An a posteriori error estimate for vertex-centered finite volume discretizations of immiscible incompressible two-phase flow

Clément Cancès, Iuliu Sorin Pop, Martin Vohralík

An a posteriori error estimate for vertex-centered finite volume discretizations of immiscible incompressible two-phase flow*

Clément Cancès† Iuliu Sorin Pop‡ Martin Vohralík†

September 13, 2011

Abstract

In this paper we derive an a posteriori error estimate for the numerical approximation of the solution of a system modeling the flow of two incompressible and immiscible fluids in a porous medium. We take into account the capillary pressure, which leads to a coupled system of two equations: parabolic and elliptic. The parabolic equation may become degenerate, i.e., the nonlinear diffusion coefficient may vanish over regions that are not known a priori. We first show that, under appropriate assumptions, the energy-type-norm differences between the exact and the approximate nonwetting phase saturations, the global pressures, and the Kirchhoff transforms of the nonwetting phase saturations can be bounded by the dual norm of the residuals. We then bound the dual norm of the residuals by fully computable a posteriori estimators. Our analysis covers a large class of conforming, vertex-centered finite volume-type discretizations with fully implicit time stepping. As an example, we focus here on two approaches: a "mathematical" scheme derived from the weak formulation, and a phase-by-phase upstream weighting "engineering" scheme. Finally, we show how the different error components, namely the space discretization error, the time discretization error, the linearization error, the algebraic solver error, and the quadrature error can be distinguished and used for making the calculations efficient.

Keywords: Two-phase flow, porous media flow, a posteriori error estimate, finite volume schemes

AMS subject classification: 65M08, 65M50, 76S05, 76T99

1 Introduction

Two-phase porous media flow models are of fundamental importance in various real life applications, such as petroleum reservoir engineering or CO_2 sequestration in the subsurface. Such processes can be modelled by a system consisting of two equations: an elliptic one for the total velocity, coupled to a parabolic one for the nonwetting phase saturation, see, e.g., [7, 12, 8]. In the latter equation, the diffusion coefficient depends nonlinearly on the unknown quantities and vanishes over regions that are not known a priori and can vary in time and space, leading to a degenerate, free boundary problem. Our aim here is to develop a rigorous a posteriori error estimate for such a model.

A large amount of publications are devoted to the mathematical and numerical analysis of two-phase flow models. In particular, the existence, uniqueness, and regularity of a (weak) solution are studied in [32, 12, 5, 6, 14, 15]. In the same spirit, much work has been carried out for developing appropriate numerical methods and proving their convergence, or a priori error estimates, like in [16] for a finite element discretization. In this paper we focus on the finite volume method [26]. In this context, the convergence of a cell-centered "mathematical" scheme involving the global pressure and the Kirchhoff transform has been obtained in [34]. Alternatively, the convergence of a cell-centered finite volume scheme with phase-by-phase upstream weighting (the so-called "engineering" scheme) has been shown in [27]. Vertex-centered finite volume methods in the

†UPMC Univ. Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, 75005, Paris, France & CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, 75005, Paris, France (cances@ann.jussieu.fr, vohralik@ann.jussieu.fr).

‡Dept. of Math. and Comp. Sci., Eindhoven University of Technology, P.O. Box 513, 5600MB, Eindhoven, the Netherlands (i.pop@tue.nl).

*This work was partly supported by the Groupement MoMaS (PACEN/CNRS, ANDRA, BRGM, CEA, EdF, IRSN) and by the ERT project "Enhanced oil recovery and geological sequestration of CO_2: mesh adaptivity, a posteriori error control, and other advanced techniques" (LJLL/IFPEN).

"mathematical" context have been studied in, e.g., [25], and in the "engineering" context in, e.g., [29], see also the references therein.

To the best of our knowledge, contrarily to the case of a priori error estimates, almost no results are available for rigorous a posteriori error estimates for the two-phase flow model. The arguments used in [17] are rather of a priori type. The results of [4] refer to the density-driven flow in porous media, whereas an posteriori error estimate for miscible displacement of one incompressible fluid by another can be found in [13]. Recently, a framework for a posteriori error estimation of the dual norm of the residuals for the two-phase flow problem has been derived in [51]. It has been applied to the cell-centered finite volume phase-by-phase upstream weighting scheme in [20]. Rigorous a posteriori error estimates for nonlinear, time-dependent problems are obtained in [22, 28, 49, 36, 38, 39, 18, 2, 19], see also the references given therein; for basic results on a posteriori error estimates, in particular for linear elliptic model problems, we refer to the textbooks [48, 1, 35, 46] and to the references therein.

The content of this paper is as follows. In Section 2 we introduce the immiscible incompressible two-phase flow model. The governing physical equations are given in Section 2.1, while Section 2.2, provides the mathematical formulation relying on the Kirchhoff transform of the nonwetting saturation (sometimes called the "complementary pressure") and on the global pressure. The physical meaning of these mathematical quantities is less obvious, but they are needed for giving a proper definition of the weak solution. The existence and uniqueness of a weak solution is guaranteed under certain assumptions on the data and on the model parameters, which are summarized in Assumption A.

In Section 3, we give the main result of the paper, Theorem 1. This theorem states that the energy-type-norm of the differences between the exact and the approximate nonwetting phase saturations, the global pressures, and the Kirchhoff transforms of the nonwetting phase saturations can be bounded by a fully computable a posteriori error estimate. This theorem is formulated as generally as possible; in particular, it does not require specifying the underlying discretization. We merely need the technical Assumption B on the data and the reconstructions $\mathbf{u}_{\mathrm{n},h\tau}$ and $\mathbf{u}_{\mathrm{w},h\tau}$ of the Darcy fluxes for each of the two phases. These are vector fields, constant on each time interval and belonging on each time interval to the functional space $\mathbf{H}(\mathrm{div},\Omega)$, with continuous normal trace over any $d-1$ dimensional manifold, and satisfying a local conservation over the mesh elements, as summarized in Assumption C. Such an approach develops those used in [50, 23, 51], see also the references therein, and relies on concepts going back to the Prager–Synge equality [43] for linear elliptic problems.

In Section 4 we apply the abstract result of Theorem 1 to particular finite volume discretizations. This implies specifying the reconstruction of the phase fluxes (in practice, $\mathbf{u}_{\mathrm{n},h\tau}$ and $\mathbf{u}_{\mathrm{w},h\tau}$ are constructed in the Raviart–Thomas–Nédélec finite-dimensional subspaces of $\mathbf{H}(\mathrm{div},\Omega)$) and verifying the Assumption C. These steps are carried out for two quite distinct vertex-centered finite volume schemes, a "mathematical" one derived from the weak formulation and a phase-by-phase upstream weighting "engineering" one.

Section 5 is devoted to the proof of the a posteriori error estimate. We first define the residuals stemming from the weak formulation in Section 5.1. Next, in Section 5.2, we show that under Assumption A the energy-type-norm of the differences between the exact and the approximate solutions can be bounded by the dual norm of the residuals. The result is stated in Theorem 2. Next, under Assumptions B, C, we show in Section 5.3 that the dual norm of the residuals is bounded by a computable a posteriori error estimate. This result is stated in Theorem 3.

Finally, in Appendix A we focus on the particular case of the "mathematical" scheme and apply the methodology developed in [31, 21, 23, 24, 51] to obtain Corollary A.3, showing how the estimators of Theorem 1 can be used to distinguish the different error components. These components are namely the space discretization error, the time discretization error, the linearization error, the algebraic solver error, and the quadrature error. We demonstrate how they can be employed to stop the various iterative procedures and to equilibrate the spatial and temporal errors in order to use the computational resources as efficiently as possible.

2 The immiscible, incompressible two-phase flow in porous media

In this section we give the mathematical model for the immiscible incompressible two-phase flow in a porous medium and bring it in a form that is more suitable for the mathematical and numerical analysis. Then we state the assumptions on the model parameters and the data, define the weak solution, and recall its existence and uniqueness.

2.1 The governing equations

For the ease of reading the model under discussion is presented in a dimensionless context. Given a porous medium occupying an open, bounded, polyhedral subset $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, consider two incompressible and immiscible phases flowing within the pores of the medium. For simplicity we restrict ourselves to the case of horizontal flow and thus neglect the gravity effects. With $\alpha \in \{\mathrm{n}, \mathrm{w}\}$ being the index for the nonwetting, respectively the wetting phase, the unknown quantities are the phase saturations s_α and pressures p_α, as well as the Darcy velocities $\mathbf{u}_\alpha$. The saturations are assumed reduced, thus taking (physical) values between 0 and 1. For each phase, the velocity and the pressure are related by the Darcy–Muskat law

$$\mathbf{u}_\alpha := -\underline{\mathbf{K}}\eta_\alpha(s_\alpha)\nabla p_\alpha, \qquad \alpha \in \{\mathrm{n}, \mathrm{w}\}. \tag{2.1}$$

Above, $\underline{\mathbf{K}}$ is the (intrinsic) permeability tensor, which is assumed symmetric and uniformly positive definite. Here we allow $\underline{\mathbf{K}}$ to be location-dependent, $\underline{\mathbf{K}} = \underline{\mathbf{K}}(\mathbf{x})$. Further, the mobilities η_α are functions of the phase saturations s_α, $\eta_\alpha = \eta_\alpha(s_\alpha)$. Their specific form depends on the medium and on the phase and is determined experimentally. In particular, these functions are continuous and increasing on $[0, 1]$, satisfying

$$\eta_\alpha(0) = 0, \qquad \alpha \in \{\mathrm{n}, \mathrm{w}\}.$$

Note that this implies the boundedness of η_α. For mathematical completeness we extend the functions η_α by constants outside the physically relevant interval $[0, 1]$,

$$\eta_\alpha(s_\alpha \le 0) := 0 \text{ and } \eta_\alpha(s_\alpha \ge 1) := \eta_\alpha(1). \tag{2.2}$$

Disregarding the porosity of the medium, which is allowed after a proper scaling of the time, the mass balance for each phase gives (see, e.g., [12, 8])

$$\partial_t s_\alpha + \nabla\cdot\mathbf{u}_\alpha = q_\alpha(s_\alpha), \qquad \alpha \in \{\mathrm{n}, \mathrm{w}\}, \tag{2.3}$$

where the source terms q_α are given functions of the phase saturations. Inserting (2.1) into (2.3) allows to eliminate the Darcy velocities $\mathbf{u}_\alpha$. Note that a vanishing mobility η_α, which is encountered whenever $s_\alpha \le 0$, leads to a degeneracy in (2.3). In this case the second term on the left becomes 0, and the equation looses its originally parabolic character.

We further assume that the volume of all pores is filled by the two phases (thus no other fluid phase is present), implying

$$s_\mathrm{n} + s_\mathrm{w} = 1. \tag{2.4}$$

Under equilibrium conditions at the pore scale, the phase pressures p_w and p_n are related by

$$p_\mathrm{n} - p_\mathrm{w} = \pi(s_\mathrm{n}), \tag{2.5}$$

where π, the capillary pressure, is an increasing function.

Defining the total velocity

$$\mathbf{u}_\mathrm{t} := -\underline{\mathbf{K}}\left(\eta_\mathrm{n}(s_\mathrm{n})\nabla p_\mathrm{n} + \eta_\mathrm{w}(s_\mathrm{w})\nabla p_\mathrm{w}\right) \tag{2.6}$$

and adding both equations (2.3) for $\alpha = \mathrm{w}, \mathrm{n}$, thanks to (2.4), one gets

$$\nabla\cdot\mathbf{u}_\mathrm{t} = q_\mathrm{n}(s_\mathrm{n}) + q_\mathrm{w}(1 - s_\mathrm{n}) =: q_\mathrm{t}(s_\mathrm{n}). \tag{2.7}$$

Using (2.6) in (2.3) for $\alpha = \mathrm{n}$ provides

$$\partial_t s_\mathrm{n} + \nabla\cdot\left(\mathbf{u}_\mathrm{t} f(s_\mathrm{n}) - \underline{\mathbf{K}}\lambda(s_\mathrm{n})\nabla\pi(s_\mathrm{n})\right) = q_\mathrm{n}(s_\mathrm{n}), \tag{2.8}$$

where the nonlinear functions f and λ are defined as

$$f(s) := \frac{\eta_\mathrm{n}(s)}{\eta_\mathrm{n}(s) + \eta_\mathrm{w}(1 - s)}, \quad \lambda(s) := \eta_\mathrm{w}(1 - s)f(s).$$

The problem is completed by initial and boundary conditions, introduced below after a suitable reformulation.

