FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Sunčica Sakić

Numerical solution of degenerate parabolic problems

Department of Numerical Mathematics

Supervisor of the doctoral thesis: Scott Congreve, Ph.D.
Study programme: Computational Mathematics
Study branch: Mathematics

Prague 2024

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date ....
Author's signature

I would like to express my deepest appreciation to my supervisor, Scott Congreve, for his continuous support, encouragement, and guidance throughout this research. My sincere gratitude goes to Prof. Vít Dolejší, who has been an unofficial supervisor to me. I sincerely appreciate his invaluable advice, availability and patience, as well as his help with the in-house code ADGFEM. I am extremely grateful to both of you for the opportunity you gave me and the immense knowledge you shared with me.

I would like to thank all at the Department of Numerical Mathematics who helped me resolve all the issues I encountered during my studies.

Thanks should also go to all at the Department of Mathematics and Informatics of the University of Novi Sad who encouraged and inspired me to pursue research.

Special thanks to all my friends who have supported me during my studies. Thank you: Mirka, for being my true friend for already fourteen years and not allowing the distance to change it; Milica, for your support and our talks, and also for showing that friends made in adulthood can grow into lifelong friendships; Katarina, for your sincere joy about any my achievement; Kristina, for welcoming me in Prague during the pandemic and being a good friend to me. I am grateful to all the people I met during this journey.

Thank you, Ivona, for helping me to embrace myself and continue to grow, personally and professionally.

I would like to thank my family and my aunt for their love and belief in me.

Finally, I am grateful to my partner for his support and care. Thank you for going through thick and thin with me and giving me a safe place to be who I am.

Also, I would like to thank all institutions that provided financial support for my research. Namely, my work was partially supported by the grant GAUK no. 28122, Research program no. PRIMUS/22/SCI/014 and the grant SVV-2023-260711 of the Charles University, and by the grant no. 20-01074S of the Czech Science Foundation.

Title: Numerical solution of degenerate parabolic problems

Author: Sunčica Sakić

Department: Department of Numerical Mathematics

Supervisor: Scott Congreve, Ph.D., Department of Numerical Mathematics

Abstract: This work is concerned with the theoretical analysis of a discontinuous Galerkin method applied to Richards' equation, one of the governing equations of porous media flow modeling. In our analysis, we admit the fast diffusion type of degeneracy, which this nonlinear degenerate parabolic equation may exhibit. Given the nonlinearities, we choose a local discontinuous Galerkin method to discretize the spatial variable. Further, we consider the resulting semidiscrete scheme, where special attention is paid to the estimation of the accumulation term that can possibly vanish. Using the continuous mathematical induction approach, we derive a priori error estimates in the L^2-norm and the jump form with respect to the spatial discretization parameter and the Hölder coefficient of the accumulation term derivative. Moreover, we present numerical experiments supporting the theoretical results. In addition, we study an anisotropic higher-order space-time discontinuous Galerkin method applied to two formulation of Richards' equation, the hydraulic head-based and the pressure head-based formulations. Namely, we examine the computational performances of this method and give a comparison of both model problems on a practial application of porous media flow.

Keywords: Richards' equation, porous media flow, degenerate parabolic equation, discontinuous Galerkin method, error estimates, *hp*-adaptation

Contents

Introduction

Porous media flow occurs in a wide range of natural and engineered systems, including groundwater flow [88], snow and soil physics [97], nuclear waste management [55], and filtration processes [6]. Modeling these flows helps to get insights into how fluids move through porous materials and interact with the surrounding environment. Porous media flow modeling also drives research and development in various fields, such as materials science, geoscience [94], and medical science [48].

To describe the flow in a variably-saturated porous media, it is common to use Richards' equation [78], which originates from the coupling of the mass conservation law and the Darcy-Buckingham law [26, 15]. This equation is a nonlinear parabolic partial differential equation (PDE) that can degenerate when the flow occurs in the transition zone between unsaturated and saturated medium. This transition region is also known as the wetting front. Moreover, in practice, it is often necessary to consider boundaries between the porous media and atmosphere; therefore, a suitable model of such boundary conditions and their numerical treatment is required. These boundary conditions are often referred to as seepage face boundary conditions (or the atmospheric conditions or the outflow boundary conditions). We also mention another porous media flow model, the two-phase flow, which roughly represents a system of two Richards' equations for two phases: a wetting phase and a non-wetting phase (usually water and air). We may say that Richards' equation model approximates the two-phase flow model.

Methods to solve Richards' equation numerically have been developing since the 1970s [56]. Since then, many methods for its spatial discretization have been proposed, namely, the finite difference method [19], conforming finite element method [67, 37], mixed finite element method [3, 77, 96, 99], finite volume method [39, 40, 63] and finite element-finite volume method [68]. We mention the multipoint flux approximation [52], a heterogeneous multiscale method [47], the relaxation scheme [50] and the lattice Boltzmann approach [45]. More recently, the discontinuous Galerkin (DG) method was used for the spatial discretization of Richards' equation in the papers [98, 34, 21, 24]. Moreover, the DG method was applied to the two-phase flow [8, 38]. Since the resulting system is stiff [34], an implicit time discretization is recommended. The lowest-order Euler method (see, e.g., [72]) and diagonally-implicit Runge-Kutta method of second or third order (see, e.g., [8]) were employed to solve the semidiscrete system. We mention the recent papers using the backward differential formula by Clément *et al.* [21] and the dual-time stepping method by Xiao *et al.* [98]. Finally, Dolejší *et al.* in [34] proposed a space-time DG (STDG) method, which discretizes both space and time variables by the DG method, to solve Richards' equation. For a more extensive overview of numerical methods for Richards' equation, we refer to [41] and [100].

The space-time discretization results in a nonlinear algebraic system that needs to be solved at each time step. The stability of a nonlinear solver that preserves the accuracy of the wetting front has been studied in [23, 71, 61, 100]. The methods to solve nonlinear algebraic system can be roughly divided into two categories: methods that uses the Jacobian matrix (e.g., Newton method) and fixed-point iteration methods that do not require the Jacobian matrix (e.g., Picard method). These two classes of methods and their modifications have been applied to solve nonlinear systems arising from Richards' equation and compared with each other. Namely, Celia *et al.* in [19] applied the Picard method to the mixed-form of Richards' equation, while in [57], it was suggested that the Picard method is not a suitable nonlinear solver for Richards' equation. Moreover, in [57] a hybrid method was proposed, which performs a few iterations of the Picard method and then switches to the Newton scheme. Casulli and Zanolli proposed a nested Newton method [18]. A quasi-Newton method, the L-scheme, was developed in [86, 74, 75] and proven unconditionally linear convergent [60]. We mention the preconditioned Newton method [11] suggested by Brenner and the parametrization technique that exploits the variable switch idea to improve the Newton method developed by Brenner and Cancès [12]. Moreover, Anderson acceleration was proposed to improve the convergence of the Picard method [95] (see also [34]).

Moving wetting fronts gives rise to the development of spatial adaptation techniques also known as h-adaptivity techniques. The enhancement of the methods is documented in [9, 59]; see also [64] for the one-dimensional Richards' equation. Moreover, high-order accuracy can be achieved using p-adaptivity, which combined with p-adaptivity gives hp-adaptivity. The hp-adaptivity has been incorporated to various methods, namely, the hp-finite element method [87, 65], the hp-local DG method [59] and the hp-adaptive STDG method [34].

The hp-adaptive STDG method discretizes space and time variables using discontinuous piecewise polynomial approximations, allowing higher-order approximation of the temporal variable, unlike the standard time integration low-order methods mentioned earlier. This method chooses the time step adaptively, allowing large time steps with sufficient accuracy. One of the most significant advantages of STDG methods is the use of unstructured grids (also nonconforming and anisotropic ones), which support mesh refinement and hp-adaptivity, yielding more accurate and efficient algorithms. Moreover, these methods are suitable for parallel implementation. Unlike the finite element method, boundary conditions are not incorporated into the definition of the approximate space, which simplifies the analysis and implementation of the method.

The anisotropic adaptive hp-STDG method proposed in [34] for the numerical solution of Richards' equation shows an excellent computational performance in terms of robustness, efficiency and accuracy. However, the supporting rigorous mathematical theory is missing, such as a priori and a posteriori analysis, the existence of the approximate solution and the convergence of the nonlinear iterative solver. In this thesis, we study a priori analysis on a class of DG methods applied to Richards' equation [24], which can be extended to the (hp-)STDG method. We mention some works on a priori analysis for various methods, namely, mixed finite element method [96], conforming finite element method [2] and adaptive DG method for two-phase flow [38]. Moreover, we study the anisotropic hp-STDG method applied to different formulations of Richards' equation.

The present work is devoted to the theoretical analysis of a local DG (LDG) method applied to Richards' equation and the numerical study of the hp-STDG method on some practical examples arising in porous media flow modeling. In particular, the organization of the thesis is the following.

In Chapter 1, we introduce the main concepts of the porous media flow modeling and Richards' equation. We define constitutive relations that shall be used later and discuss the possible degeneracies of Richards' equation.

In Chapter 2, we formulate model problems referred to as Ψ-formulation and ψ-formulation and introduce the spatial discretization by the interior penalty variant of the DG method. Moreover, we define the discretization of the domain and suitable functional spaces. Afterward, a semidiscrete solution of Richards' equation using the DG method is defined.

Chapter 3 is devoted to the numerical analysis of the LDG method applied to Richards' equation. We define the expanded mixed formulation of Richards' equation and derive the corresponding LDG method. Moreover, we derive a stability bound for the time continuous LDG scheme. Then, the a priori analysis is presented using techniques in [24], namely, the continuous mathematical induction which results in the implicit application of Gronwall's lemma commonly used in the method of lines analysis. We obtain error estimates in terms of the spatial discretization parameter and the Hölder coefficient of the water content function.

Chapter 4 contains numerical experiments supporting the previously obtained error estimates in Chapter 3.

In Chapter 5, we introduce temporal discretization using the DG method leading to a fully discrete scheme. We introduce the space-time partition and space-time dependent polynomial spaces. Moreover, we define the approximate solution using the *hp*-STDG method.

Finally, in Chapter 6, we describe the anisotropic *hp*-STDG method. We start by interpreting the *hp*-STDG scheme as a nonlinear algebraic system and defining a Newton-like method and Anderson acceleration. We present a numerical study of nonlinear solvers for Richards' equation using a numerical experiment. Moreover, we introduce the regularization of constitutive law. Lastly, we present a numerical example comparing computational performances of the anisotropic *hp*-STDG method applied to Ψ-formulation and ψ-formulation.

1. Richards' equation

In this chapter, we introduce *Richards' equation* [78], one of the governing equations for modeling porous media flows. First, we formulate laws valid in porous media flow modeling, from which Richards' equation is derived. Then, we define constitutive laws and draw attention to possible degeneracies that this equation may exhibit. Lastly, we mention boundary conditions and the balance law.

1.1 Modelling of flows in porous media

Modeling of fluid flows in a porous medium is based on the mass conservation law and the Darcy-Buckingham law [26, 15],

$$\partial_t(\rho\Phi S) + \nabla \cdot (\rho \boldsymbol{q}) = 0, \tag{1.1}$$

$$\boldsymbol{q} = -\frac{\boldsymbol{k}(S)}{\mu}\nabla(p + \rho g z), \tag{1.2}$$

where ρ is the density of the fluid, Φ is the porosity of the media, S is the saturation, $\boldsymbol{q}$ is the volume flux density, $\boldsymbol{k}$ is the permeability tensor, μ is the dinamic viscosity, p is the pressure, g is the gravity and z is the distance from the reference level. The Darcy-Buckingham law (1.2) is a common law for hydrodynamical problems that relates flow velocity to the gradient of pressure.

Furthermore, by substitution of (1.2) in (1.1) and assuming that the fluid is incompressible ($\partial_t \rho = 0$), its density is homogenuous ($\nabla\rho = 0$) and the porous media is nondeformable ($\partial_t \Phi = 0$) [89], we obtain

$$\partial_t(\Phi S) - \nabla \cdot \left(\frac{\rho g}{\mu}\boldsymbol{k}(S)\nabla\left(\frac{p}{\rho g} + z\right)\right) = 0. \tag{1.3}$$

Since the properties of the fluid and the porous medium are usually determined, this equation has two unknowns: the pressure and the saturation.

We introduce quantities often met in porous media flow modeling; namely, the water content θ, the hydraulic conductivity tensor $\mathbf{K}$, the pressure head ψ and the capillary capacity C defined as

$$\begin{aligned}\theta(S) &= \Phi S,\\ \mathbf{K}(S) &= \frac{\rho g}{\mu}\boldsymbol{k}(S),\\ \psi &= \frac{p}{\rho g},\\ C(\psi) &= \frac{\mathrm{d}\theta(\psi)}{\mathrm{d}\psi}.\end{aligned}$$

Let us note that θ and $\mathbf{K}$ depend on S; however, the water retention curve $S = p_c(p) = p_c(\rho g \psi)$, where p_c is an invertible function standing for the capillary pressure, gives the relation between S and p. This allows (1.3) to be solved with respect to p or with respect to S. Since we now have direct relations between θ and ψ, and between $\mathbf{K}$ and ψ, we can give three main formulations of Richards' equation. Namely, we define

- the pressure-based formulation

$$C(\psi)\partial_t\psi - \nabla \cdot (\mathbf{K}(\psi)\nabla(\psi + z)) = 0, \tag{1.4}$$

- the saturation-based formulation

$$\partial_t\theta - \nabla \cdot (\mathbf{D}(\theta)\nabla\theta + \mathbf{K}(\theta)\nabla z) = 0, \tag{1.5}$$

- the mixed formulation

$$\partial_t\theta(\psi) - \nabla \cdot (\mathbf{K}(\psi)\nabla(\psi + z)) = 0, \tag{1.6}$$

of the Richards' equation.

The quantity $\mathbf{D}$ (cf. (1.5)) represents the hydraulic diffusivity defined as

$$\mathbf{D} = \frac{\mathbf{K}(\theta)}{C(\theta)}.$$

We mention a modified capillary capacity

$$C(\psi) = \frac{\mathrm{d}\theta(\psi)}{\mathrm{d}\psi} + \frac{S_S}{\theta_S}\theta(\psi), \tag{1.7}$$

where S_S and θ_S are the specific aquifer storage and saturated water content, respectively. Moreover, we define the active pore volume ϑ as

$$\vartheta(\psi) = \theta(\psi) + \frac{S_S}{\theta_S}\int_{-\infty}^{\psi} \theta(s)\mathrm{d}s, \tag{1.8}$$

such that it holds

$$\partial_t\vartheta(\psi) = C(\psi)\partial_t\psi. \tag{1.9}$$

The relation (1.9) enables us to rewrite (1.4) in a divergence form as

$$\partial_t\vartheta(\psi) - \nabla \cdot (\mathbf{K}(\psi)\nabla(\psi + z)) = 0. \tag{1.10}$$

Furthermore, by introducing the new quantity, the hydraulic head

$$\Psi = \psi + z,$$

we may formulate (1.10) as in [34, 21]

$$\partial_t\vartheta(\Psi - z) - \nabla \cdot (\mathbf{K}(\Psi - z)\nabla\Psi) = 0, \tag{1.11}$$

where Ψ is the primary unknown. In the rest of the work, we shall be concerned with the formulations of Richards' equations given by the relations (1.10)–(1.11).

Table 1.1: Parameters for the van Genuchten-Mualem model.

α	n	m	θ_S	θ_r
0.8	1.2	0.167	0.55	0

1.2 Constitutive relations

In order to solve (1.4), we need to prescribe relations on θ and $\mathbf{K}$. Namely, several models have been developed depending on the hydraulic properties of the porous medium. Here, we mention two of them, the Gardner constitutive relations [44] and the van Genuchten-Mualem constitutive relations [92, 66].

The Gardener constitutive relations are given by

$$\theta(\psi) = \begin{cases} \theta_S & \text{for } \psi \geq 0 \\ \theta_R + (\theta_S - \theta_R)\exp(A\,\psi) & \text{for } \psi < 0 \end{cases}, \tag{1.12}$$

$$\mathbf{K}(\psi) = \begin{cases} K_S\mathbf{I} & \text{for } \psi \geq 0 \\ K_S\mathbf{I}\exp(A\,\psi) & \text{for } \psi < 0 \end{cases}, \tag{1.13}$$

where $\mathbf{I}$ is the identity matrix and $A > 0$, $K_S > 0$, $\theta_S > \theta_R > 0$ are material parameters. These relations are used in the Tracy problem [91] where the exact solution is given analytically; see Chapter 4.

Furthermore, to define the van Genuchten-Mualem constitutive relation, we start by rewriting the hydraulic conductivity $\mathbf{K}$ as

$$\mathbf{K}(\psi) = K_r(\psi)\mathbf{K}_S,$$

where K_r and $\mathbf{K}_S$ are the relative and saturated hydraulic conductivity, respectively. Hence, the functions θ and K_r are given by the van Genuchten-Mualem constitutive relations

$$\theta(\psi) = \begin{cases} \frac{\theta_S - \theta_r}{(1+(-\alpha\psi)^n)^m} + \theta_r, & \psi < 0, \\ \theta_S, & \psi \geq 0, \end{cases} \tag{1.14}$$

$$K_r(\psi) = \begin{cases} \frac{(1-(-\alpha\psi)^{mn}(1+(-\alpha\psi)^n)^{-m})^2}{(1+(-\alpha\psi)^n)^{m/2}}, & \psi < 0, \\ 1, & \psi \geq 0, \end{cases} \tag{1.15}$$

where θ_r is the residual water content, m and n are pore size distribution parameters and α is the inverse of the air entry value. These relations are illustrated in Fig. 1.2 with parameters specified in Table 1.1.

1.3 Degeneracies

Richards' equation is well-known due to its degeneracy. Namely, this nonlinear parabolic PDE degenerates into the elliptic one when the porous media becomes saturated, i.e., $\theta'(\psi) = 0$, $\psi \geq 0$. Also, in this case, it holds $\vartheta'(\psi) = 0$ since $S_S = 0$. This type of degeneracy is also known as *fast-diffusion*. Moreover, when $\psi \to -\infty$, the water content change becomes constant and $\mathbf{K}(\psi) \to 0$, which results in *slow-diffusion* type of degeneracy. Thus, the numerical treatment of this PDE is rather complicated.

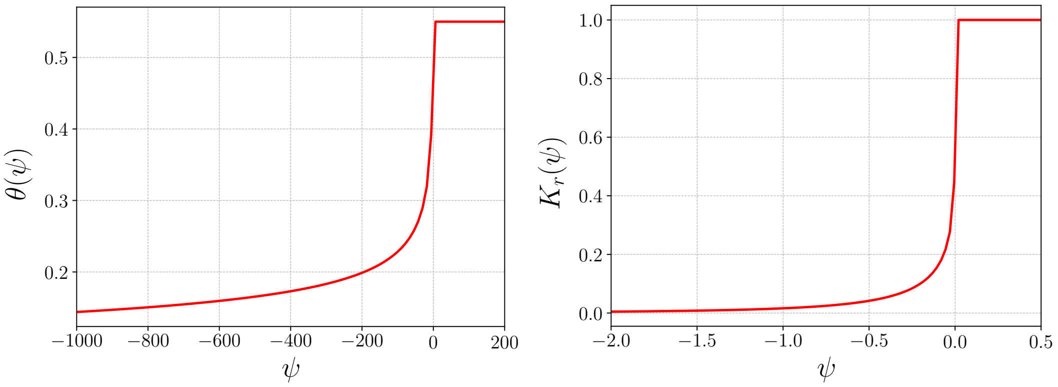

Figure 1.1: Water content and relative hydraulic conductivity curves for the van Genuchten-Mualem model.

Remark 1. We use the notation $\partial_t \vartheta(\psi) := \frac{\partial \vartheta(\psi(\boldsymbol{x},t))}{\partial t}$ for the partial derivative of $\vartheta(\psi)$ with respect to t and $\vartheta'(\psi) = \frac{\mathrm{d}\vartheta(\psi)}{\mathrm{d}\psi}$ for the derivative of ϑ with respect to ψ.

If ψ is chosen to be the primary variable (cf. (1.4)) for solving of variably-saturated porous media flows, then the Jacobian matrix is ill-conditioned around $\psi \approx 0$ since the nonlinear system contains $\theta'(\psi)$ and $\mathbf{K}'(\psi)$ (cf. Fig. 1.3). On the other hand, θ can be used as a primary variable (cf. (1.5)) only in dry regions because the saturated region cannot be described (cf. (1.14)). Several strategies have been found to overcome this issue. The methods based on the variable switch can be found in [29, 43]. More recently, this idea was extended to the parametrization method developed by Brenner and Cancès [12]. This approach seems to work successfully for the schemes where the Kirchhoff transformation is used; however, for original pressure formulation, it meets difficulties caused by vanishing diffusion, and therefore, some additional regularizations are required [7].

The Kirchhoff transformation introduces a new quantity, the global pressure

$$U(\psi) := \int_0^{\psi} \mathbf{K}(\chi)\mathrm{d}\chi,$$

which is more regular than ψ; however, it has no particular physical meaning [16]. Moreover, due to its definition, it is not always possible to derive it analytically, e.g., for the van Genuchten-Mualem model [7]. Nevertheless, this transform is advantageous for mathematical analysis [3, 96], as well as in computation since it linearizes the higher order term ($\nabla U = \mathbf{K}(\psi)\nabla\psi$) avoiding K_r' to blow up around $\psi \approx 0$.

1.4 Boundary conditions

Besides Dirichlet and Neumann boundary conditions, in porous media flow modeling, it is common to prescribe the seepage face boundary conditions. These conditions model interface between a porous medium and the open space [83, 71].

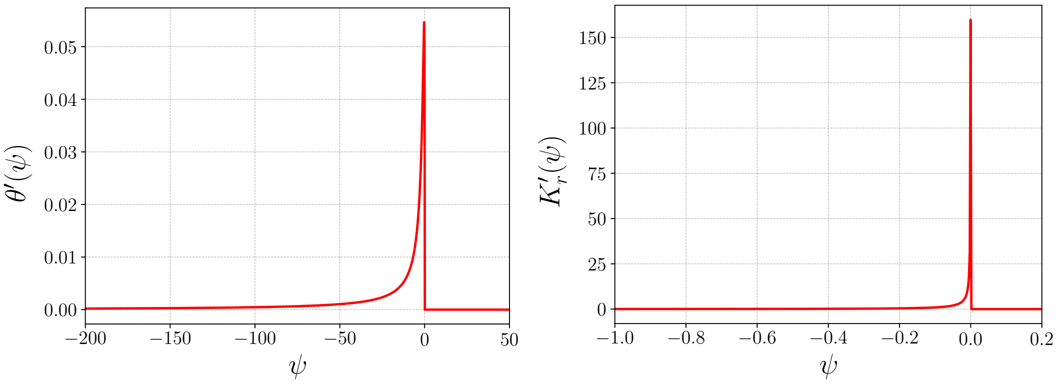

Figure 1.2: Derivatives of water content and relative hydraulic conductivity functions for the van Genuchten-Mualem model.

Namely, if the porous media is saturated, then $\psi = 0$ (or $\Psi = z$); otherwise, there is no flux.

Sometimes, they are called outflow boundary conditions [84], which assumes that the pressure head cannot be positive, fluid cannot enter the medium, and fluid exits only if the pressure is zero, or mathematically, it holds

$$\begin{aligned} \psi = \Psi - z &\leq 0, \\ -\mathbf{K}(\psi)\nabla\Psi \cdot \boldsymbol{n} &\geq 0, \\ \psi(\nabla\Psi \cdot \boldsymbol{n}) &= 0, \end{aligned}$$

where $\boldsymbol{n}$ is a unit normal.

Alternatively, this type of boundary condition can be considered as nonlinear Robin boundary conditions [84, 58, 73]

$$\mathbf{1}_E(\Psi)\psi = (1 - \mathbf{1}_E(\Psi))\mathbf{K}(\psi)\nabla\Psi \cdot \boldsymbol{n},$$

where $\mathbf{1}_E$ is the indicator function taking value 1 if $\psi \geq 0$ and $-\mathbf{K}(\psi)\nabla\Psi \cdot \boldsymbol{n} > 0$ or 0 otherwise.

In [34, 85, 53], the seepage boundary conditions are treated numerically as a switch between Dirichlet and Neumann boundary conditions

$$\psi = 0, \qquad \text{if } \psi \geq 0 \text{ and } -\mathbf{K}(\psi)\nabla\Psi \cdot \boldsymbol{n} > 0, \tag{1.16a}$$

$$-\mathbf{K}(\psi)\nabla\Psi \cdot \boldsymbol{n} = 0, \qquad \text{otherwise.} \tag{1.16b}$$

1.5 Balance of the water content

If we integrate (1.10) (or equivalently (1.11)) over a space-time domain $\Omega \times (0, T)$, $T > 0$ and use Green's theorem we obtain the *balance of the water content* given by

$$\Delta Q(t) - F(t) = 0, \tag{1.17}$$

where

$$\Delta Q(t) = Q(t) - Q(0), \tag{1.18}$$

and

$$Q(t) = \int_\Omega \vartheta(\Psi(x,t) - z)\mathrm{d}x = \int_\Omega \vartheta(\psi(x,t))\mathrm{d}x \tag{1.19}$$

is the water content at time $t \in [0, T]$ and

$$\begin{align} F(t) &= \int_0^t \int_{\partial\Omega} \mathbf{K}(\Psi(x,s) - z)\nabla\Psi(x,s) \cdot \boldsymbol{n}\,\mathrm{d}S\mathrm{d}s \tag{1.20} \\ &= \int_0^t \int_{\partial\Omega} \mathbf{K}(\psi(x,s))\nabla(\psi(x,s) + z) \cdot \boldsymbol{n}\,\mathrm{d}S\mathrm{d}s \tag{1.21} \end{align}$$

is the boundary flux on the interval $(0,t)$ and $\partial\Omega$ is the boundary of Ω.

The violation of the balance of the water content (1.17) has been studied in the literature; see, e.g., [19, 90]. In Chapter 6 we shall examine the violation of this law on a numerical example.

2. Spatial discretization

Within this chapter, we introduce the space semidiscretization of Richards' equation using the DG method. We start by defining two formulations of the problem depending on which primary variable is chosen: the hydraulic head or the pressure head. Afterward, we define corresponding time-continuous numerical schemes for both formulations.

2.1 Continuous problem

Let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain with Lipschitz-continuous boundary $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$, $\partial\Omega_D \cap \partial\Omega_N = \emptyset$. Let $T > 0$ and denote $Q_T := \Omega \times (0, T)$. We define two formulations of the Richards equation, the Ψ-formulation and the ψ-formulation, i.e., the hydraulic head and the pressure head based formulations, respectively, both already briefly mentioned in Chapter 1 (cf. (1.10)–(1.11)). Namely, we shall consider the following nonlinear problems with initial and mixed Dirichlet-Neumann boundary conditions:

- Ψ-formulation: Find $\Psi : Q_T \to \mathbb{R}$ such that

$$\partial_t \vartheta(\Psi - z) - \nabla \cdot (\mathbf{K}(\Psi - z)\nabla\Psi) = g \qquad \text{in } Q_T, \tag{2.1a}$$

$$\Psi\Big|_{\partial\Omega_D \times (0,T)} = \Psi_D, \tag{2.1b}$$

$$\mathbf{K}(\Psi - z)\nabla\Psi \cdot \boldsymbol{n}\Big|_{\partial\Omega_N \times (0,T)} = g_N, \tag{2.1c}$$

$$\Psi(\boldsymbol{x}, 0) = \Psi_0(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \tag{2.1d}$$

- ψ-formulation: Find $\psi : Q_T \to \mathbb{R}$ such that

$$\partial_t \vartheta(\psi) - \nabla \cdot (\mathbf{K}(\psi)\nabla(\psi + z)) = g \qquad \text{in } Q_T, \tag{2.2a}$$

$$\psi\Big|_{\partial\Omega_D \times (0,T)} = \psi_D, \tag{2.2b}$$

$$\mathbf{K}(\psi)\nabla(\psi + z) \cdot \boldsymbol{n}\Big|_{\partial\Omega_N \times (0,T)} = g_N, \tag{2.2c}$$

$$\psi(\boldsymbol{x}, 0) = \psi_0(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega. \tag{2.2d}$$

Here, we denote $g : Q_T \to \mathbb{R}$ as the source or sink term, $\Psi_D : \partial\Omega_D \times (0, T) \to \mathbb{R}$, $g_N : \partial\Omega_N \times (0, T) \to \mathbb{R}$ and $\Psi_0 : \Omega \to \mathbb{R}$ are the boundary and initial conditions functions, respectively. Similarly, we denote the corresponding boundary and initial functions for the pressure head, ψ_D and ψ_0. We use the notation $\boldsymbol{n} = (n_1, n_2)$ for a unit outer normal to $\partial\Omega$, and z is the vertical component such that $\boldsymbol{x} = (x_1, x_2) = (x_1, z)$. We recall the relation between the hydraulic and pressure head $\Psi = \psi + z$. The nonlinear tensor $\mathbf{K} : \mathbb{R} \to \mathbb{R}^{2\times2}$ is the conductivity tensor, and the nonlinear function $\vartheta : \mathbb{R} \to \mathbb{R}_0^+$ (the active pore volume) represents the nonlinear change of $\Psi - z = \psi$.

We assume the following [34, 77]; cf. Subsection 3.1.1.

(H1) The function $\vartheta : \mathbb{R} \to \mathbb{R}$, $\vartheta(\psi) = \vartheta(\Psi - z)$ is Hölder continuous and monotone nondecreasing, cf. the assumption (A4). Moreover, if $S_S = 0$ and $\psi > 0$ then $\vartheta'(\psi) = \vartheta'(\Psi - z) = 0$, which implies the fast-diffusion type of degeneracy of (2.1a) and (2.2a); cf. Section 1.3.

(H2) The function $\mathbf{K} : \mathbb{R} \to \mathbb{R}^{2\times 2}$ is a positive, nondecreasing and Lipschitz continuous function that can vanish when $\psi \to -\infty$ producing the slow-diffusion type of degeneracy.

(H3) The source function, and the initial and boundary data are L^2-integrable over their domain of definition.

2.1.1 Weak solution

We introduce the weak solution of (2.2); analogously, the weak solution of (2.1) can be defined. For theory on its existence and uniqueness, we refer to [1, 69, 70]. Prior to it, we establish some notation and recall dual spaces.

We shall use the standard notation $L^p(\Omega)$, $1 \leq p \leq \infty$ for the Lebesgue space of p-integrable functions on $\Omega \subset \mathbb{R}^2$. By $W^{k,p}(\Omega)$ we denote the Sobolev spaces of order k over Ω, which stand for the spaces of all functions from the space $L^p(\Omega)$ whose distributional derivatives up to order k belong to $L^p(\Omega)$. For $p = 2$, the Sobolev space $W^{k,2}(\Omega)$ is a Hilbert space, which we denote by $H^k(\Omega)$. Furthermore, we shall use the Bochner spaces $L^\infty(0, T; X)$ and $L^2(0, T; X)$, where X is a Banach space, standing for essentially bounded and square-integrable functions over the interval $[0, T]$ with values in X, repectively. Also, we use the spaces of continuous and continuously differentiable functions over $[0, T]$ with values from X denoted by $C([0, T]; X)$ and $C^1([0, T]; X)$, respectively. If B is a Banach space, then its norm is denoted by $\|\cdot\|_B$. Moreover, we use the notation $(\cdot\ , \cdot)$ for the classical scalar product in $L^2(\Omega)$, and by $(\cdot\ , \cdot)_N$ the face integral over the Neumann part of the boundary $\partial\Omega_N$. By $v(t)$ we refer to the function defined on Ω such that $v(t)(\boldsymbol{x}) = v(\boldsymbol{x}, t)$, $\boldsymbol{x} \in \Omega$.

We assume that ψ_D is a trace of some $\psi_D^* \in C([0, T]; H^1(\Omega)) \cap L^\infty(Q_T)$ on $\partial\Omega_D \times (0, T)$. We define the space

$$H^1_{0D}(\Omega) = \{v \in H^1(\Omega) : v|_{\partial\Omega_D} = 0\},$$

and denote its dual by $H^{-1}(\Omega)$.

Definition 1. *A function ψ such that*

$$\begin{aligned} \psi - \psi_D^* &\in L^2(0, T; H^1_{0D}(\Omega)), \\ \vartheta(\psi) &\in L^\infty(0, T; L^1(\Omega)) \cap H^1(0, T; H^{-1}(\Omega)), \end{aligned}$$

is called the weak solution of problem (2.2)*, if it satisfies the condition*

$$(\partial_t\vartheta(\psi), v) + (\mathbf{K}(\theta(\psi))\nabla(\psi + z), \nabla v) = (g(t), v) + (g_N(t), v)_N,$$

for all $v \in H^1_{0D}(\Omega)$, a.e. in $(0, T)$ and $\psi(0) = \psi_0$ in Ω.

2.2 Partition of the domain

Let $\mathcal{T}_h$, $h > 0$ be a partition of Ω such that

$$\Omega = \bigcup_{K\in\mathcal{T}_h} K,$$

with the spatial parameters

$$h = \max_{K \in \mathcal{T}_h} h_K, \quad h_K = \operatorname{diam}(K),\ K \in \mathcal{T}_h.$$

We denote by $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$, $\bar{h} > 0$ a family of triangulations of the domain Ω. Moreover, by $\mathcal{F}_h$ we denote the set of all edges of all elements $K \in \mathcal{T}_h$, particularly,

$$\mathcal{F}_h = \mathcal{F}_h^I \cup \mathcal{F}_h^B, \quad \mathcal{F}_h^B = \mathcal{F}_h^D \cup \mathcal{F}_h^N \quad \text{and} \quad \mathcal{F}_h^{ID} = \mathcal{F}_h^I \cup \mathcal{F}_h^D,$$

where $\mathcal{F}_h^I$ and $\mathcal{F}_h^B$ are the inner and boundary edges, respectively; additionally, $\mathcal{F}_h^D$ and $\mathcal{F}_h^N$ are edges on the boundary $\partial\Omega_D$ and $\partial\Omega_N$, respectively.

Let $\Gamma \in \mathcal{F}_h$, then we denote by $K_\Gamma^{(L)}$ and $K_\Gamma^{(R)}$ the neighboring elements such that $\Gamma \subset K_\Gamma^{(L)} \cap K_\Gamma^{(R)}$. We define the orientation of a unit normal to the edge Γ, $\boldsymbol{n}_\Gamma$, as the outer normal to $\partial K_\Gamma^{(L)}$ and the inner normal to $\partial K_\Gamma^{(R)}$. If $\Gamma \in \mathcal{F}_h^B$, we define $K_\Gamma^{(L)}$ as the adjacent element to Γ. Moreover, by $v_\Gamma^{(L)}$ and $v_\Gamma^{(R)}$, we denote the trace of $v|_{K_\Gamma^{(L)}}$ and $v|_{K_\Gamma^{(R)}}$ on Γ, respectively. If $\Gamma \in \mathcal{F}_h^I$, then we define the average and the jump on Γ as

$$\langle v \rangle_\Gamma = \frac{1}{2}\left(v_\Gamma^{(L)} + v_\Gamma^{(R)}\right), \quad [v]_\Gamma = v_\Gamma^{(L)} - v_\Gamma^{(R)}, \tag{2.3}$$

respectively. Especially, if $\Gamma \in \mathcal{F}_h^B$ and $K_\Gamma^{(L)}$ is such that $\Gamma \subset \partial K_\Gamma^{(L)} \cap \partial\Omega$, then we have

$$\langle v \rangle_\Gamma = [v]_\Gamma = v_\Gamma^{(L)}.$$

In the sequel, we omit the subscript Γ when there is no chance of confusion.

We mention some important concepts on meshes. A family of triangulations $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$, $\bar{h} > 0$

- is *shape-regular* if there exists a positive constant C_R such that

$$h_K \le C_R \rho_K \quad \forall K \in \mathcal{T}_h \quad \forall h \in (0, \bar{h}), \tag{2.4}$$

- is *quasi-uniform* if there exists a positive constant C_U such that

$$h \le C_U h_K \quad \forall K \in \mathcal{T}_h \quad \forall h \in (0, \bar{h}), \tag{2.5}$$

- satisfies the *equivalence condition* if there exist $C_T, C_G > 0$ such that

$$C_T h_K \le d(\Gamma) \le C_G h_K,\ \forall K \in \mathcal{T}_h\ \forall \Gamma \in \mathcal{F}_h\ \Gamma \subset \partial K\ \forall h \in (0, \bar{h}), \tag{2.6}$$

where ρ_K denotes the radius of the maximal inscribed two-dimensional ball in the element $K \in \mathcal{T}_h$ and $d(\Gamma) = \operatorname{diam}(\Gamma)$. We note that for a conforming triangulation $\mathcal{T}_h$ if the shape-regularity condition is satisfied then the equivalence condition holds.

2.3 Function spaces

Over a triangulation $\mathcal{T}_h$, for each $k \in \mathbb{N}$, we define the broken Sobolev space of scalar functions

$$H^k(\Omega, \mathcal{T}_h) = \{v \in L^2(\Omega) : v|_K \in H^k(K) \quad \forall K \in \mathcal{T}_h\},$$

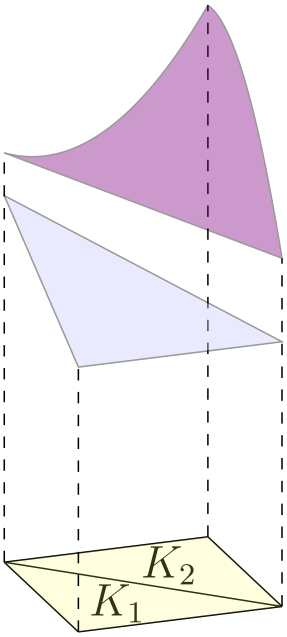

Figure 2.1: An illustration of $S_{h,p}$ spaces.

equipped with the seminorm

$$|v|_{H^k(\Omega,\mathcal{T}_h)} = \left(\sum_{K\in\mathcal{T}_h} |v|^2_{H^k(K)} \right)^{1/2},$$

where $|\cdot|_{H^k(K)}$ denotes the seminorm in the space $H^k(K)$.

The semidiscrete solution to the problem (3.1) is sought in the functional space of discontinuous polynomials of degree $p \geq 1$

$$S_{h,p} = \{v \in L^2(\Omega) : \ v|_K \in P_p(K) \ \forall K \in \mathcal{T}_h\}, \tag{2.7}$$

where $P_p(K)$ refers to the space of all polynomials defined on element K of total degree $\leq p$ (see Fig. 2.3).

For each element $K \in \mathcal{T}_h$ we denote by $\pi_{K,p}$ the L^2-projection of some $v \in L^2(K)$ to the space $P_p(K)$,

$$\int_K (\pi_{K,p}v - v)\varphi \,\mathrm{d}x = 0 \quad \forall \varphi \in P_p(K). \tag{2.8}$$

Thus, for $v \in L^2(\Omega)$ we introduce the $S_{h,p}$-interpolant $\Pi_{h,p}$ as

$$(\Pi_{h,p}v)|_K := \pi_{K,p}(v|_K) \ \ \forall K \in \mathcal{T}_h,$$

or equivalently,

$$(\Pi_{h,p}v - v, \varphi) = 0 \ \ \forall \varphi \in S_{h,p}.$$

Furthermore, we shall need the standard approximation property later in the analysis.

Lemma 1 (Approximation properties [33, Lemma 2.24, Eq. (4.100)])**.** *Let the assumption* (2.4) *be satisfied. Then, for any $v \in H^s(K)$, $K \in \mathcal{T}_h$, $h \in (0, \bar{h})$ there exists a positive constant C_A such that*

$$|\pi_{K,p}v - v|_{H^q(K)} \le C_A h_K^{\mu-q} |v|_{H^\mu(K)}, \; q = 0, 1, \tag{2.9a}$$

$$\|\partial_t(\pi_{K,p}v - v)\|_{L^2(K)} \le C_A h_K^{\mu} |\partial_t v|_{H^\mu(K)}, \tag{2.9b}$$

where $C_A > 0$, $\pi_{K,p}$ is the orthogonal L^2-projection defined by (2.8) *and $\mu = \min(p+1, s)$.*

2.4 Discretization of the problem

We multiply (2.1a) by $v \in H^1(\Omega, \mathcal{T}_h)$, integrate over $K \in \mathcal{T}_h$ and use Green's theorem. After summing over all elements $K \in \mathcal{T}_h$, we obtain the identity

$$\begin{aligned}
&\sum_{K\in\mathcal{T}_h} \int_K \partial_t \vartheta(\Psi - z) v \, \mathrm{d}x - \sum_{K\in\mathcal{T}_h} \int_{\partial K} \mathbf{K}(\Psi - z)\nabla\Psi \cdot \boldsymbol{n} v \mathrm{d}S \\
&+ \sum_{K\in\mathcal{T}_h} \int_K \mathbf{K}(\Psi - z)\nabla\Psi \cdot \nabla v \, \mathrm{d}x \\
&= \sum_{K\in\mathcal{T}_h} \int_K g v \, \mathrm{d}x.
\end{aligned} \tag{2.10}$$

We rewrite the sum of the edge integrals arising from the diffusive term as

$$\begin{aligned}
&\sum_{K\in\mathcal{T}_h} \int_{\partial K} \mathbf{K}(\Psi - z)\nabla\Psi \cdot \boldsymbol{n} v \mathrm{d}S \\
&= \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma \mathbf{K}(\Psi - z)\nabla\Psi \cdot \boldsymbol{n} v \mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^N} \int_\Gamma \mathbf{K}(\Psi - z)\nabla\Psi \cdot \boldsymbol{n} v \mathrm{d}S \\
&+ \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma [\mathbf{K}(\Psi - z)\nabla\Psi v] \cdot \boldsymbol{n} \mathrm{d}S.
\end{aligned} \tag{2.11}$$

Furthermore, we may replace the middle term on the right-hand side of (2.11) with the Neumann boundary condition function

$$\sum_{\Gamma\in\mathcal{F}_h^N} \int_\Gamma \mathbf{K}(\Psi - z)\nabla\Psi \cdot \boldsymbol{n} v \mathrm{d}S = (g_N, v)_N. \tag{2.12}$$

Assuming $\Psi(\cdot, t) \in H^2(\Omega)$, we have that

$$[\mathbf{K}(\Psi - z)\nabla\Psi] = 0,$$

and

$$\mathbf{K}(\Psi^{(L)} - z)\nabla\Psi^{(L)} = \mathbf{K}(\Psi^{(R)} - z)\nabla\Psi^{(R)} = \langle \mathbf{K}(\Psi - z)\nabla\Psi \rangle, \; \Gamma \in \mathcal{F}_h^I,$$

which implies that

$$[\mathbf{K}(\Psi - z)\nabla\Psi v] \cdot \boldsymbol{n} = \langle \mathbf{K}(\Psi - z)\nabla\Psi \rangle \cdot \boldsymbol{n}[v], \; \Gamma \in \mathcal{F}_h^I. \tag{2.13}$$

Then, the relation (2.11) becomes

$$\begin{aligned} &\sum_{K\in\mathcal{T}_h}\int_{\partial K}\mathbf{K}(\Psi - z)\nabla\Psi\cdot\boldsymbol{n}v\mathrm{d}S \\ &= \sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma\langle\mathbf{K}(\Psi - z)\nabla\Psi)\rangle\cdot\boldsymbol{n}[v]\mathrm{d}S + (g_N, v)_N. \end{aligned} \tag{2.14}$$

We introduce the interior and boundary penalty bilinear form

$$J_h(\Psi, v) = \sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma\kappa[\Psi][v]\mathrm{d}S, \tag{2.15}$$

where κ is the penalty parameter given by

$$\kappa|_\Gamma = \frac{C_W}{d(\Gamma)}, \quad \Gamma\in\mathcal{F}_h^{ID}, \tag{2.16}$$

with penalization constant $C_W > 0$.

Remark 2. In case of nonconforming triangulations with hanging nodes, it is necessary to define differently $d(\Gamma)$ in the penalty parameter given in (2.16); e.g., if $\Gamma\subset K_\Gamma^{(L)}\cap K_\Gamma^{(R)}$, $\Gamma\in\mathcal{F}_h^{ID}$, then [33]

$$d(\Gamma) = \begin{cases}\max(h_{K_\Gamma^{(L)}}, h_{K_\Gamma^{(R)}}), & \Gamma\in\mathcal{F}_h^I, \\ h_{K_\Gamma^{(L)}}, & \Gamma\in\mathcal{F}_h^D.\end{cases}$$

Finally, we define the forms as

$$\begin{aligned} a_h(\psi;\Psi, v) &= \tilde{a}_h(\psi;\Psi, v) + J_h(\Psi, v), && (2.17)\\ \tilde{a}_h(\psi;\Psi, v) &= \sum_{K\in\mathcal{T}_h}\int_K\mathbf{K}(\psi)\nabla\Psi\cdot\nabla v\,\mathrm{d}x \\ &\quad - \sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma\Big(\langle\mathbf{K}(\psi)\nabla\Psi\rangle\cdot\boldsymbol{n}[v] \\ &\quad + \Theta\,\langle\mathbf{K}(v)\nabla v\rangle\cdot\boldsymbol{n}[\Psi]\Big)\mathrm{d}S, && (2.18)\\ \ell_h(v) &= (g, v) + (g_N, v)_N - \Theta\sum_{\Gamma\in\mathcal{F}_h^D}\int_\Gamma\boldsymbol{n}\cdot\nabla v\Psi_D\mathrm{d}S. && (2.19)\end{aligned}$$

Here, for $\Theta = -1$, $\Theta = 0$ and $\Theta = 1$, the form a_h represents the nonsymmetric variant (NIPG), incomplete variant (IIPG), and symmetric variant (SIPG), respectively, of the diffusive form.

Definition 2. *We say that* $\Psi_h\in C([0, T]; S_{h,p})$ *is the semidiscrete approximate solution to* (2.1) *obtained by the DG method if*

$$(\partial_t\vartheta(\Psi_h - z), v_h) + a_h(\Psi_h - z;\Psi_h, v_h) = \ell_h(v_h) \qquad \forall v_h\in S_{h,p}\ \forall t\in(0, T), \tag{2.20}$$

$$(\Psi_h(0), v_h) = (\Psi_0, v_h) \quad \forall v_h\in S_{h,p}. \tag{2.21}$$

In a similar manner, the DG semidiscretization can be formulated for the ψ-formulation of Richards' equation. However, now in the diffusive part instead of $\nabla\Psi$ appears $\nabla(\psi + z) = \nabla\psi + \boldsymbol{e}_2$, where $\boldsymbol{e}_2 = (0, 1)$ is a unit vector. Therefore, we define the additional form b_h

$$b_h(\psi; v) = \sum_{K\in\mathcal{T}_h} \int_K \mathbf{K}(\psi)\boldsymbol{e}_2 \cdot \nabla v \,\mathrm{d}x - \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma H(\psi^{(L)}, \psi^{(R)}, \boldsymbol{n}) v \mathrm{d}S, \tag{2.22}$$

where $H : \mathbb{R} \times \mathbb{R} \times B \to \mathbb{R}$, $B = \{\boldsymbol{n} \in \mathbb{R}^2 : |\boldsymbol{n}| = 1\}$ is the numerical flux such that

$$\int_\Gamma \mathbf{K}(\theta(\psi))\boldsymbol{e}_2 \cdot \boldsymbol{n} v \mathrm{d}S \approx \int_\Gamma H(\psi^{(L)}, \psi^{(R)}, \boldsymbol{n}) v^{(L)} \mathrm{d}S, \ \Gamma \in \mathcal{F}_h.$$

In particular, we shall use the central numerical flux given by

$$H(\psi^{(L)}, \psi^{(R)}, \boldsymbol{n}) = \frac{\mathbf{K}(\psi^{(L)}) + \mathbf{K}(\psi^{(R)})}{2} \boldsymbol{e}_2 \cdot \boldsymbol{n}. \tag{2.23}$$

We mention that it is possible to prescribe other numerical fluxes, such as upwinding numerical flux, Lax-Friedrichs numerical flux, etc. We refer to [42] for more on this topic.

