa\mu\, d\sigma = 4 \int_{\Gamma_{id}} \eta - \mu\, d\sigma + 4 \int_{\Gamma_2 \backslash \Gamma_{id}} \kappa\eta - \kappa\mu\, d\sigma$$

with $\Gamma_{id} := \{(x,y) \in \Gamma_2 : -x^2 + y^2 \leq 1\}$ where $\kappa = id$. Since $u = -x^2 + 4 \leq 1$ on Γ_{id} is satisfied if and only if $-2 < x \leq -\sqrt{3}$ or $x \geq \sqrt{3} < 2$ holds, we can estimate

$$\int_{\Gamma_1 \cup \Gamma_{id}} \frac{\partial u}{\partial \mathbf{n}} \cdot (\kappa\eta - \kappa\mu)\, d\sigma \leq 4 \cdot 2 \cdot (2 - \sqrt{3}) \cdot 1 - 2\frac{2}{3} < 0$$

and due to (3.3.55) we get

$$\begin{aligned} \int_{\Gamma_2 \backslash \Gamma_{id}} \frac{\partial u}{\partial \mathbf{n}} \cdot (\kappa\eta - \kappa\mu)\, d\sigma &= \int_{-\sqrt{3}}^{\sqrt{3}} 4(a(5 - x^2) + 1 - a - 1)\, dx \\ &= a \int_{-\sqrt{3}}^{\sqrt{3}} 4(4 - x^2)\, dx \longrightarrow 0 \quad \text{for} \quad a \to 0\,. \end{aligned}$$

This example is certainly somewhat artificial, mostly because the interface Γ is assumed to be a superset of $\Gamma_1 \cup \Gamma_2$ here, i.e. it is either quite large or disconnected. Nevertheless, such a case occurs if $T = \Omega_1$ is contained in the interior of a bounded Lipschitz domain $\Omega \subset \mathbb{R}^2$ so that the complement $\Omega_2 = \Omega \backslash T$ satisfies $\partial\Omega_2 = \partial\Omega$. However, in contrast to the situation in Figure 3.2, such a decomposition of Ω provides the trace space $\Lambda = H^{1/2}(\Gamma)$. Note that in this case, $T = \Omega_1$ must be chosen as the subdomain for the Dirichlet problems in a Dirichlet–Neumann method similar to (3.3.33)–(3.3.37) since otherwise pure indeterminate Neumann problems occur.

Alternatively, one can consider $T = \Omega_1$ as a subset of a ring-shaped domain with $\overline{\Omega}_1 \cap \overline{\Omega}_2 = \Gamma_1 \cup \Gamma_2$. Note, however, that we have $\eta_{|\Gamma_2}, \mu_{|\Gamma_2} \notin H^{1/2}_{00}(\Gamma_2)$ in the counterexample. In order to still apply the considerations above, one can think of the following modification for arbitrary $\varepsilon > 0$. Choose η and μ to be constantly 0 on the edges between $(-1,1)$ and $(-2+\varepsilon, 2-\varepsilon)$ as well as

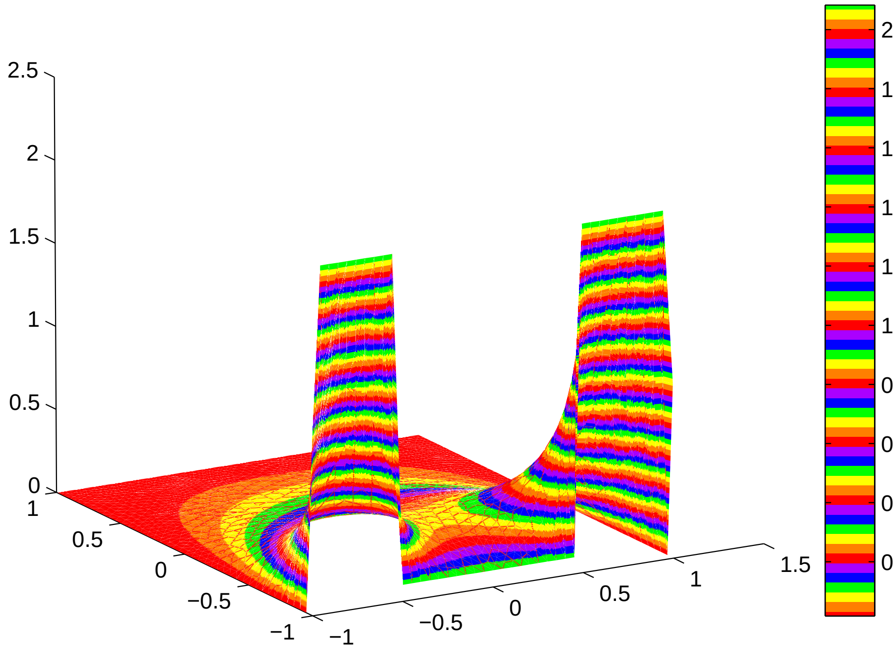


Figure 3.4: Harmonic function satisfying (3.3.53) around $(0, -1)$

between $(1, 1)$ and $(2-\varepsilon, 2-\varepsilon)$, and then to be linear between the value 0 and the value 1 which has to be assumed on the vertices $(-2, 2)$ and $(2, 2)$. At least this construction gives $\eta_{|\tilde{\Gamma}}, \mu_{|\tilde{\Gamma}} \in H_{00}^{1/2}(\tilde{\Gamma})$ with a subset $\tilde{\Gamma} \subset \partial T$ contained in an arbitrarily small neighbourhood of Γ_2.

In addition to what has been considered above, there are also counterexamples on domains $\Omega \subset \mathbb{R}^2$ for the case of connected and smaller interfaces $\Gamma \subset \partial\Omega$. Figure 3.4 illustrates such an example on $\Omega = [-1, 1] \times [-1, 1]$ in which we have chosen $\Gamma = [-1, 1] \times \{-1\}$. The plot shows an approximation of the harmonic function u on Ω with piecewise linear and continuous Dirichlet boundary values satisfying $u_{|\partial\Omega\backslash\Gamma} = 0$ and

$$u(x, -1) = \operatorname{sign}(x^2 - 0.25) + 1.1 \quad \forall x \in \tilde{\Gamma}$$

on a subset $\tilde{\Gamma} \subset \Gamma$ for which $\Gamma\backslash\tilde{\Gamma}$ is a small neighbourhood of the points $(-1, 1)$, $(-0.5, -1)$, $(0.5, -1)$ and $(1, -1)$. Although we do not have an analytical expression for u on Ω and we cannot prove that the normal derivative of u across the connected component Γ' of $\tilde{\Gamma}$ containing $(0, -1)$ is negative, it is quite obvious from the plot in Figure 3.4 that this is the case. Consequently, (3.3.53) holds on Γ', which is the basis for a similar construction as above for our analytical example, leading to the non-monotonicity (3.3.51).

3.3.5 Numerical example in 2D

The purpose of this subsection is to apply our nonlinear Dirichlet–Neumann method to a problem in two space dimensions. Although the counterexamples in Subsection 3.3.4 rule out a similar proof for the convergence of this method in a 2D-setting as done in 1D for Proposition 3.3.7, it turns out that the method works quite well in this case, too.

We consider the transmission problem (3.3.11)–(3.3.13) on the Yin Yang domain Ω within a circle of radius 1 as shown in Figure 3.5. We denote the white subdomain together with the grey circle B_1 by Ω_1 and the grey subdomain with the white circle B_2 by Ω_2.

Furthermore, we select the data

$$f(x) = (-1)^i \quad \text{on } B_i\,,\ \ i = 1, 2\,, \qquad f(x) = 0 \quad \text{elsewhere} \tag{3.3.56}$$

and the nonlinearities

$$k_i(p_i) = \begin{cases} K_h \max\{(-p_i)^{-3\lambda_i - 2},\, c\} & \text{for } p_i \leq -1 \\ 1 & \text{for } p_i \geq -1 \end{cases} \tag{3.3.57}$$

with certain parameters λ_i and $c > 0$ to be specified below.

This choice is motivated by the state equations of Brooks–Corey and Burdine for the relative permeability in saturated-unsaturated porous media with different soils that we introduced in (1.2.9)–(1.2.11), see also Subsection 1.4.3. In this way, our model problem can be regarded as a nondegenerate stationary Richards equation without gravity in a heterogeneous setting as in Subsection 3.2.3.

Note that $p_i < -1$ characterizes the unsaturated region which is separated by a free boundary from the linear, saturated regime occurring for $p_i \geq -1$. Here, the adimensional entity -1 represents the bubbling pressure p_b, see Section 1.3. In agreement with realistic hydrological data as in Subsection 1.4.1 we choose it to be equivalent to the negative pressure of 0.1 meters of a water column.

We recall from Section 1.2 that the parameters λ_1 and λ_2 in Ω_1 and Ω_2, respectively, are known as pore size distribution factors. According to Rawls et al. [77, Table 5.3.2] we choose them in an extreme manner as

$$\begin{aligned} \lambda_1 &= 1.0 \quad \text{(coarse sand)} \\ \lambda_2 &= 0.1 \quad \text{(fine clay)}\,. \end{aligned} \tag{3.3.58}$$

The factor $K_h = 0.002$ is a realistic hydraulic conductivity in the case of full saturation (see Subsection 1.4.1). The parameter $c = 0.1 > 0$ is introduced to enforce ellipticity (compare Subsection 1.4.3). The convergence results worsen if c is chosen smaller and for $c = 0$ we do not observe convergence of the method.

The choice of the data f which results in a strong sink in B_1 and a strong source in B_2 (due to the small value of K_h) and our special choice of Ω_1

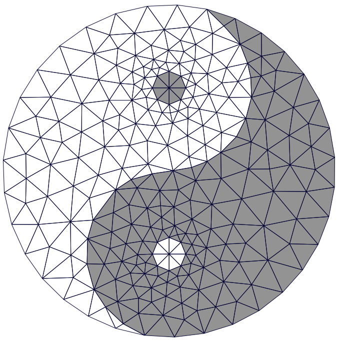


Figure 3.5: Yin Yang domain Ω

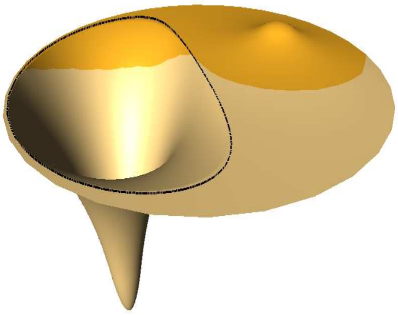


Figure 3.6: Solution p on Ω with free boundary (black line)

and Ω_2 ensure that the free boundary has a nontrivial intersection with the interface $\Gamma = \overline{\Omega}_1 \cap \overline{\Omega}_2$. Since we apply the Dirichlet–Neumann scheme (3.3.33)–(3.3.37), we hereby make sure that step (3.3.34) is nonlinear.

The free boundary can be seen in Figure 3.6 which shows the numerical solution p on Ω obtained on the 6th level with the ranges $[-56.1, 0.9]$ in Ω_1 and $[-7.3, 3.0]$ in Ω_2. Here, the pressure of one meter of a water column is used as the unit.

We discretize the problem using piecewise linear finite element spaces on each of the two subdomains, see Remark 3.3.9. The linear problems on the subdomains are solved by a linear multigrid method which occurs as the method discussed in Section 2.7 when there are no nonlinearities.

In contrast to Section 2.8, the implementation for this test case has been performed in the numerics environment DUNE [12] using the grid manager from UG [11]. This also applies to all the following numerical examples presented in this work. For the visualization of the corresponding results we make use of the toolbox AMIRA [89].

Figure 3.7 shows the average convergence rates ρ of the Dirichlet–Neumann iteration with respect to the damping parameter ϑ on the first 6 levels as indicated by numbers at the graphs of the functions. The convergence rates are given with respect to the transformed variables u_i^k and are measured in the energy norms $a_i(\cdot,\cdot)^{1/2}$ on V_i for $i = 1, 2$, which are induced by the stiffness matrices on the relevant finite element spaces.

More precisely, with initial iterates $u_i^0 = 0$ for $i = 1, 2$, the Dirichlet–Neumann iteration is carried out until the relative error satisfies

$$\frac{\left(\sum_{i=1}^{2} a_i(u_i^n - u_i^{n-1}, u_i^n - u_i^{n-1})\right)^{1/2}}{\left(\sum_{i=1}^{2} a_i(u_i^{n-1}, u_i^{n-1})\right)^{1/2}} < 10^{-12} \tag{3.3.59}$$

for some $n \geq 0$. Then we calculate ρ as the maximum of the geometric means

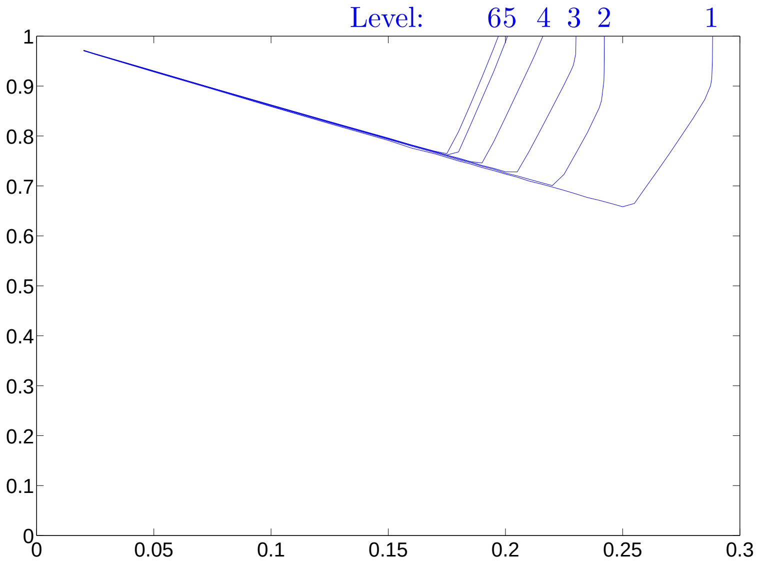


Figure 3.7: ρ vs. damping parameter ϑ on levels 1 to 6

of the rates

$$\frac{\left(\sum_{i=1}^{2} a_i(u_i^k - u_i^n, u_i^k - u_i^n)\right)^{1/2}}{\left(\sum_{i=1}^{2} a_i(u_i^{k-1} - u_i^n, u_i^{k-1} - u_i^n)\right)^{1/2}} \tag{3.3.60}$$

for $1 \leq k \leq \tilde{n}$ over all $\tilde{n} < n$ (note that we get zero for $\tilde{n} = n$). Each of the local problems on the subdomains is solved by 50 multigrid iterations which leads to numerically exact solutions.

We use a grid hierarchy of 7 levels resulting from a uniform mesh refinement of the coarse grid with 169 nodes depicted in Figure 3.5. In this way, we obtain about 940,000 nodes on the finest mesh on Ω corresponding to about 938,000 unknowns on the 7th level.

As can be seen in Figure 3.7, we need quite small damping parameters for the Dirichlet–Neumann iteration to converge. Moreover, the convergence rate as a function of the damping parameter depends on the refinement level: On higher levels more damping is necessary in order to obtain convergence at all. Furthermore, the finer the grid the smaller the optimal damping parameter ϑ_{opt} and the bigger the optimal convergence rate $\rho_{opt} = \rho(\vartheta_{opt})$ become.

However, this effect seems to stabilize on higher levels. Figures 3.8 and 3.9 show the dependency of ϑ_{opt} and ρ_{opt}, respectively, with respect to the levels 1 to 7. It turns out that we have $\vartheta_{opt} = 0.175$ on the levels 5 (with $\rho_{opt} = 0.762$) and 6 (with $\rho_{opt} = 0.765$) and $\vartheta_{opt} = 0.17$ on level 7 (with $\rho_{opt} = 0.770$).

In Berninger et al. [16], where this numerical example has first been addressed, the situation concerning the 5th level in Figure 3.7 and the constant convergence

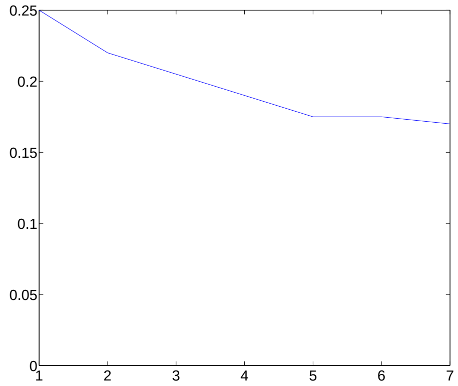


Figure 3.8: ϑ_{opt} vs. refinement level

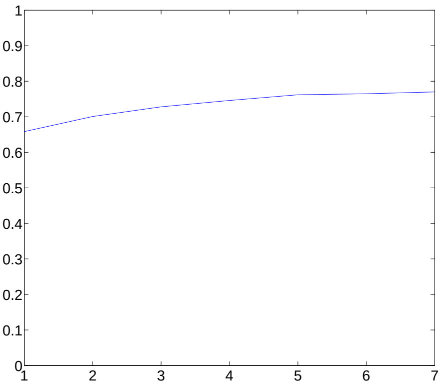


Figure 3.9: ρ_{opt} vs. refinement level

rates for $\vartheta_{opt} = 0.175$ on levels 1 to 6 have been presented. Now, Figure 3.7 indicates that level-independence of the convergence rates is obtained if and only if the damping parameter is at most ϑ_{opt} corresponding to the finest level considered.

3.4 Nonlinear Robin method

In the last section we discussed a Dirichlet–Neumann method for elliptic transmission problems related to the nondegenerate stationary Richards equation without gravity. This section is devoted to the discussion of Robin's method for a larger class of non-overlapping nonlinear domain decomposition problems on two subdomains containing the problems treated in Section 3.3 but which also includes our implicit-explicitly time-discretized Richards equation (3.2.19) in the nondegenerate case (see Subsection 1.4.3). We carry out a similar analysis for this problem class as done in Section 3.3 although our considerations need to be more sophisticated this time.

First, in Subsection 3.4.1, we present our domain decomposition problem both in the physical and in the Kirchhoff–transformed variables and we introduce the nonlinear Robin method for an iterative solution of this problem. It turns out that the Robin boundary value problems that occur in the iteration are uniquely solvable if natural conditions hold.

In Subsection 3.4.2 we derive an equivalent formulation of our domain decomposition problem in terms of an interface equation involving nonlinear Steklov–Poincaré operators. Furthermore, the Robin method can be expressed by these operators which leads us to a nonlinear ADI method for the solution of the interface equation. With regard to a convergence analysis we also introduce an altered version of this ADI method which is equal to the latter in one space dimension.

In Subsection 3.4.3 we develop a convergence analysis for the altered ADI or the altered Robin method which extends existing results for the linear case in Discacciati [33, Chapter 5]. Based on a contraction argument as in Subsection 3.3.3 we get the same sufficient conditions on the Steklov–Poincaré operators as obtained there (namely Lipschitz continuity and strong monotonicity) which ensure convergence.

Finally, in Subsection 3.4.4, we examine these sufficient conditions in more detail in one space dimension for the nondegenerate Richards equation in heterogeneous soil. This leads us to the convergence of the Robin method and the well-posedness of the original domain decomposition problem if natural conditions on the nonlinearities are satisfied. Unfortunately, the same counterexamples as given in Subsection 3.3.4 also apply in the case and the method of proof considered here.

Subsection 3.4.5 contains a short presentation of the space discretization and the numerical treatment of the subproblems occurring in the Robin iteration procedure for the Richards equation (still neglecting gravity which is discussed in Section 4.2). Basically, this can be carried out analogously as presented in Chapter 2. Moreover, analogously as in Subsections 3.4.3 and 3.4.4, a convergence result is obtained for the discrete Robin method in one space dimension.

This section ends with numerical tests for the Robin method in two space dimensions which can be found in Subsection 3.4.6. In a first part we compare the Robin method with the Dirichlet–Neumann method when applied to the Yin Yang example in Subsection 3.3.5. In a second part of Subsection 3.4.6 the Robin method is applied to a time-dependent case of the Richards equation without gravity in heterogeneous soil. We obtain reasonable results for the performance of the Robin method in both cases.

3.4.1 Robin method for elliptic problems related to the time-discretized nondegenerate Richards equation

In this subsection we introduce our class of transmission problems with a focus on the Richards equation. As already seen before, for example in Subsection 3.3.2, these problems are transformed by Kirchhoff transformations. We give both strong and weak formulations. Then we introduce the Robin method for such problems, which we discuss in the sequel and for which we give an existence and uniqueness result concerning the solvability of the subproblems. For the latter we need to apply and extend the theory on convex minimization problems which we presented in Section 2.3.

As in Subsection 3.3.2 suppose $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain and $f \in L^2(\Omega)$. We consider a non-overlapping decomposition of Ω in Ω_1 and Ω_2 as in Figure 3.2 with a Lipschitz continuous interface $\Gamma = \overline{\Omega}_1 \cap \overline{\Omega}_2$. In light of Theorem 3.2.4, Definition 3.2.7 gives sense to certain problems of finding a function p on Ω with $p_{|\partial\Omega} = 0$ such that

$$\theta(x,p) - \operatorname{div}\left(kr(x,\theta(p))\nabla p\right) = f \quad \text{in } \Omega\,. \tag{3.4.1}$$

Such problems arise from an implicitly time-discretized Richards equation with an explicit treatment of the gravitational term as in (3.2.19), and with a saturation θ and a relative permeability kr which are space-independent on the subdomains only and may jump across Γ. Analogously as in Subsection 3.3.2, we give a strong formulation first.

Given real nonnegative monotonically increasing functions θ_1 and θ_2 and functions $k_1, k_2 \in L^\infty(\mathbb{R})$ satisfying $k_1, k_2 \geq \alpha$ with an $\alpha > 0$ we consider the following domain decomposition problem:

Find a function p on Ω, $p_{|\Omega_i} = p_i \in H^1(\Omega_i)$, $i = 1, 2$, $p_{|\partial\Omega} = 0$, such that

$$\begin{aligned} \theta_i(p_i) - \operatorname{div}(k_i(p_i)\nabla p_i) &= f && \text{on } \Omega_i\,,\ i = 1,2 && (3.4.2)\\ p_1 &= p_2 && \text{on } \Gamma && (3.4.3)\\ k_1(p_1)\nabla p_1 \cdot \mathbf{n} &= k_2(p_2)\nabla p_2 \cdot \mathbf{n} && \text{on } \Gamma\,. && (3.4.4) \end{aligned}$$

If we ignore gravity and if $\theta_1 = \theta_2$ and $k_1 = k_2$, then Theorem 3.2.4 provides the equivalence of (3.4.2)–(3.4.4) and the corresponding global problem (3.4.1). If gravity is included (explicitly as always in here), one should actually consider

$$(k_1(p_1)\nabla p_1 - k_1(\tilde{p}_1)e_z) \cdot \mathbf{n} = (k_2(p_2)\nabla p_2 - k_2(\tilde{p}_2)e_z) \cdot \mathbf{n} \qquad \text{on } \Gamma \tag{3.4.5}$$

with the solutions $\tilde{p}_i$ on Ω_i, $i = 1, 2$, from the previous time step instead of (3.4.4) since we have $k_1(\tilde{p}_1) \neq k_1(\tilde{p}_1)$ in general due to $k_1 \neq k_2$ even though $\tilde{p}_1 = \tilde{p}_2$ holds on Γ. This is hydrologically reasonable and mathematically indicated (see (3.2.18)). Therefore, we point out here that all the results in this section can easily be extended to the more general transmission conditions (3.4.5) which are discussed in more detail in Subsection 3.4.5, see also Remark 3.4.3.

An application of Kirchhoff's transformation (3.3.21) to (3.4.2)–(3.4.4), separately in the two subdomains Ω_1 and Ω_2 with the transformed functions $M_i := \theta_i \circ \kappa_i^{-1}$ (see the saturation of the generalized pressure (1.3.2)), gives the following transformed problem:

Find a function u on Ω, $u_{|\Omega_i} = u_i \in H^1(\Omega_i)$, $i = 1, 2$, $u_{|\partial\Omega} = 0$, such that

$$\begin{aligned} M_i(u_i) - \Delta u_i &= f && \text{on } \Omega_i\,,\ i = 1,2 && (3.4.6)\\ \kappa_1^{-1}u_1 &= \kappa_2^{-1}u_2 && \text{on } \Gamma && (3.4.7)\\ \frac{\partial u_1}{\partial \mathbf{n}} &= \frac{\partial u_2}{\partial \mathbf{n}} && \text{on } \Gamma\,. && (3.4.8) \end{aligned}$$

Observe that just as θ_i the transformed M_i are nonnegative, monotonically increasing functions on the real line for $i = 1, 2$. Additionally, due to the definition of Kirchhoff's transformation we have $u_i = 0 \Leftrightarrow p_i = 0$.

In order to obtain weak formulations of the transmission problems (3.4.2)–(3.4.4) and (3.4.6)–(3.4.8) we use the definition (3.2.2) of the spaces V_i, V_i^0 for $i = 1, 2$

and the trace space Λ. Furthermore, we need the bilinear forms as defined in (3.3.15)–(3.3.17) and the L^2-scalar product

$$(\eta, \mu)_\Gamma := \int_\Gamma \eta\mu \, d\sigma \quad \forall \eta, \mu \in \Lambda$$

on the trace space. The norm in $L^2(\Omega_i)$ will be denoted by $\|\cdot\|_{0,\Omega_i}$, the norm in $H^1(\Omega_i)$ by $\|\cdot\|_{1,\Omega_i}$ and the norm in Λ by $\|\cdot\|_\Lambda$. Finally, as in Subsection 3.3.2, let R_i, $i = 1, 2$, be any linear continuous extension operator from Λ to V_i.

Now, with the help of Green's formula (1.5.9) it is easy to see that the weak form of (3.4.2)–(3.4.4) reads as follows:

Find $p_i \in V_i$, $i = 1, 2$, such that

$$\begin{aligned} (\theta_i(p_i), v_i)_{\Omega_i} + b_i(p_i, v_i) &= (f, v_i)_{\Omega_i} \qquad \forall v_i \in V_i^0, \;\; i = 1, 2 \qquad (3.4.9) \\ p_{1|\Gamma} &= p_{2|\Gamma} \qquad \text{in } \Lambda \qquad (3.4.10) \end{aligned}$$

$$\begin{aligned} &(\theta_1(p_1), R_1\mu)_{\Omega_1} + b_1(p_1, R_1\mu) - (f, R_1\mu)_{\Omega_1} = \\ &\qquad - (\theta_2(p_2), R_2\mu)_{\Omega_2} - b_2(p_2, R_2\mu) + (f, R_2\mu)_{\Omega_2} \quad \forall \mu \in \Lambda \,. \qquad (3.4.11) \end{aligned}$$

Analogously, the weak formulation of the transformed problem (3.4.6)–(3.4.8) reads:

Find $u_i \in V_i$, $i = 1, 2$, such that

$$\begin{aligned} (M_i(u_i), v_i)_{\Omega_i} + a_i(u_i, v_i) &= (f, v_i)_{\Omega_i} \qquad \forall v_i \in V_i^0, \;\; i = 1, 2 \qquad (3.4.12) \\ \kappa_1^{-1}(u_{1|\Gamma}) &= \kappa_2^{-1}(u_{2|\Gamma}) \qquad \text{in } \Lambda \qquad (3.4.13) \end{aligned}$$

$$\begin{aligned} &(M_1(u_1), R_1\mu)_{\Omega_1} + a_1(u_1, R_1\mu) - (f, R_1\mu)_{\Omega_1} = \\ &\qquad - (M_2(u_2), R_2\mu)_{\Omega_2} - a_2(u_2, R_2\mu) + (f, R_2\mu)_{\Omega_2} \quad \forall \mu \in \Lambda. \qquad (3.4.14) \end{aligned}$$

Proposition 3.4.1. *With our assumptions on θ_i and k_i, $i = 1, 2$, the domain decomposition problem (3.4.9)–(3.4.11) is equivalent to its transformed version (3.4.12)–(3.4.14).*

The proof is essentially the same as for Theorem 1.5.18 which already holds if we have $kr \circ \theta \in L^\infty(\Omega)$. We just recall that kr is monotonically increasing in hydrologically realistic situations and that, nonetheless, our assumptions on k_i, $i = 1, 2$, are enough for the weak chain rule (1.5.22) to hold. We also refer to Remark 3.3.2 for an explanation that Proposition 3.4.1 is not as straightforward as it seems to be since the commutativity $\kappa_i^{-1}(u_{i|\Gamma}) = \kappa_i^{-1}(u_i)_{|\Gamma}$, $i = 1, 2$, is not trivial.

For a general analysis of the problem (3.4.6)–(3.4.8), which is of course carried out for the weak form (3.4.12)–(3.4.14), we first assume that some transformations $\kappa_1, \kappa_2 : \Lambda \to \Lambda$ and monotonically increasing functions $M_1, M_2 : \mathbb{R} \to \mathbb{R}$

are given. We will impose conditions on κ_1, κ_2 and M_1, M_2 that ensure convergence of a refined version of the Robin method applied to (3.4.6)–(3.4.8). In analogy to Quarteroni and Valli [75, p. 16] the classical strong formulation of the Robin method reads as follows:

Given positive parameters γ_1 and γ_2 and an initial iterate $u_2^0 \in V_2$ find successively $u_1^{k+1} \in V_1$ and $u_2^{k+2} \in V_2$ for $k \geq 0$ such that

$$\begin{aligned} M_1(u_1^{k+1}) - \Delta u_1^{k+1} &= f && \text{on } \Omega_1 && (3.4.15) \\ \frac{\partial u_1^{k+1}}{\partial \mathbf{n}} + \gamma_1 \, \kappa_1^{-1} u_1^{k+1} &= \frac{\partial u_2^k}{\partial \mathbf{n}} + \gamma_1 \, \kappa_2^{-1} u_2^k && \text{on } \Gamma && (3.4.16) \end{aligned}$$

and then

$$\begin{aligned} M_2(u_2^{k+1}) - \Delta u_2^{k+1} &= f && \text{on } \Omega_2 && (3.4.17) \\ \frac{\partial u_2^{k+1}}{\partial \mathbf{n}} - \gamma_2 \, \kappa_2^{-1} u_2^{k+1} &= \frac{\partial u_1^{k+1}}{\partial \mathbf{n}} - \gamma_2 \, \kappa_1^{-1} u_1^{k+1} && \text{on } \Gamma \, . && (3.4.18) \end{aligned}$$

Expressed for the original nontransformed problem (3.4.2)–(3.4.4) related to the Richards equation the Robin condition (3.4.16) translates into

$$k_1(p_1^{k+1}) \nabla p_1^{k+1} \cdot \mathbf{n} + \gamma_1 \, p_1^{k+1} = k_2(p_2^k) \nabla p_2^k \cdot \mathbf{n} + \gamma_1 \, p_2^k \quad \text{on } \Gamma \qquad (3.4.19)$$

and analogously for (3.4.18). As a consequence, for a fixed point $\bar{u} = \kappa(\bar{p})$ of the iteration both the physical pressure and the part $k_i(\bar{p}_{|\Omega_i}) \nabla \bar{p}_{|\Omega_i} \cdot \mathbf{n}$, $i = 1, 2$, of the physical water flux across Γ are continuous. Therefore, $\bar{p}$ would be a solution of (3.4.2)–(3.4.4), see also Remark 3.4.3.

With the help of Green's formula (1.5.9) the weak form of the Robin method (3.4.15)–(3.4.18) reads as follows: Given a $u_2^0 \in V_2$ find successively $u_1^{k+1} \in V_1$ and $u_2^{k+2} \in V_2$ for $k \geq 0$ such that

$$(M_1(u_1^{k+1}), v_1)_{\Omega_1} + a_1(u_1^{k+1}, v_1) = (f, v_1)_{\Omega_1} \quad \forall v_1 \in V_1^0 \qquad (3.4.20)$$

$$\begin{aligned} &(M_1(u_1^{k+1}), R_1\mu)_{\Omega_1} + a_1(u_1^{k+1}, R_1\mu) - (f, R_1\mu)_{\Omega_1} + \gamma_1(\kappa_1^{-1} u_1^{k+1}, \mu)_\Gamma = \\ &\quad - (M_2(u_2^k), R_2\mu)_{\Omega_2} - a_2(u_2^k, R_2\mu) + (f, R_2\mu)_{\Omega_2} + \gamma_1(\kappa_2^{-1} u_2^k, \mu)_\Gamma \quad \forall \mu \in \Lambda \end{aligned} \qquad (3.4.21)$$

and then

$$(M_2(u_2^{k+1}), v_2)_{\Omega_2} + a_2(u_2^{k+1}, v_2) = (f, v_2)_{\Omega_2} \quad \forall v_2 \in V_2^0 \qquad (3.4.22)$$

$$\begin{aligned} &(M_2(u_2^{k+1}), R_2\mu)_{\Omega_2} + a_2(u_2^{k+1}, R_2\mu) - (f, R_2\mu)_{\Omega_1} + \gamma_2(\kappa_2^{-1} u_2^{k+1}, \mu)_\Gamma = \\ &-(M_1(u_1^{k+1}), R_1\mu)_{\Omega_1} - a_1(u_1^{k+1}, R_1\mu) + (f, R_1\mu)_{\Omega_1} + \gamma_2(\kappa_1^{-1} u_1^{k+1}, \mu)_\Gamma \quad \forall \mu \in \Lambda \, . \end{aligned} \qquad (3.4.23)$$

Note that the alternating signs in front of γ_1 and γ_2 are also respected in the weak formulation because the outward normal on Γ with respect to Ω_2 in (3.4.18) is $-\mathbf{n}$ which has to be considered for (3.4.23). Of course, we first have to check whether this method consists of well-posed subproblems generating uniquely determined iterates. In order to clarify this, we need to apply and extend the theory of convex minimization problems given in Section 2.3 to convex functionals given by "convex" superposition operators on $\Gamma \subset \partial\Omega$ instead of Ω.

Theorem 3.4.2. *Let $M_i : \mathbb{R} \to \mathbb{R}$ be monotonically increasing and either bounded or else Hölder continuous outside of a bounded interval. Furthermore, let $\kappa_i : \Lambda \to \Lambda$ be a superposition operator arising from Kirchhoff's transformation (3.3.21) with $k_i \in L^\infty(\mathbb{R})$ satisfying $k_i \geq \alpha > 0$ for $i = 1, 2$. Then the subproblems (3.4.20)–(3.4.21) and (3.4.22)–(3.4.23) of the Robin method are uniquely solvable.*

Proof. We only consider the problem (3.4.20)–(3.4.21) since it is symmetric to (3.4.22)–(3.4.23) and skip the indices k and $k+1$ for convenience.

Using the trace operator $tr_\Gamma : V_1 \to \Lambda$ we define the functional ℓ_1 on V_1 by

$$\begin{aligned}\ell_1 : v_1 \mapsto (f, v_1)_{\Omega_1} - (M_2(u_2), R_2\, tr_\Gamma v_1)_{\Omega_2} - a_2(u_2, R_2\, tr_\Gamma v_1) \\ + (f, R_2\, tr_\Gamma v_1)_{\Omega_2} + \gamma_1(\kappa_2^{-1} u_2, tr_\Gamma v_1)_\Gamma \quad \forall v_1 \in V_1 .\end{aligned}$$

ℓ_1 is linear and bounded since tr_Γ and R_2 are. Then, one can easily see that (3.4.20)–(3.4.21) is equivalent to the variational equality

$$(M_1(u_1), v_1)_{\Omega_1} + \gamma_1(\kappa_1^{-1} u_1, tr_\Gamma v_1)_\Gamma + a_1(u_1, v_1) = \ell_1(v_1) \quad \forall v_1 \in V_1 . \quad (3.4.24)$$

Now, with a primitive Ψ_1 of κ_1^{-1} we define the functional ψ_1 on V_1 by

$$\psi_1 : v_1 \mapsto \int_\Gamma \Psi_1(v_1(s))\, d\sigma(s) \quad \forall v_1 \in V_1 . \quad (3.4.25)$$

Since the real function κ_1^{-1} is monotonically increasing and Lipschitz continuous (Proposition 3.3.1), Ψ_1 is convex and differentiable (Lemma 2.3.6). The latter provides the convexity of ψ_1 (Proposition 2.3.7) and the existence of the directional derivative $\partial_v \psi_1(w)$ for any $v, w \in V_1$ with

$$\partial_v \psi_1(w) = \int_\Gamma \kappa_1^{-1}(w) \cdot v\, d\sigma = (\kappa_1^{-1} w, tr_\Gamma v)_\Gamma \quad (3.4.26)$$

analogously as in Proposition 2.3.9. Therefore, if we choose a convex primitive Φ_1 of M_1 as in Subsection 2.3.2 (possibly with M_1 as in (2.3.25)) leading to a convex functional ϕ_1 on $\mathcal{K} = V_1$ as in Subsection 2.3.3, Proposition 2.3.11 provides the equivalence of (3.4.24) and the convex minimization problem

$$u_1 \in V_1 : \quad F_1(u_1) \leq F_1(v_1) \quad \forall v_1 \in V_1 \quad (3.4.27)$$

for the convex functional

$$F_1 : v_1 \mapsto \phi_1(v_1) + \gamma_1 \psi_1(v_1) + \frac{1}{2} a_1(u_1, v_1) - \ell_1(v_1) \quad \forall v_1 \in V_1 . \quad (3.4.28)$$

Now, an obvious extension of Theorem 2.3.16 can be applied to show that (3.4.27) is uniquely solvable. This generalization can be obtained although M_1

and κ_1^{-1} are not necessarily bounded since, as indicated in (2.3.25) for M_1, Hölder continuity of M_1 and κ_1^{-1} outside of an interval is enough to ensure the coercivity of the functional F_1. For κ_1^{-1} this follows with the same reasoning as for M_1 because here, an affine estimate on the right hand side of (2.3.11) can be established after an application of the trace inequality (A.2.9). □

We point out that the unique solvability of a Robin boundary value problem for the implicit-explicitly time-discretized Richards equation (in the nondegenerate case $kr \geq \alpha > 0$ as in Theorem 3.4.2) also holds if we include additional Dirichlet, Neumann and Signorini-type boundary conditions, see Section 2.3. Furthermore, as discussed in Remark 2.3.17 and in Section 2.4, the continuity and boundedness of M_i, $i = 1, 2$, can be relaxed while still preserving the unique solvability of the convex minimization problem (3.4.27). After a suitable space discretization the numerical treatment of these Robin boundary value problems for the Richards equation can be carried out by monotone multigrid, see Subsection 3.4.5 and note Remark 3.4.28.

Unfortunately, we cannot prove Theorem 3.4.2 for general linear operators $\kappa_i : \Lambda \to \Lambda$, $i = 1, 2$, which have continuous inverses. In this case one would be tempted to deal with an additional bilinear form

$$(w, v) \mapsto (\kappa_1^{-1} tr_\Gamma w, tr_\Gamma v)_\Gamma \quad \forall v, w \in V_1 \tag{3.4.29}$$

on the right hand side of (3.4.24) instead of the contribution $\partial_v \psi_1(w)$ of the convex functional ψ_1. However, we would need to require symmetry and nonnegativity of this bilinear form as an additional assumption in order to proceed successfully as in the proof above. Note that if a $\kappa_i : \Lambda \to \Lambda$ is a superposition operator coming from a Kirchhoff transformation (3.3.21), it is linear if and only if $k_i : \mathbb{R} \to \mathbb{R}$ is a constant function and in this case the corresponding (symmetric) bilinear form (3.4.29) is nonnegative if and only if $k_i \geq 0$.

Remark 3.4.3. Finally, we address the case of the Richards equation in our time-discretized form for which (3.4.4) is replaced by the continuity (3.4.5) of the discretized water flux across Γ. Due to the explicit treatment of the gravitational term, this generalization of problem (3.4.2)–(3.4.4) easily fits into our framework in the following way. Consider (3.4.4) to be replaced by

$$k_1(p_1)\nabla p_1 \cdot \mathbf{n} + g_1 = k_2(p_2)\nabla p_2 \cdot \mathbf{n} + g_2 \quad \text{on } \Gamma$$

with $g_1, g_2 \in L^2(\Gamma)$. Then the linear and bounded functionals $\mu \mapsto -(f, R_i\mu)_{\Omega_i}$, $i = 1, 2$, on Λ in the weak formulations (3.4.11) or (3.4.14) and also in (3.4.21) and (3.4.23) need to be replaced by the functionals

$$\mu \mapsto -(f, R_i\mu)_{\Omega_i} + (g_i, \mu)_\Gamma \quad \forall \mu \in \Lambda$$

which are also linear and bounded. Of course this does not change the validity of Theorem 3.4.2. Moreover, the rest of our theory in this section, to which we turn now, can be extended to this more general case in the same way.

3.4.2 Steklov–Poincaré formulation, equivalence of Robin and ADI method, and altered versions

As done in Subsections 3.3.2 and 3.3.3 for the Dirichlet–Neumann method (3.3.33)–(3.3.37), it is possible to derive a Steklov–Poincaré formulation of the Robin method (3.4.20)–(3.4.23). This shall be carried out in the following, generalizing existing results for the linear case in Discacciati [33, Chapter 5], which also leads to the introduction of a nonlinear ADI method and an altered version of the Robin method. Since in contrast to (3.3.33)–(3.3.37), where we only had one nonlinearity on the interface, two nonlinearities M_i and κ_i are involved in each subproblem for $i = 1, 2$ now. Therefore, in the following we recapitulate the ideas given in Section 3.3.1 in order to extend them to our nonlinear case.

In general, a classical Steklov–Poincaré interface equation should provide an equivalent formulation of the given problem (3.4.6)–(3.4.8) in terms of a suitably chosen interface value λ corresponding to the solution on Γ while enforcing the continuity of the normal derivatives (3.4.8) on the interface. More concretely, applying the Steklov–Poincaré operators S_i, $i = 1, 2$, to λ means solving a Dirichlet subproblem in V_i with boundary values arising from the interface value λ and then providing the normal derivative $S_i\lambda$ on Γ of the obtained solution. Taking into account the sign of the normals, the continuity of the normal derivative then just means

$$S_1\lambda + S_2\lambda = 0 \tag{3.4.30}$$

which is already the Steklov–Poincaré formulation of the problem (3.4.6)–(3.4.8) that we desired. Now, the difficulty in the analysis is often due to the fact that the solution of a subproblem does not only depend on λ but also on the source term f. This is why in the linear setting one usually splits linearly the dependency of the solutions u_1 and u_2 on the two inhomogeneities, i.e. f and the boundary value λ on Γ, thus obtaining an equation

$$S\lambda = S_1\lambda + S_2\lambda = \chi \tag{3.4.31}$$

in which the application of S is no longer dependent on f since the dependency on f is completely contained in χ on the right hand side, as done in (3.3.5)–(3.3.9). Due to the nonlinearities M_1 and M_2 we cannot carry out such a splitting here. Therefore, we only split off parts of the solutions u_1 and u_2 which are representing the source term f with regard to the main part, i.e the partial differential operator in the equation, the Laplacian, which is linear. More concretely, analogously to (3.3.5) and for $i = 1, 2$ we consider the splitting

$$u_i = u_i(\lambda) + u_i^* \tag{3.4.32}$$

where u_i^* is the solution of the linear problem

$$\begin{aligned} -\Delta u_i^* &= f \quad \text{on } \Omega_i & (3.4.33)\\ u_i^* &= 0 \quad \text{on } \partial\Omega_i & (3.4.34)\end{aligned}$$

with homogeneous Dirichlet boundary conditions and $u_i(\lambda)$ is the solution of the nonlinear problem

$$\begin{aligned} M_i(u_i(\lambda)+u_i^*) - \Delta u_i(\lambda) &= 0 && \text{on } \Omega_i && (3.4.35)\\ u_i(\lambda) &= \kappa_i(\lambda) && \text{on } \Gamma && (3.4.36)\\ u_i(\lambda) &= 0 && \text{on } \partial\Omega_i\backslash\Gamma && (3.4.37) \end{aligned}$$

which is homogeneous in the source term. With regard to the time-discretized Richards equation we have chosen the interface value λ as the retransformed variable that has to be equal on both subdomains and that represents the physical pressure in the case of the Richards equation. Observe that the functions $u_i(\lambda)$ depend nonlinearly on λ, but they are also still dependent on f whose influence is hidden in the solutions u_i^* which occur in the problems (3.4.35)–(3.4.37) for $i = 1, 2$.

Remark 3.4.4. If Ω_i are bounded Lipschitz domains and M_i monotonically increasing, continuous and either bounded or else Hölder continuous outside of a bounded interval (see (2.3.25)), we have unique solutions $u_i \in V_i$, $i = 1, 2$, of the nonlinear problems

$$\begin{aligned} M_i(u_i) - \Delta u_i &= f && \text{on } \Omega_i && (3.4.38)\\ u_i &= \kappa_i(\lambda) && \text{on } \Gamma && (3.4.39)\\ u_i &= 0 && \text{on } \partial\Omega_i\backslash\Gamma && (3.4.40) \end{aligned}$$

whose weak forms are equivalent to convex minimization problems (see Section 2.3 and in particular Theorem 2.3.16 and Remark 2.3.17). Due to the unique solvability of (3.4.33)–(3.4.34) we also have unique solutions $u_i(\lambda)$ of the problems (3.4.35)–(3.4.37). The latter problems can be regarded as partially homogeneous, i.e. the solutions $u_i(\lambda)$ "depend mainly" on λ if M_i and κ_i, $i = 1, 2$, behave well enough, which will become clear in and is the basis of the analysis carried out in Subsections 3.4.3 and 3.4.4 (see Proposition 3.4.14, Corollary 3.4.18 and Lemma 3.4.21).

