A posteriori error estimates for the Richards equation

Koondanibha Mitra, Martin Vohralík

A posteriori error estimates for the Richards equation

K. Mitra[1] and M. Vohralík[2,3]

[1]Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
[2]Inria, 2 rue Simone Iff, 75589 Paris, France
[3]CERMICS, Ecole des Ponts, 77455 Marne-la-Vallée, France

December 6, 2022

Abstract

The Richards equation is commonly used to model the flow of water and air through soil, and it serves as a gateway equation for multiphase flows through porous media. It is a nonlinear advection–reaction–diffusion equation that exhibits both parabolic–hyperbolic and parabolic–elliptic kind of degeneracies. In this study, we provide reliable, fully computable, and locally space–time efficient a posteriori error bounds for numerical approximations of the fully degenerate Richards equation. For showing global reliability, a nonlocal-in-time error estimate is derived individually for the time-integrated $H^1(H^{-1})$, $L^2(L^2)$, and the $L^2(H^1)$ errors. A maximum principle and a degeneracy estimator are employed for the last one. Global and local space–time efficiency error bounds are then obtained in a standard $H^1(H^{-1}) \cap L^2(H^1)$ norm. The reliability and efficiency norms employed coincide when there is no nonlinearity. Moreover, error contributors such as flux nonconformity, time discretization, quadrature, linearization, and data oscillation are identified and separated. The estimates are also valid in a setting where iterative linearization with inexact solvers is considered. Numerical tests are conducted for nondegenerate and degenerate cases having exact solutions, as well as for a realistic case and a benchmark case. It is shown that the estimators correctly identify the errors up to a factor of the order of unity.

Keywords— Richards equation, a-posteriori error estimates, nonlinear degenerate problems, flow through porous media, finite element method

1 Introduction

The Richards equation models flow of water through porous medium (e.g., soil) partially filled with air [6, 20]. For a domain $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, and final time $T > 0$, with water saturation s and pressure p being the primary unknowns, it equates

$$\partial_t s - \nabla \cdot [\bar{\mathbf{K}}(\boldsymbol{x})\, \kappa(s)\, (\nabla p + \boldsymbol{g})] = f(s, \boldsymbol{x}, t) \text{ in } \Omega \times [0, T]. \tag{1.1a}$$

Here, space and time variables are denoted by $\boldsymbol{x}$ and t, respectively. The source term $f(s, \boldsymbol{x}, t)$ represents contribution due to reaction/absorption. The gravity is represented by the constant vector $-\boldsymbol{g}$. The absolute permeability tensor $\bar{\mathbf{K}}(\boldsymbol{x})$ and the relative permeability function $\kappa : [0, 1] \to [0, 1]$ are properties of the medium. Initial condition is provided for the saturation s, and homogeneous Dirichtlet boundary condition is provided for the pressure p, i.e.,

$$s(\boldsymbol{x}, 0) = s_0(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \Omega \text{ and } p = 0 \text{ on } \partial\Omega \times (0, T]. \tag{1.1b}$$

This project has received funding by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 647134)

Dirichlet–Neumann mixed boundary conditions are also considered in the numerical Section 6. To close (1.1a)–(1.1b), it is usually assumed that saturation and pressure are related algebraically (commonly referred to as the capillary pressure relationship [20]), i.e., for a function $S : \mathbb{R} \to [0, 1]$ one has

$$s = S(p). \tag{1.1c}$$

Here, we assume that the saturation s is bounded in the closed interval $[0, 1]$. Equation (1.1a) is obtained by combining the constitutive relation for the flux, stated by the Darcy law

$$\boldsymbol{\sigma} := -\bar{\mathbf{K}}(\boldsymbol{x})\kappa(s)(\nabla p + \boldsymbol{g}),$$

with the mass balance equation $\partial_t s + \nabla \cdot \boldsymbol{\sigma} = f(s, \boldsymbol{x}, t)$. The Richards equation is important in modelling groundwater flow and various chemical and biological processes. It is a nonlinear advection–reaction–diffusion equation which degenerates into an elliptic equation if $S'(p) = 0$ at some point of the domain. On the other hand, if $\kappa(s) = 0$, then the equation becomes a first order ordinary differential equation (hyperbolic) with the loss of regularity of the solution. Nonlinearity and degeneracy are the two main challenges in analysing the system (1.1).

Existence of solutions for the Richards equation was shown in [2, 3]. However, in the degenerate case when $\kappa(s) = 0$, only the existence of a weak limit can be shown [3]. Consequently, the pair (s, p) might not satisfy (1.1) in a weak sense. We give appropriate details in Section 2.4. Uniqueness of solutions is proven in [33] using the L^1-contraction method.

Different spatial discretization methods have been designed for the Richards equation. Some notable examples are [19] for finite volumes, [31] for finite elements, [4, 36] for mixed finite elements, [27] for the discontinuous Galerkin method, and [22] for multi-point flux approximations. Iterative linearization methods such as the Newton, Picard, Jäger–Kačur, and the L-schemes have been investigated in [7], [11], [21], and [28, 29], respectively, see also the references therein. An improvement of the Newton method was proposed in [9] by parametrizing both the saturation and the pressure as functions of a separate primary variable. A comprehensive review of numerical methods for the Richards equation can be found in [42].

The theory of a posteriori estimates for elliptic differential equations is well studied, see, e.g. [1, 37, 41]. A posteriori upper error bounds for the heat equation in the $L^2(H^1) \cap L^\infty(L^2)$ norm were derived in [35]. In [40], global efficiency in space on every time step together with reliability are proven for the $L^2(H^1) \cap L^\infty(L^2) \cap H^1(H^{-1})$ norm. In [17], a local efficiency estimate in space and in time is established for the norm further enriched by time jumps. A general framework for obtaining rigorous a posteriori estimates for nonlinear problems has been laid out in [38, 39]. However, the Lipschitz continuity and invertibility of the operators associated with the differential equations are assumed, which limits the scope of the estimates. A more specific result for the p-Laplacian problem is given in [24]. Using a formulation relying on the N-functions, the coercivity and Lipschitz-continuity of the flux function are shown with respect to the gradient. This makes it possible to derive a posteriori estimates for the problem. Estimators for nonlinear advection–diffusion equations were proposed in [14]. Both upper and lower bounds (reliability and efficiency) were established, robust with respect to the nonlinearities and advection dominance, but for a weaker space–time mesh-dependent norm. Moreover, it was also assumed that the solutions belong to $H^1(L^2)$, which may not be the case for degenerate problems and/or if the initial condition is discontinuous. Using entropy methods, error estimates in the L^1-norm were derived in [32] for singularly perturbed nonlinear advection–diffusion problems. Degenerate parabolic equations were considered in [30]. An $L^\infty(H^{-1})$ estimate was derived using dual equations of the diffusion problem. For problems having parabolic–hyperbolic degeneracy, a posteriori upper bounds on the $L^2(H^{-1}) \cap L^\infty(H^{-1})$ norm combined with the time-integrated $L^2(L^2)$ norm of error were derived using Green's function in [13] for a Stefan problem and on

the $L^2(H^{-1}) \cap L^2(H^1) \cap L^2(L^2)$ norm in [10] for two-phase flow through porous media. For the Richards equation, a posteriori error upper bounds in the $L^2(H^1) \cap H^1(L^2)$ norm were derived in [8]. A regularization term was introduced to avoid degeneracy and to obtain $H^1(L^2)$ estimates.

In the present paper, we provide a posteriori error estimates for the Richards equation (1.1a). The main improvements in this study are: (a) *Rigorous derivation of the upper as well as lower bounds of error* by the equivalence of the dual norm of the residual with an error metric that reduces to the $L^2(H^1) \cap H^1(H^{-1}) \cap L^\infty(L^2)$ norm in the linear case. (b) Equivalence of the dual norm of the residual with *fully computable* and *locally space–time efficient estimates.* (c) *No higher-order regularity assumptions* such as the pressure in $L^2(H^2) \cap H^1(L^2)$ or that the initial condition is in H^1. (d) *Inclusion of both the parabolic–hyperbolic and the parabolic–elliptic type of degeneracies.* This requires relaxing the assumptions on the associated functions such as $S'(p)$, $\kappa(s) > 0$, assumed for instance in [5, 8, 10] in order to avoid the blow-up due to degeneracy. It poses a challenge particularly since the parabolic–hyperbolic degeneracy, stemming from $\kappa(s) = 0$, causes a loss of regularity of the solutions. To circumvent this issue, we assume instead that the initial saturation s_0 is bounded away from the degenerate value at 0. With this assumption, a function $S_{\mathrm{m}} : [0, T] \to (0, 1]$ is computed using maximum principle such that $S_{\mathrm{m}}(t) \leq s(\boldsymbol{x}, t) \leq 1$ for all $(\boldsymbol{x}, t) \in \Omega \times [0, T]$. For the parabolic–elliptic degeneracy, a degeneracy estimator is introduced to provide an upper bound on the $L^2(H^1)$ norm of the error. (e) *Rigorous inclusion* of *linearization errors due to inexact solvers*, space and time *adaptive meshes*, and implementation of *adaptive linearization.* (f) It is shown numerically that despite nonlinearities, degeneracies, and heterogeneities, the effectivity index of the estimators lies between 1 and 3 in most cases, even locally.

The paper is organized as follows. Section 2 serves as a mathematical prologue to the Richards equation. The associated functions, relevant transformations, well-posedness results, and maximum principles are discussed in detail. In Section 3, lower and upper bounds on error by the dual norm of the residual are derived. The upper bound is provided separately for the $H^1(H^{-1})$, $L^2(L^2)$, and the $L^2(H^1)$ errors in a time-smoothened fashion, see Theorem 3.4. In Section 4, a finite element approximation to the Richards problem (1.1) is considered, and some time-interpolations are discussed. These are used in Section 5 to compute the equilibrated flux and the a posteriori estimators. Reliability and local space–time efficiency bounds are proven for the estimators. Finally, numerical results are presented in Section 6. The theoretical findings are verified and the corresponding effectivity indices are obtained using a nondegenerate as well as a degenerate case with known exact solutions. To demonstrate the prowess of the estimators, a realistic degenerate problem is analyzed in a heterogeneous, anisotropic domain, with discontinuous initial condition and mixed boundary condition, along with a benchmark case. To conclude, it is shown in Appendix A how to take into account the additional errors from iterative linearization, whereas Appendix B collects some technical proofs.

2 The Richards equation

Here, we give a brief introduction to the Richards equation and state some of its properties important for our analysis.

2.1 Basic notation

Spaces: Let $\Omega \subset \mathbb{R}^d$ be an open polytope with a Lipschitz-continuous boundary. Let $(\cdot, \cdot)$ and $\|\cdot\|$ represent respectively the $L^2(\Omega)$ inner product and norm; $(\cdot, \cdot)_\omega$ and $\|\cdot\|_\omega$ stand for the L^2-inner product and norm with respect to any Lipschitz subdomain $\omega \subset \Omega$. The Sobolev space $H^1(\Omega)$ contains all functions $u \in L^2(\Omega)$ such that the weak derivative $\nabla u \in \boldsymbol{L}^2(\Omega; \mathbb{R}^d)$, and

$H_0^1(\Omega)$ is the subspace of $H^1(\Omega)$ containing functions vanishing at the boundary $\partial\Omega$ in the trace sense. The space $H^{-1}(\Omega)$ stands for the dual of $H_0^1(\Omega)$, and $\langle\cdot,\cdot\rangle$ denotes the corresponding duality pairing. With final time $T > 0$ and $L^2(0,T;V)$ denoting the L^2 Bochner space for a Banach space V, we introduce the Hilbert spaces

$$\mathcal{X} := L^2(0,T;H_0^1(\Omega)) \text{ and } \mathcal{Y} := \{u \in L^2(0,T;H^1(\Omega)) : \partial_t u \in L^2(0,T;H^{-1}(\Omega))\}. \tag{2.1}$$

Inequalities: For a Lipschitz subdomain $\omega \subseteq \Omega$ with diameter h_ω, let $u \in H^1(\omega)$ be such that either $\int_\omega u = 0$ or the trace of u is zero on a section of $\partial\omega$ of nonzero measure. Then the Poincaré–Friedrichs inequality states that there exists a constant $C_{\mathrm{P},\omega} > 0$ such that

$$\|u\|_\omega \le C_{\mathrm{P},\omega} h_\omega \, \|\nabla u\|_\omega. \tag{2.2}$$

For a convex ω in the zero mean-value case, $C_{\mathrm{P},\omega}$ can be taken as π^{-1}.

Notation: Let $[\cdot]_+ = \max(\cdot,0)$ and $[\cdot]_- = \min(\cdot,0)$ denote the positive and negative part functions respectively. In our notation, $a \lesssim b$ will refer to the inequality $a \le Cb$, where $C > 0$ is a constant that depends solely on the shape-regularity of the spatial meshes in the space dimension d, and on the ratio $K_\mathrm{m}/K_\mathrm{M}$ (see (P3) below). In particular, it is independent of mesh-size, time-step size, the functions $\kappa(\cdot)$, $S(\cdot)$, f, and the polynomial degrees associated with the numerical scheme.

2.2 Assumptions on the data

We assume the following properties for the data in (1.1):

(P1) The relative permeability function κ is of the class $C^1([0,1])$ with $\kappa(0) \ge 0$, $\kappa(1) = 1$, and $\kappa(0) < \kappa(s) < \kappa(1)$ for all $s \in (0,1)$.

(P2) The saturation function S is of the class $\mathrm{Lip}(\mathbb{R})$ with Lipschitz constant $L_S > 0$. It is either linear, or there exists a constant $p_\mathrm{M} \in (0,\infty]$ such that $\lim_{p \searrow -\infty} S(p) = 0$, and

  (a) $S|_{(-\infty,p_\mathrm{M}]} \in C^2((-\infty,p_\mathrm{M}])$, $S'(p) > 0$ for all $p < p_\mathrm{M}$, and $\lim_{p \nearrow p_\mathrm{M}} S'(p) > 0$;

  (b) $S(p) = 1$ and consequently $S'(p) = 0$ for all $p > p_\mathrm{M}$.

(P3) The absolute permeability tensor $\bar{\mathbf{K}} : \Omega \mapsto \mathbb{R}^{d\times d}$ is piecewise constant in Ω, bounded, and satisfies the ellipticity condition, i.e., there exists positive constants $K_\mathrm{m}, K_\mathrm{M}$ such that for any $\boldsymbol{\zeta} \in \mathbb{R}^d$,

$$K_\mathrm{m}|\boldsymbol{\zeta}|^2 \le \boldsymbol{\zeta}^\mathrm{T}\bar{\mathbf{K}}(\boldsymbol{x})\boldsymbol{\zeta} \le K_\mathrm{M}|\boldsymbol{\zeta}|^2 \quad \text{for almost all } \boldsymbol{x} \in \Omega,$$

  where $|\boldsymbol{\zeta}|$ is the Euclidean norm of $\boldsymbol{\zeta}$, i.e., $|\boldsymbol{\zeta}| = (\sum_{j=1}^d \zeta_j^2)^{\frac{1}{2}}$. Consequently, there exist unique positive-definite tensor-valued functions $\bar{\mathbf{K}}^{\frac{1}{2}}$, $\bar{\mathbf{K}}^{-\frac{1}{2}}$, and $\bar{\mathbf{K}}^{-1}$.

(P4) The source term $f \in C^1([0,1]\times\Omega\times\mathbb{R})$ and there exists a function $f_\mathrm{m} \in C^1([0,1])$ such that $f_\mathrm{m}(\cdot) \le \inf_{\boldsymbol{x}\in\Omega, t\in\mathbb{R}^+} f(\cdot,\boldsymbol{x},t)$.

(P5) The initial condition $s_0 \in L^\infty(\Omega)$ satisfies

$$0 < \operatorname{ess\,inf}_{\boldsymbol{x}\in\Omega}\{s_0(\boldsymbol{x})\} \le \operatorname{ess\,sup}_{\boldsymbol{x}\in\Omega}\{s_0(\boldsymbol{x})\} \le 1.$$

These assumptions are consistent with experiments, see e.g. [20].

Remark 2.1 (Choices for the functions κ and S)**.** *Two most commonly used models for the functions $\kappa(\cdot)$ and $S(\cdot)$ [26] are the Brooks–Corey model,*

$$\kappa(s) = s^{\frac{2+3\lambda_1}{\lambda_1}}, \quad S(p) = (2 - p/p_{\mathrm{M}})^{-\lambda_1} \text{ for } p \le p_{\mathrm{M}}, \tag{2.3}$$

and the van Genuchten model,

$$\kappa(s) = \sqrt{s}\,(1 - (1 - s^{1/\lambda_2})^{\lambda_2})^2, \quad S(p) = 1/(1 + (p_{\mathrm{M}} - p)^{\frac{1}{1-\lambda_2}})^{\lambda_2} \text{ for } p \le p_{\mathrm{M}}, \tag{2.4}$$

where $\lambda_1 > 0$ and $\lambda_2 \in (0, 1)$ are parameters. These functions $\kappa(\cdot)$ and $S(\cdot)$ are plotted in Figure 1 for $\lambda_1 = 0.75$ and $\lambda_2 = 2$. Observe that both the models satisfy assumptions (P1)–(P2).

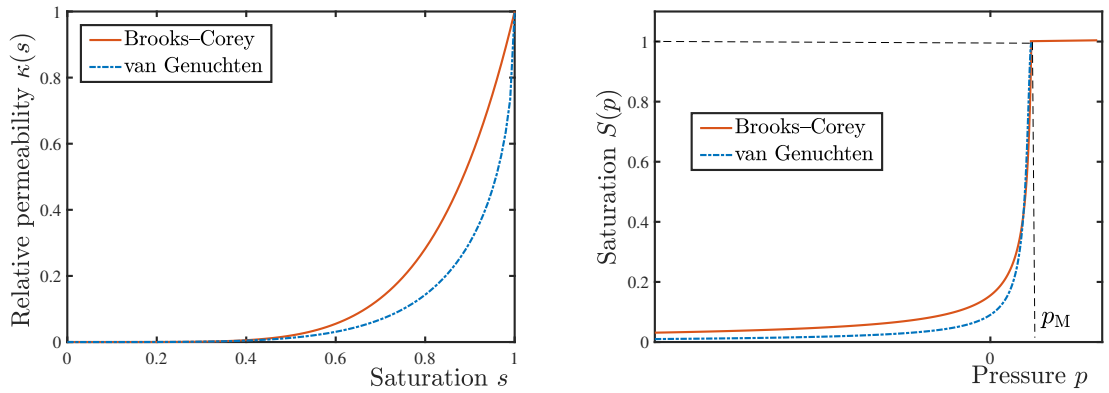

Figure 1: The functions $\kappa(s)$ (left) and $S(p)$ (right) as modeled by the Brooks–Corey (2.3) and the van Genuchten (2.4) models. The parameters are $\lambda_1 = 0.75$ and $\lambda_2 = 2$ taken from [26], which gives a rather close match between the two models. For the heat equation, for comparison, $\kappa(s) = 1$ and $S(p) = p$.

2.3 Capillary pressure, diffusivity, total pressure, and auxiliary functions

Here, we introduce some auxiliary functions that will be useful later.

2.3.1 Capillary pressure function

Since $S(\cdot)$ is a strictly increasing function in the interval $(-\infty, p_{\mathrm{M}}]$, its inverse

$$p_{\mathrm{c}}(s) := S^{-1}(s) \tag{2.5a}$$

is well-defined for $0 < s \le 1$. This is commonly known as the capillary pressure function. It is strictly increasing and $\lim_{s \searrow 0} p_{\mathrm{c}}(s) = -\infty$, see Figure 2. Using $p_{\mathrm{c}}(\cdot)$, the relation (1.1c) is alternatively stated as

$$p \begin{cases} = p_{\mathrm{c}}(s) & \text{if } 0 < s < 1, \\ \in [p_{\mathrm{M}}, \infty] & \text{if } s = 1. \end{cases} \tag{2.5b}$$

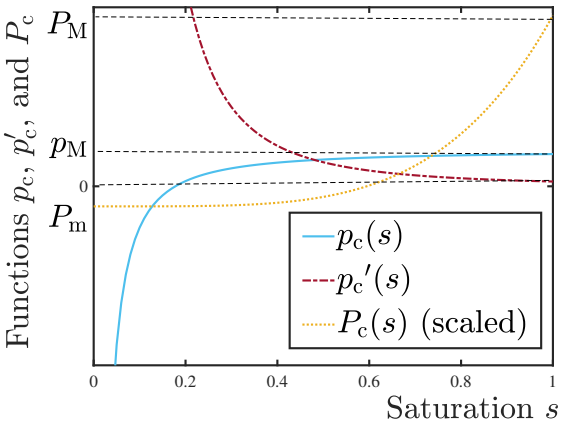

Figure 2: The p_c, p_c', and P_c functions for the Brooks–Corey model with $\lambda_1 = 0.75$.

2.3.2 Diffusivity and total pressure functions

We further introduce the diffusivity function $D : (0,1] \to \mathbb{R}^+$ as

$$D(s) := \kappa(s)\, p_c'(s), \tag{2.6}$$

and the total pressure function $P_c : (0,1] \to \mathbb{R}$ (see Figure 2) as

$$P_c(s) := \int_{S(0)}^{s} D(\varrho)\, \mathrm{d}\varrho. \tag{2.7}$$

The properties of D and P_c that follow from (P1)–(P2) are

$$D \in C^1((0,1]);\ 0 < D(s) < \infty \text{ for all } 0 < s \le 1;\ \text{and } \lim_{s \searrow 0} D(s) \ge 0; \tag{2.8}$$

whereas, $P_c \in C^1((0,1])$ is strictly increasing since

$$P_c'(s) = D(s), \tag{2.9a}$$

and there exists fixed P_m, $P_M \in [-\infty, \infty)$ depending only upon $\kappa(\cdot)$ and $p_c(\cdot)$ such that

$$P_m = \lim_{s \searrow 0} P_c(s),\ \text{and } P_M = P_c(1). \tag{2.9b}$$

Accordingly, an increasing and continuous function $\theta : \mathbb{R} \to [0,1]$ is defined by

$$\theta(\Psi) := \begin{cases} 0 & \text{if } \Psi \le P_m, \\ (P_c)^{-1}(\Psi) & \text{if } P_m < \Psi < P_M, \\ 1 & \text{if } \Psi \ge P_M. \end{cases} \tag{2.10}$$

The plots of $D(\cdot)$ and $\theta(\cdot)$ are shown in Figure 3.

Remark 2.2 (Properties of the function θ)**.** *Observe from* (2.9)–(2.10) *that,*

$$\theta'(\Psi) = \frac{1}{P_c'(\theta(\Psi))} = \frac{1}{D(\theta(\Psi))} \text{ for all } \Psi \in (P_m, P_M]. \tag{2.11}$$

Consequently, $\theta|_{(P_m, P_M]} \in C^1((P_m, P_M])$*. Moreover, it holds for all* $\Psi > P_m$ *that*

$$\Psi = P_c(\theta(\Psi)) + [\Psi - P_M]_+. \tag{2.12}$$

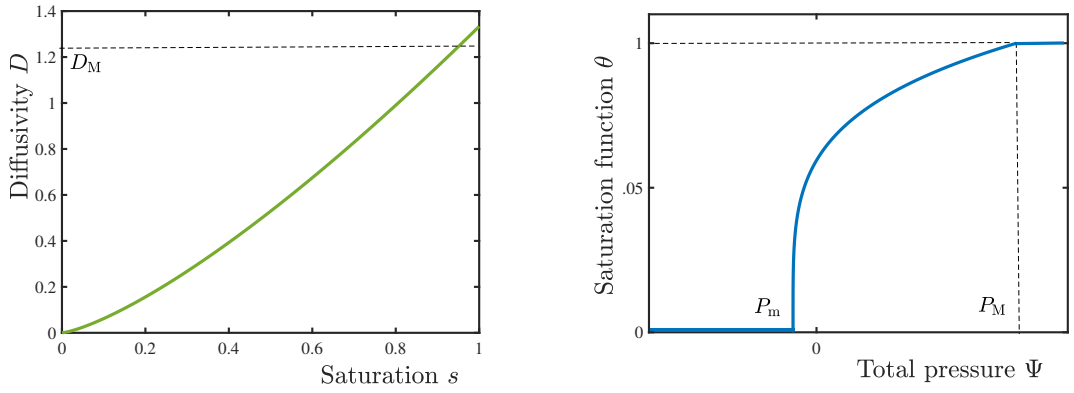

Figure 3: The functions $D(\cdot)$ and $\theta(\cdot)$ for the Brooks–Corey model with $\lambda_1 = 0.75$.

2.3.3 The Kirchhoff transform function

The well-known Kirchhoff transformation [3], $\mathcal{K} \in C^1(\mathbb{R})$, is defined by

$$\mathcal{K}(p) := \begin{cases} P_{\mathrm{c}}(S(p)) = \int_0^p \kappa(S(\varrho))\,\mathrm{d}\varrho & \text{for } p \le p_{\mathrm{M}}, \\ P_{\mathrm{M}} + \kappa(1)(p - p_{\mathrm{M}}) & \text{for } p > p_{\mathrm{M}}. \end{cases} \tag{2.13}$$

The plot of $\mathcal{K}$ is shown in Figure 4. Note from (P2) that $\mathcal{K}(p) = P_{\mathrm{c}}(S(p)) > P_{\mathrm{m}}$. This implies $\theta \circ \mathcal{K} = S$ since $\theta(\mathcal{K}(p)) = P_{\mathrm{c}}^{-1}(P_{\mathrm{c}}(S(p))) = S(p)$ if $p \le p_{\mathrm{M}}$, and $\theta(\mathcal{K}(p)) = \theta(P_{\mathrm{M}} + \kappa(1)(p - p_{\mathrm{M}})) = 1 = S(p)$ if $p > p_{\mathrm{M}}$ (see (2.10)). Consequently,

$$\text{taking } \Psi = \mathcal{K}(p) \text{ there holds } \nabla\Psi = \kappa(S(p))\nabla p, \text{ and } s = S(p) = \theta(\Psi). \tag{2.14}$$

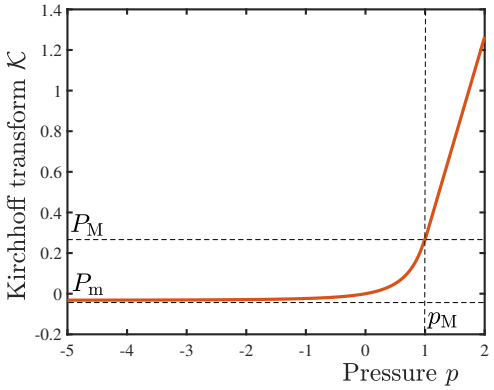

Figure 4: The Kirchhoff transform $\mathcal{K}$ for the Brooks–Corey model with $\lambda_1 = 0.75$.