We are ready now to define the semidiscrete solution to ψ-formulation.

Definition 3. *We say that $\psi_h \in C([0, T]; S_{h,p})$ is the semidiscrete approximate solution to* (2.2) *obtained by the DG method if*

$$(\partial_t\vartheta(\psi_h), v_h) + a_h(\psi_h; \psi_h, v_h) + b_h(\psi_h; v_h) = \ell_h(v_h) \qquad \forall v_h \in S_{h,p} \ \forall t \in (0, T), \tag{2.24}$$

$$(\psi_h(0), v_h) = (\psi_0, v_h) \quad \forall v_h \in S_{h,p}. \tag{2.25}$$

In Chapter 5, we shall proceed with the discretization of the time variable leading to fully discrete schemes for Definitions 2–3. Prior to it, in the next two chapters, we present the error analysis for a semidiscrete scheme for Richards' equation.

3. Error analysis for a semidiscrete scheme

In what follows, we shall be concerned with the numerical analysis of a variant of the problem introduced in the previous chapters, which excludes the gravity term of Richards' equation. In particular, we define a local DG method to discretize the spatial variable and derive error estimates for the obtained time-continuous numerical scheme. The results from this chapter can be found in [24]. For the sake of clarity, we introduce local constants denoted by $C_1, C_2, \dots,$ valid only within the specific proof.

3.1 Model problem

Let $\Omega \subset \mathbb{R}^2$ be a polygon with Lipschitz continuous boundary $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$, $\partial\Omega_D \cap \partial\Omega_N = \emptyset$ as assumed in Chapter 2 and let $T > 0$. We study the following initial-boundary value problem: Find $u : Q_T \to \mathbb{R}$ such that

$$\partial_t \vartheta(u) - \nabla \cdot (\mathbf{K}(\theta(u))\nabla u) = g \qquad \text{in } Q_T, \tag{3.1a}$$

$$u\Big|_{\partial\Omega_D \times (0,T)} = u_D, \tag{3.1b}$$

$$\mathbf{K}(\theta(u))\nabla u \cdot \boldsymbol{n}\Big|_{\partial\Omega_N \times (0,T)} = g_N, \tag{3.1c}$$

$$u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \tag{3.1d}$$

where $g : Q_T \to \mathbb{R}$ denotes the source or sink term, $u_D : \partial\Omega_D \times (0,T) \to \mathbb{R}$, $g_N : \partial\Omega_N \times (0,T) \to \mathbb{R}$ and $u_0 : \Omega \to \mathbb{R}$ are the functions corresponding to initial and boundary conditions, respectively, and $\boldsymbol{n} = (n_1, n_2)$ is a unit outer normal to $\partial\Omega$. The function $\vartheta : \mathbb{R} \to \mathbb{R}_0^+$ describes the nonlinear change of the unknown function u through time and is defined as (cf. (1.8))

$$0 \le \theta(u) \le \vartheta(u) := \theta(u) + \frac{S_S}{\theta_S} \int_{-c}^{u} \theta(s) \mathrm{d}s, \tag{3.2}$$

where $c > 0$, $S_S \ge 0$ and $\theta_S > 0$ are constants, and $\theta : \mathbb{R} \to \mathbb{R}_0^+$ is a nonlinear function. The nonlinear tensor $\mathbf{K} : \mathbb{R} \to \mathbb{R}^{2\times 2}$ denotes the diffusion flux.

Remark 3. The problem (3.1) corresponds to the problem (2.2) when the gravity term is omitted and the diffusion coefficient $\mathbf{K}$ depends on $\theta(\psi)$ not only ψ. Thus, in this chapter, we distinguish the notation u for the pressure head ψ, while the rest of the quantities we keep denoted as before.

Remark 4. We point out that the problem formulation (3.1) is independent of the choice of the parameter $c > 0$ in (3.2). Taking into account that ϑ appears in (3.1) as a derivative, if we choose some other $c' > 0$, we have that

$$\int_{-c'}^{u} \theta(s) \mathrm{d}s = \int_{-c}^{u} \theta(s) \mathrm{d}s + \int_{-c'}^{-c} \theta(s) \mathrm{d}s.$$

Here, the second term on the right-hand side is constant and therefore its derivative is vanishing. The values $c > 0$ in (3.2) have been prescribed in different ways; e.g., $c = 0$ in [96], $c = \infty$ in [34], etc. In this chapter, we keep c fixed and finite.

3.1.1 Assumptions

For the purpose of the analysis ahead, we set assumptions on the hydraulic functions (θ, ϑ, $\mathbf{K}$), and boundary and source/sink terms, which describe flow in variably saturated porous media. Namely, we suppose (see [2, 96]):

(A1) The tensor $\mathbf{K}$ is uniformly bounded, uniformly symmetric positive definite in u and Lipschitz continuous in θ; i.e., there exists constants $k_0, k_1, k_L > 0$ such that

$$k_0|\boldsymbol{\zeta}|^2 \leq \boldsymbol{\zeta} \cdot \mathbf{K}(\theta(v))\boldsymbol{\zeta}, \qquad \boldsymbol{\zeta} \in \mathbb{R}^2, v \in \mathbb{R}, \tag{3.3}$$

$$|\mathbf{K}(\theta(v))\boldsymbol{\zeta}| \leq k_1|\boldsymbol{\zeta}|, \qquad \boldsymbol{\zeta} \in \mathbb{R}^2, v \in \mathbb{R}, \tag{3.4}$$

$$|\mathbf{K}(\theta(u_1)) - \mathbf{K}(\theta(u_2))| \leq k_L|\theta(u_1) - \theta(u_2)|, \qquad u_1, u_2 \in \mathbb{R}. \tag{3.5}$$

(A2) The component-wise derivative $\frac{\mathrm{d}\mathbf{K}(\theta(u))}{\mathrm{d}u}$ is bounded; i.e., there exists $k_d > 0$ such that

$$\left|\frac{\mathrm{d}\mathbf{K}(\theta(u))}{\mathrm{d}u}\right| \leq k_d.$$

(A3) The function θ is monotone nondecreasing, uniformly bounded from above, and Lipschitz continuous; i.e., there exists a constant $L_\theta > 0$ such that

$$|\theta(u_1) - \theta(u_2)| \leq L_\theta|u_1 - u_2| \quad \forall u_1, u_2 \in \mathbb{R}.$$

(A4) The composition $\vartheta' \circ \vartheta^{-1}$ is Hölder continuous with order $1/3 < \beta \leq 1$; i.e., there exists a constant $H_\vartheta > 0$ such that for any $u_1, u_2 \in \mathbb{R}$

$$|\vartheta'(u_1) - \vartheta'(u_2)| \leq H_\vartheta|\vartheta(u_1) - \vartheta(u_2)|^\beta.$$

(A5) u_D is the trace of some $u_D^* \in C([0,T]; H^1(\Omega)) \cap L^\infty(Q_T)$ on $\partial\Omega_D \times (0,T)$.
(A6) $g_N \in L^2(0,T; L^2(\partial\Omega_N))$.
(A7) $g \in L^2(0,T; L^2(\Omega))$.

Remark 5. Let $u_1, u_2 \in \mathbb{R}$, $u_1 \geq u_2$; then from (3.2) we have that

$$\vartheta(u_1) - \vartheta(u_2) = \theta(u_1) - \theta(u_2) + \frac{S_S}{\theta_S}\int_{u_2}^{u_1} \theta(s)\mathrm{d}s \geq \theta(u_1) - \theta(u_2),$$

and thus,

$$|\theta(u_1) - \theta(u_2)| \leq |\vartheta(u_1) - \vartheta(u_2)| \quad \forall u_1, u_2 \in \mathbb{R}. \tag{3.6}$$

Furthermore, the assumption (A3) implies that ϑ is Lipschitz continuous too, with the Lipschitz constant $L_\vartheta := L_\theta + \frac{S_S}{\theta_S}\sup_{s\in\mathbb{R}} \theta(s)$. Since the integral operator is monotone, ϑ is a monotone nondecreasing function too.

Remark 6. We note that the assumptions (A3)–(A4) admit the case $\vartheta' = 0$, i.e., when the flow occurs in the saturated regime (hence, $\theta' = 0$ and $S_S = 0$).

Remark 7. Richards' equation with the van Genuchten-Mualem [92, 66] model of constitutive relations yields $\theta(u) = (1 + (C|u|)^{1/(1-m)})^{-m}$ (cf. (1.14) and [77]), where $C > 0$ and $m \in (0,1)$. Moreover, the Taylor expansion around $u \approx 0$ implies $\theta(u) \sim 1 - m(C|u|)^{1/(1-m)}$ with $\alpha = 1/(1-m)$. Therefore, the assumption (A4) is satisfied with $\beta = m$ (cf. [2, Remark 3]). However, due to the numerical analysis we have to additionally restrict $\beta > 1/3$ (cf. Lemma 21).

3.1.2 Some notation and remarks

In this chapter, we consider triangulations that are shape-regular and quasi-uniform. In particular, we assume conforming triangulations $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$, $\bar{h} > 0$, but this assumption can be relaxed (cf. Remark 2).

Since in this chapter we deal with vector-valued functions, we discuss the corresponding notation. Namely, as in (2.3), we define trace and jump operators of a vector-valued function $\boldsymbol{w}$,

$$\langle \boldsymbol{w} \rangle_\Gamma = \frac{1}{2}\Big(\boldsymbol{w}_\Gamma^{(L)} + \boldsymbol{w}_\Gamma^{(R)}\Big),\ [\boldsymbol{w}]_\Gamma = \boldsymbol{w}_\Gamma^{(L)} - \boldsymbol{w}_\Gamma^{(R)},\ \Gamma \in \mathcal{F}_h^I,$$

and

$$\langle \boldsymbol{w} \rangle_\Gamma = [\boldsymbol{w}]_\Gamma = \boldsymbol{w}_\Gamma^{(L)},\ \Gamma \in \mathcal{F}_h^B.$$

In addition, we mention the broken Sobolev space of vector-valued functions

$$\boldsymbol{H}^k(\Omega, \mathcal{T}_h) = (H^k(\Omega, \mathcal{T}_h))^2,$$

and the two-dimensional space of discontinuous polynomial functions

$$\boldsymbol{S}_{h,p} = \{\boldsymbol{w} \in (L^2(\Omega))^2 : \boldsymbol{w}|_K \in (P_p(K))^2, \forall K \in \mathcal{T}_h\}.$$

Analogously to (2.8), we denote the L^2-projection of some $\boldsymbol{w} \in (L^2(\Omega))^2$, which represents the L^2-projection of scalar components of the vector $\boldsymbol{w}$. We shall consider triangulations that are shape-regular (2.4) and quasi-uniform (2.5).

3.1.3 Weak solution

As mentioned in Remark 6, we consider a nonlinear parabolic PDE that can degenerate to elliptic one when $\vartheta' \to 0$. Since the exact solution to (3.1) often has low regularity, we introduce its weak solution.

Definition 4. *A function u is called the weak solution to the problem* (3.1) *if*

$$u - u_D^* \in L^2(0, T; H_{0D}^1(\Omega)), \tag{3.7a}$$

$$\vartheta(u) \in L^\infty(0, T; L^1(\Omega)) \cap H^1(0, T; H^{-1}(\Omega)), \tag{3.7b}$$

and the following identity is satisfied

$$(\partial_t \vartheta(u), v) + (\mathbf{K}(\theta(u))\nabla u, \nabla v) = (g(t), v) + (g_N(t), v)_N,$$

for all $v \in H_{0D}^1(\Omega)$ a.e. in $(0, T)$, and $u(0) = u_0$ in Ω.

Remark 8. We mention that the condition (3.7b) from Definition 4 is equivalent to (cf. [86])

$$\begin{aligned} \vartheta(u) &\in C([0, T]; H^{-1}(\Omega)) \cap L^\infty(0, T; L^2(\Omega)), \\ \partial_t \vartheta(u) &\in L^2(0, T; H^{-1}(\Omega)). \end{aligned}$$

More on its existence, uniqueness and regularity can be found in [1, 69] and [70].

3.2 A local discontinuous Galerkin method

In order to define the local discontinuous Galerkin (LDG) method [17, 22, 46], we introduce the expanded mixed formulation of the problem (3.1) (cf. [96]). Namely, we define two new auxiliary vector variables, $\boldsymbol{q} = \nabla u$ and $\boldsymbol{\sigma} = \boldsymbol{K}(\theta(u))\boldsymbol{q}$, and rewrite (3.1) as a system of three first order hyperbolic equations

$$\partial_t \vartheta(u) - \nabla \cdot \boldsymbol{\sigma} = g \qquad \text{in } Q_T, \tag{3.8a}$$

$$\boldsymbol{\sigma} = \boldsymbol{K}(\theta(u))\boldsymbol{q} \qquad \text{in } Q_T, \tag{3.8b}$$

$$\boldsymbol{q} = \nabla u \qquad \text{in } Q_T, \tag{3.8c}$$

$$u\Big|_{\partial\Omega_D \times (0,T)} = u_D, \tag{3.8d}$$

$$\boldsymbol{\sigma} \cdot \boldsymbol{n}\Big|_{\partial\Omega_N \times (0,T)} = g_N, \tag{3.8e}$$

$$u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega. \tag{3.8f}$$

Furthermore. we assume the exact solution $(u, \boldsymbol{q}, \boldsymbol{\sigma})$ to the problem (3.8) satisfies the following regularity conditions for some $s \geq 2$.

(B1) $u \in L^2(0,T; H^s(\Omega))$, $\partial_t u \in L^\infty(Q_T) \cap L^2(0,T; H^s(\Omega))$, $\|\partial_t u\|_{L^2(Q_T)} \leq C_X$.
(B2) $\boldsymbol{q} \in L^2(0,T; \boldsymbol{H}^s(\Omega))$, $\|\boldsymbol{q}\|_{L^\infty(\Omega)} \leq C_B$ for $t \in (0,T)$.
(B3) $\boldsymbol{\sigma} \in L^2(0,T; \boldsymbol{H}^s(\Omega))$.

Now, we start with the derivation of the scheme. Namely, we multiply (3.8a), (3.8b) and (3.8c) by $v \in H^1(\Omega, \mathcal{T}_h)$, $\boldsymbol{w} \in \boldsymbol{H}^1(\Omega, \mathcal{T}_h)$ and $\boldsymbol{z} \in \boldsymbol{H}^1(\Omega, \mathcal{T}_h)$, respectively. Then, integrate over $K \in \mathcal{T}_h$ and use Green's theorem so that

$$\int_K \partial_t \vartheta(u) v \,\mathrm{d}x + \int_K \boldsymbol{\sigma} \cdot \nabla v \,\mathrm{d}x - \int_{\partial K} \boldsymbol{\sigma} \cdot \boldsymbol{n} v \,\mathrm{d}S = \int_K g v \,\mathrm{d}x, \tag{3.9a}$$

$$\int_K \boldsymbol{K}(\theta(u))\boldsymbol{q} \cdot \boldsymbol{w} \,\mathrm{d}x - \int_K \boldsymbol{\sigma} \cdot \boldsymbol{w} = 0, \tag{3.9b}$$

$$\int_K \boldsymbol{q} \cdot \boldsymbol{z} \,\mathrm{d}x + \int_K u \nabla \cdot \boldsymbol{z} \,\mathrm{d}x - \int_{\partial K} u \boldsymbol{z} \cdot \boldsymbol{n} \,\mathrm{d}S = 0. \tag{3.9c}$$

At this point, we observe that the solution on integrals on edges should be defined carefully. Therefore, in order to define the approximate solution $(u_h, \boldsymbol{q}_h, \boldsymbol{\sigma}_h) \in S_{h,p} \times \boldsymbol{S}_{h,p} \times \boldsymbol{S}_{h,p}$, $t \in [0,T]$, we substitute $v_h \in S_{h,p}$ in (3.9a) and $\boldsymbol{w}_h, \boldsymbol{z}_h \in \boldsymbol{S}_{h,p}$ in (3.9b)–(3.9c),

$$\int_K \partial_t \vartheta(u_h) v_h \,\mathrm{d}x + \int_K \boldsymbol{\sigma}_h \cdot \nabla v_h \,\mathrm{d}x - \int_{\partial K} \hat{\boldsymbol{\sigma}} \cdot \boldsymbol{n} v_h \,\mathrm{d}S = \int_K g v_h \,\mathrm{d}x, \tag{3.10a}$$

$$\int_K \boldsymbol{K}(\theta(u_h))\boldsymbol{q}_h \cdot \boldsymbol{w}_h \,\mathrm{d}x - \int_K \boldsymbol{\sigma}_h \cdot \boldsymbol{w}_h = 0, \tag{3.10b}$$

$$\int_K \boldsymbol{q}_h \cdot \boldsymbol{z}_h \,\mathrm{d}x + \int_K u_h \nabla \cdot \boldsymbol{z}_h \,\mathrm{d}x - \int_{\partial K} \hat{u} \boldsymbol{z}_h \cdot \boldsymbol{n} \,\mathrm{d}S = 0, \tag{3.10c}$$

where $\hat{u}$ and $\hat{\boldsymbol{\sigma}}$ denote the numerical fluxes, which approximate the solution across the interfaces. Particularly, we choose the numerical fluxes such that the stability of the resulting scheme is preserved [4], [46]

$$\hat{u}(u_h) = \begin{cases} \langle u_h \rangle + \boldsymbol{\lambda} \cdot \boldsymbol{n}[u_h], & \Gamma \in \mathcal{F}_h^I, \\ u_D, & \Gamma \in \mathcal{F}_h^D, \\ u_h, & \Gamma \in \mathcal{F}_h^N, \end{cases}$$

$$\hat{\boldsymbol{\sigma}}(u_h, \boldsymbol{\sigma}_h) \cdot \boldsymbol{n} = \begin{cases} \langle \boldsymbol{\sigma}_h \rangle \cdot \boldsymbol{n} + \boldsymbol{\lambda} \cdot \boldsymbol{n}[\boldsymbol{\sigma}_h \cdot \boldsymbol{n}] - \kappa [u_h], & \Gamma \in \mathcal{F}_h^I, \\ \boldsymbol{\sigma}_h \cdot \boldsymbol{n} - \kappa(u_h - u_D), & \Gamma \in \mathcal{F}_h^D, \\ g_N, & \Gamma \in \mathcal{F}_h^N, \end{cases}$$

where $\boldsymbol{\lambda} : \cup_{\Gamma \in \mathcal{F}_h} \Gamma \to \mathbb{R}^2$ is a constant vector independent of h, while $\kappa : \cup_{\Gamma \in \mathcal{F}_h} \Gamma \to \mathbb{R}$ is the penalty parameter defined as in (2.16). Moreover, the numerical fluxes are *conservative*

$$[\hat{u}] = 0, \; [\hat{\boldsymbol{\sigma}}] = 0, \; \Gamma \in \mathcal{F}_h^I \tag{3.11}$$

and *consistent*

$$\hat{u}(u) = u, \; \hat{\boldsymbol{\sigma}}(u, \boldsymbol{\sigma}) \cdot \boldsymbol{n} = \boldsymbol{\sigma} \cdot \boldsymbol{n}. \tag{3.12}$$

Furthermore, for $v, r \in H^1(\Omega, \mathcal{T}_h)$, $\boldsymbol{w}, \boldsymbol{z} \in \boldsymbol{H}^1(\Omega, \mathcal{T}_h)$ we define the following forms

$$A_h(\boldsymbol{w}, v) = (\boldsymbol{w}, \nabla v) - \sum_{\Gamma \in \mathcal{F}_h^{ID}} \int_\Gamma \Big(\langle \boldsymbol{w} \rangle \cdot \boldsymbol{n} - \boldsymbol{\lambda} \cdot \boldsymbol{n} [\boldsymbol{w} \cdot \boldsymbol{n}] \Big) [v] \, \mathrm{d}S, \tag{3.13}$$

$$B_h(v; \boldsymbol{w}, \boldsymbol{z}) = ((\mathbf{K}(\theta(v))\boldsymbol{w}, \boldsymbol{z}), \tag{3.14}$$

$$J_h(v, r) = \sum_{\Gamma \in \mathcal{F}_h^{ID}} \kappa \int_\Gamma [v][r] \, \mathrm{d}S, \tag{3.15}$$

$$J_h^D(v) = \sum_{\Gamma \in \mathcal{F}_h^D} \kappa \int_\Gamma u_D v \, \mathrm{d}S, \tag{3.16}$$

$$F_h(v) = (g, v) + J_h^D(v) + (g_N, v)_N, \tag{3.17}$$

$$G_h(\boldsymbol{w}) = \sum_{\Gamma \in \mathcal{F}_h^D} \int_\Gamma \boldsymbol{w} \cdot \boldsymbol{n} u_D \, \mathrm{d}S, \tag{3.18}$$

where the penalty parameter κ is given by (2.16). Let us note that the bilinear form J_h is the same as in Chapter 2 (cf. (2.15))

Finally, we define the semidiscrete approximate solution obtained by the LDG method. The existence and uniqueness of the approximate solution of the LDG method for nonmonotone quasilinear equations has been studied in [46] (see also [79]).

Definition 5. *We say that the triplet*

$$(u_h, \boldsymbol{q}_h, \boldsymbol{\sigma}_h) \in C^1([0,T]; S_{h,p}) \times C^1([0,T]; \boldsymbol{S}_{h,p}) \times C^1([0,T]; \boldsymbol{S}_{h,p})$$

is a semidiscrete approximate solution to (3.8) *obtained by the LDG method, if it satisfies*

$$(\partial_t \vartheta(u_h), v_h) + A_h(\boldsymbol{\sigma}_h, v_h) + J_h(u_h, v_h) = F_h(v_h) \qquad \forall v_h \in S_{h,p}, \tag{3.19a}$$

$$B_h(u_h; \boldsymbol{q}_h, \boldsymbol{w}_h) - (\boldsymbol{\sigma}_h, \boldsymbol{w}_h) = 0 \qquad \forall \boldsymbol{w}_h \in \boldsymbol{S}_{h,p}, \tag{3.19b}$$

$$(\boldsymbol{q}_h, \boldsymbol{z}_h) - A_h(\boldsymbol{z}_h, u_h) = G_h(\boldsymbol{z}_h) \qquad \forall \boldsymbol{z}_h \in \boldsymbol{S}_{h,p}, \tag{3.19c}$$

with the initial condition $u_h(0) \equiv \Pi_{hp} u_0$ *for almost all* $t \in (0, T)$.

It can be verified that the exact solution $(u, \boldsymbol{q}, \boldsymbol{\sigma})$ satisfies (3.19a)–(3.19c); namely, from the consistency of numerical fluxes (3.12) implies the consistency of the LDG scheme

$$\begin{aligned} (\partial_t \vartheta(u), v_h) + A_h(\boldsymbol{\sigma}, v_h) + J_h(u, v_h) &= F_h(v_h) \qquad && \forall v_h \in S_{h,p}, && (3.20a)\\ B_h(u; \boldsymbol{q}, \boldsymbol{w}_h) - (\boldsymbol{\sigma}, \boldsymbol{w}_h) &= 0 && \forall \boldsymbol{w}_h \in \boldsymbol{S}_{h,p}, && (3.20b)\\ (\boldsymbol{q}, \boldsymbol{z}_h) - A_h(\boldsymbol{z}_h, u) &= G_h(\boldsymbol{z}_h) && \forall \boldsymbol{z}_h \in \boldsymbol{S}_{h,p}, && (3.20c) \end{aligned}$$

for almost all $t \in (0, T)$.

3.3 Auxiliary results

In this section, we introduce some results from the theory of the finite element method (see, e.g., [14]) and DG method (see, e.g., [33]). Furthermore, we derive the coercivity bound of the nonlinear form B_h with respect to linear arguments and its upper bound.

We use the multiplicative trace inequality (cf. [33, Lemma 2.19])

$$\|v\|^2_{L^2(\partial K)} \le C_M\Big(\|v\|_{L^2(K)}|v|_{H^1(K)} + h_K^{-1}\|v\|^2_{L^2(K)}\Big),$$

where $v \in H^1(K)$, $K \in \mathcal{T}_h$, $h \in (0, \bar{h})$, and the inverse estimates (cf. [14, Lemma 4.5.3])

$$|v_h|_{H^1(K)} \le C_I h_K^{-1}\|v_h\|_{L^2(K)} \quad \forall v \in P^p(K) \quad \forall K \in \mathcal{T}_h \quad \forall h \in (0, \bar{h}), \qquad (3.21a)$$

$$\|v_h\|_{L^\infty(K)} \le C_I h_K^{-1}\|v_h\|_{L^2(K)} \quad \forall v \in P^p(K) \quad \forall K \in \mathcal{T}_h \quad \forall h \in (0, \bar{h}). \qquad (3.21b)$$

Furthermore, if $v_h \in S_{hp}$, then the multiplicative trace inequality and the inverse inequality (3.21a) imply that

$$\sum_{K \in \mathcal{T}_h} h_K \|v_h\|^2_{L^2(\partial K)} \le C_M(C_I + 1)\|v_h\|^2_{L^2(\Omega)}. \qquad (3.22)$$

We shall use the elementary inequalities (see, e.g., [28])

$$|a-b|^2 \le 2|a|^2 + 2|b|^2 \quad \text{and} \quad |a^r - b^r| \le |a-b|^r, \quad 0 \le r \le 1, \ a, b \in \mathbb{R}. \qquad (3.23)$$

We will use some relations for the form J_h (2.15), cf. [33, Lemma 2.32]

$$|J_h(v, w)| \le J_h^{1/2}(v, v) J_h^{1/2}(w, w), \ v, w \in H^1(\Omega, \mathcal{T}_h), \ h \in (0, \bar{h}), \qquad (3.24)$$

$$J_h(v, v) \le \frac{C_W C_M}{C_T} \sum_{K \in \mathcal{T}_h} \Big(3h_K^{-2}\|v\|^2_{L^2(K)} + |v|^2_{H^1(K)}\Big), \ v \in H^1(\Omega, \mathcal{T}_h), \ h \in (0, \bar{h}). \qquad (3.25)$$

Furthermore, we define the norm on $H^1(\Omega, \mathcal{T}_h)$

$$|||v||| = \Bigg(\sum_{K \in \mathcal{T}_h} |v|^2_{H^1(K)} + J_h(v, v)\Bigg)^{1/2}. \qquad (3.26)$$

We shall use the estimate that can be obtained by applying the inverse inequality to the general version of the multiplicative trace inequality proposed in [33, Lemma 4.8]

$$\|v_h\|^2_{L^2(\partial\Omega)} \le C_N\Big(|||v_h||| \|v_h\|_{L^2(\Omega)} + \|v_h\|^2_{L^2(\Omega)}\Big), \ v_h \in S_{h,p}. \qquad (3.27)$$

We establish the relation between the L^2-norm and $|||\cdot|||$-norm of a function belonging to the $S_{h,p}$ space with the aid of the broken Poincaré inequality [13]

$$\|v_h\|_{L^2(\Omega)} \le C_P |||v_h||| \quad \forall v_h \in S_{hp}. \qquad (3.28)$$

3.3.1 Bounds on the forms A_h and B_h

We propose an estimate on a term appearing in the form A_h defined by (3.13), whose proof is analogous to [33, Lemma 2.27].

Lemma 2. *Let κ be given by* (2.16)*. Then, it holds that*

$$\sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \kappa^{-1}\Big|\langle \boldsymbol{w}\rangle - \boldsymbol{\lambda}\cdot\boldsymbol{n}[\boldsymbol{w}]\Big|^2 \,\mathrm{d}S \le R(\boldsymbol{w}), \qquad \boldsymbol{w}\in \boldsymbol{H}^1(\Omega,\mathcal{T}_h), \tag{3.29}$$

$$\sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \kappa^{-1}\Big|\langle \boldsymbol{w}_h\rangle - \boldsymbol{\lambda}\cdot\boldsymbol{n}[\boldsymbol{w}_h]\Big|^2 \,\mathrm{d}S \le C_a\|\boldsymbol{w}_h\|^2_{L^2(\Omega)}, \quad \boldsymbol{w}_h\in \boldsymbol{S}_{h,p}, \tag{3.30}$$

where $C_a > 0$ is a constant independent of h and

$$R(\boldsymbol{w}) := C_r \sum_{K\in\mathcal{T}_h} \Big(h_K^2 |\boldsymbol{w}|^2_{H^1(K)} + 3\|\boldsymbol{w}\|^2_{L^2(K)}\Big),\ \boldsymbol{w}\in \boldsymbol{H}^1(\Omega,\mathcal{T}_h), \tag{3.31}$$

with $C_r := 3C_M C_G \max(1, 2|\lambda|^2)/C_W$.

In the next lemma, we show the coercivity of the form B_h with respect to the linear terms, and its boundedness.

Lemma 3. *Let $B_h(\cdot\,;\cdot\,,\cdot\,)$ be defined by* (3.14)*. Then,*

$$B_h(v;\boldsymbol{w},\boldsymbol{w}) \ge k_0\|\boldsymbol{w}\|^2_{L^2(\Omega)}, \qquad v\in H^1(\Omega,\mathcal{T}_h),\ \boldsymbol{w}\in \boldsymbol{H}^1(\Omega,\mathcal{T}_h), \tag{3.32}$$

$$|B_h(v;\boldsymbol{w},\boldsymbol{z})| \le k_1\|\boldsymbol{w}\|_{L^2(\Omega)}\|\boldsymbol{z}\|_{L^2(\Omega)}, \quad v\in H^1(\Omega,\mathcal{T}_h),\ \boldsymbol{w},\boldsymbol{z}\in \boldsymbol{H}^1(\Omega,\mathcal{T}_h). \tag{3.33}$$

Proof. Using (A1) we get

$$\sum_{K\in\mathcal{T}_h}\int_K \mathbf{K}(\theta(\zeta))\boldsymbol{w}\cdot\boldsymbol{w}\,\mathrm{d}x \ge \sum_{K\in\mathcal{T}_h}\int_K k_0|\boldsymbol{w}|^2\,\mathrm{d}x,$$

which yields the first part of the statement. For the other part, we simply combine (A1) with the Cauchy-Schwarz inequality,

$$\sum_{K\in\mathcal{T}_h}\int_K \mathbf{K}(\theta(\zeta))\boldsymbol{w}\cdot\boldsymbol{z}\,\mathrm{d}x \le \sum_{K\in\mathcal{T}_h}\int_K k_1|\boldsymbol{w}||\boldsymbol{z}|\,\mathrm{d}x.$$

□

3.3.2 Some integral identities and inequalities

We shall use a special case of the Hölder inequality

$$\int_\Omega f^\gamma\,\mathrm{d}x \le |\Omega|^{1-\gamma}\bigg(\int_\Omega f\,\mathrm{d}x\bigg)^\gamma,\ 0\le\gamma\le 1,\ f\in L^1(\Omega),\ f\ge 0 \text{ a.e. on } \Omega, \tag{3.34}$$

where the notation $|\cdot|$ stands for the Lebesgue measure of the corresponding set in $\mathbb{R}^2$. This relation can be derived if we set $w := 1$, $v := f^\gamma$ and $p = \gamma^{-1}$ in the classical Hölder inequality, $\|vw\|_{L^1(\Omega)} \le \|v\|_{L^p(\Omega)}\|w\|_{L^q(\Omega)}$, where $v\in L^p(\Omega)$,

$w \in L^q(\Omega)$, $1/p+1/q = 1$ for $p, q \geq 1$. Furthermore, we shall need the generalized Hölder inequality

$$\int_\Omega |f_1 f_2 f_3| \,\mathrm{d}x \leq \|f_1\|_{L^2(\Omega)} \|f_2\|_{L^2(\Omega)} \|f_3\|_{L^\infty(\Omega)}, \quad f_1, f_2 \in L^2(\Omega),\ f_3 \in L^\infty(\Omega). \tag{3.35}$$

We introduce the p-triangle inequality for the quasi-Banach L^p spaces defined for $0 < p < 1$ with the quasi-norm $\|\cdot\|^p_{L^p(\Omega)} = \int_\Omega |\cdot|^p \,\mathrm{d}x$ [25]

$$\|f_1 + f_2\|_{L^p(\Omega)} \leq 2^{\frac{1-p}{p}} \Big(\|f_1\|_{L^p(\Omega)} + \|f_2\|_{L^p(\Omega)} \Big) \quad \forall f_1, f_2 \in L^p(\Omega). \tag{3.36}$$

In the error analysis we shall use a simplified version of the Leibnitz integral rule (see e.g., [51])

$$\frac{d}{dt} \int_{a(t)}^{b(t)} f(\chi) \mathrm{d}\chi = f(b(t))b'(t) - f(a(t))a'(t), \tag{3.37}$$

which can be proven by setting $F(a(t), b(t)) := \int_{a(t)}^{b(t)} f(\chi)\mathrm{d}\chi$ in (3.37) and applying the chain rule to $\frac{\mathrm{d}F}{\mathrm{d}t}$. Specially, if $a(t) = a(\boldsymbol{x}, t)$ and $b(t) = b(\boldsymbol{x}, t)$ then (3.37) holds but now with ∂_t instead of $\frac{d}{dt}$.

Lastly, we state a result proposed in [2, Proposition 1.], which can be shown with aid of fundamental calculus.

Lemma 4. *Let $f : \mathbb{R} \to \mathbb{R}$ be a monotone nondecreasing, uniformly Lipschitz and uniformly bounded function. Then, for arbitrary $v, w \in \mathbb{R}$ the following inequalities are fulfilled*

$$M_f (f(w) - f(v))^2 \leq \int_v^w (f(\chi) - f(v)) \mathrm{d}\chi, \tag{3.38a}$$

$$\int_v^w (f(w) - f(\chi)) \mathrm{d}\chi \leq (f(w) - f(v))(w - v), \tag{3.38b}$$

where $M_f > 0$ is a constant that depends on the function f defined as $M_f := \Big(2\, ess\, sup_{u\in\mathbb{R}} |f'(u)|\Big)^{-1}$.

Proof. Without loss of generality, let us assume $v \leq w$. For simplicity, we define an auxiliary function

$$g(\chi) := f(\chi) - f(v). \tag{3.39}$$

Note that the function g is nonnegative over the interval $[v, w]$ and it holds $g(v) = 0$. Let us consider a linear function passing through the point $g(w)$ with the slope $\sup_{u\in\mathbb{R}} |g'(u)|$ and intersecting the x-axis at some $z \in (v, w)$ as depicted in Fig. 3.1. Since for any point $r \in (v, w)$ it holds

$$f'(r) = g'(r) \leq \sup_{u\in\mathbb{R}} |g'(u)| = \sup_{u\in\mathbb{R}} |f'(u)|, \tag{3.40}$$

we have that the linear function is completely below the function g at all points. In mathematical terms, it holds

$$\int_v^w g(\chi)\mathrm{d}\chi \geq \frac{1}{2} g(w)(w - z),$$

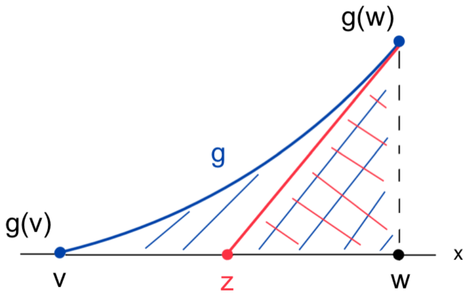

Figure 3.1: Geometrical proof of Lemma 3.38.

or using (3.39), we have that

$$\int_v^w (f(\chi) - f(v))\mathrm{d}\chi \geq \frac{1}{2}(f(w) - f(v))(w - z). \tag{3.41}$$

In other words, the area below the function g (depicted in blue) is greater or equal to the area of the triangle (depicted in red) as illustrated in the Fig. 3.1. Moreover, by the definition of the tangent line at the point $(w, g(w))$, we have that

$$g(w) = \sup_{u\in\mathbb{R}} |g'(u)|(w - z),$$

or equivivalently,

$$f(w) = \sup_{u\in\mathbb{R}} |f'(u)|(w - z) + f(v). \tag{3.42}$$

Furthermore, we rewrite (3.42) (cf. (3.40)) obtaining the identity

$$w - z = (\sup_{u\in\mathbb{R}} |f'(u)|)^{-1}(f(w) - f(v)). \tag{3.43}$$

By substituting (3.43) into (3.41), we derive the first part of the lemma.

On the other hand, the proof of the second part is more simple. Namely, the statement follows from the monotonicity of the function f

$$\int_v^w (f(w) - f(\chi))\mathrm{d}\chi \leq \int_v^w (f(w) - f(v))\mathrm{d}\chi = (f(w) - f(v))(w - v).$$

□

3.4 Stability of the approximate solution

A bound on the semidiscrete approximate solution $(u_h, \boldsymbol{q}_h, \boldsymbol{\sigma}_h)$ will play an important role in the derivation of the main error estimate; therefore, we present some results on the stability of the semidiscrete solution. We start with an auxiliary lemma.

Lemma 5. *Let the triplet $(u_h, \boldsymbol{q}_h, \boldsymbol{\sigma}_h)$ be the approximate solution given by Definition 5 and let the assumptions (A1)–(A7) be satisfied. Then,*

$$\|u_h\|_{L^2(\Omega)} \leq C_u\Big(J_h^{1/2}(u_h, u_h) + \|\boldsymbol{q}_h\|_{L^2(\Omega)} + (J_h^D(u_D))^{1/2}\Big), \tag{3.44}$$

$$\|\boldsymbol{\sigma}_h\|_{L^2(\Omega)} \leq C_\sigma \|\boldsymbol{q}_h\|_{L^2(\Omega)}, \tag{3.45}$$

where the constants $C_u, C_\sigma > 0$ are independent of the exact solutions' components $(u, \boldsymbol{q}, \boldsymbol{\sigma})$ and the discretization parameter h.

Proof. We substitute $\boldsymbol{z}_h := \nabla u_h \in \boldsymbol{S}_{h,p}$, defined as $\boldsymbol{z}_h|_K := \nabla(u_h|_K) \in P_p(K),\ K \in \mathcal{T}_h$, into (3.19c) and by use the definition of A_h (3.13) we rewrite (3.19c) as

$$\begin{aligned}&\|\nabla u_h\|^2_{L^2(\Omega)}\\ &\quad= \Bigg| \sum_{\Gamma \in \mathcal{F}_h^{ID}} \int_\Gamma \Big(\langle \nabla u_h \cdot \boldsymbol{n}\rangle - \boldsymbol{\lambda} \cdot \boldsymbol{n}[\nabla u_h \cdot \boldsymbol{n}]\Big)[u_h]\,\mathrm{d}S + (\boldsymbol{q}_h, \nabla u_h) - G_h(\nabla u_h)\Bigg|.\end{aligned} \tag{3.46}$$

Then, we estimate each term on the right-hand side of (3.46). Namely, by the Cauchy-Schwarz inequality and (3.30), we have that

$$\sum_{\Gamma \in \mathcal{F}_h^{ID}} \int_\Gamma \Big(\langle \nabla u_h \cdot \boldsymbol{n}\rangle - \boldsymbol{\lambda} \cdot \boldsymbol{n}[\nabla u_h \cdot \boldsymbol{n}]\Big)[u_h]\,\mathrm{d}S \leq C_a \|\nabla u_h\|_{L^2(\Omega)} J_h^{1/2}(u_h, u_h). \tag{3.47}$$

Furthermore, the Cauchy-Schwarz inequality implies

$$|G_h(\nabla u_h)| \leq \Bigg(\sum_{\Gamma \in \mathcal{F}_h^D} \int_\Gamma \kappa^{-1} |\nabla u_h|^2\,\mathrm{d}S\Bigg)^{1/2} (J_h^D(u_D))^{1/2}. \tag{3.48}$$

From the equivalence condition (2.6) and (3.22) we have that

$$\begin{aligned}\sum_{\Gamma \in \mathcal{F}_h^D} \int_\Gamma \kappa^{-1} |\nabla u_h|^2\,\mathrm{d}S &\leq \frac{C_W}{C_G} \sum_{K \in \mathcal{T}_h} h_K \int_{\partial K} |\nabla u_h|^2\,\mathrm{d}x\\ &\leq \frac{C_W}{C_G} C_M (C_I + 1) \|\nabla u_h\|^2_{L^2(\Omega)}.\end{aligned} \tag{3.49}$$

We substitute the relations (3.47)–(3.49) into (3.46), use the Cauchy-Schwarz inequality on the middle term on the right-hand side of (3.46), and then cancel $\|\nabla u_h\|_{L^2(\Omega)}$ on both sides of the inequality, so we get that

$$\|\nabla u_h\|_{L^2(\Omega)} \leq C_a J_h^{1/2}(u_h, u_h) + \|\boldsymbol{q}_h\|_{L^2(\Omega)} + C_1 (J_h^D(u_D))^{1/2}, \tag{3.50}$$

where we have denoted $C_1 := C_W C_M (C_I + 1)/C_G$. Then, we add the term $J_h^{1/2}(u_h, u_h)$ to both sides of (3.50) aiming to use the definition of $|\!|\!| \cdot |\!|\!|$-norm (3.26) as

$$|\!|\!| u_h |\!|\!| \leq (C_a + 1) J_h^{1/2}(u_h, u_h) + \|\boldsymbol{q}_h\|_{L^2(\Omega)} + C_1 (J_h^D(u_D))^{1/2}. \tag{3.51}$$

By application of the broken Poincaré inequality to (3.51), and by setting

$$C_u := C_P \max(C_a + 1, C_1)$$

in (3.44), we complete the proof of the first part of the lemma.

In order to prove the second part of the lemma, we insert $\boldsymbol{w}_h := \boldsymbol{\sigma}_h$ in (3.19b)

$$\|\boldsymbol{\sigma}_h\|^2 = |B_h(u_h; \boldsymbol{q}_h, \boldsymbol{\sigma}_h)|,$$

and apply (3.33) to the form B_h; hence, we obtain (3.45) for $C_\sigma := k_1$.

□

Theorem 6 (Stability of the semidiscrete approximate solution)**.** *Let the triplet* $(u_h, \boldsymbol{q}_h, \boldsymbol{\sigma}_h)$ *be the approximate solution given by Definition 5 and let the assumptions (A1)–(A7) be satisfied. Then,*

$$\begin{aligned}\|\vartheta(u_h(T)) - \vartheta(u_h(0))\|^2_{L^2(\Omega)} + \|\boldsymbol{q}_h\|^2_{L^2(0,T;L^2(\Omega))} + \int_0^T J_h(u_h, u_h)\,\mathrm{d}t \\ \le C\bigg(\|u_0\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(0,T;L^2(\Omega))} + \int_0^T J_h^D(u_D)\,\mathrm{d}t + \|g_N\|^2_{L^2(0,T;L^2(\partial\Omega_N))}\bigg),\end{aligned} \tag{3.52}$$

where the constant $C > 0$ *is independent of the exact solution* $(u, \boldsymbol{q}, \boldsymbol{\sigma})$ *and the discretization parameter* h*.*

Proof. We insert $v_h := u_h$, $\boldsymbol{w}_h := \boldsymbol{q}_h$, $\boldsymbol{z}_h := \boldsymbol{\sigma}_h$ in (3.19a)–(3.19c), sum the equations, and integrate the resulting equation over the interval $[0, T]$ obtaining

$$\begin{aligned}\int_0^T (\partial_t\vartheta(u_h), u_h)\,\mathrm{d}t + \int_0^T B_h(u_h; \boldsymbol{q}_h, \boldsymbol{q}_h)\,\mathrm{d}t + \int_0^T J_h(u_h, u_h)\,\mathrm{d}t \\ = \int_0^T \Big(F_h(u_h) + G_h(\boldsymbol{\sigma}_h)\Big)\,\mathrm{d}t.\end{aligned} \tag{3.53}$$

To bound the first term in (3.53) from below, it is convenient to rewrite the term $\partial_t\vartheta(u_h)u_h$ into a conservation form. Namely, taking into account that $u_h(t) = u_h(\boldsymbol{x})(t) \in S_{h,p}$, $t \in [0, T]$, using the Leibnitz integral rule (3.37) we get

$$\partial_t \int_{u_h(t)}^{u_h(0)} \vartheta(\chi)\mathrm{d}\chi = \vartheta(u_h(0))\partial_t u_h(0) - \vartheta(u_h(t))\partial_t u_h(t) = -\vartheta(u_h(t))\partial_t u_h(t). \tag{3.54}$$

Next, we note that by the product rule we have

$$\begin{aligned}\partial_t \int_{u_h(t)}^{u_h(0)} \vartheta(u_h(t))\mathrm{d}\chi = \partial_t\bigg(\vartheta(u_h(t)) \int_{u_h(t)}^{u_h(0)} \mathrm{d}\chi\bigg) \\ = \partial_t\vartheta(u_h(t))(u_h(0) - u_h(t)) - \vartheta(u_h(t))\partial_t u_h(t).\end{aligned} \tag{3.55}$$

We subtract (3.55) from (3.54) and rearrange the terms such that

$$\partial_t\vartheta(u_h)u_h = \partial_t\bigg(\int_{u_h}^{u_h(0)} (\vartheta(\chi) - \vartheta(u_h))\mathrm{d}\chi + \vartheta(u_h)u_h(0)\bigg). \tag{3.56}$$

Moreover, the relation (3.38a) from Lemma 4 implies the inequality

$$\int_\Omega \left(\int_{u_h(T)}^{u_h(0)} \Big(\vartheta(\chi) - \vartheta(u_h(T))\Big) \mathrm{d}\chi \right) dx \geq M_\vartheta \int_\Omega \Big(\vartheta(u_h(0)) - \vartheta(u_h(T))\Big)^2 \,\mathrm{d}x, \tag{3.57}$$

where $M_\vartheta = \Big(2 \operatorname{ess\,sup}_{u\in\mathbb{R}} |\vartheta'(u)|\Big)^{-1}$. Finally, we combine (3.56)–(3.57) to get the lower bound of the first term of (3.53)

$$\begin{aligned}
&\int_0^T (\partial_t \vartheta(u_h), u_h) \,\mathrm{d}t \\
&= \int_\Omega \left(\int_{u_h(T)}^{u_h(0)} \Big(\vartheta(\chi) - \vartheta(u_h(T))\Big) \mathrm{d}\chi + \vartheta(u_h(T)) u_h(0) - \vartheta(u_h(0)) u_h(0) \right) \mathrm{d}x \\
&\geq M_\vartheta \|\vartheta(u_h(T)) - \vartheta(u_h(0))\|^2_{L^2(\Omega)} + \Big(\vartheta(u_h(T)) - \vartheta(u_h(0)), u_h(0)\Big).
\end{aligned} \tag{3.58}$$

Furthermore, we use (3.32) to obtain the lower bound for the second term in (3.53)

$$\int_0^T B_h(u_h; \boldsymbol{q}_h, \boldsymbol{q}_h) \,\mathrm{d}t \geq k_0 \|\boldsymbol{q}_h\|^2_{L^2(0,T;L^2(\Omega))}. \tag{3.59}$$

At this moment, we apply the relations we obtained so far, i.e., (3.58)–(3.59), to (3.53),

$$\begin{aligned}
M_\vartheta \|\vartheta(u_h(T)) - \vartheta(u_h(0))\|^2_{L^2(\Omega)} + k_0 \|\boldsymbol{q}_h\|^2_{L^2(0,T;L^2(\Omega))} + \int_0^T J_h(u_h, u_h) \,\mathrm{d}t \\
\leq \Big|(\vartheta(u_h(T)) - \vartheta(u_h(0)), u_h(0))\Big| + \int_0^T \Big| F_h(u_h) + G_h(\boldsymbol{\sigma}_h)\Big| \,\mathrm{d}t.
\end{aligned} \tag{3.60}$$

In the rest of the proof, we aim to find upper bounds for the terms on the right-hand side on (3.60). The Cauchy-Schwarz inequality and the Young inequality imply

$$\begin{aligned}
&\Big|(\vartheta(u_h(T)) - \vartheta(u_h(0)), u_h(0))\Big| \\
&\leq \frac{M_\vartheta}{2} \|\vartheta(u_h(T)) - \vartheta(u_h(0))\|^2_{L^2(\Omega)} + \frac{1}{2M_\vartheta} \|u_h(0)\|^2_{L^2(\Omega)}.
\end{aligned} \tag{3.61}$$

From Definition 5, the definition of the L^2-projection (2.8) and the Cauchy-Schwarz inequality we have that

$$\|u_h(0)\|_{L^2(\Omega)} \leq \|\Pi_{h,p} u_0\|_{L^2(\Omega)} \leq \|u_0\|_{L^2(\Omega)}. \tag{3.62}$$

Furthermore, using the Cauchy-Schwarz inequality, the relation (3.44) and the Young inequality we deduce that

$$\begin{aligned}
|(g, u_h)| &\leq \|g\|_{L^2(\Omega)} \|u_h\|_{L^2(\Omega)} \\
&\leq \frac{1}{8} J_h(u_h, u_h) + \frac{k_0}{8} \|\boldsymbol{q}_h\|^2_{L^2(\Omega)} + \frac{1}{2} J_h^D(u_D) + C_1 \|g\|^2_{L^2(\Omega)},
\end{aligned} \tag{3.63}$$

where $C_1 := C_u^2(5/2 + 2/k_0)$. We estimate the jump term using (3.24)

$$|J_h^D(u_h)| \leq J_h^{1/2}(u_h, u_h) J_h^D(u_D)^{1/2} \leq \frac{1}{8} J_h(u_h, u_h) + 2 J_h^D(u_D). \tag{3.64}$$

Then, we get an estimate on the Neumann term contained in $F_h(u_h)$ by using the Cauchy-Schwarz inequality, the generalized trace inequality (3.27), and the broken Poincaré inequality (3.28)

$$|(g_N, u_h)_N| \leq \|g_N\|_{L^2(\partial\Omega_N)} \|u_h\|_{L^2(\partial\Omega_N)} \leq \sqrt{2C_N C_P} \|g_N\|_{L^2(\partial\Omega_N)} |\!|\!| u_h |\!|\!|. \tag{3.65}$$

From Lemma 5 and the relation (3.65) in combination with the Young inequality, it follows that

$$|(g_N, u_h)_N| \leq \frac{1}{4} J_h(u_h, u_h) + \frac{k_0}{8} \|\boldsymbol{q}_h\|^2_{L^2(\Omega)} + \frac{1}{2} J^D_h(u_D) + C_2 \|g_N\|^2_{L^2(\partial\Omega_N)}, \tag{3.66}$$

where $C_2 := C_N C_P C_u^2 (3/2 + 2/k_0)$. Similarly as it was showed in the proof of Lemma 5 (cf. (3.48)–(3.49), but now with $\boldsymbol{\sigma}_h$ instead of ∇u_h) and (3.45), we deduce

$$|G_h(\boldsymbol{\sigma}_h)| \leq C_1 \|\boldsymbol{\sigma}_h\|_{L^2(\Omega)} J^D_h(u_D)^{1/2} \leq \frac{k_0}{4} \|\boldsymbol{q}_h\|^2_{L^2(\Omega)} + C_3^2 J^D_h(u_D), \tag{3.67}$$

where $C_3 := C_W C_M (C_I + 1)/C_G$.