This fact turns out to be the main reason for the splitting of u_1 and u_2 whereas the occurrence of χ as the influence of the inhomogeneity f via u_1^* and u_2^* on the right hand side of (3.4.31) seems to be rather pointless here. After all, there is an influence of f on the left hand side and on $u_i(\lambda)$, $i = 1, 2$, already (compare with the advantage of χ in (3.3.32) for the proof of the well-posedness in Proposition 3.3.8, but see also the proof of Theorem 3.4.23).

Therefore, we define the Steklov–Poincaré operators in the general way as in (3.4.30) by giving the equation

$$S\lambda = S_1\lambda + S_2\lambda = \frac{\partial}{\partial \mathbf{n}}u_1 - \frac{\partial}{\partial \mathbf{n}}u_2 = \frac{\partial}{\partial \mathbf{n}}(u_1(\lambda)+u_1^*) - \frac{\partial}{\partial \mathbf{n}}(u_2(\lambda)+u_2^*) = 0 \qquad (3.4.41)$$

the following weak formulation.

Definition 3.4.5. Let $i = 1, 2$. Assume $R_i : \Lambda \to V_i$, is any linear and continuous extension operator (e.g. a harmonic extension) and M_i is as in Remark 3.4.4. Then for $\lambda \in \Lambda$ we define the functional $S_i\lambda \in \Lambda'$ by

$$\langle S_i\lambda, \mu\rangle := (M_i(u_i(\lambda) + u_i^*), R_i\mu)_{\Omega_i} + a_i(u_i(\lambda) + u_i^*, R_i\mu) - (f, R_i\mu)_{\Omega_i} \quad \forall\mu \in \Lambda \tag{3.4.42}$$

and we set $S := S_1 + S_2$.

Observe that well-definedness of S_i, $i = 1, 2$, is only guaranteed if there is a unique solution $u(\lambda)$ of (3.4.35)–(3.4.37). Therefore, we always assume that M_i, $i = 1, 2$, is chosen at least as in Remark 3.4.4 from now on.

Proposition 3.4.6. *Solving problem (3.4.12)–(3.4.14) is equivalent to finding a $\lambda \in \Lambda$ satisfying*

$$S\lambda = 0\,. \tag{3.4.43}$$

The solutions $u_i \in V_i$, $i = 1, 2$, of (3.4.12)–(3.4.14) correspond to $\lambda \in \Lambda$ by $u_{i|\Gamma} = \kappa_i(\lambda)$ or equivalently $u_i = u_i(\lambda) + u_i^*$.

Proof. By definition, $u_i = u_i(\lambda) + u_i^*$ satisfies (3.4.12)–(3.4.13) which is the weak form of (3.4.38)–(3.4.40) for $i = 1, 2$. Therefore, λ satisfies (3.4.43) by definition (3.4.42) if and only if u_i satisfies (3.4.14). □

Following Remark 3.4.3, it is clear that if (3.4.4) is replaced by (3.4.5) for the full Richards equation with gravity, then each normal derivative in (3.4.41) has to be replaced by the sum of the normal derivative and the corresponding g_i for $i = 1, 2$. Of course, (3.4.42) has to be adapted, too, and Proposition 3.4.6 also holds in this general case.

Now we are ready to derive a formulation of the Robin method (3.4.20)–(3.4.23) in terms of the Steklov–Poincaré operators. This leads to the so-called *Alternating Direction Iterative (ADI) method* which was first related to the Robin method in Discacciati [32, pp. 4–6] for the linear case, and this also works in our nonlinear setting. Using the notation

$$\langle I\eta, \mu\rangle = (\eta, \mu)_\Gamma \quad \forall\eta, \mu \in \Lambda \tag{3.4.44}$$

we obtain

Proposition 3.4.7. *Let M_i and κ_i, $i = 1, 2$, be as in Theorem 3.4.2. Then, with a given $\lambda_2^0 = \kappa_2^{-1}(u_{2|\Gamma}^0) \in \Lambda$, the Robin method (3.4.20)–(3.4.23) is equivalent to solving successively for $k \geq 0$*

$$\begin{aligned} \langle(\gamma_1 I + S_1)\lambda_1^{k+1}, \mu\rangle &= \langle(\gamma_1 I - S_2)\lambda_2^k, \mu\rangle \quad \forall\mu \in \Lambda & (3.4.45)\\ \langle(\gamma_2 I + S_2)\lambda_2^{k+1}, \mu\rangle &= \langle(\gamma_2 I - S_1)\lambda_1^{k+1}, \mu\rangle \quad \forall\mu \in \Lambda & (3.4.46)\end{aligned}$$

in the sense that

$$u_i^k = u_i(\lambda_i^k) + u_i^* \iff \lambda_i^k = \kappa_i^{-1}(u_{i|\Gamma}^k) \quad \forall k \geq 0\,,\ \ i = 1, 2\,, \tag{3.4.47}$$

holds for the iterates u_i^k from (3.4.20)–(3.4.23).

Proof. Assume that u_2^k is the initial iterate for $k = 0$ or has been obtained from (3.4.22)–(3.4.23) for $k \geq 1$. Furthermore, assume that

$$\lambda_2^k = \kappa_2^{-1}(u_{2|\Gamma}^k) \tag{3.4.48}$$

holds. Let λ_1^{k+1} be a solution of (3.4.45) and define $\bar{u}_1^{k+1} := u_1(\lambda_1^{k+1}) + u_1^*$. Then $\bar{u}_1^{k+1}$ satisfies (3.4.20) since it is the solution of (3.4.38)–(3.4.40), which also provides

$$\lambda_1^{k+1} = \kappa_1^{-1}(\bar{u}_{1|\Gamma}^{k+1})\,. \tag{3.4.49}$$

Therefore, $\bar{u}_1^{k+1}$ also solves (3.4.21) due to (3.4.48). Since (3.4.20)–(3.4.21) is uniquely solvable due to Theorem 3.4.2,

$$\bar{u}_1^{k+1} = u_1^{k+1} \tag{3.4.50}$$

is the solution of this Robin step.

As $\bar{u}_1^{k+1}$ is uniquely determined, λ_1^{k+1} must be a unique solution of (3.4.45). On the other hand, there exists a solution $\bar{\lambda}_1^{k+1}$ of (3.4.45) because with the existing solution u_1^{k+1} of (3.4.20)–(3.4.21) the function

$$\bar{\lambda}_1^{k+1} := \kappa_1^{-1}(u_{1|\Gamma}^{k+1}) \tag{3.4.51}$$

satisfies (3.4.45) if $u_1^{k+1} = u_1(\bar{\lambda}_1^{k+1}) + u_1^*$ holds. But, indeed, due to (3.4.20) and (3.4.51), u_1^{k+1} is the solution of (3.4.38)–(3.4.40) with $\lambda = \bar{\lambda}_1^{k+1}$. This shows the converse of the equivalence for the first iteration step (3.4.45).

With regard to the second step (3.4.46), one can use (3.4.50) besides (3.4.49) instead of (3.4.48) in order to derive $\bar{u}_2^{k+1} = u_2^{k+1}$ for $\bar{u}_2^{k+1} = u_2(\lambda_2^{k+1}) + u_2^*$ and the unique solvability of (3.4.46). The reasoning is the same as just seen for the first step. Therefore, since (3.4.48) holds for $k = 0$, we can conclude inductively that the iterates λ_i^k give the iterates u_i^k by (3.4.47) and vice versa. □

The iteration (3.4.45)–(3.4.46) is a nonlinear extension of the ADI method which was introduced in Peaceman and Rachford [74] on an algebraic level and further investigated e.g. in Wachspress and Habetler [95], Wachspress [94] and Varga [92]. A nonlinear ADI method (involving a linearization) for the solution of the Richards equation without gravity in homogeneous soil with Signorini-type boundary conditions is given in Hornung [50]. A convergence theory for a nonlinear version of the ADI method concerning monotone operators which act on a Hilbert space can be found in Lions and Mercier [66]. With regard to an analysis of the ADI method in connection to the Robin method for the linear case see Discacciati [32] or [33, Chapter 5], where one can also find the analysis that shall be generalized in the following. As in Lions and Mercier [66] this will lead us to monotonicity conditions for the Steklov–Poincaré operators $S_i : \Lambda \to \Lambda'$ (see Theorem 3.4.12 and Remark 3.4.26).

Due to the just derived invertibility of the operators involved, the operator $T_{\gamma_1,\gamma_2} : \Lambda \to \Lambda$ providing the iteration $\lambda_2^{k+1} = T_{\gamma_1,\gamma_2}\lambda_2^k$ in (3.4.45)–(3.4.46) is given by

$$T_{\gamma_1,\gamma_2} = (\gamma_2 I + S_2)^{-1}(\gamma_2 I - S_1)(\gamma_1 I + S_1)^{-1}(\gamma_1 I - S_2)\,. \tag{3.4.52}$$

Proposition 3.4.8. *Let M_i and κ_i, $i = 1, 2$, be as in Theorem 3.4.2. Any fixed point $u_2 \in V_2$ (corresponding to $u_1 \in V_1$) of the Robin method (3.4.20)–(3.4.23) is a solution of the domain decomposition problem (3.4.12)–(3.4.14) and vice versa, and it provides a fixed point $\lambda = u_{2|\Gamma} \in \Lambda$ of $T_{\gamma_1,\gamma_2} : \Lambda \to \Lambda$ in (3.4.52). Conversely, any fixed point $\lambda \in \Lambda$ of the operator T_{γ_1,γ_2} provides a fixed point $u_2 = u_2(\lambda) + u_2^*$ of the Robin method (3.4.20)–(3.4.23). Finally, any fixed point $\lambda \in \Lambda$ of $T_{\gamma_1,\gamma_2} : \Lambda \to \Lambda$ is a solution of the Steklov–Poincaré interface equation $S\lambda = 0$ and vice versa.*

Proof. Let $u_2 \in V_2$ (corresponding to $u_1 \in V_1$) be a fixed point of the Robin iteration (3.4.20)–(3.4.23). Then subtracting (3.4.23) from (3.4.21) with $u_1 = u_1^{k+1}$ and $u_2 = u_2^k = u_2^{k+1}$ we obtain

$$(\kappa_1^{-1} u_1, \mu)_\Gamma = (\kappa_2^{-1} u_2, \mu)_\Gamma \quad \forall \mu \in \Lambda\,, \tag{3.4.53}$$

i.e. $\kappa_1^{-1}(u_{1|\Gamma}) = \kappa_2^{-1}(u_{2|\Gamma})$ in $L^2(\Gamma)$ (in particular almost everywhere on Γ) and therefore in Λ. Now, adding $\gamma_2 \cdot$ (3.4.21) to $\gamma_1 \cdot$ (3.4.23) with $u_1 = u_1^{k+1}$ and $u_2 = u_2^k = u_2^{k+1}$ and considering (3.4.53) gives (3.4.14), i.e. $u_1 \in V_1$, $u_2 \in V_2$ is a solution of the domain decomposition problem (3.4.12)–(3.4.14).

Conversely, if $u_1 \in V_1$, $u_2 \in V_2$ satisfy (3.4.12)–(3.4.14) one obtains (3.4.53) from (3.4.13) and therefore (3.4.21) and (3.4.23) from (3.4.14) with $u_1 = u_1^{k+1}$ and $u_2 = u_2^k = u_2^{k+1}$.

Furthermore, the fixed points of the Robin method (3.4.20)–(3.4.23) and the ADI method (3.4.45)–(3.4.46) coincide in the asserted way due to Proposition 3.4.7.

Finally, since the fixed points of $T_{\gamma_1,\gamma_2} : \Lambda \to \Lambda$ and the solutions of the domain decomposition problem (3.4.12)–(3.4.14) correspond to each other in the way described, Proposition 3.4.43 entails the last assertion. □

For a rigorous analysis we replace the identity operator I on Λ inducing bounded linear functionals on Λ via (3.4.44) by the bounded linear operator $\mathcal{I} : \Lambda \to \Lambda'$ from the Riesz representation theorem (see e.g. [98, p. 222]). The latter can be defined as

$$\mathcal{I}\mu \in \Lambda' : \quad (\mathcal{I}\mu, \xi)_{\Lambda'} = \langle \xi, \mu \rangle \quad \forall \xi \in \Lambda' \tag{3.4.54}$$

where $(\cdot,\cdot)_{\Lambda'}$ denotes the scalar product in Λ'. Consequently, we have

$$\langle \mathcal{I}\eta, \mu \rangle = (\mathcal{I}\mu, \mathcal{I}\eta)_{\Lambda'} = (\eta, \mu)_\Lambda\,. \tag{3.4.55}$$

Therefore, if we replace I by $\mathcal{I}$ in (3.4.52), we replace the L^2-scalar products

$$(\kappa_i^{-1} u_i^{k+1}, \mu)_\Gamma\,, \quad i = 1, 2\,, \quad \text{and} \quad (\kappa_2^{-1} u_2^k, \mu)_\Gamma$$

in (3.4.20)–(3.4.23), $k \geq 0$, by the corresponding scalar products in Λ

$$(\kappa_i^{-1} u_i^{k+1}, \mu)_\Lambda\,, \quad i = 1, 2\,, \quad \text{and} \quad (\kappa_2^{-1} u_2^k, \mu)_\Lambda\,.$$

Consequently, we consider another iterative method instead of the classical weak formulation of the Robin method given in (3.4.20)–(3.4.23). Note that this difference is not yet reflected in the strong Robin boundary condition

$$\frac{\partial}{\partial \mathbf{n}} w + \gamma w \quad \text{on } \Gamma$$

for $\gamma > 0$. The usual weak treatment of the quantities $\frac{\partial}{\partial \mathbf{n}} w$ and w as in (3.4.20)–(3.4.23) is to interpret them as elements of Λ' induced by the L^2-scalar product. This is natural for $\frac{\partial}{\partial \mathbf{n}} w$ for which Green's formula (1.5.9) provides the weak form. In what is to come, the element of Λ' arising from the second quantity w shall be induced by the scalar product in Λ' via Riesz's operator $\mathcal{I}$ as shown above. The corresponding iterative method reads:

For a given $u_2^0 \in V_2$ find successively $u_1^{k+1} \in V_1$ and $u_2^{k+2} \in V_2$ for $k \geq 0$ such that

$$(M_1(u_1^{k+1}), v_1)_{\Omega_1} + a_1(u_1^{k+1}, v_1) = (f, v_1)_{\Omega_1} \quad \forall v_1 \in V_1^0 \tag{3.4.56}$$

$$\begin{aligned}(M_1(u_1^{k+1}), R_1\mu)_{\Omega_1} + a_1(u_1^{k+1}, R_1\mu) - (f, R_1\mu)_{\Omega_1} + \gamma_1(\kappa_1^{-1}u_1^{k+1}, \mu)_\Lambda = \\ -(M_2(u_2^k), R_2\mu)_{\Omega_2} - a_2(u_2^k, R_2\mu) + (f, R_2\mu)_{\Omega_2} + \gamma_1(\kappa_2^{-1}u_2^k, \mu)_\Lambda \quad \forall \mu \in \Lambda\end{aligned} \tag{3.4.57}$$

and then

$$(M_2(u_2^{k+1}), v_2)_{\Omega_2} + a_2(u_2^{k+1}, v_2) = (f, v_2)_{\Omega_2} \quad \forall v_2 \in V_2^0 \tag{3.4.58}$$

$$\begin{aligned}(M_2(u_2^{k+1}), R_2\mu)_{\Omega_2} + a_2(u_2^{k+1}, R_2\mu) - (f, R_2\mu)_{\Omega_1} - \gamma_2(\kappa_2^{-1}u_2^{k+1}, \mu)_\Lambda = \\ -(M_1(u_1^{k+1}), R_1\mu)_{\Omega_1} - a_1(u_1^{k+1}, R_1\mu) + (f, R_1\mu)_{\Omega_1} - \gamma_2(\kappa_1^{-1}u_1^{k+1}, \mu)_\Lambda \quad \forall \mu \in \Lambda\,.\end{aligned} \tag{3.4.59}$$

We shall call this new iterative method the *altered Robin method*.

Remark 3.4.9. We point out, that it seems unclear if (3.4.56)–(3.4.57) and (3.4.58)–(3.4.59) are equivalent to convex minimization problems or uniquely solvable at all. So on the one hand, the application of monotone multigrid for the equations arising from (3.4.56)–(3.4.57) and (3.4.58)–(3.4.59) on the discrete level might be ruled out, and on the other hand we have to *assume the*

$$\textit{unique solvability of (3.4.56)–(3.4.57) and (3.4.58)–(3.4.59)} \tag{3.4.60}$$

here. Obviously, we need to assume the well-definedness of the Steklov–Poincaré operators in the first place. With this assumption, however, we can establish a convergence analysis for the altered Robin method to which we turn in the next subsection and which leads to natural conditions on the Steklov–Poincaré operators S_i or, more specifically, on the nonlinearities M_i and κ_i, $i = 1, 2$, that guarantee convergence.

In order to see that this assumption is not artificial note that it is certainly satisfied in one space dimension. In this case Γ only contains one point such that the altered Robin method and the classical one (3.4.20)–(3.4.23) coincide with $I = \mathcal{I}$ and equality of $(\cdot,\cdot)_\Gamma$ and $(\cdot,\cdot)_\Lambda$ since here $\Lambda = L^2(\Gamma)$ is just the one dimensional Hilbert space $(\mathbb{R}, |\cdot|)$. At least in the one-dimensional case we obtain convergence of the Robin method for the implicit-explicitly time-discretized Richards equation in the nondegenerate setting (see Subsection 1.4.3) which we prove in Subsection 3.4.4.

It is straightforward to see and it will be helpful in our further analysis that Propositions 3.4.7 and 3.4.8 carry over to the altered Robin method.

Proposition 3.4.10. *With the assumption (3.4.60) the altered Robin method (3.4.56)–(3.4.59) is equivalent to the altered ADI method*

$$\begin{aligned} \langle(\gamma_1\mathcal{I} + S_1)\lambda_1^{k+1}, \mu\rangle &= \langle(\gamma_1\mathcal{I} - S_2)\lambda_2^k, \mu\rangle \qquad \forall \mu \in \Lambda \qquad (3.4.61)\\ \langle(\gamma_2\mathcal{I} + S_2)\lambda_2^{k+1}, \mu\rangle &= \langle(\gamma_2\mathcal{I} - S_1)\lambda_1^{k+1}, \mu\rangle \quad \forall \mu \in \Lambda \qquad (3.4.62) \end{aligned}$$

for given $\lambda_2^0 = \kappa_2^{-1}(u^0_{2|\Gamma}) \in \Lambda$ in the same sense as in Proposition 3.4.7.

Proposition 3.4.11. *We assume (3.4.60). Then the assertions of Proposition 3.4.8 also hold if we replace the Robin method (3.4.20)–(3.4.23) by the altered Robin method (3.4.56)–(3.4.59) and the operator $T_{\gamma_1,\gamma_2} : \Lambda \to \Lambda$ in (3.4.52) by $\mathcal{T}_{\gamma_1,\gamma_2} : \Lambda \to \Lambda$ defined as*

$$\mathcal{T}_{\gamma_1,\gamma_2} = (\gamma_2\mathcal{I} + S_2)^{-1}(\gamma_2\mathcal{I} - S_1)(\gamma_1\mathcal{I} + S_1)^{-1}(\gamma_1\mathcal{I} - S_2)\,. \qquad (3.4.63)$$

In the following section we deduce conditions for which the fixed points in this proposition are existing and unique.

3.4.3 Convergence analysis for the altered Robin method via nonlinear Steklov–Poincaré operators

In this subsection we provide conditions first on the Steklov–Poincaré operators $S_i : \Lambda \to \Lambda'$ and then on $M_i : \mathbb{R} \to \mathbb{R}$ and $\kappa_i : \Lambda \to \Lambda$, $i = 1, 2$, that guarantee convergence of the altered Robin method (3.4.56)–(3.4.59). We have introduced this altered version of (3.4.20)–(3.4.23) in the last subsection because this variant seems to be the "right" weak formulation of (3.4.15)–(3.4.18) for a successful convergence analysis using a contraction argument. The aim is to prove the convergence of a transformed sequence $(\tilde{\lambda}_2^k)_{k\geq 0}$ of the sequence $(\lambda_2^k)_{k\geq 0}$ of iterates from (3.4.61)–(3.4.62) in the Hilbert space Λ' using its natural inner product $(\cdot,\cdot)_{\Lambda'}$.

As pointed out in Remark 3.4.9 we need to assume (3.4.60). For the proof of the next theorem, where we deal with contractions on Λ', we even require the stronger condition

$$\gamma_i\mathcal{I} + S_i : \Lambda \to \Lambda' \quad \textit{is invertible for } i = 1, 2 \qquad (3.4.64)$$

if we are not in a one dimensional setting where this condition is satisfied due to Theorem 3.4.2 under reasonable conditions on M_i, κ_i, $i = 1, 2$. For the well-definedness of the Steklov–Poincaré operators in Definition 3.4.5 we need to assume the conditions on M_i, $i = 1, 2$, given in this theorem anyway. With these assumptions we can prove the following generalization of a linear result in Discacciati [33, pp. 99/100]. It is not surprising that two conditions that we already encountered in Theorem 3.3.4 (see (3.3.42) and (3.3.45)) reoccur in this case.

Theorem 3.4.12. *Assume that (3.4.64) holds and $M_i : \mathbb{R} \to \mathbb{R}$, $i = 1, 2$, is monotonically increasing, continuous and either bounded or Hölder continuous outside of a bounded interval. Let $\gamma_1 = \gamma_2 = \gamma > 0$. Then for any initial iterate $\lambda_2^0 \in \Lambda$ the operator $\mathcal{T}_\gamma : \Lambda \to \Lambda$ defined by*

$$\mathcal{T}_\gamma = (\gamma\mathcal{I} + S_2)^{-1}(\gamma\mathcal{I} - S_1)(\gamma\mathcal{I} + S_1)^{-1}(\gamma\mathcal{I} - S_2)$$

provides a sequence $(\lambda_2^k)_{k\geq 0}$ which converges in Λ to the unique fixed point of $\mathcal{T}_\gamma$ if both $S_1, S_2 : \Lambda \to \Lambda'$ are Lipschitz continuous and strongly monotone, i.e. if there are positive constants c_i and C_i such that the following holds for $i = 1, 2$:

$$\langle S_i\eta - S_i\mu, \lambda\rangle \;\leq\; C_i\|\eta - \mu\|_\Lambda\|\lambda\|_\Lambda \quad \forall \eta, \mu, \lambda \in \Lambda \tag{3.4.65}$$

$$\langle S_i\eta - S_i\mu, \eta - \mu\rangle \;\geq\; c_i\|\eta - \mu\|_\Lambda^2 \quad \forall \eta, \mu \in \Lambda\,. \tag{3.4.66}$$

Proof. First the operator $\mathcal{T}_\gamma$ is well-defined by assumption (3.4.60). For $k \geq 0$ we introduce the auxiliary variable $\tilde{\lambda}_2^k = (\gamma\mathcal{I} + S_2)\lambda_2^k$ and rewrite $\lambda_2^{k+1} = \mathcal{T}_\gamma\lambda_2^k$ as $\tilde{\lambda}_2^{k+1} = \tilde{\mathcal{T}}_\gamma\tilde{\lambda}_2^k$ with the operator $\tilde{\mathcal{T}}_\gamma : \Lambda' \to \Lambda'$ defined by

$$\tilde{\mathcal{T}}_\gamma = (\gamma\mathcal{I} - S_1)(\gamma\mathcal{I} + S_1)^{-1}(\gamma\mathcal{I} - S_2)(\gamma\mathcal{I} + S_2)^{-1}\,.$$

Then, since $\mathcal{I} : \Lambda \to \Lambda'$ is continuous, $S_2 : \Lambda \to \Lambda'$ is assumed to be Lipschitz continuous and $\gamma\mathcal{I} + S_2 : \Lambda \to \Lambda'$ is invertible with

$$\mathcal{T}_\gamma = (\gamma\mathcal{I} + S_2)^{-1}\,\tilde{\mathcal{T}}_\gamma\,(\gamma\mathcal{I} + S_2)\,, \tag{3.4.67}$$

it suffices to prove the analogous statement for $\tilde{\mathcal{T}}_\gamma$ and the auxiliary variables instead of $\mathcal{T}_\gamma$ and the original ones. In fact, it turns out that both operators

$$\tilde{\mathcal{T}}_{i,\gamma} : \Lambda' \to \Lambda'\,, \quad \tilde{\mathcal{T}}_{i,\gamma} = (\gamma\mathcal{I} - S_i)(\gamma\mathcal{I} + S_i)^{-1}\,, \quad i = 1, 2\,,$$

are contractions under the assumptions that are imposed on the operators S_i. So for any $\tilde{\eta}, \tilde{\mu} \in \Lambda'$, $\tilde{\eta} \neq \tilde{\mu}$, and $i = 1, 2$ we consider the ratio

$$\begin{aligned}\frac{\|\tilde{\mathcal{T}}_{i,\gamma}\tilde{\eta} - \tilde{\mathcal{T}}_{i,\gamma}\tilde{\mu}\|_{\Lambda'}^2}{\|\tilde{\eta} - \tilde{\mu}\|_{\Lambda'}^2} &= \frac{\|(\gamma\mathcal{I} - S_i)(\gamma\mathcal{I} + S_i)^{-1}\tilde{\eta} - (\gamma\mathcal{I} - S_i)(\gamma\mathcal{I} + S_i)^{-1}\tilde{\mu}\|_{\Lambda'}^2}{\|\tilde{\eta} - \tilde{\mu}\|_{\Lambda'}^2} \\ &= \frac{\|(\gamma\mathcal{I} - S_i)\eta - (\gamma\mathcal{I} - S_i)\mu\|_{\Lambda'}^2}{\|(\gamma\mathcal{I} + S_i)\eta - (\gamma\mathcal{I} + S_i)\mu\|_{\Lambda'}^2}\end{aligned}$$

where we have introduced the auxiliary variables $\eta = (\gamma\mathcal{I} + S_i)^{-1}\tilde{\eta} \in \Lambda$ and $\mu = (\gamma\mathcal{I} + S_i)^{-1}\tilde{\mu} \in \Lambda$, $\eta \neq \mu$. This ratio can be reformulated as

$$\frac{\gamma^2\|\mathcal{I}\eta - \mathcal{I}\mu\|_{\Lambda'}^2 - 2\gamma(S_i\eta - S_i\mu, \mathcal{I}\eta - \mathcal{I}\mu)_{\Lambda'} + \|S_i\eta - S_i\mu\|_{\Lambda'}^2}{\gamma^2\|\mathcal{I}\eta - \mathcal{I}\mu\|_{\Lambda'}^2 + 2\gamma(S_i\eta - S_i\mu, \mathcal{I}\eta - \mathcal{I}\mu)_{\Lambda'} + \|S_i\eta - S_i\mu\|_{\Lambda'}^2}$$

and the Riesz representation theorem provides $\|\mathcal{I}\eta - \mathcal{I}\mu\|_{\Lambda'} = \|\eta - \mu\|_\Lambda$ while (3.4.54) gives $(S_i\eta - S_i\mu, \mathcal{I}\eta - \mathcal{I}\mu)_{\Lambda'} = \langle S_i\eta - S_i\mu, \eta - \mu\rangle$ so that all that is left to do is to find a positive constant $c < 1$ such that

$$\frac{\gamma^2\|\eta-\mu\|_\Lambda^2 - 2\gamma\langle S_i\eta - S_i\mu, \eta-\mu\rangle + \|S_i\eta - S_i\mu\|_{\Lambda'}^2}{\gamma^2\|\eta-\mu\|_\Lambda^2 + 2\gamma\langle S_i\eta - S_i\mu, \eta-\mu\rangle + \|S_i\eta - S_i\mu\|_{\Lambda'}^2} \le c\,. \tag{3.4.68}$$

Now, exploiting first the strong monotonicity and then the Lipschitz continuity of S_i we can estimate the left hand side in (3.4.68) by

$$\begin{aligned}
\frac{\gamma^2\|\eta-\mu\|_\Lambda^2 - 2\gamma c_i\|\eta-\mu\|_\Lambda^2 + \|S_i\eta - S_i\mu\|_{\Lambda'}^2}{\gamma^2\|\eta-\mu\|_\Lambda^2 + 2\gamma c_i\|\eta-\mu\|_\Lambda^2 + \|S_i\eta - S_i\mu\|_{\Lambda'}^2} &\le \frac{\gamma^2 - 2\gamma c_i + \frac{\|S_i\eta - S_i\mu\|_{\Lambda'}^2}{\|\eta-\mu\|_\Lambda^2}}{\gamma^2 + 2\gamma c_i + \frac{\|S_i\eta - S_i\mu\|_{\Lambda'}^2}{\|\eta-\mu\|_\Lambda^2}} \\
&\le 1 - \frac{4\gamma c_i}{\gamma^2 + 2\gamma c_i + \frac{\|S_i\eta - S_i\mu\|_{\Lambda'}^2}{\|\eta-\mu\|_\Lambda^2}} \\
&\le 1 - \frac{4\gamma c_i}{\gamma^2 + 2\gamma c_i + C_i^2} < 1\,.
\end{aligned}$$

□

Observe that the stronger condition (3.4.64) instead of (3.4.60) was only necessary for $i = 2$ in the proof. Note, in addition, that in contrast to Theorem 3.3.4, which plays the same role in Section 3.3 as Theorem 3.4.12 here, the above result does not guarantee the existence of an upper bound $\rho_{\max} \in (0,1)$ for convergence rates of the convergence $\lambda_2^k \to \lambda$ since the operator $\mathcal{T}_\gamma$ providing the sequence $(\lambda_2^k)_{k\ge0}$ is only a composition (3.4.67) of continuous operators with a contraction and may not be a contraction itself. On the other hand, Theorem 3.4.12 guarantees the convergence $\lambda_2^k \to \lambda$ for *any* parameter $\gamma > 0$ whereas it is well known (and observed in Subsection 3.3.5) that for damping parameters above some threshold in $(0,1)$ the Dirichlet–Neumann method need not converge.

In what is to come we give conditions on $M_i : \mathbb{R} \to \mathbb{R}$ and $\kappa_i : \Lambda \to \Lambda$, $i = 1,2$, under which the assumptions (3.4.65) and (3.4.66) are satisfied. The following nice result will turn out to be a fundamental tool for the further considerations in this direction.

Lemma 3.4.13. *Let $M_i : \mathbb{R} \to \mathbb{R}$ be as in Theorem 3.4.12 and $H_i : \Lambda \to V_i$ the harmonic extension operators for $i = 1,2$. Then for any $\eta, \mu \in \Lambda$ we have the identity*

$$\begin{aligned}
(M_i(u_i(\eta) + u_i^*) - M_i(u_i(\mu) + u_i^*), u_i(\eta) - u_i(\mu))_{\Omega_i}& \\
+a_i(u_i(\eta) - u_i(\mu), u_i(\eta) - u_i(\mu)) &= \\
(M_i(u_i(\eta) + u_i^*) - M_i(u_i(\mu) + u_i^*), H_i\kappa_i\eta - H_i\kappa_i\mu)_{\Omega_i}& \\
+a_i(u_i(\eta) - u_i(\mu), H_i\kappa_i\eta - H_i\kappa_i\mu)&\,.
\end{aligned} \tag{3.4.69}$$

Proof. Let $i = 1, 2$. The conditions on M_i provide the unique solvability of (3.4.35)–(3.4.37). The weak form of (3.4.35) reads

$$(M_i(u_i(\lambda) + u_i^*), v_i)_{\Omega_i} + a_i(u_i(\lambda), v_i) = 0 \quad \forall v_i \in V_i^0 \tag{3.4.70}$$

and due to (3.4.36) we have $v_i(\eta, \mu) := u_i(\eta) - u_i(\mu) - (H_i\kappa_i\eta - H_i\kappa_i\mu) \in V_i^0$ for any $\eta, \mu \in \Lambda$. So setting $\lambda = \eta$ in (3.4.70) and $v_i = v_i(\eta, \mu)$ we obtain

$$\begin{aligned}(M_i(u_i(\eta) + u_i^*), u_i(\eta) - u_i(\mu))_{\Omega_i} + a_i(u_i(\eta), u_i(\eta) - u_i(\mu)) = \\ (M_i(u_i(\eta) + u_i^*), H_i\kappa_i\eta - H_i\kappa_i\mu)_{\Omega_i} + a_i(u_i(\eta), H_i\kappa_i\eta - H_i\kappa_i\mu)\end{aligned} \tag{3.4.71}$$

while setting $\lambda = \mu$ in (3.4.70) and testing again with $v_i = v_i(\eta, \mu)$ we have

$$\begin{aligned}(M_i(u_i(\mu) + u_i^*), u_i(\eta) - u_i(\mu))_{\Omega_i} + a_i(u_i(\mu), u_i(\eta) - u_i(\mu)) = \\ (M_i(u_i(\mu) + u_i^*), H_i\kappa_i\eta - H_i\kappa_i\mu)_{\Omega_i} + a_i(u_i(\mu), H_i\kappa_i\eta - H_i\kappa_i\mu)\,.\end{aligned} \tag{3.4.72}$$

Now, subtracting (3.4.72) from (3.4.71) gives the identity (3.4.69). □

At first glance (3.4.69) looks like a special "nonlinear orthogonality" of the difference $u_i(\eta) - u_i(\mu)$ of solutions to (3.4.35)–(3.4.37) and the corresponding difference $H_i\kappa_i\eta - H_i\kappa_i\mu$ of solutions to the same problem with $M_i = 0$ for $i = 1, 2$. Note, however, that in the proof no special properties of H_i are applied and that, in fact, (3.4.69) holds for arbitrary extension operators $R_i : \Lambda \to V_i$, $i = 1, 2$. Nevertheless, we will exploit the linearity and the well-known continuity

$$\|H_i\eta\|_{1,\Omega_i} \le C\|\eta\|_\Lambda \quad \forall \eta \in \Lambda$$

with a $C > 0$ of the harmonic extension operators (see Quarteroni and Valli [75, p. 8/9]) in the following analysis. Therefore, we choose $R_i = H_i$ in the representation of the Steklov–Poincaré operators (3.4.42) (which could of course be any extension operator with the same properties). Before we turn to the main propositions of this subsection we need an additional preliminary result which is of interest on its own.

Proposition 3.4.14. *Let $i = 1, 2$. If $M_i : \mathbb{R} \to \mathbb{R}$ is monotonically increasing and Lipschitz continuous and $\kappa_i : \Lambda \to \Lambda$ is Lipschitz continuous, then for any given $u_i^* \in V_i^0$ the problem (3.4.35)–(3.4.37) is well-posed with respect to λ. More specifically, there is a well-defined solution operator*

$$L_i : \Lambda \to V_i\,, \qquad L_i\lambda = u_i(\lambda)\,,$$

which is Lipschitz continuous, i.e. there is a constant $C > 0$ such that

$$\|L_i\eta - L_i\mu\|_{1,\Omega_i} \le C\|\eta - \mu\|_\Lambda \quad \forall \eta, \mu \in \Lambda\,.$$

Proof. The unique solvability of problem (3.4.35)–(3.4.37) depends on the properties of M_i and has already been addressed in Remark 3.4.4. For the continuous dependence of the solution on the data consider the identity (3.4.69) which we would like to abbreviate by

$$A = B\,.$$

In the following C_k, $k = 1, 2, 3$, are suitably chosen positive constants. Using the monotonicity of M_i (noting that $u_i(\eta) - u_i(\mu) = u_i(\eta) + u_i^* - u_i(\mu) - u_i^*$) and the Poincaré inequality (A.2.13), the left hand side A in (3.4.69) can be estimated from below by

$$A \geq C_1 \|u_i(\eta) - u_i(\mu)\|^2_{1,\Omega_i}\,. \tag{3.4.73}$$

With the Cauchy–Schwarz inequality and the Lipschitz continuity of M_i (with the Lipschitz constants $L(M_i)$), κ_i and H_i on the right hand side B of (3.4.69) can be estimated from above by

$$\begin{aligned} B &\leq \|M_i(u_i(\eta) + u_i^*) - M_i(u_i(\mu) + u_i^*)\|_{0,\Omega_i} \|H_i(\kappa_i\eta - \kappa_i\mu)\|_{0,\Omega_i} \\ &\quad + \|u_i(\eta) - u_i(\mu)\|_{1,\Omega_i} \|H_i(\kappa_i\eta - \kappa_i\mu)\|_{1,\Omega_i} \\ &\leq (L(M_i) + 1)\|u_i(\eta) - u_i(\mu)\|_{1,\Omega_i} \|H_i(\kappa_i\eta - \kappa_i\mu)\|_{1,\Omega_i} \\ &\leq C_2\|u_i(\eta) - u_i(\mu)\|_{1,\Omega_i} \|\eta - \mu\|_\Lambda\,. \end{aligned}$$

Taking into account the estimate (3.4.73) we obtain

$$\|u_i(\eta) - u_i(\mu)\|_{1,\Omega_i} \leq C_3 \|\eta - \mu\|_\Lambda$$

as claimed. □

Proposition 3.4.15. *The Steklov–Poincaré operator $S_i : \Lambda \to \Lambda'$, $i = 1, 2$, is Lipschitz continuous if $M_i : \mathbb{R} \to \mathbb{R}$ is monotonically increasing and Lipschitz continuous and $\kappa_i : \Lambda \to \Lambda$ is Lipschitz continuous.*

Proof. We choose $\eta, \lambda, \mu \in \Lambda$. Using (3.4.42) with $R_i = H_i$ and we obtain

$$\begin{aligned} \langle S_i\eta - S_i\mu, \lambda\rangle &\leq |(M_i(u_i(\eta) + u_i^*) - M_i(u_i(\mu) + u_i^*), H_i\lambda)_{\Omega_i}| \\ &\quad + |a_i(u_i(\eta) - u_i(\mu), H_i\lambda)|\,. \end{aligned} \tag{3.4.74}$$

Applying the Cauchy–Schwarz inequality and exploiting the Lipschitz continuity of M_i as in the proof of Proposition 3.4.14, the right hand side of (3.4.74) can be estimated by

$$\begin{aligned} &\|M_i(u_i(\eta) + u_i^*) - M_i(u_i(\mu) + u_i^*)\|_{0,\Omega_i} \|H_i\lambda\|_{0,\Omega_i} + \|u_i(\eta) - u_i(\mu)\|_{1,\Omega_i} \|H_i\lambda\|_{1,\Omega_i} \\ &\leq (L(M_i) + 1)\, \|u_i(\eta) - u_i(\mu)\|_{1,\Omega_i}\, \|H_i\lambda\|_{1,\Omega_i}\,. \end{aligned} \tag{3.4.75}$$

The assertion now follows by applying Proposition 3.4.14 as well as the continuity of H_i. □

Just as encountered for the transmission problem (3.3.11)–(3.3.13) discussed in Section 3.3 (compare Proposition 3.3.7), the monotonicity of the Steklov–Poincaré operators turns out to be considerably more delicate than their Lipschitz continuity in our more general case of problem (3.4.2)–(3.4.4), too. This is reflected by the next proposition which does not yet provide easily verifiable conditions that guarantee (3.4.66). But even though the estimate in this proposition seems to be a bit technical, it unveils the power of our identity (3.4.69) and makes space for more detailed monotonicity considerations, which we carry out in the next subsection.

Proposition 3.4.16. *The Steklov–Poincaré operator $S_i : \Lambda \to \Lambda'$, $i = 1, 2$, is strongly monotone if $M_i : \mathbb{R} \to \mathbb{R}$ is as in Theorem 3.4.12, $\kappa_i : \Lambda \to \Lambda$ is invertible with Lipschitz continuous κ_i^{-1} and there is a constant $C_i > 0$ such that*

$$\begin{aligned}
&(M_i(u_i(\eta) + u_i^*) - M_i(u_i(\mu) + u_i^*), H_i\eta - H_i\mu)_{\Omega_i} \\
&+ a_i(u_i(\eta) - u_i(\mu), H_i\eta - H_i\mu) \\
&\geq C_i\Big((M_i(u_i(\eta) + u_i^*) - M_i(u_i(\mu) + u_i^*), H_i\kappa_i\eta - H_i\kappa_i\mu)_{\Omega_i} \\
&\qquad + a_i(u_i(\eta) - u_i(\mu), H_i\kappa_i\eta - H_i\kappa_i\mu)\Big)\,.
\end{aligned} \tag{3.4.76}$$

Proof. Let $i = 1, 2$. Condition (2.3.25) on M_i guarantees that (3.4.35)–(3.4.37) is uniquely solvable, and therefore S_i is well-defined. In the following C_k, $k = 3, 4, 5$, are suitable positive constants. Let $\eta, \mu \in \Lambda$. Then, with $R_i = H_i$ in (3.4.42) we have

$$\begin{aligned}
\langle S_i\eta - S_i\mu, \eta - \mu\rangle \;&=\; (M_i(u_i(\eta) + u_i^*) - M_i(u_i(\mu) + u_i^*), H_i\eta - H_i\mu)_{\Omega_i} \\
&\quad + a_i(u_i(\eta) - u_i(\mu), H_i\eta - H_i\mu)\,.
\end{aligned}$$

Now, due to (3.4.76) and (3.4.69) we obtain

$$\begin{aligned}
\langle S_i\eta - S_i\mu, \eta - \mu\rangle \;&\geq\; C_i\Big((M_i(u_i(\eta) + u_i^*) - M_i(u_i(\mu) + u_i^*), u_i(\eta) - u_i(\mu))_{\Omega_i} \\
&\qquad + a_i(u_i(\eta) - u_i(\mu), u_i(\eta) - u_i(\mu))\Big)
\end{aligned}$$

which can be further estimated from below by

$$C_3\|u_i(\eta) - u_i(\mu)\|_{1,\Omega_i}^2 \geq C_4\|(u_i(\eta) - u_i(\mu))_{|\Gamma}\|_\Lambda^2 = C_4\|\kappa_i\eta - \kappa_i\mu\|_\Lambda^2 \geq C_5\|\eta - \mu\|_\Lambda^2 \tag{3.4.77}$$

using successively the monotonicity of M_i, the Poincaré inequality (A.2.13) and the trace inequality (A.2.9) as well as the Lipschitz continuity of κ_i^{-1}. □

Remark 3.4.17. Note that (3.4.76) could be more compactly written as

$$\langle S_i\eta - S_i\mu, \eta - \mu\rangle \geq C_i\langle S_i\eta - S_i\mu, \kappa_i\eta - \kappa_i\mu\rangle$$

in terms of the Steklov–Poincaré operators S_i, $i = 1, 2$, and the above proof guarantees that the right hand side of this inequality can always be bounded from below by $C_5\|\eta - \mu\|_\Lambda^2$. This should not confuse since the term on the right hand side is actually “the symmetric term” in this inequality. It depends on $\kappa_i\eta$ and $\kappa_i\mu$ in both entries of $\langle\cdot,\cdot\rangle$ due to the definition (3.4.42) of S_i, $i = 1, 2$, based on the solutions $u_i = u_i(\lambda) + u_i^*$ of (3.4.38)–(3.4.40) with $u_{i|\Gamma} = \kappa_i(\lambda)$.

With the last estimate (3.4.77) in the proof of Proposition 3.4.16 we obtain the following result which completes Proposition 3.4.14. Recall that the unique solvability of (3.4.35)–(3.4.37), i.e. the existence of L_i, $i = 1, 2$, only depends on the properties of M_i.

Corollary 3.4.18. *Let $i = 1, 2$. If $M_i : \mathbb{R} \to \mathbb{R}$ is monotonically increasing and Lipschitz continuous and $\kappa_i : \Lambda \to \Lambda$ invertible with Lipschitz continuous κ_i^{-1}, then the solution operator L_i in Proposition 3.4.14 has a Lipschitz continuous inverse on $L_i(\Lambda)$, i.e. there is a constant $c > 0$ such that*

$$\|\eta - \mu\|_\Lambda \le c\,\|u_i(\eta) - u_i(\mu)\|_{1,\Omega_i} \quad \forall \eta, \mu \in \Lambda\,.$$

Concluding Remarks 3.4.19. Summarizing the analysis in this section we can establish an abstract convergence result for the altered Robin method (3.4.56)–(3.4.59) which reads:

Let $i = 1, 2$ and $\gamma_1 = \gamma_2 > 0$. Suppose $M_i : \mathbb{R} \to \mathbb{R}$ is a monotonically increasing and Lipschitz continuous function and $\kappa_i : \Lambda \to \Lambda$ is Lipschitz continuous and has a Lipschitz continuous inverse. Moreover, we assume (3.4.64) and (3.4.76) for $C_i > 0$. Then the altered Robin method (3.4.57)–(3.4.59) provides sequences of iterates $(u_1^k)_{k\ge1}$ and $(u_2^k)_{k\ge0}$ converging in V_1 and in V_2, respectively, to the unique solution of the domain decomposition problem (3.4.12)–(3.4.14) which, in addition, is well-posed with respect to $f \in L^2(\Omega)$.

The proof is mainly based on Theorem 3.4.12 and Propositions 3.4.15 and 3.4.16 and will be carried out in detail for a more concrete one-dimensional result in Theorem 3.4.23 to which we turn in the next subsection.

3.4.4 Convergence of the Robin method in 1D for the time-discretized nondegenerate Richards equation in heterogeneous soil

The purpose of this subsection is to prove (3.4.65) and (3.4.66) in 1D for our setting in Subsection 3.4.1. As already indicated at the end of the last subsection, the main task will be the proof of (3.4.76). Note that we have already come across a similar inequality in the proof of Proposition 3.3.7, where we derived an inverse estimate for $M_i = 0$ and monotonically increasing and Lipschitz continuous $\kappa_i, \kappa_i^{-1} : \mathbb{R} \to \mathbb{R}$, $i = 1, 2$, in one space dimension. The proof of (3.4.76) in this case could be carried out with the same arguments as used there.