Explicit expressions of all the functions introduced above can be computed for the Brooks–Corey model. They are stated in Table 1.

2.4 Weak formulations

We give below two equivalent weak formulations of the problem (1.1) discussing their strong and weak points. They will both be used to derive the a posteriori error estimates.

func.	unit	Brooks–Corey expression	func.	unit	Brooks–Corey expression
$\kappa(s)$	–	$s^{\frac{2+3\lambda_1}{\lambda_1}}$	$S(p)$	–	$(2-\frac{p}{p_{\mathrm{M}}})^{-\lambda_1}$
$p_{\mathrm{c}}(s)$	[Pa]	$p_{\mathrm{M}}(2-s^{-\frac{1}{\lambda_1}})$	$D(s)$	[Pa]	$\frac{p_{\mathrm{M}}}{\lambda_1}S^{2+\frac{1}{\lambda_1}}$
$P_{\mathrm{c}}(s)$	[Pa]	$\frac{p_{\mathrm{M}}}{(1+3\lambda_1)}(s^{3+\frac{1}{\lambda_1}}-2^{-(1+3\lambda_1)})$	$\theta(\Psi)$	–	$[\frac{1+3\lambda_1}{p_{\mathrm{M}}}\Psi+2^{-(1+3\lambda_1)}]^{\frac{\lambda_1}{1+3\lambda_1}}$

Table 1: The table of the introduced functions κ, S, p_{c}, D, P_{c}, and θ with their physical units and expressions for the Brooks–Corey model. The expressions are valid for $p \le p_{\mathrm{M}}$, $s \in (0,1]$, and $\Psi \le P_{\mathrm{M}}$. In addition, $\mathcal{K}(p) = P_{\mathrm{c}}(S(p))$ for $p \le p_{\mathrm{M}}$. In the heat equation case, for comparison, $\kappa(s) = D(s) = 1$ and S, p_{c}, P_{c}, θ, and $\mathcal{K}$ are all identity functions.

2.4.1 The pressure formulation

In the pressure formulation of (1.1), the main unknown is the pressure p. It reads: solve for $p \in \mathcal{X}$ and $s = S(p) \in H^1(0,T;H^{-1}(\Omega))$ such that $s(0) = s_0$ and for all $\varphi \in \mathcal{X}$,

$$\int_0^T \langle \partial_t S(p), \varphi\rangle + \int_0^T (\bar{\mathbf{K}}\kappa(S(p))(\nabla p + \boldsymbol{g}), \nabla\varphi) = \int_0^T (f(S(p),\boldsymbol{x},t),\varphi). \tag{2.15}$$

This formulation has the advantage of generalizing to heterogeneous porous media, where the functions S and κ are defined differently in different subdomains of Ω. In particular, since p is a physical quantity that remains continuous across the interfaces of such subdomains, formulation (2.15) has a conforming nature also in such circumstances.

2.4.2 The total pressure formulation

In the total pressure formulation of (1.1), the main unknown is the total pressure $\mathcal{K}(p)$ which will henceforth be denoted by Ψ. It reads: solve for $\Psi \in \mathcal{X}$ with $s = \theta(\Psi) \in H^1(0,T;H^{-1}(\Omega))$ such that $s(0) = s_0$ and for all $\varphi \in \mathcal{X}$,

$$\int_0^T \langle \partial_t \theta(\Psi), \varphi\rangle + \int_0^T (\bar{\mathbf{K}}(\nabla\Psi + \boldsymbol{g}\kappa(\theta(\Psi))), \nabla\varphi) = \int_0^T (f(\theta(\Psi),\boldsymbol{x},t),\varphi). \tag{2.16}$$

The formulation (2.16) is derived from (2.15) using the variable transformation (2.14). The total pressure formulation has the advantage of having a linear diffusion term. However, if the definition of κ and S varies inside the domain, for instance, in the case of heterogeneous porous media, then Ψ is not uniformly defined. Moreover, the inverse transform $\Psi \mapsto p$ is often numerically expensive to compute, and Ψ lacks a physical interpretation. We emphasize that, in this study, we have refrained from using $\mathcal{K}^{-1}$.

For $S(p) < 1$, a saturation formulation is also valid, where $s = S(p)$ is the primary unknown and $D(s)$ serves as the diffusion coefficient. This formulation, however, breaks down at $s = 1$ due to the non-invertibility of $S(p)$ [3].

2.4.3 Well-posedness

Proposition 2.1 (Existence, uniqueness, and regularity)**.** *Let (P1)–(P5) hold. Then there exists a unique weak solution $p \in \mathcal{X}$ of* (2.15) *with $s = S(p) \in \mathcal{Y}$ and $s(0) = s_0$. Moreover, there exists a unique weak solution $\Psi \in \mathcal{X}$ of* (2.16) *with $\theta(\Psi) \in \mathcal{Y}$ and $\theta(\Psi(0)) = s_0$. Furthermore, the variables p, s, and Ψ are related through* (2.14)*.*

The existence of a solution of (2.15) for $p \in \mathcal{X}$ with $\partial_t S(p) \in L^2(0,T;H^{-1}(\Omega))$ has been proven in the seminal papers [2, 3], whereas uniqueness is proven in [33] using L^1-contraction. Since $p \in \mathcal{X}$, and $S(\cdot)$ is Lipschitz continuous, one automatically gets $s \in \mathcal{Y}$. From the embedding of $\mathcal{Y}$ in $C(0,T;L^2(\Omega))$, we have $s \in C(0,T;L^2(\Omega))$. The equivalence of the p and the Ψ formulations follows from the uniqueness of the solutions.

2.5 Maximum principle

In the case of the Richards equation, the saturation s is bounded in $[0,1]$, and $s \searrow 0$ causes parabolic–hyperbolic degeneracy to occur. In this section, we use the maximum principle to obtain computable lower bounds for $s(\boldsymbol{x},t)$, bounding it away from 0. For a positive initial saturation, the function $S_{\mathrm{m}} : \mathbb{R}^+ \to (0,1]$ is a lower bound function of $s \in \mathcal{Y}$, if

$$0 < S_{\mathrm{m}}(t) \leq s(\boldsymbol{x},t) \text{ for almost all } (\boldsymbol{x},t) \in \Omega \times [0,T]. \tag{2.17}$$

To ensure that a lower bound function satisfying (2.17) exists for $S(p) \in \mathcal{Y}$ when $p \in \mathcal{X}$ solves (2.15), additional restrictions have to be imposed on the source term function f. In particular, note that if $\kappa(0) = 0$ and for some $(\boldsymbol{x},t) \in \Omega \times [0,T]$ we have $s = 0$, then from (1.1a) $\partial_t s = f(0,\boldsymbol{x},t)$. Since $s < 0$ is unphysical, this forces

$$f(0,\boldsymbol{x},t) \geq 0 \text{ for all } (\boldsymbol{x},t) \in \Omega \times [0,T]. \tag{2.18}$$

This constraint will be imposed below to obtain computable maximum principle estimates. If f is independent of s, then (2.18) simply implies that $f \geq 0$. In comparison, in the context of the heat equation, s is not bounded in $[0,1]$, and hence conditions such as (2.18) are not required.

2.5.1 A time-dependent lower bound

Recalling hypothesis (P4), define a function $\bar{S}_{\mathrm{m}}(t)$ by the integral equation

$$\bar{S}_{\mathrm{m}}(t) = \min\left(\operatorname*{ess\,inf}_{\boldsymbol{x}\in\Omega}\{s_0(\boldsymbol{x})\}, S(0)\right) + \int_0^t f_{\mathrm{m}}(\bar{S}_{\mathrm{m}}(\varrho))\,\mathrm{d}\varrho. \tag{2.19}$$

Then, we have the following result:

Proposition 2.2 (Existence of $\bar{S}_{\mathrm{m}}$ satisfying (2.19))**.** *Assume (P4)–(P5). Additionally, assume that there exists a choice of the function f_{m} such that an interval $[0,J]$, $J \in (0,1)$, and a constant $C_f \geq 0$ exist for which the inequality*

$$f_{\mathrm{m}}(s) \geq -C_f s \text{ holds } \forall s \in [0,J].$$

Then, there exists a continuous function $\bar{S}_{\mathrm{m}} : \mathbb{R}^+ \to \mathbb{R}^+$ that satisfies (2.19)*.*

The existence of $\bar{S}_{\mathrm{m}}$ follows from the Picard–Lidelöf theorem by the differentiability of the function f_{m} assumed in (P4). The bound $\bar{S}_{\mathrm{m}}(t) > 0$ follows from the inequality $\frac{d}{dt}\bar{S}_{\mathrm{m}}(t) \geq -C_f \bar{S}_{\mathrm{m}}(t)$ and $\bar{S}_{\mathrm{m}}(0) > 0$. The constraint $f \geq f_{\mathrm{m}} \geq -C_f s$ embodies and generalises (2.18). In practice, $\bar{S}_{\mathrm{m}}(t)$ can be computed to arbitrary precision using numerical approaches such as the Runge–Kutta method.

Proposition 2.3 (Time-dependent lower bound of s)**.** *Let (P4)–(P5) hold and $p \in \mathcal{X}$ with $s = S(p) \in \mathcal{Y}$ and $s(0) = s_0$ be a solution of* (2.15)*. Moreover, let $\bar{\mathbf{K}}$ be constant in Ω. Then $S_{\mathrm{m}} = \min(\bar{S}_{\mathrm{m}}, S(0))$, with $\bar{S}_{\mathrm{m}}$ defined in* (2.19)*, is a lower bound function of s satisfying* (2.17)*.*

Since proving the maximum principle result is not the main focus of this paper, we postpone the proof to Appendix B, along with other proofs of this section.

2.5.2 A space-dependent lower bound

Proposition 2.3 gives a computable lower bound of $s(\boldsymbol{x},t)$ for a given $t \in [0,T]$, provided the absolute permeability $\bar{\mathbf{K}}$ is constant. The following result also gives a lower bound of s, relaxing the assumption of $\bar{\mathbf{K}}$ being constant.

Proposition 2.4 (Existence of a bounded function)**.** *Let (P1)–(P4) hold. For a constant $J \leq 0$, let $\varsigma \in H^1(\Omega)$ with $\varsigma = J$ on $\partial\Omega$ in the trace sense, solve*

$$(\bar{\mathbf{K}}\kappa(S(\varsigma))[\nabla\varsigma + \boldsymbol{g}], \nabla\varphi) = \left(\inf_{t\in\mathbb{R}^+}[f(S(\varsigma),\boldsymbol{x},t)]_-, \varphi\right), \quad \forall\varphi \in H_0^1(\Omega). \tag{2.20}$$

Assume that there exists a constant $p_\mathrm{l} \leq 0$ such that $f(S(p),\boldsymbol{x},t) \geq 0$ for all $p < p_\mathrm{l}$. Then

$$\min(p_\mathrm{l}, J) + \min_{\boldsymbol{x}\in\Omega}\{\boldsymbol{g}\cdot\boldsymbol{x}\} \leq \varsigma(\boldsymbol{x}) + \boldsymbol{g}\cdot\boldsymbol{x} \leq J + \max_{\boldsymbol{x}\in\Omega}\{\boldsymbol{g}\cdot\boldsymbol{x}\} \text{ for almost all } \boldsymbol{x} \in \Omega. \tag{2.21}$$

The existence of ς follows from [2] and the existence of $p_\mathrm{l} < 0$ is compatible with (2.18). The counterpart of Proposition 2.3 for this case is:

Proposition 2.5 (Space-dependent lower bound of s)**.** *Let (P4)–(P5) hold and $p \in \mathcal{X}$ with $s = S(p) \in \mathcal{Y}$ and $s(0) = s_0$ be a solution of* (2.15)*. For the constant*

$$J = \operatorname*{ess\,inf}_{\boldsymbol{x}\in\Omega}\left([p_\mathrm{c}(s_0(\boldsymbol{x}))]_- - \max_{\boldsymbol{x}\in\Omega}\{\boldsymbol{g}\cdot\boldsymbol{x}\} + \boldsymbol{g}\cdot\boldsymbol{x}\right) \leq 0,$$

let $\varsigma \in H^1(\Omega)$ be obtained from Proposition 2.4. Then $S_\mathrm{m}(t) = \operatorname*{ess\,inf}_{\boldsymbol{x}\in\Omega}(S(\varsigma(\boldsymbol{x})))$ for $t > 0$ is a lower bound function of s satisfying (2.17)*.*

3 Relations between the error and the residual

In this section, the dual norm of the residual will be used to bound from above and from below an error metric that we will use in place of the $\mathcal{Y}$-norm in the present nonlinear and degenerate setting.

3.1 Residual

For

$$\Psi_{h\tau} \in \mathcal{X} \text{ with } s_{h\tau} := \theta(\Psi_{h\tau}) \in \mathcal{Y}, \tag{3.1}$$

the residual $\mathcal{R}(\Psi_{h\tau}) \in L^2(0,T;H^{-1}(\Omega))$ with respect to the weak formulation (2.16) is defined as

$$\int_0^T \langle\mathcal{R}(\Psi_{h\tau}),\varphi\rangle := \int_0^T \left[(f(s_{h\tau},\boldsymbol{x},t),\varphi) - \langle\partial_t s_{h\tau},\varphi\rangle - (\bar{\mathbf{K}}(\nabla\Psi_{h\tau} + \boldsymbol{g}\kappa(s_{h\tau})),\nabla\varphi)\right] \tag{3.2}$$

for all $\varphi \in \mathcal{X}$. If $\Psi \in \mathcal{X}$ with $s = \theta(\Psi) \in \mathcal{Y}$ denotes the solution to (2.16) then $\mathcal{R}(\Psi) = 0$.

3.2 Norms

On a Lipschitz subdomain $\omega \subseteq \Omega$, we introduce equivalent (semi)norms on $H^1(\omega)$, $H^1_0(\omega)$ and $H^{-1}(\omega)$:

$$\|\varrho\|_{H^1_{\bar{\mathbf{K}}}(\omega)} := \|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla\varrho\|_\omega \text{ for } \varrho \in H^1(\omega), \tag{3.3a}$$

$$\|\varrho\|_{H^{-1}_{\bar{\mathbf{K}}}(\omega)} := \sup_{\varphi\in H^1_0(\omega)} \{\langle\varrho,\varphi\rangle_{H^{-1}(\omega),H^1_0(\omega)}/\|\varphi\|_{H^1_{\bar{\mathbf{K}}}(\omega)}\} \text{ for } \varrho \in H^{-1}(\omega). \tag{3.3b}$$

From the properties of $\bar{\mathbf{K}}$ stated in (P3), it is immediate that

$$K_{\mathrm{m}}^{\frac{1}{2}}\|\nabla\varrho\|_\omega \le \|\varrho\|_{H^1_{\bar{\mathbf{K}}}(\omega)} \le K_{\mathrm{M}}^{\frac{1}{2}}\|\nabla\varrho\|_\omega \text{ and } K_{\mathrm{M}}^{-\frac{1}{2}}\|\varrho\|_{H^{-1}(\omega)} \le \|\varrho\|_{H^{-1}_{\bar{\mathbf{K}}}(\omega)} \le K_{\mathrm{m}}^{-\frac{1}{2}}\|\varrho\|_{H^{-1}(\omega)}. \tag{3.4}$$

Let $\alpha : [0,T] \to [0,\infty)$ denote a bounded non-negative function. For a subdomain $\omega \subseteq \Omega$, and an interval $I \subseteq [0,T]$, we introduce the distance measure $\mathrm{dist}^\alpha_{\omega,I}$ on the set $\{\psi \in L^2(0,T;H^1(\omega)) : \theta(\psi) \in H^1(0,T;H^{-1}(\omega))\}$ as

$$\begin{aligned}\mathrm{dist}^\alpha_{\omega,I}(\Psi_1,\Psi_2) :=& \|\partial_t(\theta(\Psi_1)-\theta(\Psi_2))\|_{L^2(I;H^{-1}_{\bar{\mathbf{K}}}(\omega))} \\ &+ \|\alpha(\theta(\Psi_1)-\theta(\Psi_2))\|_{L^2(\omega\times I)} + \|\Psi_1-\Psi_2\|_{L^2(I,H^1_{\bar{\mathbf{K}}}(\omega))}.\end{aligned} \tag{3.5}$$

The distance measure combines the $L^2(I;H^1_{\bar{\mathbf{K}}}(\omega))$-norm of $\Psi_1-\Psi_2$ with the $H^1(I;H^{-1}_{\bar{\mathbf{K}}}(\omega)) \cap L^2(\omega\times I)$ norms of $\theta(\Psi_1)-\theta(\Psi_2)$. Note that for $\alpha = 0$, the middle term disappears.

Let $\varrho \in L^2([0,T])$. Later, we will take for ϱ the error components containing $\|s-s_{h\tau}\|$ and $\|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(\Psi-\Psi_{h\tau})\|$ and we will also use $\|\mathcal{R}(\Psi_{h\tau})\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)}$ for ϱ. We introduce the class of time-integration functionals $\mathcal{J}_\alpha : L^2([0,T]) \to [0,\infty)$

$$\mathcal{J}_\alpha(\varrho) := \left[\exp\left(-\int_0^T\alpha\right)\int_0^T\left(\varrho^2(t)+\alpha(t)\exp\left(\int_t^T\alpha\right)\int_0^t\varrho^2\right)\mathrm{d}t\right]^{\frac{1}{2}}. \tag{3.6}$$

The operator $\mathcal{J}_\alpha$ defines a norm on $L^2([0,T])$, so that it in particular satisfies the triangle inequality (since $(\int_0^t(\varrho_1+\varrho_2)^2)^{\frac{1}{2}} \le (\int_0^t\varrho_1^2)^{\frac{1}{2}}+(\int_0^t\varrho_2^2)^{\frac{1}{2}}$). It is actually equivalent to the $L^2([0,T])$ norm, since the inequality $0 \le \int_0^t\varrho^2\mathrm{d}t \le \|\varrho\|^2_{L^2([0,T])}$ for $t\in[0,T]$, and $\int_0^T\alpha\exp(\int_t^T\alpha)\mathrm{d}t = \exp(\int_0^T\alpha)-1$ directly gives for $\alpha \ge 0$ that

$$\exp\left(-\tfrac{1}{2}\int_0^T\alpha\right)\|\varrho\|_{L^2([0,T])} \le \mathcal{J}_\alpha(\varrho) \le \|\varrho\|_{L^2([0,T])}. \tag{3.7}$$

Consequently, it is equal to the $L^2([0,T])$ norm if $\alpha = 0$ and up to by the exponential factor $\exp\left(-\frac{1}{2}\int_0^T\alpha\right)$ smaller than the $L^2([0,T])$ norm. This way of measuring the error has been designed in [13] based on working with as sharp as possible form of the Gronwall Lemma, not neglecting some integral terms and avoiding the appearance of the usual factor $\exp(T)$ in the relation between the error and the residual, see Lemma 5.1 and Theorem 5.2 in this reference. The following remarks brings additional background:

Remark 3.1 (Interpretation of $\mathcal{J}_\alpha$)**.** *Let $\varrho \in C^1([0,T])$ be a non-negative function whose instantaneous rate of change in relative magnitude is given by a constant $\frac{1}{2}\bar\alpha \in \mathbb{R}$ at time $t>0$, i.e. $\frac{\mathrm{d}\varrho}{\mathrm{d}t} = \frac{1}{2}\bar\alpha\varrho$. This gives $\varrho(t)^2 = A^2\exp(\bar\alpha t)$ for some $A \ge 0$, so that ϱ grows exponentially with the final simulation time T. This is the usual setting for the errors in numerical approximation*

of the problems we are considering, see Section 6, in particular Figures 8 and 14. Evaluating the integral in (3.6) *and using* $\frac{\mathrm{d}v}{\mathrm{d}t} + \alpha v = \bar{\alpha}\, e^{\int_0^t (\bar{\alpha}-\alpha)}$ *for* $v(t) = e^{-\int_0^t \alpha}(e^{\bar{\alpha}t} - 1)$*, one has*

$$\mathcal{J}_\alpha(\varrho) = A\left(\tfrac{1}{\bar{\alpha}}\, e^{-\int_0^T \alpha}(e^{\bar{\alpha}T} - 1) + \int_0^T \tfrac{\alpha(t)}{\bar{\alpha}} e^{-\int_0^t \alpha}(e^{\bar{\alpha}t} - 1)\right)^{\frac{1}{2}} = A\left(\int_0^T e^{\int_0^t (\bar{\alpha}-\alpha)}\,\mathrm{d}t\right)^{\frac{1}{2}}, \quad (3.8)$$

Hence, considering α *also constant such that* $\alpha > \bar{\alpha}$ *and* $(\alpha - \bar{\alpha})T \gg 1$*, we get that* $\mathcal{J}_\alpha(\varrho) = A(\frac{1-\exp(-(\alpha-\bar{\alpha})T)}{\alpha-\bar{\alpha}})^{\frac{1}{2}} \approx A(\alpha-\bar{\alpha})^{-\frac{1}{2}}$*. Since in our study, the assumption* $(\alpha - \bar{\alpha})T \gg 1$ *turns out to be satisfied, starting from rather small* T*, see Section 6, the norm* $\mathcal{J}_\alpha$ *yields an almost constant value of error independent of* $T \geq 1$*, see Figures 6 and 12 for examples. We find this as a particularly suitable setting which allows us to compare the errors at different time instances during the simulation period* $[0, T]$*.*

3.3 Lower bound on the error by the residual

Extending Theorem 2.1 of [17] to the present degenerate nonlinear setting, we have

Theorem 3.2 (Lower bound on error by the dual norm of the residual)**.** *Let (P1)–(P5) hold and let* $\Psi \in \mathcal{X}$ *with* $s = \theta(\Psi) \in \mathcal{Y}$ *denote the unique solution of* (2.16)*. Let* $\omega \subseteq \Omega$ *be a Lipschitz subdomain of* Ω *and let* $I \subseteq [0, T]$ *be a time interval. Let the norms* $\|\cdot\|_{H^1_{\bar{\mathbf{K}}}}$, $\|\cdot\|_{H^{-1}_{\bar{\mathbf{K}}}}$ *and the error measure* $\mathrm{dist}^\alpha_{\omega,I}(\cdot,\cdot)$ *be defined as in* (3.3)–(3.5) *for* $\alpha(t) = C_{\mathrm{P},\omega}\, h_\omega\, K_{\mathrm{m}}^{-\frac{1}{2}} \max_{[0,1]\times\omega\times\{t\}} |\partial_s f| + |\boldsymbol{g}| K_{\mathrm{M}}^{\frac{1}{2}} \|\kappa'\|_{L^\infty([0,1])}$*. Then, for any* $\Psi_{h\tau} \in \mathcal{X}$ *with* $s_{h\tau} = \theta(\Psi_{h\tau}) \in \mathcal{Y}$*, one has*

$$\|\mathcal{R}(\Psi_{h\tau})\|_{L^2(I;H^{-1}_{\bar{\mathbf{K}}}(\omega))} \leq \mathrm{dist}^\alpha_{\omega,I}(\Psi, \Psi_{h\tau}). \quad (3.9)$$

Remark 3.3 (The linear case)**.** *Observe that in the linear case,* $\kappa = 1$ *and* $\partial_s f = 0$*, yielding* $\alpha = 0$*.*

Proof. From (3.2), one has for any $\varphi \in L^2(I; H^1_0(\omega))$, extended to $\Omega \setminus \omega$ and $[0, T] \setminus I$ by 0, that

$$\begin{aligned}\int_I \langle \mathcal{R}(\Psi_{h\tau}), \varphi\rangle = \int_I \big[&\langle \partial_t(s - s_{h\tau}), \varphi\rangle + (\bar{\mathbf{K}}\nabla(\Psi - \Psi_{h\tau}), \nabla\varphi) \\ &+ (f(s_{h\tau}, \boldsymbol{x}, t) - f(s, \boldsymbol{x}, t), \varphi) + (\bar{\mathbf{K}}\boldsymbol{g}(\kappa(s) - \kappa(s_{h\tau})), \nabla\varphi)\big]. \end{aligned} \quad (3.10)$$

Then, from the triangle inequality and definitions of the norms $\|\cdot\|_{H^1_{\bar{\mathbf{K}}}}$, $\|\cdot\|_{H^{-1}_{\bar{\mathbf{K}}}}$ we get

$$\begin{aligned}\|\mathcal{R}(\Psi_{h\tau})\|_{L^2(I;H^{-1}_{\bar{\mathbf{K}}}(\omega))} &\leq \|\partial_t(s - s_{h\tau})\|_{L^2(I;H^{-1}_{\bar{\mathbf{K}}}(\omega))} + \|\Psi - \Psi_{h\tau}\|_{L^2(I;H^1_{\bar{\mathbf{K}}}(\omega))} \\ &+ \sup_{\substack{\varphi\in L^2(I;H^1_0(\omega)),\\ \|\varphi\|_{L^2(I;H^1_{\bar{\mathbf{K}}}(\omega))}=1}} \int_I \big[(f(s_{h\tau}, \boldsymbol{x}, t) - f(s, \boldsymbol{x}, t), \varphi)_\omega + (\bar{\mathbf{K}}\boldsymbol{g}(\kappa(s) - \kappa(s_{h\tau})), \nabla\varphi)_\omega\big].\end{aligned}$$

The result then follows from the definition of $\mathrm{dist}^\alpha_{\omega,I}$ and the computation of the last two terms using

$$\begin{aligned}|(f(s_{h\tau}, \boldsymbol{x}, t) - f(s, \boldsymbol{x}, t), \varphi)_\omega| &\overset{(2.2),(3.4)}{\leq} C_{\mathrm{P},\omega}\, h_\omega\, K_{\mathrm{m}}^{-\frac{1}{2}} \max_{[0,1]\times\omega\times\{t\}} |\partial_s f| \|s_{h\tau} - s\|_\omega \|\varphi\|_{H^1_{\bar{\mathbf{K}}}(\omega)}, \\ |(\bar{\mathbf{K}}\boldsymbol{g}(\kappa(s) - \kappa(s_{h\tau})), \nabla\varphi)_\omega| &\overset{(P1),(P3)}{\leq} |\boldsymbol{g}| K_{\mathrm{M}}^{\frac{1}{2}} \|\kappa'\|_{L^\infty([0,1])} \|s_{h\tau} - s\|_\omega \|\varphi\|_{H^1_{\bar{\mathbf{K}}}(\omega)}.\end{aligned}$$