We combine (3.63), (3.64), (3.66), and (3.67) to get a bound for the integrand in (3.60),

$$\begin{aligned} &|F_h(u_h) + G_h(\boldsymbol{\sigma}_h)| \\ &\leq \frac{1}{2} J_h(u_h, u_h) + \frac{k_0}{2} \|\boldsymbol{q}_h\|^2_{L^2(\Omega)} + C_4 J^D_h(u_D) + C_2 \|g_N\|^2_{L^2(\partial\Omega_N)} + C_1 \|g\|^2_{L^2(\Omega)}, \end{aligned} \tag{3.68}$$

where $C_4 := 5/2 + C_3^2$. Finally, we get the statement of the theorem by incorporating (3.68), (3.62) and (3.61) into (3.60), and setting $C := 2C_5 / \min(M_\vartheta, k_0, 1)$, where $C_5 := \max(1/M_\vartheta, C_4, C_2, C_1)$.

□

3.5 Error estimates

In what follows we derive the error estimate for the semidiscrete numerical scheme (3.19). We start with forming the error equation. To do so, we subtract (3.20a)–(3.20c) from (3.19a)–(3.19c),

$$(\partial_t(\vartheta(u_h) - \vartheta(u)), v_h) + A_h(\boldsymbol{\sigma}_h - \boldsymbol{\sigma}, v_h) + J_h(u_h - u, v_h) = 0, \tag{3.69a}$$

$$B_h(u_h; \boldsymbol{q}_h, \boldsymbol{w}_h) - B_h(u; \boldsymbol{q}, \boldsymbol{w}_h) - (\boldsymbol{\sigma}_h - \boldsymbol{\sigma}, \boldsymbol{w}_h) = 0, \tag{3.69b}$$

$$(\boldsymbol{q}_h - \boldsymbol{q}, \boldsymbol{z}_h) - A_h(\boldsymbol{z}_h, u_h - u) = 0, \tag{3.69c}$$

where $v_h \in S_{h,p}$, $\boldsymbol{w}_h, \boldsymbol{z}_h \in \boldsymbol{S}_{h,p}$ and $t \in [0, T]$ is fixed. As in the finite element analysis, we decompose the error $(e_u, \boldsymbol{e_q}, \boldsymbol{e_\sigma})$ using the L^2-projection (2.8) as

$$(e_u, \boldsymbol{e_q}, \boldsymbol{e_\sigma}) = (\xi_u + \eta_u, \boldsymbol{\xi_q} + \boldsymbol{\eta_q}, \boldsymbol{\xi_\sigma} + \boldsymbol{\eta_\sigma}),$$

where

$$\xi_u = u_h - \Pi_{h,p} u \in S_{h,p}, \quad \eta_u = \Pi_{h,p} u - u, \tag{3.70a}$$

$$\boldsymbol{\xi_q} = \boldsymbol{q}_h - \Pi_{h,p} \boldsymbol{q} \in \boldsymbol{S}_{h,p}, \quad \boldsymbol{\eta_q} = \Pi_{h,p} \boldsymbol{q} - \boldsymbol{q}, \tag{3.70b}$$

$$\boldsymbol{\xi_\sigma} = \boldsymbol{\sigma}_h - \Pi_{h,p} \boldsymbol{\sigma} \in \boldsymbol{S}_{h,p}, \quad \boldsymbol{\eta_\sigma} = \Pi_{h,p} \boldsymbol{\sigma} - \boldsymbol{\sigma}. \tag{3.70c}$$

Let us emphasize that the variables above depend on t; however, for simplicity, we omit the time dependence. We substitute the error decomposition (3.70a)–(3.70c) into (3.69a)–(3.69c), and use the definition of the L^2-projection obtaining the error equations which will be used in the further analysis

$$(\partial_t(\vartheta(u_h) - \vartheta(u)), v_h) + A_h(\boldsymbol{\xi_\sigma}, v_h) + J_h(\xi_u, v_h) = -A_h(\boldsymbol{\eta_\sigma}, v_h) - J_h(\eta_u, v_h), \tag{3.71a}$$

$$B_h(u_h; \boldsymbol{q}_h, \boldsymbol{w}_h) - B_h(u; \boldsymbol{q}, \boldsymbol{w}_h) - (\boldsymbol{\xi_\sigma}, \boldsymbol{w}_h) = 0 \tag{3.71b}$$

$$(\boldsymbol{\xi_q}, \boldsymbol{z}_h) - A_h(\boldsymbol{z}_h, \xi_u) = A_h(\boldsymbol{z}_h, \eta_u), \tag{3.71c}$$

where $v_h \in S_{h,p},\ \boldsymbol{w}_h, \boldsymbol{z}_h \in \boldsymbol{S}_{h,p}$.

Later in Subsection 3.5.4, it will be seen that the technique that uses the standard test function $(\xi_u, \boldsymbol{\xi_q}, \boldsymbol{\xi_\sigma})$ in (3.69a)–(3.69c) gives an incomplete estimate due to the nonlinear function ϑ. In order to get a complete error bound, in Subsection 3.5.5 we introduce modified test functions based on ξ_u, $\boldsymbol{\xi_q}$ and $\boldsymbol{\xi_\sigma}$ and derive a new incomplete bound for the problematic nonlinear term from the previous estimate. Finally, in Subsection 3.5.6, we couple these estimates by continuous mathematical induction and develop the final estimate. First, we present some results that shall be used to obtain the partial estimates.

3.5.1 Properties of the form A_h

Lemma 7. *Let $A_h(\cdot\ ,\cdot)$ be defined by* (3.13)*. Then,*

$$|A_h(\boldsymbol{w}, v_h)| \le R_a(\boldsymbol{w})\|v_h\|_{L^2(\Omega)}, \qquad \boldsymbol{w} \in \boldsymbol{H}^1(\Omega, \mathcal{T}_h),\ v_h \in S_{h,p}, \tag{3.72}$$

$$|A_h(\boldsymbol{w}_h, v)| \le C_a\|\boldsymbol{w}_h\|_{L^2(\Omega)}|\!|\!|v|\!|\!|, \qquad \boldsymbol{w}_h \in \boldsymbol{S}_{h,p},\ v \in H^1(\Omega, \mathcal{T}_h), \tag{3.73}$$

with

$$R_a(\boldsymbol{w}) := C_P\Big(\|\boldsymbol{w}\|_{L^2(\Omega)} + R^{1/2}(\boldsymbol{w})\Big), \quad \boldsymbol{w} \in \boldsymbol{H}^1(\Omega, \mathcal{T}_h). \tag{3.74}$$

where C_a and $R(\boldsymbol{w})$ are the constant and the form (3.31) *from Lemma 2, respectively.*

Proof. Let $\boldsymbol{w} \in \boldsymbol{H}^1(\Omega, \mathcal{T}_h)$, $v \in H^1(\Omega, \mathcal{T}_h)$, then, by the definition (3.13) we have that

$$A_h(\boldsymbol{w}, v) = (\boldsymbol{w}, \nabla v) - \sum_{\Gamma \in \mathcal{F}_h^{ID}} \int_\Gamma \Big(\langle \boldsymbol{w} \rangle \cdot \boldsymbol{n} - \boldsymbol{\lambda} \cdot \boldsymbol{n}[\boldsymbol{w} \cdot \boldsymbol{n}]\Big)[v]\,\mathrm{d}S.$$

The Cauchy-Schwarz inequality implies

$$|A_h(\boldsymbol{w}, v)| \le \|\boldsymbol{w}\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} + \Bigg(\sum_{\Gamma \in \mathcal{F}_h^{ID}} \int_\Gamma \kappa^{-1}\Big|\langle \boldsymbol{w} \rangle - \boldsymbol{\lambda} \cdot \boldsymbol{n}[\boldsymbol{w}]\Big|^2 \mathrm{d}S\Bigg)^{1/2} J_h^{1/2}(v, v). \tag{3.75}$$

Hence, if $\boldsymbol{w} \in \boldsymbol{H}^1(\Omega, \mathcal{T}_h)$ and $v := v_h \in S_{h,p}$, then by (3.29), definition of the $|\!|\!|\cdot|\!|\!|$-norm (3.26), and the broken Poincaré inequality (3.28) we get (3.72). Conversely, if $\boldsymbol{w} := \boldsymbol{w}_h \in \boldsymbol{S}_{h,p}$ and $v \in H^1(\Omega, \mathcal{T}_h)$, then by (3.30) we obtain (3.73). □

3.5.2 Properties of the form B_h

We observe some identities related to the form B_h which can be shown by a simple rearrangement.

$$\begin{aligned}
&B_h(u_h; \boldsymbol{q}_h, \boldsymbol{w}_h) - B_h(u; \boldsymbol{q}, \boldsymbol{w}_h) \\
&\quad = B_h(u_h; \boldsymbol{\xi}_{\boldsymbol{q}}, \boldsymbol{w}_h) + B_h(u_h; \boldsymbol{\eta}_{\boldsymbol{q}}, \boldsymbol{w}_h) + \Big(B_h\Big(u_h; \Pi_{h,p}\boldsymbol{q}, \boldsymbol{w}_h) - B_h(u; \Pi_{h,p}\boldsymbol{q}, \boldsymbol{w}_h)\Big),
\end{aligned} \tag{3.76}$$

$$\begin{aligned}
&B_h(u_h; \boldsymbol{q}_h, \boldsymbol{w}_h) - B_h(u; \boldsymbol{q}, \boldsymbol{w}_h) \\
&\quad = B_h(u; \boldsymbol{\xi}_{\boldsymbol{q}}, \boldsymbol{w}_h) + B_h(u; \boldsymbol{\eta}_{\boldsymbol{q}}, \boldsymbol{w}_h) + \Big(B_h(u_h; \boldsymbol{\xi}_{\boldsymbol{q}}, \boldsymbol{w}_h) - B_h(u; \boldsymbol{\xi}_{\boldsymbol{q}}, \boldsymbol{w}_h)\Big) \\
&\qquad + \Big(B_h(u_h; \Pi_{h,p}\boldsymbol{q}, \boldsymbol{w}_h) - B_h(u; \Pi_{h,p}\boldsymbol{q}, \boldsymbol{w}_h)\Big).
\end{aligned} \tag{3.77}$$

Therefore, we propose some estimates concerning the above identities.

Lemma 8. *Let $B_h(\cdot\,;\cdot\,,\cdot)$ be defined by* (3.14) *and $\boldsymbol{\xi}_{\boldsymbol{q}}$ be given by* (3.70b)*. Then, for any $\boldsymbol{w}_h \in \boldsymbol{S}_{h,p}$ we have*

$$\begin{aligned}
&|B_h(u_h; \boldsymbol{\xi}_{\boldsymbol{q}}, \boldsymbol{w}_h) - B_h(u; \boldsymbol{\xi}_{\boldsymbol{q}}, \boldsymbol{w}_h)| \\
&\qquad \le C_b h^{-1} \|\theta(u_h) - \theta(u)\|_{L^2(\Omega)} \big\|\boldsymbol{\xi}_{\boldsymbol{q}}\big\|_{L^2(\Omega)} \|\boldsymbol{w}_h\|_{L^2(\Omega)},
\end{aligned} \tag{3.78}$$

$$\begin{aligned}
&|B_h(u_h; \Pi_{h,p}\boldsymbol{q}, \boldsymbol{w}_h) - B_h(u; \Pi_{h,p}\boldsymbol{q}, \boldsymbol{w}_h)| \\
&\qquad \le C_c \Big(\|\theta(u_h) - \theta(u)\|_{L^2(\Omega)} + \big\|\boldsymbol{\eta}_{\boldsymbol{q}}\big\|_{L^2(\Omega)}\Big) \|\boldsymbol{w}_h\|_{L^2(\Omega)},
\end{aligned} \tag{3.79}$$

where $C_b, C_c > 0$ are constants independent of the discretization parameter h.

Proof. To prove the first statement, we use the definition of B_h, the generalized Hölder inequality, Lipschitz continuity of $\mathbf{K}$ in θ (cf. (A1)) and the Cauchy-Schwarz inequality, so we have that

$$\begin{aligned}
&\Big|\Big((\mathbf{K}(\theta(u_h)) - \mathbf{K}(\theta(u)))\boldsymbol{\xi}_{\boldsymbol{q}}, \boldsymbol{w}_h\Big)\Big| \\
&\qquad \le \left(k_L^2 \sum_{K \in \mathcal{T}_h} \|\theta(u_h) - \theta(u)\|^2_{L^2(K)} \big\|\boldsymbol{\xi}_{\boldsymbol{q}}\big\|^2_{L^\infty(K)}\right)^{1/2} \|\boldsymbol{w}_h\|^2_{L^\infty(K)}.
\end{aligned} \tag{3.80}$$

Moreover, the inverse inequality (3.21b) and quasi-uniformity of the mesh $\mathcal{T}_h$ (2.5) imply that

$$\big\|\boldsymbol{\xi}_{\boldsymbol{q}}\big\|^2_{L^\infty(K)} \le C_I^2 h_K^{-2} \big\|\boldsymbol{\xi}_{\boldsymbol{q}}\big\|^2_{L^2(K)} \le C_I^2 C_U^2 h^{-2} \big\|\boldsymbol{\xi}_{\boldsymbol{q}}\big\|^2_{L^2(K)} \quad \forall K \in \mathcal{T}_h. \tag{3.81}$$

The relation (3.81) and the identity for finite sums $\sum_i a_i^2 b_i^2 \le \sum_i a_i^2 \sum_j b_j^2$ yield

$$\sum_{K \in \mathcal{T}_h} \|\theta(u_h) - \theta(u)\|^2_{L^2(K)} \big\|\boldsymbol{\xi}_{\boldsymbol{q}}\big\|^2_{L^\infty(K)} \le C_I^2 C_U^2 h^{-2} \|\theta(u_h) - \theta(u)\|^2_{L^2(\Omega)} \big\|\boldsymbol{\xi}_{\boldsymbol{q}}\big\|^2_{L^2(\Omega)}. \tag{3.82}$$

We complete the proof of the first part of the lemma by inserting (3.82) in (3.80) and setting $C_b := k_L C_I C_U$.

Conversely, to prove the second part, we rewrite

$$\begin{aligned}&|B_h(u_h;\Pi_{h,p}\boldsymbol{q}_h,\boldsymbol{w}_h)-B_h(u;\Pi_{h,p}\boldsymbol{q}_h,\boldsymbol{w}_h)|\\&\quad\le|B_h(u_h;\boldsymbol{\eta}_{\boldsymbol{q}},\boldsymbol{w}_h)-B_h(u;\boldsymbol{\eta}_{\boldsymbol{q}},\boldsymbol{w}_h)|+|B_h(u_h;\boldsymbol{q},\boldsymbol{w}_h)-B_h(u;\boldsymbol{q},\boldsymbol{w}_h)|,\end{aligned}\tag{3.83}$$

and then estimate each of the terms appearing on the right-hand side of (3.83). Namely, the assumption (A1) and the Cauchy-Schwarz inequality yield

$$|B_h(u_h;\boldsymbol{\eta}_{\boldsymbol{q}},\boldsymbol{w}_h)-B_h(u;\boldsymbol{\eta}_{\boldsymbol{q}},\boldsymbol{w}_h)|\le 2k_1\left\|\boldsymbol{\eta}_{\boldsymbol{q}}\right\|_{L^2(\Omega)}\|\boldsymbol{w}_h\|_{L^2(\Omega)}.\tag{3.84}$$

Furthermore, using the assumption (B2) and the generalized Hölder inequality, we obtain

$$|B_h(u_h;\boldsymbol{q},\boldsymbol{w}_h)-B_h(u;\boldsymbol{q},\boldsymbol{w}_h)|\le C_B\|\theta(u_h)-\theta(u)\|_{L^2(\Omega)}\|\boldsymbol{w}_h\|_{L^2(\Omega)}.\tag{3.85}$$

We get the final statement by combining the relations (3.83)–(3.85) and setting $C_c := \max(2k_1, C_B)$ in (3.79).

□

3.5.3 Relation between ξ_u, $\boldsymbol{\xi}_{\boldsymbol{q}}$ and $\boldsymbol{\xi}_{\boldsymbol{\sigma}}$

In the next lemma, we present a relation between ξ_u, $\boldsymbol{\xi}_{\boldsymbol{q}}$, and $\boldsymbol{\xi}_{\boldsymbol{\sigma}}$, which will simplify the analysis.

Lemma 9. *Let* $(\xi_u,\boldsymbol{\xi}_{\boldsymbol{q}},\boldsymbol{\xi}_{\boldsymbol{\sigma}})$ *be given by* (3.70)*. Then, there exist* $C_d, C_e > 0$ *such that*

$$\|\boldsymbol{\xi}_{\boldsymbol{\sigma}}\|_{L^2(\Omega)}\le C_d\Big(\left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|_{L^2(\Omega)}+\left\|\boldsymbol{\eta}_{\boldsymbol{q}}\right\|_{L^2(\Omega)}+\|\vartheta(u_h)-\vartheta(u)\|_{L^2(\Omega)}\Big),\tag{3.86}$$

$$\|\xi_u\|_{L^2(\Omega)}\le C_e\Big(\left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|_{L^2(\Omega)}+J_h^{1/2}(\xi_u,\xi_u)+|||\eta_u|||\Big).\tag{3.87}$$

Proof. We set $\boldsymbol{w}_h := \boldsymbol{\xi}_{\boldsymbol{\sigma}}$ in (3.71b) so that

$$\|\boldsymbol{\xi}_{\boldsymbol{\sigma}}\|^2_{L^2(\Omega)}=|B_h(u_h;\boldsymbol{q}_h,\boldsymbol{\xi}_{\boldsymbol{\sigma}})-B_h(u_h;\boldsymbol{q},\boldsymbol{\xi}_{\boldsymbol{\sigma}})|.\tag{3.88}$$

Then, we use the identity (3.76) and bound the individual terms on the right-hand side. By the relations (3.33) and (3.79), we have that

$$\begin{aligned}&|B_h(u_h;\boldsymbol{\xi}_{\boldsymbol{q}},\boldsymbol{\xi}_{\boldsymbol{\sigma}})|+|B_h(u_h;\boldsymbol{\eta}_{\boldsymbol{q}},\boldsymbol{\xi}_{\boldsymbol{\sigma}})|+|B_h(u_h;\Pi_{h,p}\boldsymbol{q},\boldsymbol{\xi}_{\boldsymbol{\sigma}})-B_h(u;\Pi_{h,p}\boldsymbol{q},\boldsymbol{\xi}_{\boldsymbol{\sigma}})|\\&\le\Big(k_1\Big(\left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|_{L^2(\Omega)}+\left\|\boldsymbol{\eta}_{\boldsymbol{q}}\right\|_{L^2(\Omega)}\Big)+C_d\Big(\|\theta(u_h)-\theta(u)\|_{L^2(\Omega)}+\left\|\boldsymbol{\eta}_{\boldsymbol{q}}\right\|_{L^2(\Omega)}\Big)\Big)\|\boldsymbol{\xi}_{\boldsymbol{\sigma}}\|_{L^2(\Omega)}.\end{aligned}\tag{3.89}$$

To prove (3.86), we combine (3.88) with (3.89), use (3.6), cancel out $\|\boldsymbol{\xi}_{\boldsymbol{\sigma}}\|_{L^2(\Omega)}$ and set $C_d := \max(k_1, C_c)$.

On the other hand, to show the other part of the lemma, we set $\boldsymbol{z}_h := \nabla\xi_u$ in (3.71c) and perform a similar procedure as in the proof of Lemma 5 for the relation (3.44),

$$\begin{aligned}\|\nabla\xi_u\|^2_{L^2(\Omega)}=\Bigg|&\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma\Big(\langle\nabla\xi_u\rangle\cdot\boldsymbol{n}-\boldsymbol{\lambda}\cdot\boldsymbol{n}[\nabla\xi_u\cdot\boldsymbol{n}]\Big)[\xi_u]\,\mathrm{d}S\\&+(\boldsymbol{\xi}_{\boldsymbol{q}},\nabla\xi_u)-A_h(\nabla\xi_u,\eta_u)\Bigg|.\end{aligned}\tag{3.90}$$

By the Cauchy-Schwarz inequality and (3.30) we have

$$\begin{aligned}
&\Big|\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma\Big(\langle\nabla\xi_u\rangle\cdot\boldsymbol{n}-\boldsymbol{\lambda}\cdot\boldsymbol{n}[\nabla\xi_u\cdot\boldsymbol{n}]\Big)[\xi_u]\,\mathrm{d}S\Big|\\
&\le\Big(\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma\kappa^{-1}\Big|\langle\nabla\xi_u\rangle-\boldsymbol{\lambda}\cdot\boldsymbol{n}[\nabla\xi_u]\Big|^2\,\mathrm{d}S\Big)^{1/2}J_h^{1/2}(\xi_u,\xi_u)\\
&\le C_a\|\nabla\xi_u\|_{L^2(\Omega)}J_h^{1/2}(\xi_u,\xi_u).
\end{aligned}\tag{3.91}$$

Then, we apply (3.91), the Cauchy-Schwarz inequality, (3.73), and cancel out $\|\nabla\xi_u\|_{L^2(\Omega)}$ so that

$$\|\nabla\xi_u\|_{L^2(\Omega)}\le C_aJ_h^{1/2}(\xi_u,\xi_u)+\Big\|\boldsymbol{\xi_q}\Big\|_{L^2(\Omega)}+C_a|\!|\!|\eta_u|\!|\!|.\tag{3.92}$$

We complete the proof by adding $J_h^{1/2}(\xi_u,\xi_u)$ to both sides of (3.92) in order to use the $|\!|\!|\cdot|\!|\!|$-norm and the broken Poincaré inequality (3.28). Lastly, we set $C_e:=C_P(C_a+1)$ in (3.87).

□

3.5.4 Partial error estimate in the $L^2(\Omega)$-norm

Since the unknown solution u in our problem (3.1) is the independent variable of the nonlinear function ϑ, the numerical analysis of the time-continuous scheme (3.71) is rather challenging. Therefore, to treat the problematic term $\partial_t\vartheta(u)$, we use a nonstandard approach from [2] and [96]. In this way, we derive an incomplete error bound with respect to the Hölder coefficient β (cf. (A2)). Before giving the first main result, we propose some lemmas. In order to later implicitly use Gronwall's lemma, the results in this subsection contain a monotone decreasing exponential function. We recall that Gronwall's lemma is a commonly used tool in analysis of method of lines; see [3, 33, 80].

Lemma 10. *Let $t\in[0,T]$ and $\delta,\bar{Q}>0$ be arbitrary. Then,*

$$\begin{aligned}
&\int_0^t\Big(\partial_s(\vartheta(u_h)-\vartheta(u)),u_h-u\Big)e^{-\bar{Q}s}\mathrm{d}s\\
&\quad\ge M_\vartheta\|\vartheta(u_h)-\vartheta(u)\|^2_{L^2(\Omega)}e^{-\bar{Q}t}-L_\vartheta\|\eta_u(0)\|^2_{L^2(\Omega)}\\
&\qquad-\int_0^t\int_\Omega\Big((\vartheta(u_h)-\vartheta(u))\partial_su-\partial_s\vartheta(u)(u_h-u)\Big)e^{-\bar{Q}s}\,\mathrm{d}x\mathrm{d}s\\
&\qquad+\bar{Q}M_\vartheta\int_0^t\|\vartheta(u_h)-\vartheta(u)\|^2_{L^2(\Omega)}e^{-\bar{Q}s}\mathrm{d}s
\end{aligned}\tag{3.93}$$

and

$$\begin{aligned}
&\Big|(\vartheta(u_h)-\vartheta(u))\partial_tu-\partial_t\vartheta(u)(u_h-u)\Big|\\
&\quad\le H_\vartheta C_X\left(\frac{1+\beta}{2\delta^{\frac{1-\beta}{1+\beta}}}\Big((\vartheta(u_h)-\vartheta(u))(u_h-u)\Big)^{\frac{2\beta}{1+\beta}}+\delta(1-\beta)\Big(|\xi_u|^2+|\eta_u|^2\Big)\right).
\end{aligned}\tag{3.94}$$

Proof. As in the proof of Theorem 6 (cf. (3.54)–(3.55)) but now with both integral limits dependent on time (and space) variable, using the Leibnitz integral rule (3.37), we obtain

$$\partial_t \int_{u_h}^{u} \vartheta(\chi)\mathrm{d}\chi = \vartheta(u)\partial_t u - \vartheta(u_h)\partial_t u_h, \tag{3.95}$$

$$\partial_t \int_{u_h}^{u} \vartheta(u_h)\mathrm{d}\chi = \partial_t\vartheta(u_h) \int_{u_h}^{u} \mathrm{d}\chi = \partial_t\vartheta(u_h)(u-u_h) + \vartheta(u_h)(\partial_t u - \partial_t u_h). \tag{3.96}$$

If we subtact (3.96) from (3.95) and use some manipulations, we get

$$\begin{aligned}\partial_t\biggl(\int_{u_h}^{u}(\vartheta(\chi)-\vartheta(u_h))\mathrm{d}\chi\biggr) &= (\vartheta(u)-\vartheta(u_h))\partial_t u - \partial_t\vartheta(u_h)(u-u_h)\\ &= -(\vartheta(u_h)-\vartheta(u))\partial_t u + \partial_t(\vartheta(u_h)-\vartheta(u))(u_h-u)\\ &\quad + \partial_t\vartheta(u)(u_h-u).\end{aligned}$$

Therefore, for arbitrary but fixed $\bar{Q} > 0$ and $t \in [0,T]$ by the product rule we deduce

$$\begin{aligned}&\partial_t(\vartheta(u_h)-\vartheta(u))(u_h-u)e^{-\bar{Q}t}\\ &= \partial_t\biggl(\int_{u_h}^{u}(\vartheta(\chi)-\vartheta(u_h))\mathrm{d}\chi e^{-\bar{Q}t}\biggr) + \bar{Q}\int_{u_h}^{u}(\vartheta(\chi)-\vartheta(u_h))\mathrm{d}\chi e^{-\bar{Q}t}\\ &\quad - \Bigl((\vartheta(u_h)-\vartheta(u))\partial_t u - \partial_t\vartheta(u)(u_h-u)\Bigr)e^{-\bar{Q}t}.\end{aligned} \tag{3.97}$$

Then, for an arbitrary fixed $t \in [0,T]$ we integrate (3.97) over $\Omega \times (0,t)$

$$\begin{aligned}&\int_0^t \Bigl(\partial_s(\vartheta(u_h)-\vartheta(u)), u_h-u\Bigr)e^{-\bar{Q}s}\mathrm{d}s\\ &= \int_\Omega\int_{u_h(t)}^{u(\cdot,t)}(\vartheta(\chi)-\vartheta(u_h))\mathrm{d}\chi e^{-\bar{Q}t}\,\mathrm{d}x - \int_\Omega\int_{u_h(0)}^{u(\cdot,0)}(\vartheta(\chi)-\vartheta(u_h))\mathrm{d}\chi\,\mathrm{d}x\\ &\quad - \int_0^t\int_\Omega\Bigl((\vartheta(u_h)-\vartheta(u))\partial_s u - \partial_s\vartheta(u)(u_h-u)\Bigr)e^{-\bar{Q}s}\,\mathrm{d}x\mathrm{d}s\\ &\quad + \bar{Q}\int_0^t\int_\Omega\int_{u_h}^{u}(\vartheta(\chi)-\vartheta(u_h))\mathrm{d}\chi e^{-\bar{Q}s}\,\mathrm{d}x\mathrm{d}s.\end{aligned} \tag{3.98}$$

In what follows, we shall find upper bounds for the terms on the right hand side of (3.98). Namely, by Lemma 4 we have that

$$\begin{aligned}&\int_\Omega\int_{u_h(t)}^{u(\cdot,t)}(\vartheta(\chi)-\vartheta(u_h))\mathrm{d}\chi e^{-\bar{Q}t}\,\mathrm{d}x - \int_\Omega\int_{u_h(0)}^{u(\cdot,0)}(\vartheta(\chi)-\vartheta(u_h))\mathrm{d}\chi\,\mathrm{d}x\\ &\geq M_\vartheta\|\vartheta(u_h)-\vartheta(u)\|^2_{L^2(\Omega)}e^{-\bar{Q}t} - (\vartheta(u_h(0))-\vartheta(u_0), u_h(0)-u_0),\end{aligned} \tag{3.99}$$

and

$$\bar{Q}\int_0^t\int_\Omega\int_{u_h}^{u}(\vartheta(\mu)-\vartheta(u_h))\mathrm{d}\mu e^{-\bar{Q}s}\,\mathrm{d}x\mathrm{d}s \geq \bar{Q}M_\vartheta\int_0^t\|\vartheta(u_h)-\vartheta(u)\|^2_{L^2(\Omega)}e^{-\bar{Q}s}\mathrm{d}s. \tag{3.100}$$

Furthermore, Definition 5 implies that $u_h(0) - \Pi_{hp}u_0 = 0$, so by virtue of the Minkowski inequality we have that

$$\|u_h(0)-u_0\|_{L^2(\Omega)} \leq \|\eta_u(0)\|_{L^2(\Omega)}. \tag{3.101}$$

Hence, the Cauchy-Schwarz inequality, Lipschitz continuity of ϑ (cf. Remark 5), and (3.101) yield

$$(\vartheta(u_h(0)) - \vartheta(u_0), u_h(0) - u_0) \leq L_\vartheta \|u_h(0) - u_0\|^2_{L^2(\Omega)} \leq L_\vartheta \|\eta_u(0)\|^2_{L^2(\Omega)}. \tag{3.102}$$

Finally, we substitute (3.99), (3.100) and (3.102) into (3.98), and prove the first part of the statement.

Conversely, to prove the second part of the lemma, we use the mean value theorem; namely, there exists some $w \in [\min(u_h, u), \max(u_h, u)]$ such that $\vartheta(u_h) - \vartheta(u) = \vartheta'(w)(u_h - u)$. Then, from monotonocity of ϑ (cf. Remark 5), and assumptions (A4) and (B1), we have that

$$\begin{aligned}\Big|(\vartheta(u_h) - \vartheta(u))\partial_t u - \partial_t \vartheta(u)(u_h - u)\Big| &= \Big|\vartheta'(w)(u_h - u)\partial_t u - \vartheta'(u)\partial_t u(u_h - u)\Big| \\ &= \Big|\Big(\vartheta'(w) - \vartheta'(u)\Big)(u_h - u)\partial_t u\Big| \\ &\leq H_\vartheta C_X |\vartheta(w) - \vartheta(u)|^\beta |u_h - u| \\ &\leq H_\vartheta C_X |\vartheta(u_h) - \vartheta(u)|^\beta |u_h - u|.\end{aligned} \tag{3.103}$$

Moreover, we use the inequality

$$|A|^\beta |B| \leq \frac{1+\beta}{2\delta^{\frac{1-\beta}{1+\beta}}} |AB|^{\frac{2\beta}{1+\beta}} + \delta \frac{1-\beta}{2} |B|^2, \quad A, B \in \mathbb{R},\ \delta > 0,\ 0 < \beta \leq 1, \tag{3.104}$$

which can be obtained by applying the standard Young inequality

$$ab \leq \frac{\delta}{p} a^p + \frac{1}{\delta^{q/p} q} b^2,\ a, b \in \mathbb{R},\ \frac{1}{p} + \frac{1}{q} = 1,\ p, q \geq 1,$$

to $|A|^\beta |B| = |AB|^\beta |B|^{1-\beta}$ by setting $a = |AB|^\beta$, $b = |B|^{1-\beta}$, $p = 2/(1+\beta)$, and $q = 2/(1-\beta)$. Ultimately, we set $A = \vartheta(u_h) - \vartheta(u)$ and $B = u_h - u$ in (3.104), and substitute into (3.103) using (3.23), which yields (3.94).

□

Lemma 11. *Let $t \in [0, T]$ and $\bar{Q} > 0$ be arbitrary. Then,*

$$\begin{aligned}&\int_0^t \Big(\partial_s(\vartheta(u_h) - \vartheta(u)), \eta_u\Big) e^{-\bar{Q}s} \mathrm{d}s \\ &\leq \|\vartheta(u_h) - \vartheta(u)\|_{L^2(\Omega)} \|\eta_u\|_{L^2(\Omega)} e^{-\bar{Q}t} + L_\vartheta \|\eta_u(0)\|^2_{L^2(\Omega)} \\ &\quad + \int_0^t \|\vartheta(u_h) - \vartheta(u)\|_{L^2(\Omega)} \|\partial_s \eta_u\|_{L^2(\Omega)} e^{-\bar{Q}s} \mathrm{d}s \\ &\quad + \bar{Q} \int_0^t \|\vartheta(u_h) - \vartheta(u)\|_{L^2(\Omega)} \|\eta_u\|_{L^2(\Omega)} e^{-\bar{Q}s} \mathrm{d}s.\end{aligned}$$

Proof. Using the partial integration we rewrite

$$\begin{aligned}&\int_0^t \Big(\partial_s\Big(\vartheta(u_h) - \vartheta(u)\Big), \eta_u\Big) e^{-\bar{Q}s} \mathrm{d}s \\ &= (\vartheta(u_h) - \vartheta(u), \eta_u) e^{-\bar{Q}t} - (\vartheta(u_h(0)) - \vartheta(u_0), \eta_u(0)) \\ &\quad - \int_0^t (\vartheta(u_h) - \vartheta(u), \partial_s \eta_u) e^{-\bar{Q}s} \mathrm{d}s + \bar{Q} \int_0^t (\vartheta(u_h) - \vartheta(u), \eta_u) e^{-\bar{Q}s} \mathrm{d}s.\end{aligned} \tag{3.105}$$

By virtue of the Lipschitz continuity of ϑ and the relation (3.101), we get

$$\|\vartheta(u_h(0)) - \vartheta(u_0)\|_{L^2(\Omega)}\|\eta_u(0)\|_{L^2(\Omega)} \leq L_\vartheta \|\eta_u(0)\|^2_{L^2(\Omega)}. \tag{3.106}$$

We get the desired result by applying the Cauchy-Schwarz inequality to (3.105) and (3.106).

□

Now we formulate the abstract error estimate; namely, the error estimate in terms of the $S_{h,p}$ and $\boldsymbol{S}_{h,p}$ interpolation error $(\eta_u, \boldsymbol{\eta_q}, \boldsymbol{\eta_\sigma})$.

Theorem 12 (Abstract error estimate). *Let the triangulation $\mathcal{T}_h$ satisfy conditions* (2.4) *and* (2.5)*,* $(u, \boldsymbol{q}, \boldsymbol{\sigma})$ *be the exact solution of* (3.8) *satisfying the assumptions (B1)–(B3),* $(u_h, \boldsymbol{q}_h, \boldsymbol{\sigma}_h)$ *be the approximate solution given by Definition 5, and let the assumptions (A1)–(A7) be fulfilled. Then, there exists constants $C^A_{E1} > 0$ and $\bar{Q} > 0$ independent of h and t such that*

$$\begin{aligned}
&\|\vartheta(u_h) - \vartheta(u)\|^2_{L^2(\Omega)} e^{-\bar{Q}t} + \int_0^t \left\|\boldsymbol{\xi_q}\right\|^2_{L^2(\Omega)} e^{-\bar{Q}s}\mathrm{d}s + \int_0^t J_h(\xi_u, \xi_u) e^{-\bar{Q}s}\mathrm{d}s \\
&\leq C^A_{E1}\left(t^{\frac{1-\beta}{1+\beta}} \left(\int_0^t \Big(\vartheta(u_h) - \vartheta(u), u_h - u\Big) e^{-\bar{Q}s\frac{1+\beta}{2\beta}}\mathrm{d}s \right)^{\frac{2\beta}{1+\beta}} + R_b(\eta_u, \boldsymbol{\eta_q}, \boldsymbol{\eta_\sigma}) \right),
\end{aligned} \tag{3.107}$$

where $t \in (0, T)$, $h \in (0, \bar{h})$ and

$$\begin{aligned}
&R_b(\eta_u, \boldsymbol{\eta_q}, \boldsymbol{\eta_\sigma}) \\
&\quad = \bar{R}_b(\eta_u) + \int_0^t \Big(\left\|\boldsymbol{\eta_q}\right\|^2_{L^2(\Omega)} + J_h(\eta_u, \eta_u) + R_a^2(\boldsymbol{\eta_\sigma}) + |||\eta_u|||^2 \Big) e^{-\bar{Q}s}\mathrm{d}s, \qquad (3.108a) \\
&\bar{R}_b(\eta_u) \\
&\quad = \|\eta_u(t)\|^2_{L^2(\Omega)} e^{-\bar{Q}t} + \|\eta_u(0)\|^2_{L^2(\Omega)} + \int_0^t \Big(\|\eta_u\|^2_{L^2(\Omega)} + \|\partial_s \eta_u\|^2_{L^2(\Omega)} \Big) e^{-\bar{Q}s}\mathrm{d}s. \qquad (3.108b)
\end{aligned}$$

Proof. We set the test function $v_h := \xi_u$, $\boldsymbol{w}_h := \boldsymbol{\xi_q}$, $\boldsymbol{z}_h := \boldsymbol{\xi_\sigma}$ in the error equations (3.71a)–(3.71c) and combine them. Then, we multiply the resulting equation by $e^{-\bar{Q}s}$, integrate over $(0, t)$, and use (3.76) so that

$$\begin{aligned}
&\int_0^t \Big(\partial_s(\vartheta(u_h) - \vartheta(u)), u_h - u\Big) e^{-\bar{Q}s}\mathrm{d}s \\
&+ \int_0^t B_h(u_h; \boldsymbol{\xi_q}, \boldsymbol{\xi_q}) e^{-\bar{Q}s}\mathrm{d}s + \int_0^t J_h(\xi_u, \xi_u) e^{-\bar{Q}s}\mathrm{d}s \\
&= \int_0^t \Big(\partial_s(\vartheta(u_h) - \vartheta(u)), \eta_u\Big) e^{-\bar{Q}s}\mathrm{d}s - \int_0^t J_h(\eta_u, \xi_u) e^{-\bar{Q}s}\mathrm{d}s \\
&- \int_0^t \Big(A_h(\boldsymbol{\eta_\sigma}, \xi_u) - A_h(\boldsymbol{\xi_\sigma}, \eta_u) \Big) e^{-\bar{Q}s}\mathrm{d}s \\
&- \int_0^t \Big(B_h(u; \boldsymbol{\eta_q}, \boldsymbol{\xi_q}) + \Big(B_h(u_h; \Pi_{h,p}\boldsymbol{q}, \boldsymbol{\xi_q}) - B_h(u; \Pi_{h,p}\boldsymbol{q}, \boldsymbol{\xi_q}) \Big) \Big) e^{-\bar{Q}s}\mathrm{d}s.
\end{aligned} \tag{3.109}$$

In the rest of the proof, we shall estimate the terms in (3.109). First, we consider the terms containing ϑ; namely, Lemma 10 implies

$$\begin{aligned}
&\int_0^t \Big(\partial_s(\vartheta(u_h) - \vartheta(u)), u_h - u\Big) e^{-\bar{Q}s} \mathrm{d}s \\
&\geq M_\vartheta \|\vartheta(u_h) - \vartheta(u)\|^2_{L^2(\Omega)} e^{-\bar{Q}t} - L_\vartheta \|\eta_u(0)\|^2_{L^2(\Omega)} \\
&\quad + \bar{Q} M_\vartheta \int_0^t \|\vartheta(u_h) - \vartheta(u)\|^2_{L^2(\Omega)} e^{-\bar{Q}s} \mathrm{d}s \\
&\quad - \int_0^t \int_\Omega \Big((\vartheta(u_h) - \vartheta(u))\partial_s u - \partial_s \vartheta(u)(u_h - u)\Big) e^{-\bar{Q}s} \, \mathrm{d}x \mathrm{d}s.
\end{aligned}$$

Using the same lemma, for some $\delta > 0$ it holds

$$\begin{aligned}
&\int_0^t \int_\Omega \Big| \Big((\vartheta(u_h) - \vartheta(u))\partial_s u - \partial_s \vartheta(u)(u_h - u)\Big) e^{-\bar{Q}s} \Big| \, \mathrm{d}x \mathrm{d}s \\
&\leq H_\vartheta C_X \frac{1+\beta}{2\delta^{\frac{1-\beta}{1+\beta}}} \int_0^t \int_\Omega \Big((\vartheta(u_h) - \vartheta(u))(u_h - u)\Big)^{\frac{2\beta}{1+\beta}} e^{-\bar{Q}s} \mathrm{d}s \, \mathrm{d}x \\
&\quad + H_\vartheta C_X \delta(1-\beta) \bigg(\int_0^t \|\xi_u\|^2_{L^2(\Omega)} e^{-\bar{Q}s} \mathrm{d}s + \int_0^t \|\eta_u\|^2_{L^2(\Omega)} e^{-\bar{Q}s} \mathrm{d}s \bigg).
\end{aligned}$$

By applying the Hölder inequality (3.34) to the space-time cylinder Q_T for $\gamma := \frac{2\beta}{1+\beta} \leq 1$ and using $1 - \gamma = \frac{1-\beta}{1+\beta}$ and $\Big|[0,t]\Big| = t$, we obtain

$$\begin{aligned}
&\int_\Omega \int_0^t \Big((\vartheta(u_h) - \vartheta(u))(u_h - u)\Big)^{\frac{2\beta}{1+\beta}} e^{-\bar{Q}s} \mathrm{d}s \, \mathrm{d}x \\
&\leq (t|\Omega|)^{\frac{1-\beta}{1+\beta}} \bigg(\int_0^t \Big(\vartheta(u_h) - \vartheta(u), u_h - u\Big) e^{-\bar{Q}s \frac{1+\beta}{2\beta}} \mathrm{d}s \bigg)^{\frac{2\beta}{1+\beta}}.
\end{aligned}$$

The Young inequality applied on Lemma 11 gives

$$\begin{aligned}
&\int_0^t \Big(\partial_s(\vartheta(u_h) - \vartheta(u)), \eta_u\Big) e^{-\bar{Q}s} \mathrm{d}s \\
&\leq \frac{M_\vartheta}{2} \|\vartheta(u_h) - \vartheta(u)\|^2_{L^2(\Omega)} e^{-\bar{Q}t} + \frac{1}{2M_\vartheta} \|\eta_u\|^2_{L^2(\Omega)} e^{-\bar{Q}t} + L_\vartheta \|\eta_u(0)\|^2_{L^2(\Omega)} \\
&\quad + \bar{Q} \frac{M_\vartheta}{4} \int_0^t \|\vartheta(u_h) - \vartheta(u)\|^2_{L^2(\Omega)} e^{-\bar{Q}s} \mathrm{d}s + \frac{1}{\bar{Q} M_\vartheta} \int_0^t \|\partial_s \eta_u\|^2_{L^2(\Omega)} e^{-\bar{Q}s} \mathrm{d}s \\
&\quad + \bar{Q} \frac{M_\vartheta}{4} \int_0^t \|\vartheta(u_h) - \vartheta(u)\|^2_{L^2(\Omega)} e^{-\bar{Q}s} \mathrm{d}s + \bar{Q} \frac{1}{M_\vartheta} \int_0^t \|\eta_u\|^2_{L^2(\Omega)} e^{-\bar{Q}s} \mathrm{d}s. \quad (3.110)
\end{aligned}$$