In the following, we prove that (3.4.76) holds in one space dimension if M_i and κ_i, $i = 1, 2$, are as in the setting of the implicit-explicitly time-discretized Richards equation (3.4.6)–(3.4.8). To achieve this we need to look more closely at the concrete monotonic behaviour of the solutions $u_i(\lambda)$ of (3.4.35)–(3.4.37) for varying λ in 1D. The latter can be carried out after we have established the following regularity result for these solutions.

Lemma 3.4.20. *Let $\Omega_1 = (a, b)$ and $\Omega_2 = (b, c)$ for $a < b < c$ and $i = 1, 2$. Suppose that $M_i : \mathbb{R} \to \mathbb{R}$ satisfies the conditions in Theorem 3.4.12. Then for any $\lambda \in \mathbb{R}$ the weak solution $u_i(\lambda) \in V_i$ of (3.4.35)–(3.4.37) is also a strong solution of (3.4.35)–(3.4.37) satisfying $u_i(\lambda) \in C^2(\overline{\Omega}_i)$.*

Proof. Let $i = 1, 2$. The conditions on M_i guarantee the unique solvability of (3.4.35)–(3.4.37) in the weak sense (see Remark 3.4.4). Since (3.4.33)–(3.4.34) is also uniquely solvable in the weak sense ($f_{|\Omega_i} \in L^2(\Omega_i)$ is always assumed) and the Sobolev embedding theorem (2.5.36) provides $u_i(\lambda), u_i^* \in C(\overline{\Omega}_i)$ in 1D, we also have

$$M_i(u_i(\lambda) + u_i^*) \in C(\overline{\Omega}_i) \qquad (3.4.78)$$

because $M_i : \mathbb{R} \to \mathbb{R}$ is continuous. The weak form of (3.4.35) reads

$$\int_{\Omega_i} M_i(u_i(\lambda) + u_i^*)\, v_i\, dx + \int_{\Omega_i} u_i(\lambda)'\, v_i'\, dx = 0 \quad \forall v_i \in V_i^0\,. \qquad (3.4.79)$$

Now consider $i = 1$. Due to (3.4.78) the primitive $\int_a^{\cdot} M_1(u_1(\lambda) + u_1^*)\, ds$ is in $C^1(\overline{\Omega}_1) \cap V_1$ such that weak partial integration (A.2.11) gives

$$\int_{\Omega_1} M_1(u_1(\lambda) + u_1^*)\, v_1\, dx = -\int_{\Omega_1} \left(\int_a^{\cdot} M_1(u_1(\lambda) + u_1^*)\, ds \right) v_1'\, dx \quad \forall v_1 \in V_1^0\,.$$

Consequently, (3.4.79) provides

$$\int_{\Omega_1} \left(u_1(\lambda)' - \int_a^{\cdot} M_1(u_1(\lambda) + u_1^*)\, ds \right) v_1'\, dx = 0 \quad \forall v_1 \in V_1^0 \qquad (3.4.80)$$

or equivalently

$$a_1(w_1, v_1) = 0 \quad \forall v_1 \in V_1^0 \qquad (3.4.81)$$

for the function $w_1 \in V_1 \subset C(\overline{\Omega}_1)$ defined by

$$w_1 = u_1(\lambda) - \int_a^{\cdot} \int_a^x M_1(u_1(\lambda) + u_1^*)\, ds\, dx\,.$$

In case of $i = 2$ we can argue in the same way as for $i = 1$ by replacing the homogeneous Dirichlet boundary a by c. Now, again for both $i = 1, 2$, it is well known that the energy scalar product $a_i(\cdot, \cdot)$ induces an equivalent norm on V_i (see (A.2.13)) such that (3.4.81) entails

$$(w_i, \varphi_i)_{\Omega_i} = 0 \quad \forall \varphi_i \in C_0^\infty(\Omega_i)$$

for the L^2-scalar product with the dense subset $C_0^\infty(\Omega_i)$ of $L^2(\Omega_i)$ (or of V_i^0). We conclude $w_i = 0$ in $L^2(\Omega_i)$, and since $w_i, u_i(\lambda) \in C(\overline{\Omega}_i)$ we obtain the identities

$$u_i(\lambda) = \int_a^{\cdot} \int_a^x M_i(u_i(\lambda) + u_i^*)\, ds\, dx \quad \text{and} \quad u_i(\lambda)'' = M_i(u_i(\lambda) + u_i^*) \quad (3.4.82)$$

pointwise on $\overline{\Omega}_i$ and with a replaced by c for $i = 2$. □

Lemma 3.4.21. *Let $\Omega_1 = (a, b)$ and $\Omega_2 = (b, c)$ for $a < b < c$ and $i = 1, 2$. Suppose that $M_i : \mathbb{R} \to \mathbb{R}$ satisfies the conditions in Theorem 3.4.12 and κ_i is a monotonically increasing real function. Then for real $\eta \geq \mu$ the solutions $u_i(\eta)$ and $u_i(\mu)$ of (3.4.35)–(3.4.37) satisfy*

$$\begin{aligned} u_i(\eta) &\geq u_i(\mu) \quad \text{on } \Omega_i & (3.4.83)\\ u_1(\eta)' &\geq u_1(\mu)' \quad \text{on } \Omega_1 & (3.4.84)\\ u_2(\eta)' &\leq u_2(\mu)' \quad \text{on } \Omega_2\,. & (3.4.85)\end{aligned}$$

Proof. For the proof of (3.4.83) we first consider $i = 1$. Suppose the functions $u_1(\eta)$ and $u_1(\mu)$ coincide in a point $x \in (a, b]$, then they coincide on the whole interval $[a, x]$ due to the unique solvability of the convex subproblem (3.4.35)–(3.4.37) for $i = 1$ restricted to $[a, x]$ which is guaranteed by the assumptions on M_1, see Remark 3.4.4. Since κ_1 is monotonically increasing and $u_1(\eta) = u_1(\mu)$ for $\kappa_1\eta = \kappa_1\mu$ we can assume $\kappa_1\eta > \kappa_1\mu$. Then, due to the continuity of $u_1(\eta)$ and $u_1(\mu)$ provided by (2.5.36), we conclude $u_1(\eta) > u_1(\mu)$ on an interval $(x, b]$ where x is either a or the maximum of the elements in $[a, b)$ in which $u_1(\eta)$ and $u_1(\mu)$ coincide. This shows $u_1(\eta) \geq u_1(\mu)$ on Ω_1.

$u_2(\eta) \geq u_2(\mu)$ on Ω_2 follows analogously.

We turn to the proof of (3.4.84) and (3.4.85). Since $u_i(\lambda)'' = M_i(u_i(\lambda) + u_i^*)$, $i = 1, 2$, holds for any real λ due to Lemma 3.4.20, we can write

$$u_1(\lambda)' = \int_a^{\cdot} M_1(u_1(\lambda) + u_1^*)\, ds \tag{3.4.86}$$

with the homogeneous boundary condition in a. Due to $u_1(\eta) \geq u_1(\mu)$ on Ω_1 we also have $u_1(\eta) + u_1^* \geq u_1(\mu) + u_1^*$ on Ω_1, and since M_1 is monotonically increasing we conclude

$$u_1(\eta)' - u_1(\mu)' = \int_a^{\cdot} (M_1(u_1(\eta) + u_1^*) - M_1(u_1(\mu) + u_1^*))\, ds \geq 0 \quad \text{on } \Omega_1\,.$$

Arguing in the same way for $i = 2$ we obtain instead

$$u_2(\eta)' - u_2(\mu)' = \int_c^{\cdot} (M_2(u_2(\eta) + u_2^*) - M_2(u_2(\mu) + u_2^*))\, ds \leq 0 \quad \text{on } \Omega_2$$

with the homogeneous boundary condition in c and the reversed direction of integration on Ω_2. □

We remark that it seems unclear whether the estimates (3.4.83) are strict inequalities on Ω_i if κ_i, $i = 1, 2$, are strictly increasing. On the other hand, in the Richards equation we usually have continuous M_i and $M_i \geq 0$ (or even $M_i > 0$). Then, due to $u_i(\lambda)'' = M_i(u_i(\lambda)+u_i^*)$ we obtain (strict) convexity of $u_i(\lambda)$ on Ω_i and therefore (strict) monotonicity of $u_i(\lambda)'$ on at most two subintervals of Ω_i. Note that this implies $u_i(0) \leq 0$ (or even $u_i(0) < 0$) on Ω_i for $i = 1, 2$. Finally, observe that the solutions $u_i = u_i(\lambda) + u_i^*$ of the inhomogeneous subproblems (3.4.38)–(3.4.40) also satisfy the inequalities (3.4.83)–(3.4.85).

The next proposition can be regarded as corresponding to Proposition 3.3.7 for the problem considered in the last section. As done there, we apply the concrete form of the harmonic extension operators H_i, $i = 1, 2$, in one space dimension in the proof of this crucial result.

Proposition 3.4.22. *Let $\Omega_1 = (a, b)$ and $\Omega_2 = (b, c)$ for $a < b < c$ and $i = 1, 2$. Suppose that $M_i : \mathbb{R} \to \mathbb{R}$ satisfies the conditions in Theorem 3.4.12 and κ_i is monotonically increasing and Lipschitz continuous. Then (3.4.76) is satisfied for $C_i = L(\kappa_i)^{-1}$ with the Lipschitz constant $L(\kappa_i)$ of κ_i.*

Proof. Suppose without loss of generality $\eta > \mu$. Then, considering first $i = 1$ we have for any $x \in (a, b)$

$$\begin{aligned}(H_1\eta - H_1\mu)(x) &= \frac{\eta-\mu}{b-a}(x-a)\\ &\geq \frac{\kappa_1\eta - \kappa_1\mu}{L(\kappa_1)(b-a)}(x-a) = \frac{1}{L(\kappa_1)}(H_1\kappa_1\eta - H_1\kappa_1\mu)(x) \geq 0\end{aligned}$$

which obviously entails

$$\begin{aligned}&(M_1(u_1(\eta)+u_1^*) - M_1(u_1(\mu)+u_1^*), H_1\eta - H_1\mu)_{\Omega_1}\\ &\qquad\geq \frac{1}{L(\kappa_1)}(M_1(u_1(\eta)+u_1^*) - M_1(u_1(\mu)+u_1^*), H_1\kappa_1\eta - H_1\kappa_1\mu)_{\Omega_1}\end{aligned}$$

because (3.4.83) gives $u_1(\eta) + u_1^* \geq u_1(\mu) + u_1^*$ and the monotonicity of M_1 provides $M_1(u_1(\eta)+u_1^*) - M_1(u_1(\mu)+u_1^*) \geq 0$ on Ω_1. Analogously, considering $i = 2$ we obtain for any $x \in (b, c)$

$$\begin{aligned}(H_2\eta - H_2\mu)(x) &= \frac{\eta-\mu}{c-b}(c-x)\\ &\geq \frac{\kappa_2\eta - \kappa_2\mu}{L(\kappa_2)(c-b)}(c-x) = \frac{1}{L(\kappa_2)}(H_2\kappa_2\eta - H_2\kappa_2\mu)(x) \geq 0\end{aligned}$$

and argue in the same way.

Furthermore, for $i = 1$ we can estimate from below

$$\nabla(H_1\eta - H_1\mu) = \frac{\eta-\mu}{b-a} \geq \frac{\kappa_1\eta-\kappa_1\mu}{L(\kappa_1)(b-a)} = \frac{1}{L(\kappa_1)}\nabla(H_1\kappa_1\eta - H_1\kappa_1\mu) \geq 0$$

which leads to

$$a_1(u_1(\eta) - u_1(\mu), H_1\eta - H_1\mu) \geq \frac{1}{L(\kappa_1)}a_1(u_1(\eta) - u_1(\mu), H_1\kappa_1\eta - H_1\kappa_1\mu)$$

since we have $\nabla(u_1(\eta) - u_1(\mu)) \geq 0$ on Ω_1 due to (3.4.84). In the case $i = 2$ we estimate from above

$$\nabla(H_2\eta - H_2\mu) = -\frac{\eta-\mu}{c-b} \leq -\frac{\kappa_2\eta-\kappa_2\mu}{L(\kappa_2)(b-a)} = \frac{1}{L(\kappa_2)}\nabla(H_2\kappa_2\eta - H_2\kappa_2\mu) \leq 0$$

which then gives again

$$a_2(u_2(\eta) - u_2(\mu), H_2\eta - H_2\mu) \geq \frac{1}{L(\kappa_2)}a_2(u_2(\eta) - u_2(\mu), H_2\kappa_2\eta - H_2\kappa_2\mu)$$

since we now have $\nabla(u_2(\eta) - u_2(\mu)) \leq 0$ on Ω_2 because of (3.4.85). □

Collecting the results of this and the previous subsections we obtain the main theorem of this section. It states the convergence of the Robin method (3.4.15)–(3.4.18) for the domain decomposition problem (3.4.6)–(3.4.8) and its original version (3.4.2)–(3.4.4) concerning the Richards equation and, in addition, it asserts the well-posedness of these problems in one space dimension.

Theorem 3.4.23. *Let $\gamma_1 = \gamma_2 > 0$, $a < b < c$ and $\Omega_1 = (a, b)$ and $\Omega_2 = (b, c)$. Suppose $M_i : \mathbb{R} \to \mathbb{R}$ are monotonically increasing and Lipschitz continuous and $\kappa_i : \mathbb{R} \to \mathbb{R}$ are monotonically increasing, Lipschitz continuous and invertible with Lipschitz continuous κ_i^{-1} for $i = 1, 2$.*

Then the Robin method (3.4.20)–(3.4.23) provides sequences $(u_1^k)_{k\geq 1}$ as well as $(u_2^k)_{k\geq 0}$ of iterates converging in V_1 and in V_2, respectively, to the unique solution of the domain decomposition problem (3.4.12)–(3.4.14).

Furthermore, the domain decomposition problem (3.4.9)–(3.4.11) has a unique solution to which the sequences $(\kappa_1^{-1} u_1^k)_{k\geq 1}$ and $(\kappa_2^{-1} u_2^k)_{k\geq 0}$ of the retransformed iterates converge in V_1 and in V_2, respectively.

Finally, both domain decomposition problems (3.4.12)–(3.4.14) and (3.4.9)–(3.4.11) are well-posed with respect to the data $f \in L^2([a, c])$.

Proof. Let $i = 1, 2$. By Theorem 3.4.2 the conditions on $M_i : \mathbb{R} \to \mathbb{R}$ guarantee the unique solvability of the subproblems (3.4.20)–(3.4.21) and (3.4.22)–(3.4.23) as well as (3.4.64). In one space dimension the Robin method (3.4.20)–(3.4.23) and the altered Robin method (3.4.56)–(3.4.59) coincide so that Theorem 3.4.12 is applicable. With the given conditions on $M_i : \mathbb{R} \to \mathbb{R}$ and $\kappa_i : \mathbb{R} \to \mathbb{R}$ we obtain the Lipschitz continuity (3.4.65) of the Steklov–Poincaré operators by Proposition 3.4.15 and their strong monotonicity (3.4.66) by Propositions 3.4.16 and 3.4.22.

Now, Theorem 3.4.12 provides a converging sequence $(\lambda_2^k)_{k\geq 0} \subset \Lambda$ for which we have $u_2^k = u_2(\lambda_2^k) + u_2^*$ for the iterates u_2^k, $k \geq 0$, from the Robin method (3.4.20)–(3.4.23) by Proposition 3.4.7. Due to Proposition 3.4.14 the convergence $\lambda_2^k \to \lambda$ for $k \to \infty$ in Λ entails the convergence $u_2^k \to u_2$ for $k \to \infty$ in V_2 to the unique solution $u_2 = u_2(\lambda) + u_2^*$ on Ω_2 of the domain decomposition problem (3.4.12)–(3.4.14) since λ is the unique fixed point of $\mathcal{T}_\gamma$ (Proposition 3.4.8 and Theorem 3.4.12).

Since $(\lambda_2^k)_{k\geq 0}$ converges in Λ, so does $(\lambda_1^k)_{k\geq 1}$ because we have

$$\lambda_1^{k+1} = (\gamma\mathcal{I} + S_1)^{-1}(\gamma\mathcal{I} - S_2)\lambda_2^k \quad \forall k \geq 0 \tag{3.4.87}$$

due to (3.4.45) and the operators $\gamma\mathcal{I} - S_2 : \Lambda \to \Lambda'$ and $(\gamma\mathcal{I} + S_1)^{-1} : \Lambda' \to \Lambda$ are continuous. For the latter this is a consequence of the strong monotonicity (3.4.66) of S_1 which provides

$$\langle(\gamma\mathcal{I} + S_1)\eta - (\gamma\mathcal{I} + S_1)\mu, \eta - \mu\rangle \geq (c_i + \gamma)\|\eta - \mu\|_\Lambda^2 \quad \forall \eta, \mu \in \Lambda$$

with $\langle\gamma\mathcal{I}(\eta - \mu), \eta - \mu\rangle = \gamma\|\eta - \mu\|_\Lambda^2$ (see (3.4.55)), i.e. the Lipschitz continuity of $(\gamma\mathcal{I} + S_1)^{-1}$. Now, with $\lambda_2^k \to \lambda$ for $k \to \infty$ and (3.4.87) we obtain a $\bar{\lambda} \in \Lambda$

such that we have $\lambda_1^k \to \bar{\lambda}$ for $k \to \infty$ and

$$(\gamma\mathcal{I} + S_1)\bar{\lambda} = (\gamma\mathcal{I} - S_2)\lambda\,. \tag{3.4.88}$$

However, we also have (3.4.46) which gives

$$\lambda_2^{k+1} = (\gamma\mathcal{I} + S_2)^{-1}(\gamma\mathcal{I} - S_1)\lambda_1^{k+1} \quad \forall k \geq 0$$

such that we obtain

$$(\gamma\mathcal{I} - S_1)\bar{\lambda} = (\gamma\mathcal{I} + S_2)\lambda \tag{3.4.89}$$

and adding (3.4.88) to (3.4.89) gives $\bar{\lambda} = \lambda$.

Now, with the same reasoning as above for $i = 2$ we obtain the convergence $u_1^k \to u_1$ for $k \to \infty$ of the iterates u_1^k from the Robin method (3.4.20)–(3.4.23) in V_1 to the unique solution $u_1 = u_1(\lambda) + u_1^*$ on Ω_1 of the domain decomposition problem (3.4.12)–(3.4.14).

Finally, the unique solvability of (3.4.12)–(3.4.14) entails the unique solvability of the original problem (3.4.9)–(3.4.11) by Proposition 3.4.1, and Theorem 1.5.15 provides the convergence of the retransformed iterates to this unique physical solution p on Ω with $p_{|\Omega_i} = p_i$ in V_i, $i = 1, 2$.

As far as the continuous dependency of u_i or p_i on the data $f \in L^2([a, c])$ is concerned for $i = 1, 2$, observe first that the convex subproblems (3.4.38)–(3.4.40) are well-posed due to Proposition 2.4.11. (This can also be obtained directly by testing the difference of (3.4.38) for two solutions corresponding to different right hand sides with the difference of these solutions). Then another application of Theorem 1.5.15 leads to the well-posedness of the retransformed subproblems and therefore of (3.4.9)–(3.4.11). □

Remark 3.4.24. With a glance at Remark 2.3.17, we note that Theorem 3.4.23 can be extended to more generalized situations in which there are space-dependent parameter functions on Ω_i like the porosity $n_i(\cdot)$ as a factor in front of $\theta_i(p_i)$ and the hydraulic conductivity $K_{h,i}(\cdot)$ as a factor in front of $k_i(p_i)$ in (3.4.2) for $i = 1, 2$. In contrast to Remark 2.3.17, however, we need to make sure that these functions fit into our one-dimensional theory, in particular into the regularity and the monotonicity results in Lemmas 3.4.20 and 3.4.21.

Going through the proofs of these lemmas, it becomes clear from (3.4.78) and the proof of (3.4.84) and (3.4.85) that we need to choose nonnegative $n_i(\cdot) \in C(\overline{\Omega}_i)$ while $K_{h,i}(\cdot)$ should be Lipschitz continuous on Ω_i, $i = 1, 2$, with $K_{h,i}(\cdot) \geq c > 0$ as always. Then $K_{h,i}(\cdot)^{-1}$ is Lipschitz continuous on Ω_i and therefore an element of $H^1(\Omega_i)$ with $(K_{h,i}(\cdot)^{-1})' \in L^\infty(\Omega_i)$ (see e.g. [15, p. 25]). Therefore, the product function

$$K_{h,1}(\cdot)^{-1} \int_a^{\cdot} n_1(s)\, M_1((u_1(\lambda) + u_1^*)(s))\, ds$$

appearing in (3.4.80) instead of the primitive of $M_1(u_1(\lambda) + u_1^*)$ is Lipschitz continuous on Ω_1 and in V_1, too. Altogether, we obtain

$$u_i(\lambda) = \int_a^{\cdot} K_{h,i}(x)^{-1} \int_a^x n_i(s)\, M_i((u_i(\lambda) + u_i^*)(s))\, ds\, dx$$

and $(K_{h,i}\, u_i(\lambda)')' = n_i\, M_i(u_i(\lambda) + u_i^*)$ instead of (3.4.82) pointwise on Ω_i for $i = 1,2$ with a replaced by c for $i = 2$. Now, with this result the proof of Lemma 3.4.21 can be carried out in the same way.

Remark 3.4.25. Unfortunately, we cannot extend our theory of the last two subsections to higher space dimensions because it is based on a contraction argument for the proof of Theorem 3.4.12, which requires the strong monotonicity (3.4.66) of the Steklov–Poincaré operators. This property may not hold in 2D since the same counterexamples as given in Subsection 3.3.4 apply in our more general situation if we set $M_i = 0$ for $i = 1,2$.

As already pointed out in Remark 3.4.17, our nonlinear Steklov–Poincaré operators S_i defined in 3.4.42 already contain the nonlinearities κ_i, $i = 1,2$, due to the definition of $u_i(\lambda)$ as the solutions of (3.4.35)–(3.4.37). Therefore, they do not reduce to the Steklov–Poincaré operators of Section 3.3 but to the composition $S_i\kappa_i$ in (3.3.31) with the definition (3.3.29) of the well-known linear operators S_i for $i = 1,2$.

Despite this lack of rigidness of the notation, it should not confuse since it is justified by the different requirements of the convergence theory for the Dirichlet–Neumann method (3.3.33)–(3.3.37) and the Robin method (3.4.20)–(3.4.23). Recall that the convergence of the former could only be proved by the linearity of the preconditioner S_2 for the unsymmetric Steklov–Poincaré interface equation (3.3.32), whose solution λ_2 is the limit of the corresponding sequence $(\lambda_2^k)_{k\geq 0}$ of transformed variables and occurs as a transformed physical pressure $\lambda_2 = u_{2|\Gamma} = \kappa_2(p_{2|\Gamma}) = \kappa_2(p_{1|\Gamma})$. In contrast, the convergence of the latter is based on the symmetric interface equation (3.3.31) to which (3.4.42) reduces for $M_i = 0$, $i = 1,2$. Here, the sequence $(\lambda_2^k)_{k\geq 0}$ of interface values in Theorem 3.4.12 consists of physical variables converging to the solution p_i, $i = 1,2$, in the form $\lambda_2^k \to p_{1|\Gamma} = p_{2|\Gamma} = \lambda$ for $k \to \infty$ with λ as in (3.3.31). And since for general $M_i \neq 0$, the Steklov–Poincaré operators are nonlinear anyway, they are defined in (3.4.42) as to contain both nonlinearities M_i and κ_i for $i = 1,2$.

Our treatment of the domain decomposition problem (3.3.11)–(3.3.13) in Proposition 3.3.7 is based on the special properties of the harmonic extension operators H_i, $i = 1,2$. Note, however, that the theory in this section and Remark 3.4.24 also guarantee the convergence of the nonlinear Dirichlet–Neumann method (3.3.33)–(3.3.37) if we replace $k_i(p_i)$ in (3.3.11) by $K_{h,i}(\cdot)\, k_i(p_i)$ with space-dependent functions $K_{h,i}(\cdot)$ as in Remark 3.4.24. This can easily be seen if we consider the function $\kappa := \kappa_1\kappa_2^{-1} : \mathbb{R} \to \mathbb{R}$ which has the same properties as κ_1 and κ_2 such that Propositions 3.4.15, 3.4.16 and 3.4.22 can be applied on the nonlinear operator $S_1\kappa_1\kappa_2^{-1} : \Lambda \to \Lambda'$ in Subsection 3.3.3. Then, well-posedness of (3.3.11)–(3.3.13) can be obtained analogously as in the proof of Theorem 3.4.23.

Remark 3.4.26. With regard to Theorem 3.4.12 we note that in Lions and Mercier [66] there has been done some nice analysis on a nonlinear ADI method given by the operator

$$T_\delta = (I + \delta B)^{-1}(I - \delta A)(I + \delta A)^{-1}(I - \delta B) : H \to H$$

with $\delta > 0$ for inclusion problems

$$u \in H : \quad C(u) \ni 0 \tag{3.4.90}$$

involving a monotone operator $C = A + B$ on a Hilbert space H on which A and B are maximal monotone. The severe restriction which is imposed in [66] is the assumption that (3.4.90) is solvable. Then, strong convergence of the sequence given by a $u^0 \in H$ and the iteration $u^{k+1} = T_\delta u^k$, $k \geq 0$, to the solution u of (3.4.90), which then turns out to be unique, can be proved e.g. if B is strongly monotone.

We remark that one can alter the proofs in [66] such that they cover the situation of (possibly multivalued monotone) operators $A, B : H \to H'$ and I replaced by the Riesz operator $\mathcal{I} : H \to H'$. Then the results are also applicable to our situation above with $H = \Lambda$, $A = S_1$ and $B = S_2$, however, as already emphasized, with the assumption that a solution of $(S_1 + S_2)\lambda = 0$ exists.

Remark 3.4.27. At the end of our discussion on the convergence of the Robin method in the continuous setting and in view of Remark 3.4.26 it seems to be in order to comment on Lions [65], which one might regard as the classical reference concerning the Robin method. In this paper, proofs for the convergence of the Robin method (in different topologies and in different spaces) can be found for linear cases and an arbitrary number of subdomains.

Nice variants of these proofs, which are based on “energy type” estimates, can be found in Quarteroni and Valli [75, pp. 135–137, 242–244] for the special case of two subdomains and a single Robin parameter $\gamma = \gamma_1 = \gamma_2 > 0$. The idea is to deduce that the sequence of error norms tends to 0 by proving that the series of these norms converges. Therefore, one needs to assume the existence of a solution and one does not obtain convergence rates.

Unfortunately, these classical proofs do not seem to work for our problems related to the Richards equation. The drawback one encounters if one follows this approach is that the main part in the spatial derivative of the Richards equation (see e.g. (3.4.1) and (3.4.2)) does not generate a monotone operator, i.e. we do not have

$$\int_{\Omega_i} \left(k_i(p_i^1)\nabla p_i^1 - k_i(p_i^2)\nabla p_i^2\right) \nabla(p_i^1 - p_i^2) \geq 0$$

for $p_i^1, p_i^2 \in H^1_{\partial\Omega\cap\partial\Omega_i}(\Omega_i)$ and positive $k_i \in L^\infty(\Omega_i)$ in general, not even if the functions k_i are monotonically increasing for $i = 1, 2$. This can be different if the relative permeabilities k_i are functions of ∇p_i rather than functions of p_i with $p_i \in H^1_{\partial\Omega\cap\partial\Omega_i}(\Omega_i)$, $i = 1, 2$, for example if they generate (possibly different) p-Laplacians. In the latter case, the corresponding operators are well-known to be strictly monotone, i.e., with a *change to the usual notation for p-Laplacians,* we have

$$\int_{\Omega_i} \left(|\nabla u_i^1|^{p_i-2}\nabla u_i^1 - |\nabla u_i^2|^{p_i-2}\nabla u_i^2\right) \nabla(u_i^1 - u_i^2) > 0$$

for real numbers $1 < p_i < \infty$, now, and functions $u_i^1, u_i^2 \in W^{1,p_i}_{\partial\Omega\cap\partial\Omega_i}(\Omega_i)$ with $u_i^1 \neq u_i^2$ for $i = 1, 2$. (We refer to [81, p. 71/72] or [102, p. 568] for more details.) The corresponding relative permeability functions

$$k_i : u_i \mapsto |\nabla u_i|^{p_i - 2}\,, \quad i = 1, 2\,,$$

lead to nonlinear versions of Darcy's law (1.2.3) in case of full saturation and have been considered in Chipot and Lyaghfouri [26]. In that sense, an analogue of the domain decomposition problem (3.3.11)–(3.3.13) for p-Laplacians related to different p_i with $1 < p_i < \infty$, $i = 1, 2$, can be given a hydrological interpretation.

Moreover, again assuming the existence of a solution to such a problem, the proof in [75, pp. 135–137], given in a differential form, can in principle be carried over to the situation of different p-Laplacians. However, one needs to assume that the "normal fluxes" of the solution and of the initial iterate across Γ belong to $L^2(\Gamma)$ as well as their traces and the traces of all iterates on Γ provided by the Robin method. The latter is guaranteed by

$$p_i \geq \frac{2d}{d+1}\,, \quad i = 1, 2\,,$$

with the space dimension d according to Theorem A.2.2. But still, the unique solvability of the Robin problems involved is not clear in all cases since the trace spaces generated by $W^{1,p_i}(\Omega_i)$, $i = 1, 2$, on Γ (and their norms) do not coincide for $p_1 \neq p_2$. Here, one probably needs to impose further regularity assumptions in order to succeed as in the proof given in [75, pp. 242–244] for the weak formulation of the problem.

3.4.5 Robin method applied to the time- and space-discretized Richards equation, convergence and numerical treatment

In this subsection we take a close look at the Robin method for the time-discretized Richards equation, its space discretization and the numerical treatment how we perform it. We obtain convergence of the discrete Robin method in one space dimension. The considerations here will be completed in Section 4.2 where we finally include gravity.

Note that for the Dirichlet and Neumann subproblems as in Section 3.3 the space discretization and the numerical treatment is clear even in the nonlinear case of the Richards equation (without gravity) since they have already been discussed in Section 2. This is not true for the Robin subproblems which are even nonlinear in the stationary case of Section 3.3 due to the contribution of the physical pressure which is the retransformed unknown on the interface. Unfortunately, we do not have a suitable discretization and a numerical treatment for the altered Robin method (3.4.56)–(3.4.59), let alone a treatment based on convex analysis. We do not even know if the corresponding continuous subproblems are well-posed in general. Nevertheless, in one space dimension the Robin method and its altered version coincide and the convergence proof in Subsections 3.4.3 and 3.4.4 can be successfully translated into discrete arguments.

Robin method for the implicit-explicitly time-discretized Richards equation

With regard to the numerical treatment of the Robin method for the implicit-explicitly time-discretized Richards equation, we consider it once again in concrete terms here since this has not been done in complete detail in Subsection 3.4.1. Neglecting constants as well as the porosity n and the hydraulic conductivity K_h, our time-discretized Richards equation in the Kirchhoff–transformed version reads

$$\frac{M_i(u_i) - M_i(\tilde{u}_i)}{\tau} - \operatorname{div}\Big(\nabla u_i - kr_i(M_i(\tilde{u}_i))e_z\Big) = 0 \quad \text{on } \Omega_i \tag{3.4.91}$$

for $i = 1, 2$ with the time step size $\tau > 0$, see (3.2.19). Here, the time-discretized (unknown) water flux for the present time step expressed in the transformed variables is

$$\mathbf{v}_i = -(\nabla u_i - kr_i(M_i(\tilde{u}_i))e_z) =: -\tilde{\mathbf{v}}_i\,, \quad i = 1, 2\,.$$

Let us consider Neumann data $\mathbf{v} \cdot \mathbf{n} = f_N \in L^2(\gamma_N)$ (which is flow of water out of Ω) on a subset $\gamma_N \subset \partial\Omega$ as well as homogeneous Dirichlet values on $\gamma_D := \partial\Omega \backslash \gamma_N$ with $\gamma_{N_i} := \partial\Omega_i \cap \gamma_N$ and nontrivial $\gamma_{D_i} := \partial\Omega_i \cap \gamma_D$ for $i = 1, 2$. (Inhomogeneous Dirichlet data and boundary conditions of Signorini's type can be included according to Chapter 2 where this case has been extensively discussed.)

In this setting, we look for $u_i \in \tilde{V}_i := H^1_{\gamma_{D_i}}(\Omega_i)$ (with the corresponding trace space Λ), where $\tilde{u}_i \in H^1(\Omega_i)$, $i = 1, 2$, is supposed to be known from the previous time step. As indicated in (3.4.5) and in Remark 3.4.3, the Robin interface conditions (3.4.16) and (3.4.18) are replaced by

$$\begin{aligned} \tilde{\mathbf{v}}_1^{k+1} \cdot \mathbf{n} + \gamma_1\, \kappa_1^{-1}(u_1^{k+1}) &= \tilde{\mathbf{v}}_2^{k} \cdot \mathbf{n} + \gamma_1\, \kappa_2^{-1}(u_2^{k}) && \text{on } \Gamma \quad (3.4.92)\\ \tilde{\mathbf{v}}_2^{k+1} \cdot \mathbf{n} - \gamma_2\, \kappa_2^{-1}(u_2^{k+1}) &= \tilde{\mathbf{v}}_1^{k+1} \cdot \mathbf{n} - \gamma_2\, \kappa_1^{-1}(u_1^{k+1}) && \text{on } \Gamma\,. \quad (3.4.93) \end{aligned}$$

We derive a weak formulation of (3.4.92) by testing (3.4.92) with $\mu \in \Lambda$ and setting $v_i = R_i\mu$ with linear continuous extension operators $R_i : \Lambda \to \tilde{V}_i$ for $i = 1, 2$. First, Green's formula (1.5.9) (or (A.2.12) in the weak sense) gives

$$\int_{\Omega_i} (\operatorname{div} \tilde{\mathbf{v}}_i)\, v_i\, dx = -\int_{\Omega_i} \tilde{\mathbf{v}}_i \nabla v_i\, dx + \int_{\gamma_{N_i}} \tilde{\mathbf{v}}_i \cdot \mathbf{n}_i\, v_i\, d\sigma + \int_\Gamma \tilde{\mathbf{v}}_i \cdot \mathbf{n}_i\, \mu\, d\sigma \tag{3.4.94}$$

for $i = 1, 2$, with the outward normal $\mathbf{n}_i$ of Ω_i and $\operatorname{div} \tilde{\mathbf{v}}_i = \tau^{-1}(M_i(u_i) - M_i(\tilde{u}_i))$ as well as $\tilde{\mathbf{v}}_i \cdot \mathbf{n}_i = -f_N$ on γ_{N_i}. Therefore, skipping the indices k and $k + 1$ in (3.4.92) as done in the proof of Theorem 3.4.2, the weak form of the left hand side in (3.4.92) reads

$$\begin{aligned} \tau \int_\Gamma \tilde{\mathbf{v}}_1 \cdot \mathbf{n}\, \mu\, d\sigma + \tau\, \gamma_1 \int_\Gamma \kappa_1^{-1}(u_1)\, \mu\, d\sigma \\ = \int_{\Omega_1} M_1(u_1)\, v_1\, dx + \tau \int_{\Omega_1} \nabla u_1 \nabla v_1\, dx + \tau\, \gamma_1 \int_\Gamma \kappa_1^{-1}(u_1)\, \mu\, d\sigma - \tilde{\ell}_1(v_1) \end{aligned}$$

with

$$\tilde{\ell}_1(v_1) := \int_{\Omega_1} M_1(\tilde{u}_1)\, v_1\, dx + \tau \int_{\Omega_1} kr_1(M_1(\tilde{u}_1)) e_z \nabla v_1\, dx - \tau \int_{\gamma_{N_1}} f_N\, v_1\, d\sigma\,.$$

Analogously, setting $\mathbf{n}_2 = -\mathbf{n}_1 = -\mathbf{n}$ in (3.4.94), we obtain the weak form of the right hand side in (3.4.92) as

$$\begin{aligned}
&\tau \int_\Gamma \tilde{\mathbf{v}}_2 \cdot \mathbf{n}\, \mu\, d\sigma + \tau\, \gamma_2 \int_\Gamma \kappa_2^{-1}(u_2)\, \mu\, d\sigma \\
&= -\Big(\int_{\Omega_2} M_2(u_2)\, v_2\, dx - \int_{\Omega_2} M_2(\tilde{u}_2)\, v_2\, dx + \tau \int_{\Omega_2} \nabla u_2 \nabla v_2\, dx \\
&\qquad - \tau \int_{\Omega_2} kr_2(M_2(\tilde{u}_2)) e_z \nabla v_2\, dx + \tau \int_{\gamma_{N_2}} f_N\, v_2\, d\sigma \Big) + \tau\, \gamma_2 \int_\Gamma \kappa_2^{-1}(u_2)\, \mu\, d\sigma
\end{aligned}$$

which can be regarded as a continuous linear functional $\tilde{\ell}_2$ on $\tilde{V}_1$ if one considers $v_2 = R_2\, tr_\Gamma v_1$ and $\mu = tr_\Gamma v_1$. We denote by

$$c_i(\tilde{u}_i, v_i) := \int_{\Omega_i} kr_i(M_i(\tilde{u}_i)) e_z \nabla v_i\, dx \tag{3.4.95}$$

the influence of the gravitation in the functionals ℓ_i, $i = 1, 2$. Altogether, setting $\ell_1 := \tilde{\ell}_1 + \tilde{\ell}_2$, the weak form of (3.4.92) reads

$$(M_1(u_1), v_1)_{\Omega_1} + \tau\gamma_1(\kappa_1^{-1} u_1, tr_\Gamma v_1)_\Gamma + \tau a_1(u_1, v_1) = \ell_1(v_1)\,, \quad v_1 = R_1 \mu\,, \ \ \forall \mu \in \Lambda\,.$$

Choosing $v_1 \in V_1^0$ in this equation gives

$$(M_1(u_1), v_1)_{\Omega_1} + \tau a_1(u_1, v_1) = (M_1(\tilde{u}_1), v_1)_{\Omega_1} + \tau c_1(\tilde{u}_1, v_1) - \tau (f_N, v_1)_{\gamma_{N_1}} \tag{3.4.96}$$

by which the weak formulation of (3.4.91) in V_1^0 is recovered (the scalar product in $L^2(\gamma_{N_1})$ is denoted by $(\cdot,\cdot)_{\gamma_{N_1}}$). Therefore, just as (3.4.24) in the proof of Theorem 3.4.2, the variational equality

$$u_1 \in \tilde{V}_1 : \ \ (M_1(u_1), v_1)_{\Omega_1} + \tau\gamma_1(\kappa_1^{-1} u_1, tr_\Gamma v_1)_\Gamma + \tau a_1(u_1, v_1) = \ell_1(v_1) \quad \forall v_1 \in \tilde{V}_1 \tag{3.4.97}$$

is equivalent to one iteration step of the Robin method for (3.4.91) in subdomain Ω_1 with the Robin transmission condition (3.4.92). Due to $\mathbf{n}_2 = -\mathbf{n}$, the weak form of the corresponding subdomain problem on Ω_2 with the Robin condition (3.4.93) turns out to be completely symmetric to (3.4.97). Therefore, we deal with (3.4.97) in the following.

Remark 3.4.28. The boundedness of $kr_i \in L^\infty(\mathbb{R})$ and the conditions on $M_i : \mathbb{R} \to \mathbb{R}$, $i = 1, 2$, in Theorem 3.4.2 guarantee the boundedness of ℓ_1 (see Propositions 2.5.11 and 2.5.12) and the unique solvability of (3.4.97). If $M_1 : [u_c, \infty) \to \mathbb{R}$ is monotonically increasing, continuous and bounded as in Chapter 2, this is still true as long as M_2, κ_2 and u_2 are such that $\ell_1 \in \tilde{V}_1'$. Note, however, if $\kappa_1^{-1} : (u_c, \infty) \to \mathbb{R}$ is as in case of the Brooks–Corey functions, i.e. unbounded around the singularity u_c (see (1.3.24) or Figure 1.9), we cannot guarantee the unique solvability of (3.4.97). In fact, we do not even have well-definedness $\kappa_1^{-1} u_1 \in L^2(\Gamma)$ with $u_1 \in \tilde{V}_1$ in this case for certain u_1 with ranges around the critical value u_c. But, again, if M_2, κ_2 and u_2 are such that $\ell_1 \in \tilde{V}_1'$, then one can still consider the related minimization problem as in (3.4.27). The convex function $\Psi_1 : [u_c, \infty) \to \mathbb{R}$ occurring in this problem is continuous in u_c because κ_1^{-1} has a finite integral on $(u_c, -1)$. Therefore, this minimization problem is uniquely solvable. We point out that, just as for the generalized saturation, this even works in the limit cases considered in Subsections 1.4.2 and 1.4.4. And again, we can solve the corresponding problems on the discrete level robustly in these limit cases (see later in this subsection). Nevertheless, we assume here that $\kappa_1 : \mathbb{R} \to \mathbb{R}$ has the properties as in Theorem 3.4.2. Then, with the additional convex functional ψ_1 on $\tilde{V}_1$ as defined in (3.4.25), which has the directional derivative as in (3.4.26), the variational equality (3.4.97) (with $\tau = 1$ for simplicity) is equivalent to finding the minimum of F_1 as given in (3.4.28) on $\tilde{V}_1$.

Space discretization of the subproblems in the Robin method

Now we turn to the space discretization of (3.4.97), more concretely its discretization with linear finite elements in a corresponding finite element space $\mathcal{S}_j^1 \subset \tilde{V}_1$, $j \geq 0$, with the set of nodes $\mathcal{N}_j^1$. This has already been dealt with in Section 2.5 except for the contribution $(\kappa_1^{-1} u_1, tr_\Gamma v_1)_\Gamma$ coming from the directional derivative of ψ_1. With a glance at (2.5.12) and (2.5.44), one finds that in light of Section 2.5, the "correct" discretization of this term must be given by the lumped $L^2(\Gamma)$-scalar product on the restrictions of the functions in $\mathcal{S}_j^1$ on Γ, which is the integral over the $\mathcal{S}_j^1$-interpolant of $\kappa_1^{-1}(u_1) \cdot v_1$ over Γ. In concrete terms, with an assumed discrete solution $u_{1,j} \in \mathcal{S}_j^1$ and $v \in \mathcal{S}_j^1$ as in Section 2.5, we discretize $(\kappa_1^{-1} u_1, tr_\Gamma v_1)_\Gamma$ by

$$\int_\Gamma \sum_{p \in \mathcal{N}_j^1} \kappa_1^{-1}(u_{1,j}(p))\, v(p)\, \lambda_p^{(j)}\, dx = \sum_{p \in \mathcal{N}_j^1 \cap \Gamma} \kappa_1^{-1}(u_{1,j}(p))\, v(p)\, h_{\Gamma,p}$$

with the weights

$$h_{\Gamma,p} := \int_\Gamma \lambda_p^{(j)}\, d\sigma \quad \forall p \in \mathcal{N}_j^1$$

which vanish for $p \notin \Gamma$. Obviously, this discretization corresponds to discretizing $\psi_1 : \tilde{V}_1 \to \mathbb{R}$, given via the convex function Ψ_1 in (3.4.25), by $\psi_{1,j} : \mathcal{S}_j^1 \to \mathbb{R}$ defined as

$$\psi_{1,j} : v \mapsto \sum_{p \in \mathcal{N}_j^1 \cap \Gamma} \Psi_1(v(p))\, h_{\Gamma,p} \quad \forall v \in \mathcal{S}_j^1\,. \tag{3.4.98}$$

It is analogous to the discretization of ϕ_1 in (3.4.28) by $\phi_{1,j}$ as in (2.5.2). With the properties of $\Psi_1' = \kappa_1^{-1} : \mathbb{R} \to \mathbb{R}$ in Theorem 3.4.2 one can first guarantee the unique solvability of the obtained discretization of (3.4.97) in $\mathcal{S}_j^1$ as in Subsection 2.5.1. Furthermore, one can derive the convergence of $u_{1,j} \in \mathcal{S}_j^1$ to the solution of (3.4.97) in the norm $\|\cdot\|_{1,\Omega_1}$ in $\tilde{V}_1$ for $j \to \infty$

analogously as in Subsection 2.5.2, which shall not be carried out here. Recall, however, that we are not really interested in this convergence result for the time-discretized Richards equation with Robin boundary values, but rather in the question whether the Robin method on the space-discretized level converges. (Note that the influence of the previous iterate u_2 in the second subdomain is still hidden in ℓ_1.) We will address this question and give a positive answer to it at the end of this subsection.

As far as the numerical realization of the discretized functional ℓ_1 is concerned, we approximate the above integrals after the finite element discretization by interpolation of $M_1(\tilde{u}_1)$ or $M_1(\tilde{u}_1)\,v$ in $\mathcal{S}_j^1$ and f_N in $\mathcal{S}_j^1$ restricted to Γ. As a consequence, a numerical treatment with the mass matrix given by the entries

$$\int_{\Omega_1} \lambda_p^{(j)} \lambda_q^{(j)}\, dx \quad \text{for } p, q \in \mathcal{N}_j^1$$

or its lumped version (with the entries $\int_{\Omega_1} \lambda_p^{(j)}\, dx$ for $p \in \mathcal{N}_j^1$ on the diagonal, and the same on Γ instead of Ω_1) is possible. The same applies to $\tilde{\ell}_2$. In case of a nonconstant porosity $n(\cdot)$ or a nonconstant hydraulic conductivity $K_h(\cdot)$ on Ω_1 quadrature rules need to be applied to the integral $(M_1(u_1), v_1)_{\Omega_1}$ and the bilinear form $a_1(u_1, v_1)$ in (3.4.97), too, see Subsection 2.5.3.