□

3.4 Upper bound on the error by the residual

For the lower bound function $S_m(t)$ satisfying (2.17), the diffusivity function D of (2.6), the saturation function θ of (2.10), and the source term f of (P4), let

$$D_{\mathrm{m}}(t) := \min\{D(\varrho) : \varrho \in [S_{\mathrm{m}}(t), 1]\}, \quad D_{\mathrm{M}}(t) := \max\{|D'(\varrho)| : \varrho \in [S_{\mathrm{m}}(t), 1]\}, \tag{3.11a}$$

$$\theta_{\partial,\mathrm{M}}(t) := \max\{\theta'(P_{\mathrm{c}}(\varrho)) : \varrho \in [S_{\mathrm{m}}(t), 1]\}, \tag{3.11b}$$

$$f_{\partial,M}(t) := \max\{|\partial_s f(\varrho, \boldsymbol{x}, t)| : \varrho \in [0,1],\ (\boldsymbol{x}, t) \in \Omega \times [0,T]\}. \tag{3.11c}$$

Recalling (2.8), we have $D_{\mathrm{m}}(t) > 0$ and $D_{\mathrm{M}}(t) < \infty$. Similarly $\theta_{\partial,\mathrm{M}}(t) < \infty$, $f_{\partial,M}(t) < \infty$. Then, inspired by [13] we propose

Theorem 3.4 (Upper bound on error by the dual norm of the residual)**.** *Let (P1)–(P5) hold and $\Psi \in \mathcal{X}$ denote the unique solution of (2.16) with $s = \theta(\Psi) \in \mathcal{Y}$. Let $\Psi_{h\tau} \in \mathcal{X}$ with $s_{h\tau} = \theta(\Psi_{h\tau}) \in \mathcal{Y}$ be arbitrary. Assume that a lower bound function $S_{\mathrm{m}}(t)$, satisfying (2.17), exists for s and $s_{h\tau}$. Recall the definitions of D_{m}, D_{M}, $f_{\partial,M}$, and $\theta_{\partial,\mathrm{M}}$ from (3.11). Let the residual $\mathcal{R}$, norms $\|\cdot\|_{H^{\pm1}_{\bar{\mathbf{K}}}}$, and the time-integrator $\mathcal{J}_\alpha$ be defined in (3.2), (3.3), and (3.6) respectively. Then, for any $\lambda : [0,T] \to \mathbb{R}^+$, the following estimates hold:* ***Estimate in the*** *$L^2(\Omega \times [0,T])$* ***and*** *$L^\infty(0,T; H^{-1}_{\bar{\mathbf{K}}}(\Omega))$* ***norms:***

$$\begin{aligned} &e^{-\int_0^T (\lambda + \mathfrak{C}_1)} \|(s - s_{h\tau})(T)\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} + \mathcal{J}_{\lambda+\mathfrak{C}_1}\left(\theta_{\partial,\mathrm{M}}^{-\frac{1}{2}} \|s - s_{h\tau}\|\right)^2 \\ \le &\|s_0 - s_{h\tau}(0)\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} + \mathcal{J}_{\lambda+\mathfrak{C}_1}(\lambda^{-\frac{1}{2}} \|\mathcal{R}(\Psi_{h\tau})\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)})^2. \end{aligned} \tag{3.12a}$$

Estimate in the *$L^2(0,T; H^1_{\bar{\mathbf{K}}}(\Omega))$* ***and*** *$L^\infty(0,T; L^2(\Omega))$* ***norms:*** *For*

$$C^\infty_{h\tau}(t) := \|\bar{\mathbf{K}}^{\frac{1}{2}} \nabla s_{h\tau}(t)\|^2_{L^\infty(\Omega)}, \quad \text{assume that } \int_0^T C^\infty_{h\tau}(t)\,\mathrm{d}t < \infty.$$

On a Lipschitz subdomain

$$\Omega^{\deg}(t) \supseteq \{s(\boldsymbol{x}, t) = 1\} \cup \{s_{h\tau}(\boldsymbol{x}, t) = 1\}$$

of Ω (possibly disconnected), let $D(s)/2 \le D(s_{h\tau}) \le 2D(s)$ hold, and define the parabolic–elliptic degeneracy estimator $\eta^{\deg} \in L^2([0,T])$ as

$$\begin{aligned} \eta^{\deg}(t) := &\sqrt{\tfrac{2}{D(1)}} \Bigg[\|[\Psi_{h\tau}(t) - P_{\mathrm{M}}]_+\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)} \\ &+ \left(\|[f(1, \boldsymbol{x}, t)]_+\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega^{\deg}(t))} + \left\| \left(\bar{\mathbf{K}}^{\frac{1}{2}} - \frac{\bar{\mathbf{K}}^{-\frac{1}{2}}}{|\Omega^{\deg}(t)|} \int_{\Omega^{\deg}(t)} \bar{\mathbf{K}} \right) \boldsymbol{g} \right\|_{\Omega^{\deg}(t)} \right)^2 \Bigg]^{\frac{1}{2}}. \end{aligned}$$

Then it holds that,

$$\begin{aligned} &e^{-\int_0^T \mathfrak{C}_2} \|(s - s_{h\tau})(T)\|^2 + \tfrac{1}{2} \mathcal{J}_{\mathfrak{C}_2}\left(\left\| D(s)^{-\frac{1}{2}} \bar{\mathbf{K}}^{\frac{1}{2}} \nabla(\Psi - \Psi_{h\tau}) \right\| \right)^2 \\ \le &\|s_0 - s_{h\tau}(0)\|^2 + \mathcal{J}_{\mathfrak{C}_2}\left(\eta^{\deg}\right)^2 + 4\,\mathcal{J}_{\mathfrak{C}_2}\left(D_{\mathrm{m}}^{-\frac{1}{2}} \|\mathcal{R}(\Psi_{h\tau})\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} \right)^2. \end{aligned} \tag{3.12b}$$

Estimate in the *$H^1(0,T; H^{-1}_{\bar{\mathbf{K}}}(\Omega))$* ***norm:***

$$\begin{aligned} &\mathcal{J}_\lambda(\|\partial_t(s - s_{h\tau})\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)})^2 \\ \le &\,3\left[\mathcal{J}_\lambda(\|\Psi - \Psi_{h\tau}\|_{H^1_{\bar{\mathbf{K}}}(\Omega)})^2 + \mathfrak{C}_3(T)\, \mathcal{J}_\lambda\left(\|s - s_{h\tau}\|\right)^2 + \mathcal{J}_\lambda(\|\mathcal{R}(\Psi_{h\tau})\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)})^2 \right]. \end{aligned} \tag{3.12c}$$

Recalling the Poincaré constant $C_{\mathrm{P},\Omega}$ from (2.2)*, the functions $\mathfrak{C}_{1,2,3} : (0,T) \to [0,\infty)$ are*

$$\mathfrak{C}_1(t) := 2\theta_{\partial,\mathrm{M}}(t)\left[K_{\mathrm{M}}|\boldsymbol{g}|^2\|\kappa'\|^2_{L^\infty([0,1])} + \tfrac{C^2_{\mathrm{P},\Omega}h^2_\Omega}{K_{\mathrm{m}}} f^2_{\partial,\mathrm{M}}(t)\right], \tag{3.13a}$$

$$\mathfrak{C}_2(t) := \tfrac{1}{D_{\mathrm{m}}(t)}\left[D^2_{\mathrm{M}}(t)\, C^\infty_{h\tau}(t) + 4K_{\mathrm{M}}|\boldsymbol{g}|^2\|\kappa'\|^2_{L^\infty([0,1])}\right] + 2f_{\partial,\mathrm{M}}(t), \tag{3.13b}$$

$$\mathfrak{C}_3(t) := (C_{\mathrm{P},\Omega}h_\Omega\, K_{\mathrm{m}}^{-\frac{1}{2}}\, \|f_{\partial,\mathrm{M}}\|_{L^\infty([0,t])} + K_{\mathrm{M}}^{\frac{1}{2}}|\boldsymbol{g}|\|\kappa'\|_{L^\infty([0,1])})^2. \tag{3.13c}$$

The function $\lambda > 0$ in (3.12a) is introduced to optimize the effectivity of the estimates. The reason as well as a possible value of λ will be explained in detail in Remark 6.1.

Remark 3.5 (Degeneracy at $s = 1$)**.** *Observe that the estimate* (3.12b) *contains the degenereacy estimator η^{deg}, despite the estimates* (3.12a), (3.12c) *not including it. This stems from the fact that proving a contraction in $L^2(0,T;H^1(\Omega))$ is generally not possible for degenerate problems. However, proving contraction in the $L^\infty(0,T;L^1(\Omega))$ and the $L^2(0,T;H^{-1}(\Omega))$ norms is possible [25, 33]. The last two components in the definition of η^{deg} represent the two reasons why the parabolic–elliptic degeneracy might occur despite the initial condition s_0 being in $(0,1]$, i.e. the positivity of f and the non-uniformity of $\bar{\mathbf{K}}$. For a specified Ω^{deg}, the estimator η^{deg} is fully computable. In the numerical examples of Sections 6.2 to 6.4 we take Ω^{deg} to be a considerably larger superset of $\{s_{h\tau} = 1\}$ to ensure the validity of Ω^{deg} also being a superset of $\{s = 1\}$. However, we admit that this is a theoretical assumption that cannot be in general verified in practice.*

Remark 3.6 (Reduction in the linear case)**.** *Observe that, in the linear heat equation case, $\kappa(s)$ and $p'_{\mathrm{c}}(s)$ are equal to 1, giving a constant $D(s) = 1$ and $D_{\mathrm{M}}(t) = 0$. Similarly $\partial_s f = 0$. Thus, one has $\mathfrak{C}_{1,2,3} = 0$. Hence, for the linear case, taking $\lambda = 0$ in* (3.12c) *exponential terms in* (3.12b)–(3.12c) *vanish, and they reduce to the estimates provided in [17, 40].*

Remark 3.7 (Bounds on $\|\partial_t(s - s_{h\tau})\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)}$ and $\mathrm{dist}^\alpha_{\Omega,[0,T]}(\Psi,\Psi_{h\tau})$)**.** *Choosing λ in* (3.12a) *such that $\lambda + \mathfrak{C}_1 = \mathfrak{C}_2$ and $\lambda = \mathfrak{C}_2$ in* (3.12c)*, we have a complete bound for $\|\partial_t(s - s_{h\tau})\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)}$ using the other components of* (3.12)*. Combining* (3.12)*, one obtains an estimate for all components of $\mathrm{dist}^\alpha_{\Omega,[0,T]}(\Psi,\Psi_{h\tau})$ defined in* (3.5)*. Hence, Theorems 3.2 and 3.4 provide both lower and upper bounds of $\mathrm{dist}^\alpha_{\Omega,[0,T]}(\Psi,\Psi_{h\tau})$ in that using* (3.7)*, one has*

$$\begin{aligned}\int_0^T \|\mathcal{R}(\Psi_{h\tau})\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} &\overset{(3.9)}{\leq} \mathrm{dist}^\alpha_{\Omega,[0,T]}(\Psi,\Psi_{h\tau})^2 \\ &\overset{(3.12)}{\lesssim} \exp\left(\int_0^T \mathfrak{C}_2\right)\left[\|s_0 - s_{h\tau}(0)\|^2 + \int_0^T \left([\eta^{\mathrm{deg}}]^2 + \|\mathcal{R}(\Psi_{h\tau})\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)}\right)^2\right].\end{aligned}$$

However, this upper bound is rather rough since it hides its dependence on D_{m}, $\mathfrak{C}_{1/3}$ and $C_{\mathrm{P},\Omega}\, h_\Omega$. Note that $\exp\left(\int_0^T \mathfrak{C}_2\right)$ may take very large values and might explode as $T \to \infty$, which is the usual consequence of using Gronwall Lemma. This is avoided in our analysis.

Proof of Theorem 3.4. In the proof, we shorten $\mathcal{R}(\Psi_{h\tau})$ to simply $\mathcal{R}$. From (3.2), we have for all $\varphi \in L^2(0,T;H^1_0(\Omega))$,

$$\begin{aligned}&\int_0^T \left[\langle \partial_t(s - s_{h\tau}), \varphi\rangle + (\bar{\mathbf{K}}\nabla(\Psi - \Psi_{h\tau}), \nabla\varphi)\right] \\ &\quad = \int_0^T \left[\langle \mathcal{R}, \varphi\rangle + (f(s,\boldsymbol{x},t) - f(s_{h\tau},\boldsymbol{x},t)), \varphi) + (\bar{\mathbf{K}}\boldsymbol{g}(\kappa(s_{h\tau}) - \kappa(s)), \nabla\varphi)\right].\end{aligned} \tag{3.14}$$

Step 1 (Estimate (3.12a)**):** Let the Green function $G^0_{h\tau} \in C(0,T;H^1_0(\Omega))$ satisfy for all $t \in [0,T]$ and $\varphi \in H^1_0(\Omega)$,

$$(\bar{\mathbf{K}}\nabla G^0_{h\tau}(t), \nabla\varphi) = \langle (s - s_{h\tau})(t), \varphi\rangle. \tag{3.15}$$

The problem is well-defined as $(s - s_{h\tau})(t) \in L^2(\Omega)$. Moreover,

$$\begin{aligned}\|G^0_{h\tau}(t)\|_{H^1_{\bar{\mathbf{K}}}(\Omega)} &= \sup_{\|\varphi\|_{H^1_{\bar{\mathbf{K}}}(\Omega)}=1} (\bar{\mathbf{K}}\nabla G^0_{h\tau}(t), \nabla\varphi) \\ &= \sup_{\|\varphi\|_{H^1_{\bar{\mathbf{K}}}(\Omega)}=1} \langle (s - s_{h\tau})(t), \varphi\rangle = \|(s - s_{h\tau})(t)\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)}.\end{aligned} \tag{3.16}$$

Since $\partial_t(s - s_{h\tau}) \in L^2(0,T;H^{-1}(\Omega))$, equation (3.15) can be differentiated in time, implying that $\partial_t G^0_{h\tau} \in \mathcal{X}$ exists satisfying

$$\int_0^T (\bar{\mathbf{K}}\nabla\partial_t G^0_{h\tau}, \nabla\varphi) = \int_0^T \langle \partial_t(s - s_{h\tau}), \varphi\rangle \text{ for all } \varphi \in \mathcal{X}. \tag{3.17}$$

We now insert the test function $\varphi = G^0_{h\tau}$ in (3.14). Using (3.17), we see

$$\begin{aligned}&\int_0^T \langle \partial_t(s - s_{h\tau}), G^0_{h\tau}\rangle = \int_0^T (\bar{\mathbf{K}}\nabla\partial_t G^0_{h\tau}, \nabla G^0_{h\tau}) = \tfrac{1}{2}\int_\Omega \left[|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla G^0_{h\tau}(T)|^2 - |\bar{\mathbf{K}}^{\frac{1}{2}}\nabla G^0_{h\tau}(0)|^2\right] \\ &\overset{(3.16)}{=} \tfrac{1}{2}\|G^0_{h\tau}(T)\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)} - \tfrac{1}{2}\|s_0 - s_{h\tau}(0)\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)}.\end{aligned} \tag{3.18}$$

Using the identity (2.12) and noting that $([\Psi - P_{\mathrm{M}}]_+ - [\Psi_{h\tau} - P_{\mathrm{M}}]_+, s - s_{h\tau}) \geq 0$ which follows from the monotonicity of $[\cdot]_+$, one further has from (3.15) that

$$\begin{aligned}&\int_0^T (\bar{\mathbf{K}}\nabla(\Psi - \Psi_{h\tau}), \nabla G^0_{h\tau}) = \int_0^T (\Psi - \Psi_{h\tau}, s - s_{h\tau}) \overset{(2.12)}{\geq} \int_0^T (P_{\mathrm{c}}(s) - P_{\mathrm{c}}(s_{h\tau}), s - s_{h\tau}) \\ &= \int_0^T \int_\Omega {P_{\mathrm{c}}}'(s - s_{h\tau})^2 \overset{(2.11),(3.11)}{\geq} \int_0^T \frac{1}{\theta_{\partial,M}(t)}\|s - s_{h\tau}\|^2.\end{aligned} \tag{3.19}$$

Recalling the Poincaré inequality (2.2) and the definitions (3.3) of $\|\cdot\|_{H^1_{\bar{\mathbf{K}}}}$, $\|\cdot\|_{H^{-1}_{\bar{\mathbf{K}}}}$ norms, we have

$$\int_0^T \langle \mathcal{R}, G^0_{h\tau}\rangle \leq \int_0^T \|\mathcal{R}\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)}\, \|G^0_{h\tau}\|_{H^1_{\bar{\mathbf{K}}}(\Omega)} \leq \int_0^T \left[\tfrac{1}{2\lambda}\|\mathcal{R}\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} + \tfrac{\lambda}{2}\|G^0_{h\tau}\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)}\right], \tag{3.20}$$

as well as

$$\begin{aligned}&\int_0^T (f(s,\boldsymbol{x},t) - f(s_{h\tau},\boldsymbol{x},t), G^0_{h\tau}) \overset{(3.11)}{\leq} \int_0^T f_{\partial,\mathrm{M}}(t)\|s - s_{h\tau}\|\|G^0_{h\tau}\| \\ &\leq \tfrac{1}{4}\int_0^T \frac{1}{\theta_{\partial,\mathrm{M}}(t)}\|s - s_{h\tau}\|^2 + \int_0^T \theta_{\partial,\mathrm{M}}(t)\, f_{\partial,\mathrm{M}}(t)^2\, \|G^0_{h\tau}\|^2 \\ &\overset{(2.2),(3.4)}{\leq} \tfrac{1}{4}\int_0^T \frac{1}{\theta_{\partial,\mathrm{M}}(t)}\|s - s_{h\tau}\|^2 + \tfrac{C^2_{\mathrm{P},\Omega}h^2_\Omega}{K_{\mathrm{m}}}\int_0^T \theta_{\partial,\mathrm{M}}(t)\, f_{\partial,\mathrm{M}}(t)^2\, \|G^0_{h\tau}\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)},\end{aligned} \tag{3.21}$$

and

$$\begin{aligned}&\int_0^T (\bar{\mathbf{K}}\boldsymbol{g}(\kappa(s_{h\tau}) - \kappa(s)), \nabla G^0_{h\tau}) \leq \tfrac{1}{4K_{\mathrm{M}}|\boldsymbol{g}|^2\|\kappa'\|^2_{L^\infty([0,1])}}\int_0^T \tfrac{1}{\theta_{\partial,\mathrm{M}}(t)}\int_\Omega \boldsymbol{g}^{\mathrm{T}}\bar{\mathbf{K}}\boldsymbol{g}(\kappa(s) - \kappa(s_{h\tau}))^2 \\ &\qquad + K_{\mathrm{M}}|\boldsymbol{g}|^2\|\kappa'\|^2_{L^\infty([0,1])}\int_0^T \theta_{\partial,\mathrm{M}}(t)\int_\Omega |\bar{\mathbf{K}}^{\frac{1}{2}}\nabla G^0_{h\tau}|^2 \\ &\overset{(P3)}{\leq} \tfrac{1}{4}\int_0^T \frac{1}{\theta_{\partial,\mathrm{M}}(t)}\|s - s_{h\tau}\|^2 + K_{\mathrm{M}}|\boldsymbol{g}|^2\|\kappa'\|^2_{L^\infty([0,1])}\int_0^T \theta_{\partial,\mathrm{M}}(t)\, \|G^0_{h\tau}\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)}.\end{aligned} \tag{3.22}$$

Combining (3.18)–(3.22) with (3.14), one has

$$\|G^0_{h\tau}(T)\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)} + \int_0^T \frac{1}{\theta_{\partial,\mathrm{M}}(t)}\|(s-s_{h\tau})(t)\|^2$$
$$\leq \|s_0 - s_{h\tau}(0)\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} + \int_0^T \tfrac{1}{\lambda}\|\mathcal{R}\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} + \int_0^T (\lambda + \mathfrak{C}_1(t))\,\|G^0_{h\tau}(t)\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)}. \tag{3.23}$$

Applying the Gronwall Lemma

$$u(t) \leq \alpha(t) + \int_0^t \beta(\varrho)u(\varrho)\mathrm{d}\varrho \implies u(t) \leq \alpha(t) + \int_0^t \beta(\varrho)\,\alpha(\varrho)\exp\left(\int_\varrho^t \beta(r)\,\mathrm{d}r\right)\mathrm{d}\varrho \tag{3.24}$$

with $u(t) = \|G^0_{h\tau}(t)\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)}$, $\alpha(t) = \|s_0 - s_{h\tau}(0)\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} + \int_0^t \lambda^{-1}\|\mathcal{R}\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} - \int_0^t \frac{1}{\theta_{\partial,\mathrm{M}}}\|(s-s_{h\tau})\|^2$, $\beta(t) = \lambda + \mathfrak{C}_1(t)$, and re-normalizing both sides by dividing with $\exp(\int_0^T(\lambda+\mathfrak{C}_1))$, we have (3.12a). Observe that the total coefficient of $\|s_0 - s_{h\tau}(0)\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)}$, after cancellation of terms and subsequent division, becomes unity.

Step 2 (Estimate (3.12b)**:** We choose the test function $\varphi = s - s_{h\tau} \in \mathcal{X}$ in (3.14). Termwise, this gives

$$\int_0^T \langle \partial_t(s-s_{h\tau}), s-s_{h\tau}\rangle = \tfrac{1}{2}\|s(T)-s_{h\tau}(T)\|^2 - \tfrac{1}{2}\|s_0 - s_{h\tau}(0)\|^2, \tag{3.25}$$
$$\int_0^T \langle \mathcal{R}, s-s_{h\tau}\rangle \overset{(3.3)}{\leq} \int_0^T \|\mathcal{R}\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)}\|s-s_{h\tau}\|_{H^1_{\bar{\mathbf{K}}}(\Omega)}$$
$$\leq \int_0^T \tfrac{2}{D_{\mathrm{m}}(t)}\|\mathcal{R}\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} + \int_0^T \tfrac{D_{\mathrm{m}}(t)}{8}\|s-s_{h\tau}\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)}$$
$$\overset{(3.11)}{\leq} 2\int_0^T \tfrac{1}{D_{\mathrm{m}}(t)}\|\mathcal{R}\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} + \tfrac{1}{8}\int_0^T\int_\Omega D(s)|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(s-s_{h\tau})|^2, \tag{3.26}$$
$$\int_0^T (f(s,\boldsymbol{x},t) - f(s_{h\tau},\boldsymbol{x},t)), s-s_{h\tau}) \overset{(3.11)}{\leq} \int_0^T f_{\partial,\mathrm{M}}(t)\|s-s_{h\tau}\|^2, \tag{3.27}$$
$$\int_0^T (\bar{\mathbf{K}}\boldsymbol{g}(\kappa(s_{h\tau})-\kappa(s)), \nabla(s-s_{h\tau})) \leq \int_0^T \left[\tfrac{2\boldsymbol{g}^{\mathrm{T}}\bar{\mathbf{K}}\boldsymbol{g}}{D_{\mathrm{m}}(t)}\|\kappa(s)-\kappa(s_{h\tau})\|^2 + \tfrac{D_{\mathrm{m}}(t)}{8}\|s-s_{h\tau}\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)}\right]$$
$$\overset{(P3)}{\leq} 2K_{\mathrm{M}}|\boldsymbol{g}|^2\|\kappa'\|^2_{L^\infty([0,1])}\int_0^T \tfrac{1}{D_{\mathrm{m}}(t)}\|s-s_{h\tau}\|^2 + \tfrac{1}{8}\int_0^T\int_\Omega D(s)|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(s-s_{h\tau})|^2. \tag{3.28}$$

To estimate $\|\Psi - \Psi_{h\tau}\|_{L^2(0,T;H^1_{\bar{\mathbf{K}}}(\Omega))}$, we need to also consider the parabolic-elliptic degeneracy. Consider the domains $\Omega^1(t) := \{\boldsymbol{x}\in\Omega : s(\boldsymbol{x},t),\ s_{h\tau}(\boldsymbol{x},t) < 1\}$, $\Omega^2(t) := \{\boldsymbol{x}\in\Omega : s(\boldsymbol{x},t) = 1, s_{h\tau}(\boldsymbol{x},t) < 1\}$, $\Omega^3(t) := \{\boldsymbol{x}\in\Omega : s(\boldsymbol{x},t) < 1, s_{h\tau}(\boldsymbol{x},t) = 1\}$, and $\Omega^4(t) := \{\boldsymbol{x}\in\Omega : s(\boldsymbol{x},t) = s_{h\tau}(\boldsymbol{x},t) = 1\}$ where the equalities and inequalities are satisfied in an almost everywhere sense inside the domains. We divide accordingly the remaining term of (3.14)

$$\int_0^T (\bar{\mathbf{K}}\nabla(\Psi-\Psi_{h\tau}), \nabla(s-s_{h\tau})) = 2T_1 + T_2 + T_3 + T_4,$$

where the terms $T_{1,2,3,4}$ are explained below.