We summarize the analysis done so far; i.e., we replace the obtained relations into (3.109)

$$\begin{aligned}
&\frac{M_\vartheta}{2}\|\vartheta(u_h)-\vartheta(u)\|^2_{L^2(\Omega)}e^{-\bar{Q}t}+\bar{Q}\frac{M_\vartheta}{2}\int_0^t\|\vartheta(u_h)-\vartheta(u)\|^2_{L^2(\Omega)}e^{-\bar{Q}s}\mathrm{d}s\\
&+\int_0^t B_h(u_h;\boldsymbol{\xi_q},\boldsymbol{\xi_q})e^{-\bar{Q}s}\mathrm{d}s+\int_0^t J_h(\xi_u,\xi_u)e^{-\bar{Q}s}\mathrm{d}s\\
&\le\int_0^t\Big(\Big|B_h(u;\boldsymbol{\eta_q},\boldsymbol{\xi_q})\Big|+\Big|B_h(u_h;\Pi_{h,p}\boldsymbol{q},\boldsymbol{\xi_q})-B_h(u;\Pi_{h,p}\boldsymbol{q},\boldsymbol{\xi_q})\Big|\Big)e^{-\bar{Q}s}\mathrm{d}s\\
&\quad+\int_0^t\Big|J_h(\eta_u,\xi_u)\Big|e^{-\bar{Q}s}\mathrm{d}s+\int_0^t\Big(\Big|A_h(\boldsymbol{\eta_\sigma},\xi_u)\Big|+\Big|A_h(\boldsymbol{\xi_\sigma},\eta_u)\Big|\Big)e^{-\bar{Q}s}\mathrm{d}s\\
&\quad+H_\vartheta C_X(1-\beta)\delta\int_0^t\|\xi_u\|^2_{L^2(\Omega)}e^{-\bar{Q}s}\mathrm{d}s+\tilde{R}_b(\eta_u)\\
&\quad+C_1t^{\frac{1-\beta}{1+\beta}}\left(\int_0^t\Big(\vartheta(u_h)-\vartheta(u),u_h-u\Big)e^{-\bar{Q}s\frac{1+\beta}{2\beta}}\mathrm{d}s\right)^{\frac{2\beta}{1+\beta}},
\end{aligned}\tag{3.111}$$

where $C_1:=H_\vartheta C_X(1+\beta)|\Omega|^{\frac{1-\beta}{1+\beta}}/(2\delta^{\frac{1-\beta}{1+\beta}})$ and $\tilde{R}_b(\eta_u):=C_2\bar{R}_b(\eta_u)$ (cf. (3.108b)) with

$$C_2:=\max\left(\frac{1}{2M_\vartheta},L_\vartheta,\frac{\bar{Q}}{M_\vartheta}+H_\vartheta C_X(1-\beta)\delta,\frac{1}{\bar{Q}M_\vartheta}\right).$$

Next we estimate the terms containing the form B_h; namely, by (3.32), (3.33), and the Young inequality

$$B_h(u_h;\boldsymbol{\xi_q},\boldsymbol{\xi_q})\ge k_0\Big\|\boldsymbol{\xi_q}\Big\|^2_{L^2(\Omega)},\tag{3.112}$$

$$\Big|B_h(u;\boldsymbol{\eta_q},\boldsymbol{\xi_q})\Big|\le\frac{4k_1^2}{k_0}\Big\|\boldsymbol{\eta_q}\Big\|^2_{L^2(\Omega)}+\frac{k_0}{16}\Big\|\boldsymbol{\xi_q}\Big\|^2_{L^2(\Omega)},\tag{3.113}$$

and the relations (3.78), (3.6), (3.23), and the Young inequality give that

$$\begin{aligned}
&\Big|B_h(u_h;\Pi_{h,p}\boldsymbol{q},\boldsymbol{\xi_q})-B_h(u;\Pi_{h,p}\boldsymbol{q},\boldsymbol{\xi_q})\Big|\\
&\le\frac{8C_c^2}{k_0}\Big(\|\vartheta(u_h)-\vartheta(u)\|^2_{L^2(\Omega)}+\Big\|\boldsymbol{\eta_q}\Big\|^2_{L^2(\Omega)}\Big)+\frac{k_0}{16}\Big\|\boldsymbol{\xi_q}\Big\|^2_{L^2(\Omega)}.
\end{aligned}\tag{3.114}$$

Moreover, we bound the jump term using the relation (3.24) and the Young inequality

$$\Big|J_h(\eta_u,\xi_u)\Big|\le J_h^{1/2}(\eta_u,\eta_u)J_h^{1/2}(\xi_u,\xi_u)\le J_h(\eta_u,\eta_u)+\frac{1}{4}J_h(\xi_u,\xi_u).\tag{3.115}$$

Then we estimate the terms containing the form A_h; namely, we use the inequality (3.72) from Lemma 7 and (3.87) from Lemma 9, define $C_m:=\min(k_0,1)$ (note $C_m\le 1$ and $C_m\le k_0$), and apply the Young inequality so that

$$\begin{aligned}
|A_h(\boldsymbol{\eta_\sigma},\xi_u)|&\le R_a(\boldsymbol{\eta_\sigma})\|\xi_u\|_{L^2(\Omega)}\\
&\le C_eR_a(\boldsymbol{\eta_\sigma})\Big(\Big\|\boldsymbol{\xi_q}\Big\|_{L^2(\Omega)}+J_h^{1/2}(\xi_u,\xi_u)+|||\eta_u|||\Big)\\
&\le C_3R_a^2(\boldsymbol{\eta_\sigma})+\frac{C_m}{8}\Big\|\boldsymbol{\xi_q}\Big\|^2_{L^2(\Omega)}+\frac{C_m}{8}J_h(\xi_u,\xi_u)+|||\eta_u|||^2\\
&\le C_3R_a^2(\boldsymbol{\eta_\sigma})+\frac{k_0}{8}\Big\|\boldsymbol{\xi_q}\Big\|^2_{L^2(\Omega)}+\frac{1}{8}J_h(\xi_u,\xi_u)+|||\eta_u|||^2,
\end{aligned}\tag{3.116}$$

where $C_3 := C_e^2(4/C_m + 1/4)$. Using again the same lemmas (cf. (3.73) and (3.86)) and the Young inequality we obtain

$$
\begin{aligned}
|A_h(\boldsymbol{\xi}_{\boldsymbol{\sigma}}, \eta_u)| &\le C_a \|\boldsymbol{\xi}_{\boldsymbol{\sigma}}\|_{L^2(\Omega)} |\!|\!|\eta_u|\!|\!| \\
&\le C_d C_a \Big(\big\|\boldsymbol{\xi}_{\boldsymbol{q}}\big\|_{L^2(\Omega)} + \big\|\boldsymbol{\eta}_{\boldsymbol{q}}\big\|_{L^2(\Omega)} + \|\vartheta(u_h) - \vartheta(u)\|_{L^2(\Omega)} \Big) |\!|\!|\eta_u|\!|\!| \\
&\le \frac{k_0}{8} \Big(\big\|\boldsymbol{\xi}_{\boldsymbol{q}}\big\|^2_{L^2(\Omega)} + \big\|\boldsymbol{\eta}_{\boldsymbol{q}}\big\|^2_{L^2(\Omega)} + \|\vartheta(u_h) - \vartheta(u)\|^2_{L^2(\Omega)} \Big) + \frac{6C_d^2 C_a^2}{k_0} |\!|\!|\eta_u|\!|\!|^2.
\end{aligned}
\tag{3.117}
$$

Moreover, we choose $\delta := C_m(8H_\vartheta C_X(1-\beta)C_f^1)^{-1}$ and use (3.87), so that

$$
\begin{aligned}
H_\vartheta C_X \delta(1-\beta) \int_0^t \|\xi_u\|^2_{L^2(\Omega)} e^{-\bar{Q}s} \mathrm{d}s &\le \frac{C_m}{8} \big\|\boldsymbol{\xi}_{\boldsymbol{q}}\big\|^2_{L^2(\Omega)} + \frac{C_m}{8} J_h(\xi_u, \xi_u) + \frac{C_m}{8} |\!|\!|\eta_u|\!|\!|^2 \\
&\le \frac{k_0}{8} \big\|\boldsymbol{\xi}_{\boldsymbol{q}}\big\|^2_{L^2(\Omega)} + \frac{1}{8} J_h(\xi_u, \xi_u) + \frac{1}{8} |\!|\!|\eta_u|\!|\!|^2.
\end{aligned}
\tag{3.118}
$$

Now we update (3.111) with the relations (3.112)–(3.118) deducing

$$
\begin{aligned}
&\frac{M_\vartheta}{2} \|\vartheta(u_h) - \vartheta(u)\|^2_{L^2(\Omega)} e^{-\bar{Q}t} + \bar{Q} \frac{M_\vartheta}{2} \int_0^t \|\vartheta(u_h) - \vartheta(u)\|^2_{L^2(\Omega)} e^{-\bar{Q}s} \mathrm{d}s \\
&+ \frac{k_0}{2} \int_0^t \big\|\boldsymbol{\xi}_{\boldsymbol{q}}\big\|_{L^2(\Omega)} e^{-\bar{Q}s} \mathrm{d}s + \frac{1}{2} \int_0^t J_h(\xi_u, \xi_u) e^{-\bar{Q}s} \mathrm{d}s \\
&\le \left(\frac{8C_c^2}{M_\vartheta k_0} + \frac{k_0}{4M_\vartheta} \right) \int_0^t \|\vartheta(u_h) - \vartheta(u)\|^2_{L^2(\Omega)} e^{-\bar{Q}s} \mathrm{d}s + \hat{R}_b(\eta_u, \boldsymbol{\eta}_{\boldsymbol{q}}, \boldsymbol{\eta}_{\boldsymbol{\sigma}}) \\
&\quad + C_1 t^{\frac{1-\beta}{1+\beta}} \left(\int_0^t \Big(\vartheta(u_h) - \vartheta(u), u_h - u \Big) e^{-\bar{Q}s \frac{1+\beta}{2\beta}} \mathrm{d}s \right)^{\frac{2\beta}{1+\beta}},
\end{aligned}
$$

where $\hat{R}_b(\eta_u, \boldsymbol{\eta}_{\boldsymbol{q}}, \boldsymbol{\eta}_{\boldsymbol{\sigma}}) := C_4 R_b(\eta_u, \boldsymbol{\eta}_{\boldsymbol{q}}, \boldsymbol{\eta}_{\boldsymbol{\sigma}})$ (cf. (3.108a)) and

$$
C_4 := \max \left(1 + \frac{1}{8} + \frac{6C_d^2 C_a^2}{k_0}, \frac{4k_1^2}{k_0} + \frac{8C_d^2}{k_0} \frac{k_0}{8}, C_3 \right).
$$

Finally, we choose $\bar{Q} := 16C_c^2(M_\vartheta k_0)^{-1} + k_0(4M_\vartheta k_0)^{-1}$ so that the second term on the left-hand side cancels with the first term on the right-hand side. We finish the proof of the theorem by defining $C_{E1}^A := 2\max(1, C_1)/\min(M_\vartheta, k_0, 1)$.

□

We present the main result of this subsection, namely, we apply approximation properties to the abstract error estimate derived in Theorem 12.

Theorem 13. *Let the triangulation $\mathcal{T}_h$ satisfy conditions* (2.4) *and* (2.5)*, $(u, \boldsymbol{q}, \boldsymbol{\sigma})$ be the exact solution of* (3.8) *satisfying the assumptions (B1)–(B3), $(u_h, \boldsymbol{q}_h, \boldsymbol{\sigma}_h)$ be the approximate solution given by Definition 5, $\mu := \min(p+1, s)$, and let the assumptions (A1)–(A7) be fulfilled. Then, there exists constants $C_{E1} > 0$ and $\bar{Q} > 0$ independent of h such that*

$$
\begin{aligned}
&\|\vartheta(u_h) - \vartheta(u)\|^2_{L^2(\Omega)} e^{-\bar{Q}t} + \int_0^t \big\|\boldsymbol{\xi}_{\boldsymbol{q}}\big\|^2_{L^2(\Omega)} e^{-\bar{Q}s} \mathrm{d}s + \int_0^t J_h(\xi_u, \xi_u) e^{-\bar{Q}s} \mathrm{d}s \\
&\le C_{E1} \left(h^{2(\mu-1)} + t^{\frac{1-\beta}{1+\beta}} \left(\int_0^t \Big(\vartheta(u_h) - \vartheta(u), u_h - u \Big) e^{-\bar{Q}s \frac{1+\beta}{2\beta}} \mathrm{d}s \right)^{\frac{2\beta}{1+\beta}} \right), \ t \in (0, T).
\end{aligned}
\tag{3.119}
$$

Proof. In this proof we shall estimate the term $R_b(\eta_u, \boldsymbol{\eta_q}, \boldsymbol{\eta_\sigma})$ given by (3.108a)–(3.108b) on the right-hand side of (3.107) in Theorem 12. Using the standard approximation properties from Lemma 1

$$|\eta_u|_{H^q(\Omega)} \leq C_A h^{\mu-q} |u|_{H^\mu(\Omega)}, \quad q = 0, 1, \tag{3.120a}$$

$$\|\partial_t \eta_u\|_{L^2(\Omega)} \leq C_A h^\mu |\partial_t u|_{H^\mu(\Omega)}, \tag{3.120b}$$

for $t \in (0, T)$ and $\mu = \min(p + 1, s)$, while for vector valued functions it holds

$$\left\|\boldsymbol{\eta_q}\right\|_{L^2(\Omega)} \leq C_A h^\mu |\boldsymbol{q}|_{H^\mu(\Omega)}, \tag{3.121a}$$

$$|\boldsymbol{\eta_\sigma}|_{H^q(\Omega)} \leq C_A h^{\mu-q} |\boldsymbol{\sigma}|_{H^\mu(\Omega)}, \quad q = 0, 1. \tag{3.121b}$$

Thus, (3.120a) for $q = 0$ yields

$$\|\eta_u(t)\|^2_{L^2(\Omega)} e^{-\bar{Q}t} \leq C_A^2 h^{2\mu} |u(t)|^2_{H^\mu(\Omega)}, \quad t \in (0, T), \tag{3.122}$$

$$\|\eta_u(0)\|^2_{L^2(\Omega)} \leq C_A^2 h^{2\mu} |u(0)|^2_{H^\mu(\Omega)}. \tag{3.123}$$

We estimate the form $\bar{R}_b(\eta_u)$ given by (3.108b) by virtue of (3.122)–(3.123), (3.120a)–(3.120b) as

$$\bar{R}_b(\eta_u) \leq C_1 h^{2\mu},$$

where $C_1 := C_A^2 \Big(|u(0)|^2_{H^\mu(\Omega)} + |u(t)|^2_{H^\mu(\Omega)} + |\partial_t u|^2_{L^2(0,t;H^\mu(\Omega))} + |u|^2_{L^2(0,t;H^\mu(\Omega))} \Big)$, $t \in (0, T)$. Moreover, (3.25) and (3.120a) give

$$\begin{aligned} &\int_0^t J_h(\eta_u, \eta_u) e^{-\bar{Q}s} \mathrm{d}s \\ &\leq \int_0^t \frac{C_W C_M}{C_T} \sum_{K \in \mathcal{T}_h} \Big(3h_K^{-2} \|\eta_u\|^2_{L^2(K)} + |\eta_u|^2_{H^1(K)} \Big) e^{-\bar{Q}s} \mathrm{d}s \leq C_2 h^{2(\mu-1)}, \end{aligned} \tag{3.124}$$

where $C_2 := C_W C_M C_T^{-1} 4 C_A^2 |u|^2_{L^2(0,t;H^\mu(\Omega))}$, $t \in (0, T)$. Therefore, the previous relation (3.124) combined (3.120a) yield

$$\int_0^t |\!|\!| \eta_u |\!|\!|^2 e^{-\bar{Q}s} \mathrm{d}s \leq C_3 h^{2(\mu-1)}, \tag{3.125}$$

where $C_3 := C_A^2 (1 + 4 C_W C_M C_T^{-1}) |u|^2_{L^2(0,t;H^\mu(\Omega))}$, $t \in (0, T)$. It remains to estimate $R_a(\boldsymbol{\eta_\sigma})$ defined by (3.74) and (3.31); namely, from (3.23) and (3.121b) we have that

$$\begin{aligned} \int_0^t R_a^2(\boldsymbol{\eta_\sigma}) \mathrm{d}s &\leq \int_0^t 2C_P^2 \Big(\|\boldsymbol{\eta_\sigma}\|^2_{L^2(\Omega)} + C_r \sum_{K \in \mathcal{T}_h} \Big(h_K^2 |\boldsymbol{\eta_\sigma}|^2_{H^1(K)} + 3 \|\boldsymbol{\eta_\sigma}\|^2_{L^2(K)} \Big) \Big) \mathrm{d}s \\ &\leq C_4 h^{2(\mu-1)}, \end{aligned} \tag{3.126}$$

where $C_4 := 2C_P^2 C_A^2 (1 + 4C_r) |\boldsymbol{\sigma}|^2_{L^2(0,t;\boldsymbol{H}^\mu(\Omega))}$, $t \in (0, T)$. Finally, we combine the relations above and get the statement for

$$C_{E1} := C_{E1}^A \max(hC_1, hC_A^2 |\boldsymbol{q}|^2_{L^2(0,t;\boldsymbol{H}^\mu(\Omega))}, C_2, C_4, C_3).$$

□

Remark 9. For the limit case $\beta = 1$, i.e., when $\vartheta' \circ \vartheta^{-1}$ is Lipschitz continuous (cf. (A4)), previous analysis simplifies. Namely, the relation (3.94) from Lemma 10 becomes

$$\Big|(\vartheta(u_h) - \vartheta(u))\partial_t u - \partial_t \vartheta(u)(u_h - u)\Big| \leq H_\vartheta C_X |\vartheta(u_h) - \vartheta(u)||u_h - u|$$

which means that the term $\int_0^t \Big(\vartheta(u_h) - \vartheta(u), u_h - u\Big) e^{-\bar{Q}s\frac{1+\beta}{2\beta}} \mathrm{d}s$ in the proof of Theorem 12 is omitted. In particular, we can avoid the steps between (3.109) and (3.110) and use the Cauchy-Schwarz and Young inequalities only. The final estimate in this case is

$$\|\vartheta(u_h) - \vartheta(u)\|^2_{L^\infty(0,T;L^2(\Omega))} + \int_0^T \Big\|\boldsymbol{\xi_q}\Big\|^2_{L^2(\Omega)} \mathrm{d}t + \int_0^T J_h(\xi_u, \xi_u)\mathrm{d}t \leq C_{ES} h^{2(\mu-1)}. \tag{3.127}$$

Hence, the analysis in Subsections 3.5.5 and 3.5.6 is not necessary. Nevertheless, the constant $C_{ES} > 0$ grows exponentially with respect to the final time T due to the factor $e^{\bar{Q}T}$ (cf. Remark 13). Therefore, the estimate (3.127) corresponds to the assertion of Theorem 24 without the assumption $\mu - 1 > (1+\beta)/(3\beta - 1) = 1$.

3.5.5 Error estimates in the nonlinear form

Within this subsection we derive a bound for the nonlinear term

$$\int_0^t \Big(\vartheta(u_h) - \vartheta(u), u_h - u\Big) e^{-\bar{Q}s\frac{1+\beta}{2\beta}} \mathrm{d}s,$$

aiming to complete the bound from Theorem 12. To do so, we shall need a different test function than the one used in the previous subsections. First, we introduce the test functions and some technical results.

Let Q be a positive real number that shall be specified later, and let $\bar{t} \in [0, T]$ be an arbitrary constant. For $v_h = v_h(t) \in S_{h,p}$, and $\boldsymbol{w}_h = \boldsymbol{w}_h(t) \in \boldsymbol{S}_{h,p}$, $t \in [0, \bar{t}]$, we define $\Upsilon(\cdot; t)$ as

$$\Upsilon(v_h; t) \equiv \Upsilon(v_h; t, \bar{t}) = \int_t^{\bar{t}} v_h(s) e^{-Qs} \mathrm{d}s \in S_{h,p}, \tag{3.128a}$$

$$\Upsilon(\boldsymbol{w}_h; t) \equiv \Upsilon(\boldsymbol{w}_h; t, \bar{t}) := \int_t^{\bar{t}} \boldsymbol{w}_h(s) e^{-Qs} \mathrm{d}s \in \boldsymbol{S}_{h,p}. \tag{3.128b}$$

We list some properties of the functions defined above.

Lemma 14. *Let $v_h = v_h(t) \in S_{h,p}$, $\boldsymbol{w}_h = \boldsymbol{w}_h(t) \in \boldsymbol{S}_{h,p}$, $t \in [0, \bar{t}]$, and let Υ be* *defined by* (3.128)*. Then,*

$$\nabla\Upsilon(v_h;t) = \Upsilon(\nabla v_h;t) \quad \textit{in } K \in \mathcal{T}_h, \tag{3.129a}$$

$$\partial_t\Big(\Big|\Upsilon(\boldsymbol{w}_h;t)\Big|^2\Big) = -2\boldsymbol{w}_h \cdot \Upsilon(\boldsymbol{w}_h;t)e^{-Qt} \quad \textit{in } K \in \mathcal{T}_h, \tag{3.129b}$$

$$\Big[\Upsilon(v_h;t)\Big] = \Upsilon([v_h];t) \quad \textit{on } \Gamma \in \mathcal{F}_h, \tag{3.129c}$$

$$[v_h]\Upsilon([v_h];t) = -\frac{1}{2}\partial_t\Big(\Upsilon^2([v_h];t)\Big)e^{Qt} \quad \textit{on } \Gamma \in \mathcal{F}_h, \tag{3.129d}$$

$$\begin{aligned}[v_h]\Upsilon([v_h];t) = &-\frac{1}{2}\partial_t\Big(\Upsilon^2([v_h];t)e^{Qt}\Big) \\ &+ \frac{1}{2}Q\Big(\Upsilon([v_h];t)\Big)^2 e^{Qt} \quad \textit{on } \Gamma \in \mathcal{F}_h,\end{aligned} \tag{3.129e}$$

$$\begin{aligned}\int_0^{\bar{t}} J_h(v_h, \Upsilon(v_h;t))\mathrm{d}t = &\frac{1}{2}J_h(\Upsilon(v_h;0), \Upsilon(v_h;0)) \\ &+ \frac{1}{2}Q\int_0^{\bar{t}} J_h(\Upsilon(v_h;t), \Upsilon(v_h;t))e^{Qt}\mathrm{d}t.\end{aligned} \tag{3.129f}$$

Proof. The relations (3.129a), (3.129c), and (3.129d) follows from the properties of the approximation spaces $S_{h,p}$ and $\boldsymbol{S}_{h,p}$, and the definition of Υ. We prove (3.129b) using (3.129a) and the chain rule

$$\partial_t\Big(\Big|\Upsilon(\boldsymbol{w}_h;t)\Big|^2\Big) = \partial_t\left(\left|\int_t^{\bar{t}} \boldsymbol{w}_h e^{-Qs}\mathrm{d}s\right|^2\right) = -2\left(\int_t^{\bar{t}} \boldsymbol{w}_h e^{-Qs}\mathrm{d}s\right) \cdot \boldsymbol{w}_h e^{-Qt}.$$

Moreover, by the product rule we have

$$\begin{aligned}-\frac{1}{2}\partial_t\left(\int_t^{\bar{t}}[v_h]e^{-Qs}\mathrm{d}s\right)^2 e^{Qt} = -\frac{1}{2}\Bigg(&\partial_t\left(\left(\int_t^{\bar{t}}[v_h]e^{-Qs}\mathrm{d}s\right)^2 e^{Qt}\right) \\ &- Q\left(\int_t^{\bar{t}}[v_h]e^{-Qs}\mathrm{d}s\right)^2 e^{Qt}\Bigg),\end{aligned} \tag{3.130}$$

and therefore, using (3.129d) we deduce (3.129e).

Finally, let us note the identity

$$\int_0^{\bar{t}} \partial_t\left(\left(\int_t^{\bar{t}}[v_h]e^{-Qs}\mathrm{d}s\right)^2 e^{Qt}\right)\mathrm{d}t = -\left(\int_0^{\bar{t}}[v_h]e^{-Qt}\,\mathrm{d}t\right)^2,$$

which together with the definition of J_h, (3.129c), (3.129d), and (3.130) imply

$$\begin{aligned}\int_0^{\bar{t}} J_h(v_h, \Upsilon(v_h;t))\mathrm{d}t &= \int_0^{\bar{t}} \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \kappa[v_h]\int_t^{\bar{t}}[v_h]e^{-Qs}\mathrm{d}s\mathrm{d}S\mathrm{d}t \\ &= -\frac{1}{2}\sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \kappa\int_0^{\bar{t}} \partial_t\left(\int_t^{\bar{t}}[v_h]e^{-Qs}\mathrm{d}s\right)^2 e^{Qt}\,\mathrm{d}t\mathrm{d}S \\ &= \frac{1}{2}\sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \kappa\left(\int_0^{\bar{t}}[v_h]e^{-Qt}\,\mathrm{d}t\right)^2\mathrm{d}S \\ &\quad + \frac{Q}{2}\int_0^{\bar{t}} \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \kappa\left[\int_t^{\bar{t}} v_h e^{-Qs}\mathrm{d}s\right]^2\mathrm{d}S\mathrm{d}t.\end{aligned}$$

Applying the definition of the form J_h to the right-hand side of (3.131), we have completed the proof of the lemma.

$\square$

Furthermore, in the next lemma we establish the relation between $\Upsilon(\xi_u;\cdot)$ and $\Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}};\cdot)$, analogously as in Lemma 9.

Lemma 15. *Let* Υ *be given by* (3.128)*. Then,*

$$\begin{aligned}&\|\Upsilon(\xi_u;t)\|_{L^2(\Omega)}\\&\leq C_e\Big(\big\|\Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}};t)\big\|_{L^2(\Omega)}+J_h^{1/2}(\Upsilon(\xi_u;t),\Upsilon(\xi_u;t))+|\!|\!|\Upsilon(\eta_u;t)|\!|\!|\Big),\ t\in(0,T),\end{aligned}\tag{3.131}$$

where $C_e>0$ *is the constant from Lemma 9.*

Proof. We multiply (3.71c) by e^{-Qs} and integrate over $(t,\bar{t})$ while keeping $\boldsymbol{z}_h\in\boldsymbol{S}_{h,p}$ fixed, so that the equation (3.71c) becomes

$$(\Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}};t),\boldsymbol{z}_h)-A_h(\boldsymbol{z}_h,\Upsilon(\xi_u;t))=A_h(\boldsymbol{z}_h,\Upsilon(\eta_u;t)).$$

We set $\boldsymbol{z}_h:=\nabla\Upsilon(\xi_u;t)\in\boldsymbol{S}_{h,p}$ since $\xi_u\in S_{h,p}$ and $\nabla\Upsilon(\xi_u;t)=\Upsilon(\nabla\xi_u;t)$; cf. (3.129a). From this point, the proof is completely analogous to the proof of (3.87) in Lemma 9.

$\square$

The following lemma provides a relationship between $\Upsilon(\eta_u;\cdot)$ and η_u in the $|\!|\!|\cdot|\!|\!|$-norm.

Lemma 16. *Let* Υ *be given by* (3.128)*. Then,*

$$|\!|\!|\Upsilon(\eta_u;0)|\!|\!|^2\leq C_{\bar{t}}\int_0^{\bar{t}}|\!|\!|\eta_u|\!|\!|^2\,\mathrm{d}t,\tag{3.132}$$

$$\int_0^{\bar{t}}|\!|\!|\Upsilon(\eta_u;t)|\!|\!|^2\,\mathrm{d}t\leq\bar{t}C_{\bar{t}}\int_0^{\bar{t}}|\!|\!|\eta_u|\!|\!|^2\,\mathrm{d}t,\tag{3.133}$$

where $C_{\bar{t}}:=(1-e^{-2Q\bar{t}})/(2Q)$*.*

Proof. The definition of the $|\!|\!|\cdot|\!|\!|$-norm reads

$$|\!|\!|\Upsilon(\eta_u;0)|\!|\!|^2=|\Upsilon(\eta_u;0)|^2_{H^1(\Omega,\mathcal{T}_h)}+J_h(\Upsilon(\eta_u;0),\Upsilon(\eta_u;0));\tag{3.134}$$

thus, we consider these terms individually. From the property of the function Υ (3.129a) we have

$$\begin{aligned}|\Upsilon(\eta_u;0)|^2_{H^1(\Omega,Th)}=\|\nabla\Upsilon(\eta_u;0)\|^2_{L^2(\Omega)}&=\|\Upsilon(\nabla\eta_u;0)\|^2_{L^2(\Omega)}\\&=\int_\Omega\left|\int_0^{\bar{t}}\nabla\eta_u e^{-Qs}\mathrm{d}s\right|^2\mathrm{d}x.\end{aligned}\tag{3.135}$$

The Cauchy-Schwarz inequality applied to the integral over time implies

$$\left|\int_0^{\bar{t}}\nabla\eta_u e^{-Qs}\mathrm{d}s\right|^2\leq\left|\left(\int_0^{\bar{t}}|\nabla\eta_u|^2\mathrm{d}s\right)^{1/2}\left(\int_0^{\bar{t}}e^{-2Qs}\mathrm{d}s\right)^{1/2}\right|^2\leq C_{\bar{t}}\int_0^{\bar{t}}|\nabla\eta_u|^2\mathrm{d}s.\tag{3.136}$$

Furthermore, the property on jumps (3.129c) from Lemma 14 yields

$$J_h(\Upsilon(\eta_u;0),\Upsilon(\eta_u;0)) = \sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \kappa[\Upsilon(\eta_u;0)]^2\,\mathrm{d}S = \sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \kappa\Upsilon^2([\eta_u];0)\,\mathrm{d}S, \tag{3.137}$$

and in the same way as in (3.136) we have

$$\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \kappa\bigg(\int_0^{\bar{t}}[\eta_u]e^{-Qs}\mathrm{d}s\bigg)^2\,\mathrm{d}S \le C_{\bar{t}}\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma\kappa\int_0^{\bar{t}}[\eta_u]^2\mathrm{d}s\,\mathrm{d}S. \tag{3.138}$$

We get the relation (3.132) if we insert (3.138) into (3.135) and (3.136) into (3.137).

We show the second part similarly as the first one; namely, we consider

$$\int_0^{\bar{t}}|\!|\!|\Upsilon(\eta_u;t)|\!|\!|^2\,\mathrm{d}t = \int_0^{\bar{t}}|\Upsilon(\eta_u;t)|^2_{H^1(\Omega,Th)}\,\mathrm{d}t + \int_0^{\bar{t}}J_h(\Upsilon(\eta_u;t),\Upsilon(\eta_u;t))\,\mathrm{d}t. \tag{3.139}$$

Analogously to (3.135), we use a property of the integral of a nonnegative function (i.e., the exponential function) over a finite set, since $(t,\bar{t})\subset(0,\bar{t})$, $t\in(0,\bar{t})$, such that

$$\begin{aligned}\int_0^{\bar{t}}\int_\Omega\bigg|\int_t^{\bar{t}}\nabla\eta_u e^{-Qs}\mathrm{d}s\bigg|^2\,\mathrm{d}x\,\mathrm{d}t &\le C_{\bar{t}}\int_0^{\bar{t}}\int_\Omega\int_0^{\bar{t}}|\nabla\eta_u|^2\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\\ &\le C_{\bar{t}}\int_0^{\bar{t}}\int_0^{\bar{t}}|\eta_u|^2_{H^1(\Omega,\mathcal{T}_h)}\mathrm{d}s\,\mathrm{d}t.\end{aligned}$$

Finally, by applying the same argument to the second term on the right-hand side of (3.139) we get the desired result.

□

Next we propose some lower bounds for the nonlinear forms.

Lemma 17. *Let* Υ *be given by* (3.128). *Then,*

$$\begin{aligned}&\int_0^{\bar{t}}\Big(\partial_t(\vartheta(u_h)-\vartheta(u)),\Upsilon(\xi_u;t)\Big)\,\mathrm{d}t\\ &\ge\int_0^{\bar{t}}\Big(\Big(\vartheta(u_h)-\vartheta(u),u_h-u\Big)-\Big(\vartheta(u_h)-\vartheta(u),\eta_u\Big)\Big)e^{-Qt}\,\mathrm{d}t\\ &\quad - L_\vartheta\|u_h(0)\|_{L^2(\Omega)}\|\Upsilon(\xi_u;0)\|_{L^2(\Omega)}.\end{aligned}$$

Proof. We use the integration by parts, the Leibnitz integral rule (3.37) on the integral $\int_t^{\bar{t}}\xi_u e^{-Qs}\mathrm{d}s$, and the Cauchy-Schwarz inequality such that

$$\begin{aligned}&\int_0^{\bar{t}}\Big(\partial_t(\vartheta(u_h)-\vartheta(u)),\int_t^{\bar{t}}\xi_u e^{-Qs}\mathrm{d}s\Big)\,\mathrm{d}t\\ &=-\Big(\vartheta(u_h(0))-\vartheta(u_0),\int_0^{\bar{t}}\xi_u e^{-Qs}\mathrm{d}s\Big)-\int_0^{\bar{t}}\Big(\vartheta(u_h)-\vartheta(u),\partial_t\int_t^{\bar{t}}\xi_u e^{-Qs}\mathrm{d}s\Big)\,\mathrm{d}t\\ &=-\Big(\vartheta(u_h(0))-\vartheta(u_0),\int_0^{\bar{t}}\xi_u e^{-Qs}\mathrm{d}s\Big)+\int_0^{\bar{t}}\Big(\vartheta(u_h)-\vartheta(u),\xi_u\Big)e^{-Qt}\,\mathrm{d}t\\ &\ge-\|\vartheta(u_h(0))-\vartheta(u_0)\|_{L^2(\Omega)}\bigg\|\int_0^{\bar{t}}\xi_u e^{-Qs}\mathrm{d}s\bigg\|_{L^2(\Omega)}\\ &\quad+\int_0^{\bar{t}}\Big(\Big(\vartheta(u_h)-\vartheta(u),u_h-u\Big)-\Big(\vartheta(u_h)-\vartheta(u),\eta_u\Big)\Big)e^{-Qt}\,\mathrm{d}t.\end{aligned} \tag{3.140}$$

To prove the statement, it only remains to apply Lipschitz continuity of ϑ and (3.101) to (3.140).

□

Lemma 18. *Let* Υ *be given by* (3.128). *Then*

$$\int_0^{\bar{t}} B_h(u;\boldsymbol{\xi_q},\Upsilon(\boldsymbol{\xi_q};t))\,\mathrm{d}t \geq \frac{k_0}{2}\left\|\Upsilon(\boldsymbol{\xi_q};0)\right\|^2_{L^2(\Omega)} + Q\frac{k_0}{2}\int_0^{\bar{t}}\left\|\Upsilon(\boldsymbol{\xi_q};t)\right\|^2_{L^2(\Omega)} e^{Qt}\,\mathrm{d}t$$
$$+\frac{1}{2}\int_0^{\bar{t}}\sum_{K\in\mathcal{T}_h}\int_K \partial_t\mathbf{K}(\theta(u))\Big|\Upsilon(\boldsymbol{\xi_q};t)\Big|^2 e^{Qt}\,\mathrm{d}x\,\mathrm{d}t.$$

Proof. We deduce from the product rule and (3.129b) the following identity

$$\mathbf{K}(\theta(u))\boldsymbol{\xi_q}\cdot\Upsilon(\boldsymbol{\xi_q};t) = -\frac{1}{2}\partial_t\Big(\mathbf{K}(\theta(u))\Big|\Upsilon(\boldsymbol{\xi_q};t)\Big|^2 e^{Qt}\Big)$$
$$+\frac{1}{2}\Big(Q\mathbf{K}(\theta(u))+\partial_t\mathbf{K}(\theta(u))\Big)\Big|\Upsilon(\boldsymbol{\xi_q};t)\Big|^2 e^{Qt}.$$

Moreover, from the assumption (A1) we have that

$$\begin{aligned}
&-\frac{1}{2}\int_0^{\bar{t}}\sum_{K\in\mathcal{T}_h}\int_K \partial_t\Big(\mathbf{K}(\theta(u))\Big|\Upsilon(\boldsymbol{\xi_q};t)\Big|^2 e^{Qt}\Big)\,\mathrm{d}x\,\mathrm{d}t\\
&=-\frac{1}{2}\int_0^{\bar{t}}\partial_t\sum_{K\in\mathcal{T}_h}\int_K\left(\mathbf{K}(\theta(u))\left|\int_t^{\bar{t}}\boldsymbol{\xi_q}e^{-Qs}\mathrm{d}s\right|^2 e^{Qt}\right)\mathrm{d}x\,\mathrm{d}t\\
&=\frac{1}{2}\sum_{K\in\mathcal{T}_h}\int_K\mathbf{K}(\theta(u_0))\left|\int_0^{\bar{t}}\boldsymbol{\xi_q}e^{-Qs}\mathrm{d}s\right|^2\mathrm{d}x\\
&\geq\frac{k_0}{2}\sum_{K\in\mathcal{T}_h}\int_K\left|\int_0^{\bar{t}}\boldsymbol{\xi_q}e^{-Qs}\mathrm{d}s\right|^2\mathrm{d}x=\frac{k_0}{2}\left\|\Upsilon(\boldsymbol{\xi_q};0)\right\|^2_{L^2(\Omega)},
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{2}Q\int_0^{\bar{t}}\sum_{K\in\mathcal{T}_h}\int_K\mathbf{K}(\theta(u))\Big(\Upsilon(\boldsymbol{\xi_q};t)\Big)^2 e^{Qt}\,\mathrm{d}x\mathrm{d}t\\
&=\frac{1}{2}Q\int_0^{\bar{t}}\sum_{K\in\mathcal{T}_h}\int_K\mathbf{K}(\theta(u))\left|\int_t^{\bar{t}}\boldsymbol{\xi_q}e^{-Qs}\mathrm{d}s\right|^2\mathrm{d}xe^{Qt}\mathrm{d}t\\
&\geq Q\frac{k_0}{2}\int_0^{\bar{t}}\left\|\Upsilon(\boldsymbol{\xi_q};t)\right\|^2_{L^2(\Omega)}e^{Qt}\mathrm{d}t.
\end{aligned}$$

We combine the relations above with (3.14), which completes the proof of the lemma.

□

In what follows, we shall state the main auxiliary result; i.e., a partial estimate of the error in the nonlinear form that shall be later used to prove the final result. However, before it, we present its abstract formulation.

Lemma 19 (Abstract error estimate in the nonlinear form)**.** *Let the triangulation* $\mathcal{T}_h$ *satisfy* (2.4)–(2.5), $(u,\boldsymbol{q},\boldsymbol{\sigma})$ *be the exact classical solution of* (3.8) *satisfying (B1)–(B3),* $(u_h,\boldsymbol{q}_h,\boldsymbol{\sigma}_h)$ *be the approximate solution given by Definition 5, and*

$\mu := \min(p+1, s)$. *Let the assumptions (A1)–(A7) be fulfilled and* Υ *be given by* (3.128). *Then, for arbitrary* $\varepsilon > 0$, *there exists a constant* $C^A_{E2}(\bar{\varepsilon}) > 0$ *independent of* h *and* t *but depending on some* $\bar{\varepsilon} > 0$ *such that*

$$\begin{aligned}
&\int_0^{\bar{t}} \Big(\vartheta(u_h) - \vartheta(u), u_h - u\Big) e^{-Qt}\,\mathrm{d}t \\
&\leq C^A_{E2}(\bar{\varepsilon}) R_c(\eta_u, \boldsymbol{\eta_q}, \boldsymbol{\eta_\sigma}) + \varepsilon h^{-2} \int_0^{\bar{t}} \left\|\boldsymbol{\xi_q}\right\|^2_{L^2(\Omega)} \|\theta(u_h) - \theta(u)\|^2_{L^2(\Omega)} e^{-Qt}\,\mathrm{d}t,
\end{aligned} \tag{3.141}$$

where

$$\begin{aligned}
R_c(\eta_u, \boldsymbol{\eta_q}, \boldsymbol{\eta_\sigma}) = \|\eta_u(0)\|^2_{L^2(\Omega)} + \lVert\!\vert \Upsilon(\eta_u; 0) \vert\!\rVert^2 + \int_0^{\bar{t}} \Big(R^2_a(\boldsymbol{\eta_\sigma}) + \lVert\!\vert \Upsilon(\eta_u; t) \vert\!\rVert^2\Big)\,\mathrm{d}t \\
+ \int_0^{\bar{t}} \Big(\|\eta_u\|^2_{L^2(\Omega)} + J_h(\eta_u, \eta_u) + R^2_a(\boldsymbol{\eta_\sigma}) + \left\|\boldsymbol{\eta_q}\right\|^2_{L^2(\Omega)}\Big) e^{-Qt}\,\mathrm{d}t.
\end{aligned} \tag{3.142}$$

Proof. We insert the test function $v_h := \Upsilon(\xi_u; t)$ in the error equation (3.71a)

$$\begin{aligned}
&(\partial_t(\vartheta(u_h) - \vartheta(u)), \Upsilon(\xi_u; t)) + A_h(\boldsymbol{\xi_\sigma}, \Upsilon(\xi_u; t)) + J_h(\xi_u, \Upsilon(\xi_u; t)) \\
&\quad = -A_h(\boldsymbol{\eta_\sigma}, \Upsilon(\xi_u; t)) - J_h(\eta_u, \Upsilon(\xi_u; t)),
\end{aligned} \tag{3.143}$$

and $\boldsymbol{w}_h := \Upsilon(\boldsymbol{\xi_q}; t)$ in (3.71b), and then apply the identity (3.77) such that

$$\begin{aligned}
&B_h(u; \boldsymbol{\xi_q}, \Upsilon(\boldsymbol{\xi_q}; t)) \\
&\quad = -B_h(u; \boldsymbol{\eta_q}, \Upsilon(\boldsymbol{\xi_q}; t)) - \Big(B_h(u_h; \boldsymbol{\xi_q}, \Upsilon(\boldsymbol{\xi_q}; t)) - B_h(u; \boldsymbol{\xi_q}, \Upsilon(\boldsymbol{\xi_q}; t))\Big) \\
&\qquad - \Big(B_h(u_h; \Pi_{h,p}\boldsymbol{q}, \Upsilon(\boldsymbol{\xi_q}; t)) - B_h(u; \Pi_{h,p}\boldsymbol{q}, \Upsilon(\boldsymbol{\xi_q}; t))\Big) - (\boldsymbol{\xi_\sigma}, \Upsilon(\boldsymbol{\xi_q}; t)).
\end{aligned} \tag{3.144}$$

We multiply (3.71c) by e^{-Qs} and integrate over $(t, \bar{t})$ while keeping $\boldsymbol{z}_h$ fixed such that

$$(\Upsilon(\boldsymbol{\xi_q}; t), \boldsymbol{z}_h) - A_h(\boldsymbol{z}_h, \Upsilon(\xi_u; t)) = A_h(\boldsymbol{z}_h, \Upsilon(\eta_u; t)). \tag{3.145}$$

Then, we set $\boldsymbol{z}_h := \boldsymbol{\xi_\sigma}$ in (3.145), sum the three equations (3.143)–(3.145), and integrate the resulting equation over $(0, \bar{t})$ deducing

$$\begin{aligned}
&\int_0^{\bar{t}} \Big((\partial_t(\vartheta(u_h) - \vartheta(u)), \Upsilon(\xi_u; t)) + J_h(\xi_u, \Upsilon(\xi_u; t)) + B_h(u; \boldsymbol{\xi_q}, \Upsilon(\boldsymbol{\xi_q}; t))\Big)\,\mathrm{d}t \\
&= -\int_0^{\bar{t}} \Big(A_h(\boldsymbol{\eta_\sigma}, \Upsilon(\xi_u; t)) + A_h(\boldsymbol{\xi_\sigma}, \Upsilon(\eta_u; t))\Big)\,\mathrm{d}t \\
&\quad - \int_0^{\bar{t}} \Big(J_h(\eta_u, \Upsilon(\xi_u; t)) + B_h(u; \boldsymbol{\eta_q}, \Upsilon(\boldsymbol{\xi_q}; t))\Big)\,\mathrm{d}t \\
&\quad - \int_0^{\bar{t}} \Big(B_h(u_h; \boldsymbol{\xi_q}, \Upsilon(\boldsymbol{\xi_q}; t)) - B_h(u; \boldsymbol{\xi_q}, \Upsilon(\boldsymbol{\xi_q}; t))\Big)\,\mathrm{d}t \\
&\quad - \int_0^{\bar{t}} \Big(B_h(u_h; \Pi_{h,p}\boldsymbol{q}, \Upsilon(\boldsymbol{\xi_q}; t)) - B_h(u; \Pi_{h,p}\boldsymbol{q}, \Upsilon(\boldsymbol{\xi_q}; t))\Big)\,\mathrm{d}t.
\end{aligned} \tag{3.146}$$

In what follows, we analyze this equation; namely, by Lemma 17 bound the first term on the left-hand side of (3.146) as

$$\begin{aligned}
&\int_0^{\bar{t}} \Big(\partial_t(\vartheta(u_h)-\vartheta(u)), \Upsilon(\xi_u;t)\Big)\,\mathrm{d}t \\
&\geq \int_0^{\bar{t}} \Big(\Big(\vartheta(u_h)-\vartheta(u), u_h-u\Big) - \Big(\vartheta(u_h)-\vartheta(u), \eta_u\Big)\Big) e^{-Qt}\,\mathrm{d}t \\
&\quad - L_\vartheta \|\eta_u(0)\|_{L^2(\Omega)} \|\Upsilon(\xi_u;0)\|_{L^2(\Omega)}.
\end{aligned} \tag{3.147}$$

The last term on the right-hand side of (3.147) can be bounded using Lemma 15 and the Young inequality for $\delta := C_m,\ C_m := \min(1,k_0)$ as

$$\begin{aligned}
&L_\vartheta \|\eta_u(0)\|_{L^2(\Omega)} \|\Upsilon(\xi_u;0)\|_{L^2(\Omega)} \\
&\leq \frac{3L_\vartheta^2 C_e^2}{2C_m} \|\eta_u(0)\|^2_{L^2(\Omega)} \\
&\quad + \frac{C_m}{2}\Big(\Big\|\Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}};0)\Big\|^2_{L^2(\Omega)} + J_h(\Upsilon(\xi_u;0),\Upsilon(\xi_u;0)) + |||\Upsilon(\eta_u;0)|||^2\Big) \\
&\leq \frac{3L_\vartheta^2 C_e^2}{2C_m} \|\eta_u(0)\|^2_{L^2(\Omega)} \\
&\quad + \frac{k_0}{2}\Big\|\Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}};0)\Big\|^2_{L^2(\Omega)} + \frac{1}{2}J_h(\Upsilon(\xi_u;0),\Upsilon(\xi_u;0)) + \frac{1}{2}|||\Upsilon(\eta_u;0)|||^2.
\end{aligned} \tag{3.148}$$

Furthermore, the relation (3.129f) from Lemma 14 yields

$$\int_0^{\bar{t}} J_h(\xi_u,\Upsilon(\xi_u;t))\mathrm{d}t = \frac{1}{2}J_h(\Upsilon(\xi_u;0),\Upsilon(\xi_u;0)) + \frac{1}{2}Q\int_0^{\bar{t}} J_h(\Upsilon(\xi_u;t),\Upsilon(\xi_u;t))e^{Qt}\mathrm{d}t. \tag{3.149}$$