For the time being we can consider the contributions (3.4.95) from the gravitation in the linear functional to be vanishing. Our further treatment of these terms in the finite element discretization will be explained in Section 4.2.

Numerical treatment of the space-discretized subproblems

In the following, we deal with the numerical treatment of the discretized variational equality which, as just seen, results in a finite-dimensional convex minimization problem of finding the minimum $u_{1,j} \in \mathcal{S}_j^1$ of the functional

$$v \mapsto \phi_{1,j}(v) + \tau\gamma_1\psi_{1,j}(v) + \frac{1}{2}\tau a_1(v,v) - \ell_1(v) \quad \forall v \in \mathcal{S}_j^1\,. \tag{3.4.99}$$

On the one hand, we have $\phi_{1,j}(v) + \tau\gamma_1\psi_{1,j}(v) = \phi_{1,j}(v)$ for all $v \in \mathcal{S}_j^1$ whose support does not intersect Γ, i.e. for all nodes in $\mathcal{N}_j^1 \backslash \Gamma$ we are in the situation of the problems dealt with in our numerical Sections 2.6 and 2.7. On the other hand, if we replace ϕ_j by $\tilde{\phi}_j := \phi_{1,j} + \tau\gamma_1\psi_{1,j}$ in these sections, we can transfer the considerations therein to the present case.

More concretely, recall by the Gauss–Seidel step (2.6.7) or (2.6.12) that the convex function Φ in (2.5.2) may well depend on the nodes $p \in \mathcal{N}_j$ in the Gauss–Seidel method as long as we have the decoupling structure in (2.6.2), which is enforced by the definition (3.4.98) of $\psi_{1,j}$ for nodes $p \in \mathcal{N}_j^1 \cap \Gamma$, too. In addition, as emphasized at the beginning of Subsection 2.7.2, this generalization Φ_p instead of constant Φ for all $p \in \mathcal{N}_j$ is also covered by the theory of monotone multigrid.

As far as the practical realization of the Gauss–Seidel method is concerned, note that we only need to write the additional summand

$$\tau\gamma_1\Psi_1'(w_{l-1}^{\nu}(p_l)+z_l)\,h_{\Gamma,p_l} \quad \text{with} \quad \Psi_1'=\kappa_1^{-1}$$

on the right hand side of (2.6.13). Consequently, the nonlinear monotonically increasing (multi-)function $\mathcal{H}_l$ in (2.6.15) (possibly given on $\mathbb{R}$ for $M_1:\mathbb{R}\to\mathbb{R}$) needs to be replaced by the function

$$F_\Gamma(\cdot):=\mathcal{H}_l(\cdot)+\tau\gamma_1\kappa_1^{-1}(\cdot)\frac{h_{\Gamma,p_l}}{h_{p_l}} \tag{3.4.100}$$

on (u_c,∞) (or $\mathbb{R}$) so that for $p_l\in\Gamma$ the intersection of the graph of F_Γ and the line given by $G_{-\bar{w}_l}$ in Figure 2.1 has to be calculated instead of (2.6.16). Obviously, this also works for unbounded $\kappa_1^{-1}:(u_c,\infty)\to\mathbb{R}$ according to Brooks–Corey as in (1.3.24), see Figure 1.9. In fact, the Gauss–Seidel iteration converges to a unique minimum of (3.4.99) in this case of the fully discrete Richards equation according to Brooks–Corey and even in the limit cases (cf. Subsections 1.4.2 and 1.4.4), too, see Glowinski [45, pp. 142–147]. Compare this with our considerations on the continuous setting as noted in Remark 3.4.28. We point out that even if $\mathcal{H}_l=0$ in the case of the stationary problems considered in Subsection 3.3.5, the subproblems in the Robin method applied to them are always nonlinear due to the convex functional $\psi_1:\tilde{V}_1\to\mathbb{R}$ such that nonlinear Gauss–Seidel steps as just described always occur for nodes $p\in\mathcal{N}_j^1\cap\Gamma$.

As far as the coarse grid corrections in the monotone multigrid for the minimization of (3.4.99) are concerned, we have already emphasized that they can be carried out with an additional convex contribution on $p\in\mathcal{N}_j^1\cap\Gamma$. A different critical point (u_α instead of u_c) occurs if altered Brooks–Corey functions (from Subsection 1.4.3) are used in the nondegenerate case. For the constrained Newton linearization as in (2.7.14) we obviously have to calculate $\Psi_1''=(\kappa_1^{-1})'$. In addition, for the estimates as in (2.7.7) we also need $\Psi_1'''=(\kappa_1^{-1})''$ in case of $p\in\mathcal{N}_j^1\cap\Gamma$ and $F_\Gamma:(u_c,\infty)\to\mathbb{R}$. More concretely, we use a global threshold $\tilde{L}$ and compute $\tilde{u}_c$ as in Subsection 2.7.4 with $L=\tilde{L}/2$. Then we compute $\overline{u}_c$ by

$$\tau\gamma_1\kappa_1^{-1}(\overline{u}_c)\frac{h_{\Gamma,p_l}}{h_{p_l}}=L \tag{3.4.101}$$

(which can be done explicitly in case of the Brooks–Corey functions) and choose $u_c'=\max(\tilde{u},\overline{u}_c)$ in order to obtain an estimate as in (2.7.7) on $[u_c',\infty)$ with the global Lipschitz constant $\tilde{L}$ and with F_Γ' instead of Φ_p''.

Remark 3.4.29. As already indicated in Remark 1.4.1, in case of different bubbling pressures $p_{b,i}$ in different subdomains for $i=1,2,\ldots,n$, one only needs to take into account the different scaling factors $u_{r,i}=-p_{b,i}(\varrho g z_0)^{-1}$ in the stiffness matrix for the numerical treatment of the subproblems (compare (1.3.15) and (1.3.18)). Otherwise, nothing else needs to be changed in the latter. In addition to that, only the Kirchhoff transformations and their inverses on each subdomain need to be altered because the physical pressure should be measured not in units of $-p_{b,i}$ but in units of p_0 which is common for all subdomains. (For convenience one can set the unit $p_0 = \varrho g z_0$, i.e. as the pressure equivalent to the pressure of a water column with the height z_0.)

As a convention introduced in Section 1.3 by the definition of the adimensional parameter functions, the generalized pressures u_i on the subdomains are always given in units of $-p_{b,i}$ and so are $\kappa_i^{-1}(u_i)$. (Recall that $p_{b,i}$ always corresponds to -1 in the scaled functions in (1.3.20)–(1.3.26).) Therefore, for $i = 1, 2, \ldots, n$ in case of n subdomains, the conversion of the units can be seen by multiplying 1 in

$$\kappa_i^{-1}(u_i)\,[-p_{b,i}] \cdot \left(\frac{[p_0]}{[-p_{b,i}]} \cdot \frac{[-p_{b,i}]}{[p_0]} \right) = \kappa_i^{-1}(u_i) \frac{[-p_{b,i}]}{[p_0]}\,[p_0]\,,$$

i.e. $\kappa_i^{-1}(u_i)$ needs to be multiplied by $-p_{b,i}/p_0 = u_{r,i}$ in order to get p_i in the unit p_0 and analogously, p_i given in this unit needs to be multiplied by $p_0/(-p_{b,i}) = u_{r,i}^{-1}$ before applying κ_i to it. Note that this has to be done for the Dirichlet and the Robin transmission conditions (3.3.34) and (3.4.16) or (3.4.18), respectively, in each iteration step. In addition, this must be done for the transformation of the initial or Dirichlet boundary conditions into the generalized pressure in each subdomain as well as for the retransformation of the solutions u_i in the subdomains into the physical pressure measured in the common unit p_0.

Convergence of the Robin method for the time- and space-discretized Richards equation

As in the discrete case for the Dirichlet–Neumann method discussed in Remark 3.3.9, we also obtain a convergence result for the discrete Robin method, the proof of which shall be sketched here. Again, the considerations follow those of the continuous case and are more complicated than for the Dirichlet–Neumann method. First, we discretize the transmission problem (3.4.12)–(3.4.14) analogously as in Remark 3.3.9 with the discretization

$$\sum_{p \in \mathcal{N}_j^i} M_i(u_{i,j}(p))\, v(p)\, h_p\,, \quad u_{i,j}, v \in \mathcal{S}_j^i$$

of $(M_i(u_i), v_i)_{\Omega_i}$ via the lumped L^2-scalar product and use the (corresponding) discretization of the Robin steps (3.4.20)–(3.4.21) and (3.4.22)–(3.4.23) as described above. Accordingly, the nonlinear discrete Steklov–Poincaré operators $S_{i,j} : \Lambda^j \to (\Lambda^j)'$ are given by

$$\begin{aligned}\langle S_{i,j}\lambda, \mu\rangle := \sum_{p \in \mathcal{N}_j^i} & M_i((u_{i,j}(\lambda) + u_{i,j}^*)(p))\,(R_{i,j}\mu)(p)\,h_p \\ & + a_i(u_{i,j}(\lambda) + u_{i,j}^*, R_{i,j}\mu) - (f, R_{i,j}\mu)_{\Omega_i} \qquad \forall \mu \in \Lambda^j\end{aligned}$$

for $i = 1, 2,\ j \geq 0$, with the notation as in Remark 3.3.9. Then Propositions 3.4.6, 3.4.7 and 3.4.8 can be established in the discrete case, too. Since the altered discrete Robin method is equal to our discretized Robin method in one space dimension (we do not have a general suitable discretization of the altered Robin method), Theorem 3.4.12 can be carried over to the discrete case in 1D. Our basic Lemma 3.4.13 also holds on a discrete level. For the proof of the discrete versions of Propositions 3.4.14 and 3.4.15 (in arbitrary dimensions) we can use j-independent H^1-estimates for lumped L^2-scalar products on Ω_i, $i = 1, 2$, which can be found e.g. in Elliott et al. [35] or in Blowey and Elliott [17]. Proposition 3.4.16 and Corollary 3.4.18 can be proved analogously in the discrete setting.

We need to take a closer look at the proof of Lemma 3.4.21, where we used "continuous arguments" based Lemma 3.4.20 and on the strong formulation of (3.4.35)–(3.4.37). The latter does not have a straightforward discrete analogue in a strong sense so that, for example, (3.4.86) is false on the discrete level. However, we can also establish the results in Lemma 3.4.21 here by translating the ideas used there into "discrete arguments". Note that the discretization of (3.4.35) with $\lambda \in \Lambda^j$ reads

$$\sum_{p \in \mathcal{N}_j^i} M_i(u_{i,j}(\lambda)(p))\, v(p)\, h_p + \int_{\Omega_i} u_{i,j}(\lambda)' v'\, dx \quad \forall v \in \mathcal{S}_j^i\,,\ \ v_{|\Gamma} = 0\,. \tag{3.4.102}$$

Let $a := p_0 < p_1 < \cdots < p_n =: b$, $n \in \mathbb{N}$, be the nodes in $\mathcal{N}_j^1 \cap [a, b]$. In order to prove

$$u_{i,j}(\eta)(p) \geq u_{i,j}(\mu)(p) \quad \forall p \in \mathcal{N}_j^i \tag{3.4.103}$$

with $\eta \geq \mu$ for $i = 1$, we carry out an induction in which one step looks as follows. We test (3.4.102) with $v \in \mathcal{S}_j^1$ satisfying $v(p) = 1$ for all nodes $p < p_k$ and vanishing on all others, once the inequality $u_{i,j}(\eta)(p) \geq u_{i,j}(\mu)(p)$ for all $p < p_k$ and k with $1 \leq k \leq n$ is known. Consequently, since

$$\sum_{p_0 \leq p \leq p_k} \Big(M_1(u_{1,j}(\eta)(p)) - M_1(u_{1,j}(\mu)(p)) \Big) v(p)\, h_p \geq 0$$

holds, we obtain

$$\int_a^{p_1} (u_{1,j}(\eta) - u_{1,j}(\mu))' v'\, dx + \int_{p_{k-1}}^{p_k} (u_{1,j}(\eta) - u_{1,j}(\mu))' v'\, dx \leq 0$$

from (3.4.102) and therefore $u_{i,j}(\eta)(p_k) \geq u_{i,j}(\mu)(p_k)$. For the proof of

$$u_{1,j}(\eta)' \geq u_{1,j}(\mu)' \quad \text{on } (p_{k-1}, p_k)\,,\ \ 1 \leq k \leq n\,, \tag{3.4.104}$$

we can also use induction. Testing (3.4.102) by λ_p, $p \in \mathcal{N}_j^i$, $\lambda_p(b) = 0$, gives

$$\Big(M_i(u_{i,j}(\eta)(p)) - M_i(u_{i,j}(\mu)(p)) \Big) v(p)\, h_p = \Delta_j \Big(u_{i,j}(\eta) - u_{i,j}(\mu) \Big)(p) \geq 0$$

for a discrete Laplacian Δ_j which provides the same discretization as central differences for the second derivatives. With this inequality one can easily deduce (3.4.104) by induction for $k = 1, \ldots, n$ while using $u_{i,j}(\eta)(a) = u_{i,j}(\mu)(a) = 0$ and (3.4.103). The proofs of (3.4.103) and (3.4.104) with $\geq$ replaced by $\leq$

for $i = 2$ are analogous for the second subdomain. In addition, they work for nonconstant porosity $n(\cdot)$ and hydraulic conductivity $K_h(\cdot)$ if we use a quadrature by replacing h_p in (3.4.102) by $n(p)h_p$ and approximate K_h between two successive nodes by a constant value (compare Remark 3.4.24).

Now, Proposition 3.4.22 can be carried over to the discrete case with the same constants which are independent of $j \geq 0$. Altogether, if we collect all the results we can prove the following discrete analogue of Theorem 3.4.23 as in the continuous setting.

Theorem 3.4.30. *Assume that the conditions as in Theorem 3.4.23 are satisfied and the domain decomposition problem (3.4.12)–(3.4.14) as well as the Robin method (3.4.20)–(3.4.23) are discretized as described above.*

Then this discretization of the Robin method provides sequences $(u_{1,j}^k)_{k\geq 1}$ and $(u_{2,j}^k)_{k\geq 0}$ of finite element iterates converging in $\mathcal{S}_j^1$ and in $\mathcal{S}_j^2$, respectively, to the unique solution of the discretized domain decomposition problem (3.4.12)–(3.4.14) which, moreover, is well-posed with respect to the data $f \in L^2([a,c])$.

The lack of a convergence result $u_{i,j} \to u_i$, $i = 1, 2$, for $j \to \infty$ and the basic problems around the connection to the untransformed discretized problem (3.4.9)–(3.4.11) have already been addressed in Remark 3.3.9.

3.4.6 Numerical tests in 2D: Robin method vs. Dirichlet–Neumann method and applied to the Richards equation without gravity

In this subsection we present some numerical results on the performance of the Robin method that we obtained for some model problems related to the Richards equation in two space dimensions. As in Subsection 3.3.5 for the Dirichlet–Neumann method, it turns out that the algorithm can also be applied successfully in cases for which we do not have a proof of its convergence.

In the first part of this subsection we test the Robin method in the same situation as done for the Dirichlet–Neumann method in Subsection 3.3.5. This leads to a comparison of these two methods in the Yin Yang case considered there which we extend by applying these iteration techniques in the same case with nonlinearities that differ from each other even more than in Subsection 3.3.5. Although both procedures can be applied successfully to these domain decomposition problems, it turns out that the Dirichlet–Neumann method shows better convergence results in these cases than the Robin method. Nevertheless, due to our analytical results from this section, we go on to test the Robin method in cases such as (3.4.2)–(3.4.4) involving two nonlinearities $\theta_i, k_i : \mathbb{R} \to \mathbb{R}$ for $i = 1, 2$ in the subproblems. Therefore, in the second part of the subsection we present numerical results obtained by the application of the Robin method to the Richards equation without gravity in a heterogeneous setting with two different soil types. The performance of the method for this two-dimensional time-dependent case is quite satisfying.

Comparison of Robin method and Dirichlet–Neumann method

With the following numerical results the performance of the Robin method and the Dirichlet–Neumann method when applied to the model problem in Subsection 3.3.5 can be compared. In the first example concerning the Robin method we use the same data, parameter functions and meshes as in that subsection. The discretization by piecewise linear finite elements is carried out as discussed in Subsection 3.4.5. Recall that in contrast to the Dirichlet–Neumann method in Section 3.3 the resulting problems on the subdomains are nonlinear for the Robin method in general. They are treated by a monotone multigrid solver which we also described in Subsection 3.4.5. More concretely, we use truncated monotone multigrid with a $V(3,3)$-cycle, i.e. containing 3 presmoothing and 3 postsmoothing steps as explained on page 123. The threshold in (3.4.101) is chosen as $L = 10^8$. As a stopping criterion for this local solver we require the relative error to satisfy

$$\frac{|u_i^j - u_i^{j-1}|_{1,\Omega_i}}{|u_i^{j-1}|_{1,\Omega_i}} \leq 10^{-12}\,, \quad i = 1,2\,, \tag{3.4.105}$$

for the last multigrid iterate u_i^j with $j \geq 1$ where $|\cdot|_{1,\Omega_i}$ is the energy norm on Ω_i induced by the bilinear form $a_i(\cdot,\cdot)$.

The convergence rates ρ for the Robin method in Figures 3.10 and 3.12 have been obtained in the same way as for the Dirichlet–Neumann method in Figures 3.7 and 3.9 (see (3.3.59)–(3.3.60)). Here, the Robin parameter γ needs to be chosen on another scale than the damping parameter ϑ for the Dirichlet–Neumann method. But as in the latter case there also seems to be an opti-

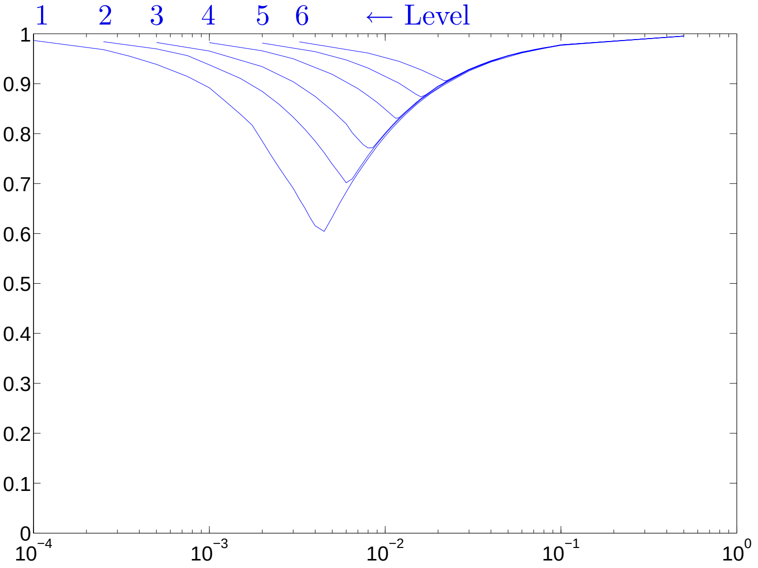


Figure 3.10: ρ vs. Robin parameter γ on levels 1 to 6
Robin method for the case in (3.3.58)

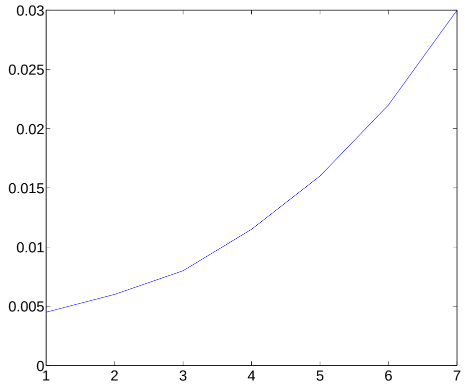


Figure 3.11: γ_{opt} vs. refinement level Robin method for (3.3.58)

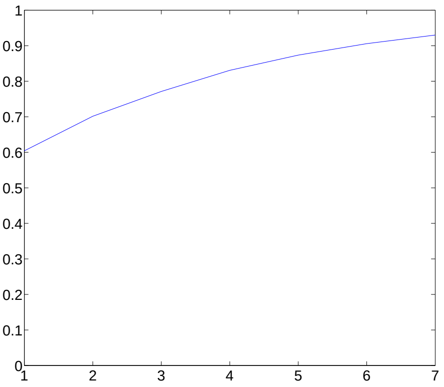


Figure 3.12: ρ_{opt} vs. refinement level Robin method for (3.3.58)

mal choice of the parameter which depends on the refinement level, and there seems to be an interval inside of which the parameter has to be chosen in order to guarantee reasonable convergence rates, see Figure 3.10. To achieve this we need larger parameters γ on higher levels which corresponds to a bigger damping needed for the Dirichlet–Neumann method on higher levels (compare Figure 3.7). Unfortunately, however, in contrast to the Dirichlet–Neumann procedure this effect does not seem to stabilize on higher levels for the Robin method (note the logarithmic scale in Figure 3.10). Figures 3.11 and 3.12 show the behaviour of the optimal Robin parameter γ_{opt} and the optimal convergence rate $\rho_{opt} = \rho(\gamma_{opt})$ on levels 1 to 7.

We remark that apart from the size of the hydraulic conductivity K_h in (3.3.57), which occurs as a factor in front of the Laplacians in (3.4.15) and (3.4.17) and therefore in front of the normal derivatives in (3.4.16) and (3.4.18), respectively, we do not have further a priori indicators for the order of magnitude of the optimal Robin parameters. (Concerning the linear theory on this topic we refer to Wachspress [93] and Discacciati [33, pp. 102–105].)

The model problem in Subsection 3.3.5 that we considered so far is heterogeneous in the sense that different pore size distribution factors (3.3.58) are given on Ω_1 and Ω_2. However, the bubbling pressure $p_b = -1$ corresponding to the discontinuity in the derivative of the nonlinearities $k_i(p_i)$ in (3.3.57) and the hydraulic conductivities K_h as an additional factor in front of k_i are the same in both subdomains. In the following, we extend this model problem such that it also contains different bubbling pressures $p_{b,i}$ (see (1.2.9), (1.2.11) and Section 1.3 for details) and hydraulic conductivities $K_{h,i}$ in Ω_i for $i = 1, 2$. More concretely, we consider the case of sand in Ω_1 and clay in Ω_2 given in Table 3.1 which we shall call *strongly heterogeneous case*.

As in Subsection 3.3.5 we choose one meter of a water column as the pressure unit for $p_{b,i}$ and $K_{h,i}$ are given in $[m/s]$. The data in Table 3.1 are hydrologically

Ω_i	λ_i	$p_{b,i}$	$K_{h,i}$
$i = 1$ (sand)	0.694	-0.073	$6.54 \cdot 10^{-5}$
$i = 2$ (clay)	0.165	-0.373	$1.67 \cdot 10^{-7}$

Table 3.1: Strongly heterogeneous case

quite extreme since they are covering the whole range of possible soil parameters according to the USDA (United States Department of Agriculture) soil texture triangle, see Rawls et al. [77, Tables 5.3.2 and 5.5.5]. In order to distinguish this strongly heterogeneous case from the heterogeneity (3.3.58) in the first model problem above we call the latter *mildly heterogeneous case* form now on.

Note that $K_{h,1}$ and $K_{h,2}$ in Table 3.1 are smaller than $K_h = 0.002$ chosen in the previous example and, in addition, differ from each other by two orders of magnitude. Therefore, we also alter the right hand side f (as in (3.3.56) given in the unit $[1/s]$) in order to obtain a solution which has approximately the same range as the one from the previous example in Figure 3.6. We define f as

$$f(x) = \begin{cases} -2.5 \cdot 10^{-3} & \text{on } B_1 \\ 5 \cdot 10^{-5} & \text{on } B_2 \\ 0 & \text{elsewhere.} \end{cases} \tag{3.4.106}$$

Moreover, we choose a smaller ellipticity constant $c = 0.01$ for the nonlinearity in (3.3.57) than in the mildly heterogeneous case.

For the numerical treatment of different bubbling pressures in different subdomains see Remarks 1.4.1 and 3.4.29. Other than that, the solvers are the same as used in the mildly heterogeneous case. Figure 3.13 shows the numerical solution p on Ω of the domain decomposition problem (3.3.11)–(3.3.13) in the strongly heterogeneous case of Table 3.1, calculated on about 235,000 nodes as in Figure 3.6 for the case (3.3.58). Here, the range of p is $[-56.1, 0.0]$ in Ω_1 and $[-36.2, 3.0]$ in Ω_2 and thus resembles the range of the solution in the previous example.

Again, but now more obvious than in Figure 3.6 (where this has been indicated by a black line), the plot contains a "crater" corresponding to a circular area in Ω in which the soil is not fully saturated (i.e. where $p_i < p_{b,i}$ holds for $i = 1, 2$). The boundary Σ of this area, which is the free boundary separating the unsaturated from the saturated regime, is given by $p_i = p_{b,i}$, $i = 1, 2$, in Ω_i. In particular, even though this can only be guessed from Figure 3.13, this free boundary has a nontrivial two-dimensional intersection with the interface $\Gamma = \overline{\Omega}_1 \cup \overline{\Omega}_2$ because we now have $p_{b,1} \neq p_{b,2}$. More concretely, the same pressure may result in a maximal saturation in Ω_2 (which contains the source and the smaller bubbling pressure) but not in Ω_1. Still, as can be seen along $\Sigma \cup \Omega_2$ and despite the strong effect of the Kirchhoff transformation across Σ,

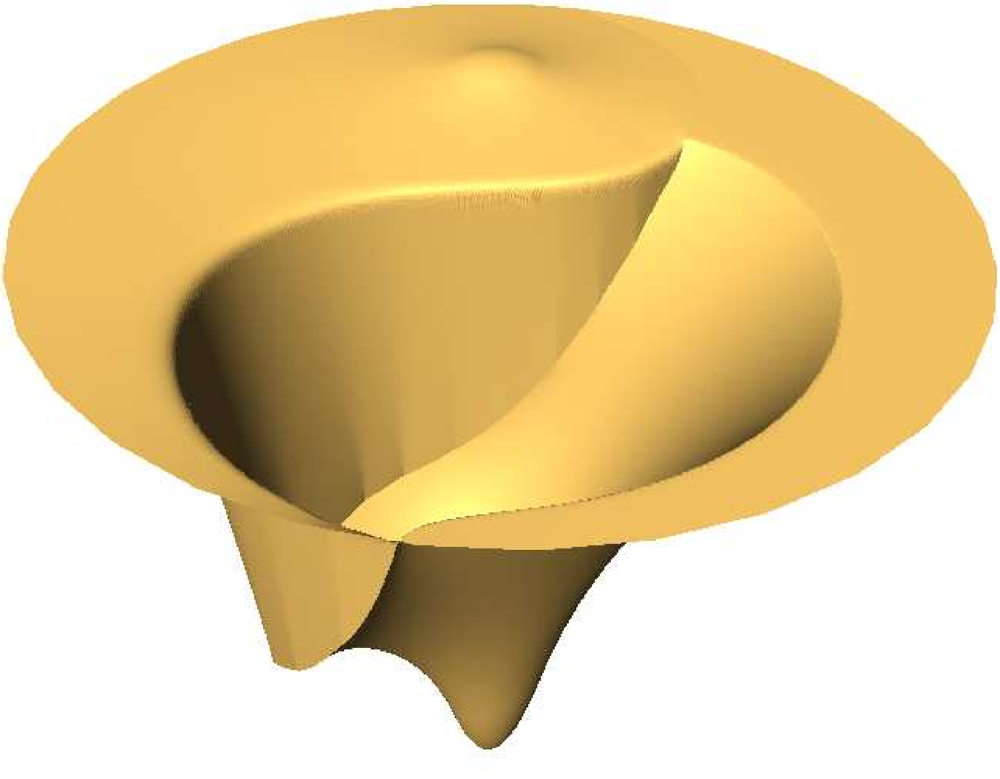


Figure 3.13: Solution p on Ω in the strongly heterogeneous case

the solution $p_i \in H^1(\Omega_i)$ is smooth across $\Sigma \cup \Omega_i$ for $i = 1, 2$. However, it is obvious that p is not smooth across Γ which reflects that (3.3.13) rather enforces the continuity of the normal fluxes across the interface such that we have

$$\frac{\partial p_1}{\partial \mathbf{n}} \neq \frac{\partial p_2}{\partial \mathbf{n}} \quad \text{on } \Gamma \text{ in general}\,.$$

This contrasts the situation in the saturated region in Figure 3.6 where (3.3.13) also entails the continuity of the normal derivatives of p across Γ because of the constant bubbling pressure and hydraulic conductivity on Ω.

Despite the strong impact of the heterogeneities in Table 3.1 on the solution in Figure 3.13 we do not obtain worse convergence results for our domain decomposition methods in this model problem compared to the mildly heterogeneous case. On the contrary, the Dirichlet–Neumann method applied to the strongly heterogeneous case shows an unexpected and surprisingly good behaviour. Figures 3.14–3.16 and Figures 3.17–3.19 were obtained for the Dirichlet–Neumann and the Robin iteration, respectively, in the same way as described above for the mildly heterogeneous case.

As can be seen in Figure 3.14, the Dirichlet–Neumann method provides extremely good optimal convergence rates ($\rho_{opt} = 0.012$) on the first two levels with almost no damping ($\vartheta_{opt} = 0.9875$). Moreover, we can allow a considerable overrelaxation (i.e. $\vartheta > 1$) and still obtain convergence. The situation on the third level is worse but still much better than in the mildly heterogeneous case in Figure 3.7. Surprisingly, the convergence results improve again on level 4, and even on levels 5 and 6 they are better than on the third level. As in the previous example, but now with considerably better values, the optimal damp-

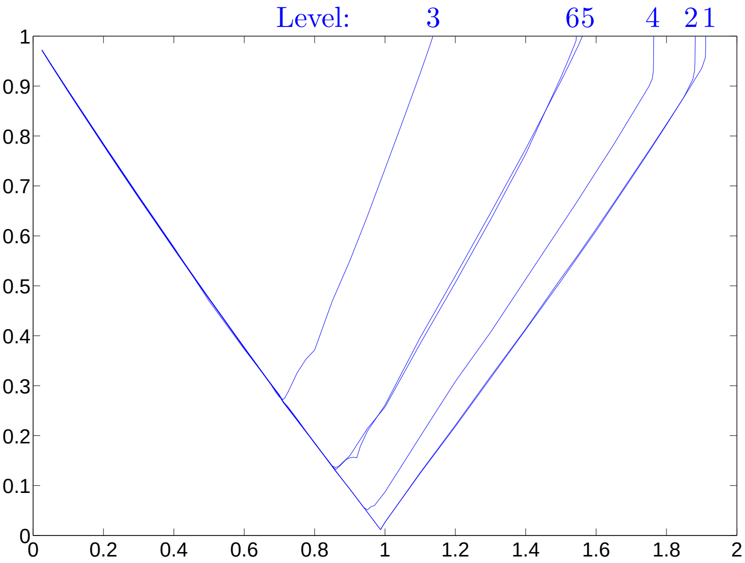


Figure 3.14: ρ vs. damping parameter ϑ on levels 1 to 6
Dirichlet–Neumann method for the case in Table 3.1

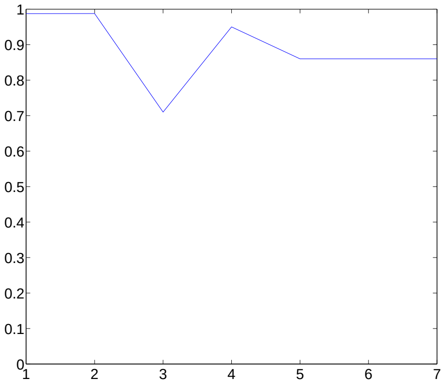


Figure 3.15: ϑ_{opt} vs. refinement level Dirichlet–Neumann for Table 3.1

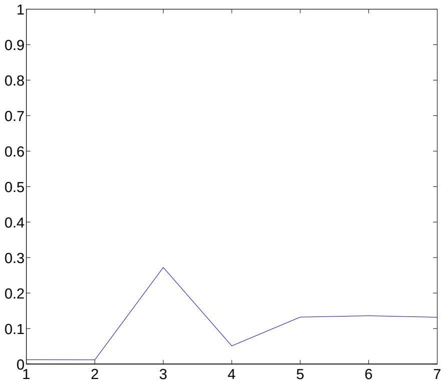


Figure 3.16: ρ_{opt} vs. refinement level Dirichlet–Neumann for Table 3.1

ing parameter ($\vartheta_{opt} = 0.86$) and the optimal convergence rates ($\rho_{opt} = 0.136$ on the 6th and $\rho_{opt} = 0.132$ on the 5th and the 7th level) stabilize on higher levels, see Figures 3.15 and 3.16.

Figures 3.17–3.19 show that the Robin method applied to the new model problem behaves similarly as in the mildly heterogeneous case. In contrast to the latter, the Robin parameter γ has to be chosen on a smaller scale now which reflects the smaller hydraulic conductivities used in Table 3.1. (As indicated above, the hydraulic conductivities $K_{h,i}$, $i = 1, 2$, are factors in front of the Laplacians in (3.4.15) and (3.4.17), which means they also appear as factors

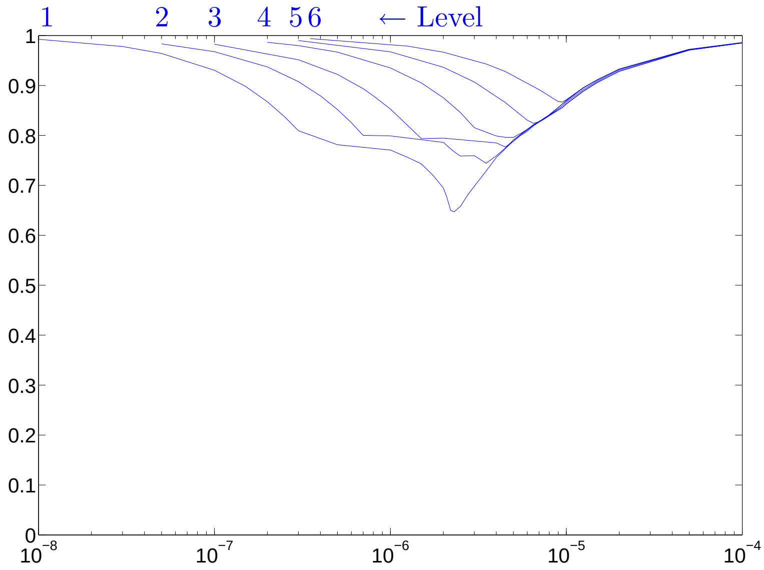


Figure 3.17: ρ vs. Robin parameter γ on levels 1 to 6 Robin method for the case in Table 3.1

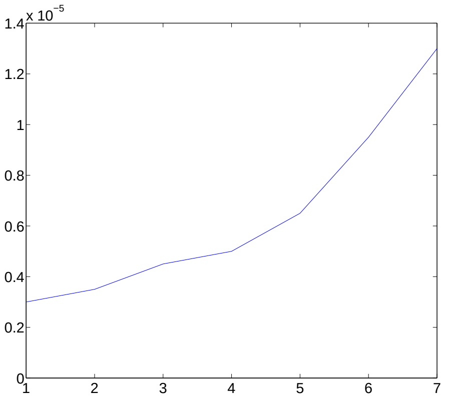


Figure 3.18: γ_{opt} vs. refinement level Robin method for Table 3.1

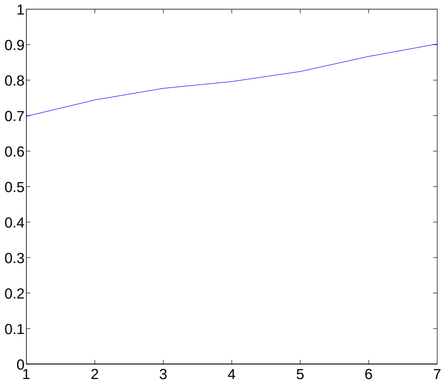


Figure 3.19: ρ_{opt} vs. refinement level Robin method for Table 3.1

in front of the normal derivatives in (3.4.16) and (3.4.18), respectively.) As already seen in Figure 3.10, level-independence of the convergence rates occurs if and only if the Robin parameter is bigger than the optimal parameter γ_{opt} related the the finest level considered. In addition, as in the mildly heterogeneous case, we also observe in Figures 3.18 and 3.19 that γ_{opt} and ρ_{opt} do not seem to stabilize on higher levels. If one regards the continuous situation as the limit of the discrete settings, these results could provoke the hypothesis that our 1D-convergence result on the Robin method in Theorem 3.4.23 cannot be generalized to higher dimensions.

With regard to the original problem (3.3.11)–(3.3.13) related to the Richards equation, one might be interested in the convergence rates ρ_p for the domain decomposition methods measured in the physical variables p_i, $i = 1, 2$, i.e. after the application of the inverse of the Kirchhoff transformation. Recall that in general this retransformation is ill-conditioned for physical pressures corresponding to the unsaturated regime or at least having a similar effect for small ellipticity constants $c > 0$ in (3.3.57) (see Subsections 1.4.1 and 1.4.3). However, the values $c = 0.1$ and $c = 0.01$ used in our two examples do not seem to be small in this sense. If we calculate the convergence rates ρ_p in the same way as the convergence rates ρ, i.e. by replacing u by p in (3.3.59) and (3.3.60), it turns out that in neither case nor method the convergence rates ρ and ρ_p seem to differ more than by a few percentage points. The optimal parameters can also vary a bit, but qualitatively the same results as above are obtained for the convergence rates ρ_p. In general the situation is different for $c = 0$, see Remark 3.4.31.

As already indicated in Subsection 3.3.5, we do not obtain convergence of our methods if we do not have uniform ellipticity, i.e. for $c = 0$. We even observe numerical instabilities in this case which occur if small generalized pressure values in iteration histories happen to be too close to the critical generalized pressure u_c corresponding to the physical pressure $p = -\infty$. Furthermore, the smaller $c > 0$ is chosen the worse the convergence rates become — again with the exception of the Dirichlet–Neumann method applied to the strongly heterogeneous case for which the convergence rates ρ (measured in u) even seem to remain stable in case of a very small ellipticity constants such as $c = 10^{-100}$. In contrast to these observations we can always choose $c = 0$ in our time-dependent cases which are still to come.

Despite the very good convergence behaviour of the Dirichlet–Neumann method in the strongly heterogeneous case we apply the Robin method to the time-dependent numerical examples concerning the Richards equation in the rest of this subsection and in Chapter 4. This decision is mainly motivated by the fact that, in contrast to the Dirichlet–Neumann method (Proposition 3.3.8), our convergence results for the Robin method in Theorem 3.4.23 include spatial problems arising from a time-discretization of the Richards equation at least in the uniformly elliptic case $c > 0$. Our aim is to test how far our 1D-convergence theory of this section can be numerically maintained in two space dimensions.

Robin method applied to the Richards equation without gravity

In the following, we present a numerical example which we obtained for the Robin method applied to the Richards equation without gravity in a heterogeneous setting with two different soil types in two subdomains. Our analytical and numerical approaches to such a problem have been provided in Sections 3.4.1 and 3.4.5. In contrast to the examples above, the spatial problems which we encounter now are transmission problems of the kind (3.4.2)–(3.4.4) or (3.4.6)–(3.4.8) which contain an additional nonlinearity related to the saturation.

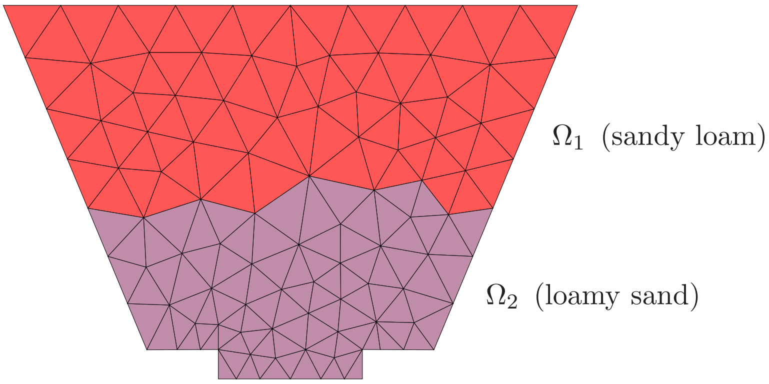


Figure 3.20: Coffee filter like domain Ω

Concretely, we consider the domain $\Omega \subset \mathbb{R}^2$ depicted in Figure 3.20 which resembles a coffee filter. We assume that the top subdomain Ω_1 is filled with sandy loam while the bottom subdomain Ω_2 contains loamy sand. As before in the strongly heterogeneous case in Table 3.1, the corresponding hydrological data are chosen according to the USDA soil texture triangle, see Rawls et al. [77, Tables 5.3.2 and 5.5.5]. Since now the saturation occurs explicitly in the equations (3.4.2) or (3.4.6) we also need to specify the residual water contents $\theta_{m,i}$ and the porosity-values n_i for $i = 1, 2$ (compare the original equation (1.2.7) and (1.2.9)). The maximal water contents $\theta_{M,i}$ are constant with $\theta_{M,i} = 1$ for $i = 1, 2$ according to Rawls et al. [77, Table 5.1.1]. However, the variation of these values is neither significant nor does it effect the performance of the Robin method in a way as done by the other soil parameters λ_i, $p_{b,i}$ and $K_{h,i}$, $i = 1, 2$, which influence the spatial derivative in (3.4.2). Altogether, the heterogeneities for our coffee filter example are given in Table 3.2. Here, we use the unaltered Brooks–Corey functions $p \mapsto \theta(p)$ and $\theta \mapsto kr(\theta)$ according to Burdine given in Sections 1.2 and 1.3 (as opposed to the ones in Subsection 1.4.3) which corresponds to the situation of (3.3.57) with $c = 0$, i.e. to the degenerate Richards equation.

Ω_i	n_i	$\theta_{m,i}$	λ_i	$p_{b,i}$	$K_{h,i}$
$i = 1$ (sandy loam)	0.453	0.091	0.378	-0.147	$6.06 \cdot 10^{-6}$
$i = 2$ (loamy sand)	0.437	0.080	0.553	-0.087	$1.66 \cdot 10^{-5}$

Table 3.2: Soil parameters for the coffee filter example

The domain Ω is situated in the quadrilateral $[-1, 1] \times [-0.74, 0.56]$ and the top boundary is $[-1, 1] \times \{-0.74\}$ (the z-axis is directed downward although we do not yet deal with gravity here). We start with a practically dry soil as the initial condition given by $p_0 = -20$ on Ω except for $p_0 = 100$ on the subset $[-0.21, 0.21] \times \{-0.74\}$ of the top boundary, see Figure 3.21. The latter is treated as a Dirichlet boundary γ_D with constant data $p_D = 100$ for all time steps (as before the pressure unit is one meter of a water column). Apart from the bottom boundary γ_S situated on $[-0.25, 0.25] \times \{0.56\}$, which is chosen as a Signorini-type boundary where outflow is possible, we assume homogeneous Neumann boundary conditions $\mathbf{v} \cdot \mathbf{n} = 0$ on $\partial\Omega \backslash (\gamma_D \cup \gamma_S)$. This situation results in an evolution process with an increasing saturation due to flow of water into Ω with possible outflow across γ_S until Ω is fully saturated and a stationary solution is obtained.

We treat the problem as described in Subsection 3.4.5 using an implicit time discretization (since there is no gravity) with the constant time step size $\tau = 1\,[s]$ and a space discretization with linear finite elements. The discrete Robin problems for the Richards equation in each time step result in convex minimization problems which are solved by monotone multigrid with a $V(3, 3)$-cycle (see page 123). We use 4 levels of a grid hierarchy with 112 nodes on the coarse grid in Figure 3.20 and about 5500 nodes on the finest grid with a mesh size of $h = (10 \cdot 2^4)^{-1} = 1/160$ obtained by uniform refinement. Moreover, a constant Robin parameter $\gamma = 3 \cdot 10^{-4}$ suggested by numerical experiments is chosen.

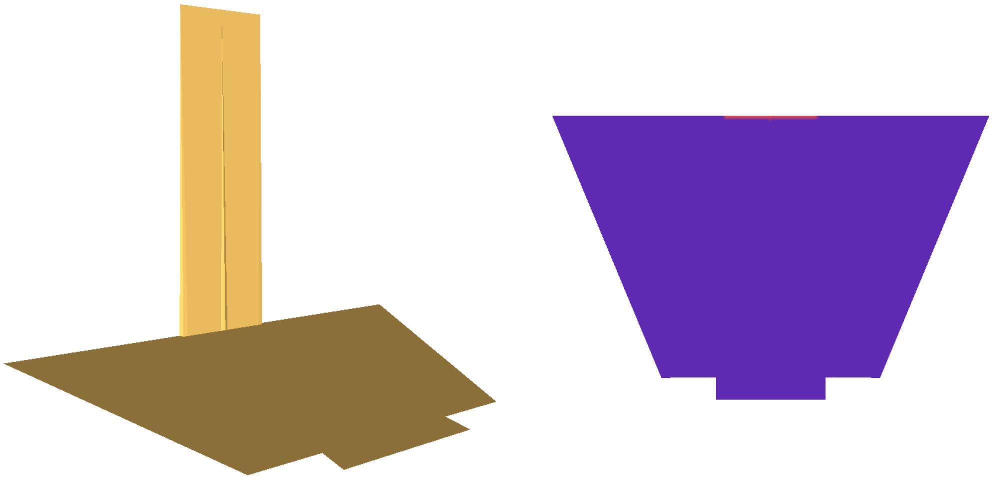


Figure 3.21: $t = 0$

In Figures 3.22–3.33 one can see the evolution of the physical pressure at equidistant time steps (except for the last one) in heightplots on the left and colourplots on the right. One can clearly detect the wetting front (cf. [13, p. 303]), where a pressure difference of almost $\Delta p = 20$ occurs, moving from the top to the bottom. More concretely, the wetting front marks the free boundary which separates the unsaturated from the fully saturated regime and, thus, around which we encounter the pressure difference between the initial condition p_0 and the bubbling pressure $p_{b,i}$ on Ω_i for $i = 1, 2$. We need 684 time steps until the stationary situation with a fully saturated Ω is reached. At about $t = 133$ the wetting front reaches the interface and starting with $t = 473$ the top subdomain Ω_1 is fully saturated. The range of p is between -20 and 100 until shortly before the last time step and in the stationary case it is in the interval [26.3, 100] on Ω_1 and [0.0, 32.2] on Ω_2. As in the solution of the strongly heterogeneous Yin Yang case in Figure 3.13, one can see in the heightplots that the physical pressure is nonsmooth across the interface, at least in the saturated regime.