- Observing that $\theta(\Psi)$, $\theta(\Psi_{h\tau}) < 1$ a.e. in $\Omega^1(t)$, the first term T_1 is divided into two

parts

$$
\begin{aligned}
T_1 &:= \tfrac{1}{2}\int_0^T (\bar{\mathbf{K}}\nabla(\Psi-\Psi_{h\tau}),\nabla(s-s_{h\tau}))_{\Omega^1} = \tfrac{1}{2}\int_0^T (\bar{\mathbf{K}}\nabla(\Psi-\Psi_{h\tau}),\nabla(\theta(\Psi)-\theta(\Psi_{h\tau})))_{\Omega^1}\\
&= \tfrac{1}{2}\int_0^T (\bar{\mathbf{K}}\nabla(\Psi-\Psi_{h\tau}),(\theta'(\Psi)\nabla\Psi-\theta'(\Psi_{h\tau})\nabla\Psi_{h\tau}))_{\Omega^1}\\
&= \tfrac{1}{2}\int_0^T (\bar{\mathbf{K}}\nabla(\Psi-\Psi_{h\tau}),\theta'(\Psi)\nabla(\Psi-\Psi_{h\tau})+(\theta'(\Psi)-\theta'(\Psi_{h\tau}))\nabla\Psi_{h\tau})_{\Omega^1}\\
&\overset{(2.11)}{=} \tfrac{1}{2}\int_0^T\int_{\Omega^1}\tfrac{|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(\Psi-\Psi_{h\tau})|^2}{D(\theta(\Psi))} + \tfrac{1}{2}\int_0^T (\bar{\mathbf{K}}\nabla(\Psi-\Psi_{h\tau}),(\theta'(\Psi)-\theta'(\Psi_{h\tau}))\nabla\Psi_{h\tau})_{\Omega^1}. \quad (3.29a)
\end{aligned}
$$

The second term on the right is estimated as

$$
\begin{aligned}
&\tfrac{1}{2}\int_0^T (\bar{\mathbf{K}}\nabla(\Psi-\Psi_{h\tau}),(\theta'(\Psi)-\theta'(\Psi_{h\tau}))\nabla\Psi_{h\tau})_{\Omega^1}\\
&\overset{(2.11)}{=} -\tfrac{1}{2}\int_0^T (\bar{\mathbf{K}}\nabla(\Psi-\Psi_{h\tau}),\left(\tfrac{D(s)-D(s_{h\tau})}{D(s)D(s_{h\tau})}\right)\nabla\Psi_{h\tau})_{\Omega^1}\\
&= -\tfrac{1}{2}\int_0^T \left(\frac{1}{D(s)}\bar{\mathbf{K}}\nabla(\Psi-\Psi_{h\tau}),(D(s)-D(s_{h\tau}))\nabla s_{h\tau}\right)_{\Omega^1}\\
&\geq -\tfrac{1}{2}\int_0^T \|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla s_{h\tau}\|_{L^\infty(\Omega_1)}\int_{\Omega^1}\frac{1}{D(\theta(\Psi))}|D(s)-D(s_{h\tau})||\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(\Psi-\Psi_{h\tau}))|\\
&\geq -\tfrac{1}{4}\int_0^T \|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla s_{h\tau}\|^2_{L^\infty(\Omega)}\int_{\Omega^1}\frac{|D(s)-D(s_{h\tau})|^2}{D(\theta(\Psi))} - \tfrac{1}{4}\int_0^T\int_{\Omega^1}\tfrac{|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(\Psi-\Psi_{h\tau})|^2}{D(\theta(\Psi))}\\
&\geq -\int_0^T \tfrac{C^\infty_{h\tau}(t)D^2_M(t)}{4D_{\mathrm{m}}(t)}\int_{\Omega^1}|s-s_{h\tau}|^2 - \tfrac{1}{4}\int_0^T\int_{\Omega^1}\tfrac{|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(\Psi-\Psi_{h\tau})|^2}{D(\theta(\Psi))}. \quad (3.29b)
\end{aligned}
$$

Hence, we have

$$
T_1 \geq \int_0^T\int_{\Omega^1}\tfrac{|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(\Psi-\Psi_{h\tau})|^2}{4D(\theta(\Psi))} - \int_0^T \tfrac{C^\infty_{h\tau}(t)D^2_M(t)}{4D_{\mathrm{m}}(t)}\|s-s_{h\tau}\|^2. \quad (3.30)
$$

- We estimate T_1 once again. Recall that s, $s_{h\tau} < 1$ a.e. in $\Omega^1(t)$ implying $\Psi_{h\tau} = P_{\mathrm{c}}(s_{h\tau})$ and $\Psi = P_{\mathrm{c}}(s)$ in $\Omega^1(t)$. Note from (2.9) that $P_{\mathrm{c}}{}' = D$. Hence, we have

$$
\begin{aligned}
T_1 &= \tfrac{1}{2}\int_0^T (\bar{\mathbf{K}}\nabla(\Psi-\Psi_{h\tau}),\nabla(s-s_{h\tau}))_{\Omega^1} = \tfrac{1}{2}\int_0^T (\bar{\mathbf{K}}\nabla(P_{\mathrm{c}}(s)-P_{\mathrm{c}}(s_{h\tau})),\nabla(s-s_{h\tau}))_{\Omega^1}\\
&= \tfrac{1}{2}\int_0^T (\bar{\mathbf{K}}(D(s)\nabla s - D(s_{h\tau})\nabla s_{h\tau}),\nabla(s-s_{h\tau}))_{\Omega^1}\\
&= \tfrac{1}{2}\int_0^T\int_{\Omega^1} D(s)|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(s-s_{h\tau})|^2 + \tfrac{1}{2}\int_0^T (\bar{\mathbf{K}}(D(s)-D(s_{h\tau}))\nabla s_{h\tau},\nabla(s-s_{h\tau}))_{\Omega^1}. \quad (3.31)
\end{aligned}
$$

Similar to (3.29b), the second term is estimated as

$$
\begin{aligned}
&\tfrac{1}{2}\int_0^T ((D(s)-D(s_{h\tau}))\bar{\mathbf{K}}\nabla s_{h\tau},\nabla(s-s_{h\tau}))_{\Omega^1}\\
&\quad\geq -\int_0^T \tfrac{C^\infty_{h\tau}(t)D^2_{\mathrm{M}}(t)}{4D_{\mathrm{m}}(t)}\|s-s_{h\tau}\|^2 - \tfrac{1}{4}\int_0^T\int_{\Omega^1} D(s)|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(s-s_{h\tau})|^2. \quad (3.32)
\end{aligned}
$$

Hence, we have so far that

$$2T_1 \geq \tfrac{1}{4}\int_0^T\int_{\Omega^1}\left[\tfrac{|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(\Psi-\Psi_{h\tau})|^2}{D(\theta(\Psi))}+D(s)|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(s-s_{h\tau})|^2\right]-\int_0^T\tfrac{C^\infty_{h\tau}(t)D_M^2(t)}{2D_{\rm m}(t)}\,\|s-s_{h\tau}\|^2. \quad (3.33)$$

- Observe that $s=1$ in $\Omega^2(t)$ and $\theta'(\Psi_{h\tau})=1/D(\theta(\Psi_{h\tau}))$ using (2.11). Also, $\int_{\Omega^2}|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla\Psi|^2\leq\int_\Omega|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla[\Psi-P_{\rm M}]_+|^2$. Moreover, $\Omega^2(t)\subseteq\Omega^{\rm deg}(t)$ implying that $D(s)/2\leq D(s_{h\tau})\leq 2D(s)$ in $\Omega^2(t)$ from the assumptions of Theorem 3.4. Using these, we have

$$\begin{aligned}
T_2 &:= \int_0^T(\bar{\mathbf{K}}\nabla(\Psi-\Psi_{h\tau}),\nabla(s-s_{h\tau}))_{\Omega^2}=\int_0^T(\theta'(\Psi_{h\tau})\bar{\mathbf{K}}\nabla(\Psi_{h\tau}-\Psi),\nabla\Psi_{h\tau})_{\Omega^2}\\
&=\tfrac{1}{2}\int_0^T\int_{\Omega^2}\left[\theta'(\Psi_{h\tau})|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla\Psi_{h\tau}|^2+\theta'(\Psi_{h\tau})|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(\Psi_{h\tau}-\Psi)|^2-\theta'(\Psi_{h\tau})|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla\Psi|^2\right]\\
&=\tfrac{1}{2}\int_0^T\int_{\Omega^2}\left[D(s_{h\tau})|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla s_{h\tau}|^2+\tfrac{|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(\Psi-\Psi_{h\tau})|^2}{D(s_{h\tau})}\right]-\tfrac{1}{2}\int_0^T\int_{\Omega^2}\tfrac{|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla\Psi|^2}{D(s_{h\tau})}\\
&\geq\tfrac{1}{4}\int_0^T\int_{\Omega^2}D(s)|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(1-s_{h\tau})|^2+\int_0^T\int_{\Omega^2}\tfrac{|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(\Psi-\Psi_{h\tau})|^2}{4D(s)}-\int_0^T\int_\Omega\tfrac{|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla[\Psi-P_{\rm M}]_+|^2}{D(1)}. \quad (3.34)
\end{aligned}$$

In the above inequality, the identity $(a-b)a=\frac{1}{2}[a^2+(a-b)^2-b^2]$ has been used.

- With the same manipulations one has (note that $D(s)\leq 2D(1)$ in $\Omega^3(t)$ from the assumptions of Theorem 3.4)

$$\begin{aligned}
T_3 &:= \int_0^T(\bar{\mathbf{K}}\nabla(\Psi-\Psi_{h\tau}),\nabla(s-s_{h\tau}))_{\Omega^3}\\
&\geq\tfrac{1}{2}\int_0^T\int_{\Omega^3}D(s)|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(1-s)|^2+\int_0^T\int_{\Omega^3}\tfrac{|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(\Psi-\Psi_{h\tau})|^2}{2D(s)}-\int_0^T\int_\Omega\tfrac{|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla[\Psi_{h\tau}-P_{\rm M}]_+|^2}{D(1)}. \quad (3.35)
\end{aligned}$$

- Finally, in $\Omega^4(t)$ one has $s=s_{h\tau}=1$, thus giving

$$T_4 := \int_0^T(\bar{\mathbf{K}}\nabla(\Psi-\Psi_{h\tau}),\nabla(s-s_{h\tau}))_{\Omega^4}=0. \quad (3.36)$$

With this, we have

$$\begin{aligned}
2T_1+T_2+T_3+T_4 &\geq \tfrac{1}{4}\int_0^T\int_\Omega\left[\tfrac{|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(\Psi-\Psi_{h\tau})|^2}{D(\theta(\Psi))}+D(s)|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(s-s_{h\tau})|^2\right]\\
&\quad-\int_0^T\tfrac{C^\infty_{h\tau}(t)D_M^2(t)}{2D_{\rm m}(t)}\,\|s-s_{h\tau}\|^2-\tfrac{1}{D(1)}\int_0^T\left(\|[\Psi-P_{\rm M}]_+\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)}+\|[\Psi_{h\tau}-P_{\rm M}]_+\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)}\right). \quad (3.37)
\end{aligned}$$

- To estimate $\|[\Psi-P_{\rm M}]_+\|_{H^1_{\bar{\mathbf{K}}}(\Omega)}$ insert $\varphi=[\Psi-P_{\rm M}]_+$ in (2.16). Note that $\partial_t\theta(\Psi)=0$ and $f(s,\boldsymbol{x},t)=f(1,\boldsymbol{x},t)$ if $\Psi>P_{\rm M}$. Also, $\int_0^T(\boldsymbol{c},\nabla[\Psi-P_{\rm M}]_+)=\int_0^T\int_{\partial\Omega}\boldsymbol{c}\cdot\hat{\boldsymbol{n}}_{\partial\Omega}[\Psi-P_{\rm M}]_+=0$ for the constant vector $\boldsymbol{c}=\int_{\Omega^{\rm deg}(t)}\bar{\mathbf{K}}\boldsymbol{g}$. Moreover, $f[\Psi-P_{\rm M}]_+\leq[f]_+[\Psi-P_{\rm M}]_+$. Using these relations leads to

$$\begin{aligned}
\int_0^T\|[\Psi-P_{\rm M}]_+\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)} &= \int_0^T(f(1,\boldsymbol{x},t),[\Psi-P_{\rm M}]_+)-\int_0^T(\bar{\mathbf{K}}\boldsymbol{g},\nabla[\Psi-P_{\rm M}]_+)\\
&\leq\int_0^T([f(1,\boldsymbol{x},t)]_+,[\Psi-P_{\rm M}]_+)-\int_0^T(\bar{\mathbf{K}}\boldsymbol{g}-\tfrac{1}{|\Omega^{\rm deg}|}\int\limits_{\Omega^{\rm deg}}\bar{\mathbf{K}}\boldsymbol{g},\nabla[\Psi-P_{\rm M}]_+)\\
&\leq\int_0^T\left(\|[f(1,\boldsymbol{x},t)]_+\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega^{\rm deg})}+\|\bar{\mathbf{K}}^{-\frac{1}{2}}(\bar{\mathbf{K}}\boldsymbol{g}-\tfrac{1}{|\Omega^{\rm deg}|}\int\limits_{\Omega^{\rm deg}}\bar{\mathbf{K}}\boldsymbol{g})\|_{\Omega^{\rm deg}}\right)\|[\Psi-P_{\rm M}]_+\|_{H^1_{\bar{\mathbf{K}}}(\Omega)}.
\end{aligned}$$

Using Young's inequality on the right hand side and recalling the definition of η^{deg} we estimate

$$\tfrac{1}{D(1)}\int_0^T \left(\|[\Psi - P_{\mathrm{M}}]_+\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)} + \|[\Psi_{h\tau} - P_{\mathrm{M}}]_+\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)} \right) \leq \tfrac{1}{2}\int_0^T [\eta^{\mathrm{deg}}]^2. \tag{3.38}$$

- Combining all the estimates above, one obtains

$$\begin{aligned} &\|s(T) - s_{h\tau}(T)\|^2 + \tfrac{1}{2}\int_0^T \int_\Omega \tfrac{|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(\Psi - \Psi_{h\tau})|^2}{D(\theta(\Psi))} \\ &\leq \|s_0 - s_{h\tau}(0)\|^2 + 4\int_0^T \tfrac{1}{D_{\mathrm{m}}(t)}\|\mathcal{R}\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} + \int_0^T [\eta^{\mathrm{deg}}]^2 + \int_0^T \mathfrak{C}_2(t)\|s - s_{h\tau}\|^2. \end{aligned} \tag{3.39}$$

Since $\mathfrak{C}_2(t) > 0$, one has (3.12b) from applying the Gronwall Lemma (3.24), where $u(t) = \|s(t) - s_{h\tau}(t)\|^2$, $\beta(t) = \mathfrak{C}_2(t)$ and

$$\alpha(t) = \|s_0 - s_{h\tau}(0)\|^2 + 4\int_0^t \tfrac{1}{D_{\mathrm{m}}(t)}\|\mathcal{R}\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} + \int_0^t [\eta^{\mathrm{deg}}]^2 - \tfrac{1}{2}\int_0^t \int_\Omega \tfrac{|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(\Psi - \Psi_{h\tau})|^2}{D(\theta(\Psi))}.$$

Step 3 (Estimate (3.12c)**):** Using the definition of $H^{\pm 1}_{\bar{\mathbf{K}}}$-norms in (3.14), we have

$$\int_0^t \|\partial_t(s - s_{h\tau})\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} \leq 3\int_0^t \left[\|\Psi - \Psi_{h\tau}\|^2_{H^1_{\bar{\mathbf{K}}}(\Omega)} + \|\mathcal{R}\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} + \mathfrak{C}_3(T)\|s - s_{h\tau}\|^2 \right],$$

for any $t \in (0, T]$. Multiplying the above inequality with $\lambda(t)\exp(\int_t^T \lambda)$, integrating on $[0, T]$, and adding the above inequality for $t = T$, we get from the first term

$$\begin{aligned} &\int_0^T \left[\|\partial_t(s - s_{h\tau})\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} + \lambda(t)\exp\left(\int_t^T \lambda\right)\int_0^t \|\partial_t(s - s_{h\tau})\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} \right] \\ &\overset{(3.6)}{=} \exp\left(\int_0^T \lambda\right)\mathcal{J}_\lambda\left(\|\partial_t(s - s_{h\tau})\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)}\right)^2, \end{aligned}$$

and similar for the other terms. The estimate (3.12c) follows then by cancelling the $\exp\left(\int_0^T \lambda\right)$ multipliers. □

Remark 3.8 (Upper bound on $\|D(s)^{\frac{1}{2}}\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(s - s_{h\tau})\|$)**.** *From the step 2 of the proof of Theorem 3.4, it is evident that the error component $\|D(s)^{\frac{1}{2}}\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(s - s_{h\tau})\|$ can be estimated as well through slight changes in coefficients of the right hand side of* (3.12b)*. However, to have symmetry between the lower and the upper bounds of Theorems 3.2 and 3.4, this has not been pursued.*

4 Finite element discretization

We describe in this section the discretization of the Richards problem (1.1) by the finite element method.

4.1 Time steps

For the time-interval $(0, T)$, we introduce $N + 1$ discrete times $\boldsymbol{t}_N = (t_n)_{n=0}^N$ where $t_0 = 0 < t_1 < \cdots < t_n < \cdots < t_N = T$. Let $I_n := (t_{n-1}, t_n]$ denote the time intervals and $\tau_n := t_n - t_{n-1}$ the lengths of the time steps for $n \in \{1, \ldots, N\}$. Note that, we allow nonuniform time stepping. Further, for a vector space V, $\mathcal{Q}_1(I_n; V)$ denotes the space of V-valued affine functions over the time-step interval I_n.

4.2 Space meshes

For the time sequence $\boldsymbol{t}_N$, let $\{\mathcal{T}_n\}_{n=1}^N$ denote the sequence of matching and uniformly shape regular simplical meshes for the domain Ω. The meshes are allowed to undergo refinement or coarsening between time steps. Henceforth, discontinuities of $\bar{\mathbf{K}}$ are only allowed to happen along internal edges of the mesh. For each element $K \in \mathcal{T}_n$, let $h_K := \mathrm{diam}\{K\}$ denote the diameter of K and let $\mathfrak{p}_n \geq 1$ denote the spatial polynomial degree associated with $\mathcal{T}_n$. Our results are generalizable to polynomial degrees depending locally on $K \in \mathcal{T}_n$. However, to keep the notation simple, we only consider $\mathfrak{p}_n$ changing between time steps here. For full *hp*-adaptive algorithm, we refer to [17].

4.3 Approximation spaces

On a time step $n \in \{1, \ldots, N\}$, we define the $H_0^1(\Omega)$-conforming *hp*-finite element space $V_{n,h}$ as:

$$V_{n,h} := \left\{ u_h \in H_0^1(\Omega),\ u_h|_K \in \mathcal{P}_{\mathfrak{p}_n}(K) \quad \forall\, K \in \mathcal{T}_n \right\}, \tag{4.1}$$

where $\mathcal{P}_{\mathfrak{p}_n}(K)$ denotes the polynomial space of degree $\mathfrak{p}_n \in \mathbb{N}$ on K. Further, let $\Pi_{n,h} : L^2(\Omega) \to V_{n,h}$ and $\Lambda_{n,h} : L^2(\Omega) \to \mathcal{P}_{\mathfrak{p}_n}(\mathcal{T}_n)$ represent the L^2-orthogonal projection operator with respect to the spaces $V_{n,h}$ and $\mathcal{P}_{\mathfrak{p}_n}(\mathcal{T}_n)$, i.e.,

$$\Pi_{n,h} u \in V_{n,h} \text{ for } u \in L^2(\Omega) \text{ is such that } (\Pi_{n,h} u, \varphi_h) = (u, \varphi_h), \text{ for all } \varphi_h \in V_{n,h}; \tag{4.2a}$$

$$\Lambda_{n,h} u \in \mathcal{P}_{\mathfrak{p}_n}(\mathcal{T}_n) \text{ for } u \in L^2(\Omega) \text{ is such that } (\Lambda_{n,h} u, \varphi_h) = (u, \varphi_h), \text{ for all } \varphi_h \in \mathcal{P}_{\mathfrak{p}_n}(\mathcal{T}_n). \tag{4.2b}$$

4.4 Finite element discretization

In Section 3, formulation (2.16) is used to derive the estimates. However, since (2.15) is the most general and commonly used formulation, we propose the finite element scheme for (2.15). We will still be able to apply the analysis of Section 3. For time discretization, we consider the backward Euler scheme. The problem for each $n \in \{1, \ldots, N\}$ and a given $S_{n-1,h} \in L^2(\Omega)$ is to find $p_{n,h} \in V_{n,h}$ which satisfies for all $\varphi_h \in V_{n,h}$,

$$\tfrac{1}{\tau_n}(S(p_{n,h}) - S_{n-1,h}, \varphi_h) + (\bar{\mathbf{K}}\kappa(S(p_{n,h}))[\nabla p_{n,h} + \boldsymbol{g}], \nabla \varphi_h) = (f(S(p_{n,h})), \boldsymbol{x}, t_n), \varphi_h). \tag{4.3}$$

For $n = 1$, we set $S_{n-1,h} := \Pi_{1,h} s_0$, whereas, for $n > 1$, $S_{n-1,h} := S(p_{n-1,h})$. The existence of $p_{n,h}$ solving (4.3) is discussed in [15] for the nondegenerate case ($\kappa(0) > 0$). The degenerate case is covered in [34] for the control volume finite element method. In practice, since the problem (4.3) is nonlinear, the exact $p_{n,h}$ is generally not known, and linear iterations have to be used to approximate $p_{n,h}$. This is discussed at length in Appendix A.

From the sequence $\{p_{n,h}\}_{n=1}^N$, we define the space–time discrete total pressure and saturation for all $n \in \{1, \ldots, N\}$ as

$$\Psi_{n,h} := \mathcal{K}(p_{n,h}) \in H_0^1(\Omega) \text{ and } S_{n,h} := \theta(\Psi_{n,h}) \overset{(2.14)}{=} S(p_{n,h}) \in H^1(\Omega). \tag{4.4}$$

The choice $\Psi_{0,h} = P_{\mathrm{c}}(S_{0,h}) = P_{\mathrm{c}}(\Pi_{1,h} s_0)$ is used for extending the definition of $\Psi_{n,h}$ to $n = 0$. We stress from the above that the inverse of the Kirchhoff transform $\mathcal{K}^{-1}$ (see (2.13)) does not need to be evaluated for the above scheme.

4.5 Time-continuous solutions

There are multiple ways to define a time-continuous total pressure $\Psi_{h\tau} \in \mathcal{X}$ and saturation $s_{h\tau} \in \mathcal{Y}$, satisfying the requirements of Theorems 3.2 and 3.4, starting from $\{S_{n,h}\}_{n=1}^N$ and $\{\Psi_{n,h}\}_{n=1}^N$ introduced in (4.4). Here, for $t \in I_n$, we choose

$$\Psi_{h\tau}(t) := P_{\mathrm{c}}\left(\tfrac{t-t_{n-1}}{\tau_n} S_{n,h} + \tfrac{t_n-t}{\tau_n} S_{n-1,h}\right) + \left[\tfrac{t-t_{n-1}}{\tau_n}\Psi_{n,h} + \tfrac{t_n-t}{\tau_n}\Psi_{n-1,h} - P_{\mathrm{M}}\right]_+, \tag{4.5a}$$

$$s_{h\tau}(t) := \theta(\Psi_{h\tau}(t)). \tag{4.5b}$$

Observe that, $\Psi_{h\tau}$ and $s_{h\tau}$ defined this way satisfy

$$\Psi_{h\tau} \in C(0,T;H_0^1(\Omega)) \subset \mathcal{X}, \qquad \text{and } s_{h\tau} \in W^{1,\infty}(0,T;H^1(\Omega)) \subset \mathcal{Y}; \tag{4.6a}$$

$$\Psi_{h\tau}(t_n) = \Psi_{n,h}, \qquad \text{and } s_{h\tau}(t_n) = S_{n,h}, \quad \forall\, n \in \{1,\dots,N\}. \tag{4.6b}$$

The relation (4.6b) even holds when $\Psi_{n,h} > P_{\mathrm{M}}$ since $P_{\mathrm{c}}(S_{n,h}) = P_{\mathrm{M}}$ in this case and the other contribution from (4.5a) adds $[\Psi_{n,h} - P_{\mathrm{M}}]_+$. Another advantage of this interpolation is that, using (2.10), if both $\Psi_{n,h}$, $\Psi_{n-1,h} \le P_{\mathrm{M}}$ (nondegenerate case), or $\Psi_{n,h}$, $\Psi_{n-1,h} \ge P_{\mathrm{M}}$ (degenerate case), i.e.,

$$\text{if either } \Psi_{h\tau} \le P_{\mathrm{M}} \text{ or } \Psi_{h\tau} \ge P_{\mathrm{M}} \text{ in } I_n, \text{ then } \partial_t s_{h\tau} = \tfrac{1}{\tau_n}(S_{n,h} - S_{n-1,h}). \tag{4.7}$$

5 A posteriori error estimates

We apply here the developments of Section 3 to perform a posteriori error analysis of the finite element discretization of Section 4.

5.1 Equilibrated flux

The objective of this section is to design an equilibrated flux $\boldsymbol{\sigma}_{n,h} \in \boldsymbol{H}(\mathrm{div},\Omega)$ that satisfies the mass balance property

$$\int_K \left[\tfrac{1}{\tau_n}(S_{n,h} - S_{n-1,h}) + \nabla\cdot\boldsymbol{\sigma}_{n,h} - f(S_{n,h},\boldsymbol{x},t_n)\right] = 0 \text{ for all } K \in \mathcal{T}_n. \tag{5.1}$$

5.1.1 Local mixed finite element spaces

For the construction of $\boldsymbol{\sigma}_{n,h}$, we introduce some standard mixed finite element spaces. For each $n \in \{1,\dots,N\}$, let $\mathcal{V}_n$ denote the set of vertices of the mesh $\mathcal{T}_n$, where we distinguish the set of interior vertices $\mathcal{V}_n^{\mathrm{int}}$ and the set of boundary vertices $\mathcal{V}_n^{\mathrm{ext}}$. For $K \in \mathcal{T}_n$, $\mathcal{V}_K \subset \mathcal{V}_n$ denotes the set of vertices of K. For each $\mathbf{a} \in \mathcal{V}_n$, let $\psi_{\mathbf{a}}$ denote the hat function associated with $\mathbf{a}$ and $\omega_{\mathbf{a}}$ the interior of the support of $\psi_{\mathbf{a}}$, with the associated diameter $h_{\omega_{\mathbf{a}}}$. Furthermore, let $\mathcal{T}_n^{\mathbf{a}}$ denote the restriction of the mesh $\mathcal{T}_n$ to $\omega_{\mathbf{a}}$.

For a polynomial degree $\mathfrak{p} \ge 0$, the local spaces $\mathcal{P}_{\mathfrak{p}}(\mathcal{T}_n^{\mathbf{a}})$ and $\mathbf{RTN}_{\mathfrak{p}}(\mathcal{T}_n^{\mathbf{a}})$ are defined by

$$\mathcal{P}_{\mathfrak{p}}(\mathcal{T}_n^{\mathbf{a}}) := \{u_h \in L^2(\omega_{\mathbf{a}}), \quad u_h|_K \in \mathcal{P}_{\mathfrak{p}}(K) \quad \forall\, K \in \mathcal{T}_n^{\mathbf{a}}\},$$
$$\mathbf{RTN}_{\mathfrak{p}}(\mathcal{T}_n^{\mathbf{a}}) := \{\boldsymbol{v}_h \in \boldsymbol{L}^2(\omega_{\mathbf{a}};\mathbb{R}^d), \quad \boldsymbol{v}_h|_K \in \mathbf{RTN}_{\mathfrak{p}}(K) \quad \forall K \in \mathcal{T}_n^{\mathbf{a}}\},$$

where $\mathbf{RTN}_{\mathfrak{p}}(K) := \mathcal{P}_{\mathfrak{p}}(K;\mathbb{R}^d) + \mathcal{P}_{\mathfrak{p}}(K)\boldsymbol{x}$ denotes the Raviart–Thomas–Nédélec space of order $\mathfrak{p}$ on K. We use a similar notation on the whole mesh $\mathcal{T}_n$, and introduce the local mixed finite

element spaces $\boldsymbol{V}_{n,h}^{\mathbf{a}}$ and $Q_{n,h}^{\mathbf{a}}$ as

$$\boldsymbol{V}_{n,h}^{\mathbf{a}} := \begin{cases} \{\boldsymbol{v}_h \in \mathbf{RTN}_{\mathfrak{p}_n+1}(\mathcal{T}_n^{\mathbf{a}}),\ \boldsymbol{v}_h \in \boldsymbol{H}(\mathrm{div}, \omega_{\mathbf{a}}),\ \boldsymbol{v}_h \cdot \boldsymbol{n} = 0 \text{ on } \partial\omega_{\mathbf{a}}\} \text{ if } \mathbf{a} \in \mathcal{V}_n^{\mathrm{int}}, \\ \{\boldsymbol{v}_h \in \mathbf{RTN}_{\mathfrak{p}_n+1}(\mathcal{T}_n^{\mathbf{a}}),\ \boldsymbol{v}_h \in \boldsymbol{H}(\mathrm{div}, \omega_{\mathbf{a}}),\ \boldsymbol{v}_h \cdot \boldsymbol{n} = 0 \text{ on } \partial\omega_{\mathbf{a}} \setminus \partial\Omega\} \text{ if } \mathbf{a} \in \mathcal{V}_n^{\mathrm{ext}}, \end{cases}$$
$$Q_{n,h}^{\mathbf{a}} := \begin{cases} \{u_h \in \mathcal{P}_{\mathfrak{p}_n+1}(\mathcal{T}_n^{\mathbf{a}}), \quad (u_h, 1)_{\omega_{\mathbf{a}}} = 0\} & \text{if } \mathbf{a} \in \mathcal{V}_n^{\mathrm{int}}, \\ \mathcal{P}_{\mathfrak{p}_n+1}(\mathcal{T}_n^{\mathbf{a}}) & \text{if } \mathbf{a} \in \mathcal{V}_n^{\mathrm{ext}}. \end{cases} \tag{5.2}$$

The projector $\Pi_{n,h}^{\mathrm{RT}} : \boldsymbol{L}^2(\Omega; \mathbb{R}^d) \to \mathbf{RTN}_{\mathfrak{p}_n}(\mathcal{T}_n)$ is then defined as:

$$\text{for } \boldsymbol{u} \in \boldsymbol{L}^2(\Omega; \mathbb{R}^d), \quad (\bar{\mathbf{K}} \Pi_{n,h}^{\mathrm{RT}} \boldsymbol{u}, \boldsymbol{v}_h) = (\bar{\mathbf{K}}\, \boldsymbol{u}, \boldsymbol{v}_h), \ \text{ for all } \boldsymbol{v}_h \in \mathbf{RTN}_{\mathfrak{p}_n}(\mathcal{T}_n). \tag{5.3}$$

Note that it is computed elementwise.