By Lemma 18, we estimate

$$\begin{aligned}
\int_0^{\bar{t}} B_h(u;\boldsymbol{\xi}_{\boldsymbol{q}},\Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}};t))\,\mathrm{d}t &\geq \frac{k_0}{2}\Big\|\Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}};0)\Big\|^2_{L^2(\Omega)} + Q\frac{k_0}{2}\int_0^{\bar{t}}\Big\|\Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}};t)\Big\|^2_{L^2(\Omega)} e^{Qt}\,\mathrm{d}t \\
&\quad + \frac{1}{2}\int_0^{\bar{t}}\sum_{K\in\mathcal{T}_h}\int_K \partial_t\mathbf{K}(\theta(u))\Big|\Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}};t)\Big|^2 e^{Qt}\,\mathrm{d}x\,\mathrm{d}t.
\end{aligned} \tag{3.150}$$

The last term in (3.150) can be estimated using the assumptions (A2) and (B1)

$$\frac{1}{2}\int_0^{\bar{t}}\sum_{K\in\mathcal{T}_h}\int_K \partial_t\mathbf{K}(\theta(u))\Big|\Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}};t)\Big|^2 e^{Qt}\,\mathrm{d}x\,\mathrm{d}t \leq \frac{k_d C_X}{2}\int_0^{\bar{t}}\Big\|\Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}};t)\Big\|^2_{L^2(\Omega)} e^{Qt}\,\mathrm{d}t. \tag{3.151}$$

Now we combine the relations above, while the terms $J_h(\Upsilon(\xi_u;0),\Upsilon(\xi_u;0))$ and

$\left\|\Upsilon(\boldsymbol{\xi_q};0)\right\|^2_{L^2(\Omega)}$ terms cancel,

$$\begin{aligned}
&\int_0^{\bar{t}} \Big(\vartheta(u_h)-\vartheta(u),u_h-u\Big)e^{-Qt}\,\mathrm{d}t + Q\frac{k_0}{2}\int_0^{\bar{t}} \left\|\Upsilon(\boldsymbol{\xi_q};t)\right\|^2_{L^2(\Omega)} e^{Qt}\,\mathrm{d}t\\
&+\frac{1}{2}Q\int_0^{\bar{t}} J_h(\Upsilon(\xi_u;t),\Upsilon(\xi_u;t))e^{Qt}\,\mathrm{d}t\\
&\le \int_0^{\bar{t}} |(\vartheta(u_h)-\vartheta(u),\eta_u)|e^{-Qt}\,\mathrm{d}t + \frac{k_dC_X}{2}\int_0^{\bar{t}} \left\|\Upsilon(\boldsymbol{\xi_q};t)\right\|^2_{L^2(\Omega)} e^{Qt}\\
&\quad+\int_0^{\bar{t}} \Big(\Big|A_h(\boldsymbol{\eta_\sigma},\Upsilon(\xi_u;t))\Big| + \Big|A_h(\boldsymbol{\xi_\sigma},\Upsilon(\eta_u;t))\Big| + \Big|J_h(\eta_u,\Upsilon(\xi_u;t))\Big|\Big)\,\mathrm{d}t\\
&\quad+\int_0^{\bar{t}} \Big(\Big|B_h(u;\boldsymbol{\eta_q},\Upsilon(\boldsymbol{\xi_q};t))\Big| + \Big|B_h(u_h;\boldsymbol{\xi_q},\Upsilon(\boldsymbol{\xi_q};t)) - B_h(u;\boldsymbol{\xi_q},\Upsilon(\boldsymbol{\xi_q};t))\Big|\Big)\,\mathrm{d}t\\
&\quad+\int_0^{\bar{t}} \Big|B_h(u_h;\Pi_{h,p}\boldsymbol{q},\Upsilon(\boldsymbol{\xi_q};t)) - B_h(u;\Pi_{h,p}\boldsymbol{q},\Upsilon(\boldsymbol{\xi_q};t))\Big|\,\mathrm{d}t\\
&\quad+\frac{3L_\vartheta^2C_f^2}{2C_m}\|\eta_u(0)\|^2_{L^2(\Omega)}. \qquad (3.152)
\end{aligned}$$

First, we consider the form A_h in (3.152); namely, by (3.72)

$$\Big|A_h(\boldsymbol{\eta_\sigma},\Upsilon(\xi_u;t))\Big| \le R_a(\boldsymbol{\eta_\sigma})\|\Upsilon(\xi_u;t)\|_{L^2(\Omega)}, \qquad (3.153)$$

where R_a is defined by (3.74). We further estimate (3.153) using Lemma 15 and the Young inequality as

$$\begin{aligned}
&\Big|A_h(\boldsymbol{\eta_\sigma},\Upsilon(\xi_u;t))\Big|\\
&\le \frac{2C_e^2}{QC_m}R_a^2(\boldsymbol{\eta_\sigma})e^{-Qt}\\
&\quad+Q\frac{C_m}{4}\Big(\left\|\Upsilon(\boldsymbol{\xi_q};t)\right\|^2_{L^2(\Omega)} + J_h(\Upsilon(\xi_u;t),\Upsilon(\xi_u;t))\Big)e^{Qt} + R_1(\eta_u,\boldsymbol{\eta_\sigma})\\
&\le \frac{2C_e^2}{QC_m}R_a^2(\boldsymbol{\eta_\sigma})e^{-Qt}\\
&\quad+Q\bigg(\frac{k_0}{4}\left\|\Upsilon(\boldsymbol{\xi_q};t)\right\|^2_{L^2(\Omega)} + \frac{1}{4}J_h(\Upsilon(\xi_u;t),\Upsilon(\xi_u;t))\bigg)e^{Qt} + R_1(\eta_u,\boldsymbol{\eta_\sigma}), \qquad (3.154)
\end{aligned}$$

where $R_1(\eta_u,\boldsymbol{\eta_\sigma}) := C_eR_a(\boldsymbol{\eta_\sigma})|\!|\!|\Upsilon(\eta_u;t)|\!|\!|$. Moreover, the relation (3.73) and the Young inequality for some $\bar{\varepsilon}>0$ imply

$$\Big|A_h(\boldsymbol{\xi_\sigma},\Upsilon(\eta_u;t))\Big| \le C_a\|\boldsymbol{\xi_\sigma}\|_{L^2(\Omega)}|\!|\!|\Upsilon(\eta_u;t)|\!|\!| \le \bar{\varepsilon}\|\boldsymbol{\xi_\sigma}\|^2_{L^2(\Omega)} + \frac{1}{4\bar{\varepsilon}}|\!|\!|\Upsilon(\eta_u;t)|\!|\!|^2. \qquad (3.155)$$

On the other hand, by (3.33) and the Young inequality we have

$$\Big|B_h(u;\boldsymbol{\eta_q},\Upsilon(\boldsymbol{\xi_q};t))\Big| \le \frac{1}{2}\left\|\boldsymbol{\eta_q}\right\|^2_{L^2(\Omega)}e^{-Qt} + \frac{k_1^2}{2}\left\|\Upsilon(\boldsymbol{\xi_q};t)\right\|^2_{L^2(\Omega)}e^{Qt}. \qquad (3.156)$$

Utilizing (3.78) and the Young inequality for some $\varepsilon>0$, we estimate the differences of the form B_h

$$\begin{aligned}
&\Big|B_h(u_h;\boldsymbol{\xi_q},\Upsilon(\boldsymbol{\xi_q};t)) - B_h(u;\boldsymbol{\xi_q},\Upsilon(\boldsymbol{\xi_q};t))\Big|\\
&\le \varepsilon h^{-2}\|\theta(u_h)-\theta(u)\|^2_{L^2(\Omega)}\left\|\boldsymbol{\xi_q}\right\|^2_{L^2(\Omega)}e^{-Qt} + \frac{C_b^2}{4\varepsilon}\left\|\Upsilon(\boldsymbol{\xi_q};t)\right\|^2_{L^2(\Omega)}e^{Qt}, \qquad (3.157)
\end{aligned}$$

while by using (3.79) and (3.6), we obtain

$$
\begin{aligned}
&\Big|B_h(u_h; \Pi_{h,p}\boldsymbol{q}, \Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}}; t)) - B_h(u; \Pi_{h,p}\boldsymbol{q}, \Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}}; t))\Big| \\
&\le \frac{1}{4L_\vartheta}\|\vartheta(u_h) - \vartheta(u)\|^2_{L^2(\Omega)}e^{-Qt} \\
&\quad + \frac{1}{2}\Big\|\boldsymbol{\eta}_{\boldsymbol{q}}\Big\|^2_{L^2(\Omega)}e^{-Qt} + \left(L_\vartheta C_c^2 + \frac{1}{2}\right)\Big\|\Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}}; t)\Big\|^2_{L^2(\Omega)}e^{Qt}. \qquad (3.158)
\end{aligned}
$$

The Cauchy-Schwarz and Young inequalities imply

$$
|(\vartheta(u_h) - \vartheta(u), \eta_u)| \le \frac{1}{4L_\vartheta}\|\vartheta(u_h) - \vartheta(u)\|^2_{L^2(\Omega)} + L_\vartheta\|\eta_u\|^2_{L^2(\Omega)}. \qquad (3.159)
$$

Lastly, we bound the jump term J_h using the relation(3.24) and the Young inequality as

$$
\Big|J_h(\eta_u, \Upsilon(\xi_u; t))\Big| \le \frac{1}{Q}J_h(\eta_u, \eta_u)e^{-Qt} + Q\frac{1}{4}J_h(\Upsilon(\xi_u; t), \Upsilon(\xi_u; t))e^{Qt}. \qquad (3.160)
$$

We substitute (3.154)–(3.160) into (3.152), where that the jump term on the left-hand side is canceled (cf. (3.160) and (3.152)), so that we have

$$
\begin{aligned}
&\int_0^{\bar{t}} \Big(\vartheta(u_h) - \vartheta(u), u_h - u\Big)e^{-Qt}\,\mathrm{d}t + Q\frac{k_0}{4}\int_0^{\bar{t}} \Big\|\Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}}; t)\Big\|^2_{L^2(\Omega)}e^{Qt}\,\mathrm{d}t \\
&\le C_1\int_0^{\bar{t}} \Big\|\Upsilon(\boldsymbol{\xi}_{\boldsymbol{q}}; t)\Big\|^2_{L^2(\Omega)}e^{Qt} + \varepsilon h^{-2}\int_0^{\bar{t}} \Big\|\boldsymbol{\xi}_{\boldsymbol{q}}\Big\|^2_{L^2(\Omega)}\|\theta(u_h) - \theta(u)\|^2_{L^2(\Omega)}e^{-Qt}\,\mathrm{d}t \\
&\quad + \frac{1}{2L_\vartheta}\int_0^{\bar{t}} \|\vartheta(u_h) - \vartheta(u)\|^2_{L^2(\Omega)}e^{-Qt} + \bar{\varepsilon}\int_0^{\bar{t}} \|\boldsymbol{\xi}_{\boldsymbol{\sigma}}\|^2_{L^2(\Omega)}\,\mathrm{d}t\,\mathrm{d}t \\
&\quad + C_2R_c(\eta_u, \boldsymbol{\eta}_{\boldsymbol{q}}, \boldsymbol{\eta}_{\boldsymbol{\sigma}}), \qquad (3.161)
\end{aligned}
$$

where R_c is given by (3.142) and

$$
\begin{aligned}
C_1 = C_1(\varepsilon) &:= \frac{k_dC_X}{2} + \frac{k_1^2}{2} + \frac{C_b^2}{4\varepsilon} + L_\vartheta C_c^2 + \frac{1}{2}, \\
C_2 = C_2(\bar{\varepsilon}) &:= \max\left(\frac{3L_\vartheta^2C_e^2}{2C_m}, \frac{1}{2}, \frac{2C_e^2}{QC_m}, \frac{C_e}{2}, \frac{1}{4\bar{\varepsilon}} + \frac{3}{2}, L_\vartheta, \frac{1}{Q}\right).
\end{aligned}
$$

Finally, by virtue of properties of ϑ, (its nonnegativity and monotonicity), we observe the identity $|\vartheta(u_h) - \vartheta(u)||u_h - u| = (\vartheta(u_h) - \vartheta(u))(u_h - u)$, and thus using the Lipschitz continuity of ϑ we have

$$
\|\vartheta(u_h) - \vartheta(u)\|^2_{L^2(\Omega)} \le L_\vartheta\Big(\vartheta(u_h) - \vartheta(u), u_h - u\Big).
$$

We choose sufficiently small $\bar{\varepsilon}$, and set $Q := 4C_1/k_0$ and $C_{E2}^A = C_{E2}^A(\bar{\varepsilon}) := 2\max(1, C_2(\bar{\varepsilon}))$ to prove the required statement.

□

Remark 10. We emphasize that we have been allowed to hide $\int_0^{\bar{t}} \|\boldsymbol{\xi}_{\boldsymbol{\sigma}}\|^2_{L^2(\Omega)}\,\mathrm{d}t$ in the line (3.161) due to the stability result presented in Theorem 6 and the assumption (B3) since $\int_0^{\bar{t}} \|\boldsymbol{\xi}_{\boldsymbol{\sigma}}\|^2_{L^2(\Omega)}\,\mathrm{d}t \le 2\int_0^{\bar{t}} \|\boldsymbol{\sigma}_h\|^2_{L^2(\Omega)}\,\mathrm{d}t + 2\int_0^{\bar{t}} \|\boldsymbol{\sigma}\|^2_{L^2(\Omega)}\,\mathrm{d}t$. However, the obtained bound (cf. (3.141)) still contains a term that is not estimated (e.g., $h^{-2} \to \infty$ when $h \to 0$; thus, we cannot apply the same procedure). This issue will be addressed in the following subsection.

Remark 11. Note that the constant in (3.141) $C_{E2}^{A}(\bar{\varepsilon}) \to \infty$ when $\bar{\varepsilon} \to 0$. This is a consequence of the application of the Young inequality and the extended mixed formulation. Nevertheless, their usage was essential to overcome analysis of nonlinearities.

In the next lemma, we interpret the previous abstract error estimate in terms of the dependency on the mesh size h.

Lemma 20 (Error estimate in the nonlinear form)**.** *Let the triangulation $\mathcal{T}_h$ satisfy* (2.4)*–*(2.5)*, $(u, \boldsymbol{q}, \boldsymbol{\sigma})$ be the exact solution of* (3.8) *satisfying (B1)–(B3), $(u_h, \boldsymbol{q}_h, \boldsymbol{\sigma}_h)$ be the approximate solution given by Definition 5, $\mu := \min(p+1, s)$, and let the assumptions (A1)–(A7) be fulfilled. Then, for arbitrary $\varepsilon > 0$ there exists a constant $C_{E2}(\bar{\varepsilon}) > 0$ independent of h but depending on some $\bar{\varepsilon} > 0$ such that*

$$\begin{aligned}
&\int_0^{\bar{t}} \Big(\vartheta(u_h) - \vartheta(u), u_h - u\Big) e^{-Qt}\,\mathrm{d}t \\
&\le C_{E2}(\bar{\varepsilon}) h^{2(\mu-1)} + \varepsilon h^{-2} \int_0^{\bar{t}} \left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|_{L^2(\Omega)}^2 \|\theta(u_h) - \theta(u)\|_{L^2(\Omega)}^2 e^{-Qt}\,\mathrm{d}t.
\end{aligned}$$

Proof. We proceed similarly as in the proof of Theorem 13; i.e., we bound the term R_c (3.142) in (3.141) in Lemma 19. From Lemma 16 ,(3.132), and (3.125) we get

$$\left\|\left|\Upsilon(\eta_u; 0)\right|\right\|^2 \le C_{\bar{t}} \int_0^{\bar{t}} \left\|\left|\eta_u\right|\right\|^2 \mathrm{d}t \le C_1 h^{2(\mu-1)}, \tag{3.162}$$

where $C_1 := C_{\bar{t}} C_A^2 (1 + 4C_W C_M C_T^{-1}) |u|_{L^2(0,\bar{t};H^\mu(\Omega))}^2$. Using the second part of the Lemma 16, (3.133), and again (3.125) we get

$$\int_0^{\bar{t}} \left\|\left|\Upsilon(\eta_u; t)\right|\right\|^2 \mathrm{d}t \le \bar{t} C_{\bar{t}} \int_0^{\bar{t}} \left\|\left|\eta_u\right|\right\|^2 \mathrm{d}t \le C_2 h^{2(\mu-1)}, \tag{3.163}$$

where $C_2 := \bar{t} C_1$. Using these inequalities and ones already shown in the proof of Theorem 13, in the following fashion (3.123), (3.162), (3.126), (3.163), (3.120a), (3.124), and (3.121a), we obtain the statement for

$$\begin{aligned}
C_{E2}(\bar{\varepsilon}) := C_{E2}^{A}(\bar{\varepsilon}) \max\Bigg(& C_A^2 h |u(0)|_{H^\mu(\Omega)}^2, C_1, 2C_P^2 C_A^2 (1 + 4C_r) |\boldsymbol{\sigma}|_{L^2(0,\bar{t};\boldsymbol{H}^\mu(\Omega))}^2, \\
& C_2, C_A^2 h |u|_{L^2(0,\bar{t};H^\mu(\Omega))}, \\
& C_W C_M C_T^{-1} 4 C_A^2 |u|_{L^2(0,\bar{t};H^\mu(\Omega))}^2, C_A^2 h |\boldsymbol{q}|_{L^2(0,\bar{t};H^\mu(\Omega)}\Bigg).
\end{aligned}$$

□

3.5.6 Main result

In this subsection, we combine the results from previous subsections to derive a priori error bounds. Since our numerical scheme is continuous in time, it comes natural to use the *continuous induction* originally developed in [20]. In fact, we may say that the classical mathematical induction is a discrete variant of the continuous one. First, we formulate a lemma that we named the ‘induction step’ since it will serve as the one to prove the main theorem.

Lemma 21 (Induction step)**.** *Let $\mathcal{T}_h$ be a triangulation such that conditions* (2.4) *and* (2.5) *are fulfilled,* $(u, \boldsymbol{q}, \boldsymbol{\sigma})$ *be the exact solution of* (3.8) *satisfying (B1)–(B3),* $(u_h, \boldsymbol{q}_h, \boldsymbol{\sigma}_h)$ *be its approximate solution given by Definition 5,* $\mu := \min(p+1, s)$, *and let the assumptions (A1)–(A7) be valid. Furthermore, let* $t \in [0, T]$ *be the largest time for which*

$$\int_0^t \left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)} \mathrm{d}s \le h^{\frac{4\beta}{3\beta-1}}, \quad h \in (0, \bar{h}], \tag{3.164}$$

for some $0 < \bar{h} \le 1$. *Then, there exists a constant* $C_{EI} > 0$ *independent of* h *such that*

$$\|\vartheta(u_h) - \vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))} + \int_0^t \left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)} \mathrm{d}s + \int_0^t J_h(\xi_u, \xi_u)\mathrm{d}s \le C_{EI} h^{2(\mu-1)\frac{2\beta}{1+\beta}}. \tag{3.165}$$

Remark 12. We point out that in Lemma 21 $t \ne 0$ due to Theorem 6. Furthermore, for $\beta \to 1/3^+$, the exponent on the right-hand side of (3.164) blows up making the right-hand side vanish. Thus, we emphasize that the presented result is valid for β strictly larger than 1/3 and cannot be used in a limit.

Proof. The main idea of the proof relies on a careful application of Lemma 20 to Theorem 13. First, we rewrite these statements; namely, let us assume that there exists some $\bar{t} \in (0, t]$ at which the essential supremum of $\|\theta(u_h) - \theta(u)\|_{L^2(\Omega)}$ is attained. Note that this statement is valid due to the assumption (A3) on θ. Moreover, by virtue of the monotonicity of ϑ (cf. Remark 5) and properties of the exponential function ($e^{\bar{Q}(\bar{t}-s)} \ge 1,\ s \le \bar{t}$), from (3.119) (for $t = \bar{t}$), we have that

$$\begin{aligned}
&\|\vartheta(u_h) - \vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))} + \int_0^{\bar{t}} \left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)} \mathrm{d}s + \int_0^{\bar{t}} J_h(\xi_u, \xi_u)\mathrm{d}s \\
&\le e^{\bar{Q}\bar{t}} \left(C_{E1} h^{2(\mu-1)} + \bar{t}^{\frac{1-\beta}{1+\beta}} \left(\int_0^{\bar{t}} \Big(\vartheta(u_h) - \vartheta(u), u_h - u\Big) e^{-\bar{Q}s\frac{1+\beta}{2\beta}} \mathrm{d}s \right)^{\frac{2\beta}{1+\beta}} \right),
\end{aligned} \tag{3.166}$$

Conversely, using similar arguments, from Lemma 20 we deduce

$$\begin{aligned}
&\int_0^{\bar{t}} \Big(\vartheta(u_h) - \vartheta(u), u_h - u\Big) e^{-\bar{Q}s\frac{1+\beta}{2\beta}} \mathrm{d}s \\
&\le C_{E2}(\bar{\varepsilon}) h^{2(\mu-1)} + \varepsilon h^{-2} \int_0^{\bar{t}} \left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)} \|\theta(u_h) - \theta(u)\|^2_{L^2(\Omega)} e^{-\bar{Q}s\frac{1+\beta}{2\beta}} \mathrm{d}s \\
&\le C_{E2}(\bar{\varepsilon}) h^{2(\mu-1)} + \varepsilon h^{-2} \int_0^{\bar{t}} \left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)} \mathrm{d}s \|\vartheta(u_h) - \vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))},
\end{aligned} \tag{3.167}$$

The p-triangle inequality (3.36) for $p := \frac{2\beta}{1+\beta} \le 1$ applied on constant functions $f_1 := C_{E2}(\bar{\varepsilon}) h^{2(\mu-1)}$ and $f_2 := \varepsilon h^{-2} \int_0^{\bar{t}} \left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)} \mathrm{d}s \|\vartheta(u_h) - \vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))}$ gives

$$\begin{aligned}
&\left(C_{E2}(\bar{\varepsilon}) h^{2(\mu-1)} + \varepsilon h^{-2} \int_0^{\bar{t}} \left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)} \mathrm{d}s \|\vartheta(u_h) - \vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))} \right)^{\frac{2\beta}{1+\beta}} \\
&\le C_1(\bar{\varepsilon}) h^{2(\mu-1)\frac{2\beta}{1+\beta}} \\
&\quad + 2^{\frac{\beta-1}{2\beta}} \varepsilon^{\frac{2\beta}{1+\beta}} \left(h^{-2} \int_0^{\bar{t}} \left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)} \mathrm{d}s \|\vartheta(u_h) - \vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))} \right)^{\frac{2\beta}{1+\beta}},
\end{aligned} \tag{3.168}$$

where $C_1(\bar\varepsilon) := 2^{\frac{\beta-1}{2\beta}}\Big(C_{E2}(\bar\varepsilon)\Big)^{\frac{2\beta}{1+\beta}}$. We use the inequality [2, p. 14]

$$|abc|^\omega \leq |b| + (|a|^{\frac{\omega}{2\omega-1}}|b|)^{\frac{2\omega-1}{\omega}}|c|, \quad \frac{1}{2} < \omega \leq 1, \tag{3.169}$$

with settings $a = h^{-2}$, $b = \int_0^{\bar t} \left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)} \mathrm{d}s$, $c = \|\vartheta(u_h) - \vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))}$ and $\omega = 2\beta/(1+\beta)$ (cf. (A4)), (implying $(2\omega-1)/\omega = (3\beta-1)/2\beta$), and the assumption (3.164) of this lemma, so that

$$\begin{aligned}
&\left(h^{-2}\int_0^{\bar t}\left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)}\mathrm{d}s\|\vartheta(u_h)-\vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))}\right)^{\frac{2\beta}{1+\beta}}\\
&\leq \int_0^{\bar t}\left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)}\mathrm{d}s + \left(h^{-\frac{4\beta}{3\beta-1}}\int_0^{\bar t}\left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)}\mathrm{d}s\right)^{\frac{3\beta-1}{2\beta}}\|\vartheta(u_h)-\vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))}\\
&\leq \int_0^{\bar t}\left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)}\mathrm{d}s + \|\vartheta(u_h)-\vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))}.
\end{aligned}\tag{3.170}$$

Now we combine Lemma 20 with Theorem 13 (cf. (3.166)) and apply the relations above (3.167), (3.168), and (3.170). Additionally, bearing in mind that $2(\mu-1)2\beta/(1+\beta) < 2(\mu-1)$ for $\beta \leq 1$ and $\bar h < 1$, we obtain

$$\begin{aligned}
&\|\vartheta(u_h)-\vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))} + \int_0^{\bar t}\left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)}\mathrm{d}s + \int_0^{\bar t} J_h(\xi_u,\xi_u)\mathrm{d}s\\
&\leq \bar C_E(\bar t)h^{2(\mu-1)\frac{2\beta}{1+\beta}}\\
&\quad + \varepsilon^{\frac{2\beta}{1+\beta}}\tilde C_E(\bar t)\left(\int_0^{\bar t}\left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)}\mathrm{d}s + \|\vartheta(u_h)-\vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))}\right),
\end{aligned}\tag{3.171}$$

where

$$\bar C_E(\bar t) := e^{\bar Q\bar t}\Big(C_{E1} + C_1(\bar\varepsilon)\bar t^{\frac{1-\beta}{1+\beta}}\Big) \quad \text{and} \quad \tilde C_E(\bar t) := 2^{\frac{\beta-1}{2\beta}}e^{\bar Q\bar t}\bar t^{\frac{1-\beta}{1+\beta}}. \tag{3.172}$$

The last term on the right-hand side of (3.171) is bounded due to $\bar t \leq t$, (3.164) and Remark 8, so we may choose ε sufficiently small such that

$$\begin{aligned}
&\frac{1}{2}\|\vartheta(u_h)-\vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))} + \frac{1}{2}\int_0^{\bar t}\left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)}\mathrm{d}s + \int_0^{\bar t} J_h(\xi_u,\xi_u)\mathrm{d}s\\
&\leq \bar C_E(\bar t)h^{2(\mu-1)\frac{2\beta}{1+\beta}};
\end{aligned}\tag{3.173}$$

more precisely, we have chosen $\varepsilon^{\frac{2\beta}{1+\beta}}\tilde C_E(\bar t) = 1/2$, i.e., $\varepsilon := \exp\frac{-\ln 2\tilde C_E(\bar t)(1+\beta)}{2\beta}$.

Now we extend the statement (3.173) to the interval $(0,t)$. Thus, let $t \in [0,T]$. We couple again (3.119) with Lemma 20 using similar arguments as before

$$\begin{aligned}
&\|\vartheta(u_h)-\vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))} + \int_0^{t}\left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)}\mathrm{d}s + \int_0^{t} J_h(\xi_u,\xi_u)\mathrm{d}s\\
&\leq \bar C_E(t)h^{2(\mu-1)\frac{2\beta}{1+\beta}}\\
&\quad + \tilde C_E(t)\varepsilon^{\frac{2\beta}{1+\beta}}\left(h^{-2}\int_0^{t}\left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|^2_{L^2(\Omega)}\mathrm{d}s\|\vartheta(u_h)-\vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))}\right)^{\frac{2\beta}{1+\beta}}.
\end{aligned}$$

Similarly, we use the inequality (3.169) for the same parameters, only differing in $b = \int_0^t \left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|_{L^2(\Omega)}^2 \mathrm{d}s$, such that

$$\begin{aligned}
&\|\vartheta(u_h) - \vartheta(u)\|_{L^\infty(0,t;L^2(\Omega))}^2 + \int_0^t \left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|_{L^2(\Omega)}^2 \mathrm{d}s + \int_0^t J_h(\xi_u, \xi_u)\mathrm{d}s \\
&\le \bar{C}_E(t) h^{2(\mu-1)\frac{2\beta}{1+\beta}} \\
&\quad + \tilde{C}_E(t)\varepsilon^{\frac{2\beta}{1+\beta}} \left(\int_0^t \left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|_{L^2(\Omega)}^2 \mathrm{d}s + \|\vartheta(u_h) - \vartheta(u)\|_{L^\infty(0,t;L^2(\Omega))}^2 \right).
\end{aligned}$$

Ultimately, we hide the last two terms similarly as in (3.171), but now for $\varepsilon := \exp \frac{-\ln 2\tilde{C}_E(t)(1+\beta)}{2\beta}$. The final result follows if we set $C_{EI} := 2\bar{C}_E(T)$, where T is the final time.

□

Remark 13. Let us note that the constant C_{EI} from the previous lemma grows exponentially with respect to the final time T.

The next step is to remove the a priori assumption (3.164) from Lemma 21. To do so, in the next lemma we define the continuous induction.

Lemma 22 (Continuous mathematical induction [54, Lemma 7.2])**.** *Let $P(t)$, $t \in [0, T]$ be a propositional function such that*

(i) Induction basis: *$P(0)$ is true;*
(ii) Induction step: *$\exists \delta_0 > 0 \quad \forall \delta \in [0, \delta_0] : t + \delta \le T \quad P(t) \implies P(t+\delta)$, $\forall t \in [0, T]$.*

Then, $P(t)$ holds for all $t \in [0, T]$.

In fact, a propositional function (or, in logic, a predicate) $P : [0, T] \to$ [True,False] represents a statement that takes either a value of True or a value of False on the given domain of definition. In the following theorem, we shall define a propositional function and prove that it satisfies the conditions (i) and (ii) of Lemma 22, meaning that the statement is valid on the interval $[0, T]$.

Theorem 23. *Let the triangulation $\mathcal{T}_h$ satisfy conditions* (2.4) *and* (2.5)*, $(u, \boldsymbol{q}, \boldsymbol{\sigma})$ be the exact solution of* (3.8) *that satisfies the regularity conditions (B1)–(B3), $(u, \boldsymbol{q}, \boldsymbol{\sigma})$ be the approximate solution given by Definition 5, and $\mu := \min(p+1, s)$. Let the assumptions (A1)–(A7) be fulfilled, $\mu - 1 > (1+\beta)/(3\beta-1)$, and $1 \ge \bar{h} > 0$ be such that $C_{EI}\bar{h}^{2(\mu-1)2\beta/(1+\beta)} = \bar{h}^{4\beta/(3\beta-1)}/2$, where C_{EI} is the constant from Lemma 21. Then, for all $h \in (0, \bar{h}]$, the following estimate is satisfied*

$$\|\vartheta(u_h) - \vartheta(u)\|_{L^\infty(0,T;L^2(\Omega))}^2 + \int_0^T \left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|_{L^2(\Omega)}^2 \mathrm{d}t + \int_0^T J_h(\xi_u, \xi_u)\,\mathrm{d}t \le C_E h^{2(\mu-1)\frac{2\beta}{1+\beta}}, \tag{3.174}$$

where $C_E > 0$ is a constant independent of h.

Proof. We define the following propositional function

$$\begin{aligned}
P(t) = \Bigg\{ &\|\vartheta(u_h) - \vartheta(u)\|_{L^\infty(0,t;L^2(\Omega))}^2 + \int_0^t \left\|\boldsymbol{\xi}_{\boldsymbol{q}}\right\|_{L^2(\Omega)}^2 \mathrm{d}s \\
&+ \int_0^t J_h(\xi_u, \xi_u)\mathrm{d}s \le C_E h^{2(\mu-1)\frac{2\beta}{1+\beta}} \Bigg\}.
\end{aligned}$$

Induction basis. The statement $P(0)$ is true because of the Lipschitz continuity of ϑ, (3.101), and (3.123),

$$\|\vartheta(u_h(0)) - \vartheta(u_0)\|^2_{L^2(\Omega)} \leq L^2_\vartheta \|\eta_u(0)\|^2_{L^2(\Omega)} \leq L^2_\vartheta C^2_A |u(0)|^2_{H^\mu(\Omega)} h^{2\mu} \\ \leq C_E h^{2(\mu-1)\frac{2\beta}{1+\beta}},$$

where $C_E := L^2_\vartheta C^2_A |u(0)|^2_{H^\mu(\Omega)}$. Here we also have used the relation $2(\mu-1)2\beta/(1+\beta) < 2\mu$, which holds for $\beta \leq 1$.

Induction step. Now let the inductive hypothesis $P(t)$ be satisfied for some $t \in (0, T]$. By the assumption of the theorem, for a fixed $h \in (0, \bar{h}]$ we have that $C_E h^{2(\mu-1)2\beta/(1+\beta)} \leq h^{4\beta/(3\beta-1)}/2$, thus,

$$\|\vartheta(u_h) - \vartheta(u)\|^2_{L^\infty(0,t;L^2(\Omega))} + \int_0^t \left\|\boldsymbol{\xi_q}\right\|^2_{L^2(\Omega)} \mathrm{d}s + \int_0^t J_h(\xi_u, \xi_u) \mathrm{d}s \\ \leq C_E h^{2(\mu-1)\frac{2\beta}{1+\beta}} \leq \frac{1}{2} h^{\frac{4\beta}{3\beta-1}}. \tag{3.175}$$

Since the interval $[0, T]$ is a compact set, we have that $\xi : [0, T] \to L^2(\Omega)$ is a uniformly continuous mapping, and therefore the function $\phi(t) := \int_0^t \left\|\boldsymbol{\xi_q}(s)\right\|^2_{L^2(\Omega)} \mathrm{d}s$ is uniformly continuous on $[0, T]$; namely, it holds

$$\forall \varepsilon > 0 \quad \exists \delta_0 > 0 \quad \forall t_1, t_2 \in [0, T] \quad |t_1 - t_2| \leq \delta_0 \implies |\phi(t_1) - \phi(t_2)| \leq \varepsilon.$$

For $\varepsilon = \frac{1}{2} h^{\frac{4\beta}{3\beta-1}}$ there exists $\delta_0 > 0$ such that if $t \in (0, T)$ and $\delta \in [0, \delta_0]$, then

$$\int_t^{t+\delta} \left\|\boldsymbol{\xi_q}(s)\right\|^2_{L^2(\Omega)} \mathrm{d}s = \left| \int_0^{t+\delta} \left\|\boldsymbol{\xi_q}(s)\right\|^2_{L^2(\Omega)} \mathrm{d}s - \int_0^t \left\|\boldsymbol{\xi_q}(s)\right\|^2_{L^2(\Omega)} \mathrm{d}s \right| \leq \frac{1}{2} h^{\frac{4\beta}{3\beta-1}}. \tag{3.176}$$

Then, for $\delta \in [0, \delta_0]$ from the validity of $P(t)$, i.e., (3.175), and (3.176), we have

$$\int_0^{t+\delta} \left\|\boldsymbol{\xi_q}\right\|^2_{L^2(\Omega)} \mathrm{d}s = \int_0^t \left\|\boldsymbol{\xi_q}\right\|^2_{L^2(\Omega)} \mathrm{d}s + \int_t^{t+\delta} \left\|\boldsymbol{\xi_q}\right\|^2_{L^2(\Omega)} \mathrm{d}s \leq \frac{1}{2} h^{\frac{4\beta}{3\beta-1}} + \frac{1}{2} h^{\frac{4\beta}{3\beta-1}} = h^{\frac{4\beta}{3\beta-1}},$$

which is the assumption of Lemma 21. Consequently, we get the relation

$$\|\vartheta(u_h) - \vartheta(u)\|^2_{L^\infty(0,t+\delta;L^2(\Omega))} + \int_0^{t+\delta} \left\|\boldsymbol{\xi_q}\right\|^2_{L^2(\Omega)} \mathrm{d}s + \int_0^{t+\delta} J_h(\xi_u, \xi_u) \mathrm{d}s \\ \leq C_{EI} h^{2(\mu-1)\frac{2\beta}{1+\beta}}.$$

In this way we have shown the induction step defined in Lemma 22 (ii); if we set $C_E := \max(L^2_\vartheta C^2_A |u(0)|^2_{H^\mu(\Omega)}, C_{EI})$ the proof is done.

□

We reformulate the theorem in terms of e_u and $e_{\boldsymbol{q}}$.

Theorem 24. *Let the triangulation $\mathcal{T}_h$ satisfy conditions* (2.4) *and* (2.5)*,* $(u, \boldsymbol{q}, \boldsymbol{\sigma})$ *be the exact solution of* (3.8) *satisfying (B1)–(B3),* $(u_h, \boldsymbol{q}_h, \boldsymbol{\sigma}_h)$ *be the approximate solution given by Definition 5, and* $\mu := \min(p + 1, s)$*. Let the assumptions (A1)–(A7) be fulfilled,* $\mu - 1 > (1 + \beta)/(3\beta - 1)$*, and* $1 \geq \bar{h} > 0$ *be such that* $C_{EI} \bar{h}^{2(\mu-1)2\beta/(1+\beta)} = \bar{h}^{4\beta/(3\beta-1)}/2$*, where* C_{EI} *is the constant from Lemma 21. Then, for all* $h \in (0, \bar{h}]$ *the following estimate is satisfied*

$$\|\vartheta(u_h) - \vartheta(u)\|^2_{L^\infty(0,T;L^2(\Omega))} + \int_0^T \|e_{\boldsymbol{q}}\|^2_{L^2(\Omega)} \,\mathrm{d}t + \int_0^T J_h(e_u, e_u) \,\mathrm{d}t \leq C_E h^{2(\mu-1)\frac{2\beta}{1+\beta}}. \tag{3.177}$$

Proof. We add $\int_0^T \Big(\big\| \boldsymbol{\eta_q} \big\|^2_{L^2(\Omega)} + J_h(\eta_u, \eta_u) + 2J_h(\xi_u, \eta_u) \Big) \,\mathrm{d}t$ to both sides of (3.174), and use (3.23) and (3.124).

□

Corollary 25. *Let the conditions from Theorem 24 be satisfied. Then, there exists $\tilde{C}_{E2}(\bar{\varepsilon})$ such that*

$$\int_0^T (\vartheta(u_h) - \vartheta(u), u_h - u) \,\mathrm{d}t \le \tilde{C}_{E2}(\bar{\varepsilon}) h^{2(\mu-1)}.$$

Proof. As in the proof of the induction step (3.167), by Lemma 20 for $t = T$ we have that

$$\begin{aligned}
&\int_0^T \Big(\vartheta(u_h) - \vartheta(u), u_h - u \Big) \,\mathrm{d}t \\
&\le C_{E2}(\bar{\varepsilon}) h^{2(\mu-1)} + \varepsilon h^{-2} \int_0^T \big\| \boldsymbol{\xi_q} \big\|^2_{L^2(\Omega)} \,\mathrm{d}t \|\vartheta(u_h) - \vartheta(u)\|^2_{L^\infty(0,T;L^2(\Omega))}.
\end{aligned}$$

We apply the result from Theorem 23 as follows

$$\int_0^T \big\| \boldsymbol{\xi_q} \big\|^2_{L^2(\Omega)} \,\mathrm{d}t \|\vartheta(u_h) - \vartheta(u)\|^2_{L^\infty(0,T;L^2(\Omega))} \le C_E^2 h^{2(\mu-1)\frac{4\beta}{1+\beta}}.$$

It remains to combine the relations above and use $h^{-2} \le h^{2(\mu-1)(1-3\beta)/(1+\beta)}$, so that the statement implies for $C_{E2}(\bar{\varepsilon}) := C_{E2}(\bar{\varepsilon}) + \varepsilon C_E^2$.

□

In the next chapter, we study the estimate (3.177) and the convergence of the DG method on numerical examples.

4. Numerical experiments

In this chapter, we accompany the theory presented in Chapter 3 with several numerical experiments, which can be found in [24, 82]. The first example is the well-known Barenblatt problem [5], which for particular parameters fulfills the problem assumptions (cf. Subsection 3.1.1). The second example is the Tracy problem with the exact analytical solution, which represents Richards' equation with the Gardner constitutive relations introduced in Section 1.2.

4.1 Barenblatt problem

In (3.1) we set $\vartheta(u) = \theta(u) = u^{1/m}$, $m > 0$ and $\mathbf{K}$ to be the identity tensor so that we consider

$$\partial_t u^{1/m} - \Delta u = 0 \quad \text{in } Q_T, \tag{4.1}$$

with the analytical solution (cf. [5])

$$u(\boldsymbol{x}, t) = \frac{1}{t+1}\left[1 - \frac{m-1}{4m^2}\frac{|\boldsymbol{x}|^2}{(t+1)^{1/m}}\right]_+^{\frac{m}{m-1}}, \tag{4.2}$$

where $[a]_+ = \max(a, 0),\ a \in \mathbb{R}$ and $|\boldsymbol{x}|^2$ denotes the magnitude of $\boldsymbol{x} \in \mathbb{R}^2$.

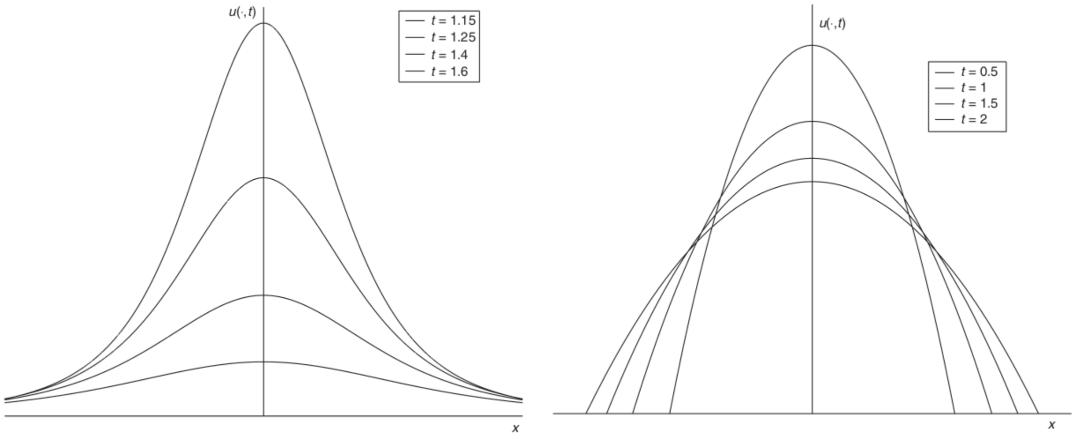

Figure 4.1: Profiles of the Barenblatt solution for $m < 1$ (left) and $m > 1$ (right) [93].

Remark 14. If we define the transformation $v := u^{1/m}$ then the equation (4.1) becomes

$$\partial_t v - \Delta(v^m) = 0,$$

which is known as a *fast-diffusion equation* for $m < 1$ or a *porous media equation* for $m > 1$. Moreover, solutions to this equation are known as the Barenblatt solutions [5] and are given by

$$v(\boldsymbol{x}, t) = \left(\frac{1}{t+1}\left[1 - \frac{m-1}{4m^2}\frac{|\boldsymbol{x}|^2}{(t+1)^{1/m}}\right]_+^{\frac{m}{m-1}}\right)^{\frac{1}{m}}.$$

In Fig. 4.1 are illustrated profiles of the Barenblatt solution in case of the fast-diffusion equation and the porous media equation.

4.1.1 Barenblatt problem for $m < 1$

Let us consider (4.1) for $m \in (0,1)$, then $\vartheta(u) = u^{1/m}$, $1/m > 1$ fulfills the assumption (A4). Moreover, from the inequality (3.23) we have that the function

$$(\vartheta' \circ \vartheta^{-1})(u) = \vartheta'(\vartheta^{-1}(u)) = \frac{1}{m}u^{1-m}$$

is Hölder continuous with the exponent $\beta := 1 - m \leq 1$. In order to fulfill the assumption (A4) ($1/3 < \beta$), we shall consider $m \in (0, 2/3)$, i.e., $1/m \in (3/2, \infty)$.

The symmetric interior penalty Galerkin method [33] earlier introduced in Section 2.4 is used for the spatial discretization of the problem (4.1). We refer to [4] for its equivalency with the LDG method used for the analysis. Furthermore, for temporal discretization we used the time discontinuous Galerkin discretization (see Chapter 5), namely, the piecewise quadratic approximation which provide sufficiently accurate approximation of the temporal variable. This space-time discretization method will be formally defined later in Chapter 5 (cf. Definition 6).

We consider the problem (4.1) in the computational domain $\Omega = (-6,6) \times (-6,6)$ and prescribe Dirichlet boundary condition on the boundary $\partial\Omega$. The domain Ω is discretized using uniform grids having 288, 1152, 4608 and 18432 elements with the corresponding mesh steps h = 1.4142, 0.7071, 0.3536 and 0.1768, respectively. The final time is set to be $T = 1$. In Tables 4.1–4.2 are presented the results for different values of m (see also Fig. 4.2–4.3 and Table 4.3 summarizing Tables 4.1–4.2); namely, we show the errors $\vartheta(u_h) - \vartheta(u)$ in the $L^\infty(0,T;L^2(\Omega))$-norm, $e_{\boldsymbol{q}} := \nabla e_u$ in the $L^2(0,T;L^2(\Omega))$-norm, and e_u in $\|\cdot\|_J := (\int_0^T J_h(\cdot,\cdot)\,\mathrm{d}t)^{1/2}$, denoted by $\|e_u\|_\vartheta$, $\|e_{\boldsymbol{q}}\|_{L^2}$, and $\|e_u\|_J$, respectively. These terms correspond to the ones appearing on the left-hand side of the estimate (3.177) from Theorem 24. In addition, we include the corresponding experimental order of convergence (EOC) calculated as $\log(e_h/e_{h'})/\log(h/h')$ for each consecutive pair of meshes with mesh steps h and h'.

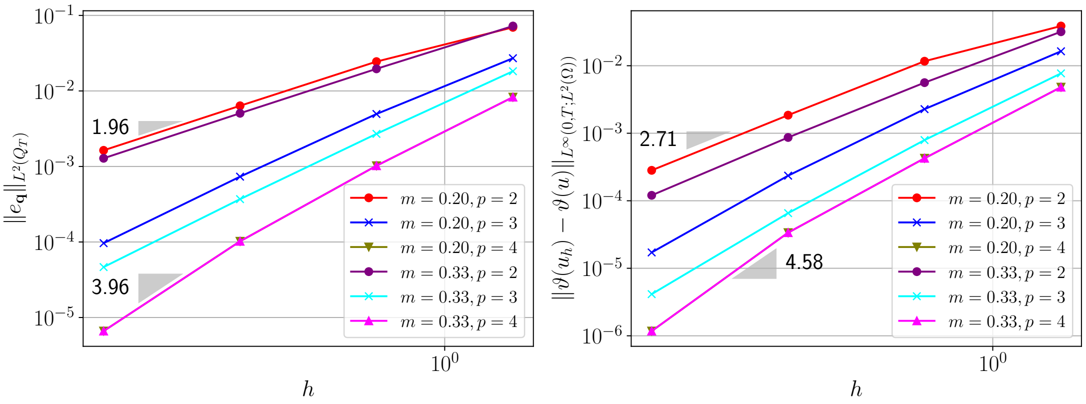

Figure 4.2: Barenblatt problem: errors *versus* h, for $m = 0.2$ and $m = 0.33$.

Table 4.1: Barenblatt problem: errors and EOC, for $m = 0.2$ (left) and $m = 0.33$ (right).