The stopping criteria for the multigrid solver and the Robin method are given according to (3.4.105) and (3.3.59), respectively, at each time step, now of course with the initial condition given by the previous time step. Figures 3.34 and 3.35 show averaged (as well as maximal) multigrid convergence rates $\rho_{m,1}$ and $\rho_{m,2}$ (as well as $\rho_{M,1}$ and $\rho_{M,2}$) per time step for Ω_1 and Ω_2, respectively, which are determined in the following way. For each domain decomposition step $l \in \mathbb{N}$ in a fixed Ω_i, $i = 1, 2$, the geometric mean

$$\bar{\rho}_{l,j} = \left(\prod_{k=2}^{j} \rho_k \right)^{1/j}$$

of the approximated rates

$$\rho_k = \frac{|u_i^k - u_i^{k-1}|_{1,\Omega_i}}{|u_i^{k-1} - u_i^{k-2}|_{1,\Omega_i}}\,, \qquad i = 1, 2\,,$$

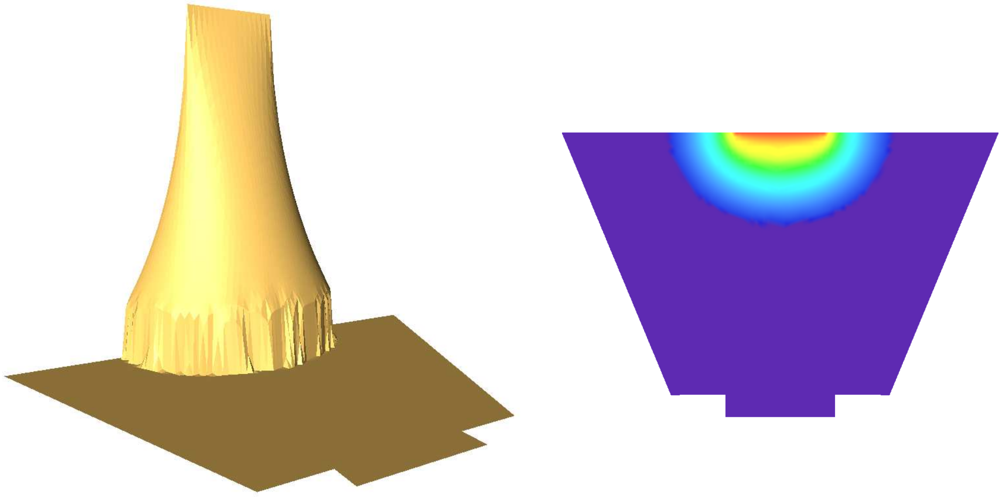


Figure 3.22: $t = 60$

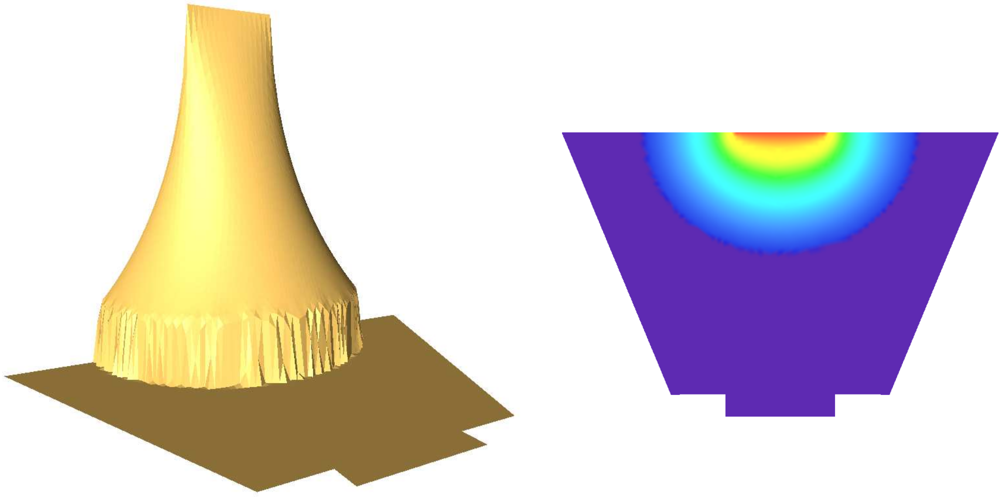


Figure 3.23: $t = 120$

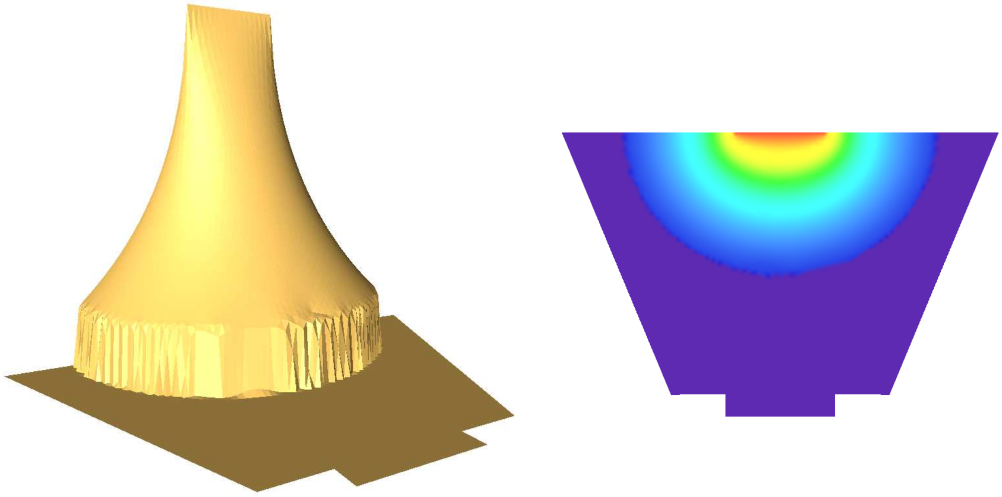


Figure 3.24: $t = 180$

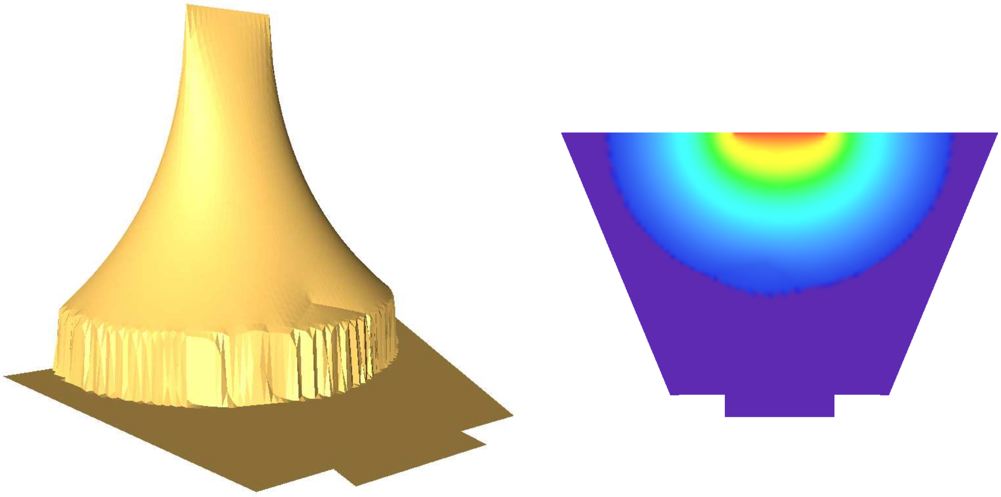


Figure 3.25: $t = 240$

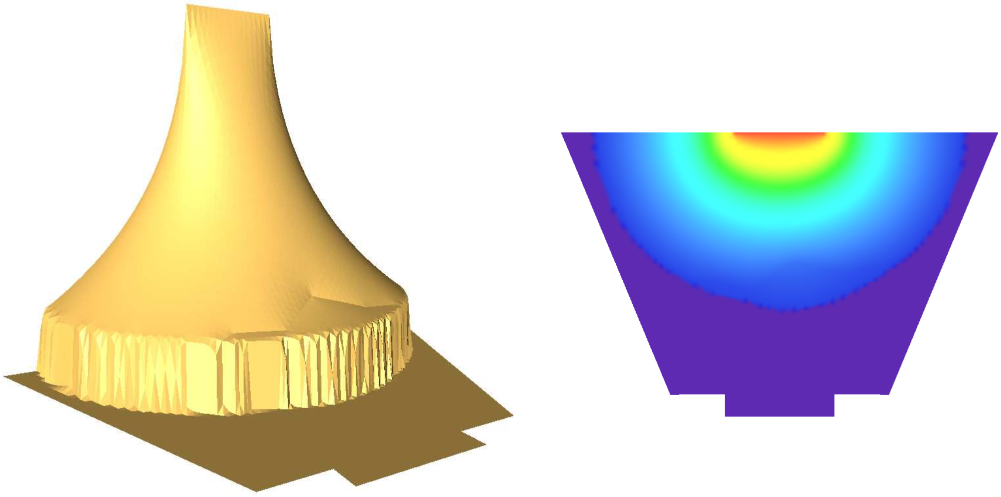


Figure 3.26: $t = 300$

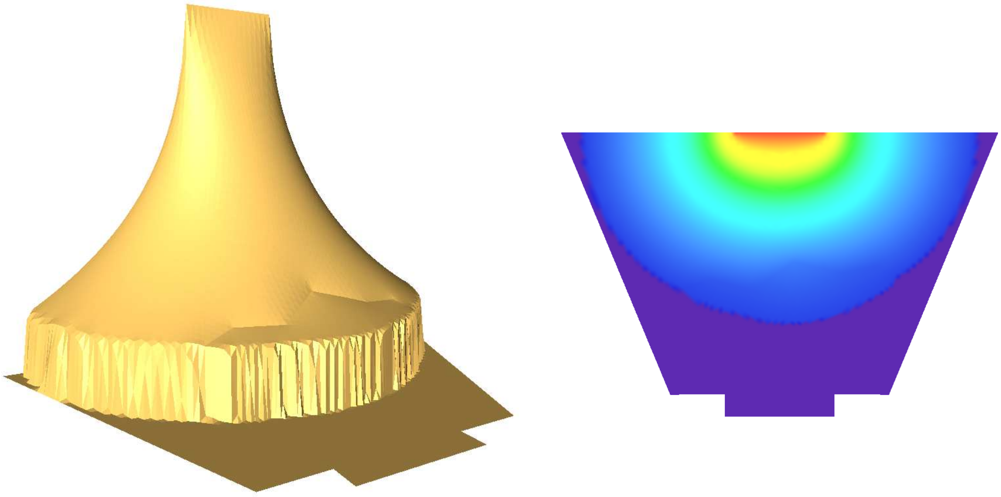


Figure 3.27: $t = 360$

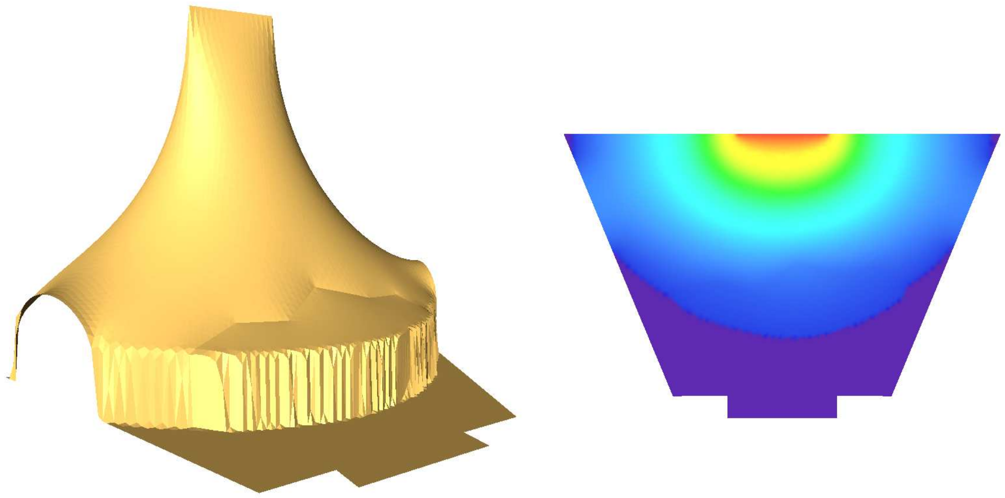


Figure 3.28: $t = 420$

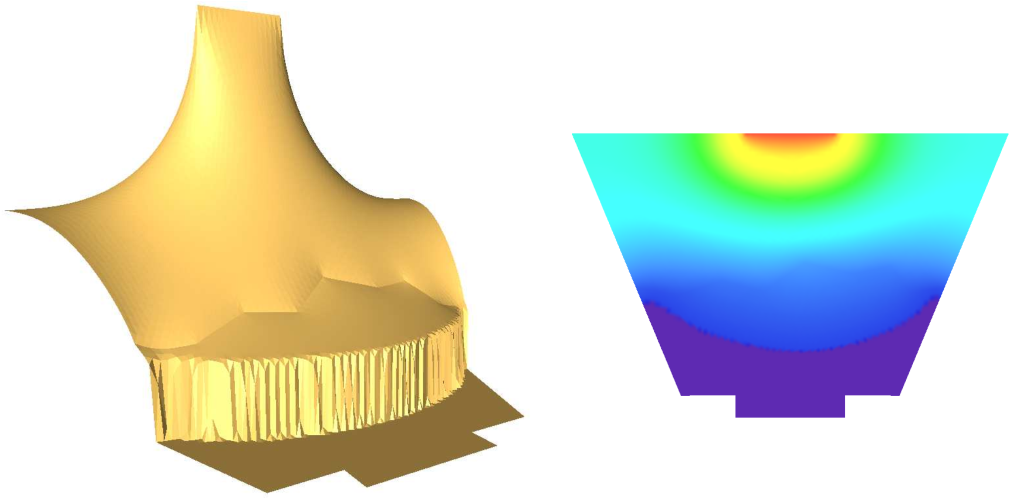


Figure 3.29: $t = 480$

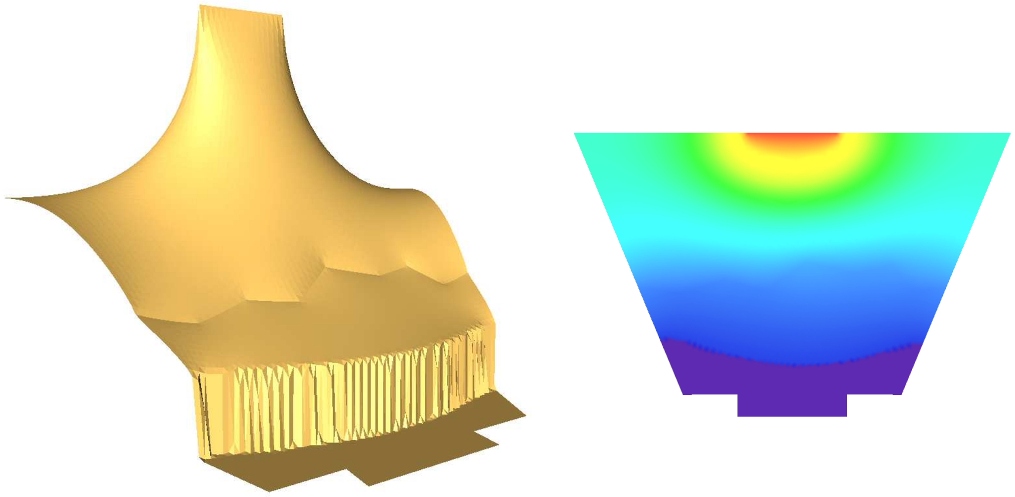


Figure 3.30: $t = 540$

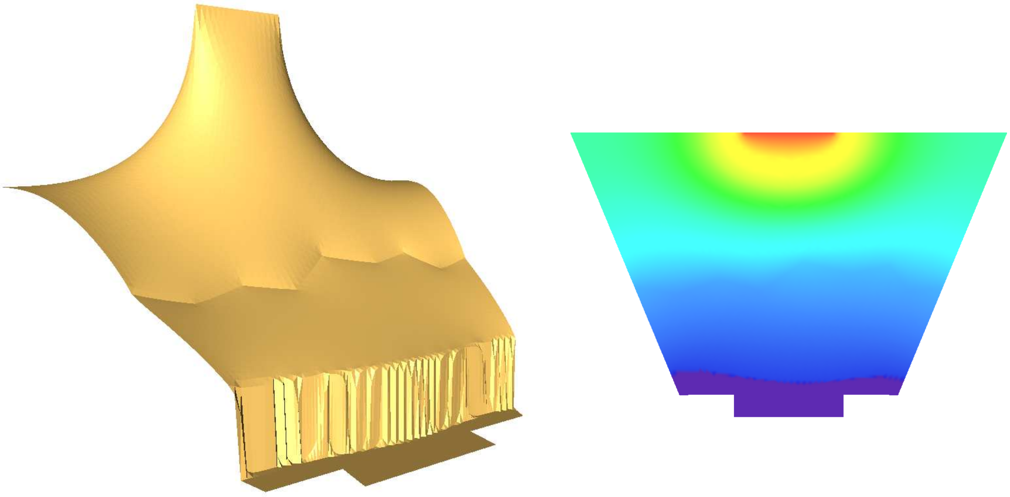


Figure 3.31: $t = 600$

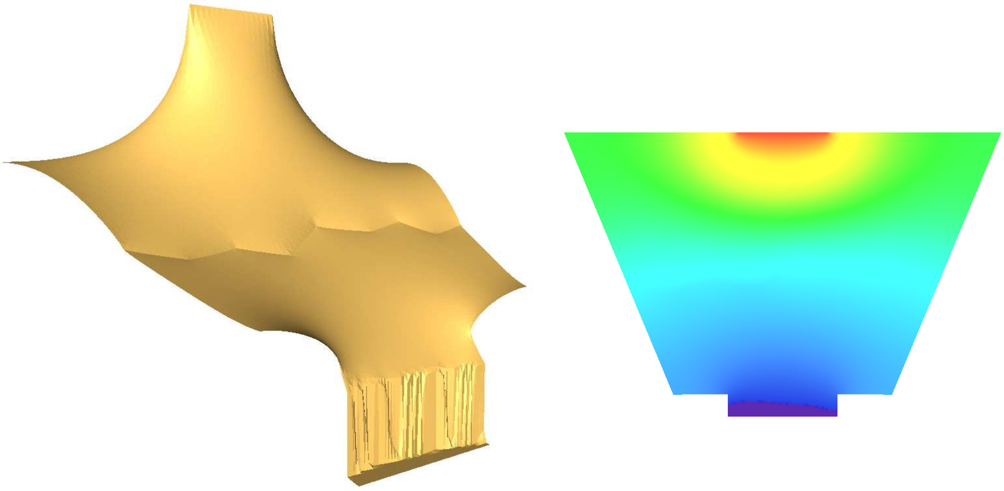


Figure 3.32: $t = 660$

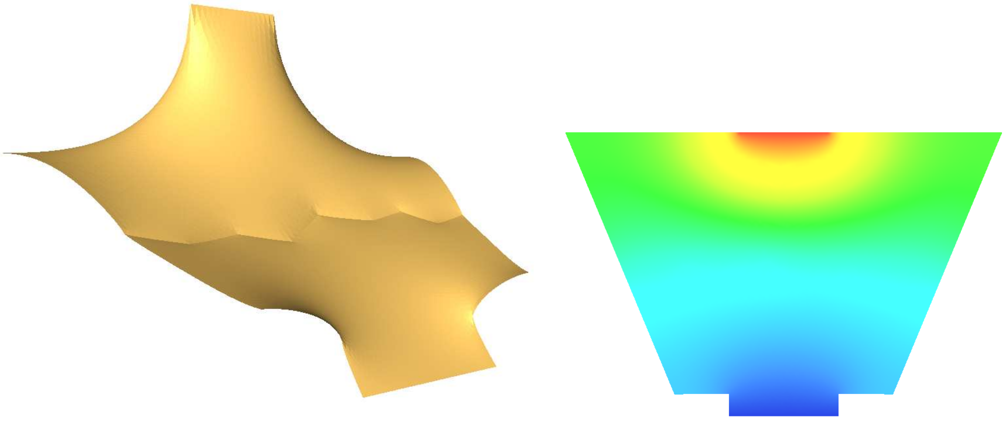


Figure 3.33: $t = 684$

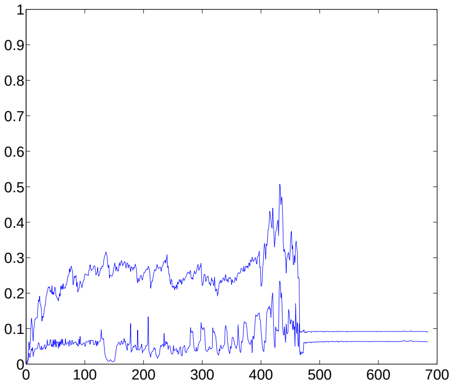


Figure 3.34: Multigrid convergence rates per time step in Ω_1 (top: maximal, bottom: averaged)

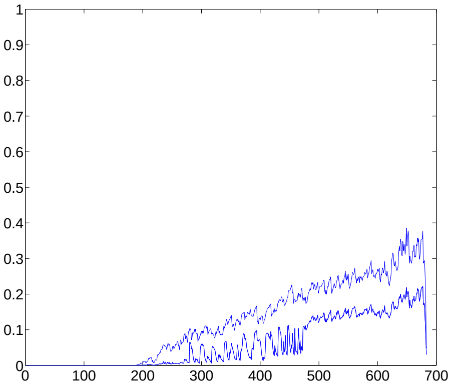


Figure 3.35: Multigrid convergence rates per time step in Ω_2 (top: maximal, bottom: averaged)

with the multigrid iterates u_i^k is calculated for $k \geq 2$ as long as $\bar{\rho}_{l,j}$ increases. (Here, again, $|\cdot|_{1,\Omega_i}$ is the energy norm on Ω_i, and we set $\bar{\rho}_{l,1} = 0$ if one multigrid step is needed only.) With the maximum obtained in this way, which we call $\bar{\rho}_l$, we determine $\rho_{m,i}$ and $\rho_{M,i}$ as

$$\rho_{m,i} = \frac{1}{n}\sum_{l=1}^{n} \bar{\rho}_l \qquad \text{and} \qquad \rho_{M,i} = \max_{1\leq l \leq n} \bar{\rho}_l\,, \qquad i = 1, 2\,,$$

where n is the number of Robin steps needed for the corresponding time step. Since we use the solution from the previous time step as the initial condition for the next time step, we already have a good approximation for the solution at that time step. With this choice we obtain fast multigrid convergence as one can see from Figures 3.34 and 3.35. Here, $\rho_{M,i}$ often occurs in the first few Robin steps, and the multigrid convergence rates can improve quite a lot for higher accuracies in the domain decomposition iteration history (where finally often one multigrid step is enough). This explains that the difference between $\rho_{M,i}$ and $\rho_{m,i}$ can be quite considerable.

Figure 3.36 displays the average convergence rates ρ for the domain decomposition iteration given by the Robin method at each time step. The convergence rates are calculated as in the examples on the Yin Yang domain, see (3.3.59) and (3.3.60), now of course with the initial condition given by the previous time step. For $t \in [138, 472]$, when the location of the wetting front has a nontrivial intersection with the interface Γ, the convergence rates vary quite a lot between around 0.3 and 0.9. These big variations can also be observed in the Robin iteration history at various time steps. In Figures 3.37 and 3.38 we illustrate two examples of such cases for $t = 197$ and $t = 443$ where we have the average convergence rates 0.56 and 0.63, respectively. One can see that different error reduction rates (i.e. convergence rates) are obtained

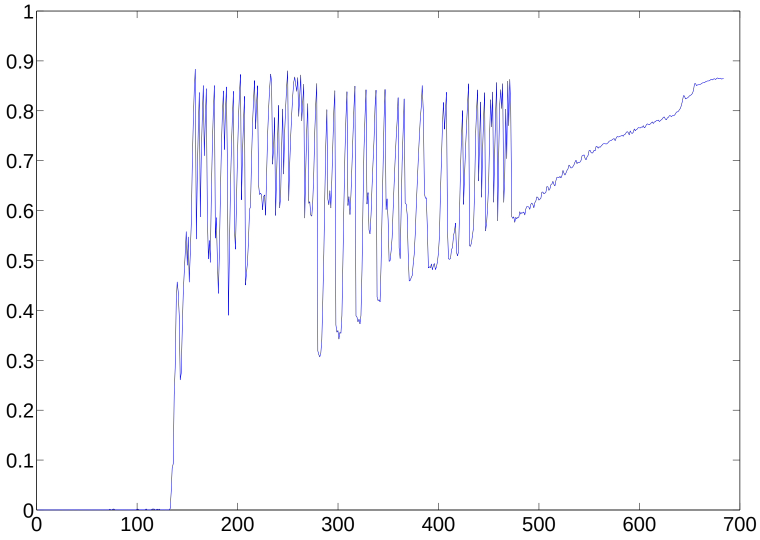


Figure 3.36: Convergence rates ρ per time step for the Robin method

for different accuracies, i.e. absolute errors

$$\left(\sum_{i=1}^{2} a_i(u_i^k - u_i^n, u_i^k - u_i^n) \right)^{1/2} , \quad k = 0, ..., n-1 , \tag{3.4.107}$$

(u_i^n being the last Robin iterate) in the iteration history. We assume that these effects occur because the pressure values for nodes directly at the wetting front probably depend quite sensitively on the solution of the previous time step, the precise Robin conditions at the interface and the required accuracy given by the stopping criterion. In addition, our measuring (3.3.59)–(3.3.60) of

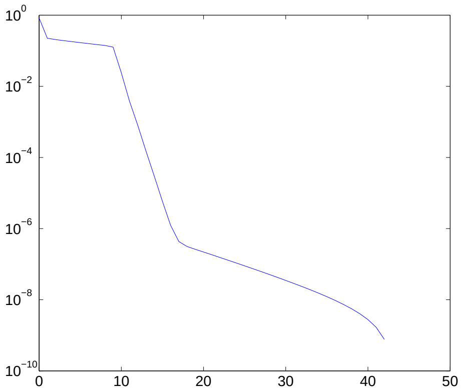


Figure 3.37: Error (3.4.107) vs. Robin iteration step at time $t = 197$

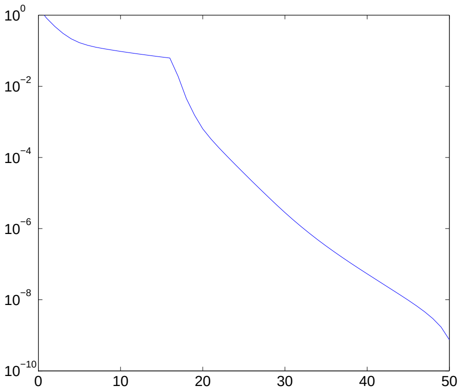


Figure 3.38: Error (3.4.107) vs. Robin iteration step at time $t = 443$

the convergence rates in the generalized variables u_i seems to be particularly sensitive in this respect, see Remark 3.4.31 below.

In general the first convergence rate in the iteration history is considerably smaller than the following ones for any time step. In particular, for time steps $t < 133$ the error reduction in the first Robin step is such that it already provides an almost vanishing average convergence rate. For $t \geq 473$, when Ω_1 is fully saturated and the wetting front is entirely located in Ω_2, the convergence rates do no longer oscillate, neither with respect to t nor in the iteration history for fixed t. Furthermore, they increase as the wetting front approaches the Signorini-type boundary until the stationary solution is attained at $t = 684$ (for $t > 684$ vanishing convergence rates are observed as expected).

Remark 3.4.31. As in the first part of this subsection concerning the examples on the Yin Yang domain, one might again be interested in the convergence rates measured in the physical variables p_i rather than in the generalized pressure variables u_i, $i = 1, 2$, as done so far. Since we have chosen $c = 0$ in the coffee filter example, one needs to be more careful here than in the Yin Yang examples because the inverse transformations κ_i^{-1}, by which u_i is transformed into p_i for $i = 1, 2$, are ill-conditioned for small generalized pressure values now, compare Figures 1.7 and 1.8. As can be seen in these figures, small perturbations in u_i can result in big variations of p_i in the unsaturated regime.

Therefore, the stopping criterion (3.3.59) expressed in p_i may correspond to a much more restrictive stopping criterion in u_i which might require a higher accuracy than provided by the local solvers given by (3.4.105). In the example above, a certain absolute error (3.4.107) in u_i usually corresponds to a much bigger absolute error calculated in p_i, $i = 1, 2$. In fact, they can differ by several orders of magnitude. We even observe numerical instabilities if we choose the same accuracy 10^{-12} in the stopping criterion (3.3.59) with u_i replaced by p_i. If, instead, we choose the stopping criterion

$$\frac{\left(\sum_{i=1}^{2} a_i(p_i^n - p_i^{n-1}, p_i^n - p_i^{n-1})\right)^{1/2}}{\left(\sum_{i=1}^{2} a_i(p_i^{n-1}, p_i^{n-1})\right)^{1/2}} < 10^{-9} \tag{3.4.108}$$

rather than (3.3.59) and measure the convergence rates as in (3.3.60) with u_i replaced by p_i, we obtain a time evolution (with 684 time steps) which practically does not differ from the one above (i.e., the first few digits of the obtained pressure values usually coincide). Interestingly, however, the convergence rates ρ_p per time step measured in the physical pressure p and displayed in Figure 3.39 do not show as big oscillations as the ones measured in u in Figure 3.36. Considerable oscillations only occur shortly before time step $t = 473$ when Ω_1 is fully saturated. In addition, the convergence rates measured p are more stable in the iteration history for fixed time steps than the ones measured in u.

As in the stationary case on the Yin Yang domain the convergence rates deteriorate on higher levels. They also deteriorate if we choose more extreme

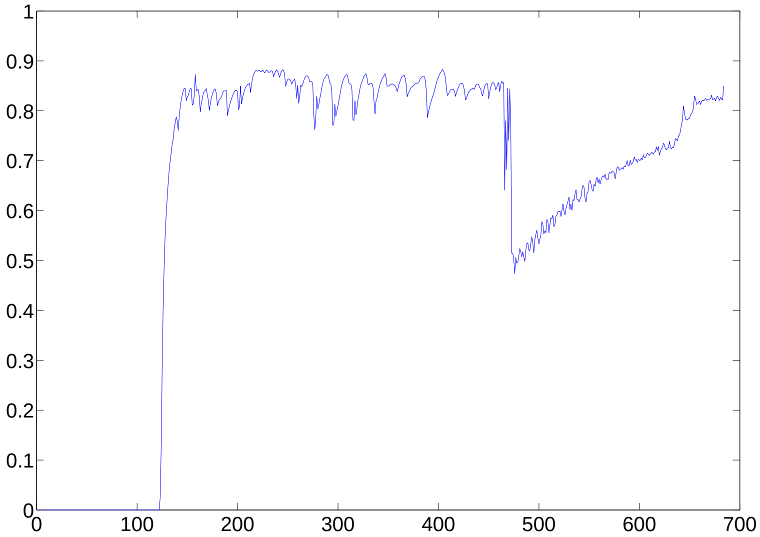


Figure 3.39: Convergence rates ρ_p per time step for the Robin method measured in physical variables with stopping criterion (3.4.108)

soil parameters such as the ones in Table 3.1. Furthermore, the convergence rates depend on the choice of the time step size for the variation of which we have to alter γ as well. (Observe that the time step size occurs as an additional factor in front of the water flux in the Robin conditions (3.4.16) or (3.4.19), compare (2.3.2) and (3.4.15).)

In addition, we observe deteriorating convergence rates if the pressure difference Δp of the wetting front is too big. Note, however, that the situation $\Delta p = 20$ in the coffee filter example with the Dirichlet value $p = 100$, measured in meters of a water column, already seems quite extreme. Therefore, we have hope that the Robin method can at least be successfully applied in reasonable hydrological settings which are not too extreme. The next chapter is devoted to the construction of such an example.

Chapter 4

Numerical results for the Richards equation with gravity, various soils and surface water in 2D

4.1 Introduction

In this last chapter we extend our solution method for the Richards equation in heterogeneous soil such that it also takes the gravitational impact into account. As a result we obtain a new solver for the Richards equation, which does not rely on smooth parameter functions or regularization such as the ones based on Newton's method which were mentioned in Section 2.2. Furthermore, we introduce a discrete version for the coupling of the saturated-unsaturated flow through a porous medium with surface water in a reservoir. At the end of this chapter we solve the Richards equation in a 2D-setting containing realistic hydrological ingredients.

The underlying idea of our approach to solve the Richards equation is to separate the occurring difficulties and to treat them in different steps. First of all, by Kirchhoff transformation (see Section 1.3) we "got rid of" the nonlinearity $kr(\theta(p))$ in the spatial derivative of the Richards equation (1.5.1) with the effect that a robust solver exploiting convexity rather than regularity (see Section 2.6 and 2.7) could be applied to the spatial problem in case of homogeneous soil. The inverse Kirchhoff transformation, which is in general ill-conditioned if $kr(\theta(p))$ can be arbitrarily small, and therefore the effect of a very small factor $kr(\theta(p))$ in the Richards equation (1.5.1), is then separated from the solution process and only occurs in the retransformation in order to obtain the physical pressure from the generalized pressure variable in which the solution was calculated.

Of course, with regard to the full Richards equation in heterogeneous soil, this approach still entails two difficulties that have to be dealt with. First, the spatial solver can only be applied to homogeneous soil. Therefore, we introduced non-overlapping domain decomposition methods (Dirichlet–Neumann and Robin method) in Section 3 for further treatment of the occurring nonlinear spatial problems in heterogeneous settings as in Definition 3.2.7. Secondly, we needed to treat the gravitational impact in the Richards equation (1.5.1) explicitly in the time discretization (2.3.2) in order to obtain a convex minimization problem to which our monotone multigrid solver can be applied.

It is well known that an explicit treatment of a convective term usually leads to instabilities in the numerical solution of the spatial problems which can be addressed by upwind techniques for small time step sizes. We will apply such a technique, more concretely an artificial viscosity method, in the framework of our finite element discretization of the Richards equation in Section 4.2. As a result, our numerical example therein shows that the problem posed by our explicit treatment of the gravitational term can be successfully addressed by this technique. Furthermore, the theoretical restriction on the time step size can be violated in this numerical example without encountering instabilities.

Finally, in Section 4.3 we apply our developed solver in a 2D-setting to the Richards equation with 4 different soils, boundary conditions of Signorini's type and surface water which is included via the reservoir model discussed in Remark 1.5.1. This final example contains both inflow from the surface water into the domain of the porous medium and outflow into the surface water and shall illustrate the applicability of our solver to realistic hydrological situations.

4.2 Treatment of the Richards equation with gravity in homogeneous soil

This section is devoted to the concrete treatment of the gravitational term $\operatorname{div}\big(kr(M(u))e_z\big)$ in our numerical solution of the homogeneous Richards equation (1.5.2). The theoretical presentation of this approach will be given in Subsection 4.2.1 followed by a numerical test of our method in Subsection 4.2.2. The method is needed in order to achieve stability of the numerical solution and it is based on an appropriate finite element discretization of the explicitly time-discretized gravitational term in (2.3.2). Such a discretization is obtained by adding a diffusion matrix to the straightforward discretization of that term which leads to upwind differences when interpreted as a finite difference scheme in special uniform settings. Our approach exploits the special direction of the convection given by the earth's gravitation and is refined by a lumping of the occurring convection matrix. It turns out that the restriction on the time step size, i.e. the CFL condition, imposed by the explicit treatment of the gravitation requires the time step size to be an order of magnitude smaller than the mesh size in realistic hydrological situations. The numerical example in Subsection 4.2.2, however, demonstrates that the method can be applied successfully in case of realistic hydrological data without the occurrence of numerical instabilities even if the CFL condition is violated considerably.

4.2.1 Upwind finite element discretization of the gravitational term by an artificial viscosity method

In this subsection we present an approach how one can easily and appropriately deal with the gravitational term $\operatorname{div}\big(kr(M(u))e_z\big)$ in the numerical treatment of the Richards equation (1.5.2). Recall that we already decided to treat the gravitational impact in a relatively simple way, i.e. explicitly in the time discretization of (1.5.2), on the right hand side in the linear functional of a variational inequality, see (2.3.2) and (3.4.96).

However, it is well known in the linear case already that a numerical treatment of an elliptic equation with a convective term is unstable if the latter is not space-discretized properly, which means that the sum of the stiffness matrix and the matrix coming from the convective part should result in an M-matrix (see e.g. Kornhuber and Schütte [60, p. 53] and Fuhrmann and Langmach [41]). In this sense we seek an appropriate space discretization of

$$\int_\Omega kr(M(u^{n-1}))e_z\nabla v\,dx = \int_\Omega kr(M(u^{n-1}))\,v_z\,dx = (kr(M(u^{n-1})),v_z)_{L^2(\Omega)} \tag{4.2.1}$$

in the linear functional (2.3.7) on the right hand side of (2.3.2). Note that the treatment of this term should lead to a stable scheme in particular if it has a considerable impact in the Richards equation (1.5.2) compared to the diffusion term $\operatorname{div}\nabla u$. Therefore, we take a look at the scalar conservation law

$$w_t + kr(w)_z = 0 \tag{4.2.2}$$

to which the Richards equation (1.5.2) reduces if we ignore the diffusion term and set $w := M(u)$.

Equations like (4.2.2) have been intensively studied both on the continuous and the discrete level, see for example Johnson [52] or Kröner [62] to which we refer in the following. It is well known for $kr \in C^1(\mathbb{R})$ that classical C^1-solutions w of (4.2.2) are (at least locally) constant on characteristics which are curves $t \mapsto (\gamma(t),t)$ in the z-t-plane with the property $\gamma'(t) = kr'(w(\gamma(t),t))$ for $t > 0$, see [62, p. 16]. Since we can assume $kr' > 0$ in case of the Brooks–Corey model (1.2.9) and (1.2.10), this already suggests that the information transport given by the equation (4.2.2) goes from the left to the right on the z-axis within time, see Figure 4.1. This fact should also be reflected by a discretization of (4.2.2) as a necessary condition for stability and convergence, compare [62, Ex. 2.1.20]. Therefore, in the framework of finite difference discretizations, one should choose *backward differences* or equivalently an *upwind discretization* of the spatial derivative $kr(w)_z$ in (4.2.2), see Kröner [62, Ex. 2.2.7] and Fuhrmann and Langmach [41, Sec. 6]. More concretely, with $w_i^n := w(z_i,t_n)$, $i = 0,\dots,K$, $K \in \mathbb{N}$, $z_0 < z_1 < \cdots < z_K$ and $n = 0,\dots,N$, $t_0 < t_1 < \ldots < t_N$

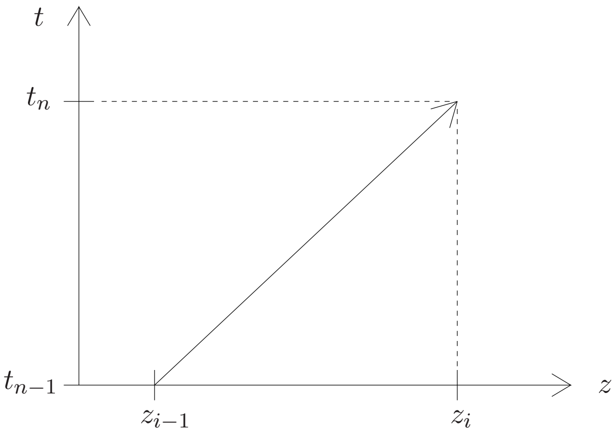


Figure 4.1: Upwind discretization

as in Subsection 2.3.1 such a finite difference discretization would read

$$\frac{w_i^n - w_i^{n-1}}{t_n - t_{n-1}} + \frac{kr\big(w_i^{n-1}\big) - kr\left(w_{i-1}^{n-1}\right)}{z_i - z_{i-1}} = 0 \tag{4.2.3}$$

together with certain initial and boundary conditions, see Figure 4.1.

Our aim is to reproduce (4.2.3) on the level of our finite element discretization of (1.5.2) in (2.3.2) and (3.4.96). As a first step to achieve this note that, as just seen, the convective part in the Richards equation (1.5.2) has a given fixed direction along the z-axis, independently of the dimension of Ω. Therefore, in 2D one should choose triangulations $\mathcal{T}_j$, $j \geq 0$, of Ω in which each triangle has an edge on a vertical line parallel to the z-axis in the y-z-plane. In other words, the triangulations should be arranged within vertical stripes in the y-z-plane as depicted in Figure 4.2.

With the notation as in Subsection 2.5.1 a straightforward finite element discretization of (4.2.1) could be given by $\mathcal{S}_j$-interpolations of the known function $kr(M(u^{n-1}))$ and of the test function v in this integral. This would lead to the matrix

$$C := (c_{pq})_{p,q\in\mathcal{N}_j} = \left(\lambda_q \frac{\partial}{\partial z}\lambda_p\right)_{p,q\in\mathcal{N}_j} . \tag{4.2.4}$$

Note that (4.2.1) was obtained after an application of Green's formula which was not the case for (4.2.2). However, this just leads to a rearrangement of terms by the application of the negative transposed matrix $-C^T$ instead of C and additionally occurring boundary terms. We come back to this fact later in this section. Since the integral (4.2.1) is on the right hand side of the equation corresponding to (4.2.2) we seek upwind difference quotients for $-kr(M(u^{n-1}(\cdot)))$ coming from suitable modifications in the discretization of that term.