2.2 A mathematical formulation

The mathematical results below are expressed in terms of the nonwetting phase saturation, denoted from now on by s, i.e.,

$$s := s_{\mathrm{n}}.$$

Clearly, the wetting phase saturation is then given by $s_{\mathrm{w}} = 1 - s$. Next we reformulate the equations (2.6) and (2.8) in terms of more convenient unknowns. This involves the following constructions. First, as in [3], we define the Kirchhoff transform as

$$\varphi(s) := \int_0^s \lambda(a)\pi'(a)\,\mathrm{d}a, \tag{2.9}$$

and observe that φ is increasing on $[0,1]$. Next, we follow [12, 5] and introduce the global pressure P, defined by

$$P := P(s, p_{\mathrm{w}}) := p_{\mathrm{w}} + \int_0^{\pi(s)} \frac{\eta_{\mathrm{n}}(\pi^{-1}(a))}{\eta_{\mathrm{n}}(\pi^{-1}(a)) + \eta_{\mathrm{w}}(1-\pi^{-1}(a))}\,\mathrm{d}a \tag{2.10a}$$

$$= P(s, p_{\mathrm{n}}) := p_{\mathrm{n}} - \int_0^{\pi(s)} \frac{\eta_{\mathrm{w}}(1-\pi^{-1}(a))}{\eta_{\mathrm{n}}(\pi^{-1}(a)) + \eta_{\mathrm{w}}(1-\pi^{-1}(a))}\,\mathrm{d}a. \tag{2.10b}$$

Using these definitions in (2.6) gives

$$\mathbf{u}_{\mathrm{t}} = -\underline{\mathbf{K}} M(s) \nabla P, \tag{2.11}$$

where

$$M(s) := \eta_{\mathrm{w}}(1-s) + \eta_{\mathrm{n}}(s).$$

The equation (2.7) then becomes

$$-\nabla\cdot(\underline{\mathbf{K}} M(s)\nabla P) = q_{\mathrm{t}}(s). \tag{2.12}$$

Similarly to the extension (2.2) of η_{n} and η_{w}, the functions f, λ, and M defined above are extended continuously by constants outside of $[0,1]$. Clearly, M is uniformly bounded away from 0 over the entire real axis. From now on, the function η_{w} will not appear explicitly anymore. For the ease of reading we therefore remove the subscript n in η_{n}, i.e., we use

$$\eta(s) := \eta_{\mathrm{n}}(s).$$

This allows rewriting (2.8) into

$$\partial_t s - \nabla\cdot(\underline{\mathbf{K}}(\eta(s)\nabla P + \nabla\varphi(s))) = q_{\mathrm{n}}(s). \tag{2.13}$$

After having done these steps, we consider the problem on the time interval $(0,T]$ for some $T > 0$ and prescribe the initial data

$$s(\cdot, 0) = s^0. \tag{2.14}$$

For the sake of simplicity, only Dirichlet boundary conditions for the saturation and the global pressure are considered, i.e.,

$$s|_{\partial\Omega\times(0,T)} = \overline{s}, \quad P|_{\partial\Omega\times(0,T)} = \overline{P}, \tag{2.15}$$

where $\overline{s}$ and $\overline{P}$ are given functions. The generalization to inhomogeneous Dirichlet conditions on a part of $\partial\Omega$ and to inhomogeneous Neumann condition on its complement is possible, by following the steps described in [16]. However, this leads to more technicalities and notations that would affect the clarity of the exposition.

For any $t \in (0,T]$ we use the notations:

$$Q_t := \Omega\times(0,t], \quad \text{and} \quad \mathbf{1}_{(0,t)}(\tau) := \begin{cases} 1 & \text{if } \tau\in(0,t), \\ 0 & \text{otherwise.}\end{cases}$$

To define a solution in the weak sense, we make use of common notations in the functional analysis. In particular, $H^{-1}(\Omega)$ is the dual of $H_0^1(\Omega)$ and $\langle\cdot,\cdot\rangle$ denotes the corresponding duality pairing. Let

$$\begin{aligned}\mathcal{E} :=& \big\{(s,P) \mid s\in\mathcal{C}([0,T];L^2(\Omega)),\ \partial_t s\in L^2((0,T);H^{-1}(\Omega)),\\ & \varphi(s)-\varphi(\overline{s})\in L^2((0,T);H_0^1(\Omega)),\ P-\overline{P}\in L^2((0,T);H_0^1(\Omega))\big\}.\end{aligned} \tag{2.16}$$

Then a weak solution of (2.12), (2.13) with the initial and boundary condition (2.14), (2.15) is introduced by

Definition 2.1 (Weak solution). *A weak solution is a pair $(s,P)\in\mathcal{E}$ such that $s(\cdot,0)=s^0$ and for all $\psi\in L^2((0,T);H^1_0(\Omega))$, there holds*

$$\int_0^T \langle \partial_t s(\cdot,\theta);\psi(\cdot,\theta)\rangle_{H^{-1},H^1_0}\,\mathrm{d}\theta + \iint_{Q_T} \underline{\mathbf{K}}\,(\eta(s)\nabla P+\nabla\varphi(s))\cdot\nabla\psi\,\mathrm{d}\mathbf{x}\mathrm{d}\theta = \iint_{Q_T} q_{\mathrm{n}}(s)\psi\,\mathrm{d}\mathbf{x}\mathrm{d}\theta, \tag{2.17a}$$

$$\iint_{Q_T} \underline{\mathbf{K}} M(s)\nabla P\cdot\nabla\psi\,\mathrm{d}\mathbf{x}\mathrm{d}\theta = \iint_{Q_T} q_{\mathrm{t}}(s)\psi\,\mathrm{d}\mathbf{x}\mathrm{d}\theta. \tag{2.17b}$$

The results in this paper are obtained under the following assumptions on the model:

Assumption A (Data and weak solution).

1. *The functions $M,\eta:\mathbb{R}\to\mathbb{R}$ are continuous and there exist positive constants c_M, C_M, and C_η such that, for all $a\in[0,1]$,*
$$c_M\le M(a)\le C_M,\qquad \eta(a)\le C_\eta. \tag{2.18}$$
2. *The diffusion tensor $\underline{\mathbf{K}}\in[L^\infty(\Omega)]^{d\times d}$ is symmetric and uniformly positive definite.*
3. *The function $\overline{P}$ in (2.15) belongs to $L^\infty((0,T);H^{1/2}(\partial\Omega))$. Thus there exists an extension, still denoted by $\overline{P}$, such that*
$$\overline{P}\in L^\infty((0,T);H^1(\Omega)).$$
Similarly, the function $\overline{s}$ in (2.15) belongs to $L^\infty(\partial\Omega\times(0,T))$ with $0\le\overline{s}\le 1$. Moreover, $\overline{s}$ can be extended on Q_T into a measurable function, still denoted by $\overline{s}$, such that
$$\partial_t\overline{s}\in L^1(Q_T),\quad \varphi(\overline{s})\in L^2((0,T);H^1(\Omega)),\quad \overline{s}(\cdot,0)=s^0.$$
4. *Concerning the sources $q_{\mathrm{n}},q_{\mathrm{w}}$ (and $q_{\mathrm{t}}=q_{\mathrm{w}}+q_{\mathrm{n}}$), we assume that for all $(\mathbf{x},t)\in Q_T$, the functions*
$$q_\alpha(\cdot;\mathbf{x},t):\left\{\begin{array}{rcl}[0,1] & \to & \mathbb{R},\\ s & \mapsto & q_\alpha(s;\mathbf{x},t),\end{array}\right.\qquad(\alpha\in\{\mathrm{n},\mathrm{w}\})$$
are Lipschitz continuous. Hence, for all $v\in L^\infty(Q_T)$ with $0\le v\le 1$ a.e. in Q_T, one has
$$q_{\mathrm{n}}(v)\in L^2(Q_T),\qquad q_{\mathrm{t}}(v)\in L^\infty((0,T);L^2(\Omega)).$$
We moreover assume that
$$q_{\mathrm{n}}(0;\cdot,\cdot)\ge 0,\qquad q_{\mathrm{w}}(1;\cdot,\cdot)\ge 0.$$
5. *The initial saturation satisfies*
$$s^0\in L^\infty(\Omega)\ \text{ with }\ 0\le s^0\le 1\ \text{ a.e. in }\Omega.$$
6. *The Kirchhoff transform function φ defined in (2.9) is increasing in $[0,1]$ and L_φ-Lipschitz continuous.*
7. *There exists a positive constant C_0 such that, for all $(a,b)\in\mathbb{R}$ and for almost all $(\mathbf{x},t)\in Q_T$,*
$$\begin{aligned}&(\eta(a)-\eta(b))^2+(M(a)-M(b))^2\\ &\quad+(q_{\mathrm{n}}(a;\mathbf{x},t)-q_{\mathrm{n}}(b;\mathbf{x},t))^2+(q_{\mathrm{t}}(a;\mathbf{x},t)-q_{\mathrm{t}}(b;\mathbf{x},t))^2\le C_0(a-b)(\varphi(a)-\varphi(b)).\end{aligned} \tag{2.19}$$
8. *There exists a weak solution (s,P) in the sense of Definition 2.1 which is such that $\nabla P\in[L^\infty(Q_T)]^d$.*

As for η, see (2.2), φ is extended on $\mathbb{R}$ by

$$\varphi(s)=\left\{\begin{array}{ll}L_\varphi s & \text{if } s<0,\\ L_\varphi(s-1)+\varphi(1) & \text{if } s>1.\end{array}\right. \tag{2.20}$$

Here L_φ is the minimal Lipschitz constant of φ on $[0,1]$. In this way the properties assumed above for the interval $[0,1]$ extend trivially to $\mathbb{R}$.

The assumptions stated above deserve some comments. Points 1 and 6 are satisfied by most of the one- or two-phase porous media flow models currently used in oil engineering. Point 3 is natural and does not impose any severe restrictions on the boundary data. As mentioned above, one can apply the techniques in [16] to extend the present results to inhomogeneous Neumann boundary conditions that are prescribed on some parts of the boundary.

For point 4, since $q_{\rm t}(s)$ belongs to $L^\infty((0,T);L^2(\Omega))$, the total velocity $\mathbf{u}_{\rm t}$ is essentially bounded in $\mathbf{H}(\mathrm{div},\Omega)$ with respect to time. Moreover, the last assertion of this point is nothing but claiming that one cannot extract a missing phase.

The condition (2.19) appearing in point 7 is similar to Assumption (A7) in [16] (see also [14]). For scalar degenerate parabolic equations, it ensures the uniqueness of a solution (see [3, 40]). This condition can further be employed for deriving a priori error estimates (see [41, 45]), and is mainly relevant for the behavior of η close to the degeneracy values, 0 and 1. For example, referring to the van Genuchten curves relating the permeability and the dynamic capillarity to the saturation (see, e.g., [8]), (2.19) holds if the van Genuchten parameters m and n are such that $n = 1/(1-m)$ and $m \in [2/3, 1)$.

Concerning point 8, it obviously requires more analysis since a weak solution as introduced in Definition 2.1 does not necessary fulfill the requirement on the pressure gradient. For domains Ω having a smooth boundary, [14, Theorem 4.5] provides the essential boundedness of ∇P. This result is, however, not usable here as we assume Ω as polyhedral.

Finally, it is worth mentioning that the assumptions in the last two points are not needed for the existence of a solution (see, e.g., [5, 6, 14]), but are stated here since these will be used later. Essentially we use the following existence and uniqueness result proved in [14].

Corollary 2.2 (Existence and uniqueness)**.** *Under Assumption A, there exists a unique weak solution to the problem (2.12)–(2.15) in the sense of Definition 2.1.*

Remark 2.3 (Continuity in time of the saturation)**.** *The space $\mathcal{E}$ in (2.16) requires that s is continuous in time. To justify this we recall (2.13), (2.8), and (2.7) and note that if (s,P) is a weak solution, the equation (2.17a) can formally be written as*

$$\partial_t s + \nabla\cdot(\mathbf{u}_{\rm t} f(s) - \underline{\mathbf{K}}\nabla\varphi(s)) = q_{\rm n}(s).$$

For fixed $\mathbf{u}_{\rm t}$, this operator involves a L^1-contraction semi-group with a comparison principle [11, 33, 40]. Thanks to Assumption A4, $s=0$ is a sub-solution, while $s=1$ is a super-solution. Therefore $0\le s\le 1$ a.e. in Q_T. The fact that $s\in\mathcal{C}([0,T];L^2(\Omega))$ then follows from [10].

3 The a posteriori error estimate

This section provides the main result, an abstract a posteriori estimate on the difference between the exact and the approximate solutions. This is obtained in the context of an Euler implicit time stepping, whereas the spatial discretization is left unspecified.

3.1 Time mesh and some additional notations and assumptions

We consider a strictly increasing sequence of discrete times $\{t^n\}_{0\le n\le N}$ such that $t^0=0$ and $t^N=T$. For all $1\le n\le N$, we define the time interval $I_n := (t^{n-1},t^n]$ and the time step $\tau^n := t^n - t^{n-1}$. For each $0\le n\le N$, we consider a partition $\mathcal{D}_h^n$ of Ω. We denote by $\mathcal{D}_h^{\mathrm{ext},n}$ the volumes from $\mathcal{D}_h^n$ having an intersection with $\partial\Omega$ of nonzero measure and by $\mathcal{D}_h^{\mathrm{int},n}$ the remaining elements of $\mathcal{D}_h^n$. An example is given in Section 4.1 below. The following weighted norm on subsets D of Ω, for $\mathbf{v}\in[L^2(D)]^d$, will be used often below:

$$\|\mathbf{v}\|_{\underline{\mathbf{K}}^{-\frac12};L^2(D)} := \left\{\int_D \left|\underline{\mathbf{K}}^{-\frac12}(\mathbf{x})\mathbf{v}(\mathbf{x})\right|^2 \mathrm{d}\mathbf{x}\right\}^{\frac12}.$$

We now define the following space:

$$V_\tau := \{v\in\mathcal{C}([0,T];L^2(\Omega)),\ v \text{ is affine in time on each } I_n \text{ for all } 1\le n\le N\}.$$

Further, for $0\le n\le N$, we let v^n stand for the function $v(\cdot,t^n)$. Note that for functions $v\in V_\tau$, $\partial_t v|_{I_n} = (v^n - v^{n-1})/\tau^n$, where $v|_{I_n}$ denotes the restriction of v to the time interval I_n.

In addition to Assumption A, we now make the following:

Assumption B (Boundary conditions and sources)**.**

1. *The boundary condition for the saturation* $\overline{s}$ *is continuous and piecewise affine in time,* $\overline{s} \in V_\tau$.
2. *The source functions* q_n *and* q_w *are piecewise constant in time, with values in* $L^2(\Omega)$.

Since q_α, $\alpha \in \{\mathrm{n}, \mathrm{t}\}$, are assumed piecewise constant in time, we set $q_\alpha^n := q_\alpha|_{I_n}$ for all $n = 1, \dots N$.