$p = 2$				$p = 2$			
h	$\|e_u\|_\vartheta$	$\|e_{\mathbf{q}}\|_{L^2}$	$\|e_u\|_J$	h	$\|e_u\|_\vartheta$	$\|e_{\mathbf{q}}\|_{L^2}$	$\|e_u\|_J$
1.4142	3.874E-02	6.953E-02	7.671E-03	1.4142	3.188E-02	7.238E-02	8.477E-03
0.7071	1.166E-02	2.446E-02	2.842E-03	0.7071	5.617E-03	1.962E-02	2.489E-03
	1.73	1.51	1.43		2.50	1.88	1.77
0.3536	1.852E-03	6.361E-03	8.202E-04	0.3536	8.660E-04	5.061E-03	6.821E-04
	2.65	1.94	1.79		2.70	1.95	1.87
0.1768	2.823E-04	1.635E-03	2.219E-04	0.1768	1.204E-04	1.283E-03	1.766E-04
	2.71	1.96	1.89		2.85	1.98	1.95

$p = 3$				$p = 3$			
h	$\|e_u\|_\vartheta$	$\|e_{\mathbf{q}}\|_{L^2}$	$\|e_u\|_J$	h	$\|e_u\|_{L^2}$	$\|e_{\mathbf{q}}\|_{L^2}$	$\|e_u\|_J$
1.4142	1.636E-02	2.710E-02	2.785E-03	1.4142	7.706E-03	1.823E-02	2.154E-03
0.7071	2.276E-03	4.951E-03	6.097E-04	0.7071	7.922E-04	2.702E-03	3.755E-04
	2.85	2.45	2.19		3.28	2.75	2.52
0.3536	2.354E-04	7.330E-04	9.896E-05	0.3536	6.561E-05	3.692E-04	4.587E-05
	3.27	2.76	2.62		3.59	2.87	3.03
0.1768	1.717E-05	9.649E-05	1.164E-05	0.1768	4.141E-06	4.654E-05	5.328E-06
	3.78	2.93	3.09		3.99	2.99	3.11

$p = 4$				$p = 4$			
h	$\|e_u\|_\vartheta$	$\|e_{\mathbf{q}}\|_{L^2}$	$\|e_u\|_J$	h	$\|e_u\|_{L^2}$	$\|e_{\mathbf{q}}\|_{L^2}$	$\|e_u\|_J$
1.4142	4.814E-03	8.276E-03	1.040E-03	1.4142	1.444E-03	3.898E-03	5.769E-04
0.7071	4.261E-04	1.022E-03	1.445E-04	0.7071	1.326E-04	4.524E-04	5.745E-05
	3.50	3.02	2.85		3.45	3.11	3.33
0.3536	3.370E-05	1.021E-04	1.316E-05	0.3536	5.215E-06	3.006E-05	4.334E-06
	3.66	3.32	3.46		4.67	3.91	3.73
0.1768	1.175E-06	6.614E-06	9.816E-07	0.1768	2.187E-07	1.925E-06	3.034E-07
	4.84	3.95	3.74		4.58	3.97	3.84

First, we study the cases $m = 0.2$ and $m = 0.33$. The results in Table 4.1 show higher EOC than one proposed by theory; namely, the values EOC are close to p which is the rate of the convergence for regular problems. We suppose that the superiority of the numerical experiments is not caused by the suboptimality of the theoretical estimates but due to the fact that the lower regularity of the solution appears only locally in this case. We remark that for $m = 0.2, 0.33$, the assumption $\mu - 1 > (1 + \beta)/(3\beta - 1)$ from Theorem 24 holds for $p \geq 2$.

Table 4.2: Barenblatt problem: errors and EOC, for $m = 0.6$.

$p = 3$				$p = 4$			
h	$\|e_u\|_\vartheta$	$\|e_\mathbf{q}\|_{L^2}$	$\|e_u\|_J$	h	$\|e_u\|_\vartheta$	$\|e_\mathbf{q}\|_{L^2}$	$\|e_u\|_J$
1.4142	3.749E-03	1.469E-02	1.220E-03	1.4142	6.825E-04	2.849E-03	1.093E-04
0.7071	3.096E-04	2.011E-03	1.362E-04	0.7071	2.817E-05	1.955E-04	7.525E-06
	3.60	2.87	3.16		4.60	3.86	3.86
0.3536	2.016E-05	2.531E-04	1.595E-05	0.3536	9.811E-07	1.256E-05	5.089E-07
	3.94	2.99	3.09		4.84	3.96	3.89
0.1768	1.243E-06	3.167E-05	1.953E-06	0.1768	1.941E-07	7.929E-07	3.389E-08
	4.02	3.00	3.03		2.34	3.99	3.91

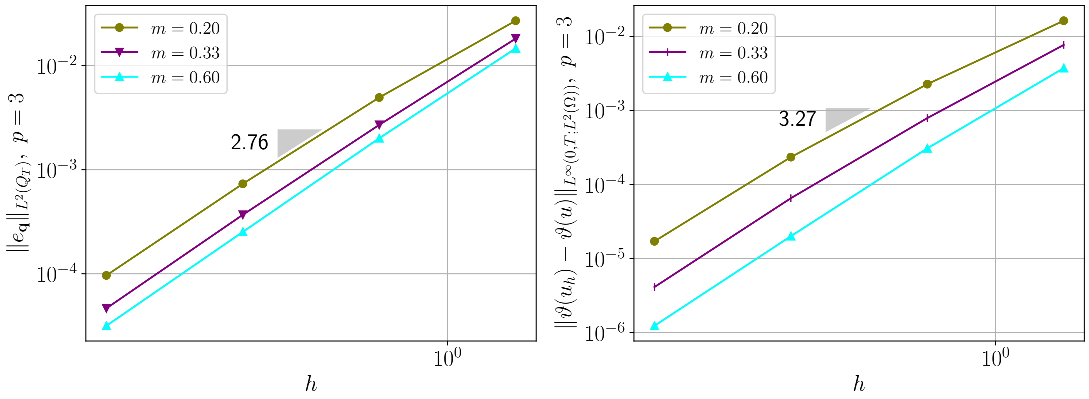

Figure 4.3: Barenblatt problem: errors *versus* h, for $m = 0.2$, $m = 0.33$ and $m = 0.6$.

Table 4.3: Barenblatt problem: expected rates according to Theorem 24 and EOC for cases $m = 0.2, 0.33, 0.6$.

m	$p = 2$			$p = 3$			$p = 4$		
	$p\frac{2\beta}{1+\beta}$	$\|e_\boldsymbol{q}\|_{L^2}$	$\|e_u\|_\vartheta$	$p\frac{2\beta}{1+\beta}$	$\|e_\boldsymbol{q}\|_{L^2}$	$\|e_u\|_\vartheta$	$p\frac{2\beta}{1+\beta}$	$\|e_\boldsymbol{q}\|_{L^2}$	$\|e_u\|_\vartheta$
0.2	1.78	1.80	2.36	2.67	2.71	3.30	3.56	3.43	4.00
0.33	1.60	1.94	2.68	2.40	2.87	3.62	3.20	3.66	4.23
0.6	–	–	–	–	2.95	3.85	–	3.94	3.93

In addition, we consider the case $m = 0.6$ where the assumption from Theorem 23 is fulfilled starting from $p \geq 7$. Nevertheless, we have included the results for $p = 3$ and $p = 4$ (cf. Tables 4.2–4.3 and Fig. 4.3) showing a higher EOC similarly to the cases satisfying the assumption of the theorem.

4.1.2 Barenblatt problem for $m > 1$

In Table 4.4 and Figs. 4.4–4.5 we demonstrate some experiments for $m > 1$ where the regularity of the Barenblatt solution is lower and ϑ is only Hölder continuous, which does not meet our assumptions (cf. (A4)). Moreover, no evidence on convergence of the DG method has been found in the literature.

Here, we note a significant decrease of the EOC in comparison to the previous case. Namely, for $m = 1.5$ we observe that the convergence is limited by the rate 2.5 starting from the polynomial approximation $p = 3$ (cf. Fig. 4.4). Moreover, for $m = 2$ we notice the rate around $(m + 1)/m$ as obtained in [77] for mixed finite element method. In the last two columns of Table 4.4, we indicated our observations on rates of convergence for several values of $m > 1$, where EOC gradually drops (see Fig. 4.5).

Table 4.4: Barenblatt problem: EOC for cases $m = 1.5, 2, 3, 4, 6$ with proposed rates of convergence.

m	$p=1$		$p=2$		$p=3$		$\|e_{\boldsymbol{q}}\|_{L^2}$ rate	$\|e_u\|_\vartheta$ rate
	$\|e_{\boldsymbol{q}}\|_{L^2}$	$\|e_u\|_\vartheta$	$\|e_{\boldsymbol{q}}\|_{L^2}$	$\|e_u\|_\vartheta$	$\|e_{\boldsymbol{q}}\|_{L^2}$	$\|e_u\|_\vartheta$		
1.5	1.00	1.97	2.01	2.43	2.52	2.46	$\min(p, m+1)$	$\min(p+1, m+1)$
2	1.02	1.47	1.50	1.44	1.50	1.45	$\min(p, \frac{m+1}{m})$	$\min(p+1, \frac{m+1}{m})$
3	0.91	1.00	1.00	0.99	0.99	0.97	$\min(p, 1)$	$\min(p+1, 1)$
4	0.81	0.78	0.84	0.88	0.83	0.86	$\min(p, \frac{m}{m+1})$	$\min(p+1, \frac{m}{m+1})$
6	0.70	0.68	0.69	0.72	0.70	0.74	$\min(p, 0.70)$	$\min(p+1, 0.70)$

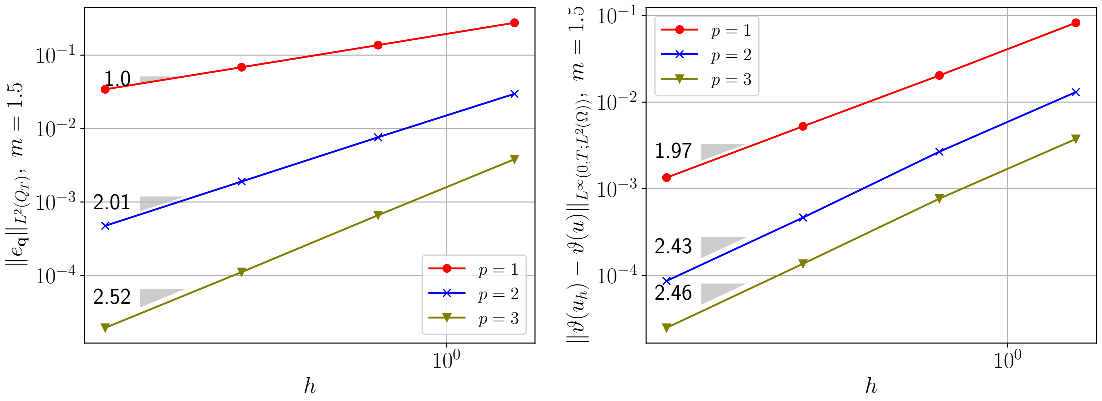

Figure 4.4: Barenblatt problem: errors *versus* h, for $m = 1.5$.

4.2 Tracy problem

We consider the Richards' equation (3.1) with $g = 0$ and with the Gardner constitutive relations [44] where θ and $\mathbf{K}$ are defined by (1.12) and (1.13), respectively, with parameters given in Table 4.5. We remark that it is not clear if this problem fulfills the assumptions (A1)–(A7); however we study it since it is a more practical example than the previous ones.

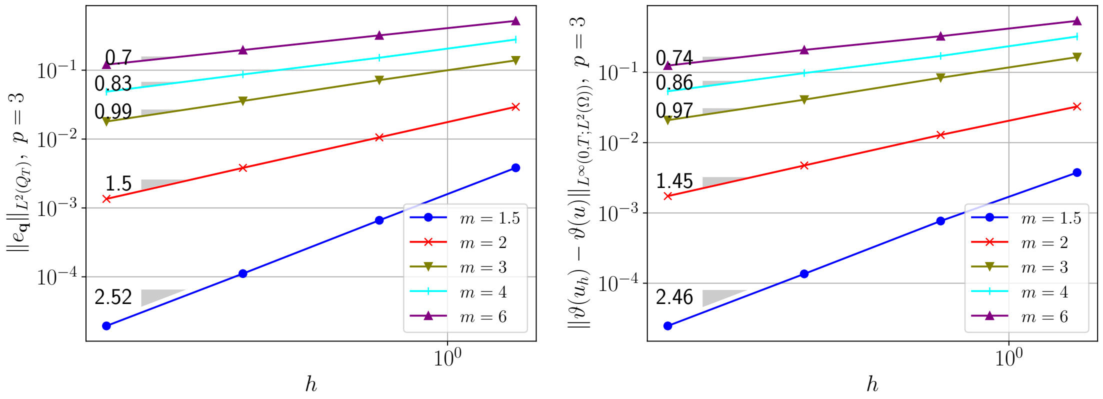

Figure 4.5: Barenblatt problem: errors *versus* h, for $m > 1$ and $p = 3$.

Table 4.5: Parameters for the Gardner model.

A	K_S	θ_S	θ_R	S_S
0.1	1.1	0.5	0	0

Let Ω be a rectangular domain $(0,1)\times(0,2)$. We prescribe the initial condition $u_0 = -10$ and the same value for the boundary conditions with an exception for the edge $(0,1)\times\{2\}$ denoted by Γ_3 in Fig. 4.6 where we set

$$u = \log(\exp(Au_0) + (1 - \exp(Au_0))\sin(\pi/A)).$$

Since there is an inconsistency between the initial and boundary conditions we set the final time to be sufficiently small $T = 10^{-4}$. We refer to [91, 87] for the exact solution to the Tracy problem obtained by the Fourier method.

The computations are carried out on a sequence of uniform triangular mesh with 100, 400, 1600 and 6400 elements with the mesh parameters $h = 0.2828$, 0.1414, 0.0707 and 0.0353, respectively. Exceptionally, for $p = 3$ the finest mesh is omitted. The time variable is approximated by use of cubic polynomials and the time-step is selected using adaptation. In Table 4.6 and Fig. 4.7, we present errors in the $L^\infty(0,T;L^2(\Omega))$-norm and the $L^2(0,T;L^2(\Omega))$-norm. We emphasize that the error in the $L^\infty(0,T;L^2(\Omega))$-norm is calculated as the maximum over the time integration nodes, which may not be sufficiently accurate in this example with inconsistent initial and boundary conditions. Thus, the errors calculated in the $L^2(0,T;L^2(\Omega))$-norm are more reliable for the numerical study. We observe a lower order of convergence comparing with sufficiently regular problems.

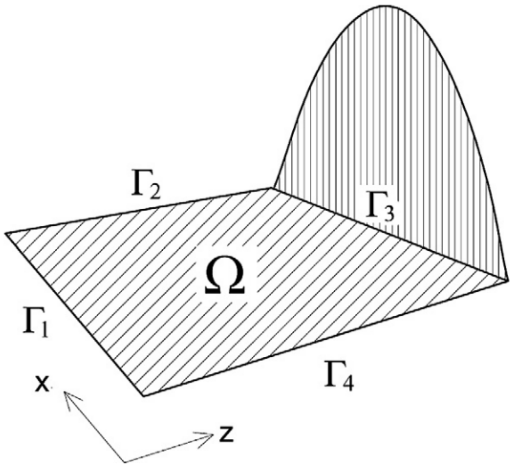

Figure 4.6: Tracy problem: geometry of the domain [87].

Table 4.6: Tracy problem: errors and EOC.

h	p	$\|e_u\|_{L^2(L^2)}$	EOC	$\|e_u\|_{L^\infty(L^2)}$	EOC
0.2828	1	1.0535E-02	–	6.2586E-03	–
0.1414	1	4.2024E-03	1.32	3.4608E-03	0.85
0.0707	1	1.5476E-03	1.44	1.4368E-03	1.26
0.0353	1	5.3082E-04	1.54	5.0843E-04	1.49
0.2828	2	3.4108E-03	–	2.8445E-03	–
0.1414	2	1.2480E-03	1.45	1.1796E-03	1.27
0.0707	2	4.0830E-04	1.61	3.9637E-04	1.57
0.0353	2	8.9429E-05	2.20	8.4325E-05	2.23
0.2828	3	1.5227E-03	–	1.3411E-03	–
0.1414	3	5.2839E-04	1.53	4.9602E-04	1.43
0.0707	3	1.3576E-04	1.96	1.3262E-04	1.90

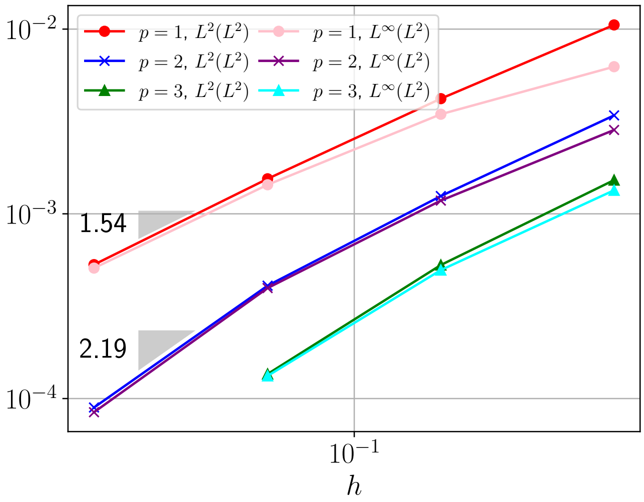

Figure 4.7: Tracy problem, errors *versus* h.

5. Temporal discretization

Within this chapter, we derive fully discrete numerical schemes for the problems introduced in Chapter 2. Namely, we discretize the temporal variable using the DG method, obtaining the space-time discontinuous Galerkin (STDG) method. To do so, we first define the suitable space-time partition and functional spaces. Moreover, we define the space of polynomial functions that have a varying polynomial degree on different elements, by virtue of which we define the higher-order STDG.

5.1 Space-time partition and function spaces

For $r > 1,\ r \in \mathbb{N}$ we define a partition

$$0 = t_0 < t_1 < \cdots < t_r = T$$

and divide the time interval $[0, T]$ into subintervals

$$I_m = (t_{m-1}, t_m).$$

Additionally, we denote the closure of a subintermal I_m as

$$\bar{I}_m = [t_{m-1}, t_m],$$

and we introduce temporal parameters

$$\tau_m = t_m - t_{m-1},\ \tau = \max_{m=1,\dots,r} \tau_m.$$

Therefore, the partition of the time interval is given by

$$[0, T] = \cup_{m=1}^{r} \bar{I}_m,$$

where $I_m \cap I_n = \emptyset$ for $m \neq n$, $m, n = 1, \dots, r$.

For a function v defined in $[0, T]$, we introduce the notation of the jump with respect to the time variable as

$$\{v\}_m = v_m^+ - v_m^-,$$

where

$$v_m^{\pm} = v(t_m^{\pm}) = \lim_{t \to t_m^{\pm}} v(t).$$

For each time instant t_m, $m = 0, 1, \dots, r$ and interval I_m, $m = 1, \dots, r$, we define a triangulation $\mathcal{T}_{h,m}$ of the domain Ω as described in Section 2.2. We may notice that the triangulations $\mathcal{T}_{h,m}$ may differ for different time levels; a one-dimensional example is depicted in Fig. 5.1.

In the sequel, we update the notation from Section 2.2 by adding a subscript m as we have denoted the time dependent grids $\mathcal{T}_{h,m}$ (previously $\mathcal{T}_h$). Namely, we denote by $\mathcal{F}_{h,m}$ the system of all faces of all elements $K \in \mathcal{T}_{h,m}$; in particular,

$$\mathcal{F}_{h,m} = \mathcal{F}_{h,m}^{I} \cup \mathcal{F}_{h,m}^{B}, \quad \mathcal{F}_{h,m}^{B} = \mathcal{F}_{h,m}^{D} \cup \mathcal{F}_{h,m}^{N} \quad \text{and} \quad \mathcal{F}_{h,m}^{ID} = \mathcal{F}_{h,m}^{I} \cup \mathcal{F}_{h,m}^{D},$$

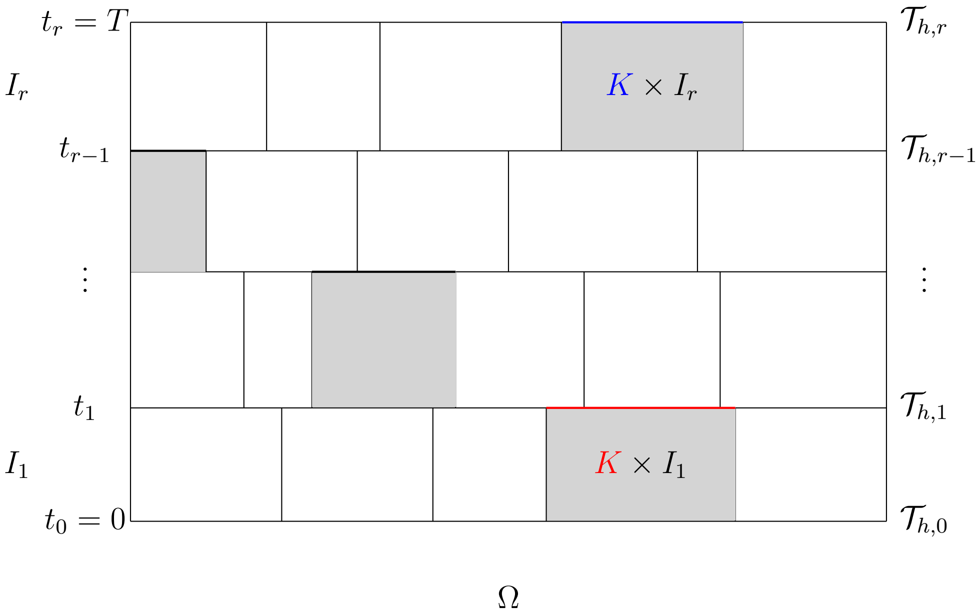

Figure 5.1: An illustration of the space-time partition in 1D.

where $\mathcal{F}^I_{h,m}$ and $\mathcal{F}^B_{h,m}$ are the inner and boundary edges, respectively. Moreover, $\mathcal{F}^D_{h,m}$ and $\mathcal{F}^N_{h,m}$ are edges on the boundary $\partial\Omega_D$ and $\partial\Omega_N$, respectively. Furthermore, we define the spatial parameters

$$h = \max_{m=1,\dots,r} h_m, \quad h_m = \max_{K\in\mathcal{T}_{h,m}} h_K, \quad h_K = \mathrm{diam}(K) \text{ for } K \in \mathcal{T}_{h,m}.$$

Now, we may proceed with defining the functional space over the space-time partitions. Over a triangulation $\mathcal{T}_{h,m}$, for each $k \in \mathbb{N}$, we define the broken Sobolev space of scalar functions

$$H^k(\Omega, \mathcal{T}_{h,m}) = \{v \in L^2(\Omega) : v|_K \in H^k(K) \ \forall K \in \mathcal{T}_{h,m}\},$$

equipped with the seminorm

$$|v|_{H^k(\Omega,\mathcal{T}_{h,m})} = \left(\sum_{K\in\mathcal{T}_{h,m}} |v|^2_{H^k(K)} \right)^{1/2}.$$

We keep the notation of jump and average value (cf. (2.3)) of $v \in H^k(\Omega, \mathcal{T}_{h,m})$

Analogously to (2.7), for $p \geq 1$ we define the finite-dimensional space at each time level t_m

$$S_{h,p,m} = \{v \in L^2(\Omega) : \ v|_K \in P_p(K) \ \forall K \in \mathcal{T}_{h,m}\}.$$

For each element $K \in \mathcal{T}_{h,m}$ we denote by $\pi_{K,p,m}$ the L^2-projection of some $v \in L^2(K)$ to the space $P_p(K)$,

$$\int_K (\pi_{K,p,m} v - v)\varphi \, \mathrm{d}x = 0 \quad \forall \varphi \in P_p(K). \tag{5.1}$$

Thus, for $v \in L^2(\Omega)$ we have that

$$(\Pi_{h,p,m} v)|_K := \pi_{K,p}(v|_K) \ \ \forall K \in \mathcal{T}_{h,m}.$$

We may define the space of space-time dependent discontinious polynomials on $\Omega \times (0,T)$

$$S_{h,p}^{\tau,q} = \{v \in L^2(Q_T) : v(x,t)|_{I_m} = \sum_{i=0}^{q} t^i v_{m,i}(x),$$
$$v_{m,i} \in S_{h,p,m},\ i = 0,\dots,q,\ m = 1,\dots,r\},$$

and the space of piecewise polynomial functions on a time layer

$$S_{h,p,m}^{\tau,q} = \{v \in L^2(\Omega \times I_m) : v(x,t) = \sum_{i=0}^{q} t^i v_{m,i}(x),$$
$$v_{m,i} \in S_{h,p,m},\ i = 0,\dots,q\},\ m = 1,\dots,r.$$

Moreover, to define an adaptive algorithm that allows polynomial approximation degrees to vary on different elements, we shall need to modify the polynomial spaces above; namely, let us denote the set of local polynomial degrees as

$$\boldsymbol{p} = \{p_K, K \in \mathcal{T}_{h,m}\}.$$

Thus, we define the space of space dependent polynomials

$$S_{h,\boldsymbol{p},m} = \{v \in L^2(\Omega) :\ v|_K \in P_{p_K}(K)\ \forall K \in \mathcal{T}_{h,m}\},$$

and the space of piecewise polynomial functions on $\Omega \times (0,T)$

$$S_{h,\boldsymbol{p}}^{\tau,q} = \{v \in L^2(Q_T) : v(x,t)|_{I_m} = \sum_{i=0}^{q} t^i v_{m,i}(x),$$
$$v_{m,i} \in S_{h,\boldsymbol{p},m},\ i = 0,\dots,q,\ m = 1,\dots,r\}.$$

Additionally, we define the space of piecewise polynomial functions on a time layer

$$S_{h,\boldsymbol{p},m}^{\tau,q} = \{v \in L^2(\Omega \times I_m) : v(x,t) = \sum_{i=0}^{q} t^i v_{m,i}(x),$$
$$v_{m,i} \in S_{h,\boldsymbol{p},m},\ i = 0,\dots,q\},\ m = 1,\dots,r. \tag{5.2}$$

Clearly, it holds

$$v|_{\Omega\times I_m} \in S_{h,\boldsymbol{p},m}^{\tau,q}\ \forall v \in S_{h,\boldsymbol{p}}^{\tau,q},\ m = 1,\dots,r.$$

5.2 STDG discretization

First, we shall derive the fully discrete scheme of the Ψ-formulation of Richards' equation (2.1), i.e., we complete the discretization of the scheme given by Definition 2. Prior to it, we need to interpret the forms given by (2.17)–(2.19), (2.15) in terms of the triangulations $\mathcal{T}_{h,m}$, namely,

$$a_{h,m}(\psi;\Psi,v) = \tilde{a}_{h,m}(\psi;\Psi,v) + J_{h,m}(\Psi,v), \tag{5.3}$$

$$\begin{aligned}\tilde{a}_{h,m}(\psi;\Psi,v) = &\sum_{K\in\mathcal{T}_{h,m}} \int_K \mathbf{K}(\psi)\nabla\Psi\cdot\nabla v\,\mathrm{d}x \\ &- \sum_{\Gamma\in\mathcal{F}_{h,m}^{ID}} \int_\Gamma \Big(\langle \mathbf{K}(\psi)\nabla\Psi\rangle \cdot \boldsymbol{n}[v] \\ &+ \Theta\,\langle \mathbf{K}(v)\nabla v\rangle \cdot \boldsymbol{n}[\Psi]\Big)\mathrm{d}S,\end{aligned} \tag{5.4}$$

$$J_{h,m}(\Psi,v) = \sum_{\Gamma\in\mathcal{F}_{h,m}^{ID}} \int_\Gamma \kappa[\Psi][v]\mathrm{d}S \tag{5.5}$$

$$\ell_{h,m}(v) = (g,v) + (g_N,v)_N - \Theta \sum_{\Gamma\in\mathcal{F}_{h,m}^{D}} \int_\Gamma \boldsymbol{n}\cdot\nabla v\Psi_D\mathrm{d}S, \tag{5.6}$$

where κ is defined by (2.16), and for $\Theta = -1$, $\Theta = 0$ and $\Theta = 1$, the form $a_{h,m}$ is the nonsymmetric (NIPG), incomplete (IIPG), and symmetric (SIPG) approximation of the diffusive form.

Due to the consistency of the semidiscrete DG scheme (cf. Definition 2), for the exact solution Ψ (where $\psi = \Psi - z$) and some $v \in S_{h,p}^{\tau,q}$ it holds

$$(\partial_t\vartheta(\psi),v) + a_{h,m}(\psi;\Psi,v) = \ell_{h,m}(v).$$

We integrate the equation above over a time interval I_m, $m = 1,\dots,r$, so that

$$\int_{I_m} (\partial_t\vartheta(\psi),v)\,\mathrm{d}t + \int_{I_m} a_{h,m}(\psi;\Psi,v)\,\mathrm{d}t = \int_{I_m} \ell_{h,m}(v)\,\mathrm{d}t.$$

Then, we apply the integration by parts to the first integral above

$$\begin{aligned}\int_{I_m} (\partial_t\vartheta(\psi),v)\,\mathrm{d}t = &-\int_{I_m} (\vartheta(\psi),\partial_t v)\,\mathrm{d}t \\ &+ (\vartheta(\psi)|_m^-, v|_m^-) - (\vartheta(\psi)|_{m-1}^+, v|_{m-1}^+).\end{aligned} \tag{5.7}$$

Due to continuity of ϑ (cf. the assumption (H1)), and assuming $\Psi \in H^2(\Omega;\mathcal{T}_{h,m})$ (hence, $\psi \in H^2(\Omega;\mathcal{T}_{h,m})$), $m = 0,\dots,r$, we have that

$$\vartheta(\psi)|_{m-1}^+ = \vartheta(\psi)|_{m-1}^-,\ m = 1,\dots,r.$$

Thus, (5.7) can be rewritten as

$$\int_{I_m} (\partial_t\vartheta(\psi),v)\,\mathrm{d}t = -\int_{I_m} (\vartheta(\psi),\partial_t v)\,\mathrm{d}t \tag{5.8}$$

$$+ (\vartheta(\psi)|_m^-, v|_m^-) - (\vartheta(\psi)|_{m-1}^-, v|_{m-1}^+). \tag{5.9}$$

We define the form

$$\begin{aligned}A_{h,m}(\psi;\Psi,v) = &\int_{I_m} \Big(-(\vartheta(\psi),\partial_t v) + a_{h,m}(\psi;\Psi,v) - \ell_{h,m}(v)\Big)\,\mathrm{d}t \\ &+ (\vartheta(\psi)|_m^-, v|_m^-) - (\vartheta(\psi)|_{m-1}^-, v|_{m-1}^+).\end{aligned} \tag{5.10}$$

Definition 6. *We say that a function* $\Psi_{h\tau} \in S_{h,p}^{\tau,q}$ *is a STDG approximate solution of problem* (2.1)*, if*

$$A_{h,m}(\Psi_{h\tau} - z; \Psi_{h\tau}, v) = 0 \quad \forall v \in S_{h,p,m}^{\tau,q}, \; m = 1, \ldots, r \tag{5.11}$$

with $\Psi_{h\tau}|_0^- = \Pi_{h,p,0}\Psi_0$.

Furthermore, if we choose different polynomial space, we get an *hp*-STDG method.

Definition 7. *We say that a function* $\Psi_{h\tau} \in S_{h,\boldsymbol{p}}^{\tau,q}$ *is an hp-STDG approximate solution of problem* (2.1)*, if*

$$A_{h,m}(\Psi_{h\tau} - z; \Psi_{h\tau}, v) = 0 \quad \forall v \in S_{h,\boldsymbol{p},m}^{\tau,q}, \; m = 1, \ldots, r \tag{5.12}$$

with $\Psi_{h\tau}|_0^- = \Pi_{h,p,0}\Psi_0$.

In the similar manner, we discretize the ψ-formulation of Richards' equation (2.2). Namely, we introduce the forms (cf. (2.22))

$$\bar{A}_{h,m}(\psi; \Psi, v) = A_{h,m}(\psi; \Psi, v) + \int_{I_m} b_{h,m}(\psi; v)\,\mathrm{d}t, \tag{5.13}$$

$$b_{h,m}(\psi; v) = \sum_{K \in \mathcal{T}_{h,m}} \int_K \mathbf{K}(\psi)\boldsymbol{e}_2 \cdot \nabla v \,\mathrm{d}x - \sum_{\Gamma \in \mathcal{F}_{h,m}^{ID}} \int_\Gamma H(\psi^{(L)}, \psi^{(R)}, \boldsymbol{n}) v \mathrm{d}S. \tag{5.14}$$

Definition 8. *We say that a function* $\psi_{h\tau} \in S_{h,\boldsymbol{p}}^{\tau,q}$ *is an hp-STDG approximate solution of problem* (2.2)*, if*

$$\bar{A}_{h,m}(\psi_{h\tau}; \psi_{h\tau}, v) = 0 \quad \forall v \in S_{h,\boldsymbol{p},m}^{\tau,q}, \; m = 1, \ldots, r \tag{5.15}$$

with $\psi_{h\tau}|_0^- = \Pi_{h,p,0}\psi_0$.

Now, when we have discretized the problems (2.1) and (2.2), we may proceed with finding their solutions. Namely, in the next chapter, we propose a suitable solution strategy.

6. Solution strategy

In this chapter, we proceed with finding the solution of the discrete schemes given by Definitions 7–8. We give their algebraic interpretations and then define a Newton-like method and the Anderson acceleration. In addition, we provide a numerical study of these methods on a hydraulic dam example. Moreover, we introduce the concept of mesh adaptation and define the anisotropic *hp*-STDG method. Finally, we make a comparison between the Ψ-formulation and ψ-formulation of Richards' equation using a single ring infiltration experiment.

6.1 Algebraic system

The equations (5.12) and (5.15) represent r-systems of nonlinear algebraic equations, i.e., on each time subinterval I_m, $m = 1, \dots, r$ there is a system of nonlinear algebraic equations. Each of the algebraic systems has N_m equations, where

$$N_m = \dim(S^{\tau,q}_{h,\boldsymbol{p},m}) = (q+1) \sum_{K \in \mathcal{T}_{h,m}} \frac{(p_K+1)(p_K+2)}{2}.$$

Furthermore, we denote

$$\Psi^m_{h\tau} = \Psi_{h\tau}|_{\Omega \times I_m} \in S^{\tau,q}_{h,\boldsymbol{p},m}, \ m = 1, \dots, r.$$

Let $B_{h,m} = \{\varphi_i(x,t)\}_{i=1}^{N_m}$ be a basis of the space $S^{\tau,q}_{h,\boldsymbol{p},m}$, $m = 1, \dots, r$. Then we may express the solution $\Psi^m_{h\tau} \in S^{\tau,q}_{h,\boldsymbol{p},m}$ in terms of the basis $B_{h,m}$ as

$$\Psi^m_{h\tau} = \sum_{j=1}^{N_m} \xi^{m,j} \varphi_j(x,t),$$

where $\xi^{m,j}$, $j = 1, \dots, N_m$ are the basis coefficients of $\Psi^m_{h\tau}$ with respect to the basis $B_{h,m}$. We define a vector valued mapping $\boldsymbol{F}_{h,m} : \mathbb{R}^{N_m} \to \mathbb{R}^{N_m}$ as

$$\boldsymbol{F}_{h,m}(\boldsymbol{\xi}_m) = \{A_{h,m}(\Psi^m_{h\tau} - z; \Psi^m_{h\tau}, \varphi_i)\}_{i=1}^{N_m}, \quad \boldsymbol{\xi}_m = \{\xi^{m,j}\}_{j=1}^{N_m}, \quad m = 1, \dots, r.$$

Hence, the system (5.12) from Definition 7 is equivalent to the problem: Find $\boldsymbol{\xi}_m$ such that

$$\boldsymbol{F}_{h,m}(\boldsymbol{\xi}_m) = \boldsymbol{0}, \quad m = 1, \dots, r. \tag{6.1}$$

On the other hand, let us denote

$$\psi^m_{h\tau} = \psi_{h\tau}|_{\Omega \times I_m} \in S^{\tau,q}_{h,\boldsymbol{p},m}, \ m = 1, \dots, r,$$

where $\psi_{h\tau}$ is the solution to (5.15), and let $\bar{\boldsymbol{\xi}}_m$ be the algebraic representation of it. Thus, we define

$$\bar{\boldsymbol{F}}_{h,m}(\bar{\boldsymbol{\xi}}_m) = \{\bar{A}_{h,m}(\psi^m_{h\tau}; \psi^m_{h\tau}, \varphi_i)\}_{i=1}^{N_m}, \quad \bar{\boldsymbol{\xi}}_m = \{\bar{\xi}^{m,j}\}_{j=1}^{N_m}, \quad m = 1, \dots, r.$$

Therefore, instead of (5.15), we may consider: Find $\bar{\boldsymbol{\xi}}_m$ such that

$$\bar{\boldsymbol{F}}_{h,m}(\bar{\boldsymbol{\xi}}_m) = \boldsymbol{0}, \quad m = 1, \dots, r. \tag{6.2}$$

The systems (6.1)–(6.2) are strongly nonlinear, and thus, it is often challenging to solve them numerically. Namely, for the considered degenerate parabolic problems, the Newton method fails due to singularity of the Jacobian matrix caused by vanishing $\mathrm{d}\mathbf{K}(\psi)/\mathrm{d}\psi$ and $\vartheta'(\psi)$. A study on iterative methods for Richards' equation is done in [60]. Other methods were proposed to deal with the nonlinear algebraic system, e.g., modified Picard method [60], L-scheme [86, 74, 75], parametrization techniques [12, 7], etc. We follow the approach from [34] that uses a Newton-like method and the Anderson acceleration originally designed to improve the convergence of the Picard method [95].

6.1.1 Formal linearization

In what follows we formally linearize the nonlinear system $\boldsymbol{F}_{h,m}$ given by (6.1) (also (6.2)). If we apply again the partial integration on the first term on the right-hand side of (5.7), we obtain the identity

$$\int_{I_m} (\partial_t \vartheta(\psi), v)\,\mathrm{d}t = \int_{I_m} (\partial_t \vartheta(\psi), v)\,\mathrm{d}t + (\{\vartheta(\psi)\}_{m-1}, v|^+_{m-1}). \tag{6.3}$$

Moreover, we approximate the last term from above as

$$\{\vartheta(\psi)\}_{m-1} = \int_{\psi|^-_{m-1}}^{\psi|^+_{m-1}} \vartheta(s)\mathrm{d}s \approx \vartheta'(\psi)(\psi|^+_{m-1} - \psi|^-_{m-1}) = \vartheta'(\psi)\{\psi\}_{m-1}. \tag{6.4}$$

Therefore, by virtue of (6.3), (5.8), (6.4), and noting that $\partial_t\psi = \partial_t(\Psi - z) = \partial_t\Psi$, and $\{\psi\}_{m-1} = \{\Psi\}_{m-1}$, we obtain

$$\begin{aligned}
&-\int_{I_m} (\vartheta(\psi), \partial_t v)\,\mathrm{d}t + (\vartheta(\psi)|^-_m, v|^-_m) - (\vartheta(\psi)|^-_{m-1}, v|^+_{m-1}) \\
&= \int_{I_m} (\partial_t\vartheta(\psi), v)\,\mathrm{d}t + (\{\vartheta(\psi)\}_{m-1}, v|^+_{m-1}) \\
&= \int_{I_m} (\vartheta'(\psi)\partial_t\psi, v)\,\mathrm{d}t + (\{\vartheta(\psi)\}_{m-1}, v|^+_{m-1}) \\
&= \int_{I_m} (\vartheta'(\psi)\partial_t\psi, v)\,\mathrm{d}t + (\vartheta'(\psi)\{\psi\}_{m-1}, v|^+_{m-1}) \qquad (6.5)\\
&= \int_{I_m} (\vartheta'(\psi)\partial_t\Psi, v)\,\mathrm{d}t + (\vartheta'(\psi)\{\Psi\}_{m-1}, v|^+_{m-1}). \qquad (6.6)
\end{aligned}$$

We define the forms

$$A^L_{h,m}(\psi; \Psi, v) = \int_{I_m} \Big((\vartheta'(\psi)\partial_t\Psi, v) + a_{h,m}(\psi; \Psi, v) \Big)\,\mathrm{d}t + (\vartheta'(\psi)\Psi|^+_{m-1}, v|^+_{m-1}) \tag{6.7}$$

$$L_{h,m}(\psi; \Psi, v) = \int_{I_m} l_{h,m}(v)\,\mathrm{d}t - (\vartheta'(\psi)\Psi|^-_{m-1}, v|^+_{m-1}). \tag{6.8}$$

We have an approximation of the form $A_{h,m}$

$$A_{h,m}(\psi; \Psi, v) \approx A^L_{h,m}(\psi; \Psi, v) - L_{h,m}(\psi; \Psi, v). \tag{6.9}$$

Furthermore, we define the flux matrix $\boldsymbol{C}_{h,m}$ and the vector $\boldsymbol{q}_{h,m}$ as

$$\begin{aligned}
\boldsymbol{C}_{h,m}(\bar{\boldsymbol{\xi}}) &= \{A^L_{h,m}(\psi; \varphi_i, \varphi_j)\}_{i,j=1}^{N_m}, \qquad (6.10)\\
\boldsymbol{q}_{h,m}(\bar{\boldsymbol{\xi}}) &= \{L_{h,m}(\psi; \psi, \varphi_i)\}_{i=1}^{N_m}, \qquad (6.11)
\end{aligned}$$

where φ_i, $i = 1, \dots, N_m$ are the basis functions from $B_{h,m}$ and $\bar{\boldsymbol{\xi}} \in \mathbb{R}^{N_m}$ is the algebraic representation of $\psi \in S^{\tau,q}_{h,\boldsymbol{p},m}$ using the basis $B_{h,m}$. Hence, we have that

$$\boldsymbol{F}_{h,m}(\boldsymbol{\xi}_m) \approx \boldsymbol{C}_{h,m}(\boldsymbol{\xi}_m)\boldsymbol{\xi_m} - \boldsymbol{q}_{h,m}(\boldsymbol{\xi}_m), \ m = 1, \dots, r. \tag{6.12}$$

The flux matrix $\boldsymbol{C}_{h,m}$ is sparse and has a block structure; namely, each block-row of $\boldsymbol{C}_{h,m}$ corresponds to an element $K \in \mathcal{T}_{h,m}$. Moreover, the sparsity of $\boldsymbol{C}_{h,m}$ is equal to sparsity of the Jacobian matrix $D\boldsymbol{F}_{h,m}(\boldsymbol{\xi})/D\boldsymbol{\xi}$. Therefore, we shall use the approximation following from (6.12)

$$\boldsymbol{C}_{h,m}(\boldsymbol{\xi}) \approx \frac{D\boldsymbol{F}_{h,m}(\boldsymbol{\xi})}{D\boldsymbol{\xi}}. \tag{6.13}$$

In the similar way, we linearize the nonlinear system (6.2). Namely, we define

$$\begin{aligned} \bar{A}^L_{h,m}(\psi;\Psi,v) &= \int_{I_m} \Big((\vartheta'(\psi)\partial_t\Psi, v) + a_{h,m}(\psi;\Psi,v) + b_{h,m}(\psi;\Psi,v) \Big)\,\mathrm{d}t \\ &\quad + (\vartheta'(\psi)\Psi|^+_{m-1}, v|^+_{m-1}) \qquad (6.14) \\ \bar{\boldsymbol{C}}_{h,m}(\bar{\boldsymbol{\xi}}) &= \{\bar{A}^L_{h,m}(\psi;\varphi_i,\varphi_j)\}_{i,j=1}^{N_m}, \qquad (6.15) \end{aligned}$$

so that the resulting approximation of the problem (6.2) is

$$\bar{\boldsymbol{F}}_{h,m}(\bar{\boldsymbol{\xi}}_m) \approx \bar{\boldsymbol{C}}_{h,m}(\bar{\boldsymbol{\xi}}_m)\bar{\boldsymbol{\xi}}_m - \boldsymbol{q}_{h,m}(\bar{\boldsymbol{\xi}}_m), \ m = 1, \dots, r. \tag{6.16}$$

6.1.2 Damped Newton-like method

We define a damped Newton-like method [27] which generates a sequence of approximations $\{\boldsymbol{\xi}^l_m\}_l$ to the solution $\boldsymbol{\xi}_m$ of the nonlinear system (6.1) using Algorithm 1. Analogously, the algorithm is defined for the system (6.2).

Algorithm 1 Newton-like method

```
Let $\boldsymbol{\xi}^0_m \in \mathbb{R}^{N_m}$ be given.
for $l = 0, 1, \dots$ do
    (a) Find $\boldsymbol{d}^l \in \mathbb{R}^{N_m}$ such that
            $\boldsymbol{C}_{h,m}(\boldsymbol{\xi}^l_m)\boldsymbol{d}^l = -\boldsymbol{F}_{h,m}(\boldsymbol{\xi}^l_m).$   (6.17)
    (b) Set
            $\boldsymbol{\xi}^{l+1}_m = \boldsymbol{\xi}^l_m + \lambda^l \boldsymbol{d}^l,$   (6.18)
        where $\lambda^l \in (0, 1]$ is a damping parameter such that
            $\delta^l := \frac{\left\| \boldsymbol{F}_{h,m}(\boldsymbol{\xi}^{l+1}_m) \right\|}{\left\| \boldsymbol{F}_{h,m}(\boldsymbol{\xi}^l_m) \right\|} < 1.$   (6.19)
    (c) If the stopping criterion is met, then STOP.
```

The role of the damping parameter is to improve the convergence of the Newton method when the initial guess is far from the solution $\boldsymbol{\xi}_m$. At each iteration, we set $\lambda^l = 1$, and if needed we multiply it by 0.75 until the criterion (6.19) is fulfilled. Hence, to obtain the next iteration $\boldsymbol{\xi}_m^{l+1}$ it is usually necessary to evaluate the matrix $\boldsymbol{F}_{h,m}$ several times.

Furthermore, if we set $\lambda = 1$ in (6.18) and combine (6.18) with (6.17), we get

$$\boldsymbol{\xi}_m^{l+1} = \boldsymbol{\xi}_m^l - \boldsymbol{C}_{h,m}(\boldsymbol{\xi}_m^l)\boldsymbol{F}_{h,m}(\boldsymbol{\xi}_m^l) =: G(\boldsymbol{\xi}_m^l). \tag{6.20}$$

Especially, if the mapping G is contractive, the relation (6.20) is a Picard iteration. As mentioned earlier, the Anderson acceleration has been developed to improve the convergence of this method. We present this technique in Algorithm 2.

Algorithm 2 Anderson acceleration

```
Let $n \in \mathbb{N}$ and $\boldsymbol{\xi}_m^0 \in \mathbb{R}^{N_m}$ be given.
Set $\boldsymbol{\xi}_m^1 := G(\boldsymbol{\xi}_m^0)$.
for $l = 0, 1, \dots$ do
    (a) Set $n_l := \min(l, n)$.
    (b) Set $g_i := G(\boldsymbol{\xi}_m^i) - \boldsymbol{\xi}_m^i$, $i = l - n_l, \dots, l$.
    (c) Find $\alpha_i$, $i = 0, \dots, n_l$ such that $\sum_{i=0}^{n_l} \alpha_i = 1$ and
            $(\alpha_0, \dots, \alpha_{n_l}) = \arg\min_{(\beta_0, \dots, \beta_{n_l})} \left\| \sum_{i=0}^{n_l} \beta_i g_{l-i} \right\|.$
    (d) Set $\boldsymbol{\xi}_m^{l+1} = \sum_{i=0}^{n_l} \alpha_i g_{l-i}$.
    (e) If the stopping criterion is met, then STOP.
```

Clearly, Algorithm 1 with $\lambda^l = 1$, $l = 0, 1, \dots$ is equivalent to Algorithm 2 with $n = 1$. In Subsection 6.1.4, we study the Newton method with and without Anderson acceleration on a numerical example.