In cases with congruent equilateral orthogonal triangles around an inner node $p \in \mathcal{N}_j$ in Figure 4.2 one can easily check that the matrix (4.2.4) produces central differences (up to a constant factor depending on j) of $-kr(M(u^{n-1}(\cdot)))$ for p on the vertical lines in z-direction but also "side effects" coming from

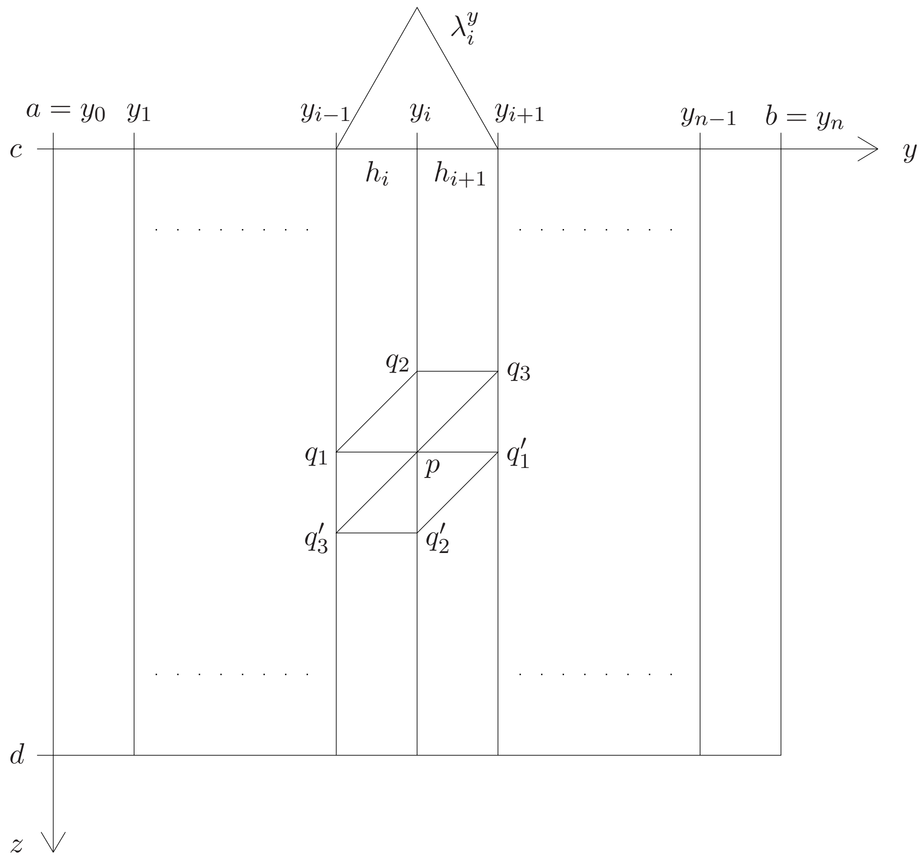


Figure 4.2: Vertical stripes for $\mathcal{T}_j$ and horizontal lumping in $\Omega = (a,b) \times (c,d)$

(neighbouring) nodes q on the neighbouring vertical lines. To "eliminate" the latter, one can carry out a lumping of the L^2-scalar product (4.2.1) in y-direction which we describe now. Here we need to assume that Ω is a rectangle with edges parallel to the y-axis or the z-axis. Then, by Fubini's theorem (see e.g. [82, pp. 164–167]), setting $w := M(u^{n-1})$ now, we can write

$$\int_\Omega kr(w)\, v_z\, dx = \int_a^b F(y)\, dy \tag{4.2.5}$$

with

$$F(y) := \int_c^d (kr(w)\, v_z)(y,z)\, dz \quad \forall y \in [a,b] \text{ a.e.}$$

and some real $a < b$ and $c < d$. We approximate the integral on the right hand side of (4.2.5) by an interpolation of F in a space of linear finite elements on the y-axis with nodes y_i, $i = 0, 1, \ldots, L$, $L \in \mathbb{N}$, given by and located on the vertical lines corresponding to our triangulation, see Figure 4.2. The resulting nodal basis functions in one space dimension which are independent of z are denoted by λ_i^y. We obtain

$$\int_a^b F(y)\, dy \approx \int_a^b \sum_{i=0}^{L} F(y_i)\, \lambda_i^y(y)\, dy = \sum_{i=0}^{L} F(y_i) \cdot \frac{1}{2}(h_i + h_{i+1})$$

if we denote $h_i := y_i - y_{i-1}$ for all $i = 1, \ldots, L$ and set $h_0 = h_{n+1} = 0$. Now, for y_i, $i = 0, 1, \ldots, L$, we approximate the integral $F(y_i)$ by interpolating both $kr(w)$ and v in $\mathcal{S}_j$ restricted to the vertical line parallel to the z-axis through the point (y_i, c). The corresponding nodal basis functions in 1D with their support on this line (and related to the nodes (y_i, z_{ik}) with the order $z_{ik} < z_{ik+1}$ for $k = 0, \ldots, K-1$) shall be denoted by λ_{ik}, $k = 0, \ldots, K$, with $K \in \mathbb{N}$ as above. The related nodes (y_i, z_{ik}) shall be ordered by $z_{ik-1} < z_{ik}$ with $h_{ik} := z_{ik} - z_{ik-1}$ for $k = 1, \ldots, K$, as for example $q_2 = (y_i, z_{ik-1})$, $p = (y_i, z_{ik})$ and $q_2' = (y_i, z_{ik+1})$ in Figure 4.2. Consequently, with $v = \lambda_{il}$, $l = 0, \ldots, K$, we have

$$F(y_i) = \int_c^d (kr(w)(\lambda_{il})_z)(y_i, z)\, dz \approx \sum_{k=0}^{K} kr(w(y_i, z_{ik})) \int_c^d \lambda_{ik} (\lambda_{il})_z \, dz$$

with

$$\int_c^d \lambda_{ik}(\lambda_{il})_z \, dz = \begin{cases} \frac{1}{2} & \text{for } l = k+1 \\ -\frac{1}{2} & \text{for } l = k-1 \\ 0 & \text{else}\,. \end{cases}$$

Altogether, by this lumping we replace the matrix C in (4.2.4) by

$$\tilde{C} := (\tilde{c}_{pq})_{p,q \in \mathcal{N}_j}$$

with

$$\tilde{c}_{pq} = \begin{cases} \frac{1}{4}(h_i + h_{i+1}) & \text{for } q = q_2 \\ -\frac{1}{4}(h_i + h_{i+1}) & \text{for } q = q_2' \\ 0 & \text{else} \end{cases} \tag{4.2.6}$$

where q_2 and q_2' are given according to Figure 4.2. Just as C, the matrix $\tilde{C}$ gives central differences (up to a j-independent factor) for inner nodes $p \in \mathcal{N}_j$, but now without additional contributions from vertical lines on which p is not situated. More concretely, for such a point this lumping is obtained by the definition

$$\begin{aligned} \tilde{c}_{pq_2} &:= c_{pq_2} + c_{pq_3} \\ \tilde{c}_{pq_2'} &:= c_{pq_2'} + c_{pq_3'} \\ \tilde{c}_{pp} &:= c_{pp} + c_{pq_1} + c_{pq_1'} \end{aligned} \tag{4.2.7}$$

with the setting and the notation as in Figure 4.2 if we have a constant mesh size $h = h_i = h_{ik}$ for all $i = 1, \ldots, L$ and $k = 1, \ldots, K$.

Now, it is well known that one can get one-sided differences from central differences by adding an artificial *viscosity* term, i.e. a diffusion term or central differences for second derivatives, see [62, Ex. 2.2.6]. Therefore, we introduce the one-dimensional diffusion matrix

$$D := \left(\frac{\partial}{\partial z} \lambda_q \frac{\partial}{\partial z} \lambda_p \right)_{p,q \in \mathcal{N}_j} \tag{4.2.8}$$

which (after multiplication with h^{-2}) leads to central differences for second derivatives of $-kr(M(u^{n-1}(\cdot)))$ in z-direction for inner nodes $p \in \mathcal{N}_j$, without further "side effects" coming from neighbouring points if we are in the uniform situation as in Figure 4.2 with a constant mesh size $h = h_i = h_{ik}$ for $i = 1, \dots, L$ and $k = 1, \dots, K$. In this setting one can easily verify that the sums of matrices

$$K := C - \frac{h}{3} D \tag{4.2.9}$$

and

$$\tilde{K} := \tilde{C} - \frac{h}{2} D \tag{4.2.10}$$

give upwind difference quotients for $-kr(M(u^{n-1}(\cdot)))$ in z-direction, i.e. the discretization (4.2.3), for inner nodes $p \in \mathcal{N}_j$ when multiplied with $3/(2h^2)$ (and with "side effects" in case of (4.2.9)).

For nodes $p \in \mathcal{N}_j$ on the leftmost and the rightmost vertical line on $\partial\Omega$ half of the upwind difference quotients from the inner nodes occur. Unfortunately, this is not the case for nodes on the top and the bottom horizontal line on $\partial\Omega$. Interestingly, however, the latter can be interpreted after an application of Green's formula to (4.2.1). Then we obtain an additional integral on $\partial\Omega$ but the contributions for $p \in \mathcal{N}_j$ in the corresponding matrix $-K^T$ or $-\tilde{K}^T$ on the bottom line lead to upwind difference quotients for $-kr(M(u^{n-1}(\cdot)))$ in this node now, whereas they vanish for $p \in \mathcal{N}_j$ on the top line (while the situation for all the other nodes does not change).

Both discretizations of (4.2.1) given by (4.2.9) and (4.2.10) are also applicable for domains $\Omega \subset \mathbb{R}^2$ which are not rectangles if a certain value for h related to the mesh size and big enough to ensure stability of the scheme is chosen. (We even choose h locally in case of non-uniform meshes.) Note, however, that (4.2.10) could only be derived by Fubini's theorem for the special rectangular cases considered above. Nevertheless, one can in principle use both definition (4.2.6) and (4.2.7) in more general cases, too. Whatever possibility is pursued, one should always choose vertical stripes for the triangulations of Ω in order to account for the direction of gravity in the space discretization of (4.2.1). Finally, we remark that our method of adding artificial viscosity terms given by the one-dimensional diffusion matrix (4.2.8) in (4.2.9) and (4.2.10) is just a special example of the streamline upwind Petrov–Galerkin method described in Johnson [52]. Since in our case the direction of convection is parallel to a coordinate axis and fixed for all times, that method can be more easily realized here.

Finally, it is well known that upwind discretizations (4.2.3) for the solution of (4.2.2) with $kr' > 0$ lead to so-called *monotone schemes* if the *CFL (Courant, Friedrichs, Lewy) condition* is satisfied, see [62, pp. 50/51, 66]. This condition is a restriction on the time step size $\tau := t_n - t_{n-1}$ in (4.2.3) which must be chosen small enough compared to the mesh size $h = z_i - z_{i-1}$ (we assume constant τ and h here). It reads

$$\tau < h \left(\sup_{w \in \mathbb{R}} |kr'(w)| \right)^{-1} \tag{4.2.11}$$

and is the price that has to be paid if explicit time discretizations are applied to convective terms as done in (4.2.3) and for the Richards equation in (2.3.2). If (4.2.11) is satisfied, however, the monotone scheme (4.2.3) is stable and convergent, see [62, pp. 72, 91].

In order to get an impression of the size of the factor $(\sup_{w\in\mathbb{R}} |kr'(w)|)^{-1}$ in (4.2.11) for realistic situations, we take a look at concrete hydrological examples for the relative permeability function $kr(\cdot)$. First, recall that the Brooks–Corey model (1.2.9) and (1.2.10) provides

$$kr'(\theta) = e(\lambda) \left(\frac{\theta - \theta_m}{\theta_M - \theta_m} \right)^{e(\lambda)-1}$$

and therefore

$$\sup_{w\in\mathbb{R}} |kr'(w)| = e(\lambda) \cdot 1 = 3 + \frac{2}{\lambda} = \begin{cases} \ 5 & \text{for } \ \lambda = 1 \\ 23 & \text{for } \ \lambda = 0.1 \end{cases} \tag{4.2.12}$$

if we choose $e(\lambda)$ according to Burdine in (1.2.10), compare Figure 1.2. Here we have chosen extreme values $\lambda = 1$ (coarse sand) and $\lambda = 0.1$ (fine clay) for the pore size distribution factor, compare Rawls et al. [77, Table 5.3.2]. However, our numerical example to which we turn in the next subsection, suggests that the theoretical bounds in (4.2.11) given by (4.2.12) are quite pessimistic in practical situations and that larger time step sizes can be used without visible drawbacks concerning the numerical stability.

4.2.2 Numerical test: Richards equation with gravity

In the following, we present a numerical example which shall illustrate the performance of the artificial viscosity method described in the previous subsection, more concretely the lumped version given by the matrix in (4.2.10). As the do-

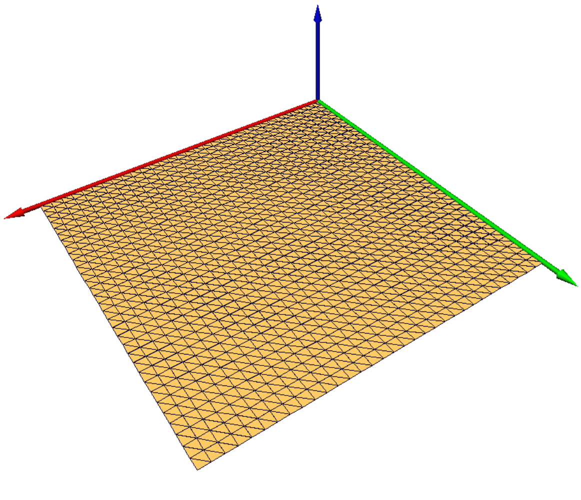


Figure 4.3: Initial condition $p = -20$, finest grid, gravity along right axis

main we choose the unit square $\Omega = [0,1]^2$ in the y-z-plane where the z-axis (i.e. the direction of gravity), is directed downwards (to the right in Figure 4.3) and the y-axis is directed to the left as in Figure 4.3. Therein, one can see the initial condition $p = -20$ (in meters of a water column, i.e. a practically dry soil) on the finest grid with $h = 1/32$ on the fourth refinement level. The coarse grid and uniform refinement are chosen such that equilateral orthogonal triangles occur within vertical stripes as proposed above (compare Figure 4.2).

We assume homogeneous soil parameters in Ω, more concretely we choose the parameters of sand as given by the USDA soil texture triangle in Rawls et al. [77, Tables 5.3.2 and 5.5.5], see Table 4.1. With regard to the parameter functions the (unaltered) Brooks–Corey model (1.2.9) and (1.2.10) shall be applied here.

$\Omega = [0,1]^2$	n	θ_m	λ	p_b	K_h
soil: sand	0.437	0.046	0.694	-0.073	$6.54 \cdot 10^{-5}$

Table 4.1: Soil parameters of sand for the Richards equation with gravity

We choose a constant inflow $-\mathbf{v} \cdot \mathbf{n} = 0.002\,[m/s]$ on $\gamma_i := \{0\} \times [0.25, 0.5]$ (i.e. on the right face of Ω) and homogeneous Neumann boundary data $\mathbf{v} \cdot \mathbf{n} = 0$ on $\partial\Omega \backslash \gamma_i$ which can already be resolved on the coarse grid. As a consequence, the time evolution depicted in Figures 4.4–4.12 in heightplots on the left and colourplots on the right (with gravity directed downwards) shows an increasing physical pressure while the saturated regime in Ω extends more and more.

Due to the mass matrix arising from the spatial discretization of the saturation term (see for example (2.5.44)) the occurring spatial Neumann problems are uniquely solvable for each time step $t \leq 810$. Our spatial solver is the monotone multigrid as used in Subsection 3.4.6 for the local problems in the coffee filter example with the same stopping criterion given in (3.4.105). Recall that the gravitational impact only occurs in the right hand side of the discrete spatial problems. As usual the solution of such a problem serves as the initial iterate for the next time step.

Starting with the solution for the first time step $t = 10$, the evolution of the physical pressure in Figures 4.4–4.12 is given at equidistant time steps until for $t = 810$ the domain is almost fully saturated. As in the coffee filter example in Subsection 3.4.6, we obtain a quite sharp moving wetting front, where a pressure difference of almost 20 occurs and which represents the interface between the saturated and the unsaturated regime of Ω. Moreover, we observe a pressure decline from the face γ_i where inflow is imposed to the wetting front which increases in time. For $t = 810$ we have $p_{\max} = 10.9$ in the central node of γ_i and $p_{\min} = -14.5$ in the left bottom corner of Ω while the pressure practically vanishes for all nodes on the left face of Ω where the domain is fully saturated. Note that we have a constant flow of water into the domain and we do not allow outflow. Therefore, it is no surprise that the multigrid no longer converges

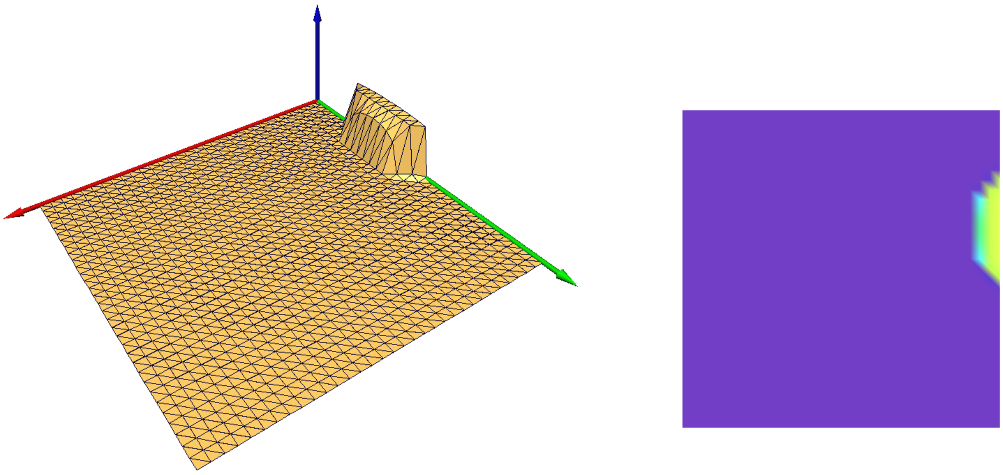


Figure 4.4: $t = 10$

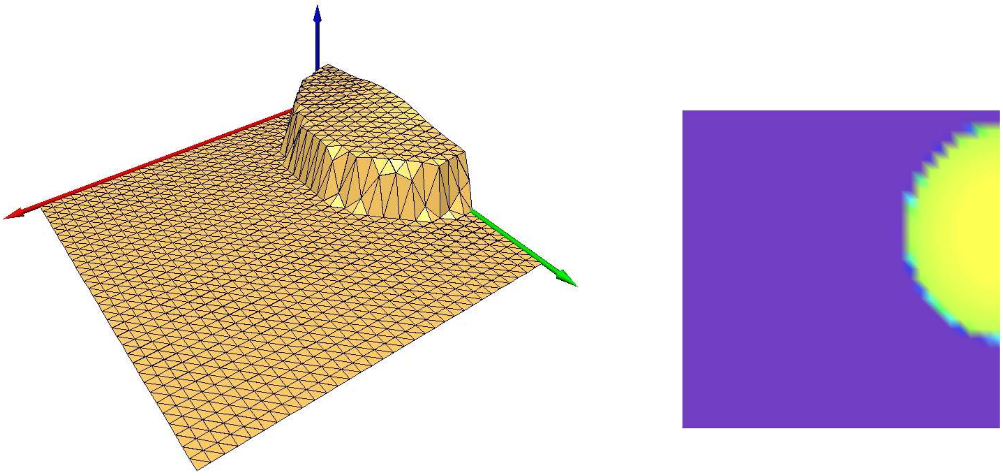


Figure 4.5: $t = 110$

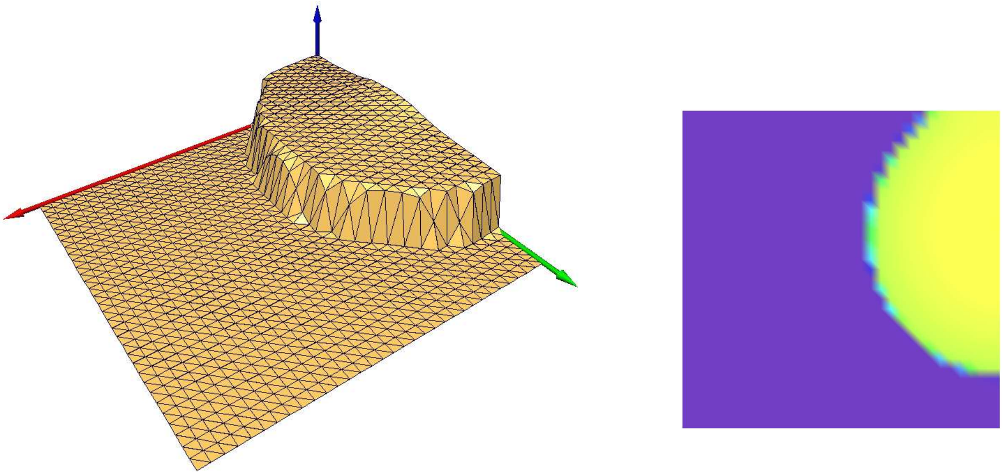


Figure 4.6: $t = 210$

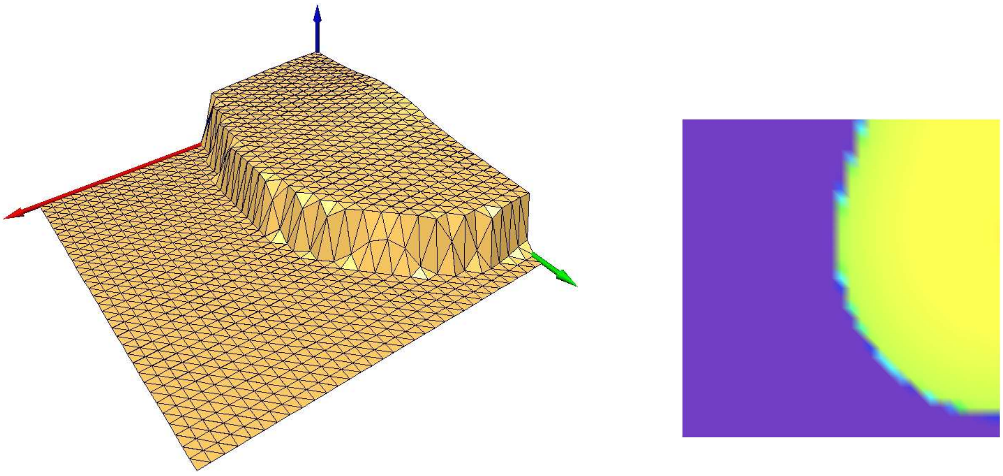


Figure 4.7: $t = 310$

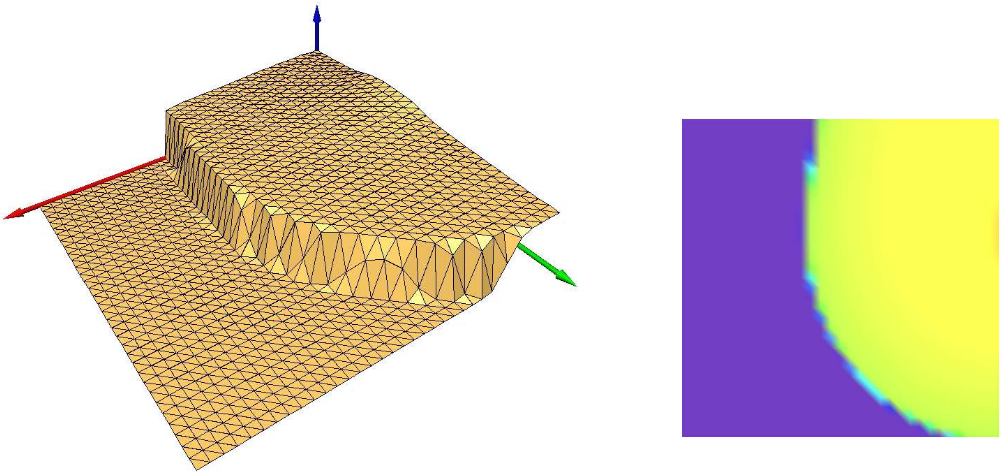


Figure 4.8: $t = 410$

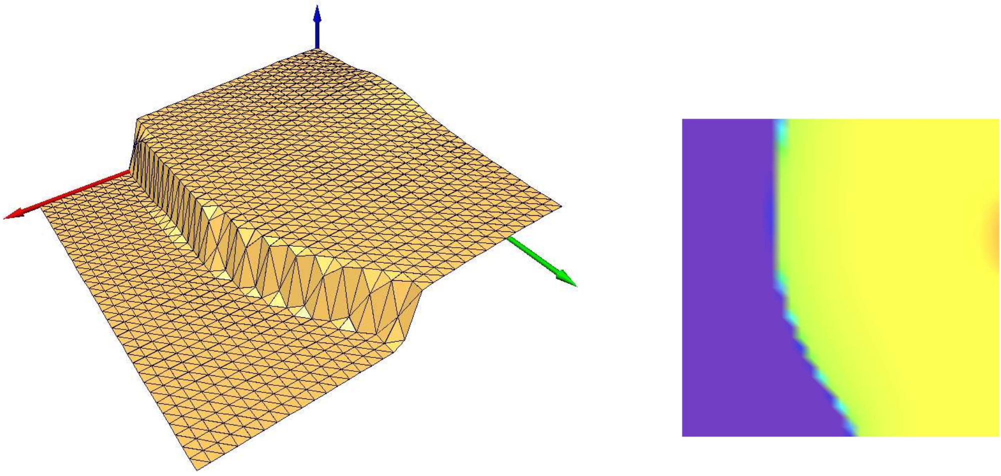


Figure 4.9: $t = 510$

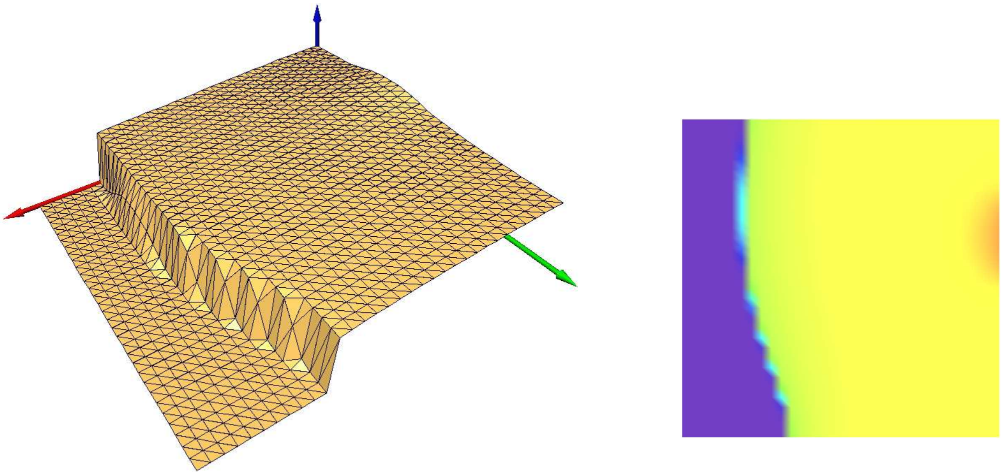


Figure 4.10: $t = 610$

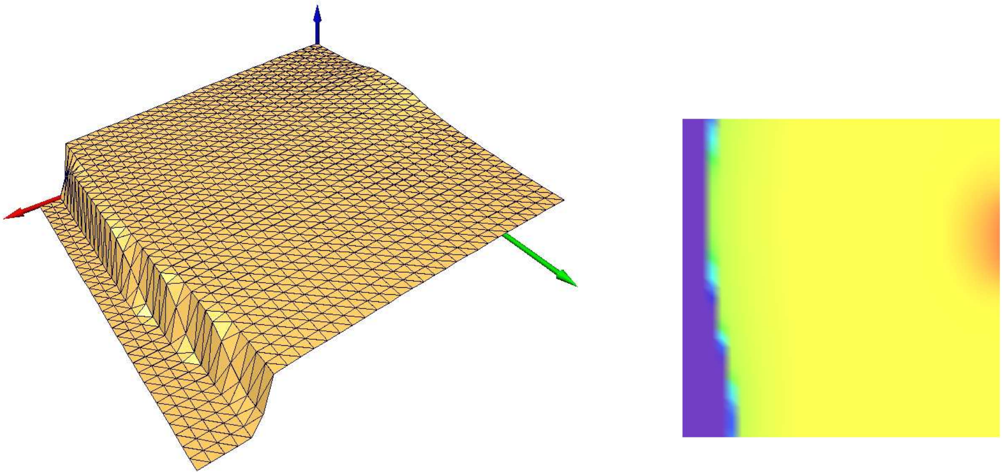


Figure 4.11: $t = 710$

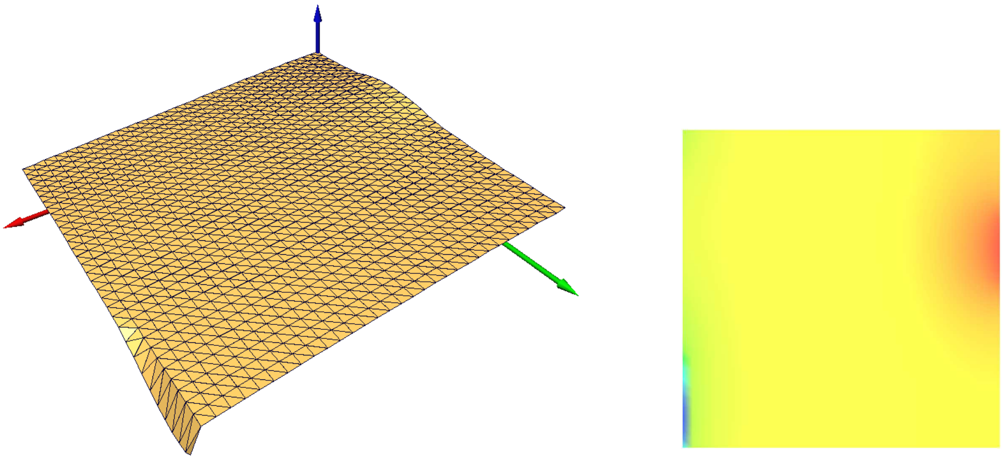


Figure 4.12: $t = 810$

for the next time step $t = 820$, for which the maximal amount of fluid in the domain would be exceeded due to the constant inflow.

The (constant) time step size $\tau = 10$ $[s]$ used in this example is three orders of magnitude larger than the upper bound given by the CFL condition (4.2.11) which is

$$h\left(3 + \frac{2}{\lambda}\right)^{-1} \approx \frac{1}{32} \cdot \frac{1}{5.88} \approx \frac{1}{188}$$

in our case (see (4.2.12) and Table 4.1). Surprisingly, we do not observe any visible numerical instabilities in the corresponding graphics if we further increase the time step size τ and decrease the mesh size h. For example, we can still solve the problem with $h = 2^{-10}$, i.e. on the 9th level, and varying big time step sizes up to $\tau = 810$ without encountering instabilities. Remarkably, these effects are the same if we use clay (compare Table 3.1) instead of sand in the model problem. In that case, we do not obtain a wetting front as in Figures 4.4–4.12 but rather a uniform decline of the pressure from the right face of Ω to the boundary of the dry regime $p = -20$ in Ω. However, this decline is steeper than the wetting front in sand since the range of the pressure is two orders of magnitude larger in clay. Observe that we obtained a stronger theoretical time step restriction (4.2.12) for clay than for sand. Altogether, the results show the practicability and the stabilizing effect of our upwind method for the gravitation as discussed in the previous subsection. In addition, the theoretical bounds for the time step sizes given by the CFL condition (4.2.11) seem to be too pessimistic at least for realistic hydrological data.

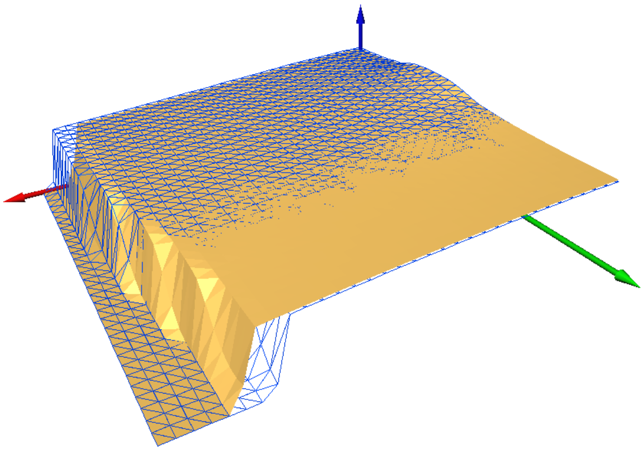


Figure 4.13: Solutions with gravity (surface graph) and without gravity (lines) at time $t = 710$

In Figure 4.13 one can see the effect of the gravitation in our model problem. The surface graph in this figure represents the above solution of our problem with gravity at the time $t = 710$. The graph given by lines was obtained at the same time for the same problem without gravity. With regard to the next numerical example, we remark that we obtain a solution for which the pressure values coincide in the first few digits with the corresponding ones in Figures 4.4–4.12 if we apply the Robin method to this homogeneous problem. In the following last section we consider a heterogeneous problem for the Richards equation with gravity and surface water.

4.3 Numerical example: Richards equation in four different soils with surface water

The following last section of this work is devoted to the presentation of a numerical example for the Richards equation in a heterogeneous setting in two space dimensions including surface water. This example simulates a situation which one may call hydrologically realistic in principle. The results suggest that our algorithm might well be suitable for the solution of more practical groundwater flow problems.

Before we turn to the presentation of our example we explain how we treat the coupling of the Richards equation with surface water numerically. Recall that in Remark 1.5.1 we already introduced the model we use for this coupling in the continuous setting. It is based on the assumption of mass conservation (1.5.8) and uses a simple reservoir model for the surface water, compare Figure 1.16. Loosely speaking, the flow of water out of (or into) the domain increases (or decreases) the height of the water reservoir, which in turn has an effect on the shape of the Dirichlet boundary and the size of the Dirichlet boundary values given by the hydrostatic pressure.

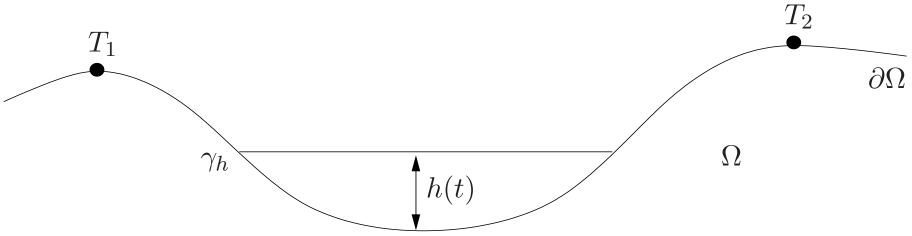


Figure 4.14: Flow across $\gamma_h \subset \partial\Omega$ between T_1 and T_2 affects lake height $h(t)$.

More concretely, with a glance at Figure 4.14, we call γ_h the part of the boundary between the top points T_1 and T_2 through which (independent of time) flow of water contributes to the height $h(t)$ of a lake between T_1 and T_2. Now, choosing a suitable $\tilde{\gamma}_h \subset \partial\Omega$ with $\gamma_h \subset \tilde{\gamma}_h$ and a $v \in H^1_{\partial\Omega\setminus\tilde{\gamma}_h}(\Omega)$ with the trace $tr_{\gamma_h} v = 1$, we can approximate the integral on the right hand side of (1.5.8) as

$$\int_{\tilde{\gamma}_h} \mathbf{v}\cdot\mathbf{n}\, v\, d\sigma = -\int_\Omega M(u)_t\, v\, dx - \int_\Omega \nabla u \nabla v\, dx + \int_\Omega kr(M(u))e_z \nabla v\, dx$$

if $\tilde{\gamma}_h \setminus \gamma_h$ has a small Hausdorff measure and the functions involved are smooth enough, compare (1.5.13). This gives rise to the following explicit time discretization of (1.5.8). With a slight abuse of notation we call $u(t_n)$ the semi-discrete solution for a time step t_n, $n \geq 0$, corresponding to u^n in (2.3.2). We denote $u(t_{-1}) := u^0$ with the initial condition u^0 and choose a constant time step size $\tau = t_n - t_{n-1}$, $n = 1, \ldots, N$. Then, with given solutions $u(t_n)$ and $u(t_{n-1})$, $n \geq 0$, the new approximated volume $V(t_{n+1})$ of the lake is defined as

$$V(t_{n+1}) = V(t_n) - \int_\Omega (M(u(t_n)) - M(u(t_{n-1})))\, v\, dx - \tau \int_\Omega \nabla u(t_n) \nabla v\, dx \\ + \tau \int_\Omega kr(M(u(t_n)))e_z \nabla v\, dx\,. \quad (4.3.1)$$

Finally, using the same notation as in Subsection 2.5.1, we carry out the space discretization of (4.3.1) in the finite element space $\mathcal{S}_j$ for a $j \geq 0$. As a test function we choose

$$v = \sum_{p \in \mathcal{N}_j \cap \gamma_h} \lambda_p$$

which has a support in a neighbourhood of γ_h only, and the discretization of the three integrals on the right hand side of (4.3.1) is carried out as already described earlier. The discretization of the first integral is given via the mass matrix as in (2.5.44), the second one is given with the help of the stiffness matrix as usual and the third integral is discretized via the matrix $\tilde{K}$ in (4.2.10) as explained in Subsection 4.2.1.

In order to translate the volume $V(t_{n+1})$ (now obtained by the fully discrete version of (4.3.1)) into the new height $h(t_{n+1})$ of the lake in our implementation we use the following simplified geometry model. First, we assume that γ_h can be approximated by a part of a semi-circle line with a certain radius $r > 0$ which should be sufficiently accurate at least for small heights of a lake. Then we can write

$$V(t_{n+1}) = \int_0^{h(t_{n+1})} 2\sqrt{r^2 - (r-s)^2}\, ds$$

and we just approximate the right hand side by numerical integration. Concretely, for a fixed small interval length Δs and $k \geq 1$ with $k\,\Delta s \leq r$ we approximate

$$\int_{(k-1)\,\Delta s}^{k\,\Delta s} 2\sqrt{r^2 - (r-s)^2}\, ds \approx I_k$$

where I_k is obtained by the trapezoidal rule applied to the integral. Then we set $h(t_{n+1}) := k'\Delta s$ if

$$\sum_{k=1}^{k'} I_k \geq V(t_{n+1}) > \sum_{k=1}^{k'-1} I_k$$

and $h(t_{n+1}) := 0$ if $V(t_{n+1}) \leq 0$. Otherwise $V(t_{n+1})$ is too big for our geometry model which does not occur in our example below. Now, with $h(t_{n+1})$ we obtain the Dirichlet boundary $\gamma_D(t_{n+1}) \subset \gamma_h$ together with the Dirichlet boundary values given by the hydrostatic pressure of the water in the lake above each Dirichlet node. The rest of the top boundary, which is not covered by the lake, is treated as a boundary of Signorini's type, compare Subsection 1.5.1.

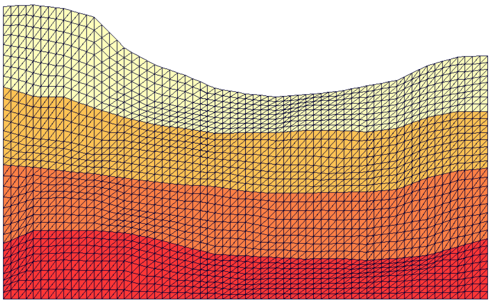


Figure 4.15: Fine grid, four layers of soil from top to bottom as in Table 4.2

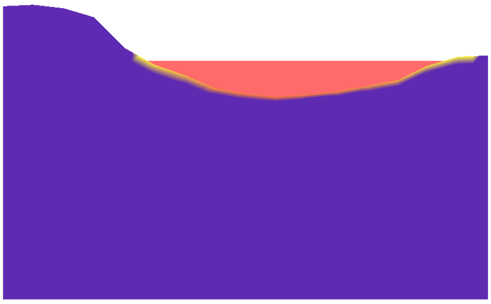


Figure 4.16: Initial condition: dry soil and surface water

With regard to our concrete example, Figure 4.15 shows the domain $\Omega \subset \mathbb{R}^2$ decomposed into the four subdomains Ω_1, Ω_2, Ω_3 and Ω_4 (from the top to the bottom) which we use for our model problem. The width of the domain is $2\,[m]$ and the height from the bottom to the highest point of Ω is approximately $1.214\,[m]$. The z-axis points downwards in gravitational direction. As always in this work we apply the Brooks–Corey model according to Burdine for the equations of state (1.2.9) and (1.2.10). We choose the soil parameters of sand, loamy sand, sandy loam and loam given in Table 4.2 (compare Rawls et al. [77, Tables 5.3.2 and 5.5.5]) as the parameters in the layers of soil corresponding to Ω_1, Ω_2, Ω_3 and Ω_4. Figure 4.15 already shows the finest grid (with the mesh size $h = 0.038$) which we obtain on the third refinement level with 585 nodes in each subdomain.

Ω_i	n_i	$\theta_{m,i}$	λ_i	$p_{b,i}$	$K_{h,i}$
$i = 1$ (sand)	0.437	0.046	0.694	-0.073	$6.54 \cdot 10^{-5}$
$i = 2$ (loamy sand)	0.437	0.080	0.553	-0.087	$1.66 \cdot 10^{-5}$
$i = 3$ (sandy loam)	0.453	0.091	0.378	-0.147	$6.06 \cdot 10^{-6}$
$i = 4$ (loam)	0.463	0.058	0.252	-0.112	$3.67 \cdot 10^{-6}$

Table 4.2: Soil parameters for the heterogeneous problem with surface water

As the initial condition depicted in Figure 4.16 in a colourplot we choose $p = -10$ (meters of a water column) corresponding to an initially dry soil in Ω except for the nodes on the top boundary which are covered by surface water (red in Figure 4.16) where a hydrostatic pressure from the lake is given. The height of the lake at the time $t = 0$ is $0.1686\,[m]$, the radius of the circle line by which we approximate γ_h is $r = 1.2\,[m]$.

As already indicated above, the part of the top boundary which is not covered by the lake is treated as a boundary of Signorini's type for all time steps (compare also the situation in Theorem 3.2.4 and Remark 3.2.5). In the time evolution we impose a constant inflow of $-\mathbf{v}\cdot\mathbf{n} = 3\cdot 10^{-4}\,[m/s]$ across the lower half of the left boundary of Ω_1. On the rest of the left boundary of Ω as well as on the right and the bottom part of $\partial\Omega$ homogeneous Neumann boundary conditions $\mathbf{v}\cdot\mathbf{n} = 0$ are assumed for all time steps. For the time evolution we choose the constant time step size $\tau = 10\,[s]$.

The spatial problems for each time step are solved by the Robin method in which one iteration step for the four subdomains looks as follows. First, for any iterate u^k, $k \geq 1$, provided by the method on Ω, and the solution u^0 from the previous time step we abbreviate

$$u_i^k := u^k_{|\Omega_i}\,, \quad i = 1,2,3,4\,.$$

Now, given an iterate u^k, $k \geq 0$, on Ω, we determine u_1^{k+1} according to (3.4.15)–(3.4.16) and u_2^{k+1} according to (3.4.17)–(3.4.18), in which the interface with the Robin boundary condition for the unknown u_2^{k+1} also contains the connected component $\partial\Omega_2 \cap \partial\Omega_3$ of $\partial\Omega_2\backslash\partial\Omega$, where the previous iterate u_3^k on Ω_3 contributes to the right hand side in (3.4.18) as well. The subsequent iterates u_3^{k+1} and then u_4^{k+1} are obtained analogously as u_2^{k+1} and u_1^{k+1}, respectively. We choose the constant parameter $\gamma = 10^{-4}$ suggested by numerical experiments for the Robin method at all time steps.

As the inner solver for the homogeneous problem on each subdomain Ω_i for $i = 1,2,3,4$, we apply the monotone multigrid method described in Subsection 3.4.5 and already used for the coffee filter example in Subsection 3.4.6. Again, the stopping criterion for the multigrid is given by (3.4.105).

In order to account for the realistic nature of our model problem in this section we use another stopping criterion and another way of measuring the performance of the Robin method than in Subsection 3.4.6. To this end, we first estimate a global convergence rate ρ of the Robin method for all time steps, carrying out the whole calculation for the time evolution with the quite restrictive stopping criterion

$$\frac{\left(\sum_{i=1}^4 a_i(p_i^n - p_i^{n-1}, p_i^n - p_i^{n-1})\right)^{1/2}}{\left(\sum_{i=1}^4 a_i(p_i^{n-1}, p_i^{n-1})\right)^{1/2}} < 10^{-8} \tag{4.3.2}$$

in which we denote $p_i^n := p^n_{|\Omega_i}$ for the nth iterate p^n of the domain decomposition iteration. We refer to Remark 3.4.31 for a discussion on a stopping criterion given in terms of the physical pressure p rather than the generalized pressure u. As before, the corresponding bilinear forms $a_i(\cdot,\cdot)$, $i = 1,2,3,4$, are induced via the related stiffness matrices given by the problem on the subdomains Ω_i. In what is to come, the norm arising from these forms on $\mathcal{S}_j$ shall be denoted by

$$\|p\| := \left(\sum_{i=1}^4 a_i(p_i, p_i)\right)^{1/2}, \quad p \in \mathcal{S}_j\,, \quad p_i := p_{|\Omega_i}\,, \ i = 1,2,3,4\,.$$

In the computation where the stopping criterion (4.3.2) was used we observed that

$$\|p^{k+1} - p^k\| \leq \rho \, \|p^k - p^{k-1}\| \, , \quad 1 \leq k \leq n-1 \, , \tag{4.3.3}$$

usually holds with the maximal rate $\rho = 0.95$ for succeeding iterates of the Robin iteration at any time step. Now, in addition to (4.3.3) we assume that this convergence rate is globally valid for this problem in the sense that for any time step we have

$$\|p - p^{k+1}\| \leq \rho \, \|p - p^k\| \, , \quad k \geq 0 \, ,$$

with the exact solution p of the discrete spatial problem. Then we can write

$$(1-\rho)\|p - p^k\| \leq \|p - p^k\| - \|p - p^{k+1}\| \leq \|p^{k+1} - p^k\| \, , \quad k \geq 0 \, ,$$

and assuming $\|p\| \approx \|p^k\|$, which is justified if initial iterates are chosen in a neighbourhood of p, we get

$$\frac{\|p - p^k\|}{\|p\|} \leq \frac{1}{1-\rho} \cdot \frac{\|p^{k+1} - p^k\|}{\|p^k\|} \, , \quad k \geq 0 \, .$$

Now, in order to determine p at least up to a relative accuracy of 1% we choose the stopping criterion

$$\frac{\|p^{k+1} - p^k\|}{\|p^k\|} < 0.0005 \tag{4.3.4}$$

and obtain

$$\frac{\|p - p^k\|}{\|p\|} < \frac{1}{1 - 0.95} \cdot 0.0005 = 0.01 \, . \tag{4.3.5}$$

Figure 4.41 shows the number of iterations obtained with this stopping criterion per time step (recall $\tau = 10$) for the time evolution displayed in Figures 4.17–4.40. Note that in these figures the evolution is given columnwise in time. Except for the last column we have chosen constant time intervals Δt in each column which are, however, increased considerably later in the evolution.

One can observe a quite fast evolution in the first 20 time steps, in which the lake loses more than half of its height while its water flows quickly into the first soil layer which is sand. With regard to the fast evolution at the beginning as compared to later time steps observe that the hydraulic conductivity of the soil gets smaller from layer to layer if we go downwards. Between $t = 800$ and $t = 4320$ the water level is below 0.01, and it is equal to 0 (e.g. for $t \in [940, 1160]$) or oscillating in the interval $[0, 0.0012]$ between $t = 930$ and $t = 2410$. Later the water level rises again slowly while the saturated area of the soil increases gradually (the colour blue in the graphics represents the initial pressure $p = -10$ while we have $p \approx 0$ in the yellow regions, orange for $p \approx 1$ and red for $p \approx 2$). At $t = 12790$ (Figure 4.39) the surface water has reached its initial height $h = 0.1686$ again, and at $t = 13420$ (Figure 4.40) the pressure has been equalized in the right bottom corner of Ω and the domain is fully saturated with the range $p \in [0, 2.2]$ while the lake can already be regarded as overflowing.