5.1.2 Flux reconstruction

For each $n \in \{1, \dots, N\}$, we unify the numerical source-like terms and flux-like terms of (4.3) in $\mathcal{G}_{n,h} \in L^2(\Omega)$ and $\boldsymbol{F}_{n,h} \in \boldsymbol{L}^2(\Omega; \mathbb{R}^d)$,

$$\mathcal{G}_{n,h} := f(S_{n,h}, \boldsymbol{x}, t_n) - \tfrac{1}{\tau_n}(S_{n,h} - S_{n-1,h}), \quad \boldsymbol{F}_{n,h} := \nabla \Psi_{n,h} + \boldsymbol{g}\kappa(S_{n,h}). \tag{5.4}$$

Observe that the terms defined above are constant in time in I_n. Recalling the projection operators $\Pi_{n,h}$, $\Lambda_{n,h}$, and $\Pi_{n,h}^{\mathrm{RT}}$ from (4.2) and (5.3), the scalar function $g_{n,h}^{\mathbf{a}} \in \mathcal{P}_{\mathfrak{p}_n+1}(\mathcal{T}_n^{\mathbf{a}})$ and the vector field $\boldsymbol{\tau}_{n,h}^{\mathbf{a}} \in \mathbf{RTN}_{\mathfrak{p}_n+1}(\mathcal{T}_n^{\mathbf{a}})$ are defined as

$$g_{n,h}^{\mathbf{a}} := (\psi_{\mathbf{a}}\, \Lambda_{n,h} \mathcal{G}_{n,h} - \nabla \psi_{\mathbf{a}} \cdot\, \bar{\mathbf{K}}\, \Pi_{n,h}^{\mathrm{RT}}\, \boldsymbol{F}_{n,h})|_{\omega_{\mathbf{a}}}, \quad \boldsymbol{\tau}_{n,h}^{\mathbf{a}} := -(\psi_{\mathbf{a}} \bar{\mathbf{K}}\, \Pi_{n,h}^{\mathrm{RT}}\, \boldsymbol{F}_{n,h})|_{\omega_{\mathbf{a}}}. \tag{5.5}$$

Since $\psi_{\mathbf{a}} \in V_{n,h}$, using $\varphi_h = \psi_{\mathbf{a}}$ in (4.3) we get directly for all $\mathbf{a} \in \mathcal{V}_n^{\mathrm{int}}$ that $(g_{n,h}^{\mathbf{a}}, 1)_{\omega_{\mathbf{a}}} = 0$.

Definition 5.1 (Equilibrated flux $\boldsymbol{\sigma}_{n,h}$)**.** For a given time-step $n \in \{1, \dots, N\}$ and for each vertex $\mathbf{a} \in \mathcal{V}_n$, let the mixed finite element spaces $\boldsymbol{V}_{n,h}^{\mathbf{a}}$ and $Q_{n,h}^{\mathbf{a}}$ be defined by (5.2). For the time discrete solutions introduced in Section 4.4, let $\mathcal{G}_{n,h}$ and $\boldsymbol{F}_{n,h}$ be defined in (5.4). Let $g_{n,h}^{\mathbf{a}}$ and $\boldsymbol{\tau}_{n,h}^{\mathbf{a}}$ be defined by (5.5). Furthermore, let $\boldsymbol{\sigma}_{n,h}^{\mathbf{a}} \in \boldsymbol{V}_{n,h}^{\mathbf{a}}$ be defined by

$$\boldsymbol{\sigma}_{n,h}^{\mathbf{a}} := \operatorname*{argmin}_{\substack{\boldsymbol{v}_h \in \boldsymbol{V}_{n,h}^{\mathbf{a}}, \\ \nabla \cdot \boldsymbol{v}_h = g_{n,h}^{\mathbf{a}}}} \|\bar{\mathbf{K}}^{-\frac{1}{2}}(\boldsymbol{v}_h - \boldsymbol{\tau}_{n,h}^{\mathbf{a}})\|_{\omega_{\mathbf{a}}}. \tag{5.6}$$

Then, after extending $\boldsymbol{\sigma}_{n,h}^{\mathbf{a}}$ by zero from $\omega_{\mathbf{a}}$ to Ω for each $\mathbf{a} \in \mathcal{V}_n$, we define the equilibriated flux as

$$\boldsymbol{\sigma}_{n,h} := \sum_{\mathbf{a} \in \mathcal{V}_n} \boldsymbol{\sigma}_{n,h}^{\mathbf{a}}. \tag{5.7}$$

The well-posedness of $\boldsymbol{\sigma}_{n,h}$ follows from Theorem 4.2 of [17], see also references therein, and it satisfies (5.1) since

$$\nabla \cdot \boldsymbol{\sigma}_{n,h} \overset{(5.7)}{=} \sum_{\mathbf{a} \in \mathcal{V}_n} \nabla \cdot \boldsymbol{\sigma}_{n,h}^{\mathbf{a}} \overset{(5.5)}{=} \sum_{\mathbf{a} \in \mathcal{V}_n} [(\psi_{\mathbf{a}}\, \Lambda_{n,h} \mathcal{G}_{n,h} - \nabla \psi_{\mathbf{a}} \cdot\, \bar{\mathbf{K}}\, \Pi_{n,h}^{\mathrm{RT}}\, \boldsymbol{F}_{n,h})] = \Lambda_{n,h} \mathcal{G}_{n,h}, \tag{5.8a}$$

$$\text{and consequently } (\nabla \cdot \boldsymbol{\sigma}_{n,h} - \mathcal{G}_{n,h}, \varphi_h)_K \overset{(4.2)}{=} 0, \quad \forall K \in \mathcal{T}_n \text{ and } \varphi_h \in \mathcal{P}_{\mathfrak{p}_n}(\mathcal{T}_n). \tag{5.8b}$$

Here, the partition of unity property, $\sum_{\mathbf{a} \in \mathcal{V}_K} \psi_{\mathbf{a}} = 1$ is used. Practically, $\boldsymbol{\sigma}_{n,h}^{\mathbf{a}}$ are computed by solving the following mixed finite element problems [17] locally in $\omega_{\mathbf{a}}$: find $\boldsymbol{\sigma}_{n,h}^{\mathbf{a}} \in \boldsymbol{V}_{n,h}^{\mathbf{a}}$ and $r_{n,h}^{\mathbf{a}} \in Q_{n,h}^{\mathbf{a}}$ such that

$$\begin{aligned} &(\bar{\mathbf{K}}^{-1} \boldsymbol{\sigma}_{n,h}^{\mathbf{a}}, \boldsymbol{v}_h)_{\omega_{\mathbf{a}}} - (\nabla \cdot \boldsymbol{v}_h, r_{n,h}^{\mathbf{a}}) = (\bar{\mathbf{K}}^{-1} \boldsymbol{\tau}_{n,h}^{\mathbf{a}}, \boldsymbol{v}_h)_{\omega_{\mathbf{a}}}, \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_{n,h}^{\mathbf{a}}, \\ &(\nabla \cdot \boldsymbol{\sigma}_{n,h}^{\mathbf{a}}, u_h)_{\omega_{\mathbf{a}}} = (g_{n,h}^{\mathbf{a}}, u_h)_{\omega_{\mathbf{a}}}, \quad \forall u_h \in Q_{n,h}^{\mathbf{a}}. \end{aligned}$$

5.2 A posterori error estimators

Recalling the definition of time-continuous solutions $(\Psi_{h\tau}, s_{h\tau})$ from Section 4.5, we introduce the following a posteriori error estimators: Take $n \in \{1, \dots, N\}$, an open polytope $\omega \subseteq \Omega$, and $t \in I_n$. Then,

$$\eta^{\mathrm{F}}_{n,h,\omega}(t) := \|\bar{\mathbf{K}}^{-\frac{1}{2}} \boldsymbol{\sigma}_{n,h} + \bar{\mathbf{K}}^{\frac{1}{2}} (\nabla \Psi_{h\tau} + \boldsymbol{g}\, \kappa(s_{h\tau}))(t)\|_{\omega} \tag{5.9a}$$

measures the lack of $\boldsymbol{H}(\mathrm{div}, \Omega)$-conformity of the numerical flux $\bar{\mathbf{K}}(\nabla \Psi_{h\tau} + \boldsymbol{g}\, \kappa(s_{h\tau}))$. The quadrature error estimator arising from $\mathcal{G}_{n,h}$ not being polynomial (see (5.4)) is

$$\eta^{\mathrm{qd},\mathcal{G}}_{n,h,\omega} := \frac{h_\omega}{\sqrt{K_{\mathrm{m}}}\pi} \|\mathcal{G}_{n,h} - \Lambda_{n,h}\mathcal{G}_{n,h}\|_{\omega}. \tag{5.9b}$$

The time-quadrature error of $\partial_t s_{h\tau}$ is measured by the estimator

$$\eta^{\mathrm{qd},t}_{n,h,\omega}(t) := \|\partial_t s_{h\tau} - \tfrac{1}{\tau_n}(S_{n,h} - S_{n-1,h})\|_{H^{-1}_{\bar{\mathbf{K}}}(\omega)}. \tag{5.9c}$$

Observe that it estimates quadrature since $\int_{I_n} \partial_t s_{h\tau} \overset{(4.6\mathrm{b})}{=} S_{n,h} - S_{n-1,h}$, and it vanishes in both purely degenerate and nondegenerate regimes due to (4.7). The temporal oscillation in data f is measured by

$$\eta^{\mathrm{osc}}_{n,\omega}(t) := \|f(s_{h\tau}(t_n), \boldsymbol{x}, t_n) - f(s_{h\tau}(t), \boldsymbol{x}, t)\|_{H^{-1}_{\bar{\mathbf{K}}}(\omega)}. \tag{5.9d}$$

The errors in the approximation of the initial condition s_0 are accounted by

$$\eta^{\mathrm{ini},L^2} := \|s_0 - \Pi_{1,h} s_0\|, \quad \eta^{\mathrm{ini},H^{-1}} := \|s_0 - \Pi_{1,h} s_0\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)}. \tag{5.9e}$$

The projectors $\Pi_{n,h}$ and $\Lambda_{n,h}$ were defined in (4.2), and the norm $\|\cdot\|_{H^{-1}_{\bar{\mathbf{K}}}(\cdot)}$ was introduced in (3.3a). With the above definitions, the total estimator is computed as

$$\eta_{\mathcal{R}}(t) := \left[\sum_{K \in \mathcal{T}_n} [\eta^{\mathrm{F}}_{n,h,K}(t) + \eta^{\mathrm{qd},\mathcal{G}}_{n,h,K}]^2\right]^{\frac{1}{2}} + \eta^{\mathrm{qd},t}_{n,h,\Omega}(t) + \eta^{\mathrm{osc}}_{n,\Omega}(t). \tag{5.9f}$$

Remark 5.2 (Inverse of the Kirchhoff transform)**.** *We observe from the above that the inverse of the Kirchhoff transform $\mathcal{K}^{-1}$ (see* (2.13)*) does not need to be evaluated for computing the estimators.*

5.3 Global reliability

Complementing Theorem 3.4, our a posteriori error estimate on the error in the finite element discretization (4.3) of the Richards equation (1.1) is

Theorem 5.3 (Global reliability)**.** *Recall the definitions and assumptions stated in Theorem 3.4. Let $\{\Psi_{n,h}\}_{n=1}^N \subset H^1_0(\Omega)$ and $\{S_{n,h}\}_{n=1}^N \subset H^1(\Omega)$ be defined using the finite element discretization* (4.3)–(4.4) *and let $\Psi_{h\tau} \in C(0,T; H^1_0(\Omega)) \subset \mathcal{X}$ with $s_{h\tau} = \theta(\Psi_{h\tau}) \in W^{1,\infty}(0,T; H^1(\Omega)) \subset \mathcal{Y}$ be their time-continuous interpolates as defined in* (4.5)*. Let the a posteriori error estimators be defined in* (5.9)*. Then, for any time $t \in [0,T]$,*

$$\|\mathcal{R}(\Psi_{h\tau}(t))\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} \le \eta_{\mathcal{R}}(t). \tag{5.10a}$$

Consequently, the errors of $s_{h\tau}$ and $\Psi_{h\tau}$ satisfy:

$$\begin{aligned}\mathcal{E}_{L^2}^2 &:= e^{-\int_0^T(\lambda+\mathfrak{C}_1)}\|(s-s_{h\tau})(T)\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} + \mathcal{J}_{\lambda+\mathfrak{C}_1}(\theta_{\partial,\mathrm{M}}^{-\frac{1}{2}}\|s-s_{h\tau}\|)^2\\ &\leq [\eta^{\mathrm{ini},H^{-1}}]^2 + \mathcal{J}_{\lambda+\mathfrak{C}_1}(\lambda^{-\frac{1}{2}}\eta_{\mathcal{R}})^2 =: \eta_{L^2}^2, \qquad (5.10b)\\ \mathcal{E}_{H^1}^2 &:= e^{-\int_0^T \mathfrak{C}_2}\|(s-s_{h\tau})(T)\|^2 + \tfrac{1}{2}\mathcal{J}_{\mathfrak{C}_2}(\|D(s)^{-\frac{1}{2}}\bar{\mathbf{K}}^{\frac{1}{2}}\nabla(\Psi-\Psi_{h\tau})\|)^2\\ &\leq [\eta^{\mathrm{ini},L^2}]^2 + \mathcal{J}_{\mathfrak{C}_2}\left(\eta^{\mathrm{deg}}\right)^2 + 4\,\mathcal{J}_{\mathfrak{C}_2}\left(D_{\mathrm{m}}^{-\frac{1}{2}}\eta_{\mathcal{R}}\right)^2 =: \eta_{H^1}^2. \qquad (5.10c)\end{aligned}$$

Proof. From the regularity of $\Psi_{h\tau}$ and $\partial_t s_{h\tau} \in L^\infty(0,T;L^2(\Omega))$ one immediately has that $\mathcal{R}(\Psi_{h\tau}) \in L^\infty(0,T;H^{-1}(\Omega))$. Hence, for all $t \in I_n$, $n \in \{1,\dots,N\}$, adding and subtracting $(\boldsymbol{\sigma}_{n,h}, \nabla\varphi)$,

$$\begin{aligned}&\langle \mathcal{R}(\Psi_{h\tau}),\, \varphi\rangle = (f(s_{h\tau},\boldsymbol{x},t) - \partial_t s_{h\tau} - \nabla\cdot\boldsymbol{\sigma}_{n,h}, \varphi) - (\boldsymbol{\sigma}_{n,h} + \bar{\mathbf{K}}[\nabla\Psi_{h\tau} + \boldsymbol{g}\kappa(s_{h\tau})], \nabla\varphi)\\ &= (\mathcal{G}_{n,h} - \nabla\cdot\boldsymbol{\sigma}_{n,h}, \varphi) + (f(s_{h\tau},\boldsymbol{x},t) - f(S_{n,h},\boldsymbol{x},t_n), \varphi)\\ &\quad + (\tfrac{1}{\tau_n}(S_{n,h}-S_{n-1,h}) - \partial_t s_{h\tau}, \varphi) - \sum_{K\in\mathcal{T}_n}(\boldsymbol{\sigma}_{n,h} + \bar{\mathbf{K}}[\nabla\Psi_{h\tau} + \boldsymbol{g}\kappa(s_{h\tau})], \nabla\varphi)_K, \qquad (5.11)\end{aligned}$$

where $\mathcal{G}_{n,h}$ is defined in (5.4). For the first term on the right, following (5.8), we use

$$\begin{aligned}(\mathcal{G}_{n,h} - \nabla\cdot\boldsymbol{\sigma}_{n,h}, \varphi) &= \sum_{K\in\mathcal{T}_n}(\mathcal{G}_{n,h} - \nabla\cdot\boldsymbol{\sigma}_{n,h}, \varphi)_K \overset{(5.8b)}{=} \sum_{K\in\mathcal{T}_n}(\mathcal{G}_{n,h} - \nabla\cdot\boldsymbol{\sigma}_{n,h},\, \varphi - \tfrac{1}{|K|}\fint_K \varphi)_K\\ &\overset{(5.8a)}{=} \sum_{K\in\mathcal{T}_n}(\mathcal{G}_{n,h} - \Lambda_{n,h}\mathcal{G}_{n,h},\, \varphi - \tfrac{1}{|K|}\fint_K \varphi)_K \overset{(2.2),\,(3.4)}{\leq} \sum_{K\in\mathcal{T}_n}[\eta^{\mathrm{qd},\mathcal{G}}_{n,h,K}]\,\|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla\varphi\|_K. \qquad (5.12a)\end{aligned}$$

For the rest of the terms in (5.11), we note

$$\begin{aligned}&(f(s_{h\tau},\boldsymbol{x},t) - f(S_{n,h},\boldsymbol{x},t_n), \varphi) \leq [\eta^{\mathrm{osc}}_{n,\Omega}(t)]\,\|\varphi\|_{H^1_{\bar{\mathbf{K}}}(\Omega)}, \qquad (5.12b)\\ &(\tfrac{1}{\tau_n}(S_{n,h}-S_{n-1,h}) - \partial_t s_{h\tau}, \varphi) \leq [\eta^{\mathrm{qd},t}_{n,h,\Omega}(t)]\,\|\varphi\|_{H^1_{\bar{\mathbf{K}}}(\Omega)}, \qquad (5.12c)\\ &\sum_{K\in\mathcal{T}_n}(\boldsymbol{\sigma}_{n,h} + \bar{\mathbf{K}}[\nabla\Psi_{h\tau} + \boldsymbol{g}\kappa(s_{h\tau})], \nabla\varphi)_K \leq \sum_{K\in\mathcal{T}_n}[\eta^{\mathrm{F}}_{n,h,K}(t)]\,\|\bar{\mathbf{K}}^{\frac{1}{2}}\nabla\varphi\|_K. \qquad (5.12d)\end{aligned}$$

Combining the inequalities of (5.12) in (5.11) and using the Cauchy–Schwarz inequality, one arrives at (5.10a). Estimates (5.10b)–(5.10c) then follow from inserting (5.10a) in Theorem 3.4. □

5.4 Quadrature and temporal discretization estimators

For providing the efficiency bound, a few more estimators need to be introduced. For $n \in \{1,\dots,N\}$, an open polytope $\omega \subseteq \Omega$, and $t \in I_n$, the quadrature estimator for the numerical flux is defined as

$$\eta^{\mathrm{qd},\boldsymbol{F}}_{n,h,\omega} := \|\bar{\mathbf{K}}^{\frac{1}{2}}(\boldsymbol{F}_{n,h} - \Pi^{\mathrm{RT}}_{n,h}\boldsymbol{F}_{n,h})\|_\omega. \qquad (5.13a)$$

To measure the temporal discretization error of the numerical solutions $\Psi_{n,h}$ and $S_{n,h}$, we further introduce for $t \in I_n$ the estimators:

$$\eta^{\mathrm{J},H^1}_{n,h,\omega}(t) := \|\Psi_{h\tau}(t) - \Psi_{n,h}\|_{H^1_{\bar{\mathbf{K}}}(\omega)}, \quad \eta^{\mathrm{J},L^2}_{n,h,\omega}(t) := \|s_{h\tau}(t) - S_{n,h}\|_\omega. \qquad (5.13b)$$

5.5 Local-in-space and in-time efficiency

Theorem 5.4 (Local and global efficiency)**.** *Let $\Psi \in \mathcal{X}$ with $s = \theta(\Psi) \in \mathcal{Y}$ be the weak solution of* (2.16)*. Let $\{\Psi_{n,h}\}_{n=1}^N \subset H_0^1(\Omega)$ and $\{S_{n,h}\}_{n=1}^N \subset H^1(\Omega)$ be defined using the finite element discretization* (4.3)–(4.4) *and let $\Psi_{h\tau} \in C(0,T;H_0^1(\Omega)) \subset \mathcal{X}$ with $s_{h\tau} = \theta(\Psi_{h\tau}) \in W^{1,\infty}(0,T;H^1(\Omega)) \subset \mathcal{Y}$, be their time-continuous interpolates as defined in* (4.5)*. Let $\boldsymbol{\sigma}_{n,h}$ denote the equilibrated flux of Definition 5.1. Let the a posteriori error estimators be defined in* (5.9) *and* (5.13)*. Let* $\operatorname{dist}^{\alpha}_{\omega,I}$ *be defined in* (3.5) *for* $\alpha(t) = \max_{\mathbf{a}\in\mathcal{V}_n}\{C_{\mathrm{P},\omega_\mathbf{a}}\, h_{\omega_\mathbf{a}}\}\, K_{\mathrm{m}}^{-\frac{1}{2}} \max_{[0,1]\times\Omega\times\{t\}} |\partial_s f|$ $+|\boldsymbol{g}| K_{\mathrm{M}}^{\frac{1}{2}} \|\kappa'\|_{L^\infty([0,1])}$ *and $t \in I_n$. Then, for each discrete time step $n \in \{1,\dots,N\}$ and mesh element $K \in \mathcal{T}_n$, the indicators satisfy the following local-in-space and in-time efficiency bound:*

$$\begin{aligned}
&\int_{I_n} ([\eta^{\mathrm{F}}_{n,h,K}]^2 + [\eta^{\mathrm{J},H^1}_{n,h,K}]^2) \\
&\lesssim \sum_{\mathbf{a}\in\mathcal{V}_K} \left(\int_{I_n} \left[\sum_{j\in\{\mathcal{G},\boldsymbol{F},t\}} [\eta^{\mathrm{qd},j}_{n,h,\omega_\mathbf{a}}]^2 + [\eta^{\mathrm{osc}}_{n,\omega_\mathbf{a}}]^2 + \alpha^2\, [\eta^{\mathrm{J},L^2}_{n,h,\omega_\mathbf{a}}]^2 + [\eta^{\mathrm{J},H^1}_{n,h,\omega_\mathbf{a}}]^2 \right] + \operatorname{dist}^{\alpha}_{\omega_\mathbf{a},I_n}(\Psi,\Psi_{h\tau})^2 \right).
\end{aligned} \tag{5.14}$$

Furthermore, we have the following global-in-space efficiency bound:

$$\begin{aligned}
[\eta^n_{\mathrm{LB}}]^2 &:= \int_{I_n} ([\eta^{\mathrm{F}}_{n,h,\Omega}]^2 + [\eta^{\mathrm{J},H^1}_{n,h,\Omega}]^2) \\
&\lesssim \int_{I_n} \left(\sum_{j\in\{\mathcal{G},\boldsymbol{F},t\}} [\eta^{\mathrm{qd},j}_{n,h,\Omega}]^2 + [\eta^{\mathrm{osc}}_{n,\Omega}]^2 + \alpha^2 [\eta^{\mathrm{J},L^2}_{n,h,\Omega}]^2 + [\eta^{\mathrm{J},H^1}_{n,h,\Omega}]^2 \right) + \operatorname{dist}^{\alpha}_{\Omega,I_n}(\Psi,\Psi_{h\tau})^2.
\end{aligned} \tag{5.15}$$

Remark 5.5 (The linear heat equation case)**.** ***Quadrature:*** *In the absence of non-lineaities, $\eta^{\mathrm{qd},\boldsymbol{F}}_{n,h,K} = \eta^{\mathrm{qd},t}_{n,h,K} = 0$. To see this, note that if $S(p) = p$ and κ is linear with respect to s, then the numerical solutions $\Psi_{n,h}$ and $S_{n,h}$ are in the same polynomial space as $p_{n,h}$. Thus, the quadrature terms above vanish. Moreover, $\eta^{\mathrm{qd},\mathcal{G}}_{n,h,K} = \frac{h_K}{\sqrt{K_{\mathrm{m}}}\pi} \|f(\cdot,t_n) - \Lambda_{n,h} f(\cdot,t_n)\|$ becomes a data oscillation term.* ***Equivalence with estimates in [17]:*** *The bounds* (5.14)–(5.15) *are equivalent to the efficiency bounds presented in [17, Theorem 5.2] for the linear heat equation since $\int_{I_n} [\eta^{\mathrm{J},H^1}_{n,h,K}]^2$ is equivalent to $\int_{I_n} \|\nabla(\mathcal{I}u_{h\tau} - u_{h\tau})\|_K^2$ defined in [17, Section 5]. Additionally, in the linear case $\alpha = 0$. The grouping of terms in* (5.14)–(5.15) *is particularly useful since the quantity $(\sum_{n=1}^N (\int_{I_n} [\eta^{\mathrm{J},H^1}_{n,h,\Omega}]^2 + \operatorname{dist}^0_{\Omega,I_n}(\Psi,\Psi_{h\tau})^2))^{\frac{1}{2}}$ directly relates to the $\|\Psi - \Psi_{h\tau}\|_{\mathcal{E}_Y}$ error introduced in [17, Section 5] which provides an estimate of $\operatorname{dist}^0_{\Omega,[0,T]}(\Psi,\Psi_{h\tau})$ as proved in [17, Theorem 5.1]. Thus, the $(\int_{I_n} [\eta^{\mathrm{J},H^1}_{n,h,\Omega}]^2 + \operatorname{dist}^0_{\Omega,I_n}(\Psi,\Psi_{h\tau})^2)^{\frac{1}{2}}$ terms can themselves be considered error measures.*