We discuss the stopping criterion mentioned in Algorithms 1–2 in the subsequent subsection. Finally, we mention that the linear system (6.17) is solved using GMRES method with the block ILU(0) preconditioner [81].

6.1.3 Stopping criteria

We specify the stopping criterion in Algorithms 1–2 and determine the length of the time step τ_m, $m = 1, \dots, r$ in (5.12)–(5.15). To preserve the accuracy and efficiency of the computations we should avoid too large or too small time steps and too strong or too weak stopping criteria in Algorithms 1–2. In particular, we balance three types of error, namely, errors arising from

- space discretization,
- time discretization,
- nonlinear algebraic system computation.

We use a technique from [35] that uses approximation of the mentioned errors in a dual norm.

Let $\boldsymbol{\xi}^l_m$ be the output of Algorithm 1 (or Algorithm 2) and $\tilde{\Psi}_{h\tau} \in S^{\tau,q}_{h,\boldsymbol{p}}$ be the solution of (5.12) that corresponds to $\boldsymbol{\xi}^l_m$. Moreover, we denote

$$\tilde{\Psi}_{h\tau}|_{I_m} =: \tilde{\Psi}^m_{h\tau} \in S^{\tau,q}_{h,\boldsymbol{p},m}, \ m = 1, \dots, r. \tag{6.21}$$

Let us note that the solution above does not fulfill the relation (5.12). Hence, we define

- algebraic estimator

$$\eta^m_A(\tilde{\Psi}^m_{h\tau}, v) = \max_{\substack{v \in S^{\tau,q}_{h,\boldsymbol{p},m} \\ v \neq 0}} \frac{A_{h,m}(\tilde{\Psi}^m_{h\tau} - z; \tilde{\Psi}^m_{h\tau}, v)}{\|v\|_X}, \tag{6.22}$$

- space-algebraic estimator

$$\eta^m_{SA}(\tilde{\Psi}^m_{h\tau}, v) = \max_{\substack{v \in S^{\tau,q}_{h,\boldsymbol{p}+1,m} \\ v \neq 0}} \frac{A_{h,m}(\tilde{\Psi}^m_{h\tau} - z; \tilde{\Psi}^m_{h\tau}, v)}{\|v\|_X}, \tag{6.23}$$

- time-algebraic estimator

$$\eta^m_{TA}(\tilde{\Psi}^m_{h\tau}, v) = \max_{\substack{v \in S^{\tau,q+1}_{h,\boldsymbol{p},m} \\ v \neq 0}} \frac{A_{h,m}(\tilde{\Psi}^m_{h\tau} - z; \tilde{\Psi}^m_{h\tau}, v)}{\|v\|_X}, \tag{6.24}$$

where

$$\|v\|_X := \left(\int_{I_m} \sum_{K \in \mathcal{T}_{h,m}} \left(\|v\|^2_{L^2(K)} + \|\nabla v\|^2_{L^2(K)} + \|\partial_t v\|^2_{L^2(K)} \right) \mathrm{d}t \right)^{1/2},$$

and the spaces $S^{\tau,q}_{h,\boldsymbol{p}+1,m}$ and $S^{\tau,q+1}_{h,\boldsymbol{p},m}$ are defined analogously to (5.2). The maxima in (6.22)–(6.24) can be obtained using the Lagrange multipliers (see [32]).

The estimators (6.22)–(6.24) are used to define the stopping criterion in Algorithms 1–2 and to set the length of the time step. Namely, at each time level,

- we solve (5.12) using either Algorithm 1 or Algorithm 2 until satisfying the following condition

$$\eta^m_A \leq c_A \min(\eta^m_{SA}, \eta^m_{TA}), \ m = 1, \dots, r, \tag{6.25}$$

where $c_A > 0$ is a constant,

- we set the condition

$$\eta^m_{TA} \leq c_T \eta^m_{SA}, \ m = 1, \dots, r, \tag{6.26}$$

where $c_T \in [0.1, 1)$. If this condition is not fulfilled then the computation is repeated for a smaller time step, otherwise, we set

$$\tau^{opt}_m := \tau_m \left(c_T \frac{\eta^m_{SA}}{\eta^m_{TA}} \right)^{\frac{1}{q+1}}. \tag{6.27}$$

Table 6.1: Landfill dam: Parameters for the van Genuchten-Mualem model corresponding to different materials.

Parameters	Gravel	Clay	Silt clay
α	250.0	0.8	2.0
n	1.41	1.20	1.41
m	0.291	0.167	0.291
K_S	4.630×10^{-7}	5.556×10^{-9}	1.238×10^{-8}
θ_S	0.60	0.38	0.45
θ_r	0.0000	0.0600	0.0067
S_S	0.01	0.01	0.01

6.1.4 Numerical study of nonlinear solvers

In what follows, we compare Algorithms 1–2 for the solution of the nonlinear algebraic system (6.1) on a practical example. Namely, we consider a simulation of a nonhomogeneous landfill dam, which is consisted of three materials: gravel, clay and silt clay (see Fig. 6.1). For the closure law, we choose the van Genuchten-Mualem model [92, 66] defined by (1.14)–(1.15) with parameters specified in Table 6.1. The computations within this example are performed on the Karlin cluster [101] using the in-house code ADGFEM [31].

The computational domain Ω with the boundary $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N \cup \partial\Omega_E$, where by $\partial\Omega_E$ is denoted the seepage face boundary (cf. Section 1.4), is depicted in Fig. 6.1. In particular, on $\partial\Omega_D$ we prescribe the Dirichlet boundary condition $\Psi_D = 15$, while on $\partial\Omega_N$ we have the homogeneous Neumann boundary condition. As described in Chapter 1, the seepage boundary condition is approximated using these two type of boundary conditions (cf. (1.16)). The domain Ω is discretized using a quasi-uniform grid presented in Fig. 6.2. The initial condition is set to be $\Psi_0 = 0$. We remark that there is an inconsistency between boundary and initial conditions, which may affect computations around $t \approx 0$. We set the final time to be $T = 5$ days.

We use the Ψ-formulation of Richards' equation to model this experiment. In particular, we use the STDG method (cf. Definition 6) to solve (2.1) with the IIPG variant of the DG method and fixed polynomial degree $p = 2$. The time variable is approximated using piecewise linear functions and the time step is chosen adaptively. In Fig. 6.3, the evolution of the pressure head ψ at different time instances is presented.

The arising nonlinear algebraic systems (6.1) are solved using either Algorithm 1 or Algorithm 2. Table 6.2 (see also Figs. 6.4–6.5) shows the number of nonlinear iterations produced by these two algorithms, where the constant c_A from the stopping criterion (6.25) is set to be $5 \cdot 10^{-3}$. In this example, we note the superiority of the Anderson acceleration applied to the Newton method, i.e., Algorithm 2. Moreover, in Figs. 6.4–6.5, the error estimators, namely, the algebraic estimator η_A^m, space-algebraic estimator η_{SA}^m, and time-algebraic estimator η_{TA}^m (cf. (6.22)–(6.24)), together with the size of the time step τ_m, are plotted with respect to iterations in time. In particular, Fig. 6.4 shows the mentioned quantities for Algorithm 1 with respect to time steps, where each dot corresponds

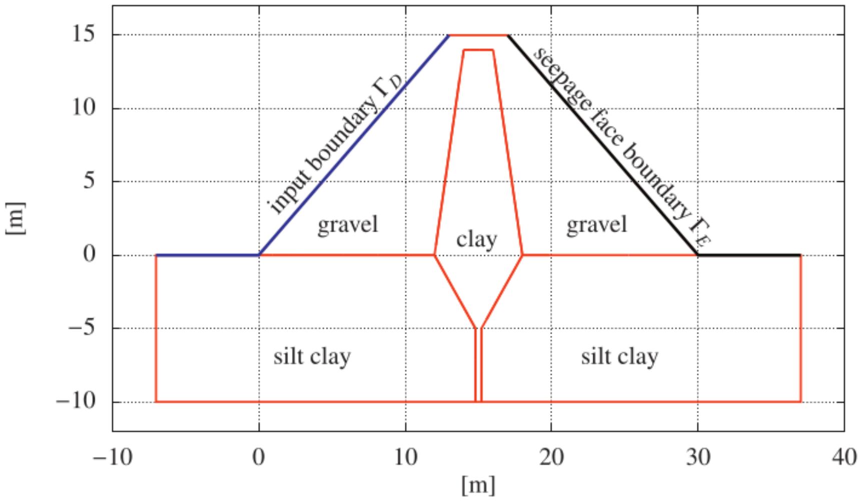

Figure 6.1: Landfill dam: domain geometry [34].

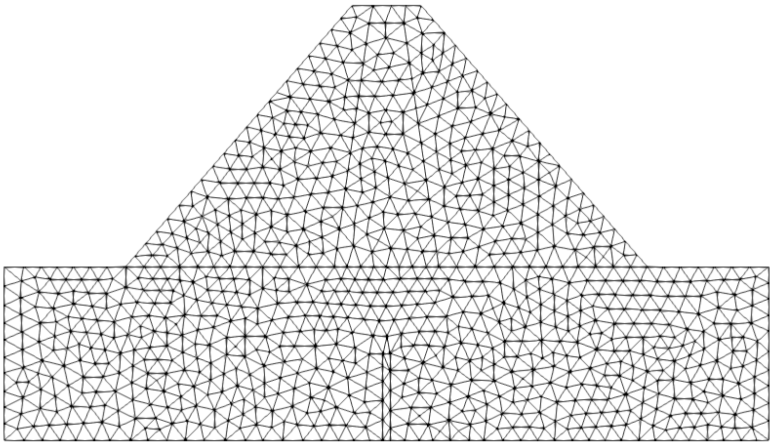

Figure 6.2: Landfill dam: the quasi-uniform mesh used in computations.

Table 6.2: Landfill dam: Computational data obtained using different nonlinear solvers.

Method	$\#\tau_m$	#refused τ_m	#nonlinear it.	CPU[s]
Algorithm 1	434	222	9 649	28757.60
Algorithm 2	179	11	2 405	6102.40

to one nonlinear iteration. In the first third of time iterations, we note the increase of the size of time step at each iteration. After this, the convergence of Algorithm 1 is not always achieved; therefore the evaluations are repeated with a smaller size of time step. On contrary, Fig. 6.5 shows that Algorithm 2 is more efficient in this case, since almost at each time step the condition (6.25) is fulfilled.

On the other hand, if we decrease the constant c_A from (6.25), then these two algorithms exhibit different behavior. Namely, in our second test case (see Figs. 6.6–6.7), we set $c_A = 10^{-3}$ and consider the same settings. Algorithm 2 seems to fail since after several iterations the condition for the convergence is not satisfied and the time step is set to be minimal (around 10^{-6}) in the rest of computations leading to an incomplete simulation. However, using Algorithm 1 (the Newton method without the Anderson acceleration) the simulation is successfully completed in 65 040 seconds using 422 iterations in time and 327 384 total accumulated nonlinear iterations, which is computationally more expensive comparing to Table 6.2.

By this example, we show that the technique presented in Subsection 6.1.3, originally developed for numerical simulation of time dependent compressible flow, does not always work for the porous media flow application. Namely, the key difference is that in our example we have a nonlinearity in the time dependent term '$\vartheta(\psi)$' instead of the linear term as in the mentioned compressible flow model; therefore, the structure of the flux matrices $\boldsymbol{C}_{h,m}$ (6.10) for these two problems are different, where the contribution of the nonlinear term is nonnegligible.

6.2 Regularization of constitutive relations

As mentioned in Section 1.3, Richards' equation degenerates when the flow is transiting from unsaturated to saturated regions, which makes derivatives of the constitutive laws vanish. In this section, we recall the van Genuchten-Mualem model (1.14)–(1.15) from Section 1.2 and define a modification of these constitutive relations in critical regions, particularly, when $\psi \to 0$.

In the case of fast diffusion type of degeneracy, we have that $S_S = 0$ and $\psi \geq 0$, which yields $C(\psi) = \vartheta'(\psi) = 0$ and the resulting equation becomes elliptic. To avoid this, we set $S_S > 0$. Another approach [76] suggests redefining $\theta(\psi)$ as $\theta(\psi) + \epsilon\psi$, where $\epsilon > 0$ is a small regularization parameter. Conversely, in the case of the slow diffusion type of degeneracy, higher values of α and n produce steep gradients in (1.14) when $\psi \to 0$ (cf. Fig. 1.3). The point $\psi = \alpha^{-1}$ appears as a numerical singularity and affects the convergence [10].

We follow the regularization proposed in [34]; on a interval $(-\epsilon, 0)$, for some

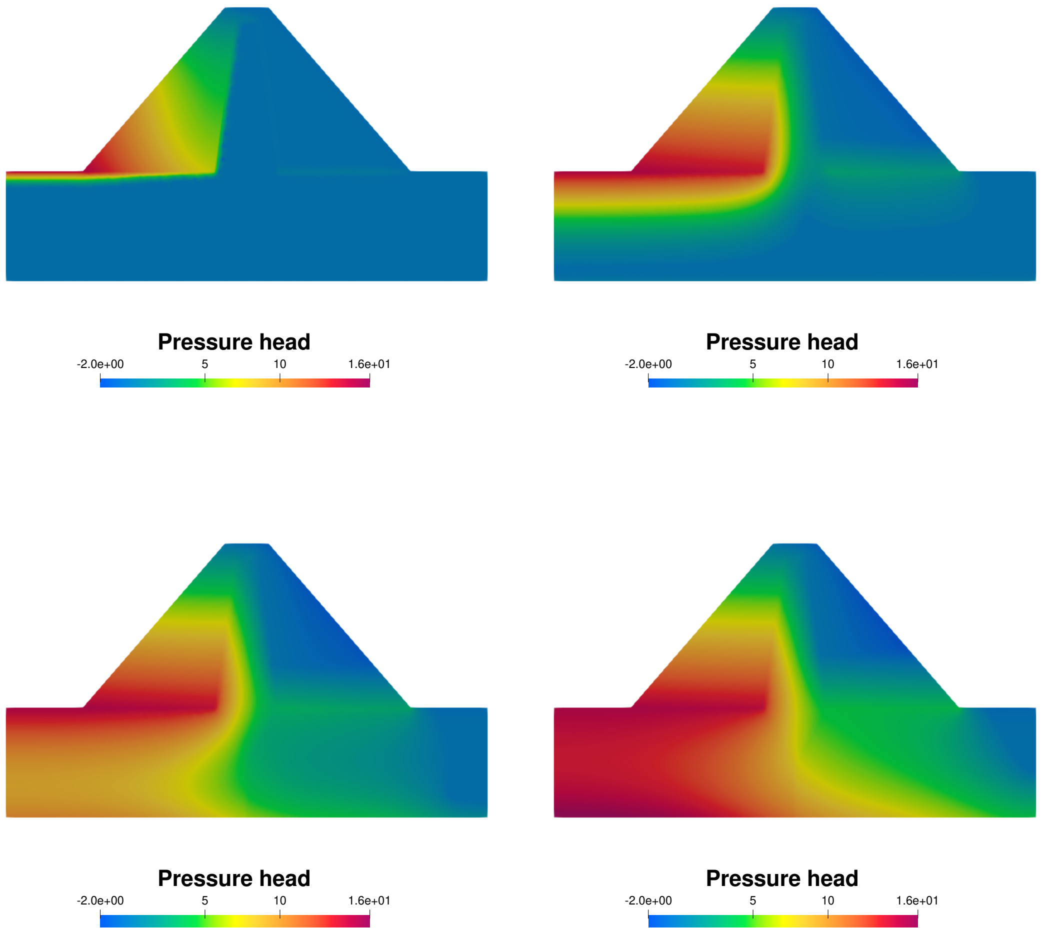

Figure 6.3: Landfill dam: the pressure head at $t = 0.04$ day (1st row left), $t = 1$ day (1st row right), $t = 3$ days (2nd row left) and $t = 5$ days (2nd row right).

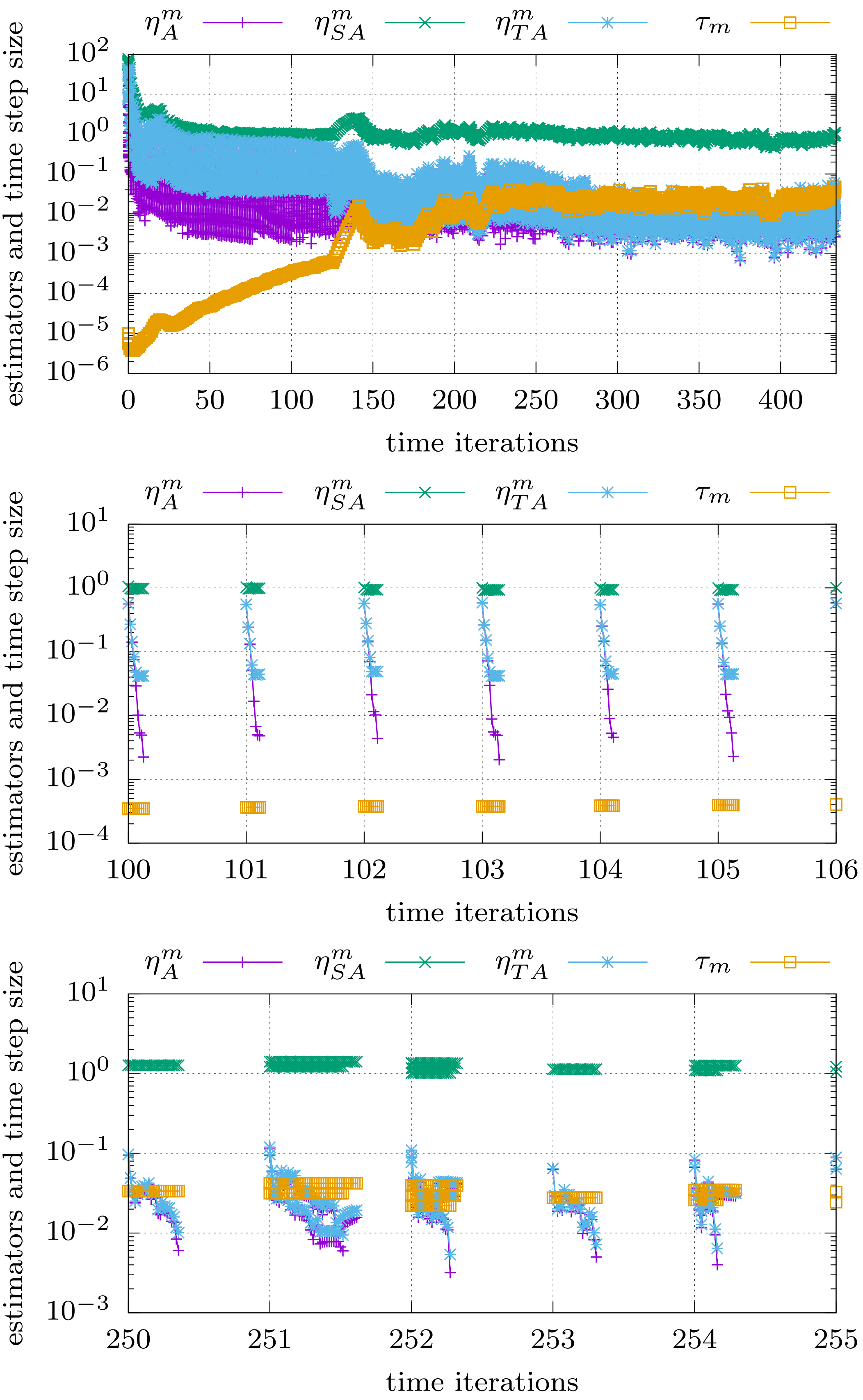

Figure 6.4: Landfill dam: error estimators η_A^m, η_{SA}^m, η_{TA}^m defined by (6.22)–(6.24) and the size of the time step τ_m *versus* time steps for Algorithm 1 with $c_A = 5 \cdot 10^{-3}$.

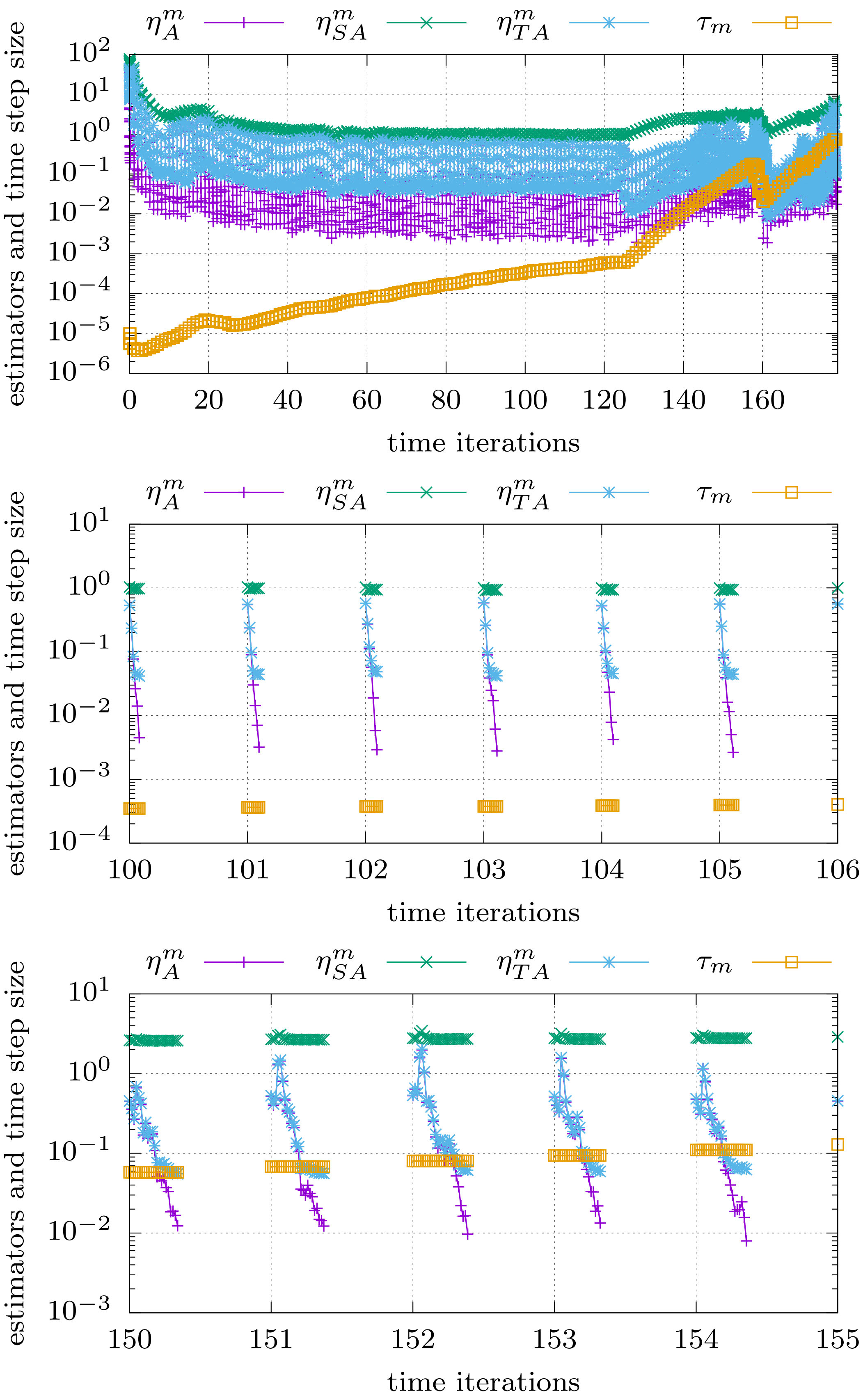

Figure 6.5: Landfill dam: error estimators η_A^m, η_{SA}^m, η_{TA}^m defined by (6.22)–(6.24) and the size of the time step τ_m *versus* time steps for Algorithm 2 with $c_A = 5 \cdot 10^{-3}$.

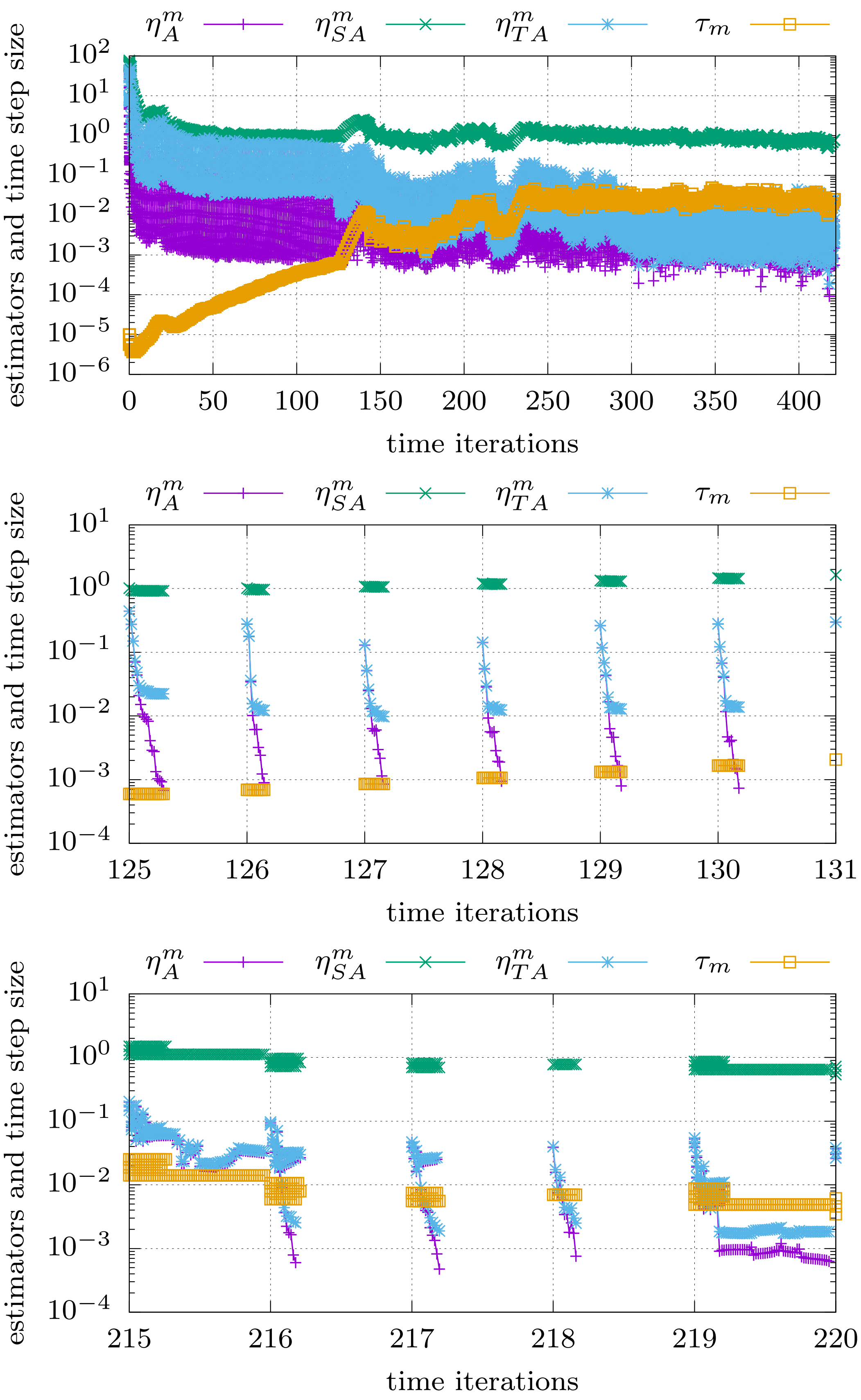

Figure 6.6: Landfill dam: error estimators η_A^m, η_{SA}^m, η_{TA}^m defined by (6.22)–(6.24) and the size of the time step τ_m *versus* time steps for Algorithm 1 with $c_A = 10^{-3}$.

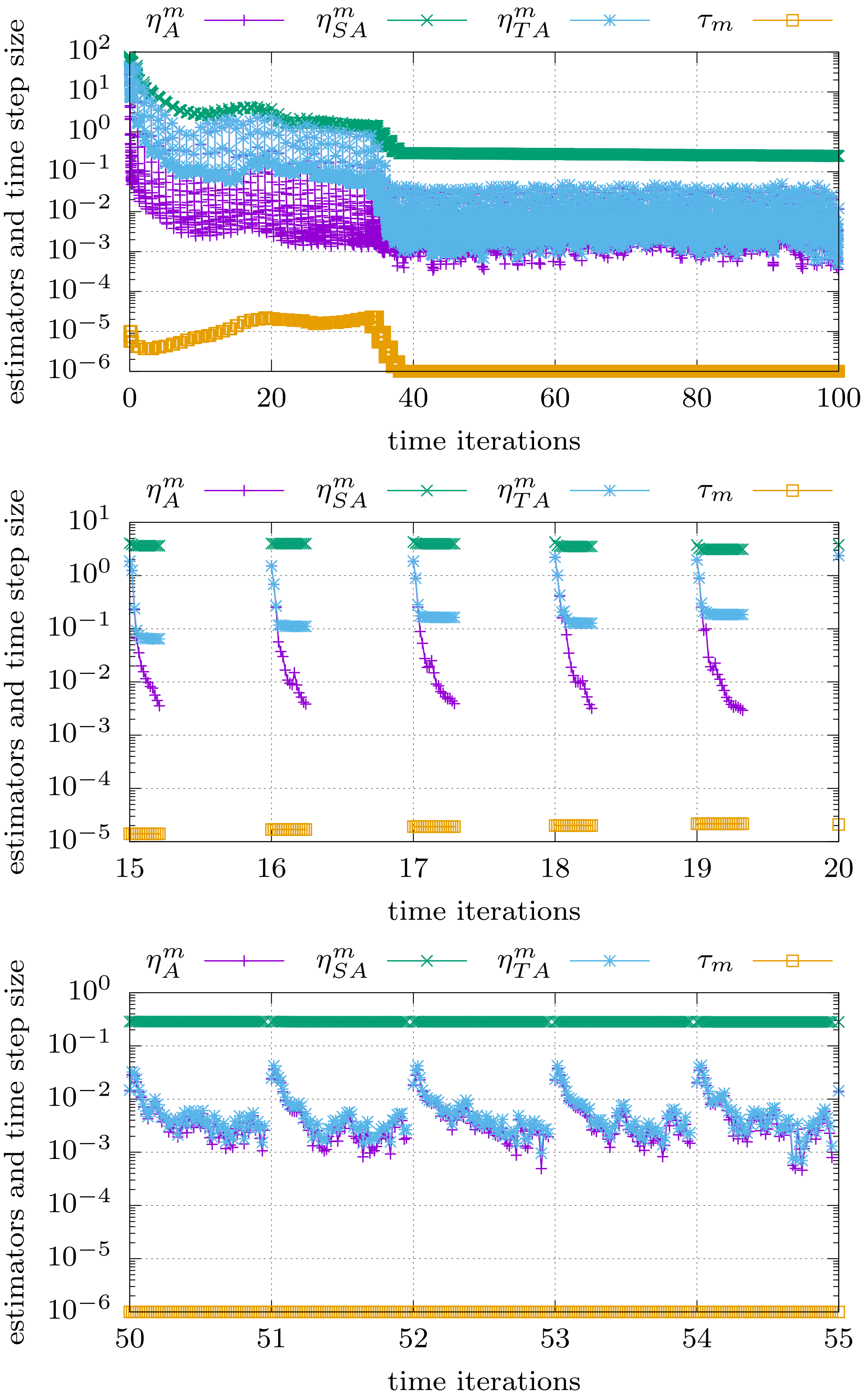

Figure 6.7: Landfill dam: error estimators η_A^m, η_{SA}^m, η_{TA}^m defined by (6.22)–(6.24) and the size of the time step τ_m *versus* time steps for Algorithm 2 with $c_A = 10^{-3}$.

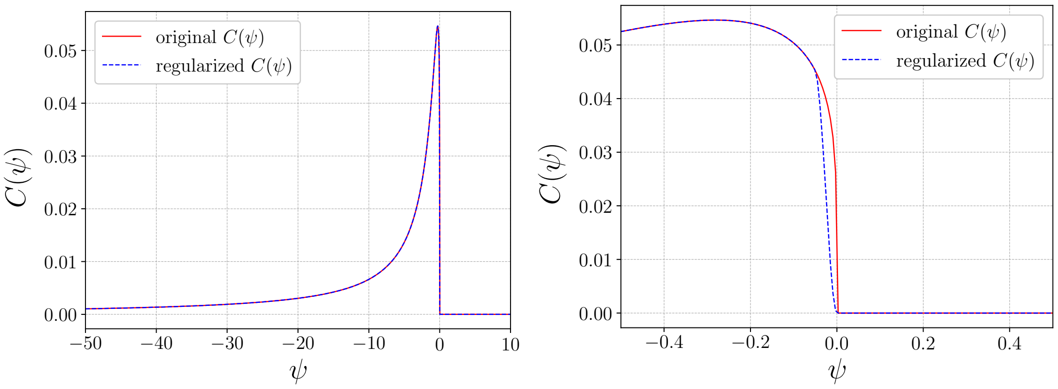

Figure 6.8: The original retention water capacity $C(\psi)$ for the van Genuchten model (1.14) with their regularization for the parameter $\epsilon = 0.05$ and with the parameters given in Table 1.1 and $S_S = 0$.

small $\epsilon > 0$, we replace the capillary capacity function with a cubic polynomial function which all together is a continuously differentiable function over whole its domain. The cubic polynomial function is defined using values $C(0)$, $C(-\epsilon)$, $C'(0)$, and $C'(-\epsilon)$. We set $C'(0) = 0$, and approximate $C'(-\epsilon)$ using central difference formula. We define a modified ϑ such that (1.9) is fulfilled. An example of such regularization for $\epsilon = 0.05$ is illustrated in Fig. 6.2. Analogously, we define a regularization for the constitutive law $K_r(\psi)$. In the next chapter, we investigate the influence of regularization using numerical experiments. In our examples, we do not include the case $\psi \to -\infty$; we refer to [67, 50, 62] for the regularization techniques in this case.

6.3 Mesh adaptation

Within this section we introduce an *hp*-adaptation technique developed in [30] and formulate an algorithm that shall be used later in numerical examples. Namely, the anisotropic *hp*-mesh adaptation technique uses a high order degree polynomial approximation on anisotropic elements, which results in the reduction of degrees of freedom without loss of accuracy.

This technique is based on minimizing degrees of freedom for the prescribed tolerance $\omega > 0$ for the interpolation error in the $L^\infty(0, T; L^2(\Omega))$-norm. We define the system of triangulations $\{\mathcal{T}_{h,m}\}_{m=1,\dots,r}$ and set of polynomial degrees $\boldsymbol{p}$ (consequently, $S_{h,\boldsymbol{p},m}$) such that

$$\|\Psi(t_m) - \Pi_{h,\boldsymbol{p},m}\Psi(t_m)\|_{L^2(\Omega)} \le \omega, \quad m = 1, \dots, r, \tag{6.28}$$

$$N'_m = \dim(S_{h,\boldsymbol{p},m}) \text{ is minimal}, \quad m = 1, \dots, r, \tag{6.29}$$

where Ψ is the exact solution. Clearly, the exact solution is not known, however, it can be approximated from the approximate solution using a high order least-square reconstruction [36].

We refer to [30] for the detailed definitions and construction of anisotropic elements and meshes. Here, we mention the choice of the polynomial approximation degree p on an anisotropic element at node $\bar{\boldsymbol{x}} \in \Omega$ is made. Namely, we first define the density of the number of degrees of freedom at $\bar{\boldsymbol{x}}$ as

$$\eta_{\bar{\boldsymbol{x}},p} := \frac{1}{|K_{\bar{\boldsymbol{x}},p}|}\frac{(p+1)(p+2)}{2},$$

where $K_{\bar{\boldsymbol{x}},p}$ is the theoretical size of a mesh element with barycenter at $\bar{\boldsymbol{x}}$ such that the interpolation error on the interpolant of degree $p+1$ is equal to the constant. This constant is set from the equidistribution principle such that

$$\|\Psi - \Pi_{h,\boldsymbol{p},m}\Psi\|_{L^2(K)} \approx \frac{\omega}{\sqrt{\#\mathcal{T}_{h,m}}} \quad \forall K \in \mathcal{T}_{h,m}.$$

Then, we set the polynomial degree such that the density of the number of degrees of freedom is minimal, i.e., mathematically,

$$p_{\bar{\boldsymbol{x}}} := \arg\min_{p\in\mathbb{N}} \eta_{\bar{\boldsymbol{x}},p}.$$

The optimization technique of the shape and orientation of mesh elements is carried out by a minimization of the interpolation error where the size of the element is kept fixed.

Algorithm 3 Anisotropic *hp*-STDG method

```
Let $\omega > 0$, $\mathcal{T}_{h,1}$, and $\tau_1$ be given.
Set $\bar{t} = 0$.
for $m = 1, \dots$ do
    (a) Perform one time step, i.e., solve (6.17) (or (5.12)) using Algorithm 2
        until the condition (6.25) is fulfilled. Set $\tilde{\Psi}^m_{h\tau}$ using (6.21).
    (b) If the condition (6.26) is violated decrease $\tau_m$ and repeat (a).
    (c) Set optimal value $\tau_m$.
    (d) From $\tilde{\Psi}^m_{h\tau}(t_m^-)$ reconstruct $\tilde{\Psi}_m$ and verify
            $\left\|\tilde{\Psi}(t_m) - \Pi_{h,\boldsymbol{p},m}\tilde{\Psi}(t_m)\right\|_{L^\infty(\Omega)} \in [\omega/2, 2\omega].$    (6.30)
    (e) If (6.30) is violated, then create a new mesh $\mathcal{T}_{h,m}$ with the corresponding
        set $\boldsymbol{p}$ and repeat (a).
    (f) Set
            $\bar{t} := \bar{t} + \tau_m.$
        If $\bar{t} \geq T$, then STOP.
    (g) Set
            $\mathcal{T}_{h,m+1} := \mathcal{T}_{h,m},$
            $\tau_{m+1} := \tau_m.$
```

In practice, the condition $\left\|\tilde{\Psi}(t_m) - \Pi_{h,\boldsymbol{p},m}\tilde{\Psi}(t_m)\right\|_{L^\infty(\Omega)} \approx \omega$ (cf. (6.28)) is unlikely to be fulfilled, therefore, the condition (6.30) is suggested in Algorithm 3. In addition, too small values of the interpolation error are not considered and the re-meshing is performed in this case.

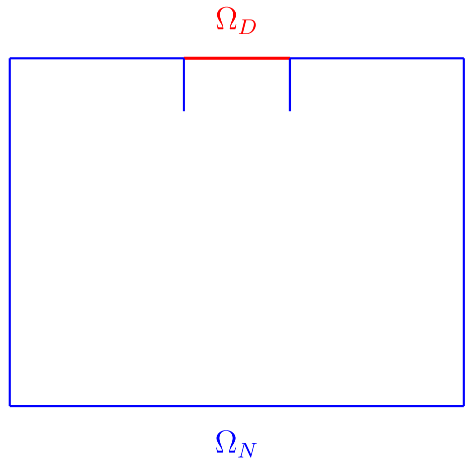

Figure 6.9: Single ring infiltration: geometry of the domain.

6.4 Numerical example

In what follows, we investigate the performance of Algorithm 3 applied on Ψ-formulation and ψ-formulation of Richards' equation. The experiments are run on the CPU architecture Intel(R) Core(TM) i7 using the in-house code ADGFEM [31].

A common application of the flow through a variably saturated medium is a single ring infiltration process. This process consists of the insertion of a solid ring into the soil to a given depth and adding of water inside of the pipe.

We simulate this process on a rectangular domain $\Omega = (0, 1.3) \times (0, 1)$. The geometry of the domain is given in Fig. 6.9; namely, the Dirichlet boundary condition $\Psi_D = 1.05$ is prescribed on the border colored in red mimicking constant flow of the water through the soil, while homogeneous Neumann boundary condition is imposed on the rest of the boundary (colored in blue) meaning that fluid cannot enter through the region. The initial condition is set to be $\Psi_0 = -2$.

We apply Algorithm 3 to Ψ-formulation and ψ-formulation of Richards' equation. Particularly, for the spatial discretization we use the IIPG method (cf. Definition 2); at the initial time step the polynomial degree is set to $p = 2$, while in the rest of the computation varying polynomial degrees are obtained using *hp*-adaptivity technique described in Section 6.3. We use the fixed polynomial degree in time $q = 1$. The initial mesh used at the step (a) of Algorithm 3 is plotted in Fig. 6.10; it is a priori refined around the boundary Ω_D due to the inconsistency between boundary and initial conditions. The final time is set to be 2 hours. We complement Richards' equation with the van Genuchten-Mualem constitutive laws with parameters from Table 1.1.

In Fig. 6.11 and Fig. 6.12 the simulations of the single ring experiment using Ψ-formulation and ψ-formulation are presented, respectively. In particular, we presented the hydraulic head on the left-hand side of Figs. 6.11–6.12 and the polynomial distribution on grids at time levels $t = 0.5$, $t = 1$ and $t = 2$ hours is shown on the right-hand side. The resulting hydraulic head simulation using both formulations seems to be identical; however, we note different polynomial degree distribution and, obviously, different grids generated by Algorithm 3.

Tables 6.3–6.4 (cf. Fig. 6.13) demonstrate the computational performances of

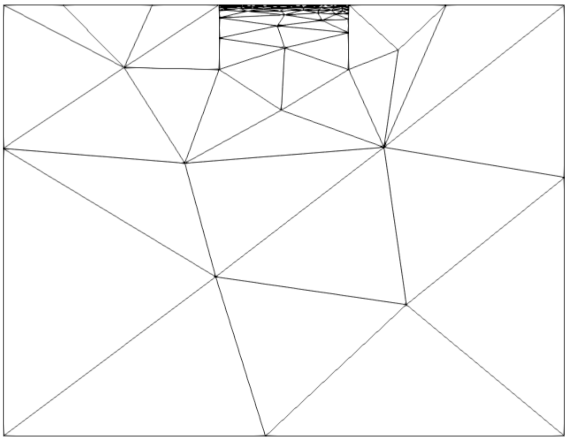

Figure 6.10: Single ring infiltration: the mesh used at $t = 0$.

Algorithm 3 applied to Ψ-formulation and ψ-formulation, respectively, for several tolerances $\omega > 0$ for the interpolation error (6.28). We observe that the adaptive algorithm leads to reduction of the number of degrees of freedom as well as the number of accumulated nonlinear and linear iterations. Moreover, in Figs. 6.14–6.15 we show the values for the interpolation error with respect to the physical time, noting similar pattern in the behavior of the interpolation error in both formulations. We mention that the tolerance for the interpolation error (6.28) is set to be $\omega = 10^{-2}$ in Figs. 6.13–6.15.

We recall the balance of the water content (1.17) introduced in Section 1.5, and the corresponding quantities, the water content at time t denoted by $\Delta Q(t)$ (cf. (1.18)) and the boundary flux on the interval $(0, t)$ denoted by $F(t)$ (cf. (1.20) (or (1.21))). Let us note that in this example the boundary flux through $\partial\Omega$ is equal to the flux through $\partial\Omega_D$. In Fig. 6.16, we plotted the boundary flux (1.20) for Ψ-formulation and the boundary flux for ψ-formulation (1.21) for several tolerances $\omega > 0$. The figures on the left-hand side correspond to Ψ-formulation, while on the right-hand side to ψ-formulation; moreover, the first row corresponds to $F(t)$ on the whole interval $[0, T]$, while the second and third rows correspond to detailed view near to $t = 0$ and $t = T$, respectively. We note that Algorithm 3 with lower tolerances exhibits a small oscillation around $t = 0$, which is less notable in the case of ψ-formulation.

Finally, in Tables 6.5– 6.6 (cf. Fig. 6.17) we studied the influence of using regularization of the constitutive laws mentioned in Section 6.2. We introduce the quantity

$$V(t) = \frac{|\Delta Q(t) - F(t)|}{\max(|\Delta Q(t)|, |F(t)|)},$$

standing for the relative violation of the balance of the water content, which measures the conservativity of the numerical method. Hence, Tables 6.5– 6.6

show the quantities $\Delta Q(t)$, $F(t)$ and $V(t)$, number of time steps, number of accumulated nonlinear iterations and the time consumed for the computations at time $t = t_r = 0.1$ for different values of the regularization parameter $\epsilon > 0$. We observe similar behaviours for both formulations. In Fig. 6.17 we may see the increase in the water content $\Delta Q(t)$ and the boundary flux $F(t)$ with respect to the increase of the regularization parameter $\epsilon > 0$. As before, on the left-hand side are shown the results for Ψ-formulation, while on the right-hand side the ψ-formulation; in the first row are shown mentioned quantities on $[0, t_r]$, in the second and third row are presented details near $t = 0$ and $t = 0.1$, respectively.

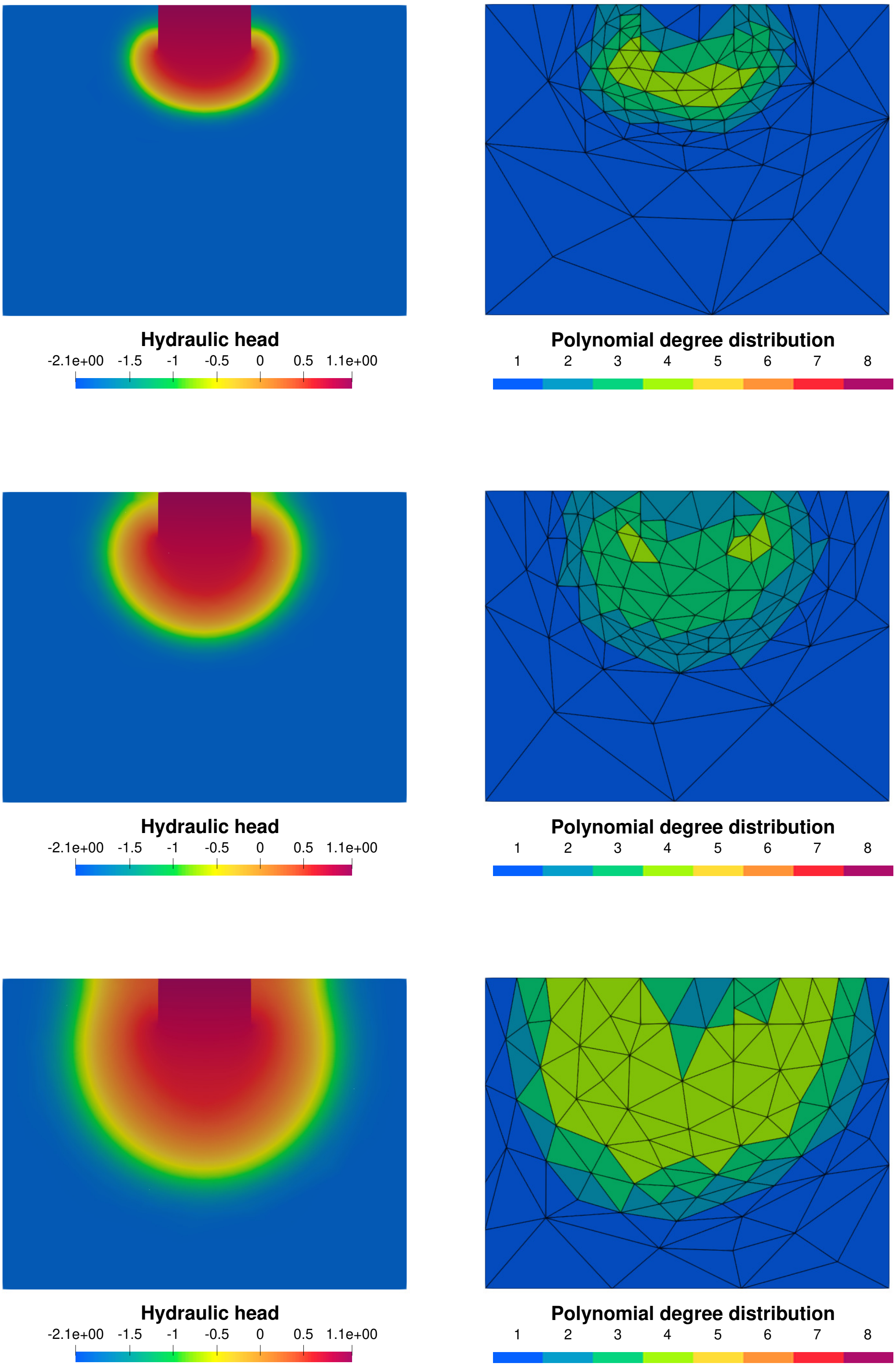

Figure 6.11: Single ring infiltration: the hydraulic head at $t = 0.5$ (1st row), $t = 1$ (2nd row) and $t = 2$ (3rd row) obtained by solving the Ψ-formulation.