Remark 3.1 (Boundary conditions and sources)**.** *Assumption B is made only for the clarity of presentation. More general boundary conditions and source terms can be taken into account, giving rise to additional error terms in the analysis carried out below.*

Having in mind the time discretization introduced above, relying on the space V_τ, we consider the following restriction of the set $\mathcal{E}$ introduced in (2.16)

$$\begin{aligned}\mathcal{E}_\tau :=\big\{&(s,P) \mid s \in V_\tau,\ \partial_t s \in L^2(Q_T),\\ &\varphi(s) - \varphi(\overline{s}) \in L^2((0,T);H_0^1(\Omega)),\ P - \overline{P} \in L^2((0,T);H_0^1(\Omega))\big\}.\end{aligned} \tag{3.1}$$

3.2 Reconstructions of the phase fluxes

Let an arbitrary pair $(s_{h\tau}, P_{h\tau}) \in \mathcal{E}_\tau$ be given. In order to proceed in a fairly general manner, particularly without specifying the discretization scheme, we make the following assumption:

Assumption C (Locally conservative fluxes reconstructions)**.** *There exist two vector fields* $\mathbf{u}_{\mathrm{n},h\tau}$ *and* $\mathbf{u}_{\mathrm{t},h\tau}$*, piecewise constant in time, such that*

$$\mathbf{u}_{\mathrm{n},h}^n := \mathbf{u}_{\mathrm{n},h\tau}|_{I_n},\ \mathbf{u}_{\mathrm{t},h}^n := \mathbf{u}_{\mathrm{t},h\tau}|_{I_n} \in \mathbf{H}(\mathrm{div},\Omega) \qquad \textit{for all } n \in \{1,\dots,N\}$$

and such that

$$\int_D \left(\frac{s_h^n - s_h^{n-1}}{\tau^n} + \nabla\cdot\mathbf{u}_{\mathrm{n},h}^n\right)\mathrm{d}\mathbf{x} = \int_D q_\mathrm{n}^n(s_h^n)\,\mathrm{d}\mathbf{x} \qquad \textit{for all } n \in \{1,\dots,N\} \textit{ and for all } D \in \mathcal{D}_h^{\mathrm{int},n}, \tag{3.2a}$$

$$\int_D \nabla\cdot\mathbf{u}_{\mathrm{t},h}^n\,\mathrm{d}\mathbf{x} = \int_D q_\mathrm{t}^n(s_h^n)\,\mathrm{d}\mathbf{x} \qquad \textit{for all } n \in \{1,\dots,N\} \textit{ and for all } D \in \mathcal{D}_h^{\mathrm{int},n}. \tag{3.2b}$$

The function $\mathbf{u}_{\mathrm{n},h}^n$ will be called *nonwetting phase flux reconstruction*, whereas the function $\mathbf{u}_{\mathrm{t},h}^n$ will be called *total flux reconstruction*. These two functions are discrete counterparts of the nonwetting phase flux $\mathbf{u}_\mathrm{n}$ in (2.1) (with $\alpha = \mathrm{n}$), respectively of the total flux $\mathbf{u}_\mathrm{t}$ in (2.6). These fluxes need to be constructed from the given numerical scheme, see Sections 4.2.2 and 4.3.2 below for two examples. Remark that (3.2a)–(3.2b) represents a discrete weak form of the continuous mass balance equation (2.3) for $\alpha = \mathrm{n}$, and of (2.7). Finally, note that with $\mathbf{u}_{\mathrm{w},h\tau} := \mathbf{u}_{\mathrm{t},h\tau} - \mathbf{u}_{\mathrm{n},h\tau}$, one gets from (3.2a)–(3.2b)

$$\int_D \left(-\frac{s_h^n - s_h^{n-1}}{\tau^n} + \nabla\cdot\mathbf{u}_{\mathrm{w},h}^n\right)\mathrm{d}\mathbf{x} = \int_D q_\mathrm{w}^n(1 - s_h^n)\,\mathrm{d}\mathbf{x} \qquad \text{for all } n \in \{1,\dots,N\} \text{ and for all } D \in \mathcal{D}_h^{\mathrm{int},n}, \tag{3.3}$$

which is a fully discrete counterpart of (2.3) for $\alpha = \mathrm{w}$.

3.3 The estimators

We can now define the a posteriori error estimators. For given $n \in \{1,\dots,N\}$, $t \in I_n$, and $D \in \mathcal{D}_h^n$, define the *diffusive flux estimators*

$$\eta_{\mathrm{DF},\mathrm{n},D}^n(t) := \|\mathbf{u}_{\mathrm{n},h}^n + \underline{\mathbf{K}}\left(\eta(s_{h\tau})\nabla P_{h\tau} + \nabla\varphi(s_{h\tau})\right)(t)\|_{\underline{\mathbf{K}}^{-\frac{1}{2}};L^2(D)}, \tag{3.4a}$$

$$\eta_{\mathrm{DF},\mathrm{t},D}^n(t) := \|\mathbf{u}_{\mathrm{t},h}^n + \underline{\mathbf{K}}M(s_{h\tau})\nabla P_{h\tau}(t)\|_{\underline{\mathbf{K}}^{-\frac{1}{2}};L^2(D)} \tag{3.4b}$$

and the *residual estimators*

$$\eta_{\mathrm{R},\mathrm{n},D}^n := m_D\|\partial_t s_{h\tau} + \nabla\cdot\mathbf{u}_{\mathrm{n},h}^n - q_\mathrm{n}^n(s_h^n)\|_{L^2(D)}, \tag{3.5a}$$

$$\eta_{\mathrm{R},\mathrm{t},D}^n := m_D\|\nabla\cdot\mathbf{u}_{\mathrm{t},h}^n - q_\mathrm{t}^n(s_h^n)\|_{L^2(D)}. \tag{3.5b}$$

Here $m_D = C_{\mathrm{P},D} h_D \, c_{\underline{\mathbf{K}},D}^{-\frac{1}{2}}$ if $D \in \mathcal{D}_h^{\mathrm{int},n}$, respectively $m_D = C_{\mathrm{F},D,\partial\Omega} h_D \, c_{\underline{\mathbf{K}},D}^{-\frac{1}{2}}$ if $D \in \mathcal{D}_h^{\mathrm{ext},n}$. The notation h_D stands for the diameter of the volume D, whereas $c_{\underline{\mathbf{K}},D}$ stands for the smallest eigenvalue of the tensor $\underline{\mathbf{K}}$ on the volume D. The constant $C_{\mathrm{P},D}$, $D \in \mathcal{D}_h^{\mathrm{int},n}$, appears in the Poincaré–Wirtinger inequality

$$\|\varphi - \varphi_D\|_{L^2(D)} \le C_{\mathrm{P},D} h_D \|\nabla\varphi\|_{L^2(D)} \qquad \forall \varphi \in H^1(D), \tag{3.6}$$

where φ_D is the mean value of the function φ over D given by $\varphi_D := \int_D \varphi \, \mathrm{d}\mathbf{x}/|D|$ ($|D|$ is the measure of D). For a convex D, $C_{\mathrm{P},D}$ can be evaluated as $1/\pi$. Similarly, $C_{\mathrm{F},D,\partial\Omega}$, $D \in \mathcal{D}_h^{\mathrm{ext},n}$, appears in the Poincaré–Friedrichs inequality

$$\|\varphi\|_{L^2(D)} \le C_{\mathrm{F},D,\partial\Omega} h_D \|\nabla\varphi\|_{L^2(D)} \quad \forall \varphi \in H^1(D) \text{ such that } \varphi = 0 \text{ on } \partial\Omega \cap \partial D; \tag{3.7}$$

$C_{\mathrm{F},D,\partial\Omega}$ can be typically taken equal to 1. We refer for more details to [50] and the references therein. Finally, the *time quadrature estimators* are given by

$$\eta_{\mathrm{Q,n},D}^n(t) := C_{\mathrm{F},\Omega} h_\Omega c_{\underline{\mathbf{K}},\Omega}^{-\frac{1}{2}} \|q_{\mathrm{n}}^n(s_h^n) - q_{\mathrm{n}}^n(s_{h\tau})(t)\|_{L^2(D)}, \tag{3.8a}$$

$$\eta_{\mathrm{Q,t},D}^n(t) := C_{\mathrm{F},\Omega} h_\Omega c_{\underline{\mathbf{K}},\Omega}^{-\frac{1}{2}} \|q_{\mathrm{t}}^n(s_h^n) - q_{\mathrm{t}}^n(s_{h\tau})(t)\|_{L^2(D)}, \tag{3.8b}$$

where h_Ω is the diameter of Ω and $c_{\underline{\mathbf{K}},\Omega}$ the smallest eigenvalue of the tensor $\underline{\mathbf{K}}$ on Ω. As above, $C_{\mathrm{F},\Omega}$ is the constant from the Poincaré–Friedrichs inequality

$$\|\varphi\|_{L^2(\Omega)} \le C_{\mathrm{F},\Omega} h_\Omega \|\nabla\varphi\|_{L^2(\Omega)} \qquad \forall \varphi \in H_0^1(\Omega), \tag{3.9}$$

and we can take $C_{\mathrm{F},\Omega} = 1$.

3.4 The a posteriori error estimate

We are now ready to state the main result of this paper.

Theorem 1 (A posteriori error estimate for problem (2.12)–(2.15))**.** *Let (s, P) be the weak solution introduced in Definition 2.1 and $(s_{h\tau}, P_{h\tau}) \in \mathcal{E}_\tau$ be an arbitrary approximate solution. Under Assumptions A, B, and C, there exists a generic constant $C > 0$, depending neither on the approximate solution nor on the space–time discretization of Q_T, such that*

$$\begin{aligned} & C \left(\|s_{h\tau} - s\|_{L^2(0,T;H^{-1}(\Omega))}^2 + \|P_{h\tau} - P\|_{L^2(0,T;H_0^1(\Omega))}^2 + \|\varphi(s_{h\tau}) - \varphi(s)\|_{L^2(Q_T)}^2 \right) \\ & \quad \le \quad \|s_{h\tau}(\cdot,0) - s^0\|_{H^{-1}(\Omega)}^2 \\ & \qquad + \sum_{n=1}^N \sum_{\alpha\in\{\mathrm{n,t}\}} \int_{I_n} \left(\left\{ \sum_{D\in\mathcal{D}_h^n} (\eta_{\mathrm{DF},\alpha,D}^n(t) + \eta_{\mathrm{R},\alpha,D}^n)^2 \right\}^{\frac{1}{2}} + \left\{ \sum_{D\in\mathcal{D}_h^n} (\eta_{\mathrm{Q},\alpha,D}^n(t))^2 \right\}^{\frac{1}{2}} \right)^2 \mathrm{d}t. \end{aligned}$$

Moreover, if φ^{-1} belongs to $\mathcal{C}^{0,r}(\mathbb{R})$, the term $\|\varphi(s_{h\tau}) - \varphi(s)\|_{L^2(Q_T)}^2$ can be replaced by $\|s_{h\tau} - s\|_{L^{1+r}(Q_T)}^{1+r}$.

The assumption that φ has a Hölder continuous inverse holds for most of the retention curves used in the subsurface (see, e.g., [7]). For example, considering again the van Genuchten framework with the parameters m and $n = 1/(1-m)$ provides a φ having a Hölder continuous inverse with exponent $2m/(3m+2)$. As follows from above, this provides better a priori estimates for the saturation (see also [45]), and the situation remains unchanged for a posteriori estimates.

Theorem 1 is an immediate consequence of the estimates in Theorems 2 and 3 below. Its application to two examples of finite volume schemes is illustrated in the next section. Appendix A deals with the additional errors that are due to the numerical quadrature, the iterative linearization, and the iterative algebraic solver, which are taken into account explicitly. Furthermore, the spatial and temporal errors are identified and adaptive stopping criteria are proposed in Appendix A.

4 An application to two types of vertex-centered finite volume discretizations

In this section we consider two relatively distinct vertex-centered finite volume discretizations of problem (2.12)–(2.15), and show how Theorem 1 can be applied in both situations.

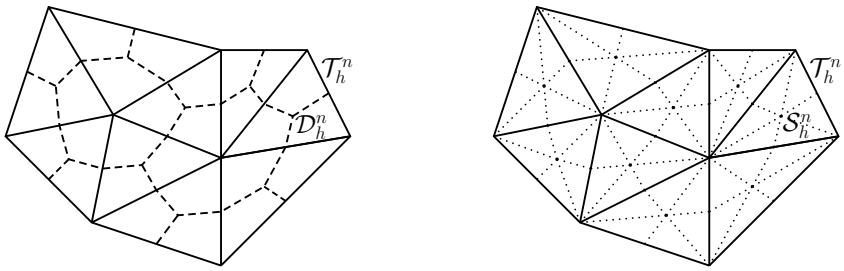

Figure 1: Simplicial mesh $\mathcal{T}_h^n$ and the dual mesh $\mathcal{D}_h^n$ (left); simplicial submesh $\mathcal{S}_h^n$ (right)

4.1 The spatial meshes and the discrete functional spaces

Let $0 \le n \le N$ be fixed. We denote by $\mathcal{T}_h^n$ the partition of Ω (the mesh) involved in the numerical calculation of the approximate solution at time t^n; $\mathcal{T}_h^0$ is the initial mesh. All partitions $\mathcal{T}_h^n$ $(0 \le n \le N)$ consist of d-dimensional simplices and are matching. This means that the intersection of two elements K and L is either empty, or a common vertex, or an ℓ-dimensional face with $1 \le \ell \le d-1$.