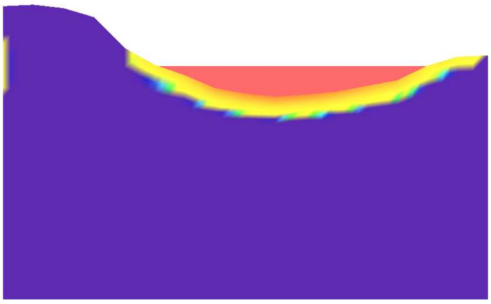


Figure 4.17: $t = 10$

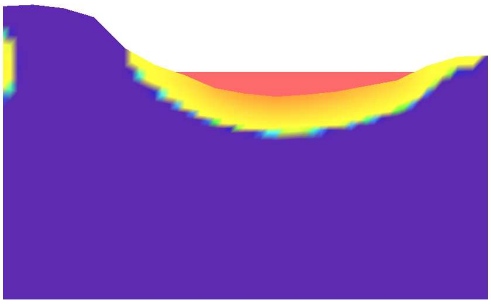


Figure 4.18: $t = 40$

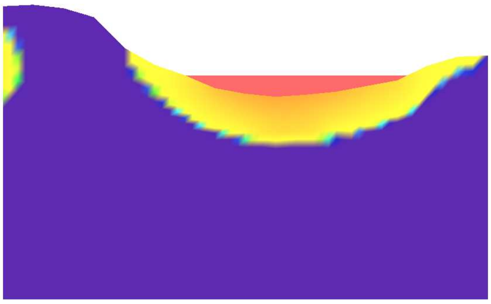


Figure 4.19: $t = 70$

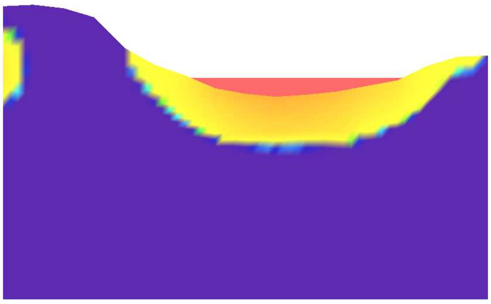


Figure 4.20: $t = 100$

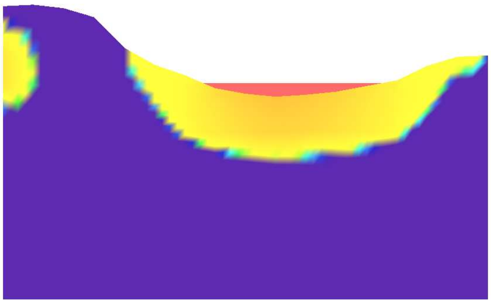


Figure 4.21: $t = 200$

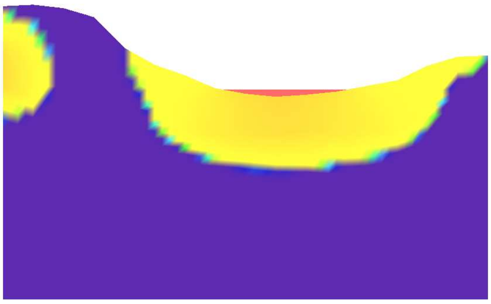


Figure 4.22: $t = 400$

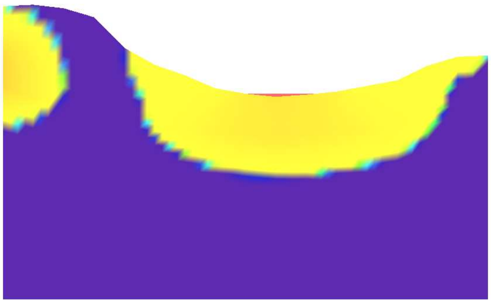


Figure 4.23: $t = 600$

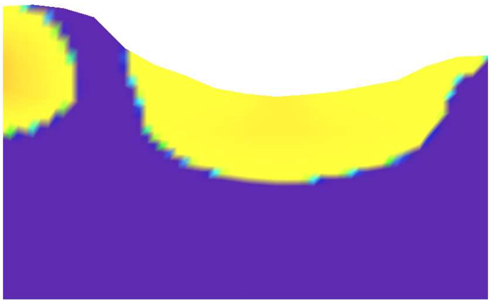


Figure 4.24: $t = 800$

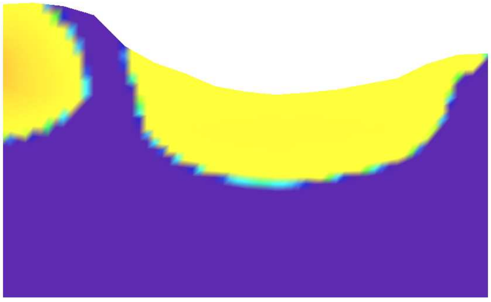


Figure 4.25: $t = 1000$

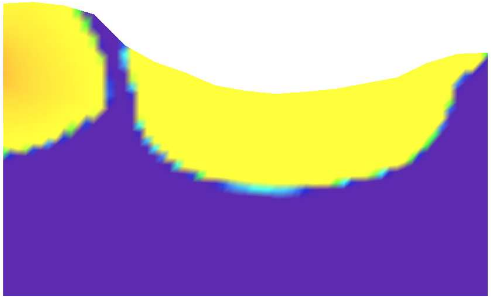


Figure 4.26: $t = 1500$

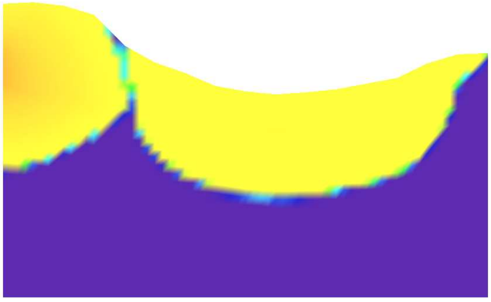


Figure 4.27: $t = 2000$

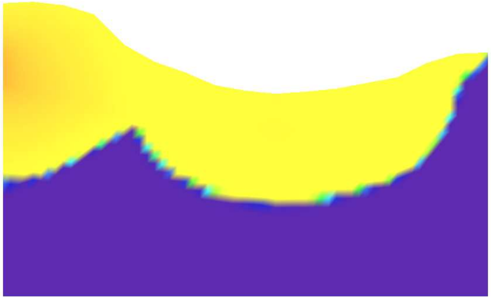


Figure 4.28: $t = 2500$

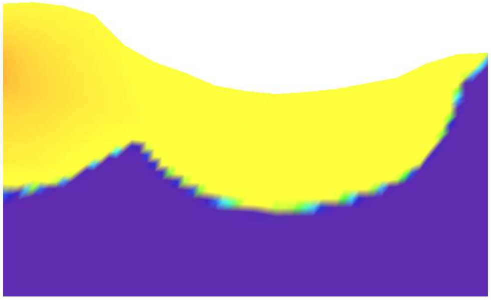


Figure 4.29: $t = 3000$

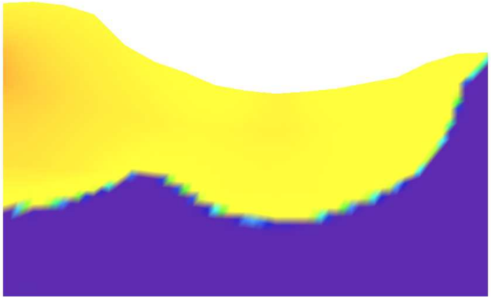


Figure 4.30: $t = 4000$

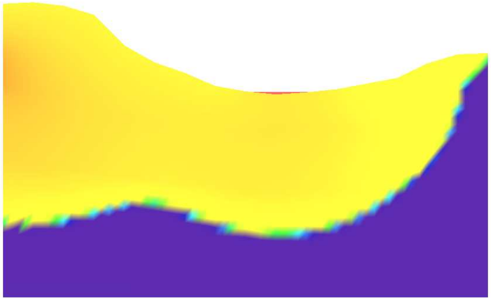


Figure 4.31: 5000

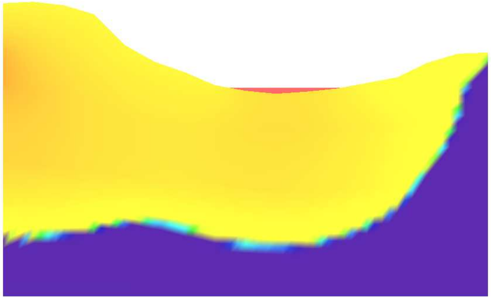


Figure 4.32: $t = 6000$

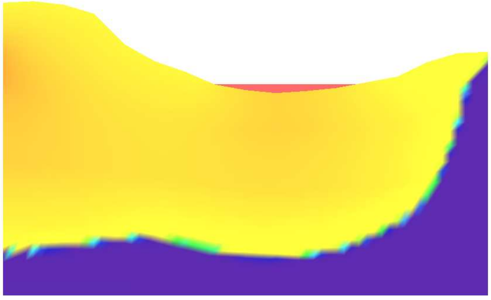


Figure 4.33: $t = 7000$

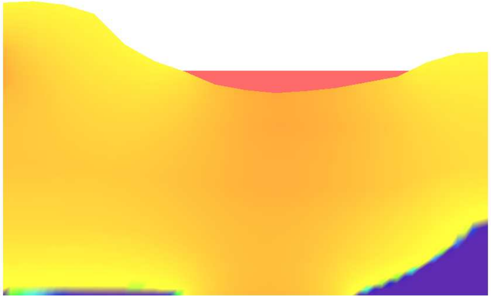


Figure 4.37: $t = 11000$

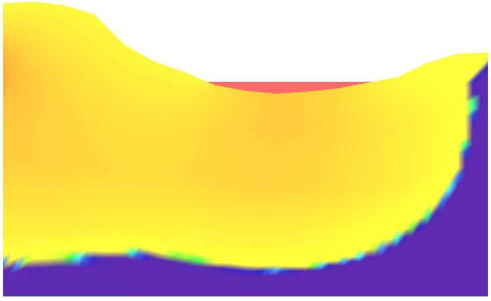


Figure 4.34: $t = 8000$

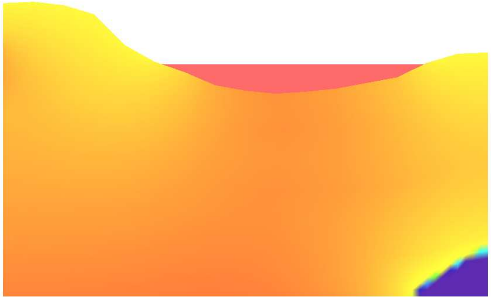


Figure 4.38: $t = 12000$

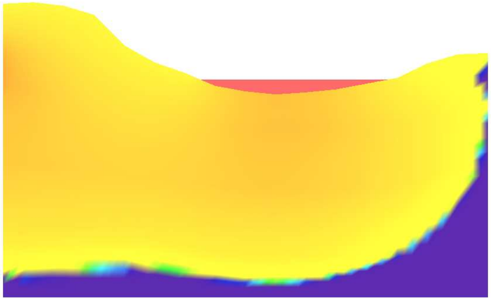


Figure 4.35: $t = 9000$

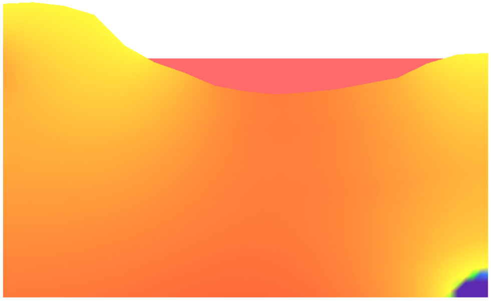


Figure 4.39: $t = 12790$

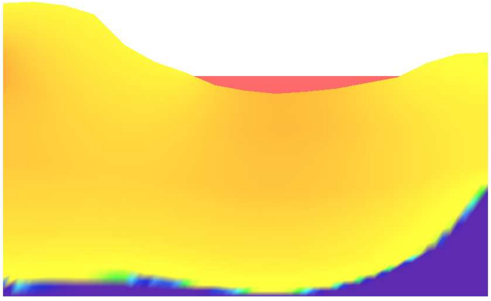


Figure 4.36: $t = 10000$

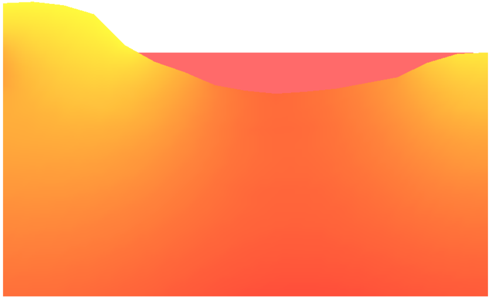


Figure 4.40: $t = 13420$

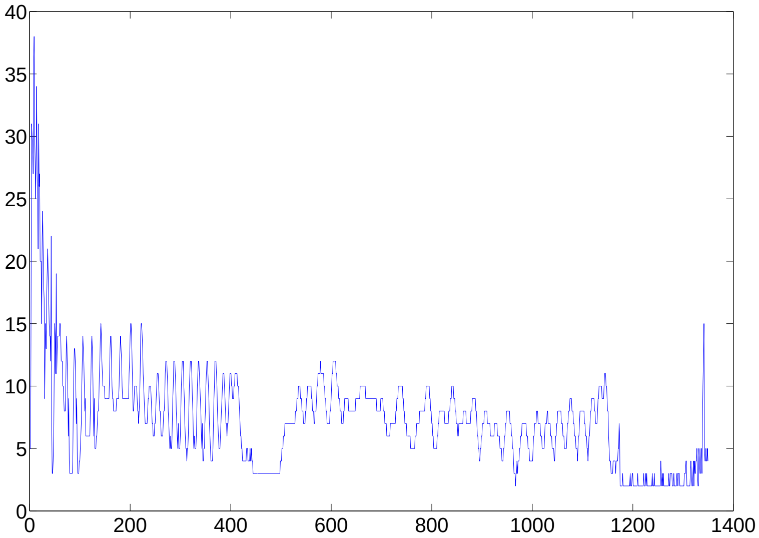


Figure 4.41: Number of iterations per time step of the Robin method with stopping criterion (4.3.4) for a relative accuracy (4.3.5) of 1%

The time evolution is also reflected by Figure 4.41 where the number of Robin iterations needed per time step is given for the stopping criterion (4.3.4). Mostly we obtain iteration numbers below 15. In the first few time steps, however, they are considerably bigger, for example at least 20 for $t \in [40, 240]$ with a maximum of 38 for $t = 100$. With regard to this fact, we point out that starting with $t = 40$ the wetting front coming from the lake already crosses the interface between the first and the second layer. (Recall that the wetting front is the border between the fully saturated and the unsaturated regime.) Furthermore, a solution from the previous time step as the initial iterate for the next time step is certainly further away from the next solution at small time steps, where we observe a fast evolution.

We remark that at $t = 1300$ the wetting front coming from the surface water reaches the third layer, and at around $t = 5000$ the wetting front reaches the bottom layer at the left boundary of the domain. At around $t = 2300$ a topology change of the previously disconnected parts of the saturated regime of Ω takes place. Starting at about $t = 10000$ the layers Ω_1 and Ω_2 are fully saturated, and so is Ω_3 at around $t = 11700$. The last peak (with the iteration number 15) in Figure 4.41 is obtained at $t = 13420$ (see Figure 4.40).

Our time step size $\tau = 10$ is chosen three orders of magnitude larger than prescribed by the CFL condition (4.2.11) for linear cases, which would require $\tau < 0.0015$ if we set h as the smallest side length of a triangle in z-direction (see Figure 4.15) and $\lambda = \lambda_4$ from Table 4.2 in (4.2.12). With a glance at Figures 4.17–4.40 and Figure 4.41, the choice of $\tau = 0.0015$ would require about 10^6 time steps for the same evolution process leading to a time resolution which seems far too fine for this problem. In addition, we would need unrealistically long calculation times of several weeks for the computation of this problem with the stopping criterion (4.3.4) on a PC. Note that we also solved our problem above (using $\tau = 10$) with the much bigger accuracy (4.3.2) which took about a day of computation while the calculation with the stopping criterion (4.3.4) only lasted a few hours.

We do observe some numerical instabilities (i.e. unrealistic physical pressure values $p < -10$) in the solution of some spatial problems in our example above. They occur in nodes of certain triangles when the wetting front crosses that area, for example at the left boundary or in the right top corner of Ω. These critical triangles contain angles which are (possibly much) bigger than $\pi/2$. We point out that we also observe such instabilities if we choose bigger mesh sizes or smaller time steps such as $\tau = 0.02$ (CFL time step for the first level). They even occur if we carry out numerical experiments with a wetting front around critical triangles using the CFL time step $\tau = 0.0015$. Moreover, we also observe such instabilities if we compute the same example without gravity. These observations suggest that the instabilities which we obtained in our model problem are not due to the gravitational term but rather due to the grid quality which we have not optimized here. We assume that such problems might require grids which satisfy the Delaunay or a similar property, see Delaunay [31] and Fuhrmann and Langmach [41].

Altogether, the nature of our example and the numerical results we obtained demonstrate that the solution method we propose for the Richards equation in heterogeneous soil with surface water can be successfully applied to a realistic hydrological model problem.

Zusammenfassung (deutsch)

Dreh- und Angelpunkt der vorliegenden Arbeit ist die Richardsgleichung. Diese ist eine nichtlineare elliptisch-parabolische partielle Differentialgleichung zur Beschreibung des Grundwasserflusses in porösen Medien im gesättigten wie im ungesättigten Fall. Eine der beiden Nichtlinearitäten, die relative Permeabilität, führt zu einer Degeneriertheit in der Ortsableitung, da sie für kleine Druckwerte beliebig klein werden kann. Weiterhin können die Parameterfunktionen große Steigungen enthalten und für extreme Bodenparameter zu Sprungfunktionen degenerieren. Da zudem die Parameterfunktionen vom Bodentyp in Teilbereichen des Rechengebiets abhängen, erhalten wir heterogene Probleme mit springenden Nichtlinearitäten.

Angesichts dieser Eigenschaften der Nichtlinearitäten streben wir eine Lösungsmethode für die Richardsgleichung an, welche völlig ohne Linearisierung auskommt. Dazu ist es nötig, das Problem in Teilprobleme zu zerlegen, welche getrennt voneinander gelöst werden können. Als erstes wird dafür der homogene Fall ortsunabhängiger Parameterfunktionen betrachtet, für den wir durch Kirchhoff–Transformation die quasilineare Richardsgleichung in eine semilineare Gleichung transformieren können, welche nun gleichmäßig elliptisch im Ort ist. Die erhaltene Gleichung wird dann implizit im Hauptteil und explizit im Gravitationsanteil der Ortsableitung in der Zeit diskretisiert. Dies führt auf Ortsprobleme, welche äquivalent zu eindeutig lösbaren konvexen Minimierungsproblemen sind. Letztere lassen eine Diskretisierung mit linearen finiten Elementen zu, für die Existenz und Eindeutigkeit sowie Konvergenz der diskreten Lösungen gezeigt werden kann. Die entstandenen diskreten Probleme können mit monotonen Mehrgitter–Methoden effizient und robust bezüglich extrem variierender Bodenparameter gelöst werden.

Im heterogenen Fall führen verschiedene Kirchhoff–Transformationen in den Teilgebieten mit homogenem Boden auf ein Gebietszerlegungsproblem. In diesem sind konvexe Minimierungsprobleme auf den Teilgebieten mit nichtlinearen Übergangsbedingungen auf den Gebietsgrenzen gekoppelt. Konkret fordern die Übergangsbedingungen die Stetigkeit des physikalischen Drucks sowie des Normalenflusses über die Gebietsgrenze. Mittels dieser Größen lassen sich nichtlineare iterative Verfahren wie die Dirichlet–Neumann– und die Robin–Methode definieren. Ohne weitere Linearisierung werden diese Verfahren durch eine Weiterentwicklung der linearen Steklov–Poincaré–Theorie analysiert. Im Falle von nichtdegenerierenden relativen Permeabilitäten auf zwei Teilgebieten im

Eindimensionalen erhalten wir folgende Konvergenzaussagen sowohl im Kontinuierlichen als auch im Diskreten. Die genügend gedämpfte Dirichlet–Neumann–Methode sowie die Robin–Methode konvergieren für die stationäre Richardsgleichung ohne Gravitation, und das zugehörige Gebietszerlegungsproblem ist wohlgestellt. Weiterhin konvergiert die Robin–Methode für die zeitdiskretisierte Richards–Gleichung, und auch hier erhalten wir die Wohlgestelltheit des Gebietszerlegungsproblems. Obwohl unsere auf einem Kontraktionsargument basierende Beweismethode im Zweidimensionalen versagt, erhalten wir befriedigende numerische Resultate für die Anwendung der Methoden auf die untersuchten Probleme in zwei Raumdimensionen.

Schließlich wird eine geeignete Upwind–Diskretisierung für den explizit zeitdiskretisierten Gravitationsterm mittels finiter Elemente entwickelt. Damit lassen sich die Ortsprobleme der zeitdiskretisierten Richardsgleichung numerisch stabil lösen, wobei Zeitschrittbeschränkungen in der Praxis akzeptabel bleiben. Ein numerisches Beispiel in zwei Dimensionen zur Lösung der Richardsgleichung mit vier verschiedenen Böden und Oberflächenwasser sowie realistischen hydrologischen Daten zeigt die Anwendbarkeit unserer Lösungsmethode.

Appendix

In this appendix we collect some basic definitions and well-known results that we use in this work. For more information we give some literature on the subjects mentioned here. In general the definitions and theorems can be applied to $\Omega \subset \mathbb{R}^d$ for any $d \in \mathbb{N}$. In Section 1.5 we specify the properties of the domains $\Omega \subset \mathbb{R}^d$ we would like to consider. Here we give the relevant definitions, one of which (C^1-polyhedron) is needed in Gauss's theorem whereas the other one (Lipschitz domain) is most prominent in the theory of Sobolev spaces.

A.1 Gauss's theorem

One central ingredient in the whole theory of partial differential equations is certainly Gauss's theorem, from which Green's formulas of partial integration are obtained. We quote this theorem from [55, p. 380] where the following conditions on the boundary $\partial\Omega$ are imposed, see [55, p. 376].

Definition A.1.1. Let Ω be an open set in $\mathbb{R}^d$.

a) A point $a \in \partial\Omega$ is called a *regular point* of $\partial\Omega$ if there is a neighbourhood $U \subset \mathbb{R}^n$ of a and a C^1-function $q : U \to \mathbb{R}$ with $\nabla q(x) \neq 0$ for all $x \in U$ such that

$$\Omega \cap U = \{x \in U : q(x) < 0\}\,.$$

An element of $\partial\Omega$ is called a *singular point* if it is not a regular point. The collection of all regular points of $\partial\Omega$ is called the *regular* or *smooth boundary* $\partial_r\Omega$ of Ω. Analogously, $\partial_s\Omega := \partial\Omega\backslash\partial_r\Omega$ is called the *singular boundary.*

b) Ω is called a *C^1-polyhedron* if its singular boundary $\partial_s\Omega$ is a $(d-1)$-nullset.

See [55, pp. 376/377] for the proof of the fact that the smooth boundary of a C^1-polyhedron is an orientable C^1-hypersurface. In this context we also refer to [55, pp. 115/116] for the definition of a (differentiable) manifold, to [55, pp. 360–369] for measurability of subsets of manifolds and the definition of a (Hausdorff) nullset, furthermore to [3, p. 13] for the $(d-1)$-dimensional Hausdorff measure of a smooth surface in $\mathbb{R}^d$.

Now we can state Gauss's theorem, also known as the divergence theorem. Later on we will encounter it again in a weak form (see (A.2.12)).

Theorem A.1.2. *Let $\Omega \subset \mathbb{R}^d$ be a bounded C^1-polyhedron and $F : \overline{\Omega} \to \mathbb{R}^d$ a vector field with the following properties:*

a) F is continuous on $\overline{\Omega}$ and continuously differentiable on Ω,

b) $\operatorname{div} F$ is integrable on Ω,

c) F is integrable on $\partial\Omega$.

Then the following holds:

$$\int_\Omega \operatorname{div} F \, dx = \int_{\partial\Omega} F \cdot \mathbf{n} \, d\sigma \, .$$

Note that property b) is satisfied if $\operatorname{div} F$ is bounded on Ω and property c) holds if $\partial\Omega$ is a measurable hyperface. As a consequence of this theorem we give the following version of partial integration in $\mathbb{R}^d$.

Theorem A.1.3. *Let $\Omega \subset \mathbb{R}^d$ be a bounded C^1-polyhedron with Hausdorff measurable $\partial\Omega$, $G : \overline{\Omega} \to \mathbb{R}^d$ a continuous vector field and $v : \overline{\Omega} \to \mathbb{R}$ a continuous scalar function, both continuously differentiable on Ω. Furthermore, let $\operatorname{div} G$ and the partial derivatives of v be bounded on Ω. Then the following holds:*

$$\int_\Omega \operatorname{div}(G(x))\, v(x)\, dx = -\int_\Omega G(x) \nabla v(x)\, dx + \int_{\partial\Omega} (G(x) \cdot \mathbf{n}(x))\, v(x)\, d\sigma(x) \, .$$

Proof. The result follows immediately by considering the vector field $F := G \cdot v$ in Theorem A.1.2 and applying the product rule $\operatorname{div}(G \cdot v) = \operatorname{div}(G)\, v + G \nabla v$. □

Note that for $G = \nabla u$ with a sufficiently smooth scalar function u we obtain Green's first formula. Instead of requiring properties a) and b) in Theorem A.1.2, one often confines oneself to vector fields with coordinate functions from the well-known space $C^1(\overline{\Omega})$ (as done in Section 1.5), which we define in the following (see [98, p. 7]). Hausdorff measurability of $\partial\Omega$ provided, Gauss's theorem clearly holds for such vector fields. In the following we define these spaces along with some other well-known function spaces.

Definition A.1.4. For any $\Omega \subset \mathbb{R}^d$ we denote by $C(\Omega)$ the space of all continuous functions $f : \Omega \to \mathbb{R}$.
If Ω is compact, then $C(\Omega)$ equipped with the norm $\|f\|_\infty := \sup_{x \in \Omega} |f(x)|$ is a Banach space.
Now let Ω be open. We call $\alpha = (\alpha_1, ..., \alpha_d) \in \mathbb{N}_0^d$ a multi-index and define $|\alpha| := \alpha_1 + \cdots + \alpha_d$. By the expression

$$D^\alpha f := \frac{\partial^{\alpha_1} \ldots \partial^{\alpha_d}}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} f$$

we abbreviate the corresponding partial derivative of a function $f : \Omega \to \mathbb{R}$ if it exists.
In case $D^\alpha f$ exists and is continuous for any multi-index α with $|\alpha| \leq k \in \mathbb{N}_0$, the function f is called k times continuously differentiable. The space of all these functions is denoted by $C^k(\Omega)$ (note $C^0(\Omega) = C(\Omega)$).
In addition, we define the space of all infinitely differentiable functions with compact support as

$$C_0^\infty(\Omega) := \{f \in \cap_{k\in\mathbb{N}} C^k(\Omega) : \operatorname{supp} f \text{ is compact}\} .$$

For open and bounded $\Omega \subset \mathbb{R}^d$ we define

$$C^k(\overline{\Omega}) := \{f \in C^k(\Omega) : D^\alpha f \text{ has a continuous extension on } \overline{\Omega} \text{ for all } |\alpha| \leq k\} .$$

Equipped with the norm

$$\|f\|_{C^k(\overline{\Omega})} := \sum_{|\alpha|\leq k} \|D^\alpha f\|_\infty$$

$C^k(\overline{\Omega})$ is a Banach space. Finally, we define the sets

$$C^\infty(\Omega) := \cap_{k\in\mathbb{N}} C^k(\Omega) \quad \text{and} \quad C^\infty(\overline{\Omega}) := \cap_{k\in\mathbb{N}} C^k(\overline{\Omega}) .$$

It makes sense not to define $C^k(\overline{\Omega})$ as the space of all k times continuously differentiable functions $f : \Omega \to \mathbb{R}$, whose one-sided partial derivatives $D^\alpha f$ exist on $\partial\Omega$. This is because there may be points in $\partial\Omega$ where, according to the shape of Ω, certain one-sided partial derivatives $D^\alpha f$ cannot even be defined (e.g. in the north pole of a circle in $\mathbb{R}^2$). However, if a one-sided partial derivative $D^\alpha f$ of an $f \in C^k(\overline{\Omega})$ can be defined in a point in $\partial\Omega$, by the mean value theorem it is equal to the extension of $D^\alpha f$ on $\overline{\Omega}$ in that point.

A.2 Sobolev spaces

For what is to come, we refer to the standard literature Adams [1], Lions and Magenes [64] and Wloka [101] as well as to Alt [3], Werner [98] and Ziemer [104]. In particular, we follow the exposition in the compendium by Brezzi and Gilardi [21] and the appendix in Quarteroni and Valli [75] according to which most of our notation is chosen.

L^p-spaces

Let $1 \leq p < \infty$ and $\Omega \subset \mathbb{R}^d$ be open and bounded. By $L^p(\Omega)$ we denote the well-known Banach space of all equivalence classes f of Lebesgue measurable functions on Ω which coincide almost everywhere in Ω and for which $|f|^p$ is Lebesgue integrable. $L^p(\Omega)$ is endowed with the norm

$$\|f\|_{L^p(\Omega)} := \left(\int_\Omega |f(x)|^p \, dx\right)^{1/p} .$$

For $p = 2$ this gives the Hilbert space $L^2(\Omega)$ with the scalar product

$$(f, g)_{L^2(\Omega)} := \int_\Omega f(x)\, g(x)\, dx\,.$$

Considering functions only except for a Lebesgue nullset, it is commonplace not to distinguish between the functions and their equivalence class. The space $L^\infty(\Omega)$ is defined as the space of all *essentially bounded* functions f on Ω with

$$\|f\|_\infty := \inf\{M > 0 : |f(x)| \leq M \text{ almost everywhere in } \Omega\} < \infty\,.$$

Now, for a multi-index α (see Definition A.1.4) and a function $f \in L^p(\Omega)$, a function $g \in L^p(\Omega)$ with the property

$$\int_\Omega g\, v\, dx = (-1)^{|\alpha|} \int_\Omega f\, D^\alpha v\, dx \quad \forall v \in C_0^\infty(\Omega) \tag{A.2.1}$$

is called the weak derivative $D^\alpha f$ of f. If such a g exists, it is unique, and if we have $f \in C^k(\overline{\Omega})$ for $|\alpha| \leq k$ and a C^1-polyhedron Ω, integration by parts (see Theorem A.1.3) gives $g = D^\alpha f$ in the classical sense of Definition A.1.4. We note that in general, by considering f as a continuous linear functional in the sense of the right hand side of (A.2.1) on the space $C_0^\infty(\Omega)$, equipped with a certain locally convex topology τ, $D^\alpha f$ can always be defined as an element of the dual $(C_0^\infty(\Omega), \tau)'$.

Sobolev spaces of natural order

For $k \in \mathbb{N}_0$ the *Sobolev space* $W^{k,p}(\Omega)$ is the space of all functions in $L^p(\Omega)$ whose weak derivatives up to the order k also belong to $L^p(\Omega)$, i.e.

$$W^{k,p}(\Omega) := \{v \in L^p(\Omega) : D^\alpha v \in L^p(\Omega) \text{ for all } |\alpha| \leq k\}\,.$$

$W^{k,p}(\Omega)$ is a Banach space with respect to the norm

$$\|v\|_{k,p} := \left(\sum_{|\alpha| \leq k} \|D^\alpha v\|^p_{L^p(\Omega)} \right)^{1/p}$$

for $1 \leq p < \infty$ and

$$\|v\|_{k,\infty} := \max_{|\alpha| \leq k} \|D^\alpha v\|_\infty$$

for $p = \infty$. If $p < \infty$ then $W^{k,p}(\Omega)$ is reflexive.

For $p = 2$ we write $H^k(\Omega)$ instead of $W^{k,2}(\Omega)$ and $\|\cdot\|_k$ instead of $\|\cdot\|_{k,2}$. $H^k(\Omega)$ is a Hilbert space with the scalar product

$$(v, w)_{H^k(\Omega)} := \sum_{|\alpha| \leq k} (D^\alpha v, D^\alpha w)_{L^2(\Omega)}\,.$$

It is crucial to have a notion of boundary values for Sobolev functions, i.e. for elements of $W^{k,p}(\Omega)$. As a first step towards that, one defines the space $W_0^{k,p}(\Omega)$

as the closure of $C_0^\infty(\Omega)$ with respect to the norm $\|\cdot\|_{k,p}$ for $1 \leq p < \infty$. For $k = 0$ we obtain $W_0^{0,p}(\Omega) = W^{0,p}(\Omega) = L^p(\Omega)$ (see [1, p. 31]), but for $k \in \mathbb{N}$ and $p > 1$, $W_0^{k,p}(\Omega)$ is always a proper subset of $W^{k,p}(\Omega)$ if $\mathbb{R}^d \backslash \Omega$ has a positive Lebesgue measure (see [1, pp. 56–63]). Again, for $p = 2$, we write $H_0^k(\Omega)$ instead of $W_0^{k,p}(\Omega)$. The most prominent case of the Hilbert spaces $H_0^k(\Omega) \subset H^k(\Omega)$ is obtained for $k = 1$.

The dual space of $W_0^{k,p}(\Omega)$ is denoted by $W^{-k,p'}(\Omega)$ with the conjugate exponent p' defined by

$$\frac{1}{p} + \frac{1}{p'} = 1$$

for $1 \leq p \leq \infty$ setting $\frac{1}{0} := \infty$ and $\frac{1}{\infty} := 0$. The reason for this notation is that every element of $W_0^{k,p}(\Omega)'$ can be interpreted as a sum of distributional derivatives up to the order k of certain functions coming from the product space $(L^{p'}(\Omega))^N$ where N is the number of multi-indices α with $0 \leq |\alpha| \leq k$ (see [1, pp. 46–51] for details). However, a similar identification for $W^{k,p}(\Omega)'$ is not possible if $W_0^{k,p}(\Omega) \neq W^{k,p}(\Omega)$ (compare [21, pp. 1.48, 1.56]). Again, for $p = 2$, we write $H^{-k}(\Omega) = H_0^k(\Omega)'$.

It has to be pointed out that the definitions of the spaces $H^k(\Omega)$ (sometimes also $H^{k,p}(\Omega)$) are by no means consistent in the literature (compare the references given at the beginning of this section). In Adams [1, p. 44] for example, $H^{k,p}(\Omega)$ is defined as the completion of the functions in $C^k(\Omega)$ with finite $\|\cdot\|_{k,p}$-norm with respect to this norm. For $1 \leq p < \infty$, this is shown to be $W^{k,p}(\Omega)$ (see [1, p. 52]). In Werner [98, p. 204], $H^k(\Omega)$ is the closure of the $C^k(\overline{\Omega})$-functions in $W^{k,p}(\Omega)$ with respect to $\|\cdot\|_{k,p}$, and in Wloka [101, pp. 96–98], $H^k(\Omega)$ is defined via the Fourier transform (see (A.2.2)). In these two cases, $H^k(\Omega)$ can be a proper subspace of $W^{k,2}(\Omega)$, however, we always have $H^k(\Omega) = W^{k,2}(\Omega)$ if the boundary $\partial\Omega$ is smooth enough (see [1, pp. 54–56]), for example to allow a continuous linear extension operator $F_\Omega : W^{k,2}(\Omega) \to W^{k,2}(\mathbb{R}^d)$ (consult [101, pp. 99/100]). For bounded $\Omega \subset \mathbb{R}^d$ a sufficient condition to obtain this is the Lipschitz boundary (see [1, pp. 66/67] and [19, p. 31]), which we define in the following.

Lipschitz boundary

Recall that a subset S of the boundary $\partial\Omega$ is called a graph of a function $f : V \subset \mathbb{R}^{d-1} \to \mathbb{R}$ if for a fixed $i \in \{1, \dots, d\}$ the i-th coordinate x_i of any $x \in S$ can be written as f applied to the other coordinates of x, considered as a vector in V.

Definition A.2.1. A set $\Omega \subset \mathbb{R}^d$ is called a *Lipschitz domain* and $\partial\Omega$ a *Lipschitz (continuous) boundary* if for any $x \in \partial\Omega$ there is a neighbourhood U_x of x such that $U_x \cap \partial\Omega$ is the graph of a scalar Lipschitz function f on an open subset of R^{d-1}. Lipschitz continuous subsets or submanifolds $\Sigma \subset \partial\Omega$ are defined accordingly.

This definition is important, in particular since we consider bounded domains $\Omega \subset \mathbb{R}^d$. For bounded Lipschitz domains, the so-called *strong local Lipschitz property* and consequently the *cone property* hold. The latter is the essential property of Ω required in the well-known embedding theorems of Sobolev [1, p. 97] and Rellich [1, p. 144]. Lipschitz continuous boundaries are also needed for the trace theorem A.2.3 and the Poincaré inequality (A.2.13) which we note below.

Real order Sobolev spaces

For the trace theorem Sobolev spaces "of fractional order" $s \in \mathbb{R}_0^+$ are needed. For $p = 2$ this can be achieved via the Fourier transform $\hat{u}$ of $u \in L^2(\mathbb{R}^d)$ by extending the result

$$H^k(\mathbb{R}^d) = \left\{ u \in L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} |\hat{u}(\xi)|^2 (1 + |\xi|^2)^k < \infty \right\} \tag{A.2.2}$$

(see [98, p. 218]) to all nonnegative reals and then by considering restrictions of these functions on $\Omega \subset \mathbb{R}^d$. Using so-called tempered distributions this can even be done for all $s \in \mathbb{R}$ (see e.g. [64, pp. 30–32] and [101, pp. 95–100]). Another abstract approach for $1 \leq p < \infty$ is to interpolate between the spaces $W^{k,p}(\Omega)$ and $W^{k+1,p}(\Omega)$ in order to get $W^{s,p}(\Omega)$ for $k < s < k+1$. If Ω is bounded and $\partial\Omega$ is smooth enough (see [1, pp. 67, 214]), $W^{s,p}(\Omega)$ with $s = k + \sigma$ and $0 < \sigma < 1$ turns out to be isomorphic to the space of all functions in $W^{k,p}(\Omega)$ with finite *Sobolev–Slobodeckij–norm*

$$\|u\|_{s,p} = \left(\|u\|_{k,p}^p + \sum_{|\alpha|=k} \int_\Omega \int_\Omega \frac{|D^\alpha u(x) - D^\alpha u(y)|^p}{|x-y|^{d+\sigma p}} \, dx \, dy \right)^{1/p} . \tag{A.2.3}$$

The spaces $W_0^{s,p}(\Omega)$ are again defined as the closure of the $C_0^\infty(\Omega)$-functions in $W^{s,p}(\Omega)$ and the duals $W^{s,p}(\Omega)'$ are called $W^{-s,p'}(\Omega)$. In case of $p = 2$ and the definition via Fourier transform, the latter is a theorem. For $p = 2$ again, Hilbert spaces $H^s(\Omega) := W^{s,2}(\Omega)$ and $H_0^s(\Omega) := W_0^{s,2}(\Omega)$ and $H^{-s}(\Omega) = W^{-s,2}(\Omega)$ are defined. It turns out that $C^\infty(\Omega)$ is dense in $H^s(\Omega)$ regardless of how regular the boundary $\partial\Omega$ is (consult [101, p. 74]). If Ω is a Lipschitz domain we also have the density of $C^\infty(\overline{\Omega})$ in $H^s(\Omega)$ (see [27, p. 114]).

Trace spaces

If $\partial\Omega$ is a sufficiently smooth hypersurface, trace spaces $W^{s,p}(\partial\Omega)$ can be defined using spaces $W^{s,p}(\tilde{\Omega})$ with $\tilde{\Omega} \subset \mathbb{R}^{d-1}$ and parameter functions constituting $\partial\Omega$ via parameter regions $\tilde{\Omega}$ (see [64, pp. 34–37] and [1, 214–217]). For polygons in $\mathbb{R}^2$ this is done in [21, pp. 1.56–1.60], where $W^{s,p}(\Sigma)$ for smooth subsets $\Sigma \subset \partial\Omega$ are defined accordingly. Finally, $W_0^{s,p}(\Sigma)$ is defined as the closure of traces $v_{|\Sigma}$ of $C^\infty(\overline{\Omega})$-functions v vanishing in a neighbourhood of $\partial\Omega\backslash\Sigma$. Now, unfortunately, for $s \geq 0$ and $1 < p < \infty$ such that $s - 1/p \in \mathbb{N}_0$, functions in

$W_0^{s,p}(\Sigma)$ do not necessarily have extensions in $W^{s,p}(\partial\Omega)$ that vanish on $\partial\Omega\backslash\Sigma$. Therefore, the space of all functions $v \in W^{s,p}(\Sigma)$ allowing trivial extensions $\tilde{v} \in W^{s,p}(\partial\Omega)$ is defined as

$$W_{00}^{s,p}(\Sigma) := \{v \in W^{s,p}(\Sigma) : \tilde{v} \in W^{s,p}(\partial\Omega)\} \tag{A.2.4}$$

with the norm

$$\|v\|_{W_{00}^{s,p}(\Sigma)} := \|\tilde{v}\|_{W^{s,p}(\partial\Omega)} \,. \tag{A.2.5}$$

If $\partial\Omega\backslash\Sigma$ has a positive Hausdorff measure, the space $W_{00}^{s,p}(\Sigma)$ is strictly contained in $W_0^{s,p}(\Sigma)$ if and only if $s \geq 0$ and $1 < p < \infty$ such that $s - 1/p \in \mathbb{N}_0$. In this case (compare [21, pp. 1.57–1.60] and [75, p. 7]) we have

$$\|v\|_{W^{s,p}(\Sigma)} \leq \|v\|_{W_{00}^{s,p}(\Sigma)} \quad \forall v \in W_{00}^{s,p}(\Sigma) \,. \tag{A.2.6}$$

In general the spaces $W^{s,p}(\Sigma)$ and $W_{00}^{s,p}(\Sigma)$ are reflexive Banach spaces for $s \geq 0$ and $1 \leq p < \infty$. In case of $W^{s,p}(\Sigma)$ the Sobolev–Slobodeckij–norm, analogously defined as in (A.2.3), provides an equivalent norm. Furthermore, we always have $W_0^{s,p}(\partial\Omega) = W^{s,p}(\partial\Omega)$, and for $s \geq 0$ and $1 \leq p < \infty$ the dual space $W_0^{s,p}(\Sigma)'$ is denoted by $W^{-s,p'}(\Sigma)$. For $p = 2$ we obtain again Hilbert spaces $H^s(\Sigma)$, $H_0^s(\Sigma)$, $H_{00}^s(\Sigma)$ and $H^{-s}(\Sigma)$.

Embedding and trace theorems

Now, with all these definitions and facts, we can state the main theorems which we use in this work. Unless otherwise stated they apply for bounded and connected $\Omega \subset \mathbb{R}^d$ with a *smooth* boundary $\partial\Omega$, i.e. a C^∞-manifold, or a polygon $\Omega \in \mathbb{R}^2$ (see [21, pp. 1.29, 1.30, 1.56]). Furthermore, we assume $\Sigma \subset \partial\Omega$ to be a connected smooth or polygonal subset of $\partial\Omega$ throughout in what is to come. First we note a version of *Sobolev's embedding theorem* (consult [21, pp. 1.52, 1.55, 1.56, 1.60]), which we need for the trace spaces $W^{s,p}(\Sigma)$. As usual, by $A \hookrightarrow B$ between two vector spaces A and B we indicate that A is naturally included in B and the corresponding linear map is continuous, i.e. a continuous *embedding*.

Theorem A.2.2. *Let $0 \leq r \leq s$, $1 < p \leq q < \infty$ and $\Omega \subset \mathbb{R}^d$, $\Sigma \subset \partial\Omega$ as above. Then $s - \frac{d-1}{p} \geq r - \frac{d-1}{q}$ implies the continuous embedding*

$$W^{s,p}(\Sigma) \hookrightarrow W^{r,q}(\Sigma)$$

and $s > \frac{d-1}{p}$ implies the compact embedding

$$W^{s,p}(\Sigma) \hookrightarrow C(\overline{\Sigma}) \,.$$

The following theorem is called *trace theorem* and gives a precise meaning to the notion of the restriction on $\Sigma \subset \partial\Omega$ of a Sobolev-function on Ω. It consists of two parts giving also an extension result and thus making clear the space of all restrictions.

Theorem A.2.3. *Let $1 \leq p \leq \infty$, $s > \frac{1}{p}$ and $s - \frac{1}{p} \notin \mathbb{N}$. Then the following holds:*

a) There is a unique continuous linear map $tr_\Sigma : W^{s,p}(\Omega) \to W^{s-1/p,p}(\Sigma)$, such that $tr_\Sigma v = v_{|\Sigma}$ for each $v \in C(\overline{\Omega}) \cap W^{s,p}(\Omega)$.

b) There is a continuous linear map $R_\Sigma : W^{s-1/p,p}(\Sigma) \to W^{s,p}(\Omega)$, such that $tr_\Sigma R_\Sigma \mu = \mu$ for each $\mu \in W^{s-1/p,p}(\Sigma)$.