Proof. Observe from (5.9), (5.13), and the definition of $\boldsymbol{F}_{n,h}$ in (5.4) that

$$\begin{aligned}
[\eta^{\mathrm{F}}_{n,h,K}] &\le \|\bar{\mathbf{K}}^{-\frac{1}{2}} \boldsymbol{\sigma}_{n,h} + \bar{\mathbf{K}}^{\frac{1}{2}} \boldsymbol{F}_{n,h}\|_K + \|\Psi_{h\tau} - \Psi_{n,h}\|_{H^1_{\bar{\mathbf{K}}}(K)} + \|\bar{\mathbf{K}}^{\frac{1}{2}} \boldsymbol{g}(\kappa(s_{h\tau}) - \kappa(S_{n,h}))\|_K \\
&\overset{(5.9),(5.13)}{\le} \left(\|\bar{\mathbf{K}}^{-\frac{1}{2}} \boldsymbol{\sigma}_{n,h} + \bar{\mathbf{K}}^{\frac{1}{2}} \Pi^{\mathrm{RT}}_{n,h} \boldsymbol{F}_{n,h}\|_K + [\eta^{\mathrm{qd},\boldsymbol{F}}_{n,h,K}] \right) + [\eta^{\mathrm{J},H^1}_{n,h,K}] + \alpha\, [\eta^{\mathrm{J},L^2}_{n,h,K}].
\end{aligned} \tag{5.16}$$

Note that, $[\eta^{\mathrm{qd},\boldsymbol{F}}_{n,h,K} + \eta^{\mathrm{J},H^1}_{n,h,K} + \alpha\, \eta^{\mathrm{J},L^2}_{n,h,K}] \lesssim \sum_{\mathbf{a}\in\mathcal{V}_K} [\eta^{\mathrm{qd},\boldsymbol{F}}_{n,h,\omega_\mathbf{a}} + \eta^{\mathrm{J},H^1}_{n,h,\omega_\mathbf{a}} + \alpha\, \eta^{\mathrm{J},L^2}_{n,h,\omega_\mathbf{a}}]$. For the first term

on the right-hand side of (5.16), one has

$$\|\bar{\mathbf{K}}^{-\frac{1}{2}}\boldsymbol{\sigma}_{n,h}+\bar{\mathbf{K}}^{\frac{1}{2}}\Pi^{\mathrm{RT}}_{n,h}\boldsymbol{F}_{n,h}\|_K=\left\|\sum_{\mathbf{a}\in\mathcal{V}_K}(\bar{\mathbf{K}}^{-\frac{1}{2}}\boldsymbol{\sigma}^{\mathbf{a}}_{n,h}+\psi_{\mathbf{a}}\bar{\mathbf{K}}^{\frac{1}{2}}\,\Pi^{\mathrm{RT}}_{n,h}\boldsymbol{F}_{n,h})\right\|_K$$
$$\leq\sum_{\mathbf{a}\in\mathcal{V}_K}\|\bar{\mathbf{K}}^{-\frac{1}{2}}\boldsymbol{\sigma}^{\mathbf{a}}_{n,h}+\psi_{\mathbf{a}}\bar{\mathbf{K}}^{\frac{1}{2}}\,\Pi^{\mathrm{RT}}_{n,h}\boldsymbol{F}_{n,h}\|_K\leq\sum_{\mathbf{a}\in\mathcal{V}_K}\|\bar{\mathbf{K}}^{-\frac{1}{2}}\boldsymbol{\sigma}^{\mathbf{a}}_{n,h}+\psi_{\mathbf{a}}\bar{\mathbf{K}}^{\frac{1}{2}}\,\Pi^{\mathrm{RT}}_{n,h}\boldsymbol{F}_{n,h}\|_{\omega_{\mathbf{a}}}. \tag{5.17}$$

By denoting $R^{\mathbf{a}}(\varphi):=(\mathcal{G}_{n,h},\varphi)_{\omega_{\mathbf{a}}}-(\bar{\mathbf{K}}\boldsymbol{F}_{n,h},\nabla\varphi)_{\omega_{\mathbf{a}}}$, we apply Theorem 1.2 of [16] (also see Lemma 10 of [17]) to get from (5.6) that

$$\|\bar{\mathbf{K}}^{-\frac{1}{2}}\boldsymbol{\sigma}^{\mathbf{a}}_{n,h}+\psi_{\mathbf{a}}\,\bar{\mathbf{K}}^{\frac{1}{2}}\Pi^{\mathrm{RT}}_{n,h}\boldsymbol{F}_{n,h}\|_{\omega_{\mathbf{a}}}\lesssim\sup_{\varphi\in H^1_0(\omega_{\mathbf{a}}),\ \|\varphi\|_{H^1_{\bar{\mathbf{K}}}(\omega_{\mathbf{a}})}=1}[(\Lambda_{n,h}\mathcal{G}_{n,h},\varphi)_{\omega_{\mathbf{a}}}-(\bar{\mathbf{K}}\,\Pi^{\mathrm{RT}}_{n,h}\boldsymbol{F}_{n,h},\nabla\varphi)_{\omega_{\mathbf{a}}}],$$
$$=\sup_{\varphi\in H^1_0(\omega_{\mathbf{a}}),\ \|\varphi\|_{H^1_{\bar{\mathbf{K}}}(\omega_{\mathbf{a}})}=1}\big[(\Lambda_{n,h}\mathcal{G}_{n,h}-\mathcal{G}_{n,h},\varphi)_{\omega_{\mathbf{a}}}-(\bar{\mathbf{K}}(\Pi^{\mathrm{RT}}_{n,h}\boldsymbol{F}_{n,h}-\boldsymbol{F}_{n,h}),\nabla\varphi)_{\omega_{\mathbf{a}}}+R^{\mathbf{a}}(\varphi)\big]$$
$$\overset{(4.2)}{=}\sup_{\substack{\varphi\in H^1_0(\omega_{\mathbf{a}}),\\ \|\varphi\|_{H^1_{\bar{\mathbf{K}}}(\omega_{\mathbf{a}})}=1}}\left[\sum_{K\in\mathcal{T}_n^{\mathbf{a}}}\left(\Lambda_{n,h}\mathcal{G}_{n,h}-\mathcal{G}_{n,h},\varphi-\tfrac{1}{|K|}\int_K\varphi\right)_K-(\bar{\mathbf{K}}(\Pi^{\mathrm{RT}}_{n,h}\boldsymbol{F}_{n,h}-\boldsymbol{F}_{n,h}),\nabla\varphi)_{\omega_{\mathbf{a}}}+R^{\mathbf{a}}(\varphi)\right]$$
$$\lesssim\eta^{\mathrm{qd},\mathcal{G}}_{n,h,\omega_{\mathbf{a}}}+\eta^{\mathrm{qd},\boldsymbol{F}}_{n,h,\omega_{\mathbf{a}}}+\sup_{\varphi\in H^1_0(\omega_{\mathbf{a}}),\ \|\varphi\|_{H^1_{\bar{\mathbf{K}}}(\omega_{\mathbf{a}})}=1}R^{\mathbf{a}}(\varphi). \tag{5.18}$$

Focusing on the final term, and recalling (5.4), one obtains for $\|\varphi\|_{H^1_{\bar{\mathbf{K}}}(\omega_{\mathbf{a}})}=1$ and $t\in I_n$ that

$$\begin{aligned}
&R^{\mathbf{a}}(\varphi)=(\mathcal{G}_{n,h},\varphi)_{\omega_{\mathbf{a}}}-(\bar{\mathbf{K}}\boldsymbol{F}_{n,h},\nabla\varphi)_{\omega_{\mathbf{a}}}\\
&=(f(S_{n,h},\boldsymbol{x},t_n)-\tfrac{1}{\tau_n}(S_{n,h}-S_{n-1,h}),\varphi)_{\omega_{\mathbf{a}}}-(\bar{\mathbf{K}}[\nabla\Psi_{n,h}+\boldsymbol{g}\kappa(S_{n,h})],\nabla\varphi)_{\omega_{\mathbf{a}}}\\
&\overset{(3.2)}{=}\langle\mathcal{R}(\Psi_{h\tau}),\varphi\rangle+(f(S_{n,h},\boldsymbol{x},t_n)-f(s_{h\tau},\boldsymbol{x},t),\varphi)_{\omega_{\mathbf{a}}}+(\partial_t s_{h\tau}-\tfrac{1}{\tau_n}(S_{n,h}-S_{n-1,h}),\varphi)_{\omega_{\mathbf{a}}}\\
&\qquad+(\bar{\mathbf{K}}[\nabla(\Psi_{h\tau}-\Psi_{n,h})+\boldsymbol{g}(\kappa(s_{h\tau})-\kappa(S_{n,h}))],\nabla\varphi)_{\omega_{\mathbf{a}}}\\
&\overset{(5.9)}{\leq}\|\mathcal{R}(\Psi_{h\tau})\|_{H^{-1}_{\bar{\mathbf{K}}}(\omega_{\mathbf{a}})}+\eta^{\mathrm{osc}}_{n,\omega_{\mathbf{a}}}+\eta^{\mathrm{qd},t}_{n,h,\omega_{\mathbf{a}}}+\eta^{\mathrm{J},H^1}_{n,h,\omega_{\mathbf{a}}}+\alpha\,\eta^{\mathrm{J},L^2}_{n,h,\omega_{\mathbf{a}}}.
\end{aligned}\tag{5.19}$$

Recall from Theorem 3.2 that $\int_{I_n}\|\mathcal{R}(\Psi_{h\tau})\|^2_{H^{-1}_{\bar{\mathbf{K}}}(\omega_{\mathbf{a}})}\leq\mathrm{dist}^{\alpha}_{\omega_{\mathbf{a}},I_n}(\Psi,\Psi_{h\tau})^2$. Thus combining (5.16)–(5.19), squaring both sides, and integrating over I_n, we have (5.14). To get the global efficiency bound (5.15) we sum (5.14) over all mesh elements and note that $\sum_{\mathbf{a}\in\mathcal{V}^n}\|u\|^2_{\omega_{\mathbf{a}}}\lesssim\|u\|^2$ and $\sum_{\mathbf{a}\in\mathcal{V}_n}\|u\|^2_{H^{\pm1}_{\bar{\mathbf{K}}}(\omega_{\mathbf{a}})}\lesssim\|u\|^2_{H^{\pm1}_{\bar{\mathbf{K}}}(\Omega)}$, see [12, Lemma 3.5]. □

6 Numerical results

We consider the following numerical test cases:

- Section 6.1: Nonlinear but nondegenerate problem with known exact solution.
- Section 6.2: Nonlinear and degenerate problem in the total pressure formulation (2.16) with known exact solution.
- Section 6.3: Realistic case, nonlinear, degenerate with heterogeneous and anisotropic $\bar{\mathbf{K}}$, mixed boundary conditions (Neumann + Dirichlet), discontinuous initial condition, non-uniform mesh, and no known exact solution.

- Section 6.4: Benchmark case [23] of groundwater reservoir recharging from a drainage trench.

For the first three test cases, we choose the unit square $\Omega = (0,1)^2$ as the simulation domain, $T = 1$ as the final time, and both uniform and non-uniform triangulations $\mathcal{T}_n$ with the discretization levels:

$$(h, \tau) = (h_0, \tau_0)/\ell \;\text{ where } \ell \in \{1, 2, 4\},\; h_0 = 0.2,\; \tau_0 = 0.04. \tag{6.1}$$

The mesh and the time-step size remain fixed between time steps. Piecewise linear finite elements are used for obtaining the solutions throughout, i.e., $\mathfrak{p}_n = 1$ in Section 4.2. Iterative linearization is discussed in Appendix A.5. The code is implemented in FreeFem++ and can be accessed through this link.

Remark 6.1 (Choice of λ in Theorems 3.4 and 5.3)**.** *The choice of $\lambda : [0,T] \to \mathbb{R}^+$ in* (3.12a) *is important in our simulations since in* (3.20)*, the $\|\mathcal{R}\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)}$ term is much larger than $\|G^0_{h\tau}\|_{H^1_{\bar{\mathbf{K}}}(\Omega)} = \|s - s_{h\tau}\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)}$. Hence, choosing $\lambda = 1$ leads to a significant overestimation of the error. Here, we have used $\lambda = 200$ in Section 6.1 and $\lambda = 100$ in Section 6.2. These yield close to minimum values of the effectivity indices defined in* (6.4)*. The optimal value of λ can also be roughly estimated by minimizing the right hand side of the Young's inequality in* (3.20)*. This gives $\lambda \sim \|\mathcal{R}\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)}/\|s - s_{h\tau}\|_{H^{-1}_{\bar{\mathbf{K}}}(\Omega)} \sim \eta_{\mathcal{R}}/\eta^{\mathrm{ini},H^{-1}}$ (see* (5.9)*) which yields λ in the same order of magnitude as the λ chosen in our simulations.*

6.1 Nonlinear nondegenerate case with known solution

For this case, $\bar{\mathbf{K}} = \mathbb{I}$, and with $\hat{\boldsymbol{e}}_x$ representing the unit vector along x-axis, we specify

$$\boldsymbol{g} = -\hat{\boldsymbol{e}}_x, \quad \kappa(s) = s^3, \text{ and } S(p) = \begin{cases} \frac{1}{(2-p)^{\frac{1}{3}}} & \text{if } p < 1, \\ 1 & \text{if } p \geq 1. \end{cases} \tag{6.2}$$

These nonlinearities resemble the Brooks–Corey parametrization (2.3). An exact solution

$$p_{\mathrm{exact}}(x, y, t) = 2 - \exp(16\,(1 + t^2)\,x\,y\,(1 - x)\,(1 - y)) \tag{6.3}$$

is fixed, see Figure 5 (left). The source term f is independent of s, and is adjusted together with the initial condition s_0, and the inhomogeneous Dirichlet boundary condition so that p_{exact} indeed solves (2.15).

Evolution of the different estimators for the case $\ell = 2$ is presented in Figure 5 (center), which shows that $\eta^{\mathrm{F}}_{n,h,K}$ is the dominant estimator followed by $\eta^{\mathrm{J},H^1}_{n,h,\Omega}(t)$ for this test. The spatial distribution of $\eta^{\mathrm{F}}_{n,h,K}$ is shown in Figure 5 (right). The time-quadrature and the degeneracy estimators, $\eta^{\mathrm{qd},t}_{n,h,\Omega}$ and η^{deg}, vanish in this case as the problem is nondegenerate.

Next, we numerically investigate the quality of the upper bound from Theorem 5.3. For this purpose, we introduce the effectivity index defined as

$$\text{effectivity index} := \text{upper bound/error} = \begin{cases} \eta_{L^2}/\mathcal{E}_{L^2}, \\ \eta_{H^1}/\mathcal{E}_{H^1}. \end{cases} \tag{6.4}$$

Effectivity index close to 1 is desirable. Figure 6 (left) shows the evolution of η_{L^2} (see (5.10b)) as a function of time and discretization level ℓ. The upper bound η_{L^2} reaches a constant state after an initial transition period. This is since $\mathfrak{C}_1(t)$ is almost constant for this case, and the

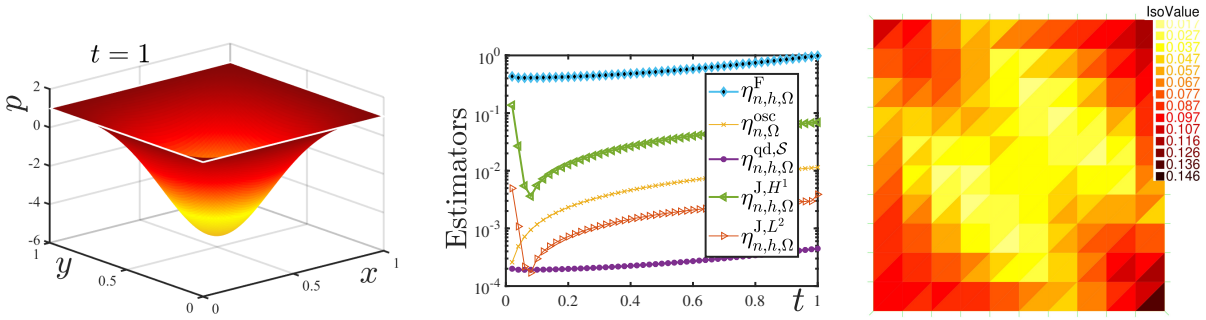

Figure 5: [Section 6.1] Exact solution p_{exact} of (6.3) at time $t = 1$ (left). Evolution of the 5 most significant estimators for $\ell = 2$ (center). The elementwise flux estimators $\eta^{\text{F}}_{n,h,K}$ for $\ell = 2$ and $t_n = 1$ (right).

error $\|\mathcal{R}\|_{H^{-1}_{\mathbf{K}}(\Omega)}$ increases exponentially with a rate much smaller than $\lambda + \mathfrak{C}_1$. Hence, a near constant η_{L^2} is expected from Remark 3.1. The (right) plot shows the effectivity indices of η_{L^2}. The effectivity varies between 1.4 and 3.1, and improves with the discretization level ℓ. Figure 7 is the same plot presented for η_{H^1}. The estimator η_{H^1} increases with t as $\mathfrak{C}_2$ increases rapidly with time. The effectivity index again improves as the discretization is refined.

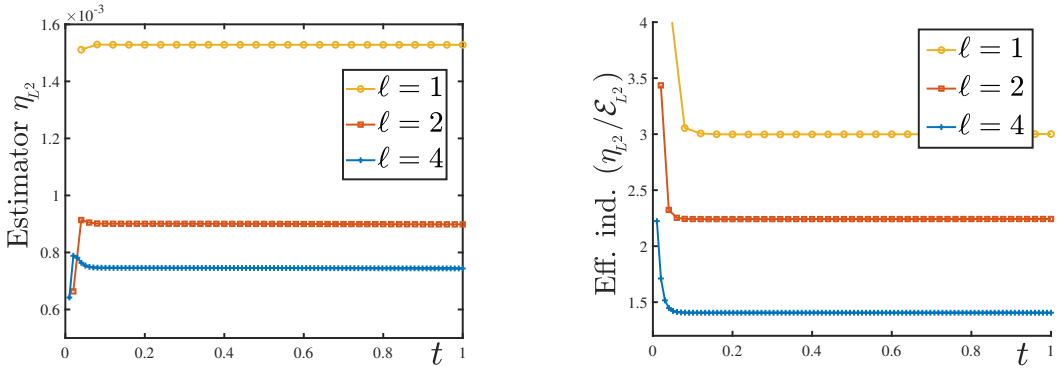

Figure 6: [Section 6.1] Estimator η_{L^2} of (5.10b) as a function of the final time (left), and the corresponding effectivity index (right).

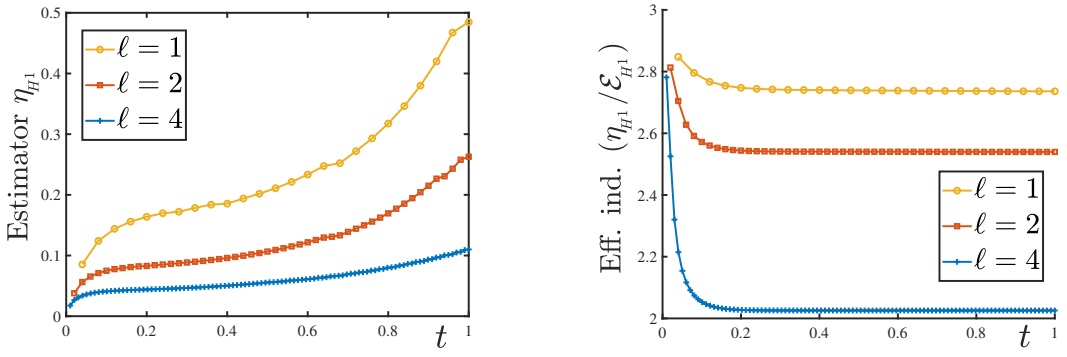

Figure 7: [Section 6.1] Estimator η_{H^1} of (5.10c) as a function of the final time (left), and the corresponding effectivity index (right).

We now turn to the lower bound of Theorem 5.4. The effectivity index in this context is

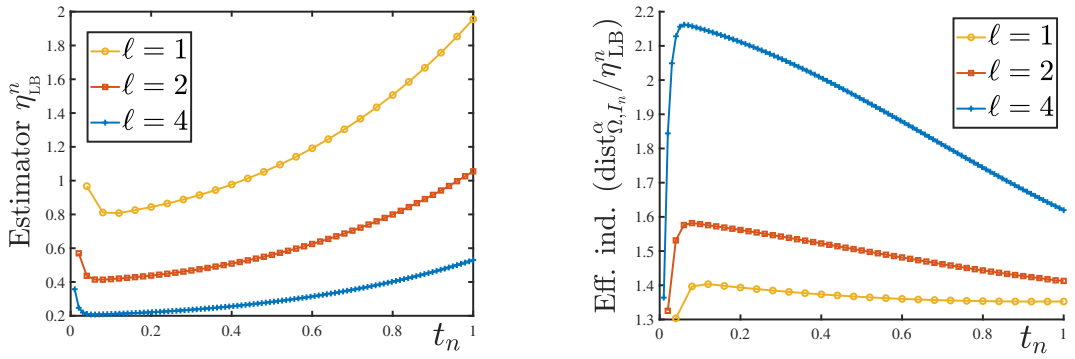

Figure 8: [Section 6.1] Estimator $\eta^n_{\rm LB}$ of (5.15) as a function of t_n (left). Its effectivity indices computed using (6.5) (right).

computed as

$$\text{effectivity index} := \text{error/lower bound} = \mathrm{dist}^{\alpha}_{\Omega,I_n}(\Psi, \Psi_{h\tau})/\eta^n_{\rm LB}, \tag{6.5}$$

where $\eta^n_{\rm LB}$ is given by (5.15). The reversed order is to make the effectivity index comparable to the effectivity index of the upper bound. Figure 8 shows the lower bound $\eta^n_{\rm LB}$ and its effectivity indices. The effectivity increases with ℓ in this case, though only varying between 1.4 and 2.2. A higher effectivity index close to $t = 0$ is also observed. This is explained by the fact that the lower bound estimator $\eta^n_{\rm LB}$ does not incorporate the initial errors, and thus, is more susceptible to inaccuracies close to $t = 0$. Figure 9 shows the variation of the effectivity indices with ℓ at the final time $T = 1$, for both the reliability and the efficiency estimates.

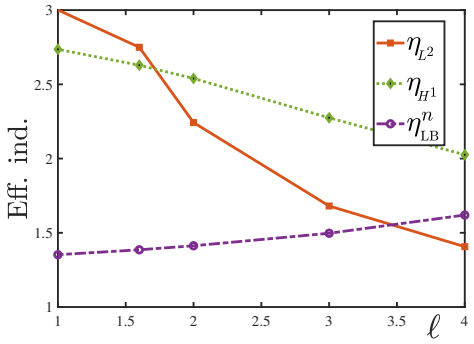

Figure 9: [Section 6.1] Effectivity indices of the estimators η_{L^2}, η_{H^1}, and $\eta^n_{\rm LB}$ at the final time $T = 1$ varying with ℓ.

Inspired by Theorem 5.4, the local-in-space and in-time effectivity indices are computed as

$$(\text{eff. ind.})_{n,K} := \mathrm{dist}^{\alpha}_{K,I_n}(\Psi, \Psi_{h\tau}) / \left(\textstyle\int_{I_n} ([\eta^{\rm F}_{n,h,K}]^2 + [\eta^{{\rm J},H^1}_{n,h,K}]^2) \right)^{\frac{1}{2}}, \tag{6.6}$$

for all $K \in \mathcal{T}_n$. From Figure 5 (right), it is observed that $\eta^{\rm F}_{n,h,K}$ varies with the mesh elements K by a factor of about 10. However, Figure 10 shows that the local effectivity indices are in the range 0.6–1.8 for $\ell = 1$, 0.8–2.4 for $\ell = 2$, and 0.8–3.8 for $\ell = 4$, which we consider excellent. Observe that, $(\text{eff. ind.})_{n,K} < 1$ does not violate Theorem 5.4 since the error $\mathrm{dist}^{\alpha}_{\omega_{\mathbf{a}},I_n}(\Psi, \Psi_{h\tau})$ and the sign '$\lesssim$' (up to a constant) was used there.

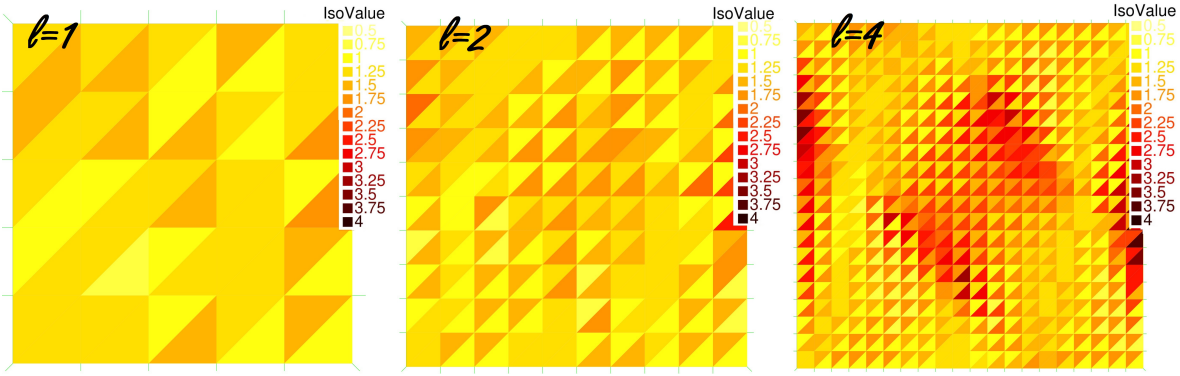

Figure 10: [Section 6.1] Local effectivity indices (6.6) at the final time $T = 1$ for different values of ℓ.