For any $0 \le n \le N$, we define the space

$$V_h^n := \{v_h : \Omega \to \mathbb{R},\ v_h \text{ is continuous and piecewise affine on } \mathcal{T}_h^n\}.$$

We will also need

$$\begin{aligned} V_{h\tau} :=& \{v_{h\tau} : \Omega \times [0,T] \to \mathbb{R},\ v_{h\tau} \text{ is affine in time on each } I_n \text{ for all } 1 \le n \le N, \\ & v_h^n := v_{h\tau}(\cdot, t^n) \in V_h^n \text{ for all } 0 \le n \le N\}. \end{aligned}$$

In reinforcement of Assumption B1, we need (cf. Remark 3.1) that $\overline{s} \in V_{h\tau}$ and $\overline{P} \in V_{h\tau}$. We then define

$$\begin{aligned} V_{h;\overline{s}}^n &:= \{v_h \in V_h^n,\ v_h = \overline{s}(\cdot, t^n) \text{ on } \partial\Omega\}, & V_{h;\overline{P}}^n &:= \{v_h \in V_h^n,\ v_h = \overline{P}(\cdot, t^n) \text{ on } \partial\Omega\}, \\ V_{h\tau;\overline{s}} &:= \{v_{h\tau} \in V_{h\tau},\ v_{h\tau} = \overline{s} \text{ on } \partial\Omega \times (0,T]\}, & V_{h\tau;\overline{P}} &:= \{v_{h\tau} \in V_{h\tau},\ v_{h\tau} = \overline{P} \text{ on } \partial\Omega \times (0,T]\}. \end{aligned}$$

For each $\mathcal{T}_h^n$, we next consider a dual mesh $\mathcal{D}_h^n$. Every element (dual volume) $D \in \mathcal{D}_h^n$ is associated with one vertex of $\mathcal{T}_h^n$, and constructed around this vertex by joining the face and element barycenters as indicated in the left picture of Figure 1 for $d = 2$. The set $\mathcal{D}_h^{\mathrm{int},n}$ contains the dual volumes associated with the interior vertices of $\mathcal{T}_h^n$; similarly, $\mathcal{D}_h^{\mathrm{ext},n}$ consists of the dual volumes associated with the boundary vertices of $\mathcal{T}_h^n$. We emphasize that the meshes $\mathcal{T}_h^n$ (and consequently $\mathcal{D}_h^n$) may change in time, typically by refining or coarsening of some elements of the previous mesh. The discrete times and meshes are typically constructed by a space–time adaptive time-marching algorithm, following, e.g., Section A.3 below.

In addition to the meshes $\mathcal{T}_h^n$ and $\mathcal{D}_h^n$, we will also need below a third mesh for each $1 \le n \le N$. This mesh is called $\mathcal{S}_h^n$, consists of d-dimensional simplices, and is matching. It is constructed by joining the barycenters of the elements of $\mathcal{T}_h^n$ with the vertices and the barycenters of the corresponding ℓ-dimensional faces $(1 \le \ell \le d-1)$, see the right picture in Figure 1 for $d = 2$. Note that $\mathcal{S}_h^n$ are submeshes of both $\mathcal{T}_h^n$ and $\mathcal{D}_h^n$; given a volume $D \in \mathcal{D}_h^n$, we denote by $\mathcal{S}_D$ the restriction of $\mathcal{S}_h^n$ onto D. On $\mathcal{S}_h^n$, we define the lowest order Raviart–Thomas–Nédélec space of vector functions, cf. [9, 47],

$$\mathbf{RTN}_0(\mathcal{S}_h^n) := \left\{\mathbf{v}_h \in \mathbf{H}(\mathrm{div}, \Omega)\,;\, \mathbf{v}_h|_K \in \mathbf{RTN}_0(K) \quad \text{for all} \quad K \in \mathcal{S}_h^n\right\}. \tag{4.1}$$

Here, $\mathbf{RTN}_0(K) := [\mathbb{P}_0(K)]^d + \mathbf{x}\mathbb{P}_0(K)$, where K is a given simplex and $\mathbb{P}_0(K)$ is the space of constants on K. In particular, $\mathbf{v}_h \in \mathbf{RTN}_0(\mathcal{S}_h^n)$ is such that $\nabla\cdot\mathbf{v}_h \in \mathbb{P}_0(K)$ for all elements K of $\mathcal{S}_h^n$, $\mathbf{v}_h\cdot\mathbf{n}_F \in \mathbb{P}_0(F)$ for all $(d-1)$-dimensional faces F of $\mathcal{S}_h^n$, and such that its normal trace is continuous.

4.2 A "mathematical" scheme

We first discuss a scheme stemming from the Definition 2.1 of the weak solution. We call it here "mathematical" since it makes use of the Kirchhoff transform. This provides the unknown $\varphi(s)$ that has more regularity, but no particular physical meaning.

4.2.1 The scheme

Let $s_h^0 \in V_{h;\overline{s}}^0$ denote the discretization of the initial condition s^0. Then the "mathematical" vertex-centered finite volume discretization of problem (2.12)–(2.15) reads:

Definition 4.1 ("Mathematical" finite volume scheme)**.** *Find a pair $(s_{h\tau}, P_{h\tau}) \in V_{h\tau;\overline{s}} \times V_{h\tau;\overline{P}}$ such that for all $1 \le n \le N$ and all $D \in \mathcal{D}_h^{\mathrm{int},n}$, $(s_h^n, P_h^n) \in V_{h;\overline{s}}^n \times V_{h;\overline{P}}^n$ are solutions of*

$$\int_D \left(\frac{s_h^n - s_h^{n-1}}{\tau^n}\right) \mathrm{d}\mathbf{x} - \int_{\partial D} \underline{\mathbf{K}}(\eta(s_h^n)\nabla P_h^n + \nabla\varphi(s_h^n))\cdot\mathbf{n}_D \,\mathrm{d}\sigma = \int_D q_{\mathrm{n}}^n(s_h^n)\,\mathrm{d}\mathbf{x}, \tag{4.2a}$$

$$-\int_{\partial D} \underline{\mathbf{K}} M(s_h^n)\nabla P_h^n\cdot\mathbf{n}_D \,\mathrm{d}\sigma = \int_D q_{\mathrm{t}}^n(s_h^n)\,\mathrm{d}\mathbf{x}. \tag{4.2b}$$

Formally, to construct $P_{h\tau}$ on the first time interval I_1 one needs an approximation P_h^0 at the initial time. Since the initial saturation s_h^0 is known, one possibility is to solve (4.2b) for $n = 0$. However, the particular construction of P_h^0 has no influence on the final approximation.

Remark 4.2 (A scheme based on the Kirchhoff transform)**.** *As follows from the analysis of the continuous problem presented in Section 2, the Kirchhoff transform $\varphi(s)$ has better regularity than the nonwetting saturation s. This motivates the following adjustment of the scheme (4.2a)–(4.2b). Let $\overline{\Theta} := \varphi(\overline{s})$ and $\Theta^0 := \varphi(s^0)$. Let $V_{h;\overline{\Theta}}^n$ and $V_{h\tau;\overline{\Theta}}$ be as $V_{h;\overline{s}}^n$ and $V_{h\tau;\overline{s}}$ with $\overline{s}$ replaced by $\varphi(\overline{s})$. Let finally $\Theta_h^0 \in V_{h;\overline{\Theta}}^0$ denote the discretization of the initial condition Θ^0. Then a Kirchhoff transform vertex-centered finite volume discretization of problem (2.12)–(2.15) is a pair $(\Theta_{h\tau}, P_{h\tau}) \in V_{h\tau;\overline{\Theta}} \times V_{h\tau;\overline{P}}$, such that for all $1 \le n \le N$ and all $D \in \mathcal{D}_h^{\mathrm{int},n}$, $(\Theta_h^n, P_h^n) \in V_{h;\overline{\Theta}}^n \times V_{h;\overline{P}}^n$ are solutions of*

$$\int_D \left(\frac{\varphi^{-1}(\Theta_h^n) - \varphi^{-1}(\Theta_h^{n-1})}{\tau^n}\right) \mathrm{d}\mathbf{x} - \int_{\partial D} \underline{\mathbf{K}}(\eta(\varphi^{-1}(\Theta_h^n))\nabla P_h^n + \nabla\Theta_h^n)\cdot\mathbf{n}_D \,\mathrm{d}\sigma = \int_D q_{\mathrm{n}}^n(\varphi^{-1}(\Theta_h^n))\,\mathrm{d}\mathbf{x}, \tag{4.3a}$$

$$-\int_{\partial D} \underline{\mathbf{K}} M(\varphi^{-1}(\Theta_h^n))\nabla P_h^n\cdot\mathbf{n}_D \,\mathrm{d}\sigma = \int_D q_{\mathrm{t}}^n(\varphi^{-1}(\Theta_h^n))\,\mathrm{d}\mathbf{x}. \tag{4.3b}$$

We then define the approximate saturations $s_h^n := \varphi^{-1}(\Theta_h^n)$, $0 \le n \le N$.

The above approach applies whenever φ is strictly increasing and thus the function $\varphi^{-1}(\cdot)$ is well defined (this being satisfied for most of the parameterizations commonly used for modeling porous media flows). If φ is not invertible, a regularization step can be employed, considering, e.g., a small number $\varepsilon > 0$ and approximating φ by φ_ε satisfying for all $s \in \mathbb{R}$

$$\varepsilon \le \varphi'(s) \le L_\varphi, \qquad |\varphi(s) - \varphi_\varepsilon(s)| \le C\varepsilon, \tag{4.4}$$

for some constant $C > 0$. This approach is often used in analyzing degenerate problems and leads to effective numerical algorithms (see, e.g., [37, 45]).

Note that the two schemes, (4.2a)–(4.2b) and (4.3a)–(4.3b), only differ by a numerical quadrature, see Remark A.1 below. Therefore from now on we only focus on the scheme (4.2a)–(4.2b).

4.2.2 The reconstruction of the fluxes

Here we show how to obtain, from the scheme (4.2a)–(4.2b), the flux reconstructions $\mathbf{u}_{\mathrm{n},h\tau}$, $\mathbf{u}_{\mathrm{t},h\tau}$ satisfying Assumption C. To do so we let $1 \le n \le N$ and $D \in \mathcal{D}_h^n$ be given and construct $\mathbf{u}_{\mathrm{n},h}^n, \mathbf{u}_{\mathrm{t},h}^n \in \mathbf{RTN}_0(\mathcal{S}_h^n)$ as follows. For each face F of the mesh $\mathcal{S}_D$ included in ∂D but not in $\partial\Omega$, we take

$$\mathbf{u}_{\mathrm{n},h}^n\cdot\mathbf{n}_F := -\frac{1}{|F|}\int_F \underline{\mathbf{K}}\big(\eta(s_h^n)\nabla P_h^n + \nabla\varphi(s_h^n)\big)\cdot\mathbf{n}_F \,\mathrm{d}\sigma, \tag{4.5a}$$

$$\mathbf{u}_{\mathrm{t},h}^n\cdot\mathbf{n}_F := -\frac{1}{|F|}\int_F \big(\underline{\mathbf{K}} M(s_h^n)\nabla P_h^n\cdot\mathbf{n}_F \,\mathrm{d}\sigma. \tag{4.5b}$$

Observe that in (4.5a)–(4.5b), the degrees of freedom of $\mathbf{u}_{\mathrm{n},h}^n$ and $\mathbf{u}_{\mathrm{t},h}^n$ are not prescribed on all faces of $\mathcal{S}_h^n$. So, equations (4.5a)–(4.5b) do not specify $\mathbf{u}_{\mathrm{n},h}^n$ and $\mathbf{u}_{\mathrm{t},h}^n$ completely. The remaining degrees of freedom can be specified in various ways, as discussed in [50, 21, 23] solution of local (Dirichlet–)Neumann problems by the mixed finite element method, direct prescription. By the Green theorem, from (4.5a)–(4.5b) we immediately get:

Lemma 4.3 (Assumption C for the scheme (4.2a)–(4.2b)). *Let* $\mathbf{u}_{\mathrm{n},h}^n$ *and* $\mathbf{u}_{\mathrm{t},h}^n$ *satisfy* (4.5a)–(4.5b). *Then Assumption C holds true.*

Lemma 4.3 guarantees the validity of the a posteriori error estimate of Theorem 1 for $(s_{h\tau}, P_{h\tau})$ provided by the scheme (4.2a)–(4.2b). For identifying the error components and for the stopping criteria, we refer to Appendix A below.

4.3 A phase-by-phase upstream weighting "engineering" scheme

We now turn to a scheme that is often used in the industrial setting, see, e.g., [29]. Compared to (4.2a)–(4.2b), it solves the mass balance for both phases explicitly, and involves a stabilizing upwinding term.

4.3.1 The scheme

Let $s_h^0 \in V_{h;\overline{s}}^0$ denote the discretization of the initial condition s^0, as in Section 4.2.1. Then the "engineering", phase-by-phase upstream weighting, vertex-centered finite volume discretization of problem (2.12)–(2.15) reads:

Definition 4.4 ("Engineering" finite volume scheme). *Find a pair* $(s_{h\tau}, p_{\mathrm{w},h\tau}) \in V_{h\tau;\overline{s}} \times V_{h\tau}$ *such that* $P(s_{h\tau}, p_{\mathrm{w},h\tau})|_{\partial\Omega} = \overline{P}$ *with* $P(\cdot,\cdot)$ *the function of* (2.10), *and for all* $1 \le n \le N$ *and all* $D \in \mathcal{D}_h^{\mathrm{int},n}$, $(s_h^n, p_{\mathrm{w},h}^n)$ *are solutions of*

$$-\int_D \left(\frac{s_h^n - s_h^{n-1}}{\tau^n}\right) \mathrm{d}\mathbf{x} - \int_{\partial D} [\underline{\mathbf{K}}(\eta_{\mathrm{w}}(1-s_h^n)\nabla(p_{\mathrm{w},h}^n))]^{\mathrm{upw}} \cdot \mathbf{n}_D \,\mathrm{d}\sigma = \int_D q_{\mathrm{w}}^n(1-s_h^n)\,\mathrm{d}\mathbf{x}, \tag{4.6a}$$

$$\int_D \left(\frac{s_h^n - s_h^{n-1}}{\tau^n}\right) \mathrm{d}\mathbf{x} - \int_{\partial D} [\underline{\mathbf{K}}\eta(s_h^n)\nabla(p_{\mathrm{w},h}^n + \pi(s_h^n))]^{\mathrm{upw}} \cdot \mathbf{n}_D \,\mathrm{d}\sigma = \int_D q_{\mathrm{n}}^n(s_h^n)\,\mathrm{d}\mathbf{x}. \tag{4.6b}$$

Here, the superscript upw *denotes the fact that the concerned quantity is evaluated using the values at the vertices in the upstream direction.*

4.3.2 The reconstruction of the fluxes

Although the scheme (4.6a)–(4.6b) is quite different from the scheme (4.2a)–(4.2b), the flux reconstructions $\mathbf{u}_{\mathrm{n},h\tau}$, $\mathbf{u}_{\mathrm{t},h\tau}$ satisfying Assumption C are obtained here in the same easy way as in Section 4.2.2.