In this work the trace theorem is frequently applied, mainly for $s = 1$ and $p = 2$. In this case the theorem is also valid for bounded and open Ω and Lipschitz continuous $\partial\Omega$ or Σ (see [75, p. 339]). Now, for s and p as in Theorem A.2.3, the following definition of Banach spaces of Sobolev functions vanishing on Σ is necessary for our purposes:

$$W^{s,p}_\Sigma(\Omega) := \{v \in W^{s,p}(\Omega) : tr_\Sigma v = 0\} \quad \text{and} \quad H^s_\Sigma(\Omega) := W^{s,2}_\Sigma(\Omega)\,. \tag{A.2.7}$$

It turns out that $W^{s,p}_0(\Omega) = W^{s,p}_{\partial\Omega}(\Omega)$ (see [21, p. 1.66]).

Note in Theorem A.2.3 that tr_Σ is surjective and R_Σ is injective. Consequently, the trace theorem provides a useful characterization of $W^{s-1/p,p}(\Sigma)$ as the space of all restrictions $v_{|\Sigma} := tr_\Sigma v$ of the functions $v \in W^{s,p}(\Omega)$. Furthermore, observe that by definition (A.2.4) of $W^{s,p}_{00}(\Sigma)$ we have

$$v \in W^{s,p}_{\partial\Omega\backslash\Sigma}(\Omega) \quad \Longleftrightarrow \quad v_{|\Sigma} \in W^{s,p}_{00}(\Sigma)\,. \tag{A.2.8}$$

The *trace operator* tr_Σ gives the *trace inequality*

$$\|v_{|\Sigma}\|_{W^{s-1/p,p}(\Sigma)} \leq C_1 \|v\|_{s,p} \quad \forall v \in W^{s,p}(\Omega)$$

with $C_1 = \|tr_\Sigma\|$. In addition, considering (A.2.5) and (A.2.8) it provides

$$\|v_{|\Sigma}\|_{W^{s-1/p,p}_{00}(\Sigma)} = \|\widetilde{v_{|\Sigma}}\|_{W^{s-1/p,p}(\partial\Omega)} \leq C_2 \|v\|_{s,p} \quad \forall v \in W^{s,p}_{\partial\Omega\backslash\Sigma}(\Omega) \tag{A.2.9}$$

with $C_2 = \|tr_{\partial\Omega}\|$, thus a trace inequality for $W^{s-1/p,p}_{00}(\Sigma)$, too (compare this result with (A.2.6)). However, although the *extension operator* R_Σ gives

$$\|R_\Sigma \mu\|_1 \leq C_3 \|\mu\|_{W^{s-1/p,p}(\Sigma)} \leq C_3 \|\mu\|_{W^{s-1/p,p}_{00}(\Sigma)} \quad \forall \mu \in W^{s-1/p,p}_{00}(\Sigma)$$

with $C_3 = \|R_\Sigma\|$, this does not entail the equivalence of the norms $\|\cdot\|_{W^{s-1/p,p}_{00}(\Sigma)}$ and $\|\cdot\|_{W^{s-1/p,p}(\Sigma)}$ on $W^{s-1/p,p}_{00}(\Sigma)$ if $s - 1/p \notin \mathbb{N}$. More specifically, we cannot set $\mu = v_{|\Sigma}$ in (A.2.9) and assume $\tilde{\mu} = tr_{\partial\Omega} R_\Sigma \mu$ $\forall \mu \in W^{s-1/p,p}_{00}(\Sigma)$. In fact, the latter must be false since this would provide the equivalence. But the $C^\infty(\overline{\Omega})$-functions vanishing in a neighbourhood of $\partial\Omega\backslash\Sigma$ are contained both in $W^{s-1/p,p}_0(\Sigma)$ and $W^{s-1/p,p}_{00}(\Sigma)$. Therefore, if $s - 1/p \notin \mathbb{N}_0$, then $W^{s-1/p,p}_{00}(\Sigma)$ is a non-closed dense subspace of $W^{s-1/p,p}_0(\Sigma)$.

Finally, using the inequalities just mentioned, it is easy to see that

$$\|\mu\| := \inf\{\|u\|_1 : u \in W^{s,p}_{\partial\Omega\backslash\Sigma}(\Omega),\ \mu = tr_\Sigma u\} \quad \forall \mu \in W^{s-1/p,p}_{00}(\Sigma) \tag{A.2.10}$$

provides an equivalent norm on $W^{s-1/p,p}_{00}(\Sigma)$.

Normal components in a weak sense

There is also a trace theorem for normal components which we can use for the notion of weak normal derivatives. For this purpose we define the space

$$L^p_{\operatorname{div}}(\Omega) := \{w \in (L^p(\Omega))^d : \operatorname{div} w \in L^p(\Omega)\}$$

for $1 \le p < \infty$ with the graph norm

$$\|w\|_{\operatorname{div}} := \left(\|w\|^p_{(L^p(\Omega))^d} + \|\operatorname{div} w\|^p_{L^p(\Omega)}\right)^{1/p},$$

wherein we write $\|w\|^p_{(L^p(\Omega))^d} := \sum_{i=1}^d \|w_i\|^p_{L^p(\Omega)}$ and $\operatorname{div} w = \sum_{i=1}^d \frac{\partial}{\partial x_i} w_i$ with weak derivatives $\frac{\partial}{\partial x_i} w_i \in L^p(\Omega)$ for $w = (w_1, \dots, w_d)$. In case of $p = 2$ this space is called $H(\operatorname{div} w, \Omega)$. We have $(W^{1,p}(\Omega))^d \subset L^p_{\operatorname{div}}(\Omega)$ with equality only if $d = 1$, and $(C^\infty(\overline{\Omega}))^d$ is dense in $L^p_{\operatorname{div}}(\Omega)$.

Theorem A.2.4. *Let $1 < p < \infty$. Then the following holds:*

a) There is a unique continuous linear operator $tr_n : L^p_{\operatorname{div}}(\Omega) \to W^{-1/p,p}(\partial\Omega)$ such that $tr_n w = (w \cdot \mathbf{n})_{|\partial\Omega}$ for each $w \in (C(\overline{\Omega}))^d \cap L^p_{\operatorname{div}}(\Omega)$.

b) There is a continuous linear operator $R_n : W^{-1/p,p}(\partial\Omega) \to L^p_{\operatorname{div}}(\Omega)$, such that $tr_n R_n \mu = \mu$ for each $\mu \in W^{-1/p,p}(\partial\Omega)$.

Here, as always, $\mathbf{n}$ denotes the unit normal vector on $\partial\Omega$ directed outward. We point out that $\mathbf{n}$ exists almost everywhere on $\partial\Omega$ since $\partial\Omega$ is a Lipschitz boundary (see Ciarlet [28, pp. 32–37]). As in Theorem A.2.3 above, Theorem A.2.4 also holds in case of $\Sigma \subset \partial\Omega$. Then for $p = 2$, however, we need to replace $W^{-1/p,p}(\Sigma)$ by $H^{1/2}_{00}(\Sigma)'$ which is larger than $H^{-1/2}(\Sigma)$ because $H^{1/2}_{00}(\Sigma)$ is strictly contained in $H^{1/2}(\Sigma)$ for nontrivial $\partial\Omega\backslash\Sigma$. In this case Theorem A.2.4 also holds for C^1-polyhedra Ω with a Lipschitz boundary (see [75, p. 339]).

Now, if for $u \in H^1(\Omega)$ we have $\operatorname{div}(\nabla u) \in L^2(\Omega)$ as an additional regularity condition, then $\nabla u \in H(\operatorname{div}, \Omega)$ holds and we obtain $\nabla u \cdot \mathbf{n} \in H^{-1/2}(\partial\Omega)$ or $\nabla u \cdot \mathbf{n} \in H^{1/2}_{00}(\Sigma)'$, defined according to Theorem A.2.4. If the regularity of u is even higher, one can prove more, for example there is an analogous trace theorem that assigns each $u \in H^2(\Omega)$ a unique normal derivative $\frac{\partial u}{\partial \mathbf{n}} \in H^{1/2}(\partial\Omega)$ or more general $\frac{\partial u}{\partial \mathbf{n}} \in H^{1/2}(\Sigma)$ (see [21, pp. 1.62–1.65]).

Gauss's theorem in a weak sense and Poincaré inequality

Using the trace theorems and the density of $C^\infty(\overline{\Omega})$ in $W^{1,p}(\Omega)$, one can generalize Gauss's theorem and Green's formulas of partial integration to the weak setting for C^1-polyhedra Ω with a Lipschitz boundary. In particular, we have

$$\int_\Omega \frac{\partial}{\partial x_i} w\, v\, dx = -\int_\Omega w \frac{\partial}{\partial x_i} v\, dx + \int_{\partial\Omega} tr_{\partial\Omega} w\, tr_{\partial\Omega} v\, n_i\, d\sigma(x) \quad \forall w, v \in H^1(\Omega) \tag{A.2.11}$$

for $i = 1, \ldots, d$ with $\mathbf{n} = (n_1, \ldots, n_d)$. More generally, for $1 < p < \infty$ one obtains

$$\int_\Omega \operatorname{div} w \, v \, dx = -\int_\Omega w \nabla v \, dx + \langle tr_n w, tr_{\partial\Omega} v \rangle \quad \forall w \in L^p_{\mathrm{div}}(\Omega) \ \forall v \in W^{1,p'}(\Omega) \tag{A.2.12}$$

as a generalization of Theorem A.1.3, which gives a generalization of Gauss's theorem A.1.2 by setting $v = 1$. Here, $\langle \cdot, \cdot \rangle$ denotes the duality pairing between the relevant trace spaces $W^{-1/p,p}(\partial\Omega)$ and $W^{1/p,p'}(\partial\Omega)$ which can be written as the usual boundary integral if the first entry has a representation in $L^{p'}(\Omega)$.

In the following, we state the well-known Poincaré inequality (see for example [30, pp. 127–130] and [75, p. 340]) which is used to prove coercivity of bilinear forms or equivalence of the energy-norm and the $\|\cdot\|_1$-norm in the space $H^1_\Sigma(\Omega)$. In [30, pp. 129–130] the notion of a positive *capacity* (which differs from a measure) of $\Sigma \subset \partial\Omega$ is required (see [29, p. 447]), but the proof applies for all Σ with a positive Hausdorff measure which are regular enough such that the trace theorem A.2.3 holds and $H^1_\Sigma(\Omega)$ can be defined as in (A.2.7).

Theorem A.2.5. *Assume that $\Omega \subset \mathbb{R}^d$ is bounded, connected and open and $\Sigma \subset \partial\Omega$ is a Lipschitz manifold with a positive Hausdorff measure. Then there exists a $C_\Omega > 0$ such that*

$$\int_\Omega |\nabla v(x)|^2 \, dx \geq C_\Omega \int_\Omega v^2(x) \, dx \quad \forall v \in H^1_\Sigma(\Omega) \,. \tag{A.2.13}$$

Spaces of vector-valued functions

For weak formulations of parabolic equations on a time cylinder $Q = \Omega \times (0,T)$ with $T > 0$ one usually considers function spaces on $(0,T)$ or $[0,T]$ with values in a Banach space $(V, \|\cdot\|_V)$, at least almost everywhere on $(0,T)$. Here, we give some basic definitions of the spaces that we address in Section 1.6. We refer to [21, pp. 1.67–1.75] for a survey of these spaces and to [101, pp. 364–390] for a more extended presentation of the topic.

The space of all continuous functions $v : (0,T) \to V$ is denoted by $C((0,T);V)$, the space of all k times continuously differentiable functions $v : (0,T) \to V$ (defined straightforwardly) is denoted by $C^k((0,T);V)$ for $k \in \mathbb{N}$. Analogously, the Banach spaces $C([0,T];V)$ and $C^k([0,T];V)$ are defined with norms just as for real-valued functions as well as the spaces $C^\infty((0,T);V)$, $C^\infty([0,T];V)$ and $C_0^\infty((0,T);V)$.

A function $v : (0,T) \to V$, defined almost everywhere on $(0,T)$, is called *measurable* if there is a sequence $(v_n)_{n\in\mathbb{N}} \in C([0,T];V)$ such that $v_n(t) \to v(t)$, $n \to \infty$, holds in V for almost all $t \in (0,T)$. For such measurable functions v, one can show that $\|v(\cdot)\|_V$ is Lebesgue measurable. For $1 \leq p < \infty$ we denote by $L^p(0,T;V)$ the space of all measurable functions $v : (0,T) \to V$ with

$$\|v\|_{L^p(0,T;V)} := \left(\int_0^T \|v(t)\|_V^p \right)^{1/p} < \infty \,.$$

Furthermore, $L^\infty(0,T;V)$ is the space of all measurable functions $v:(0,T)\to V$ with

$$\|v\|_{L^\infty(0,T;V)} := \|\, \|v(\cdot)\|_V \,\|_\infty < \infty\,.$$

If V is a Hilbert space with the scalar product $(\,\cdot\,,\,\cdot\,)_V$, then $L^2(0,T;V)$ is a Hilbert space with the scalar product

$$(u,v) := \int_0^T (u(t),v(t))_V\,dt \quad \forall u,v\in L^2(0,T;V)\,.$$

For functions $v\in L^1(0,T;V)$ one can define the integral $\int_0^T v(t)\,dt$ (Bochner integral, see [101, pp. 364–377]), which is an element of V satisfying

$$\left\|\int_0^T v(t)\,dt\right\|_V \le \int_0^T \|v(t)\|_V\,dt\,.$$

If V is separable, the dual space $L^p(0,T;V)'$ is naturally isometrically isomorphic to $L^{p'}(0,T;V')$.

Now, for $k\in\mathbb{N}$ and $1\le p<\infty$, the Sobolev space $W^{k,p}(0,T;V)$ is defined as the completion of $C^\infty([0,T];V)$ with respect to the norm

$$\|v\|_{W^{k,p}(0,T;V)} := \left(\sum_{j=0}^{k} \|v^{(j)}\|_{L^p(0,T;V)}\right)^{1/p}\,,$$

$v^{(j)}\in C^\infty([0,T];V)$ denoting the j-th derivative of v. The closure of the space $C_0^\infty((0,T);V)$ in $W^{k,p}(0,T;V)$ is denoted by $W_0^{k,p}(0,T;V)$. Analogous definitions are possible for indices $s\ge 0$ instead of $k\in\mathbb{N}$.

If V is separable and reflexive and $1<p<\infty$, the dual space of $W_0^{k,p'}(0,T;V')$ is denoted by $W^{-k,p}(0,T;V)$. Accordingly, for $p=2$ we define Hilbert spaces $H^1(0,T;V):=W^{k,2}(0,T;V)$, $H_0^1(0,T;V):=W_0^{k,2}(0,T;V)$ and the dual spaces $H^{-k}(0,T;V):=W^{-k,2}(0,T;V)$ of the latter.

It can be proved that for a sequence $(v_n)_{n\in\mathbb{N}}\in C^\infty([0,T];V)$ converging to $v\in W^{k,p}(0,T;V)$, the sequence of derivatives $(v_n^{(j)})_{n\in\mathbb{N}}$, for $j=1,\dots,k$, converges to a function in $L^p(0,T;V)$, which is then defined as the *weak derivative* of v. Thus, as for real-valued functions, $W^{k,p}(0,T;V)$ can be identified as the space of all functions in $L^p(0,T;V)$ for which weak derivatives $v^{(j)}$ exist in $L^p(0,T;V)$ up to the order k.

We remark that weak derivatives can also be defined in the distributional sense generalizing (A.2.1), see [101, pp. 378–382]. In particular, for parabolic problems with weak solutions $u\in L^p(0,T;V)$, $u'(=u_t)$ is no longer a function in general and needs to be defined by giving sense to a notion of partial integration in spaces of vector-valued Sobolev spaces. We define u' as the continuous linear operator $L\in W_0^{1,p'}(0,T;V')'=W^{-1,p}(0,T;V)$ given by

$$L(v) := -\int_0^T {}_{V'}\langle v'(t),u(t)\rangle_V\,dt \quad \forall v\in W_0^{1,p'}(0,T;V')\,,$$

where ${}_{V'}\langle \cdot\, , \, \cdot \rangle_V$ is the duality pairing between V' and V. It can be proved that $\|L\| \le \|u\|_{L^p(0,T;V)}$ holds.

Finally, we note that Sobolev embeddings and trace theorems can be obtained for Sobolev spaces of vector-valued functions as well. For example, we have the continuous embedding $W^{1,p}(0,T;V) \hookrightarrow C([0,T];V)$ such that in case of $u \in W^{1,p}(0,T;V)$ initial values $u(0) \in V$ make sense for parabolic problems.

List of Symbols

$a(\cdot,\cdot)$, $a_i(\cdot,\cdot)$	bilinear form (on Ω_i)	61, 145
$\langle\cdot,\cdot\rangle$	duality pairing between Λ' and Λ	147
A_s	scaling matrix	15
$b_i(\cdot,\cdot)$	nonlinear form on Ω_i	145
$\mathcal{C}_j^\nu$	coarse grid correction	109, 112
$C^k(\overline{\Omega})$	space of k times differentiable functions	244
$C_0^\infty(\Omega)$	space of smooth functions with compact support	244
χ, χ_1, χ_2	right hand side in Steklov–Poincaré equations	144, 148, 169
$\chi_\mathcal{K}$	characteristic function of $\mathcal{K}$	74
$\partial_v F$	directional derivative of the functional F	64
∂F	subdifferential of the functional F	71
$\operatorname{dom}\partial F$	domain of ∂F	71
$e(\lambda)$	exponent in Brooks–Corey functions, depending on pore size distribution factor λ	10
e_z	direction of gravity and of z-axis	14
η	interface value in trace space Λ	134, 146
f	right hand side in partial differential equation	8, 145, 164
f_N	Neumann values	31
F_w	translated map F	75
g	gravitational constant	9
$\mathcal{G}_1$, $\mathcal{G}_2$	Green operators on Ω_1, Ω_2	147
γ, γ_i, γ_{opt}	acceleration parameters in Robin's method	166, 199
γ_D	Dirichlet boundary	31
γ_N	Neumann boundary	31
γ_S	boundary of Signorini's type	31
Γ	interface $\overline{\Omega}_1 \cap \overline{\Omega}_2$	131
h_j	mesh size on refinement level $j \in \mathbb{N}_0$	83
h_p, $h_{\Gamma,p}$	weighting factor w.r.t. a node p ($\in \Gamma$)	80, 192
H_1, H_2	harmonic extension operators on Γ	147
$H^k(\Omega)$	Sobolev spaces	246
$H_0^1(\Omega)$	Sobolev space (with homogeneous Dirichlet boundary conditions on $\partial\Omega$)	246
$H_{\gamma_D}^1(\Omega)$	Sobolev space (with homogeneous Dirichlet boundary conditions on γ_D)	250

$H^{1/2}(\Gamma)$	trace space	248
$H_{00}^{1/2}(\Gamma)$	trace space (corresp. to homogeneous Dirichlet boundary conditions on $\partial\Omega$)	249
I	identity operator	171
$\mathcal{I}$	Riesz operator	173
$I_{\mathcal{S}_j}$	piecewise linear interpolation operator	84
$\mathcal{J}$	quadratic functional on a Sobolev space	66
$k_i(\cdot)$	nonlinearities related to Ω_i, $i = 1, 2$, in domain decomposition problems	144, 164
$kr(\cdot)$	relative permeability of the soil w.r.t. saturation $\theta(p) = M(u)$	9
κ, κ_Ω, κ_X	Kirchhoff transformation, as superposition operator on Ω, X	12, 43
κ_i	transformations on Λ, corresp. to Ω_i, $i = 1, 2$, in particular Kirchhoff transformations	146, 152, 165
$K(\cdot)$	permeability of the soil	9
K_h	hydraulic conductivity	9
$\mathcal{K}$	convex subset of a Sobolev space	61
$\mathcal{K}_j$	discretization of $\mathcal{K}$ in $\mathcal{S}_j$	79
$\mathcal{K}_{\gamma_D}$	translated convex set $\mathcal{K}$ with homogeneous Dirichlet boundary conditions on γ_D	77
$\mathcal{K}_j^D$	translated convex set $\mathcal{K}_j$ with homogeneous Dirichlet boundary conditions on γ_D	82
$\ell(\cdot)$	linear functional	62
λ, λ_i (> 0)	pore size distribution factor, corresp. to Ω_i (in Chapter 1 and in numerical examples)	10
λ $(\in [0, 1])$	convexity parameter (in Chapter 2)	62
λ, λ_i $(\in \Lambda)$	unknown in Steklov–Poincaré equations, corresponding to Ω_i (in Chapter 3)	143, 148, 171
λ^k, λ_i^k $(\in \Lambda)$	iterates in Dirichlet–Neumann or (altered) Robin method, corresponding to Ω_i	149, 171, 175
λ_p, $\lambda_p^{(j)}$	nodal basis function at the node $p \in \mathcal{N}_j$	79
Λ_j, Λ_j^D	nodal basis of $\mathcal{S}_j$, $\mathcal{S}_j^D$	79
Λ	trace space	134
Λ^j	discrete trace space	153, 195
$L^p(\Omega)$	space of Lebesgue integrable functions	245
μ	viscosity of water (in Chapter 1),	9
	interface value in trace space Λ (in Chapter 3)	137, 147, 165
$M = M(u)$	saturation (w.r.t. generalized pressure u)	12, 13, 21
$M_i(\cdot)$	saturation w.r.t. the generalized pressure on Ω_i	164
$\tilde{M} = \tilde{M}(u)$	multifunction corresponding to M	71
m_{u_j}	generalized discrete saturation corresp. to u_j	95
$\mathcal{M}_j$	Gauss–Seidel iteration operator	102
$n(\cdot)$	porosity of the soil	8
n_j	cardinality of $\mathcal{N}_j$	102
$\mathbf{n}$	normal on $\partial\Omega$ or $\partial\Omega_1$, directed outwards	31, 140

$\mathcal{N}_j$	set of nodes on refinement level $j \in \mathbb{N}_0$	79
$\mathcal{N}_j^D$, $\mathcal{N}_j^S$	set of nodes in γ_D, γ_S	79
$\mathcal{N}_j^\bullet(v)$	set of critical nodes of v	110
$\mathcal{N}_j^\circ(v)$	set of regular nodes of v	110
$\mathcal{N}_j^i(v)$	discrete phases of v	110
$\\|\cdot\\|_k$	norm on $H^k(\Omega)$, $k \geq 0$	40, 246
$\\|\cdot\\|_{k,\Omega_i}$	norm on $H^k(\Omega_i)$, $k = 0, 1$, $i = 1, 2$	132, 165
$\\|\cdot\\|_X$	norm on the space X	40, 249
Ω, Ω_1, Ω_2	domains in $\mathbb{R}^d$ (and subdomains)	31, 131
$\partial\Omega$	boundary of domain Ω	31
$p \ (= p(x,t))$	physical pressure in the water	9, 49, 132
p_i	physical pressure (solution) on Ω_i	137, 145, 164
p, $p_l \ (\in \mathcal{N}_j)$	node given by the triangulation $\mathcal{T}_j$	79, 102
p_0	pressure unit	14, 24, 194
p_b	bubbling pressure	10
p_r	scaling factor	15
p_D	Dirichlet boundary values for p	32, 34, 49
Φ, Ψ_l^S, Ψ_1	convex functions on $\mathbb{R}$	62, 105, 167
$\partial\Phi$, $\partial\Psi_l^S$	subdifferential of convex functions Φ, Ψ_l^S	71, 105
ϕ, ψ^S, ψ_1	convex functionals on a Sobolev space	64, 75, 167
$\partial_{v-u}\phi$, $\partial_v\psi_1$	directional derivatives of functionals ϕ, ψ_1	65, 167
$\partial\phi$, $\partial\psi^S$	subdifferentials of convex functionals ϕ, ψ^S	71, 71
ϕ_j, ψ^S, $\psi_{1,j}$	discretizations of convex functionals ϕ, ψ^S, ψ_1	80, 82, 192
$\partial\phi_j$, $\partial\psi^S$	subdifferentials of functionals ϕ_j, ψ^S on $\mathcal{S}_j$	82
Q	time cylinder $\Omega \times (0,T)$	52
R_Σ	extension operator w.r.t. $\Sigma \subset \partial\Omega$	250
R_1, R_2	extension operators related to Ω_1, Ω_2	165
ϱ	density of water	8
ρ, ρ_{opt}	convergence rate	125, 161, 198
$(\cdot,\cdot)_j$	lumped L^2-scalar product on $\mathcal{S}_j$	96
$(\cdot,\cdot)_\Gamma$	L^2-scalar product on Γ	165
$(\cdot,\cdot)_{\Omega_i}$	scalar product in $L^2(\Omega_i)$	145
$(\cdot,\cdot)_H$	scalar product in the Hilbert space H	71
S, S_1, S_2	Steklov–Poincaré operators (w.r.t. Ω_1, Ω_2)	143, 147, 171
$\mathcal{S}_j$	space of linear finite elements at level $j \in \mathbb{N}_0$	79
$\mathcal{S}_j^D$	subset of $\mathcal{S}_j$ with homogeneous Dirichlet boundary conditions on γ_D	79
$t \ (\in [0,T])$	time variable	8
$t \ (\in \mathcal{T}_j)$	triangle in $\mathcal{T}_j$	79
t_n	time step for $n \in \mathbb{N}_0$	60, 220
τ, τ_n	time step size (for $n \in \mathbb{N}_0$)	60
T	end time of evolution	31
T_{γ_1,γ_2}	operator providing Robin iteration step	172
$\mathcal{T}_{\gamma_1,\gamma_2}$, $\mathcal{T}_\gamma$	operator providing ADI iteration step	175, 176
$\mathcal{T}_j$	triangulation at refinement level $j \in \mathbb{N}_0$	79
$\theta = \theta(p)$	saturation (w.r.t. physical pressure p)	10, 16, 20
θ_m, θ_M	residual and maximal saturation	10

$\theta_i(\cdot)$	saturation w.r.t. the physical pressure on Ω_i	164
ϑ, ϑ_{opt}	damping factor in Dirichlet–Neumann method	144, 149, 162
tr_Σ	trace operator (w.r.t. $\Sigma \subset \partial\Omega$)	250
$u\ (= u(x,t))$	generalized pressure (solution)	12, 41, 54
u_i	generalized pressure (solution) on Ω_i	147, 164
u_0	unit of generalized pressure	14, 24, 194
u_c	critical generalized pressure	13
u_r	scaling factor	15
u_D	Dirichlet boundary values for u	31, 39, 54
u_j, $\tilde{u}_j$ $\bar{u}_j$	finite element solutions	81, 82, 95
u_j^ν	Gauss–Seidel or multigrid iterate	102, 108, 112
u_i^*	solution of a split off linear problem on Ω_i depending only on f	143, 147, 169
$u_i(\lambda)$	solution of an equation on Ω_i depending on a boundary value $\lambda \in \Lambda$ (also: $u_i(\eta)$, $u_i(\mu)$)	169, 177
u_i^k	domain decomposition iterate, $i = 1, 2$, $k \geq 0$	148, 166
v, w	(test) functions in some function space	36, 54, 61
v_1, v_2	test functions on Ω_1, Ω_2	137, 146, 165
v_j	(test) function in $\mathcal{S}_j$ (satisfying $v_j \rightharpoonup v$ or $v_j \to v$ in Section 2.5)	80, 85
$w\ ({}_{\vert\gamma_D} = u_D)$	Sobolev function satisfying Dirichlet boundary conditions (in Subsections 2.4.2–2.5.2)	74
$w_j\ (= I_{\mathcal{S}_j} w)$	function satisfying discrete Dirichlet boundary conditions (in Subsections 2.5.1, 2.5.2)	81, 84
w_u	generalized saturation	72
$w_{\bar{u}_j}$	discrete approximation of w_u	95
w_l^ν	intermediate iterate in a Gauss–Seidel step or in a multigrid step	102, 112, 115
$\mathbf{v}$	water flux	8, 10, 15
V, V_i, V_i^0	Sobolev spaces on Ω_i	134
$W^{k,p}(\Omega)$	Sobolev space	47, 188, 246
z	z-coordinate (in direction of gravity)	9
z_0	unit length in z-direction	14, 24, 194

Bibliography

[1] R.A. Adams. *Sobolev Spaces.* Academic Press, 3rd edition, 1980.

[2] H.W. Alt. The dam problem. In *Free boundary problems, theory and applications, Proc. interdisc. Symp.*, volume I of *Res. Notes Math. 78*, pages 52–68, Montecatini/Italy 1981, 1983.

[3] H.W. Alt. *Lineare Funktionalanalysis.* Springer, 1985.

[4] H.W. Alt and S. Luckhaus. Quasilinear elliptic–parabolic differential equations. *Math. Z.*, 183:311–341, 1983.

[5] H.W. Alt, S. Luckhaus, and A. Visintin. On nonstationary flow through porous media. *Ann. Math. Pura Appl.*, 136:303–316, 1984.

[6] J. Appell and P.P. Zabrejko. *Nonlinear superposition operators.* Cambridge University Press, 1990.

[7] T. Arbogast, M. Wheeler, and N.-Y. Zhang. A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media. *SIAM J. Numer. Anal.*, 33:1669–1687, 1996.

[8] F. Bagagiolo and A. Visintin. Porous media filtration with hysteresis. *Adv. Math. Sci. Appl.*, 14(2):379–403, 2004.

[9] V. Barbu. *Nonlinear semigroups and differential equations in Banach spaces.* Noordhoff International Publishing, 1976.

[10] P. Bastian, O. Ippisch, F. Rezanezhad, H.J. Vogel, and K. Roth. Numerical simulation and experimental studies of unsaturated water flow in heterogeneous systems. In W. Jäger, R. Rannacher, and J. Warnatz, editors, *Reactive Flows, Diffusion and Transport*, pages 579–598. Springer, 2005.

[11] P. Bastian, K. Birken, K. Johannsen, S. Lang, N. Neuß, H. Rentz-Reichert, and C. Wieners. UG – A flexible software toolbox for solving partial differential equations. *Comput. Vis. Sci.*, 1(1):27–40, 1997.

[12] P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, R. Kornhuber, M. Ohlberger, and O. Sander. A generic grid interface for parallel and adaptive scientific computing. Part II: Implementation and tests in DUNE. Preprint 404, MATHEON, 2007, Berlin. Submitted to *Computing*.

[13] J. Bear. *Dynamics of Fluids in Porous Media.* Dover Publications, 1988.

[14] R. Beck, B. Erdmann, and R. Roitzsch. Kaskade manual, version 3.0. Technical Report TR95-4, Konrad-Zuse-Zentrum (ZIB), Berlin, 1995.

[15] H. Berninger. Lipschitzräume. Diploma thesis, Freie Universität Berlin, February 2001.

[16] H. Berninger, R. Kornhuber, and O. Sander. On nonlinear Dirichlet–Neumann algorithms for jumping nonlinearities. In O.B. Widlund and D.E. Keyes, editors, *Domain Decomposition Methods in Science and Engineering XVI*, volume 55 of *LNCSE*, pages 483–490. Springer, 2007.

[17] J.F. Blowey and C.M. Elliot. The Cahn–Hilliard gradient theory for phase separation with non-smooth free energy. Part II: Numerical analysis. *Eur. J. Appl. Math.*, 3:147–179, 1992.

[18] F. Bonani and G. Ghione. On the application of the Kirchhoff transformation to the steady-state thermal analysis of semiconductor devices with temperature-dependent and piecewise inhomogeneous thermal conductivity. *Solid-St. Electron.*, 38(7):1409–1412, 1995.

[19] D. Braess. *Finite Elemente.* Springer, 2nd edition, 1997.

[20] J. Breuer. *Schnelle Randelementmethoden zur Simulation von elektrischen Wirbelstromfeldern sowie ihrer Wärmeproduktion und Kühlung.* PhD thesis, Universität Stuttgart, 2005.

[21] F. Brezzi and G. Gilardi. Functional spaces. In H. Kardestuncer and D.H. Norrie, editors, *Finite Element Handbook*, chapter 2 (part 1), pages 1.29–1.75. Springer, 1987.

[22] N. Calvo, J. Durany, and C. Vázquez. Mathematical analysis of a Stefan problem with Dirichlet–Signorini boundary conditions appearing in polythermic ice sheet modeling. *J. Math. Anal. Appl.*, 262(2):577–600, 2001.

[23] J. Carrillo and M. Chipot. The dam problem with leaky boundary conditions. *Appl. Math. Optimization*, 28(1):57–85, 1993.

[24] G. Chavent and J. Jaffré. *Dynamics of Fluids in Porous Media.* Elsevier Science, 1986.

[25] T.-C. Chen, S.-J. Hwang, and C.-Q. Chen. Nonlinear time-dependent thermoelastic response in a multilayered anisotropic medium. *J. Appl. Mech.*, 69(4):556–563, 2002.

[26] M. Chipot and A. Lyaghfouri. The dam problem for non-linear Darcy's laws and non-linear leaky boundary conditions. *Math. Methods Appl. Sci.*, 20(12):1045–1068, 1997.

[27] P.G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North–Holland, 1978.

[28] P.G. Ciarlet. *Mathematical Elasticity. Volume I: Three-Dimensional Elasticity.* North–Holland, 1988.

[29] R. Dautray and J.-L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology*, volume 1, Physical Origins and Classical Methods. Springer, 2000.

[30] R. Dautray and J.-L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology*, volume 2, Functional and Variational Methods. Springer, 2000.

[31] B.N. Delaunay. Sur la sphére vide. *Bull. Acad. Sci. URSS*, 7(6):793–800, 1934.

[32] M. Discacciati. An operator–splitting approach to non–overlapping domain decomposition methods. Technical Report 14.2004, Ecole Polytechnique Fédérale de Lausanne, Section Mathématique, May 2004.

[33] M. Discacciati. *Domain Decomposition Methods for the Coupling of Surface and Groundwater Flows.* PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2004.

[34] I. Ekeland and R. Temam. *Convex Analysis and Variational Problems.* North–Holland, 1976.

[35] C.M. Elliott, D.A. French, and F.A. Milner. A second order splitting method for the Cahn–Hilliard equation. *Numer. Math.*, 54:575–590, 1989.

[36] R. Eymard, M. Gutnic, and D. Hilhorst. The finite volume method for Richards equation. *Comput. Geosci.*, 3(3–4):259–294, 1999.

[37] R. Eymard, T. Gallouët, R. Herbin, M. Gutnic, and D. Hilhorst. Approximation by the finite volume method of an elliptic-parabolic equation arising in environmental studies. *Math. Models Methods Appl. Sci.*, 11 (9):1505–1528, 2001.

[38] P.A. Forsyth and M.C. Kropinski. Monotonicity considerations for saturated-unsaturated subsurface flow. *SIAM J. Sci. Comput.*, 18(5): 1328–1354, 1997.

[39] J. Fuhrmann. *Zur Verwendung von Mehrgitterverfahren bei der numerischen Behandlung elliptischer partieller Differentialgleichungen mit variablen Koeffizienten.* PhD thesis, TU Chemnitz–Zwickau, 1994.

[40] J. Fuhrmann. On numerical solution methods for nonlinear parabolic problems. In R. Helmig, W. Jäger, W. Kinzelbach, P. Knabner, and G. Wittum, editors, *Modeling and Computation in Environmental Sciences, First GAMM-Seminar at ICA Stuttgart*, pages 170–180. Vieweg, 1997.

[41] J. Fuhrmann and H. Langmach. Stability and existence of solutions of time-implicit finite volume schemes for viscous nonlinear conservation laws. *Appl. Numer. Math.*, 37(1–2):201–230, 2001.

[42] S. Gerbi, R. Herbin, and E. Marchand. Existence of a solution to a coupled elliptic system with a Signorini condition. *Adv. Differ. Equ.*, 4 (2):225–250, 1999.

[43] G. Gilardi. The evolution dam problem. In *Free boundary problems, Proc. Semin. Pavia 1979*, volume I, pages 209–217, 1980.

[44] D. Gilbarg and N.S. Trudinger. *Elliptic Partial Differential Equations of Second Order.* Springer, 1977.

[45] R. Glowinski. *Numerical Methods for Nonlinear Variational Problems.* Springer, 1984.

[46] K. Gustafson and T. Abe. The third boundary condition — was it Robin's? *Math. Intell.*, 20(1):63–71, 1998.

[47] K. Gustafson and T. Abe. (Victor) Gustave Robin: 1855–1897. *Math. Intell.*, 20(2):47–53, 1998.

[48] U. Hornung. Numerische Simulation von gesättigt-ungesättigten Wasserflüssen in porösen Medien. (German). In *Numer. Behandl. Differ.-Gleich. Randwertaufg., Tag. Oberwolfach*, volume 39 of *Int. Ser. Numer. Math.*, pages 214–232. Birkhäuser, 1978.

[49] U. Hornung. A finite element method for unsteady saturated/unsaturated flow in porous media. In *The mathematics of finite elements and applications III, Proc. Conf.*, pages 305–309, Brunel Univ., Uxbridge/U.K., 1979.

[50] U. Hornung. ADI-methods for nonlinear variational inequalities of evolution. In *Iterative solution of nonlinear systems of equations, Proc. Meeting Oberwolfach*, volume 953 of *Lect. Notes Math.*, pages 138–148. Springer, 1982.

[51] J.W. Jerome. *Approximation of Nonlinear Evolution Systems.* Academic Press, 1983.

[52] C. Johnson. *Numerical solution of partial differential equations by the finite element method.* Cambridge University Press, 1994.

[53] P. Knabner and E. Schneid. Numerical solution of unsteady saturated/unsaturated flow through porous media. In M. Feistauer, K. Kozel, and R. Rannacher, editors, *Numerical Modelling in Continuum Mechanics,* Part II, pages 337–343, Prag, 1997. Matfyzpress.

[54] G. Koethe. *Topologische lineare Räume*, volume I. Springer, 2nd edition, 1966.

[55] K. Königsberger. *Analysis 2*. Springer, 3rd edition, 2000.

[56] K. Königsberger. *Analysis 1*. Springer, 1990.

[57] R. Kornhuber. Nonlinear multigrid techniques. In J.F. Blowey, J.P. Coleman, and A.W. Craig, editors, *Theory and Numerics of Differential Equations*, pages 179–229. Springer, 2001.

[58] R. Kornhuber. On constrained Newton linearization and multigrid for variational inequalities. *Numer. Math.*, 91:699–721, 2002.

[59] R. Kornhuber. *Adaptive Monotone Multigrid Methods for Nonlinear Variational Problems*. Teubner, 1997.

[60] R. Kornhuber and C. Schütte. Numerik von Differentialgleichungen (Numerik II). Lecture notes, Freie Universität Berlin, 2001.

[61] R. Krause. *Monotone Multigrid Methods for Signorini's Problem with Friction*. PhD thesis, Freie Universität Berlin, 2001.

[62] D. Kröner. *Numerical Schemes for Conservation Laws*. Wiley–Teubner, 1997.

[63] G. Leoni and M. Morini. Necessary and sufficient conditions for the chain rule in $W^{1,1}_{\mathrm{loc}}(\mathbb{R}^N;\mathbb{R}^d)$ and $BV_{\mathrm{loc}}(\mathbb{R}^N;\mathbb{R}^d)$. *J. Eur. Math. Soc. (JEMS)*, 9 (2):219–252, 2007.

[64] J.-L. Lions and E. Magenes. *Non-Homogeneous Boundary Value Problems and Applications*, volume I. Springer, 1972.

[65] P.L. Lions. On the Schwarz alternating method. III: A variant for nonoverlapping subdomains. In T.F. Chan, R. Glowinski, J. Periaux, and O.B. Widlund, editors, *Domain Decomposition Methods for Partial Differential Equations, Proc. 3rd Int. Symp.*, pages 202–223. SIAM, 1990.

[66] P.L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16:964–979, 1979.

[67] E. MacCurdy, editor. *The Notebooks of Leonardo da Vinci*, volume 1. Harcourt, Brace & World, Inc., 1938.

[68] J. Mandel. A multilevel iterative method for symmetric, positive definite linear complementarity problems. *Appl. Math. Optim.*, 11:77–95, 1984.

[69] M. Marcus and V.J. Mizel. Complete characterization of functions which act, via superposition, on Sobolev spaces. *Trans. Amer. Math. Soc.*, 251: 187–218, 1979.

[70] M. Marcus and V.J. Mizel. Every superposition operator mapping one Sobolev space into another is continuous. *J. Funct. Anal.*, 33:217–229, 1979.

[71] L.D. Marini and A. Quarteroni. A relaxation procedure for domain decomposition methods using finite elements. *Numer. Math.*, 55(5):575–598, 1989.

[72] F. Otto. L^1–contraction and uniqueness for quasilinear elliptic–parabolic equations. *J. Differ. Equations*, 131(1):20–38, 1996.

[73] F. Otto. L^1–contraction and uniqueness for unstationary saturated–unsaturated porous media flow. *Adv. Math. Sci. Appl.*, 7(2):537–553, 1997.

[74] D.W. Peaceman and H.H. Rachford. The numerical solution of parabolic and elliptic differential equations. *J. Soc. Ind. Appl. Math.*, 3:28–41, 1955.

[75] A. Quarteroni and A. Valli. *Domain Decomposition Methods for Partial Differential Equations.* Oxford Science Publications, 1999.

[76] F. Radu, I.S. Pop, and P. Knabner. Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation. *SIAM J. Numer. Anal.*, 42(4):1452–1478, 2004.

[77] W.J. Rawls, L.R. Ahuja, D.L. Brakensiek, and A. Shirmohammadi. Infiltration and soil water movement. In D.R. Maidment, editor, *Handbook of Hydrology*, chapter 5. McGraw–Hill, 1993.

[78] L.A. Richards. The usefulness of capillary potential to soil-moisture and plant investigators. *Jour. Agr. Res.*, 37(12):719–742, 1928.

[79] L.A. Richards. Capillary conduction of liquids through porous mediums. *Physics*, 1:318–333, 1931.

[80] R.T. Rockafellar and R.J.-B. Wets. *Variational Analysis.* Springer, 1998.

[81] M. Růžička. *Nichtlineare Funktionalanalysis.* Springer, 2004.

[82] W. Rudin. *Real And Complex Analysis.* McGraw-Hill, 3rd. edition, 1986.

[83] T. Runst and W. Sickel. *Sobolev Spaces of Fractional Order, Nemytskij Operators, and Nonlinear Partial Differential Equations.* De Gruyter, 1996.

[84] E. Schneid, P. Knabner, and F. Radu. A priori error estimates for a mixed finite element discretization of the Richards' equation. *Numer. Math.*, 98 (2):353–370, 2004.

[85] J. Serrin and D. Varberg. A general chain rule for derivatives and the change of variables formula for the Lebesgue integral. *Am. Math. Mon.*, pages 514–520, 1969.

[86] A. Signorini. Sopra alcune questioni di elastostatica. *Atti Soc. It. Progr. Sc.*, 21(II):143–148, 1933.

[87] M. Slodička. A robust and efficient linearization scheme for doubly nonlinear and degenerate parabolic problems arising in flow in porous media. *SIAM J. Sci. Comput.*, 23(5):1593–1614, 2002.

[88] C.A. San Soucie. *Mixed finite element methods for variably saturated subsurface flow.* PhD thesis, Rice University, Houston, Texas, April 1996.

[89] D. Stalling, M. Westerhoff, and H.-C. Hege. Amira: A highly interactive system for visual data analysis. In C. Hansen and C. Johnson, editors, *The Visualization Handbook*, chapter 38, pages 749–767. Elsevier, 2005.

[90] C.J. van Duyn and L.A. Peletier. Nonstationary filtration in partially saturated porous media. *Arch. Ration. Mech. Anal.*, 78:173–198, 1982.

[91] M.T. van Genuchten. A closed–form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Sci. Soc. Am. J.*, 44:892–898, 1980.

[92] R.S. Varga. *Matrix iterative analysis.* Springer, 2nd edition, 2000.

[93] E.L. Wachspress. Optimum alternating-direction-implicit iteration parameters for a model problem. *J. Soc. Ind. Appl. Math.*, 10(2):339–350, 1962.

[94] E.L. Wachspress. Extended application of alternating direction implicit iteration model problem theory. *J. Soc. Ind. Appl. Math.*, 11(4):994–1016, 1963.

[95] E.L. Wachspress and G.J. Habetler. An alternating-direction-implicit iteration technique. *J. Soc. Ind. Appl. Math.*, 3:403–424, 1960.

[96] C. Wagner, G. Wittum, R. Fritsche, and H.-P. Haar. Diffusions–Reaktionsprobleme in ungesättigten porösen Medien. In K.-H. Hoffmann, W. Jäger, T. Lohmann, and H. Schunck, editors, *Mathematik: Schlüsseltechnologie für die Zukunft. Verbundprojekte zwischen Universität und Industrie.*, pages 243–253. Springer, 1997.

[97] W. Walter. *Analysis 2.* Springer, 5th edition, 2002.

[98] D. Werner. *Funktionalanalysis.* Springer, 5th edition, 2005.

[99] D. Werner. Partielle Differentialgleichungen. Lecture notes, 1997.

[100] G. Wittum. On the robustness of ILU smoothing. *SIAM J. Sci. Stat. Comput.*, 10(4):699–717, 1989.

[101] J. Wloka. *Partielle Differentialgleichungen.* Teubner, 1982.

[102] E. Zeidler. *Nonlinear Functional Analysis and its Applications II/B: Nonlinear Monotone Operators.* Springer, 1990.

[103] H. Zheng, D.F. Liu, C.F. Lee, and L.G. Tham. A new formulation of Signorini's type for seepage problems with free surfaces. *Int. J. Numer. Methods Eng.*, 64(1):1–16, 2005.

[104] W.P. Ziemer. *Weakly Differentiable Functions.* Springer, 1989.

Lebenslauf (deutsch)

For reasons of data privacy the online version of the dissertation does not contain the curriculum vitae.