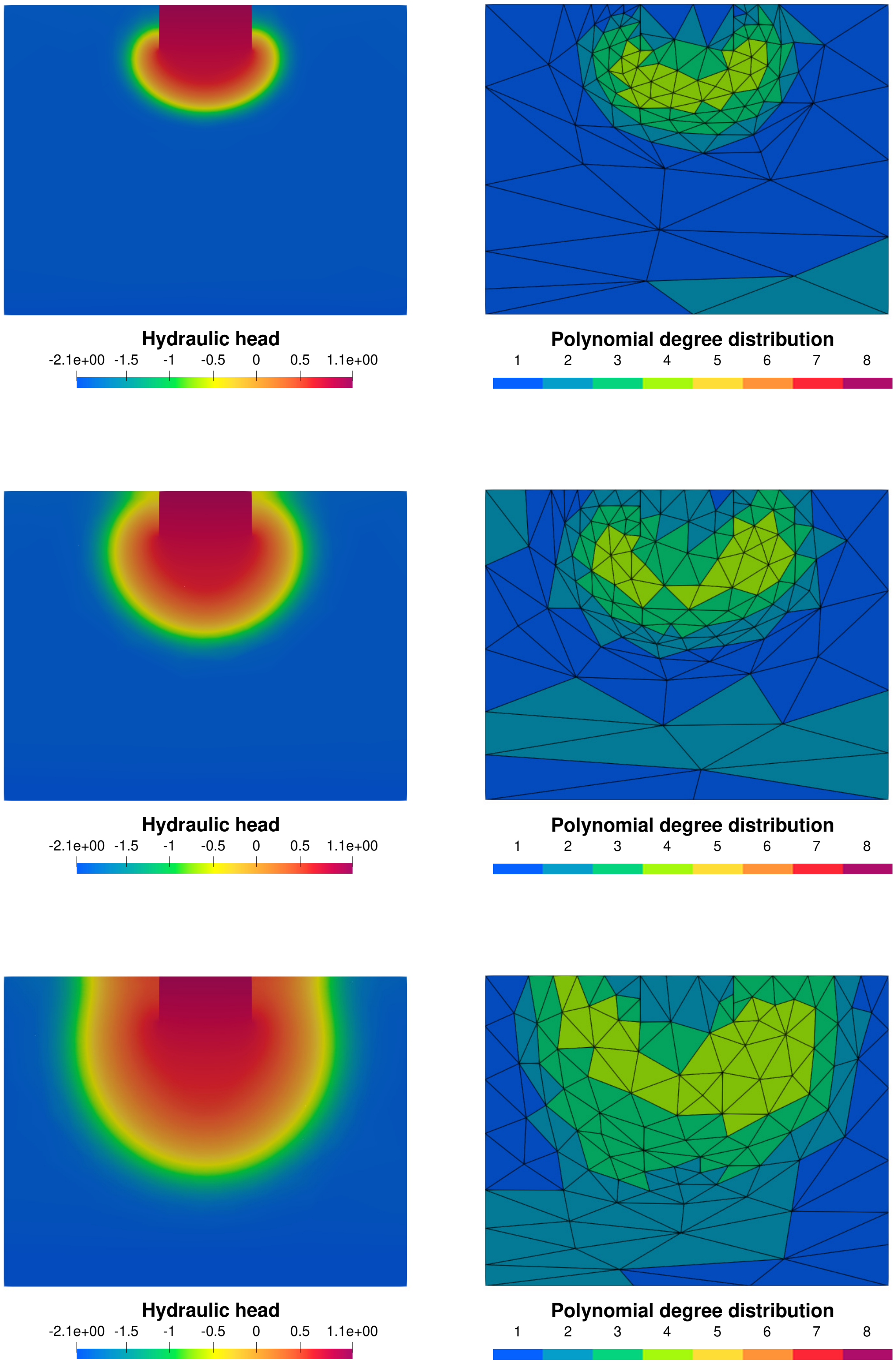

Figure 6.12: Single ring infiltration: the hydraulic head at $t = 0.5$ (1st row), $t = 1$ (2nd row) and $t = 2$ (3rd row) obtained by solving the ψ-formulation.

Table 6.3: Single ring infiltration: Ψ-formulation.

ω	DoF	$\#\Delta t$	#refused Δt	#Newton it.	#GMRES it.	CPU[s]
4E-02	1723	422	10	9 745	1 080 282	2140.90
2E-02	1786	510	21	12 620	1 738 890	2719.13
1E-02	2078	723	46	15 774	1 987 535	4334.95
8E-03	3743	1575	327	21 622	1 980 462	22151.64

Table 6.4: Single ring infiltration: ψ-formulation.

ω	DoF	$\#\Delta t$	#refused Δt	#Newton it.	#GMRES it.	CPU[s]
4E-02	1598	308	16	8 426	927 338	1349.83
2E-02	1761	407	23	12 183	1 370 133	2175.34
1E-02	2181	464	42	12 868	1 641 220	3777.91
8E-03	2966	735	165	12 068	1 495 364	6473.43

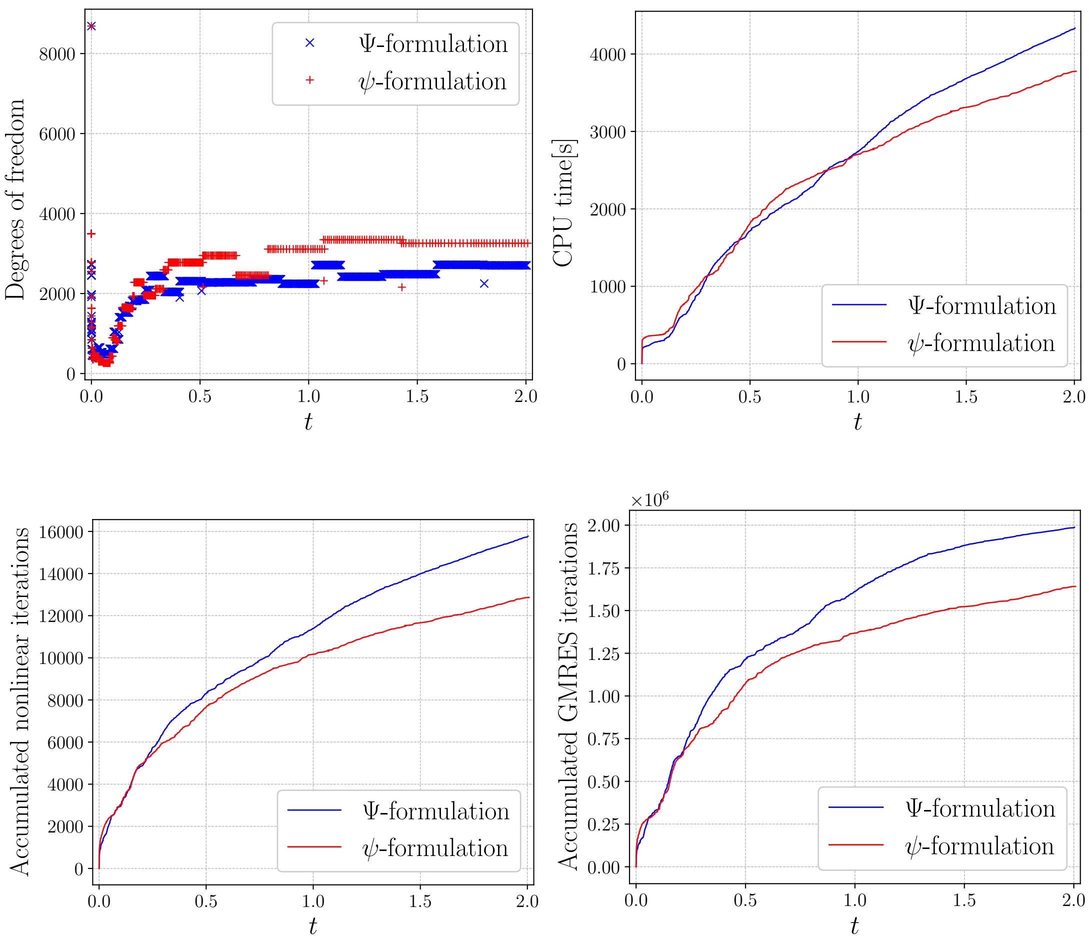

Figure 6.13: Single ring infiltration: comparison of computational performance between the Ψ-formulation and ψ-formulation for $\omega = 10^{-2}$.

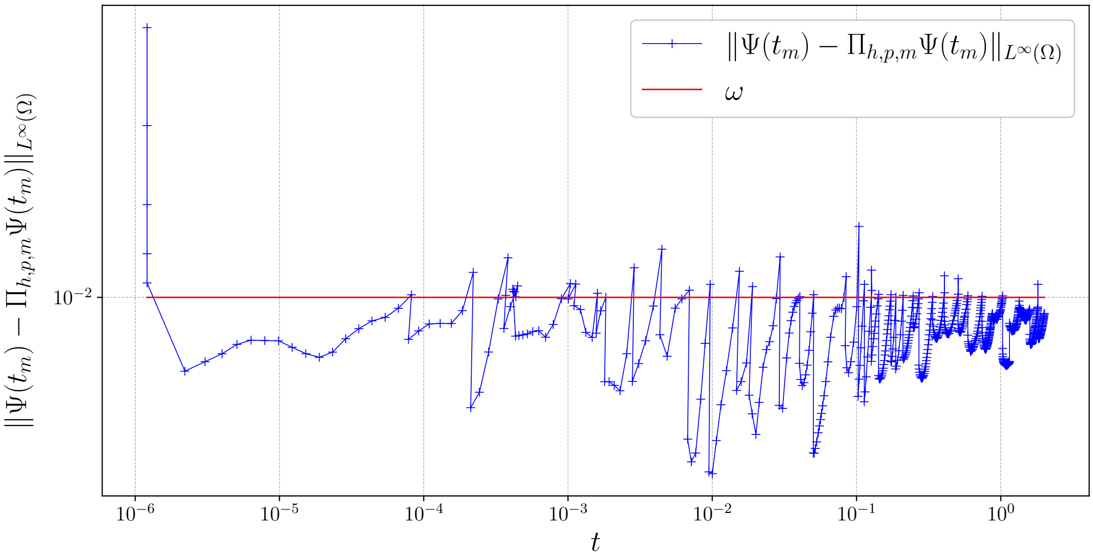

Figure 6.14: Single ring infiltration: the dependence of the error estimates on $t_m \in [0, T]$ for the Ψ-formulation.

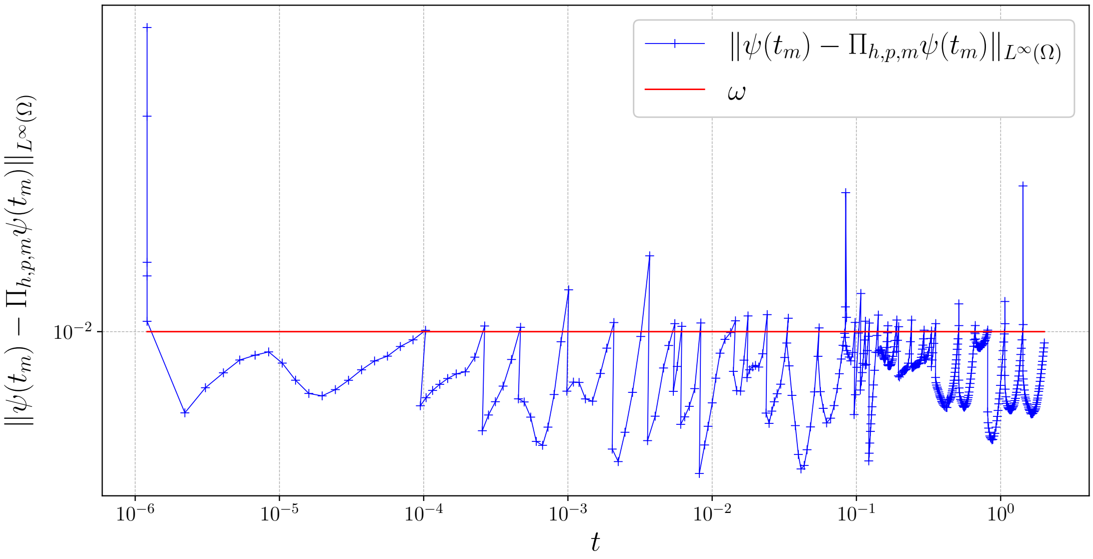

Figure 6.15: Single ring infiltration: the dependence of the error estimates on $t_m \in [0, T]$ for the ψ-formulation.

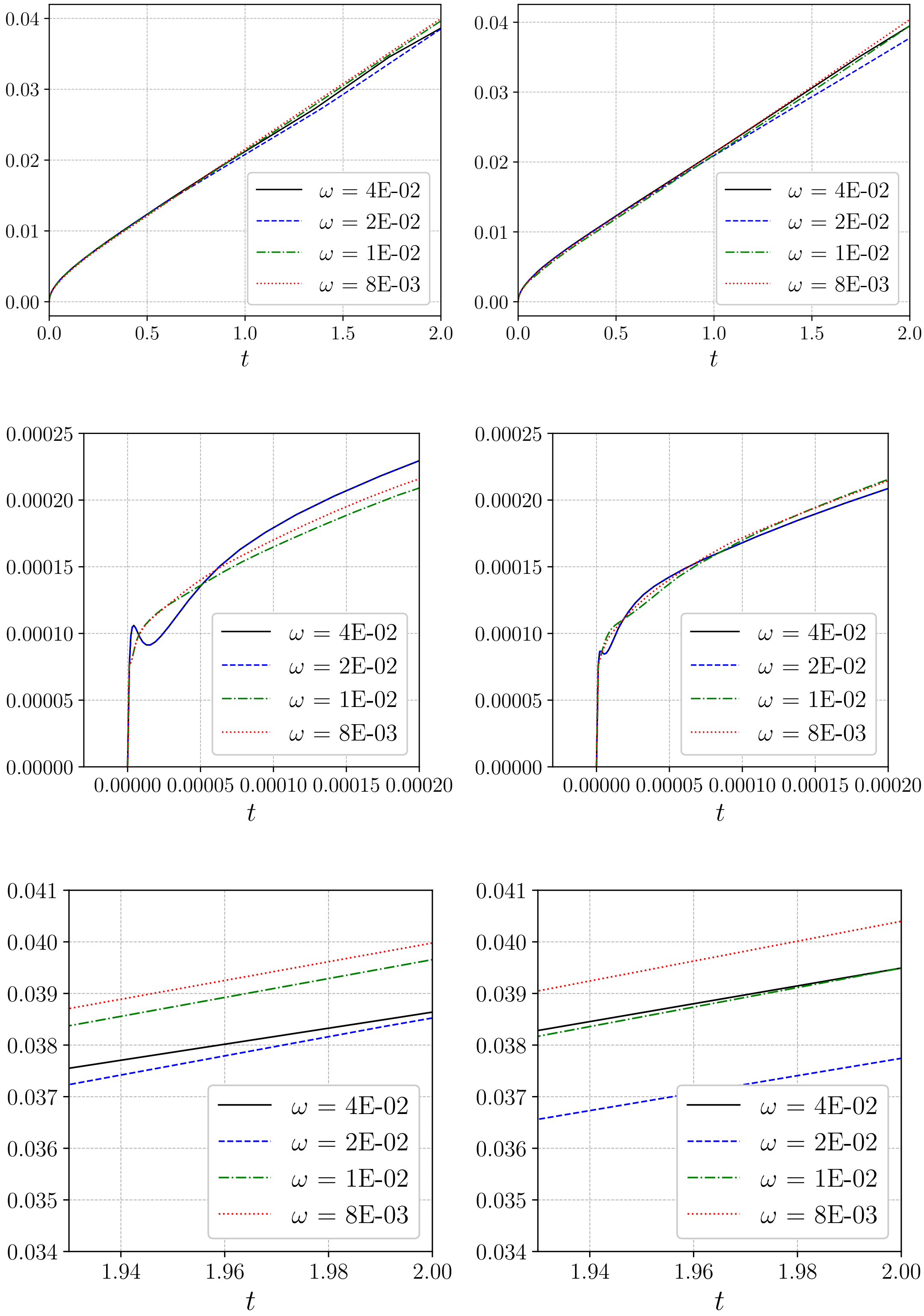

Figure 6.16: Single ring infiltration: the dependence of the actual flux $F(t)$, $t \in (0, T)$ with respect to different tolerances for the Ψ-formulation (left) and the ψ-formulation (right): total view (1st row), the details near $t = 0$ (2nd row) and $t = T$ (3rd row).

Table 6.5: Single ring infiltration: the water content ΔQ, the boundary flux F and the relative violation of the balance of the water content V at the time $t_r = 0.01$ for regularization parameters ϵ for Ψ-formulation.

ϵ	$F(t_r)$	$\Delta Q(t_r)$	$V(t_r)$	$\#\tau_m$	#nonlinear it.	CPU[s]
1E-02	1.23E-03	1.16E-03	0.06	73	1 094	209.50
1E-03	1.20E-03	1.14E-03	0.05	80	1 350	325.61
1E-04	1.18E-03	1.14E-03	0.03	81	1 865	238.78

Table 6.6: Single ring infiltration: the water content ΔQ, the boundary flux F and the relative violation of the balance of the water content V at the time $t_r = 0.01$ for regularization parameters ϵ for ψ-formulation.

ϵ	$F(t_r)$	$\Delta Q(t_r)$	$V(t_r)$	$\#\tau_m$	#nonlinear it.	CPU[s]
1E-02	1.23E-03	1.15E-03	0.06	75	1 259	221.31
1E-03	1.19E-03	1.14E-03	0.04	76	1 548	280.50
1E-04	1.21E-03	1.14E-03	0.06	83	1 369	233.69

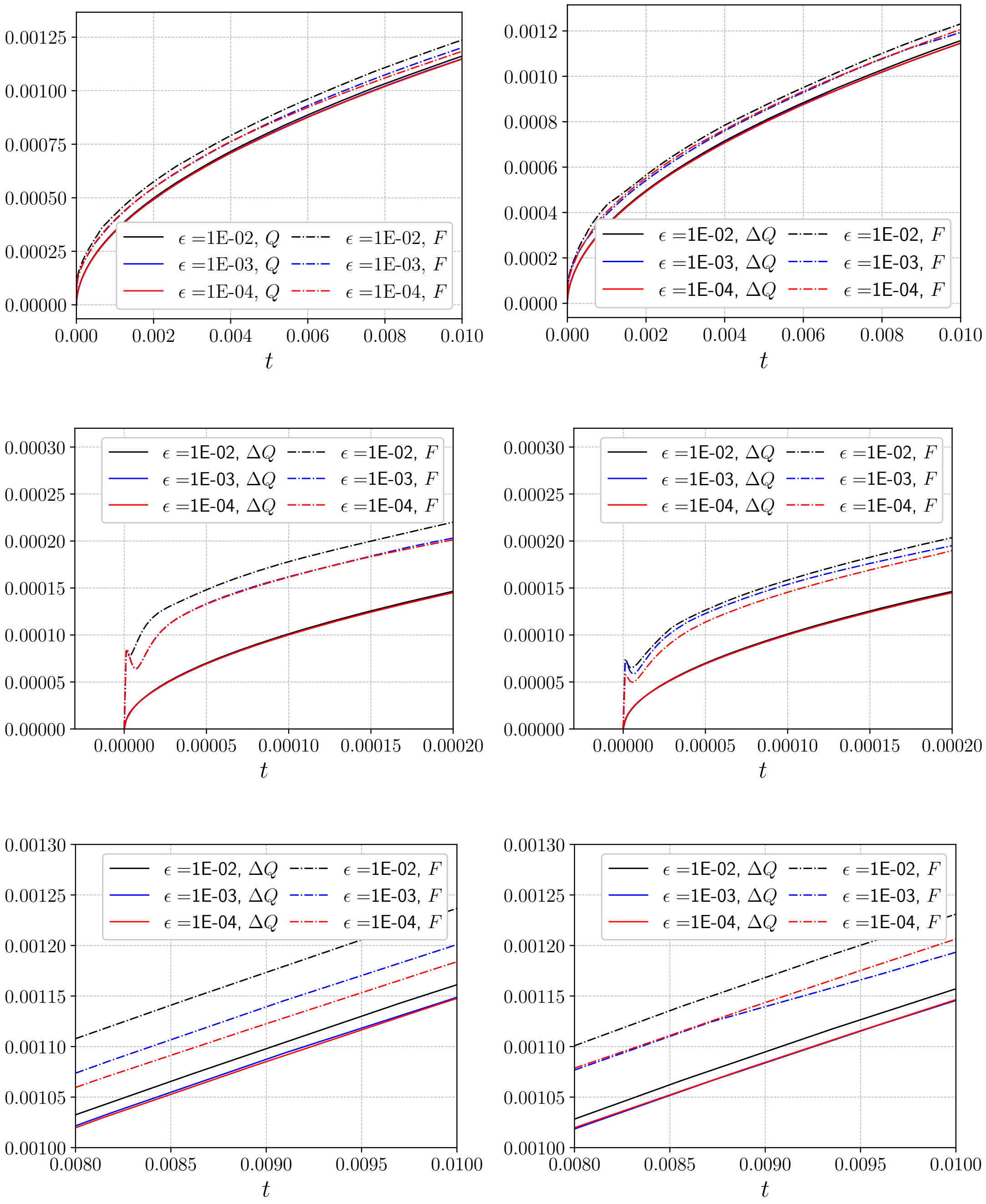

Figure 6.17: Single ring infiltration: the dependence of the actual flux $F(t)$ and water content $\Delta Q(t)$, $t \in (0, T)$, with respect to different regularization parameters ϵ for the Ψ-formulation (left) and the ψ-formulation (right): total view (1st row), the details near $t = 0$ (2nd row) and $t = T$ (3rd row).

Conclusion

In this thesis, we studied theoretically and numerically the DG method applied to a porous media flow model. Chapter 1 introduces the notion of the porous media flow, providing its principle laws and formulating its governing equation — Richards' equation. This equation is a nonlinear degenerate parabolic PDE whose possible degeneracies have been addressed. Moreover, we defined the closure laws and the balance law.

Within Chapter 2, we defined two formulations of Richards' equation: Ψ-formulation and ψ-formulation, whose primary variable is the hydraulic head Ψ and the pressure head ψ, respectively. Afterward, we discretized the spatial variable using the DG method and defined the semidiscrete schemes for both formulations.

Chapters 3–4 are devoted to error analysis of the semidiscrete solution of Richards' equation obtained by the LDG method. Here, we assumed that the active pore volume function ϑ is Lipschitz continuous with $\vartheta' \circ \vartheta^{-1}$ Hölder continuous, which is aligned with the fast-diffusion type of degeneracy of the considered PDE. Due to the presence of the nonlinearities, we considered the expanded mixed formulation of Richards' equation, by virtue of which we defined the LDG method. Moreover, the results on the stability of the semidiscrete solution were obtained. Then, the error analysis is performed in a rather nonstandard way. Namely, the standard approach resulted in an incomplete error estimate, where a bound for the arising nonlinear term was further obtained using Gronwall's lemma implicitly. Finally, these two results are combined using the continuous mathematical induction technique. The final a priori error estimate in L^2-norm and the jump form depends on the spatial parameter h and the Hölder coefficient of the composition $\vartheta' \circ \vartheta^{-1}$ indicating a lower order of convergence comparing with the regular problems. The theoretical results are obtained for the 2D case; however, an extension to the 3D case can be done straightforwardly. In addition, an extension of a fully discrete scheme, such as Euler scheme or STDG scheme, can be derived analogously. Future work may include derivation of optimal error estimates or considering some other assumptions, e.g., the case when ϑ is Hölder continuous or slow diffusion case of degeneracy.

In Chapter 4, we studied the convergence of the numerical method using numerical examples, which showed a higher experimental order of convergence. We suppose that the suboptimality of the theoretical rates is caused by the fact that the low regularity of the solution appears locally in the given example. In addition, we provided some examples outside the scope of the theoretical results. A future extension to this can be done by finding other numerical examples matching the assumption of our analysis.

The rest of the thesis is devoted to a practical application of the porous media flow. First, in Chapter 5, we introduced temporal discretization using the DG method and define the *hp*-STDG method that uses higher-order polynomial approximations including varying polynomial degrees with respect to the spatial variable. Then, we formulated the fully discrete scheme. In Chapter 6, the numerical schemes are interpreted as nonlinear algebraic systems and a Newton-like solver and the Anderson acceleration technique are defined. Moreover, we presented a numerical study of the Newton-like method with and without the Anderson acceleration using a landfill dam simulation. Furthermore, we defined the anisotropic *hp*-STDG method that performs mesh refinement by minimizing degrees of freedom for the prescribed tolerance for the interpolation error. Finally, we consider this method applied to the Ψ-formulation and ψ-formulation of Richards' equation on a single ring infiltration experiment. The computational performance of both formulations has been compared, where we noted slight superiority of the ψ-formulation.

The future work may focus on derivation of a posteriori error estimates. There are several types of a posteriori error estimates, e.g., a posteriori error estimates based on estimation of residual in dual norm; however, their use for practical problems is questionable. It seems to be more relevant to develop the goal oriented error estimates, which can provide practically useful information about the accuracy. Moreover, the research can be focused on adaptation of domain decomposition techniques or parallelization of the algorithm, leading to more efficient computation.

Bibliography

[1] H. W. Alt, S. Luckhaus, Quasilinear elliptic-parabolic differential equations, Math. Z., **183**, 311–341, (1983)

[2] T. Arbogast, An error analysis for Galerkin approximations to an equation of mixed elliptic-parabolic type, Technical report TR90-33, Department of Computational and Applied Mathematics, Rice University, Houston, TX., (1990)

[3] T. Arbogast, M.F. Wheeler, 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**(4), 1669–1687, (1996)

[4] D.N. Arnold, F. Brezzi, B. Cockburn, L.D. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems, SIAM J. Numer. Anal., **39**(5), 1749–1779, (2002)

[5] G.I. Barenblatt, On some unsteady motions of a liquid and gas in a porous medium. Akad. Nauk SSSR. Prikl. Mat. Meh., **16**, 67–78, (1952)

[6] G.I. Barenblatt, T.W. Patzek, D.B. Silin, The mathematical model of non-equilibrium filtration, Journal of Petroleum Science and Engineering, **38**(3-4), 155–169, (2003)

[7] S. Bassetto, C. Cancès, G. Enchéry, Q. H. Tran, On several numerical strategies to solve Richards' equation in heterogeneous media with Finite Volumes, Comput. Geosci., **26**, 1297–1322, (2022)

[8] P. Bastian, A fully-coupled discontinuous Galerkin method for two-phase flow in porous media with discontinuous capillary pressure, Comput. Geosci., **18**(5), 779–796, (2014)

[9] M. Bause, P. Knabner, Computation of variably saturated subsurface flow by adaptive mixed hybrid finite element methods, Adv. Water Resour. **27** (6), 565–581, (2004)

[10] A. Binley, K. Beven, Vadose zone flow model uncertainty as conditioned on geophysical data, Ground Water **41** (2) (2003) 119–127

[11] K. Brenner, Acceleration of Newton's method using nonlinear Jacobi preconditioning, in Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples, R. Klöfkorn, E. Keilegavlen, F. A. Radu, and J. Fuhrmann, eds., Cham, 2020, Springer International Publishing, pp. 395–403

[12] K. Brenner and C. Cancès, Improving Newton's method performance by parametrization: The case of the Richards equation, SIAM J. Numer. Anal., **55**, 1760–1785, (2017)

[13] S.C. Brenner, Poincaré-Friedrichs inequalities for piecewise H-1 functions, SIAM J. Numer. Anal., **41**(1) 306–324, (2003)

[14] S.C. Brenner, R.L. Scott, The Mathematical Theory of Finite Element Methods, Springer, New York, (1994)

[15] E. Buckingham, Studies on the Movement of Soil Moisture, USDA Bureau of Soils – Bulletin, 38, (1907)

[16] C. Cances, I.S. Pop, M. Vohralík, An a posteriori error estimate for vertex-centered finite volume discretizations of immiscible incompressible two-phase flow, Math. Comput. **83** (285), 153–188, (2014)

[17] P. Castillo, A review of the Local Discontinuous Galerkin (LDG) method applied to elliptic problems, Appl. Numer. Math., **56**, 1307–1313, (2006)

[18] V. Casulli and P. Zanolli, A nested Newton-type algorithm for finite volume methods solving Richards' equation in mixed form, SIAM J. Sci. Comput., **32**, pp. 2255–2273, (2010)

[19] M. Celia, E. Bouloutas, R. Zarba, A general mass-conservative numerical-solution for the unsaturated flow equation, Water Resour. Res. **26** (7), 1483–1496, (1990)

[20] Y.R. Chao, A note on 'Continuous mathematical induction', Bull. Amer. Math. Soc., **46**, 17–18, (1919)

[21] J.B. Clément, F. Golay, M. Ersoy, D. Sous, An adaptive strategy for discontinuous Galerkin simulations of Richards' equation: Application to multi-materials dam wetting, Adv. Water Resour., **151**, 661–668, (2021)

[22] B. Cockburn, C.-W. Shu, The local discontinuous Galerkin method for time-dependent convection-diffusion systems, SIAM J. Numer. Anal. **35**, 2440–2463, (1998)

[23] R.L. Cooley, Some new procedures for numerical solution of variably saturated flow problems, Water Resour. Res. **19** (5), 1271–1285, (1983)

[24] S. Congreve, V. Dolejší, S. Sakić, Error analysis for local discontinuous Galerkin semidiscretization of Richards' equation, IMA J. Numer. Anal. (2024)

[25] M.M. Day, The spaces L^p with $0 < p < 1$, Bull. Amer. Math. Soc., **46**, 816–823, (1940)

[26] H. Darcy, Les fontaines publiques de la ville de Dijon, Librairie des corps impériaux des ponts et des chaussées et des mines, (1856)

[27] P. Deuflhard, Newton Methods for Nonlinear Problems, Springer Series in Computational Mathematics, 35, Springer, 2004.

[28] E. DiBenedetto, Degenerate Parabolic Equations, Springer, New York, (1993)

[29] H.-J. G. Diersch and P. Perrochet, On the primary variable switching technique for simulating unsaturated–saturated flows, Adv. Water Resour., **23**, 271–301, (1999)

[30] V. Dolejší, Anisotropic *hp*-adaptive method based on interpolation error estimates in the L^q-norm, Appl. Numer. Math. 82 (2014) 80–114.

[31] V. Dolejší, in: ADGFEM – Adaptive discontinuous Galerkin finite element method, in-house code, 2014. Charles University, Prague, Faculty of Mathematics and Physics, `http://atrey.karlin.mff.cuni.cz/dolejsi/adgfem/`.

[32] V. Dolejší, *hp*-DGFEM for nonlinear convection-diffusion problems, Math. Comput. Simul. **87**, 87–118, (2013)

[33] V. Dolejší, M. Feistauer, Discontinuous Galerkin Method – Analysis and Applications to Compressible Flow, Springer Series in Computational Mathematics 48. Springer, Cham., (2015)

[34] V. Dolejší, M. Kuráž, P. Solin, Adaptive Higher-Order Space-Time Discontinuous Galerkin Method for the Computer Simulation of Variably-Saturated Porous Media Flows, Appl. Math. Model., **72**, 276–305, (2019)

[35] V. Dolejší, F. Roskovec, M. Vlasák, Residual based error estimates for the space-time discontinuous Galerkin method applied to the compressible flows, Comput. Fluids **117**, 304–324, (2015)

[36] V. Dolejší, P. Solin, *hp*-discontinuous Galerkin method based on local higher order reconstruction, Appl. Math. Comput. 279 (2016) 219–235.

[37] C. Ebmeyer, Error estimates for a class of degenerate parabolic equations, SIAM J. Numer. Anal., **35**(3), 1095–1112, (1998)

[38] Y. Epshteyn, B. Rivière, Analysis of *hp* discontinuous Galerkin methods for incompressible two-phase flow, J. Comput. Appl. Math., **225**(2), 487–509, (2009)

[39] R. Eymard, M. Gutnic, D. Hilhorst, The finite volume method for Richards equation, Comput. Geosci., **3**(3–4), 259–294, (1999)

[40] R. Eymard, D. Hilhorst, M. Vohralík, A combined finite volume-nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems, Numer. Math., **105**(1), 73–131, (2006)

[41] M.W. Farthing, F.L. Ogden, Numerical Solution of Richards' Equation: A Review of Advances and Challenges Soil Sci. Soc. Am. J., **81**(6), 1257–1269, (2017)

[42] M. Feistauer, J., Felcman, I., Straškraba, Mathematical and Computational Methods for Compressible Flow. Clarendon Press, Oxford (2003)

[43] P. A. Forsyth, Y. S. Wu, and K. Pruess, Robust numerical methods for saturated-unsaturated flow with dry initial conditions in heterogeneous media, Adv. Water Resour., 18 (1995), pp. 25–38,

[44] W. R. Gardner, Some Steady State Solutions of the Unsaturated Moisture Flow Equation with Application to Evaporation from a Water Table, Soil Sci., **85**, 228-232, (1958)

[45] I. Ginzburg, J.-P. Carlier, C. Kao, Lattice Boltzmann approach to Richards' equation, in: Computational Methods in Water Resources: Volume 1, Elsevier, 2004, pp. 583–595.

[46] T. Gudi, N. Nataraj, A. Pani, An *hp*-local discontinuous Galerkin method for some quasilinear elliptic boundary value problems of nonmonotone type, Math. Comput., **77**, 731–756, (2008)

[47] P. Henning, M. Ohlberger, B. Schweizer, Adaptive heterogeneous multiscale methods for immiscible two-phase flow in porous media, Comput. Geosci., **19**(1), 99–114, (2015)

[48] D. Howard, L.D. Buttery, K.M. Shakesheff, S.J. Roberts, Tissue engineering: strategies, stem cells and scaffolds. Journal of anatomy, **213**(1), 66–72, (2008)

[49] P.S. Huyakorn, S.D. Thomas, B.M. Thompson, Techniques for making finite elements competitive in modeling flow in variably saturated porous media, Water Resour. Res., **20** (8) 1099–1115, (1984)

[50] W. Jäger, J. Kačur, Solution of doubly nonlinear and degenerate parabolic problems by relaxation schemes, RAIRO Modél. Math. Anal. Numér., **29**(5), 605–627, (1995)

[51] W. Kaplan, W. Advanced calculus, Pearson Education India, (1952)

[52] R.A. Klausen, F.A. Radu, G.T. Eigestad, Convergence of MPFA on triangulations and for Richards' equation, Int. J. Numer. Methods Fluids, **58**(12), 1327–1351, (2008)

[53] P. Kordulová, M. Beneš, Solutions to the seepage face model for dual porosity flows with hysteresis, Nonlinear Anal. Theory Methods Appl. **75** (18), 6473–6484, (2012)

[54] V. Kučera, Finite element error estimates for nonlinear convective problems, J. Numer. Math., **24**(3), 143–165, (2016)

[55] M. Kuraz, P. Mayer, V. Havlicek, P. Pech, J. Pavlasek, Dual permeability variably saturated flow and contaminant transport modeling of a nuclear waste repository with capillary barrier protection, Appl. Math. Comput. **219** (13), 7127–7138, (2013)

[56] L. Lam, D.G. Fredlund, Saturated-unsaturated transient finite element seepage model for geotechnical engineering, Adv. Water Resour., **7**(3), 132–136, (1973)

[57] F. Lehmann, P. Ackerer, Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media, Transp. Porous Media **31** (3), 275–292, (1998)

[58] M. Lenzinger, B. Schweizer, Two-phase flow equations with outflow boundary conditions in the hydrophobic–hydrophilic case, Nonlinear Anal. Theory Methods Appl. **73** (4), 840–853, (2010)

[59] H. Li, M.W. Farthing, C.T. Miller, Adaptive local discontinuous Galerkin approximation to Richards' equation, Adv. Water Resour., **30**(9), 1883–1901, (2007)

[60] F. List, F.A. Radu, A study on iterative methods for solving Richards' equation, Comput. Geosci. **20** (2), 341–353, (2016)

[61] P. Lott, H. Walker, C. Woodward, U. Yang, An accelerated Picard method for nonlinear systems related to variably saturated flow, Adv. Water Resour. 38, 92–101, (2012)

[62] E. Magenes, R. Nochetto, C. Verdi, Energy error-estimates for a linear scheme to approximate nonlinear parabolic problems,

[63] G. Manzini, S. Ferraris, Mass-conservative finite volume methods on 2-D unstructured grids for the Richards' equation, Advances in Water Resources **27** (12) (2004) 1199–1215.

[64] C.T. Miller, C. Abhishek, M.W. Farthing, A spatially and temporally adaptive solution of Richards' equation, Adv. Water Resour. **29** (4), 525–545, (2006)

[65] K. Mitra, M. Vohralík, A posteriori error estimates for the Richards equation, Math. Comp. **93**, 347, 1053–1096, (2024)

[66] Y. Mualem, A new model for predicting the hydraulic conductivity of unsaturated porous media, Water Resources Research, **12**(3), 513–522, (1976)

[67] R.H. Nochetto, C. Verdi, Approximation of Degenerate Parabolic Problems Using a Numerical Integration, SIAM J. Numer. Anal., **25**(4), 784–814, (1988)

[68] M. Ohlberger, Convergence of a mixed finite elements-finite volume method for the two phase flow in porous media, East-West J. Numer. Math., **5**(3), 183–210, (1997)

[69] F. Otto, L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations, J. Differential Equations, **131**(1), 20–38, (1996)

[70] F. Otto, L^1–contraction and uniqueness for unstationary saturated-unsaturated porous media flow, Adv. Math. Sci. Appl., **7**(2), 537–553, (1997)

[71] C. Paniconi, M. Putti, A comparison of Picard and Newton iteration in the numerical solution of multidimensional variably saturated flow problems, Water Resources Research **30** (12), 3357–3374,(1994)

[72] I.S. Pop, Error estimates for a time discretization method for the Richards' equation, Comput. Geosci., **5**(2), 141–160, (2002)

[73] I. Pop, B. Schweizer, Regularization schemes for degenerate Richards equations and outflow conditions, Math. Models Methods Appl. Sci. **21** (8), 1685–1712, (2011)

[74] I. Pop, F. Radu, P. Knabner, Mixed finite elements for the Richards' equation: linearization procedure, J. Comput. Appl. Math. **168** (1-2, SI) 365–373, (2004)

[75] F.A. Radu, K. Kumar, J.M. Nordbotten, I. Pop, Robust, mass conservative scheme for two-phase flow in porous media including Holder continuous nonlinearities, IMA J. Numer. Anal **38**, 884–920, (2018)

[76] F. Radu, I. Pop, P. Knabner, On the convergence of the Newton method for the mixed finite element discretization of a class of degenerate parabolic equation, in: A.C. Bermudez, D. Gómez, P. Quintela, P. Salgado (Eds.), Numerical Mathematics and Advanced Applications, Springer, 2006, pp. 1194–1200.

[77] F.A. Radu, I.S. Pop, I.S., P. Knabner, Error estimates for a mixed finite element discretization of some degenerate parabolic equations, Numer. Math., **109**(2), 285–311, (2008)

[78] L.A. Richards, Capillary conduction of liquids through porous mediums, J. Appl. Phys., **1**(5), 318–333, (1931)

[79] B. Rivière, Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation, Frontiers in applied mathematics. SIAM., (2008)

[80] B. Rivière, M.F. Wheeler, A Discontinuous Galerkin Method Applied to Nonlinear Parabolic Equations, In: B. Cockburn, G.E. Karniadakis, CW. Shu (eds) Discontinuous Galerkin Methods. Lecture Notes in Computational Science and Engineering, vol 11. Springer, Berlin, Heidelberg. (2000)

[81] Y. Saad, M.H. Schultz, GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems, SIAM J. Sci. Stat. Comput. **7**, 856–869 (1986)

[82] S. Sakić, S. Congreve, Numerical Study of a Discontinuous Galerkin Method for a Degenerate Parabolic Equation, Numerical Mathematics and Advanced Applications ENUMATH 2023 (submitted)

[83] C. Scudeler, C. Paniconi, D. Pasetto, M. Putti, Examination of the seepage face boundary condition in subsurface and coupled surface/subsurface hydrological models, Water Resources Research **53** (3), 1799–1819, (2017)

[84] B. Schweizer, Regularization of outflow problems in unsaturated porous media with dry regions, J. Differ. Equ. **237** (2), 278–306 (2007)

[85] J. Šimůnek, M. Šejna, H. Saito, M.T. van Genuchten, in: The HYDRUS-1D software package for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media, version 4.08, Department Of Environmental Sciences University Of California Riverside, Riverside, CA, USA., 2009.

[86] M. Slodicka, A robust and efficient linearization scheme for doubly nonlinear and degenerate parabolic problems arising in flow in porous media, SIAM J. Sci. Comput., **23**, 1593–1614, (2002)

[87] P. Solin, M. Kuraz, Solving the nonstationary Richards equation with adaptive *hp*-FEM, Adv. Water Resour., **34**, 1062–1081, (2011)

[88] M.A. Sophocleous, Interactions between groundwater and surface water: The state of the science, Hydrogeology Journal, **10**(1), 52–67, (1998)

[89] A. Szymkiewicz, Modelling Water Flow in Unsaturated Porous Media, Springer Berlin Heidelberg, (2013)

[90] M.D. Tocci, C. Kelley, C.T. Miller, Accurate and economical solution of the pressure-head form of Richards' equation by the method of lines, Adv. Water Resour. **20** (1) 1–14, (1997)

[91] F.T. Tracy, Clean two- and three-dimensional analytical solutions of Richards equation for testing numerical solvers, Water Resour. Re., **42**(8), 1062–1081, (2006)

[92] M.T. van Genuchten, Closed-form equation for predicting the hydraulic conductivity of unsaturated soils, Soil Science Society of America Journal, **44** (5): 892–898, (1980)

[93] J.L. Vázquez, The porous medium equation: mathematical theory. Oxford University Press, (2007)

[94] V. Vilarrasa, D. Bolster, S. Olivella, J. Carrera, Coupled hydromechanical modeling of CO_2 sequestration in deep saline aquifers, Int. Greenh. Gas Control, **4**, 910–919, (2010)

[95] H.F. Walker, P. Ni, Anderson acceleration for fixed-point iterations, SIAM J. Numer. Anal. **49** (4), 1715–1735, (2011)

[96] C.S. Woodward, C.N. Dawson, Analysis of expanded mixed finite element methods for a nonlinear parabolic equation modeling flow into variably saturated porous media, SIAM J. Numer. Anal., **37**(3), 701–724, (2000)

[97] S. Würzer, N. Wever, R. Juras, M. Lehning, T. Jonas, Modelling liquid water transport in snow under rain-on-snow conditions — considering preferential flow, Hydrol. Earth Syst. Sci. **21** (3), 1741–1756, (2017)

[98] Y. Xiao, E.J. Kubatko, C.J. Conroy, A one-dimensional local discontinuous Galerkin Richards' equation solution with dual-time stepping, Comput Geosci **26**, 171–194, (2022)

[99] I. Yotov, Mixed finite element discretization on non-matching multiblock grids for a degenerate parabolic equation arising in porous media flow, East-West J. Numer. Math., **55**(4), 1760–1785, (1997)

[100] Y. Zha, J. Yang, J. Zeng, C.-H. M. Tso, W. Zeng, L. Shi, Review of numerical solution of Richardson–Richards equation for variably saturated flow in soils, Wiley Interdisciplinary Reviews: Water 6 (5) (2019): e1364.

[101] The Karlin cluster. https://cluster.karlin.mff.cuni.cz/snehurka/hardware/

List of Figures

List of Tables

List of Abbreviations

CPU	**C**entral **P**rocessing **U**nit
DG	**D**iscontinuous **G**alerkin
EOC	**E**xperimental **O**rder of **C**onvergence
GMRES	**G**eneralized **M**inimal **RES**idual
hp-**STDG**	higher-order adaptive **S**pace-**T**ime **D**iscontinuous **G**alerkin
IIPG	**I**ncomplete **I**nterior **P**enalty **G**alerkin
ILU	**I**ncomplete **L**ower-**U**pper (factorization)
LDG	**L**ocal **D**iscontinuous **G**alerkin
NIPG	**N**onsymmetric **I**nterior **P**enalty **G**alerkin
PDE	**P**artial **D**ifferential **E**quation
SIPG	**S**ymmetric **I**nterior **P**enalty **G**alerkin
STDG	**S**pace-**T**ime **D**iscontinuous **G**alerkin
1D	One **D**imension
2D	Two **D**imensions
3D	Three **D**imensions

List of publications

Journals

- S. Congreve, V. Dolejší, S. Sakić, Error analysis for local discontinuous Galerkin semidiscretization of Richards' equation. IMA Journal of Numerical Analysis (2024), published on-line DOI: `https://doi.org/10.1093/imanum/drae013`

- A. Araújo, S. Barbeiro, R. Bernardes, M. Morgado, S. Sakić, A mathematical model for the corneal transparency problem. Journal of Mathematics in Industry **12**(1), 10, (2022), DOI: `https://doi.org/10.1186/s13362-022-00125-y`

Conference proceedings

- S. Sakić, S. Congreve, Numerical Study of a Discontinuous Galerkin Method for a Degenerate Parabolic Equation. Numerical Mathematics and Advanced Applications ENUMATH 2023 (submitted).

- A. Araújo, S. Sakić, Stability and convergence of a class of RKDG methods for Maxwell's equation. In: Ehrhardt M, Günther M. Progress in industrial mathematics at ECMI 2021, vol. 39, pp. 443–499, Cham: Springer; 2022, DOI: `https://doi.org/10.1007/978-3-031-11818-0_64`