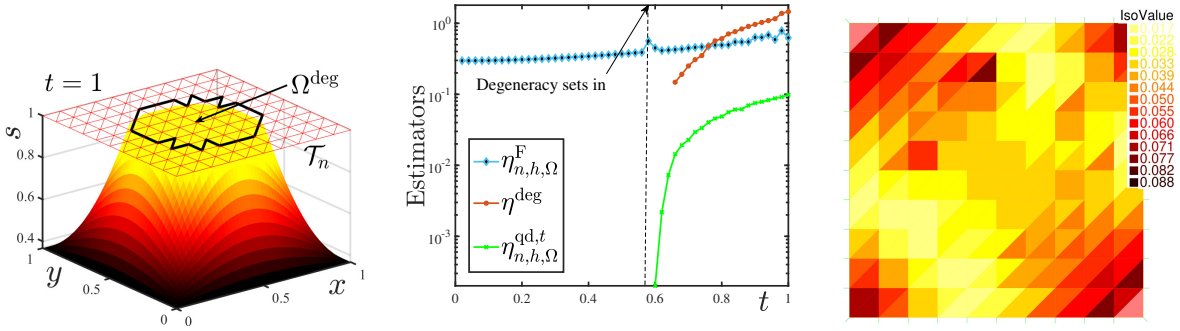

Figure 11: [Section 6.2] Saturation of the exact solution p_{exact} and the domain $\Omega^{\text{deg}}(t)$ at $t = 1$ (left). The principal estimators $\eta^{\text{F}}_{n,h,\Omega}(t)$, $\eta^{\text{deg}}(t)$, and $\eta^{\text{qd},t}_{n,h,\Omega}(t)$ for $\ell = 2$ (center), and the elementwise estimators $\eta^{\text{F}}_{n,h,K}$ at $t_n = 1$ (right).

6.2 Nonlinear degenerate case with known solution

This test on purpose focuses on a system where degeneracy is the dominant effect. For a change, the total pressure formulation (2.16) is used here. The nonlinearities are set as

$$\kappa(s) = 1, \quad S(p) = \begin{cases} \exp(p-1) & \text{if } p < 1, \\ 1 & \text{if } p \geq 1, \end{cases} \tag{6.7}$$

with $\bar{\mathbf{K}} = \mathbb{I}$. This choice implies that the pressure and the total pressure formulations are essentially the same since $\Psi = \mathcal{K}(p) = p$. The exact solution used is

$$p_{\text{exact}}(x, y, t) = 12\,(1 + t^2)\,x\,y\,(1 - x)\,(1 - y). \tag{6.8}$$

Appropriate source function f (independent of s), initial and boundary conditions are again imposed.

The solution is initially nondegenerate and contains a degenerate region after $t = 0.58$, see Figure 11 (left). This is caused by the source term f since $\bar{\mathbf{K}}$ here is uniform. The domain $\Omega^{\text{deg}}(t)$ is approximately computed for $t \in I_n$ as

$$\Omega^{\text{deg}}(t) = \cup\{K \in \mathcal{T}_n : K \cap \{\Psi_{h\tau}(t) > P_{\text{M}} = 1\} \neq \emptyset\}, \tag{6.9}$$

pointed out in Figure 11 (left). This replaces here the generally unknown Ω^{deg} from Theorem 3.4. Figure 11 (center) shows the estimators $\eta^{\text{F}}_{n,h,\Omega}$, η^{deg}, and $\eta^{\text{qd},t}_{n,h,\Omega}(t)$ for the case $\ell = 2$. The

degeneracy estimator η^{deg} defined in Theorem 3.4 quickly rises in value as degeneracy sets in. In Figure 11 (right) we see the distribution of $\eta^{\text{F}}_{n,h,K}$. The flux estimator $\eta^{\text{F}}_{n,h,K}$ stays relatively unaffected by the onset of degeneracy.

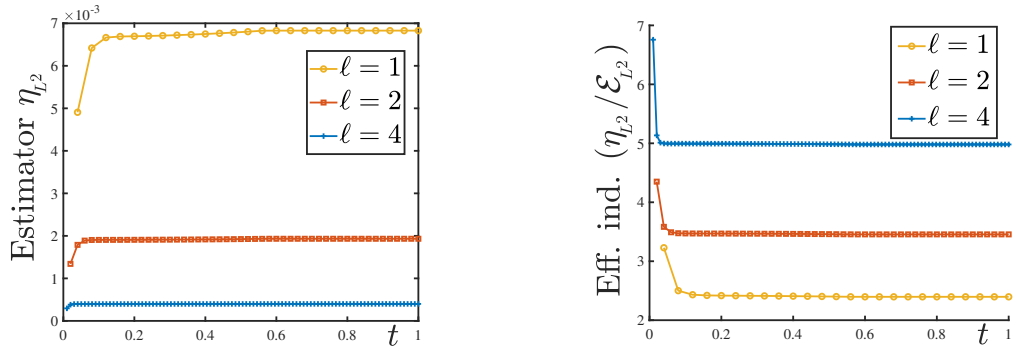

Figure 12: [Section 6.2] Reliability estimator η_{L^2} (left) and its effectivity index (right).

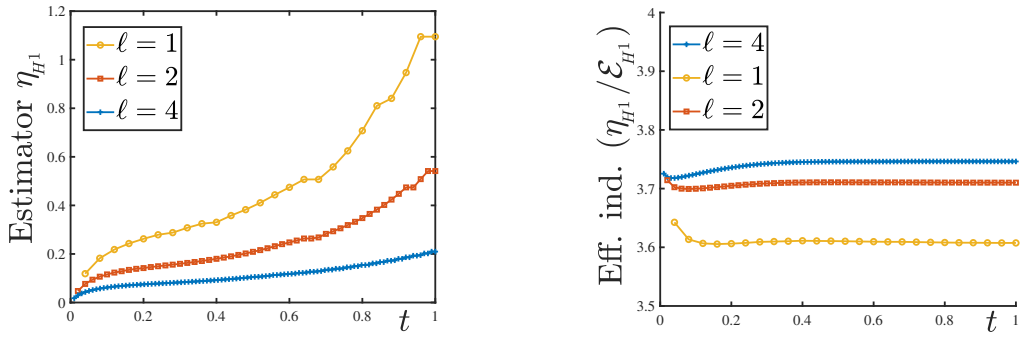

Figure 13: [Section 6.2] Reliability estimator η_{H^1} (left) and its effectivity index (right).

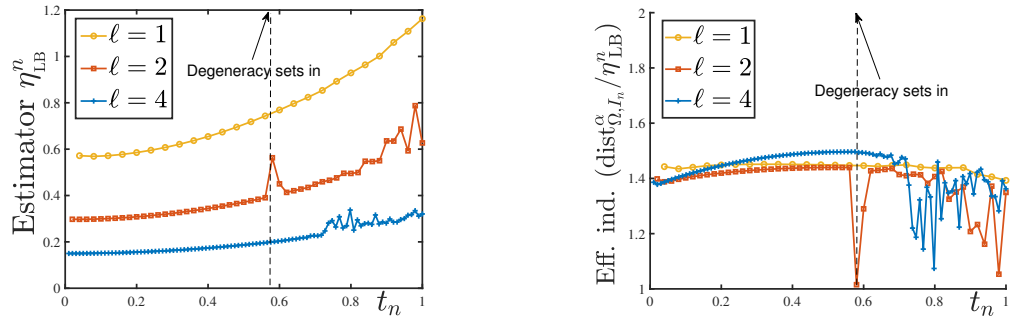

Figure 14: [Section 6.2] Lower bound η^n_{LB} (left) and its effectivity index (right).

The effectivity indices for Theorems 5.3 and 5.4 are defined as before. Figure 12 shows the estimate η_{L^2} and its effectivity. The effectivity index increases with ℓ, despite η_{L^2} decreasing monotonically, possibly since $s = s_{h\tau} = 1$ in a major portion of the domain towards the end of the simulation. Figure 13 shows the results for η_{H^1}. The effectivity remains more stable in this case. The effectivity indices for the lower bound are shown in Figure 14 (right). An oscillation in the lower bound is observed for higher values of ℓ. The reason for this behaviour is not clear. Figure 15 shows the distribution of local space–time effectivity indices. They are close to 1 in

most regions and only take a lower value close to the free-boundary $s = 1$. Overall, we find these results satisfactory.

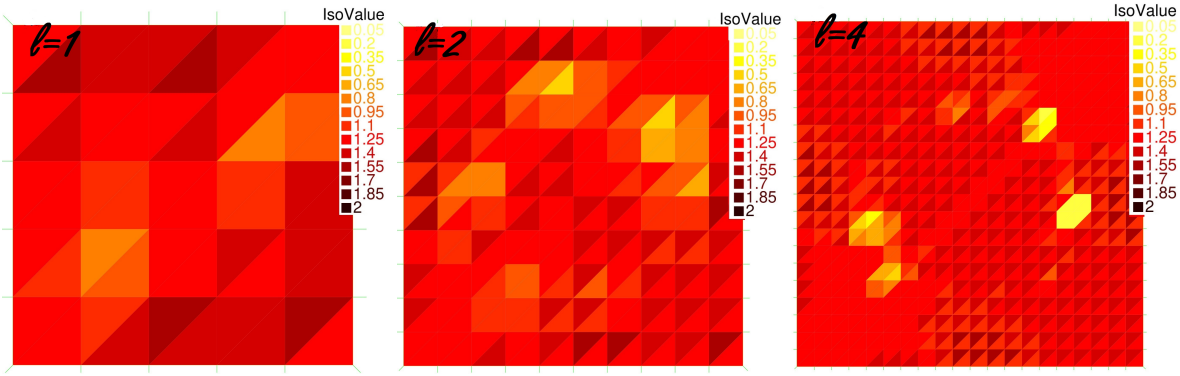

Figure 15: [Section 6.2] Local effectivity indices (6.6) for $t_n = 1$ and $\ell = 1, 2, 4$.

6.3 Realistic case

In this case, the domain and the functions given in (6.2) are kept unchanged. The source term f is made 0. However, the medium used is heterogeneous and anisotropic with

$$\bar{\mathbf{K}} = \begin{cases} \bar{\mathbf{K}}_1 & \text{for } x < 0.5, \\ K_\phi \, \boldsymbol{Q}^{\mathrm{T}} \bar{\mathbf{K}}_1 \boldsymbol{Q} & \text{for } x \geq 0.5, \end{cases} \quad \text{where } \bar{\mathbf{K}}_1 := \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix}, \; \boldsymbol{Q} := \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{6.10}$$

Here, θ represents a tilted alignment of the principle axes of $\bar{\mathbf{K}}$, and K_ϕ represents a factor stemming from the change in porosity. The choice of $\theta = \pi/3$ and $K_\phi = 0.1$ is fixed. Both Neumann and inhomogeneous Dirichlet boundary conditions are used for the computation with an input pressure $p = p_{\text{in}} = 0.8$ prescribed on $\{0\} \times (0, 0.5)$, an output pressure $p = p_{\text{out}} = -3$ prescribed on $(0.5, 1) \times \{1\}$, and no flux condition prescribed on the rest of the boundary. The initial condition used is discontinuous, i.e. $s_0 = S(p_{\text{in}})$ for $x < 0.5$ and $s_0 = S(p_{\text{out}})$ for $x \geq 0.5$. Figure 16 (left) shows these conditions inside the domain. A nonuniform mesh is used for the computation. No exact solution is known for this system.

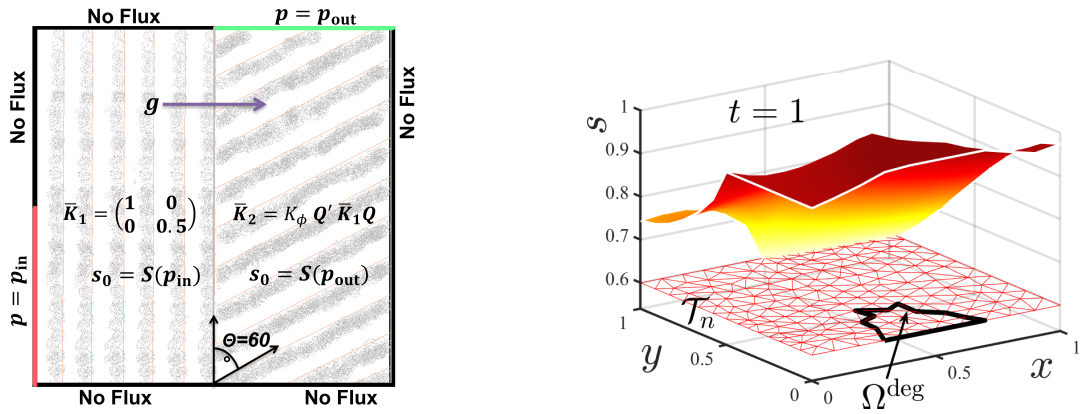

Figure 16: [Section 6.3] Computational domain showing heterogeneities, initial, and boundary conditions (left). Saturation of the numerical solution ($S_{n,h}$) for $\ell = 2$ at $t = 1$. The mesh and the domain $\Omega^{\text{deg}}(1)$ containing the degenerate region is shown (right).

Degeneracy occurs in the system close to the interface $x = 0.5$ at $y = 0$. This is caused by the jump in $\bar{\mathbf{K}}$, but also partly by the no-flux boundary condition. The error caused by this additional component is estimated by adding to $[\eta^{\mathrm{deg}}(t)]^2$,

$$\frac{2}{D(1)|\Omega^{\mathrm{deg}}|}\int_{\partial\Omega}\hat{\boldsymbol{n}}^{\mathrm{T}}\left(\int_{\Omega^{\mathrm{deg}}}\bar{\mathbf{K}}\boldsymbol{g}\right)[\Psi_{h\tau}(t)-P_{\mathrm{M}}]_+.$$

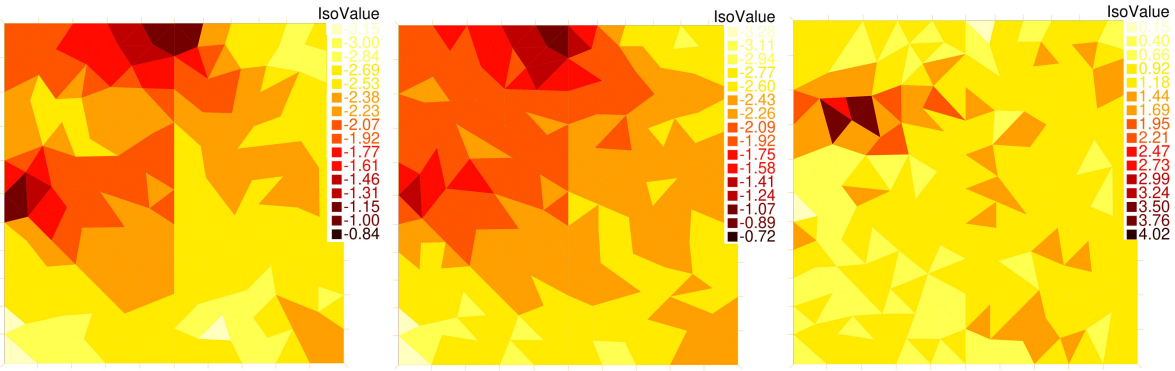

Figure 17: [Section 6.3] Elementwise distribution of $\eta^{\mathrm{F}}_{n,h,K}$ at time $t = 1$ in $\log_{10}$-scale for $\ell = 2$ (left). Estimators take larger values near the inlet and the outlet. The error distribution (center) and the local effectivity indices defined in (6.6) (right). Here, the numerical solution for $\ell = 4$ is used in the place of the exact solution.

Figure 16 (right) shows the saturation distribution and degenerate zone for this problem at $t = 1$. Figure 18 (left) shows the main estimators for $\ell = 2$. The flux estimator is still the largest component, followed by η^{deg}. Figure 17 (left) plot shows the spatial distribution of $\eta^{\mathrm{F}}_{n,h,K}$ indicating high error concentrations located around the inlet and the outlet. The right plot shows the local effectivity indices, where the numerical solution for $\ell = 4$ is used as the reference solution. Although the estimators vary by almost 3 orders of magnitude, the effectivity varies between 0.15–4 with most of the region having effectivity close to 1. The overall effectivity index (6.5) of the estimator is 2.053. We find this again quite satisfactory.

6.4 Benchmark case

Finally, we also consider a benchmark problem, proposed in [23] and used e.g. in [28]. It describes the infiltration of water in the vadose layer of the soil from a water body. The domain is $\Omega = (0, 2) \times (0, 3)$, and the porous medium is uniform slit loam. The van Genuchten parametrization (2.4) is used for the capillary curves with $\lambda_2 = 1-1/2.06$ and $p_{\mathrm{M}} = 1$. Moreover, $\bar{\mathbf{K}} = K_\phi\,\mathbb{I}$ with $K_\phi = 4.96 \times 10^{-2}$, and $\boldsymbol{g} = \hat{\boldsymbol{e}}_y$. The boundary conditions are

$$p(x,y,t) = p_{\mathrm{in}}(t) := \begin{cases} -1 + 35.2\,t & \text{if } t < \frac{1}{16}, \\ 1.2 & \text{if } t \geq \frac{1}{16}, \end{cases} \text{ on } (0,1)\times\{3\},$$
$$p(x,y,t) = 2 - y \text{ on } \{2\}\times(0,1),$$

and no flux conditions are set on the rest of the boundary. The initial condition is $s_0(x, y) = S(2 - y)$. A pictorial representation of the numerical setting is shown in Figure 19 (left). Figure 19 (right) shows the pressure distribution for a reference solution with $h = 1/15$ and $\tau = 1/144$. The a posteriori estimators are computed with respect to this reference solution for a coarser simulation with $h = 1/4$ and $\tau = 1/48$. Note that a considerable fraction of the

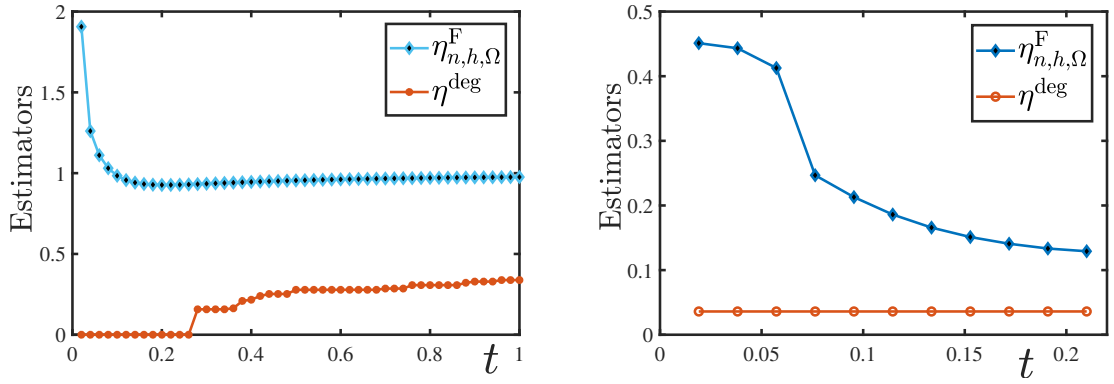

Figure 18: [Sections 6.3 and 6.4] The main estimators for the realistic and the benchmark cases. Realistic case (left): The flux estimator $\eta^{\mathrm{F}}_{n,h,\Omega}(t)$ contributes the most to the error, along with the degeneracy estimator $\eta^{\mathrm{deg}}(t)$ which becomes non-zero only after the onset of degeneracy. Benchmark case (right).

domain is saturated ($p \geq 1$ and thus $s = 1$), and the Ω^{deg} domain completely covers this region in our simulations. Despite this, Figure 18 (right) shows that the flux estimator is still the dominant estimator followed by the degeneracy estimator η^{deg}.

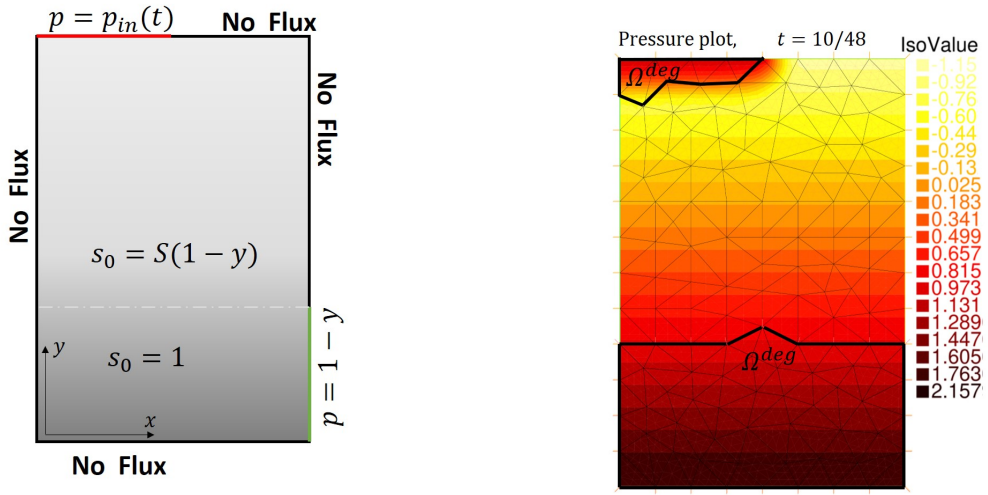

Figure 19: [Section 6.4] (left) Computational setting for the Benchmark case. (right) Pressure distribution of the reference simulation at $t = 10/48$. The mesh skeleton and the domain $\Omega^{\mathrm{deg}}(t)$ for the coarser simulation are superimposed.

Figure 20 shows the estimator $\eta^{\mathrm{F}}_{n,h,K}$, error $\mathrm{dist}^{\alpha}_{K,I_n}(\Psi, \Psi_{h\tau})$, and local effectivity indices for this simulation. The error varies for more than 5 orders of magnitude in the domain and is concentrated near the inlet. The estimator correctly predicts this trend. The local effectivity indices are close to 1 in the nondegenerate domain, but immediately shoot up to over 100 in the degenerate domain. Observe that in the degenerate domain, $\|\Psi - \Psi_{h\tau}\|_{L^2(I_n, H^1_{\widetilde{\mathbf{K}}}(K))}$ is the only non-zero component in $\mathrm{dist}^{\alpha}_{K,I_n}(\Psi, \Psi_{h\tau})$ (see (3.5)n), since the other two components vanish. Hence, the numerical results bolster our claim that to have a reliable bound over $\|\Psi - \Psi_{h\tau}\|_{L^2(I_n, H^1_{\widetilde{\mathbf{K}}}(K))}$ in the degenerate case, we need to consider both $\eta_{\mathcal{R}}$ and the degeneracy estimator η^{deg} in the upper bound. The error in the degenerate domain is minuscule compared

to the error in the nondegenerate domain. As a consequence, the overall effectivity index (6.5) of the estimator remains close to unity, more precisely it is 1.050.

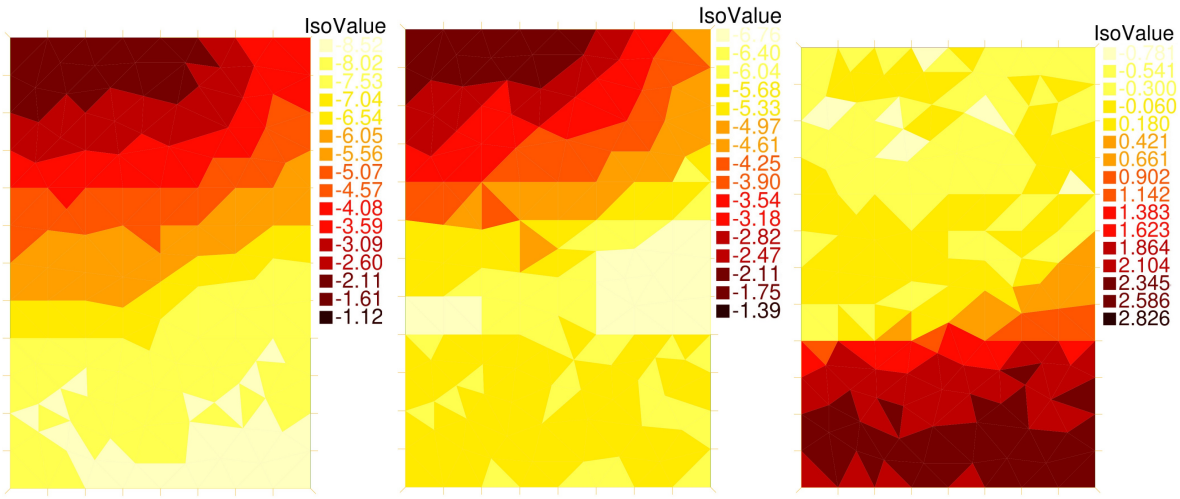

Figure 20: [Section 6.4] Estimator $\eta^{\mathrm{F}}_{n,h,K}$ (left), error $\mathrm{dist}^{\alpha}_{K,I_n}(\Psi, \Psi_{h\tau})$ (center), and local effectivity indices (6.6) (right) ($K \in \mathcal{T}_n$) in a $\log_{10}$-scale plot for the benchmark case at $t = 10/28$.

Appendix A Iterative linearization

In practice, since the problem (4.3) is nonlinear, its solution $p_{n,h}$ cannot be directly enumerated, and linearization iterations have to be used. We address this issue here.

A.1 Linearization

Set $S^{\bar{i}}_{0,h} = \Pi_{1,h} s_0$. For $n \in \{1, \dots, N\}$, let $S^{\bar{i}}_{n-1,h}$ be an approximation of $S_{n-1,h}$. Let $p^i_{n,h} \in V_{n,h}$ denote the pressure at iteration $i \in \mathbb{N}$. Then, for a given $p^{i-1}_{n,h} \in V_{n,h}$, we look for $\eth p^i_{n,h} := p^i_{n,h} - p^{i-1}_{n,h} \in V_{n,h}$ satisfying for all $\varphi_h \in V_{n,h}$,

$$\begin{aligned} &\tfrac{1}{\tau_n}(L\,\eth p^i_{n,h}, \varphi_h) + (\bar{\mathbf{K}}[\kappa(S(p^{i-1}_{n,h}))\nabla p^i_{n,h} + \boldsymbol{\xi}\,\eth p^i_{n,h}], \nabla\varphi_h) \\ &= -\tfrac{1}{\tau_n}(S(p^{i-1}_{n,h}) - S^{\bar{i}}_{n-1,h}, \varphi_h) - (\bar{\mathbf{K}}\boldsymbol{g}\kappa(S(p^{i-1}_{n,h})), \nabla\varphi_h) + (f(S(p^{i-1}_{n,h}), \boldsymbol{x}, t_n), \varphi_h), \end{aligned} \tag{A.1}$$

Here, $(L, \boldsymbol{\xi}) \in \mathbf{L}^\infty(\Omega; \mathbb{R}^{d+1})$ with $L \geq 0$, depends on the specific scheme used. Since (A.1) is linear with respect to $p^i_{n,h}$, it is directly computable. Observe that $\eth p^i_{n,h} = 0$ if and only if $p^{i-1}_{n,h}$ solves (4.3) (provided $S^{\bar{i}}_{n-1,h} = S_{n-1,h}$) which shows that the schemes are consistent.