Letting $1 \le n \le N$ and $D \in \mathcal{D}_h^n$ given, we construct $\mathbf{u}_{\mathrm{w},h}^n, \mathbf{u}_{\mathrm{n},h}^n \in \mathbf{RTN}_0(\mathcal{S}_h^n)$ as follows. For each face F of the mesh $\mathcal{S}_D$ included in ∂D but not in $\partial\Omega$, we take

$$\mathbf{u}_{\mathrm{w},h}^n \cdot \mathbf{n}_F := -\frac{1}{|F|}\int_F ([\underline{\mathbf{K}}(\eta_{\mathrm{w}}(1-s_h^n)\nabla(p_{\mathrm{w},h}^n))]^{\mathrm{upw}} \cdot \mathbf{n}_F \,\mathrm{d}\sigma, \tag{4.7a}$$

$$\mathbf{u}_{\mathrm{n},h}^n \cdot \mathbf{n}_F := -\frac{1}{|F|}\int_F ([\underline{\mathbf{K}}\eta(s_h^n)\nabla(p_{\mathrm{w},h}^n + \pi(s_h^n))]^{\mathrm{upw}} \cdot \mathbf{n}_F \,\mathrm{d}\sigma. \tag{4.7b}$$

Then we define $\mathbf{u}_{\mathrm{t},h}^n := \mathbf{u}_{\mathrm{w},h}^n + \mathbf{u}_{\mathrm{n},h}^n$. Once again, the Green theorem readily implies:

Lemma 4.5 (Assumption C for the scheme (4.6a)–(4.6b)). *Let* $\mathbf{u}_{\mathrm{w},h}^n$ *and* $\mathbf{u}_{\mathrm{n},h}^n$ *satisfy* (4.7a)–(4.7b) *and set* $\mathbf{u}_{\mathrm{t},h}^n := \mathbf{u}_{\mathrm{w},h}^n + \mathbf{u}_{\mathrm{n},h}^n$. *Then Assumption C holds true.*

As before, Lemma 4.5 ensures that the error estimate in Theorem 1 holds true for $(s_{h\tau}, P(s_{h\tau}, p_{\mathrm{w},h\tau}))$ provided by the scheme (4.6a)–(4.6b).

5 Proof of the a posteriori error estimate

In this section, we introduce the residuals of the weak formulation (2.17a)–(2.17b), show that the error between the exact solution $(s, P) \in \mathcal{E}$ given by Definition 2.1 and an arbitrary approximate solution $(s_{h\tau}, P_{h\tau}) \in \mathcal{E}$ can be bounded by the dual norm of the residuals, and finally show how to bound from above this dual norm by computable a posteriori error estimates when $(s_{h\tau}, P_{h\tau}) \in \mathcal{E}_\tau$. This altogether gives the proof of Theorem 1.

5.1 Definition of the residuals

Recall the set $\mathcal{E}$ from (2.16). We start by the following definition:

Definition 5.1 (Residuals). *Let $(s_{h\tau}, P_{h\tau}) \in \mathcal{E}$ by an arbitrary pair. Define the following continuous linear forms $\mathcal{R}_{\rm n}(s_{h\tau}, P_{h\tau})$, $\mathcal{R}_{\rm t}(s_{h\tau}, P_{h\tau})$ on $L^2((0,T); H_0^1(\Omega))$: for all $\psi, \xi \in L^2((0,T); H_0^1(\Omega))$,*

$$\langle \mathcal{R}_{\rm n}(s_{h\tau}, P_{h\tau}), \psi\rangle := \int_0^T \langle \partial_t s_{h\tau}(\cdot,\theta); \psi(\cdot,\theta)\rangle_{H^{-1},H_0^1}\,\mathrm{d}\theta + \iint_{Q_T} \underline{\mathbf{K}}\,(\eta(s_{h\tau})\nabla P_{h\tau} + \nabla\varphi(s_{h\tau}))\cdot\nabla\psi\,\mathrm{d}\mathbf{x}\mathrm{d}\theta - \iint_{Q_T} q_{\rm n}(s_{h\tau})\psi\,\mathrm{d}\mathbf{x}\mathrm{d}\theta, \tag{5.1a}$$

$$\langle \mathcal{R}_{\rm t}(s_{h\tau}, P_{h\tau}), \xi\rangle := \iint_{Q_T} \underline{\mathbf{K}} M(s_{h\tau})\nabla P_{h\tau}\cdot\nabla\xi\,\mathrm{d}\mathbf{x}\mathrm{d}\theta - \iint_{Q_T} q_{\rm t}(s_{h\tau})\xi\,\mathrm{d}\mathbf{x}\mathrm{d}\theta. \tag{5.1b}$$

Clearly, for any pair $(s_{h\tau}, P_{h\tau}) \in \mathcal{E}$ with $s_{h\tau}(\cdot,0) = s^0$, one has

$$(s_{h\tau}, P_{h\tau}) \text{ is a weak solution} \quad \Leftrightarrow \quad \mathcal{R}_{\rm n}(s_{h\tau}, P_{h\tau}) = \mathcal{R}_{\rm t}(s_{h\tau}, P_{h\tau}) = 0. \tag{5.2}$$

In obtaining the estimates, we let $\|\cdot\|_{H_0^1(\Omega)}$ stand for the energy norm on $H_0^1(\Omega)$,

$$\|v\|_{H_0^1(\Omega)} := \left\{\int_\Omega \left|\underline{\mathbf{K}}^{\frac12}(\mathbf{x})\nabla v(\mathbf{x})\right|^2 \mathrm{d}\mathbf{x}\right\}^{\frac12}, \tag{5.3}$$

and $\|\cdot\|_{L^2(0,T;H_0^1(\Omega))}$ for the energy norm on $L^2(0,T;H_0^1(\Omega))$ given by

$$\|v\|_{L^2(0,T;H_0^1(\Omega))} := \left\{\iint_{Q_T} \left|\underline{\mathbf{K}}^{\frac12}(\mathbf{x})\nabla v(\mathbf{x},\theta)\right|^2 \mathrm{d}\mathbf{x}\mathrm{d}\theta\right\}^{\frac12}.$$

These norms are equivalent to the usual $H^1(\Omega)$-norms due to the boundary conditions and the properties of $\underline{\mathbf{K}}$. The corresponding norm in $H^{-1}(\Omega)$ is defined as

$$\|\zeta\|_{H^{-1}(\Omega)} = \sup_{\psi\in H_0^1(\Omega),\, \|\psi\|_{H_0^1(\Omega)}=1} \langle \zeta, \psi\rangle_{H^{-1},H_0^1}.$$

Further, for the functionals introduced in Definition 5.1, in a standard way we define

$$|||\mathcal{R}_{\rm n}(s_{h\tau}, P_{h\tau})||| := \sup_{\substack{\psi\in L^2(0,T;H_0^1(\Omega))\\ \|\psi\|_{L^2(0,T;H_0^1(\Omega))}=1}} \langle \mathcal{R}_{\rm n}(s_{h\tau}, P_{h\tau}), \psi\rangle,$$

$$|||\mathcal{R}_{\rm t}(s_{h\tau}, P_{h\tau})||| := \sup_{\substack{\xi\in L^2(0,T;H_0^1(\Omega))\\ \|\xi\|_{L^2(0,T;H_0^1(\Omega))}=1}} \langle \mathcal{R}_{\rm t}(s_{h\tau}, P_{h\tau}), \xi\rangle.$$

Finally, for proving the results below, the following elementary inequality will be used often: for all $a, b \in \mathbb{R}$ and all $\delta > 0$,

$$ab \le \frac{a^2}{2\delta} + \delta\frac{b^2}{2}. \tag{5.4}$$

5.2 Bounding the error by the dual norm of the residuals

In this part we show that the error between the exact and approximate solutions can be bounded by the dual norms of the residuals. The results are obtained under Assumption A, employing a duality technique.

Let $(s, P) \in \mathcal{E}$ be the weak solution introduced in Definition 2.1, and satisfying in particular $\nabla P \in [L^\infty(Q_T)]^d$ (cf. Assumption A8). Consider an arbitrary pair $(s_{h\tau}, P_{h\tau}) \in \mathcal{E}$. For any given $t \in (0,T]$, we denote by $G_{h\tau}(\cdot,t) \in H_0^1(\Omega)$ the function satisfying

$$\int_\Omega \underline{\mathbf{K}}\nabla G_{h\tau}(\cdot,t)\cdot\nabla\psi\,\mathrm{d}\mathbf{x} = \int_\Omega (s_{h\tau} - s)(\cdot,t)\psi\,\mathrm{d}\mathbf{x} \tag{5.5}$$

for all $\psi \in H_0^1(\Omega)$. For any $t \in (0,T]$, the existence and uniqueness of $G_{h\tau}(\cdot,t)$ is guaranteed by standard arguments. Moreover, since $s_{h\tau}$ and s are in $C([0,T];L^2(\Omega))$, we obtain $G_{h\tau} \in L^2(0,T;H_0^1(\Omega))$.

We have the following

Lemma 5.2. *Under Assumption A, there exists a constant $C_1 > 0$ such that, for all $t \in (0,T]$, one has*

$$\begin{aligned} |||\mathcal{R}_{\rm n}(s_{h\tau}, P_{h\tau})|||^2 \;\geq\; & \|(s_{h\tau}-s)(\cdot,t)\|^2_{H^{-1}(\Omega)} - \|s_{h\tau}(\cdot,0)-s^0\|^2_{H^{-1}(\Omega)} \\ & + \iint_{Q_t} (s_{h\tau}-s)(\varphi(s_{h\tau})-\varphi(s))\,{\rm d}\mathbf{x}{\rm d}\theta - C_1\|s-s_{h\tau}\|^2_{L^2(0,t;H^{-1}(\Omega))} \\ & + 2\iint_{Q_t} \eta(s_{h\tau})\underline{\mathbf{K}}(\nabla P_{h\tau}-\nabla P)\cdot\nabla G_{h\tau}\,{\rm d}\mathbf{x}{\rm d}\theta. \end{aligned} \tag{5.6}$$

Proof. The H^1_0 norm (5.3) (and consequently the H^{-1} norm) are involving the symmetric, positive definite tensor $\underline{\mathbf{K}}$. Proceeding as for the standard norms in H^{-1}, respectively H^1_0, for all $t \in [0,T]$, the definition (5.5) gives

$$\begin{aligned} \|G_{h\tau}(\cdot,t)\|_{H^1_0(\Omega)} &= \sup_{\psi\in H^1_0(\Omega),\|\psi\|_{H^1_0(\Omega)}=1} \int_\Omega \underline{\mathbf{K}}\nabla G_{h\tau}(\cdot,t)\cdot\nabla\psi\,{\rm d}\mathbf{x} \\ &= \sup_{\psi\in H^1_0(\Omega),\|\psi\|_{H^1_0(\Omega)}=1} \int_\Omega (s_{h\tau}-s)(\cdot,t)\psi\,{\rm d}\mathbf{x} \\ &= \|(s_{h\tau}-s)(\cdot,t)\|_{H^{-1}(\Omega)}. \end{aligned} \tag{5.7}$$

Note that, thanks to (5.2), for all $\psi \in L^2(0,T;H^1_0(\Omega))$, one has

$$\langle \mathcal{R}_{\rm n}(s_{h\tau},P_{h\tau}),\psi\rangle = \langle \mathcal{R}_{\rm n}(s_{h\tau},P_{h\tau}),\psi\rangle - \langle \mathcal{R}_{\rm n}(s,P),\psi\rangle.$$

In particular, choosing $\psi = G_{h\tau}\mathbf{1}_{(0,t)}$ as the test function in this relation provides

$$\langle \mathcal{R}_{\rm n}(s_{h\tau},P_{h\tau}),G_{h\tau}\rangle = A_1+A_2+A_3+A_4+A_5, \tag{5.8}$$

where

$$\begin{aligned} A_1 &:= \int_0^t \langle \partial_t(s_{h\tau}-s)(\cdot,\theta);G_{h\tau}(\cdot,\theta)\rangle_{H^{-1},H^1_0}\,{\rm d}\theta, \\ A_2 &:= \iint_{Q_t} \eta(s_{h\tau})\underline{\mathbf{K}}(\nabla P_{h\tau}-\nabla P)\cdot\nabla G_{h\tau}\,{\rm d}\mathbf{x}{\rm d}\theta, \\ A_3 &:= \iint_{Q_t} (\eta(s_{h\tau})-\eta(s))\,\underline{\mathbf{K}}\nabla P\cdot\nabla G_{h\tau}\,{\rm d}\mathbf{x}{\rm d}\theta, \\ A_4 &:= \iint_{Q_t} \underline{\mathbf{K}}\,(\nabla\varphi(s_{h\tau})-\nabla\varphi(s))\cdot\nabla G_{h\tau}\,{\rm d}\mathbf{x}{\rm d}\theta, \\ A_5 &:= -\iint_{Q_t} (q_{\rm n}(s_{h\tau})-q_{\rm n}(s))\,G_{h\tau}\,{\rm d}\mathbf{x}{\rm d}\theta. \end{aligned}$$

Recalling (5.5), $\partial_t G_{h\tau}$ solves

$$\left\{\begin{array}{rcll} -\nabla\cdot(\underline{\mathbf{K}}\nabla(\partial_t G_{h\tau})) &=& \partial_t(s_{h\tau}-s) & \text{in } \Omega, \\ \partial_t G_{h\tau} &=& 0 & \text{on } \partial\Omega, \end{array}\right.$$

for a.e. $t \in (0,T]$, in a weak sense. Since $\partial_t(s_{h\tau}-s) \in L^2(0,T;H^{-1}(\Omega))$, we have $\partial_t G_{h\tau} \in L^2(0,T;H^1_0(\Omega))$, ensuring that $G_{h\tau} \in \mathcal{C}([0,T];H^1_0(\Omega))$. Thus, it follows from the definition (5.5) of $G_{h\tau}$ that

$$A_1 = \iint_{Q_t} \underline{\mathbf{K}}(\partial_t\nabla G_{h\tau})\cdot\nabla G_{h\tau}\,{\rm d}\mathbf{x}{\rm d}\theta = \frac{1}{2}\left(\|G_{h\tau}(\cdot,t)\|^2_{H^1_0(\Omega)} - \|G_{h\tau}(\cdot,0)\|^2_{H^1_0(\Omega)}\right).$$

Hence, using (5.7), we obtain that

$$A_1 = \frac{1}{2}\left(\|(s_{h\tau}-s)(\cdot,t)\|^2_{H^{-1}(\Omega)} - \|s_{h\tau}(\cdot,0)-s^0\|^2_{H^{-1}(\Omega)}\right). \tag{5.9}$$

Further, with C denoting a positive constant, not necessarily the same at each occurrence, since $\underline{\mathbf{K}}^{\frac{1}{2}}\nabla P \in [L^\infty(Q_t)]^d$ we get

$$\begin{aligned} A_3 &\geq -C\|\eta(s_{h\tau})-\eta(s)\|_{L^2(Q_t)}\|G_{h\tau}\|_{L^2(0,t;H_0^1(\Omega))} \\ &\geq -\frac{1}{2C_0}\|\eta(s_{h\tau})-\eta(s)\|^2_{L^2(Q_t)} - \frac{C^2C_0}{2}\|G_{h\tau}\|^2_{L^2(0,t;H_0^1(\Omega))}, \end{aligned}$$

where C_0 is the constant appearing in relation (2.19).

Third, one has

$$A_5 \geq -\frac{1}{2C_0}\|q_{\rm n}(s_{h\tau})-q_{\rm n}(s)\|^2_{L^2(Q_t)} - \frac{C_0}{2}\|G_{h\tau}\|^2_{L^2(Q_t)}.$$

Thanks to the Poincaré–Friedrichs inequality (3.9), there exists a $C>0$ such that, for almost all $\theta\in(0,t]$,

$$\|G_{h\tau}(\cdot,\theta)\|_{L^2(\Omega)} \leq C\|G_{h\tau}(\cdot,\theta)\|_{H_0^1(\Omega)} = C\|(s_{h\tau}-s)(\cdot,\theta)\|_{H^{-1}(\Omega)}.$$

Therefore, there exists a $C>0$ such that

$$A_5 \geq -\frac{1}{2C_0}\|q_{\rm n}(s_{h\tau})-q_{\rm n}(s)\|^2_{L^2(Q_t)} - C\|s_{h\tau}-s\|^2_{L^2(0,t;H^{-1}(\Omega))}.$$

By (5.7) and Assumption A7,

$$A_3+A_5 \geq -\frac{1}{2}\iint_{Q_t}(s_{h\tau}-s)(\varphi(s_{h\tau})-\varphi(s))\,{\rm d}\mathbf{x}{\rm d}\theta - C\|s_{h\tau}-s\|^2_{L^2(0,t;H^{-1}(\Omega))}. \tag{5.10}$$

Fourth, recalling (5.5), since $\varphi(s)-\varphi(s_{h\tau})\in L^2(0,T;H_0^1(\Omega)))$, we obtain

$$A_4 = \iint_{Q_t}(s_{h\tau}-s)(\varphi(s_{h\tau})-\varphi(s))\,{\rm d}\mathbf{x}{\rm d}\theta. \tag{5.11}$$

Finally, using (5.7) gives

$$\begin{aligned} \langle \mathcal{R}_{\rm n}(s_{h\tau},P_{h\tau}),G_{h\tau}\rangle &\leq |||\mathcal{R}_{\rm n}(s_{h\tau},P_{h\tau})|||\,\|s_{h\tau}-s\|_{L^2(0,t;H^{-1}(\Omega))} \\ &\leq \frac{1}{2}|||\mathcal{R}_{\rm n}(s_{h\tau},P_{h\tau})|||^2 + \frac{1}{2}\|s_{h\tau}-s\|^2_{L^2(0,t;H^{-1}(\Omega))}. \end{aligned} \tag{5.12}$$

Employing (5.9)–(5.12) into (5.8) provides (5.6). □

Lemma 5.3. *Under Assumption A, there exist the constants $C_2,C_3,C_4>0$ such that, for all $t\in(0,T]$, one has*

$$\begin{aligned} |||\mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau})|||^2 &\geq C_2\|P_{h\tau}-P\|^2_{L^2(0,t;H_0^1(\Omega))} - \frac{C_3}{2}\iint_{Q_t}(s_{h\tau}-s)(\varphi(s_{h\tau})-\varphi(s))\,{\rm d}\mathbf{x}{\rm d}\theta \\ &\quad -2C_3\iint_{Q_t}\eta(s_{h\tau})\underline{\mathbf{K}}(\nabla P_{h\tau}-\nabla P)\cdot\nabla G_{h\tau}\,{\rm d}\mathbf{x}{\rm d}\theta - C_4\|s_{h\tau}-s\|^2_{L^2(0,t;H^{-1}(\Omega))}. \end{aligned} \tag{5.13}$$

Proof. For any $t\in(0,T]$, we denote by $\widehat{G}_{h\tau}(\cdot,t)$ the function in $H_0^1(\Omega)$ satisfying

$$\int_\Omega M(s_{h\tau}(\cdot,t))\underline{\mathbf{K}}\nabla\widehat{G}_{h\tau}(\cdot,t)\cdot\nabla\psi\,{\rm d}\mathbf{x} = -\int_\Omega \eta(s_{h\tau}(\cdot,t))\underline{\mathbf{K}}\nabla G_{h\tau}(\cdot,t)\cdot\nabla\psi\,{\rm d}\mathbf{x} \tag{5.14}$$

for all $\psi\in H_0^1(\Omega)$, where $G_{h\tau}(\cdot,t)\in H_0^1(\Omega)$ solves (5.5). The existence and uniqueness of $\widehat{G}_{h\tau}(\cdot,t)$ is again guaranteed by standard arguments.

Choosing $\widehat{G}_{h\tau}(\cdot,t)$ as test function in (5.14) and using (2.18) and (2.19) gives

$$\|\widehat{G}_{h\tau}(\cdot,t)\|_{H_0^1(\Omega)} \leq \frac{C_\eta}{c_M}\|G_{h\tau}(\cdot,t)\|_{H_0^1(\Omega)} = \frac{C_\eta}{c_M}\|(s_{h\tau}-s)(\cdot,t)\|_{H^{-1}(\Omega)}. \tag{5.15}$$

With $\lambda>0$ an arbitrary parameter that will be fixed later, choosing $\xi_{h\tau} := \left(P_{h\tau}-P+\lambda\widehat{G}_{h\tau}\right)\mathbf{1}_{(0,t)}$ as test function in (5.1b) yields

$$\langle\mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau}),\xi_{h\tau}\rangle = \langle\mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau}),\xi_{h\tau}\rangle - \langle\mathcal{R}_{\rm t}(s,P),\xi_{h\tau}\rangle = B_1+B_2+B_3+B_4+B_5+B_6, \tag{5.16}$$

where

$$\begin{aligned}
B_1 &:= \iint_{Q_t} M(s_{h\tau})\underline{\mathbf{K}}\,(\nabla P_{h\tau}-\nabla P)\cdot(\nabla P_{h\tau}-\nabla P)\,\mathrm{d}\mathbf{x}\mathrm{d}\theta,\\
B_2 &:= \iint_{Q_t} (M(s_{h\tau})-M(s))\underline{\mathbf{K}}\nabla P\cdot(\nabla P_{h\tau}-\nabla P)\,\mathrm{d}\mathbf{x}\mathrm{d}\theta,\\
B_3 &:= \lambda\iint_{Q_t} M(s_{h\tau})\underline{\mathbf{K}}\,(\nabla P_{h\tau}-\nabla P)\cdot\nabla\widehat{G}_{h\tau}\,\mathrm{d}\mathbf{x}\mathrm{d}\theta,\\
B_4 &:= \lambda\iint_{Q_t} (M(s_{h\tau})-M(s))\underline{\mathbf{K}}\nabla P\cdot\nabla\widehat{G}_{h\tau}\,\mathrm{d}\mathbf{x}\mathrm{d}\theta,\\
B_5 &:= -\iint_{Q_t} (q_{\mathrm{t}}(s_{h\tau})-q_{\mathrm{t}}(s))\,(P_{h\tau}-P)\,\mathrm{d}\mathbf{x}\mathrm{d}\theta,\\
B_6 &:= -\lambda\iint_{Q_t} (q_{\mathrm{t}}(s_{h\tau})-q_{\mathrm{t}}(s))\,\widehat{G}_{h\tau}\,\mathrm{d}\mathbf{x}\mathrm{d}\theta.
\end{aligned}$$

First, thanks to (2.18), one has

$$B_1 \geq c_M\|P_{h\tau}-P\|^2_{L^2(0,t;H^1_0(\Omega))}. \tag{5.17}$$

Second, since $\underline{\mathbf{K}}^{\frac{1}{2}}\nabla P\in[L^\infty(Q_t)]^d$, we get, for some $C>0$,

$$\begin{aligned}
B_2 &\geq -C\|M(s_{h\tau})-M(s)\|_{L^2(Q_t)}\|P_{h\tau}-P\|_{L^2(0,t;H^1_0(\Omega))}\\
&\geq -C\|M(s_{h\tau})-M(s)\|^2_{L^2(Q_t)}-\frac{c_M}{4}\|P_{h\tau}-P\|^2_{L^2(0,t;H^1_0(\Omega))}.
\end{aligned}$$

By Assumption A7, there exists a $C>0$, not depending on $(s_{h\tau},P_{h\tau})$, such that

$$B_2\geq -C\iint_{Q_t}(s_{h\tau}-s)(\varphi(s_{h\tau})-\varphi(s))\,\mathrm{d}\mathbf{x}\mathrm{d}\theta-\frac{c_M}{4}\|P_{h\tau}-P\|^2_{L^2(0,t;H^1_0(\Omega))}. \tag{5.18}$$

Third, it follows from the definition (5.14) of $\widehat{G}_{h\tau}$ that

$$B_3=-\lambda\iint_{Q_t}\eta(s_{h\tau})\underline{\mathbf{K}}(\nabla P_{h\tau}-\nabla P)\cdot\nabla G_{h\tau}\,\mathrm{d}\mathbf{x}\mathrm{d}\theta. \tag{5.19}$$

Fourth, since $\underline{\mathbf{K}}^{\frac{1}{2}}\nabla P\in[L^\infty(Q_t)]^d$, by Assumption A7 and (5.15) we get

$$\begin{aligned}
B_4 &\geq -C\lambda\|M(s_{h\tau})-M(s)\|_{L^2(Q_t)}\|\widehat{G}_{h\tau}\|_{L^2(0,t;H^1_0(\Omega))}\\
&\geq -C\|M(s_{h\tau})-M(s)\|^2_{L^2(Q_t)}-\lambda^2\|\widehat{G}_{h\tau}\|^2_{L^2(0,t;H^1_0(\Omega))}\\
&\geq -C\iint_{Q_t}(s_{h\tau}-s)(\varphi(s_{h\tau})-\varphi(s))\,\mathrm{d}\mathbf{x}\mathrm{d}\theta-C\lambda^2\|s_{h\tau}-s\|_{L^2(0,t;H^{-1}(\Omega))}.
\end{aligned} \tag{5.20}$$

Fifth, by (5.4), for all $\mu>0$ one has

$$B_5\geq-\frac{1}{4\mu}\|q_{\mathrm{t}}(s_{h\tau})-q_{\mathrm{t}}(s)\|^2_{L^2(Q_t)}-\mu\|P-P_{h\tau}\|^2_{L^2(Q_t)}.$$

Therefore, using the Poincaré–Friedrichs inequality (3.9), a convenient choice of μ, and Assumption A7 lead to

$$B_5\geq -C\iint_{Q_t}(s_{h\tau}-s)(\varphi(s_{h\tau})-\varphi(s))\,\mathrm{d}\mathbf{x}\mathrm{d}\theta-\frac{c_M}{4}\|P-P_{h\tau}\|^2_{L^2(0,t;H^1_0(\Omega))}. \tag{5.21}$$

Sixth, using Assumption A7,

$$B_6\geq-\lambda^2\|\widehat{G}_{h\tau}\|^2_{L^2(Q_t)}-C\iint_{Q_t}(s_{h\tau}-s)(\varphi(s_{h\tau})-\varphi(s))\,\mathrm{d}\mathbf{x}\mathrm{d}\theta.$$

The Poincaré–Friedrichs inequality (3.9) and (5.15) give

$$B_6\geq-\lambda^2C\|s_{h\tau}-s\|^2_{L^2(0,t;H^{-1}(\Omega))}-C\iint_{Q_t}(s_{h\tau}-s)(\varphi(s_{h\tau})-\varphi(s))\,\mathrm{d}\mathbf{x}\mathrm{d}\theta. \tag{5.22}$$

From (5.16)–(5.22), one gets

$$\begin{aligned}\langle \mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau}),\xi_{h\tau}\rangle \;\geq\; & \frac{c_M}{2}\|P_{h\tau}-P\|^2_{L^2(0,t;H^1_0(\Omega))} - C\iint_{Q_t}(s_{h\tau}-s)(\varphi(s_{h\tau})-\varphi(s))\,\mathrm{d}\mathbf{x}\mathrm{d}\theta \\ & -\lambda\iint_{Q_t}\eta(s_{h\tau})\underline{\mathbf{K}}(\nabla P_{h\tau}-\nabla P)\cdot\nabla G_{h\tau}\,\mathrm{d}\mathbf{x}\mathrm{d}\theta - \lambda^2 C\|s_{h\tau}-s\|^2_{L^2(0,t;H^{-1}(\Omega))}. \end{aligned} \tag{5.23}$$