Scheme	L, $\boldsymbol{\xi}$	Convergence
Picard	0, $\mathbf{0}$	
modified Picard [11]	$S'(p_{n,h}^{i-1})-\tau_n\partial_s f(S(p_{n,h}^{i-1}),\boldsymbol{x},t_n)$, $\mathbf{0}$	Linear
Newton [7]	$S'(p_{n,h}^{i-1})-\tau_n\partial_s f(S(p_{n,h}^{i-1}),\boldsymbol{x},t_n)$, $\kappa'(S(p_{n,h}^{i-1}))[\nabla p_{n,h}^{i-1}+\boldsymbol{g}]$	Quadratic
Jäger–Kačur [21]	$\sup_{p\in\mathbb{R}}\{(S(p)-S(p_{n,h}^{i-1})-\tau_n(f(S(p))-f(S(p_{n,h}^{i-1}))))/(p-p_{n,h}^{i-1})\}$, $\mathbf{0}$	Linear
L–scheme [28]	L (constant)$\geq\frac{1}{2}\sup\{S'-\tau_n\partial_s f\}$, $\mathbf{0}$	Linear
modified L–scheme [29]	$S'(p_{n,h}^{i-1})-\tau_n\partial_s f(S(p_{n,h}^{i-1}),\boldsymbol{x},t_n)+M\tau_n$ ($M>0$ constant), $\mathbf{0}$	Linear

Table 2: Different iterative linearization schemes commonly used for Richards equation (1.1a). They fit into the common framework (A.1). The corresponding L and $\boldsymbol{\xi}$ quantities are displayed along with the convergence characteristics of the schemes.

Table 2 lists the commonly used schemes and the L and $\boldsymbol{\xi}$ associated with them. The Picard scheme is generally unstable for the Richards equation. The modified Picard scheme [11] is linearly converging and the Newton method [7] is quadratically converging for an initial guess close to the solution of the nonlinear problem. However, the convergence is not guaranteed for degenerate cases. The Jäger–Kačur scheme [21] and the L–scheme [28] are unconditionally stable, meaning that they converge linearly, independent of the initial guess even in degenerate cases and for discontinuous initial conditions. However, a global supremum has to be computed for the Jäger–Kačur scheme, whereas, L–scheme converges slowly compared to the schemes mentioned above. The modified L–scheme [29] preserves the stability of the L–scheme while being faster than the modified Picard scheme.

For a given $n\in\{1,\dots,N\}$, let the linear iteration (A.1) be terminated at $i=\bar{i}$. From the sequence $\{p_{n,h}^{\bar{i}}\}_{n=1}^{N}$, the space–time discrete total pressure and saturation are defined as

$$\Psi_{n,h}^{\bar{i}}:=\mathcal{K}(p_{n,h}^{\bar{i}}),\ \text{and}\ S_{n,h}^{\bar{i}}:=\theta(\Psi_{n,h}^{\bar{i}})\overset{(2.14)}{=}S(p_{n,h}^{\bar{i}})\quad\forall n\in\{1,\dots,N\},\tag{A.2}$$

analogous to (4.4). Replacing $(\Psi_{n,h},S_{n,h})$ by $(\Psi_{n,h}^{\bar{i}},S_{n,h}^{\bar{i}})$, we compute the time continuous solutions $\Psi_{h\tau}$ and $s_{h\tau}$ by following the steps of Section 4.5.

A.2 Equilibrated flux

The terms $\mathcal{G}_{n,h}$ and $\boldsymbol{F}_{n,h}$ from (5.4) are redefined as

$$\mathcal{G}_{n,h}:=\Big(f(S_{n,h}^{\bar{i}-1},\boldsymbol{x},t_n)-\tfrac{1}{\tau_n}(S_{n,h}^{\bar{i}-1}-S_{n-1,h}^{\bar{i}})-L\,\eth p_{n,h}^{\bar{i}}\Big)\big|_{I_n},\tag{A.3a}$$

$$\boldsymbol{F}_{n,h}:=\Big[\kappa(S_{n,h}^{\bar{i}-1})\nabla p_{n,h}^{\bar{i}}+\boldsymbol{g}\kappa(S_{n,h}^{\bar{i}-1})+\boldsymbol{\xi}\,\eth p_{n,h}^{\bar{i}}\Big]\big|_{I_n}.\tag{A.3b}$$

Observe that upon rearranging (A.1), $\mathcal{G}_{n,h}$ and $\boldsymbol{F}_{n,h}$ play the role of the source-like and flux-like terms, just as in (5.4). Moreover, $\mathcal{G}_{n,h}$ and $\boldsymbol{F}_{n,h}$ converge to their definitions in (5.4) if the iterate $p_{n,h}^{\bar{i}}$ converges to $p_{n,h}$.

The equilibrated flux $\boldsymbol{\sigma}_{n,h}\in\boldsymbol{H}(\mathrm{div},\Omega)$ is then constructed as stated in Definition 5.1.

A.3 Estimators

The estimators in (5.9a)–(5.9e) and (5.13) are defined exactly the same way replacing $(\Psi_{n,h},S_{n,h})$ by $(\Psi_{n,h}^{\bar{i}},S_{n,h}^{\bar{i}})$. The linearization estimators for the source-like and flux-like terms are introduced for $\omega\subseteq\Omega$ and $n\in\{1,\dots,N\}$ as

$$\eta_{n,\omega}^{\mathrm{lin},1}:=C_{\mathrm{P},\omega}\,h_\omega\,\Big\|\tfrac{1}{\tau_n}(S_{n,h}^{\bar{i}}-S_{n,h}^{\bar{i}-1}-L\,\eth p_{n,h}^{\bar{i}})-(f(S_{n,h}^{\bar{i}})-f(S_{n,h}^{\bar{i}-1}))\Big\|_\omega,\tag{A.4a}$$

$$\eta_{n,\omega}^{\mathrm{lin},2}:=\Big\|\bar{\mathbf{K}}^{\frac{1}{2}}\Big((\kappa(S_{n,h}^{\bar{i}})-\kappa(S_{n,h}^{\bar{i}-1}))\,[\nabla p_{n,h}^{\bar{i}}+\boldsymbol{g}]+\boldsymbol{\xi}\,\eth p_{n,h}^{\bar{i}}\Big)\Big\|_\omega,\tag{A.4b}$$

($C_{\mathrm{P},\omega} > 0$ is the Poincaré constant). The new total estimator becomes, for $t \in I_n$,

$$\eta_{\mathcal{R}}(t) := \left[\sum_{K\in\mathcal{T}_n} [\eta^{\mathrm{F}}_{n,h,K}(t) + \eta^{\mathrm{qd},\mathcal{G}}_{n,h,K}]^2\right]^{\frac{1}{2}} + \eta^{\mathrm{qd},t}_{n,h,\Omega}(t) + \eta^{\mathrm{osc}}_{n,\Omega}(t) + \eta^{\mathrm{lin},1}_{n,\Omega}. \tag{A.5}$$

Remark A.1. *Observe that in* (A.4) *to define* $\eta^{\mathrm{lin},1}_{n,\Omega}$*, we have used the* L^2 *norm instead of the* $H^{-1}_{\mathbf{K}}$*-norm, which is costly to evaluate at every iteration. Observe also that only* $\eta^{\mathrm{lin},1}_{n,\Omega}$ *appears in* (A.5)*.*

A.4 Adaptive linearization

Inspired by [18], we propose the following adaptive algorithm for the linearization:

Algorithm A.1 (Adaptive linearization)**.** For a fixed $\gamma \in (0,1)$ and $n \in \mathbb{N}$, let $S^{\bar{i}}_{n-1,h} \in L^2(\Omega)$ and $p^0_{n,h} \in H^1(\Omega)$ be given. Then, for each $i \in \mathbb{N}$, solve (A.1) until for some $i = \bar{i}$, upon computation of $\eta^{\mathrm{F}}_{n,h,\Omega}$ from (5.9a) and $\eta^{\mathrm{lin},1}_{n,\Omega}$, $\eta^{\mathrm{lin},2}_{n,\Omega}$ from (A.4), the following holds

$$\eta^{\mathrm{lin},1}_{n,\Omega} + \eta^{\mathrm{lin},2}_{n,\Omega} \le \gamma\, \eta^{\mathrm{F}}_{n,h,\Omega}. \tag{A.6}$$

With Algorithm A.1, Theorems 5.3 and 5.4 are restated as

Proposition A.1 (Reliability and efficiency with linearization)**.** *Let* $\{\Psi^{\bar{i}}_{n,h}\}_{n=1}^N \subset H^1_0(\Omega)$ *and* $\{S^{\bar{i}}_{n,h}\}_{n=1}^N \subset H^1(\Omega)$ *be defined using the numerical scheme* (A.1)–(A.2) *with stopping criteria set by Algorithm A.1. Let* $\Psi_{h\tau} \in C(0,T;H^1_0(\Omega))$ *with* $s_{h\tau} = \theta(\Psi_{h\tau}) \in W^{1,\infty}(0,T;H^1(\Omega))$*, be their time-continuous interpolates as defined in* (4.5)*, with* $(S_{n,h}, \Psi_{n,h})$ *replaced by* $(S^{\bar{i}}_{n,h}, \Psi^{\bar{i}}_{n,h})$*. Let the estimators* $\eta^{\mathrm{lin},j}_{n,\Omega}$*,* $j = 1,2$*, be defined in* (A.4) *and* $\eta_{\mathcal{R}}$ *in* (A.5)*. Then*

(a) ***Reliability:*** *Under the assumptions of Theorem 5.3, the estimates* (5.10) *hold.*

(b) ***Efficiency:*** *Under the assumptions of Theorem 5.4 and given that* γ *is smaller than a threshold independent of the discretization, the estimate* (5.15) *holds.*

The proofs are simple extensions to the proofs of Theorems 5.3 and 5.4.

A.5 Numerical study

We present the results for the test cases from Section 6.1. For this purpose, we use the modified L–scheme because of its stability and speed as discussed in Section A.1. The expression taken from Table 2 becomes $L = S'(p^{i-1}_{n,h}) + M\tau_n$, $\boldsymbol{\xi} = \mathbf{0}$ with $M = 1$ fixed throughout. Linear iterations are stopped when $\|p^{\bar{i}}_{n,h} - p^{\bar{i}-1}_{n,h}\|_{H^1_{\mathbf{K}}(\Omega)} \le 10^{-4}$. This *fixed error approach* is compared with the *adaptive approach* which follows Algorithm A.1. For this purpose, $\gamma = 0.1$ is chosen. Figure 21 and Table 3 show that the adaptive approach requires much fewer iterations while having negligible impact on the quality of solutions. Moreover, the number of iterations required is stable as opposed to the fixed error approach.

Appendix B Proofs of Section 2.5

We collect here the proofs of the statements of Section 2.5.

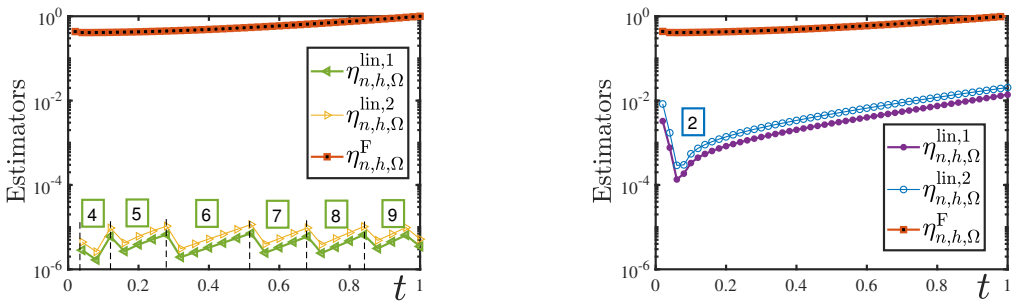

Figure 21: [Section 6.1, with adaptive linearization] The usual fixed error vs. the adaptive approach for linearization. Here $\ell = 2$. The linearization estimators from (A.4) are plotted along with $\eta^{\mathrm{F}}_{n,h,\Omega}$. Fixed error approach using $\|p^{\bar{i}}_{n,h} - p^{\bar{i}-1}_{n,h}\|_{H^1_{\bar{\mathbf{K}}}(\Omega)} \leq 10^{-4}$ as the stopping criterion (left). Adaptive approach using Algorithm A.1 with $\gamma = 0.1$ (right). The iterations required per time step are mentioned in the square boxes. They increase with time for the fixed approach and remain constant at 2 for the adaptive approach.

	Fixed error approach				Adaptive approach			
ℓ	avg. iter.	$\eta^{\mathrm{lin},1}_{n,h,\Omega}$	$\eta^{\mathrm{lin},2}_{n,h,\Omega}$	$\eta_{\mathcal{R}}$	avg. iter.	$\eta^{\mathrm{lin},1}_{n,h,\Omega}$	$\eta^{\mathrm{lin},2}_{n,h,\Omega}$	$\eta_{\mathcal{R}}$
1	7.72	3.4e-6	5.2e-6	1.859	2.00	0.021	0.038	1.869
2	6.74	5.7e-6	5.6e-6	0.998	2.00	0.014	0.020	1.088
4	5.72	1.4e-6	9.6e-7	0.497	1.98	0.007	0.009	0.506

Table 3: [Section 6.1, with adaptive linearization] Average iterations required per time step together with $\eta^{\mathrm{lin},1}_{n,h,\Omega}$, $\eta^{\mathrm{lin},2}_{n,h,\Omega}$, and $\eta_{\mathcal{R}}$ at $t_n = 1$ for the usual fixed error (left) and the adaptive linearization (right) approaches.

Proof of Proposition 2.3. We claim that $(S_{\mathrm{m}}, p_{\mathrm{c}}(S_{\mathrm{m}}))$ serves as a subsolution of (s, p) for a constant $\bar{\mathbf{K}}$. From (2.19) we have that $\bar{S}_{\mathrm{m}}(t) > S(0)$ for some $t > 0$ only if $f_{\mathrm{m}}(S(0)) \geq 0$ or $f(S(0), \boldsymbol{x}, t) \geq 0$ a.e. in $(\boldsymbol{x}, t) \in \Omega \times \mathbb{R}^+$. Hence, if $S_{\mathrm{m}}(t) = \min(\bar{S}_{\mathrm{m}}(t), S(0)) = S(0)$ in some interval $I \subseteq \mathbb{R}^+$ then $\partial_t S_{\mathrm{m}} - f(S_{\mathrm{m}}, \boldsymbol{x}, t) \leq 0 - f_{\mathrm{m}}(S(0)) \leq 0$. On the other hand, if $S_{\mathrm{m}}(t) = \bar{S}_{\mathrm{m}}(t)$ then $\partial_t S_{\mathrm{m}} - f(S_{\mathrm{m}}, \boldsymbol{x}, t) \leq \partial_t \bar{S}_{\mathrm{m}} - f_{\mathrm{m}}(\bar{S}_{\mathrm{m}}) = 0$. Hence,

$$\partial_t S_{\mathrm{m}} - \nabla \cdot [\bar{\mathbf{K}}\, \kappa(S_{\mathrm{m}})\, (\nabla p_{\mathrm{c}}(S_{\mathrm{m}}) + \boldsymbol{g})] - f(S_{\mathrm{m}}, \boldsymbol{x}, t) = \partial_t S_{\mathrm{m}} - f(S_{\mathrm{m}}, \boldsymbol{x}, t) \leq 0.$$

Moreover, $p_{\mathrm{c}}(S_{\mathrm{m}}) \leq p_{\mathrm{c}}(S(0)) = 0$ in relation to the boundary. Thus, invoking the comparison principle [33], we conclude that $(S_{\mathrm{m}}, p_{\mathrm{c}}(S_{\mathrm{m}}))$ is a subsolution of (s, p). □

Proof of Proposition 2.4. Let $J_1 := \min(J, p_l) < 0$. For the sake of simplicity, let the space coordinate be translated such that $\min\{\boldsymbol{g} \cdot \boldsymbol{x}\} = 0$. To show the lower bound of ς we use $\nabla(\boldsymbol{g} \cdot \boldsymbol{x}) = \boldsymbol{g}$, and rewrite (2.20) as

$$(\bar{\mathbf{K}}\kappa(S(\varsigma))\nabla[\varsigma + \boldsymbol{g} \cdot \boldsymbol{x}], \nabla\varphi) = \left(\inf_{t \in \mathbb{R}^+} [f(S(\varsigma), \boldsymbol{x}, t)]_-, \varphi\right). \tag{B.1}$$

Selecting the test function $\varphi = [\varsigma - J_1 + \boldsymbol{g} \cdot \boldsymbol{x}]_- \in H^1_0(\Omega)$ (observe that $\varphi = 0$ on $\partial\Omega$ since $\boldsymbol{g} \cdot \boldsymbol{x} - J_1 \geq 0$ for all $\boldsymbol{x} \in \Omega$) one then obtains in the left hand side of (B.1),

$$(\bar{\mathbf{K}}\kappa(S(\varsigma))\nabla[\varsigma + \boldsymbol{g} \cdot \boldsymbol{x}], \nabla[\varsigma - J_1 + \boldsymbol{g} \cdot \boldsymbol{x}]_-) \geq \int_\Omega \kappa(S(\varsigma)) \left|\bar{\mathbf{K}}^{\frac{1}{2}} \nabla[\varsigma - J_1 + \boldsymbol{g} \cdot \boldsymbol{x}]_-\right|^2.$$

Observe that φ is nonzero only when $\varsigma \leq J_1 - \boldsymbol{g}\cdot\boldsymbol{x} \leq p_l$, implying $f(S(\varsigma),\boldsymbol{x},t) \geq 0$. Hence, the right hand side of (B.1) yields

$$\left(\inf_{t\in\mathbb{R}^+}[f(S(\varsigma),\boldsymbol{x},t)]_-, [\varsigma - J_1 + \boldsymbol{g}\cdot\boldsymbol{x}]_-\right) = 0.$$

Hence, from (B.1), one obtains $\varsigma \geq J_1 - \boldsymbol{g}\cdot\boldsymbol{x}$. We obtain the upper bound by testing with $\varphi = [\varsigma - J + \boldsymbol{g}\cdot\boldsymbol{x} - \max\{\boldsymbol{g}\cdot\boldsymbol{x}\}]_+$ and following the arguments as before. □

Proof of Proposition 2.5. Observe that the choice of J implies from Proposition 2.4 that $\varsigma \leq 0$ and $S(\varsigma) \leq s_0$ a.e. in Ω. Moreover, from (2.20),

$$\begin{aligned}&\partial_t\varsigma - \nabla\cdot(\bar{\mathbf{K}}\kappa(S(\varsigma))[\nabla\varsigma + \boldsymbol{g}]) - f(S(\varsigma),\boldsymbol{x},t)\\ \leq& 0 - \nabla\cdot(\bar{\mathbf{K}}\kappa(S(\varsigma))[\nabla\varsigma + \boldsymbol{g}]) - \inf_{\zeta\in\mathbb{R}^+}[f(S(\varsigma),\boldsymbol{x},\zeta)]_- = 0,\end{aligned}$$

since $f(S(\varsigma),\boldsymbol{x},t) \geq \inf_{\zeta\in\mathbb{R}^+}[f(S(\varsigma),\boldsymbol{x},\zeta)]_-$. Hence, similar to the proof of Proposition 2.3, the result follows from applying the comparison principle. □

References

[1] M. Ainsworth and J.T. Oden. A posteriori error estimation in finite element analysis. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2000.

[2] H.W. Alt and S. Luckhaus. Quasilinear elliptic-parabolic differential equations. Mathematische Zeitschrift, 183(3):311–341, 1983.

[3] H.W. Alt, S. Luckhaus, and A. Visintin. On nonstationary flow through porous media. Annali di Matematica Pura ed Applicata, 136(1):303–316, 1984.

[4] T. Arbogast and M.F. Wheeler. A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media. SIAM Journal on Numerical Analysis, 33(4):1669–1687, 1996.

[5] V. Baron, Y. Coudière, and P. Sochala. Adaptive multistep time discretization and linearization based on a posteriori error estimates for the Richards equation. Applied Numerical Mathematics, 112:104–125, 2017.

[6] J. Bear. Dynamics of flow in porous media. NY: Dover, 1972.

[7] L. Bergamaschi and M. Putti. Mixed finite elements and Newton-type linearizations for the solution of Richards' equation. International Journal for Numerical Methods in Engineering, 45(8):1025–1046, 1999.

[8] C. Bernardi, L. El Alaoui, and Z. Mghazli. A posteriori analysis of a space and time discretization of a nonlinear model for the flow in partially saturated porous media. IMA Journal of Numerical Analysis, 34(3):1002–1036, 2014.

[9] K. Brenner and C. Cancès. Improving Newton's method performance by parametrization: The case of the Richards equation. SIAM Journal on Numerical Analysis, 55(4):1760–1785, 2017.

[10] C. Cancès, I.S. Pop, and M. Vohralík. An a posteriori error estimate for vertex-centered finite volume discretizations of immiscible incompressible two-phase flow. Mathematics of Computation, 83(285):153–188, 2014.

[11] M.A. Celia, E.T. Bouloutas, and R.L. Zarba. General mass-conservative numerical solution for the unsaturated flow equation. Water Resources Research, 26(7):1483–1496, 1990.

[12] A. Cohen, R. DeVore, and R.H. Nochetto. Convergence rates of AFEM with H^{-1} data. Foundations of Computational Mathematics, 12(5):671–718, 2012.

[13] D.A. Di Pietro, M. Vohralík, and S. Yousef. Adaptive regularization, linearization, and discretization and a posteriori error control for the two-phase Stefan problem. Mathematics of Computation, 84(291):153–186, 2015.
[14] V. Dolejší, A. Ern, and M. Vohralík. A framework for robust a posteriori error control in unsteady nonlinear advection-diffusion problems. SIAM Journal on Numerical Analysis, 51(2):773–793, 2013.
[15] J. Douglas Jr. and T. Dupont. Galerkin methods for parabolic equations. SIAM Journal on Numerical Analysis, 7(4):575–626, 1970.
[16] A. Ern, I. Smears, and M. Vohralík. Discrete p-robust $\boldsymbol{H}(\mathrm{div})$-liftings and a posteriori estimates for elliptic problems with H^{-1} source terms. Calcolo, 54(3):1009–1025, 2017.
[17] A. Ern, I. Smears, and M. Vohralík. Guaranteed, locally space-time efficient, and polynomial-degree robust a posteriori error estimates for high-order discretizations of parabolic problems. SIAM Journal of Numerical Analysis, 55(6):2811–2834, 2017.
[18] A. Ern and M. Vohralík. Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs. SIAM Journal on Scientific Computing, 35(4):A1761–A1791, 2013.
[19] R. Eymard, M. Gutnic, and D. Hilhorst. The finite volume method for Richards equation. Computational Geosciences, 3(3-4):259–294, 1999.
[20] R. Helmig. Multiphase flow and transport processes in the subsurface: a contribution to the modeling of hydrosystems. Springer-Verlag, 1997.
[21] W. Jäger and J. Kačur. Solution of porous medium type systems by linear approximation schemes. Numerische Mathematik, 60(1):407–427, 1991.
[22] R.A. Klausen, F.A. Radu, and G.T. Eigestad. Convergence of MPFA on triangulations and for Richards' equation. International Journal for Numerical Methods in Fluids, 58(12):1327–1351, 2008.
[23] P. Knabner. Finite element simulation of saturated-unsaturated flow through porous media. In Large Scale Scientific Computing, pages 83–93. Springer, 1987.
[24] C. Kreuzer. Reliable and efficient a posteriori error estimates for finite element approximations of the parabolic p-Laplacian. Calcolo, 50(2):79–110, 2013.
[25] M. Kubo and Q. Lu. Nonlinear degenerate parabolic equations with Neumann boundary condition. Journal of Mathematical Analysis and Applications, 307(1):232–244, 2005.
[26] R.J. Lenhard, J.C. Parker, and S. Mishra. On the correspondence between Brooks-Corey and van Genuchten models. Journal of Irrigation and Drainage Engineering, 115(4):744–751, 1989.
[27] H. Li, M.W. Farthing, C.N. Dawson, and C.T. Miller. Local discontinuous Galerkin approximations to Richards' equation. Advances in Water Resources, 30(3):555–575, 2007.
[28] F. List and F.A. Radu. A study on iterative methods for solving Richards' equation. Computational Geosciences, 20(2):341–353, 2016.
[29] K. Mitra and I.S. Pop. A modified L-scheme to solve nonlinear diffusion problems. Computers & Mathematics with Applications, 77(6):1722 – 1738, 2019. 7th International Conference on Advanced Computational Methods in Engineering (ACOMEN 2017).
[30] R. H. Nochetto, A. Schmidt, and C. Verdi. A posteriori error estimation and adaptivity for degenerate parabolic problems. Mathematics of Computation, 69(229):1–24, 2000.
[31] R.H. Nochetto and C. Verdi. Approximation of degenerate parabolic problems using numerical integration. SIAM Journal on Numerical Analysis, 25(4):784–814, 1988.
[32] M. Ohlberger. A posteriori error estimate for finite volume approximations to singularly perturbed nonlinear convection–diffusion equations. Numerische Mathematik, 87(4):737–761, 2001.
[33] F. Otto. L^1-contraction and uniqueness for quasilinear elliptic–parabolic equations. Journal

of Differential Equations, 131(1):20–38, 1996.

[34] A.A.H. Oulhaj, C. Cancès, and C. Chainais-Hillairet. Numerical analysis of a nonlinearly stable and positive control volume finite element scheme for Richards equation with anisotropy. ESAIM: Mathematical Modelling and Numerical Analysis, 52(4):1533–1567, 2018.

[35] M. Picasso. Adaptive finite elements for a linear parabolic problem. Computer Methods in Applied Mechanics and Engineering, 167(3-4):223–237, 1998.

[36] F.A. Radu and W. Wang. Convergence analysis for a mixed finite element scheme for flow in strictly unsaturated porous media. Nonlinear Analysis: Real World Applications, 15:266–275, 2014.

[37] Sergey Repin. A posteriori estimates for partial differential equations, volume 4 of Radon Series on Computational and Applied Mathematics. Walter de Gruyter GmbH & Co. KG, Berlin, 2008.

[38] R. Verfürth. A posteriori error estimates for nonlinear problems. $L^r(0,T;L^\rho(\Omega))$-error estimates for finite element discretizations of parabolic equations. Mathematics of Computation, 67(224):1335–1360, 1998.

[39] R. Verfürth. A posteriori error estimates for nonlinear problems: $L^r(0,T;W^{1,\rho}(\Omega))$-error estimates for finite element discretizations of parabolic equations. Numerical Methods Partial Differential Equations, 14(4):487–518, 1998.

[40] R. Verfürth. A posteriori error estimates for finite element discretizations of the heat equation. Calcolo, 40(3):195–212, 2003.

[41] R. Verfürth. A posteriori error estimation techniques for finite element methods. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2013.

[42] Y. Zha, J. Yang, J. Zeng, C.H.M. Tso, W. Zeng, and L. Shi. Review of numerical solution of Richardson–Richards equation for variably saturated flow in soils. Wiley Interdisciplinary Reviews: Water, 6(5):e1364, 2019.