On the other hand, we deduce from (5.15) that

$$\|\xi_{h\tau}\|_{L^2(0,t;H^1_0(\Omega))} \leq \|P_{h\tau}-P\|_{L^2(0,t;H^1_0(\Omega))} + \lambda\frac{C_\eta}{c_M}\|s_{h\tau}-s\|_{L^2(0,t;H^{-1}(\Omega))},$$

leading to

$$\langle \mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau}),\xi_{h\tau}\rangle \leq |||\mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau})|||\left(\|P_{h\tau}-P\|_{L^2(0,t;H^1_0(\Omega))} + \lambda C\|s_{h\tau}-s\|_{L^2(0,t;H^{-1}(\Omega))}\right).$$

With (5.4), we can prove that

$$\frac{2}{c_M}|||\mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau})|||^2 \geq \langle \mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau}),\xi_{h\tau}\rangle - \frac{c_M}{4}\|P_{h\tau}-P\|^2_{L^2(0,t;H^1_0(\Omega))} - \lambda^2 C\|s_{h\tau}-s\|_{L^2(0,t;H^{-1}(\Omega))}.$$

Using the relation (5.23), this provides

$$\begin{aligned}|||\mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau})|||^2 \;\geq\; & \frac{c_M^2}{8}\|P_{h\tau}-P\|^2_{L^2(0,t;H^1_0(\Omega))} - \frac{C_3}{2}\iint_{Q_t}(s_{h\tau}-s)(\varphi(s_{h\tau})-\varphi(s))\,\mathrm{d}\mathbf{x}\mathrm{d}\theta \\ & -\frac{\lambda c_M}{2}\iint_{Q_t}\eta(s_{h\tau})\underline{\mathbf{K}}(\nabla P_{h\tau}-\nabla P)\cdot\nabla G_{h\tau}\,\mathrm{d}\mathbf{x}\mathrm{d}\theta - C\lambda^2\|s_{h\tau}-s\|_{L^2(0,t;H^{-1}(\Omega))}. \end{aligned}$$

Choosing $\lambda = 4\frac{C_3}{c_M}$ leads to (5.13). □

Note that the fifth term on the right in (5.6) and the third term on the right in (5.13) differ by a constant. Therefore, a straightforward consequence of Lemmas 5.2 and 5.3 is

Lemma 5.4. *Under Assumption A, there exist the constants $C_5, C_6, C_7 > 0$ such that, for all $t \in (0,T)$, one has*

$$\begin{aligned} & |||\mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau})|||^2 + C_3|||\mathcal{R}_{\rm n}(s_{h\tau},P_{h\tau})|||^2 \\ & \geq\; C_3\left(\|(s_{h\tau}-s)(\cdot,t)\|^2_{H^{-1}(\Omega)} - \|s_{h\tau}(\cdot,0)-s^0\|^2_{H^{-1}(\Omega)}\right) \\ & \quad + C_5\iint_{Q_t}(s_{h\tau}-s)(\varphi(s_{h\tau})-\varphi(s))\,\mathrm{d}\mathbf{x}\mathrm{d}\theta \\ & \quad + C_6\|P_{h\tau}-P\|_{L^2(0,t;H^1_0(\Omega))} - C_7\|s_{h\tau}-s\|_{L^2(0,t;H^{-1}(\Omega))}. \end{aligned} \tag{5.24}$$

The last term on the right-hand side of (5.24) appears with a negative sign. We bound it as follows:

Lemma 5.5. *Under Assumption A, there exists a constant $C_8 > 0$ such that*

$$\|s_{h\tau}-s\|^2_{L^\infty(0,T;H^{-1}(\Omega))} \leq C_8\left(\|s_{h\tau}(\cdot,0)-s^0\|^2_{H^{-1}(\Omega)} + |||\mathcal{R}_{\rm n}(s_{h\tau},P_{h\tau})|||^2 + |||\mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau})|||^2\right).$$

Proof. Since φ is increasing on $\mathbb{R}$, for all $t \in (0,T]$,

$$\iint_{Q_t}(s_{h\tau}-s)(\varphi(s_{h\tau})-\varphi(s))\,\mathrm{d}\mathbf{x}\mathrm{d}\theta \geq 0.$$

By Lemma 5.4, for all $t \in (0,T]$ one has

$$\begin{aligned}\|(s_{h\tau}-s)(\cdot,t)\|^2_{H^{-1}(\Omega)} \;\leq\; & \|s_{h\tau}(\cdot,0)-s^0\|^2_{H^{-1}(\Omega)} + |||\mathcal{R}_{\rm n}(s_{h\tau},P_{h\tau})|||^2 \\ & + \frac{1}{C_3}|||\mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau})|||^2 + \frac{C_7}{C_3}\int_0^t\|(s_{h\tau}-s)(\cdot,\theta)\|^2_{H^{-1}(\Omega)}\,\mathrm{d}\theta. \end{aligned}$$

The Gronwall lemma yields for all $t \in [0,T]$

$$\|(s_{h\tau}-s)(\cdot,t)\|^2_{H^{-1}(\Omega)} \le e^{\frac{C_7 T}{C_3}} \left(\|s_{h\tau}(\cdot,0)-s^0\|^2_{H^{-1}(\Omega)} + |||\mathcal{R}_{\rm n}(s_{h\tau},P_{h\tau})|||^2 + \frac{1}{C_3}|||\mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau})|||^2 \right).$$

The conclusion follows with $C_8 = e^{\frac{C_7 T}{C_3}} \max\{1, 1/C_3\}$. □

We now give the following lemma, which is a straightforward consequence of Lemma 5.5.

Lemma 5.6. *Under Assumption A, there exists a constant $C_9 > 0$ such that*

$$\|s_{h\tau}-s\|_{L^2((0,T);H^{-1}(\Omega))} \le C_9 \left(\|s_{h\tau}(\cdot,0)-s^0\|^2_{H^{-1}(\Omega)} + |||\mathcal{R}_{\rm n}(s_{h\tau},P_{h\tau})|||^2 + |||\mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau})|||^2 \right).$$

Having proved all the results above we can now state the main result of this section:

Theorem 2 (Upper bound on the error by the residuals)**.** *Let $(s,P) \in \mathcal{E}$ be the weak solution introduced in Definition 2.1 and let $(s_{h\tau},P_{h\tau}) \in \mathcal{E}$ be arbitrary. Under Assumption A, there exists a constant $C>0$ such that*

$$\begin{aligned} &\|s_{h\tau}-s\|^2_{L^2(0,T;H^{-1}(\Omega))} + \|P_{h\tau}-P\|^2_{L^2(0,T;H^1_0(\Omega))} + \|\varphi(s_{h\tau})-\varphi(s)\|^2_{L^2(Q_T)} \\ &\qquad \le C\left(\|s_{h\tau}(\cdot,0)-s^0\|^2_{H^{-1}(\Omega)} + |||\mathcal{R}_{\rm n}(s_{h\tau},P_{h\tau})|||^2 + |||\mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau})|||^2 \right). \end{aligned} \tag{5.25}$$

Moreover, if φ^{-1} belongs to $\mathcal{C}^{0,r}(\mathbb{R})$, then there exists $C>0$ such that

$$\begin{aligned} &\|s_{h\tau}-s\|^2_{L^2(0,T;H^{-1}(\Omega))} + \|P_{h\tau}-P\|^2_{L^2(0,T;H^1_0(\Omega))} + \|s_{h\tau}-s\|^{1+r}_{L^{1+r}(Q_T)} \\ &\qquad \le C\left(\|s_{h\tau}(\cdot,0)-s^0\|^2_{H^{-1}(\Omega)} + |||\mathcal{R}_{\rm n}(s_{h\tau},P_{h\tau})|||^2 + |||\mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau})|||^2 \right). \end{aligned} \tag{5.26}$$

Proof. Since φ is increasing and L_φ-Lipschitz continuous, one has, for all $(a,b) \in \mathbb{R}$

$$(a-b)(\varphi(a)-\varphi(b)) \ge \frac{1}{L_\varphi}(\varphi(a)-\varphi(b))^2.$$

As a consequence,

$$\iint_{Q_T} (s_{h\tau}-s)\,(\varphi(s_{h\tau})-\varphi(s))\,\mathrm{d}\mathbf{x}\mathrm{d}\theta \ge \frac{1}{L_\varphi}\|\varphi(s_{h\tau})-\varphi(s)\|^2_{L^2(Q_T)}. \tag{5.27}$$

On the other hand, if φ^{-1} is r-Hölder continuous, for all $(a,b) \in \mathbb{R}$ one has

$$(a-b)(\varphi(a)-\varphi(b)) \ge C(a-b)^{1+r}.$$

This gives

$$\iint_{Q_T} (s_{h\tau}-s)\,(\varphi(s_{h\tau})-\varphi(s))\,\mathrm{d}\mathbf{x}\mathrm{d}\theta \ge C\|s_{h\tau}-s\|^{1+r}_{L^{1+r}(Q_T)}. \tag{5.28}$$

Choosing $t=T$ in (5.24), one by (5.27) obtains

$$\begin{aligned} &C\left(\|\varphi(s_{h\tau})-\varphi(s)\|^2_{L^2(Q_T)} + \|P-P_{h\tau}\|^2_{L^2(0,T;H^1_0(\Omega))} \right) \\ &\quad \le \|s_{h\tau}(\cdot,0)-s^0\|^2_{H^{-1}(\Omega)} + |||\mathcal{R}_{\rm n}(s_{h\tau},P_{h\tau})|||^2 + |||\mathcal{R}_{\rm t}(s_{h\tau},P_{h\tau})|||^2 + C\|s_{h\tau}-s\|^2_{L^2(0,T;H^{-1}(\Omega))}. \end{aligned}$$

The first result now follows from Lemma 5.6. The second one can be shown in the same way, by using (5.28) instead of (5.27). □

Remark 5.7 (Uniqueness and continuous dependence on the initial data)**.** *Let (s,P) and $(\tilde{s},\tilde{P})$ be two weak solutions following Definition 2.1 for the initial data s^0, respectively $\tilde{s}^0$. Thanks to (5.2), (5.25) gives*

$$\|s-\tilde{s}\|_{L^2(0,T;H^{-1}(\Omega))} + \|P-\tilde{P}\|_{L^2(0,T;H^1_0(\Omega))} + \|\varphi(s)-\varphi(\tilde{s})\|_{L^2(Q_T)} \le C\|s^0-\tilde{s}^0\|_{H^{-1}(\Omega)}.$$

This provides the uniqueness of the weak solution for a given initial data, as well as the continuous dependence with respect to the initial data for the above topology.

Remark 5.8 (Hölder continuity of φ^{-1})**.** *The estimate in (5.26) is obtained assuming additionally that φ^{-1} is Hölder continuous. This is fulfilled by parameterizations that are commonly encountered in the literature, like, e.g., the Brooks–Corey or van Genuchten–Mualem models (see [7]).*

5.3 Bounding the dual norm of the residuals by the a posteriori estimate

We now finally bound the dual norm of the residuals by a fully computable a posteriori error estimate. Recall the definitions (2.16) and (3.1). Herein, we need to assume $(s_{h\tau}, P_{h\tau}) \in \mathcal{E}_\tau$ instead of merely $(s_{h\tau}, P_{h\tau}) \in \mathcal{E}$.

Theorem 3 (Upper bound on the residuals)**.** *Let Assumptions A1–A3 and B hold. Let $(s_{h\tau}, P_{h\tau}) \in \mathcal{E}_\tau$ be arbitrary. Let the estimators be defined by (3.4), (3.5), and (3.8). Under Assumption C, there holds*

$$\begin{aligned}
&|||\mathcal{R}_\mathrm{n}(s_{h\tau}, P_{h\tau})|||^2 + |||\mathcal{R}_\mathrm{t}(s_{h\tau}, P_{h\tau})|||^2 \\
&\le \sum_{n=1}^N \sum_{\alpha\in\{\mathrm{n},\mathrm{t}\}} \int_{I_n} \left(\left\{ \sum_{D\in\mathcal{D}_h^n} (\eta_{\mathrm{DF},\alpha,D}^n(t) + \eta_{\mathrm{R},\alpha,D}^n)^2 \right\}^{\frac12} + \left\{ \sum_{D\in\mathcal{D}_h^n} (\eta_{\mathrm{Q},\alpha,D}^n(t))^2 \right\}^{\frac12} \right)^2 \mathrm{d}t.
\end{aligned}$$

Proof. Let $\psi, \xi \in L^2(0,T;H_0^1(\Omega))$ with $\|\psi\|_{L^2(0,T;H_0^1(\Omega))} = \|\xi\|_{L^2(0,T;H_0^1(\Omega))} = 1$ be given. Using the definition (5.1a), adding and subtracting $\mathbf{u}_{\mathrm{n},h\tau}\cdot\nabla\psi$, and employing the Green theorem and (3.2a) leads to

$$\begin{aligned}
\langle \mathcal{R}_\mathrm{n}(s_{h\tau}, P_{h\tau}), \psi\rangle &= \iint_{Q_T} (\partial_t s_{h\tau} + \nabla\cdot\mathbf{u}_{\mathrm{n},h\tau} - q_\mathrm{n}(s_{h\tau}))\psi \,\mathrm{d}\mathbf{x}\mathrm{d}t \\
&\quad + \iint_{Q_T} (\underline{\mathbf{K}}\,(\eta(s_{h\tau})\nabla P_{h\tau} + \nabla\varphi(s_{h\tau})) + \mathbf{u}_{\mathrm{n},h\tau})\cdot\nabla\psi \,\mathrm{d}\mathbf{x}\mathrm{d}t \\
&= \sum_{n=1}^N \int_{I_n} \left\{ \sum_{D\in\mathcal{D}_h^{\mathrm{int},n}} \int_D (\partial_t s_{h\tau} + \nabla\cdot\mathbf{u}_{\mathrm{n},h}^n - q_\mathrm{n}^n(s_h^n))(\psi - \psi_D)\,\mathrm{d}\mathbf{x} \right. \\
&\quad \left. + \sum_{D\in\mathcal{D}_h^{\mathrm{ext},n}} \int_D (\partial_t s_{h\tau} + \nabla\cdot\mathbf{u}_{\mathrm{n},h}^n - q_\mathrm{n}^n(s_h^n))\psi \,\mathrm{d}\mathbf{x} \right\} \mathrm{d}t \\
&\quad + \sum_{n=1}^N \int_{I_n} \sum_{D\in\mathcal{D}_h^n} \int_D (q_\mathrm{n}^n(s_h^n) - q_\mathrm{n}^n(s_{h\tau}))\psi \,\mathrm{d}\mathbf{x}\mathrm{d}t \\
&\quad + \sum_{n=1}^N \int_{I_n} \sum_{D\in\mathcal{D}_h^n} \int_D (\underline{\mathbf{K}}\,(\eta(s_{h\tau})\nabla P_{h\tau} + \nabla\varphi(s_{h\tau})) + \mathbf{u}_{\mathrm{n},h}^n)\cdot\nabla\psi \,\mathrm{d}\mathbf{x}\mathrm{d}t.
\end{aligned}$$

Here ψ_D stands for the mean value of the function ψ over the volume D. Let $1 \le n \le N$ and $t \in I_n$ be fixed. For any $D \in \mathcal{D}_h^{\mathrm{int},n}$, the Cauchy–Schwarz inequality, the Poincaré–Wirtinger inequality (3.6), the properties of $\underline{\mathbf{K}}$, and the definition (3.5a) give

$$\int_D (\partial_t s_{h\tau} + \nabla\cdot\mathbf{u}_{\mathrm{n},h}^n - q_\mathrm{n}^n(s_h^n))(\psi - \psi_D)\,\mathrm{d}\mathbf{x} \le \eta_{\mathrm{R},\mathrm{n},D}^n \|\underline{\mathbf{K}}^{\frac12}\nabla\psi\|_{L^2(D)}(t).$$

Similarly, the Poincaré–Friedrichs inequality (3.7) gives, for any $D \in \mathcal{D}_h^{\mathrm{ext},n}$,

$$\int_D (\partial_t s_{h\tau} + \nabla\cdot\mathbf{u}_{\mathrm{n},h}^n - q_\mathrm{n}^n(s_h^n))\psi\,\mathrm{d}\mathbf{x} \le \eta_{\mathrm{R},\mathrm{n},D}^n \|\underline{\mathbf{K}}^{\frac12}\nabla\psi\|_{L^2(D)}(t).$$

In the same manner, for any $D \in \mathcal{D}_h^n$, recalling (3.4a), one can use the Cauchy–Schwarz inequality to obtain

$$\int_D (\underline{\mathbf{K}}\,(\eta(s_{h\tau})\nabla P_{h\tau} + \nabla\varphi(s_{h\tau})) + \mathbf{u}_{\mathrm{n},h}^n)\cdot\nabla\psi\,\mathrm{d}\mathbf{x} \le \eta_{\mathrm{DF},\mathrm{n},D}^n(t)\|\underline{\mathbf{K}}^{\frac12}\nabla\psi\|_{L^2(D)}(t),$$

whereas definition (3.8a), the Cauchy–Schwarz inequality, and the Poincaré–Friedrichs inequality (3.9) give

$$\sum_{D\in\mathcal{D}_h^n} \int_D (q_\mathrm{n}^n(s_h^n) - q_\mathrm{n}^n(s_{h\tau}))\psi\,\mathrm{d}\mathbf{x} \le \left\{ \sum_{D\in\mathcal{D}_h^n} (\eta_{\mathrm{Q},\mathrm{n},D}^n(t))^2 \right\}^{\frac12} \|\underline{\mathbf{K}}^{\frac12}\nabla\psi\|_{L^2(\Omega)}(t).$$

This leads to

$$|||\mathcal{R}_\mathrm{n}(s_{h\tau}, P_{h\tau})|||^2 \le \sum_{n=1}^N \int_{I_n} \left(\left\{ \sum_{D\in\mathcal{D}_h^n} (\eta_{\mathrm{DF},\mathrm{n},D}^n(t) + \eta_{\mathrm{R},\mathrm{n},D}^n)^2 \right\}^{\frac12} + \left\{ \sum_{D\in\mathcal{D}_h^n} (\eta_{\mathrm{Q},\mathrm{n},D}^n(t))^2 \right\}^{\frac12} \right)^2 \mathrm{d}t.$$

Similarly, from (5.1b) and (3.2b), we obtain

$$\begin{aligned}\langle \mathcal{R}_{\mathrm{t}}(s_{h\tau}, P_{h\tau}), \xi\rangle &= \iint_{Q_T} (\underline{\mathbf{K}} M(s_{h\tau})\nabla P_{h\tau} + \mathbf{u}_{\mathrm{t},h\tau})\cdot\nabla\xi \,\mathrm{d}\mathbf{x}\mathrm{d}t + \iint_{Q_T} (\nabla\cdot\mathbf{u}_{\mathrm{t},h\tau} - q_{\mathrm{t}}(s_{h\tau}))\xi \,\mathrm{d}\mathbf{x}\mathrm{d}t \\ &= \sum_{n=1}^{N}\int_{I_n}\Bigg\{\sum_{D\in\mathcal{D}_h^{\mathrm{int},n}}\int_D (\nabla\cdot\mathbf{u}_{\mathrm{t},h}^n - q_{\mathrm{t}}^n(s_h^n))(\xi-\xi_D)\,\mathrm{d}\mathbf{x} \\ &\quad + \sum_{D\in\mathcal{D}_h^{\mathrm{ext},n}}\int_D (\nabla\cdot\mathbf{u}_{\mathrm{t},h}^n - q_{\mathrm{t}}^n(s_h^n))\xi\,\mathrm{d}\mathbf{x}\Bigg\}\mathrm{d}t \\ &\quad + \sum_{n=1}^{N}\int_{I_n}\sum_{D\in\mathcal{D}_h^n}\int_D (q_{\mathrm{t}}^n(s_h^n) - q_{\mathrm{t}}^n(s_{h\tau}))\xi\,\mathrm{d}\mathbf{x}\mathrm{d}t \\ &\quad + \sum_{n=1}^{N}\int_{I_n}\sum_{D\in\mathcal{D}_h^n}(\underline{\mathbf{K}} M(s_{h\tau})\nabla P_{h\tau} + \mathbf{u}_{\mathrm{t},h}^n)\cdot\nabla\xi\,\mathrm{d}\mathbf{x}\mathrm{d}t.\end{aligned}$$

Using the same arguments as above to bound $|||\mathcal{R}_{\mathrm{t}}(s_{h\tau}, P_{h\tau})|||^2$, the assertion of the theorem follows. □

A Distinguishing the error components and stopping criteria

Here we consider the scheme of Section 4.2.1 and show how the estimators of Theorem 1 can be further developed to distinguish between the different error components, derive stopping criteria for iterative linearizations and algebraic solvers, and show how to equilibrate the principal error components. We follow the approach introduced in [21, 23, 24, 31] and extended to the context of two-phase flows in [51].

A.1 Numerical quadrature, linearization, and algebraic solver

We start by describing the steps taken in the practical implementation of (4.2a)–(4.2b). For the sake of clarity, we only consider some simple but illustrative examples, but mention that other choices are also possible.

A.1.1 Numerical quadrature

In the practical calculations, one does not solve (4.2a)–(4.2b) exactly. This is because of the particular nature of the (nonlinear) functions $\eta(s_h^n)$, $M(s_h^n)$, $\varphi(s_h^n)$, $q_{\mathrm{n}}^n(s_h^n)$, and $q_{\mathrm{t}}^n(s_h^n)$, making an exact evaluation of the integrals over elements difficult or even impossible. For this reason, one typically uses a numerical quadrature.

In this sense, consider an arbitrary (nonlinear) function $f:\Omega\to\mathbb{R}$. Given $1\le n\le N$ and $v_h\in V_h^n$, $f(v_h)$ does not necessary belong to V_h^n. Therefore we define the operator $f_h: V_h^n\to V_h^n$ by

$$(f_h(v_h))(\mathbf{x}) = (f(v_h))(\mathbf{x}) \tag{A.1}$$

for all $v_h\in V_h^n$ and for all vertices $\mathbf{x}$ of $\mathcal{T}_h^n$. Clearly, $f_h(v_h)$ is a quadrature-based approximation of $f(v_h)$.

A practical implementation of the "mathematical" vertex-centered finite volume method in (4.2a)–(4.2b) is employing the numerical quadrature (A.1). This leads to the problem of finding a pair $(s_{h\tau}, P_{h\tau})\in V_{h\tau;\overline{s}}\times V_{h\tau;\overline{P}}$ such that for all $1\le n\le N$ and all $D\in\mathcal{D}_h^{\mathrm{int},n}$, $(s_h^n, P_h^n)\in V_{h;\overline{s}}^n\times V_{h;\overline{P}}^n$ are solutions of

$$\int_D\left(\frac{s_h^n - s_h^{n-1}}{\tau^n}\right)\mathrm{d}\mathbf{x} - \int_{\partial D}\underline{\mathbf{K}}(\eta_h(s_h^n)\nabla P_h^n + \nabla\varphi_h(s_h^n))\cdot\mathbf{n}_D\,\mathrm{d}\sigma = \int_D (q_{\mathrm{n}}^n)_h(s_h^n)\,\mathrm{d}\mathbf{x}, \tag{A.2a}$$

$$-\int_{\partial D}\underline{\mathbf{K}} M_h(s_h^n)\nabla P_h^n\cdot\mathbf{n}_D\,\mathrm{d}\sigma = \int_D (q_{\mathrm{t}}^n)_h(s_h^n)\,\mathrm{d}\mathbf{x}. \tag{A.2b}$$

Remark A.1 (Link to the Kirchhoff transform scheme of Remark 4.2)**.** *Remark 4.2 gives the scheme (4.3a)–(4.3b), which is similar to (4.2a)–(4.2b), but makes use of the Kirchhoff transform. As above, in the practical implementation of (4.3a)–(4.3b), one can consider the numerical quadrature in (A.1). The resulting scheme*

leads to the same system of nonlinear algebraic equations as (A.2a)–(A.2b). Consequently, the resulting nodal values of s_h^n, P_h^n, and $\Theta_h^n = \varphi(s_h^n)$ are the same and the two schemes only differ in the interpretation of the results: the first scheme is given in terms of $s_{h\tau} \in V_{h\tau;\overline{s}}$, whereas $\Theta_{h\tau} \in V_{h\tau;\overline{\Theta}}$ appears in the Kirchhoff-based approach.

A.1.2 Linearization

At each time step n, (A.2a)–(A.2b) represents a system of nonlinear algebraic equations. Solving it requires an iterative linearization procedure. Here we only give a simple illustrative example of a fixed point approach; fixed point or Newton-type linearizations for the equivalent Kirchhoff transform based scheme (4.3a)–(4.3b) are considered, together with the rigorous convergence proof, in [30, 42, 44].

For a given $1 \le n \le N$, let $s_h^{n,0}$ be a given initial guess for the saturation s_h^n. A typical choice is $s_h^{n,0} = s_h^{n-1}$. We consider the following fixed point linearization of (A.2a)–(A.2b). Starting with $k = 1$, at each step k we determine the pair $(s_h^{n,k}, P_h^{n,k}) \in V_{h;\overline{s}}^n \times V_{h;\overline{P}}^n$ such that, for all $D \in \mathcal{D}_h^{\mathrm{int},n}$,

$$\int_D \left(\frac{s_h^{n,k} - s_h^{n-1}}{\tau^n} \right) \mathrm{d}\mathbf{x} - \int_{\partial D} \underline{\mathbf{K}} \big(\eta_h(s_h^{n,k-1}) \nabla P_h^{n,k} + \nabla \varphi_h(s_h^{n,k-1}) \big) \cdot \mathbf{n}_D \, \mathrm{d}\sigma = \int_D (q_{\mathrm{n}}^n)_h(s_h^{n,k-1}) \, \mathrm{d}\mathbf{x}, \tag{A.3a}$$

$$- \int_{\partial D} \underline{\mathbf{K}} M_h(s_h^{n,k-1}) \nabla P_h^{n,k} \cdot \mathbf{n}_D \, \mathrm{d}\sigma = \int_D (q_{\mathrm{t}}^n)_h(s_h^{n,k-1}) \, \mathrm{d}\mathbf{x}. \tag{A.3b}$$

A.1.3 Algebraic solver

At each time step n and each linearization step k, (A.3a)–(A.3b) represents a system of linear algebraic equations, which is typically solved by an iterative solver (i being the corresponding iteration index). Here we keep the discussion general without specifying any particular solver.

For a given $1 \le n \le N$ and $k \ge 1$, let $s_h^{n,k,0}$ be a given initial guess for the saturation $s_h^{n,k}$. Typically, $s_h^{n,k,0} = s_h^{n,k-1}$. Starting from $i = 1$, on each step i an iterative algebraic solver for (A.3a)–(A.3b) provides the pair $(s_h^{n,k,i}, P_h^{n,k,i}) \in V_{h;\overline{s}}^n \times V_{h;\overline{P}}^n$ such that, for all $D \in \mathcal{D}_h^{\mathrm{int},n}$,

$$\int_D \left(\frac{s_h^{n,k,i} - s_h^{n-1}}{\tau^n} \right) \mathrm{d}\mathbf{x} - \int_{\partial D} \underline{\mathbf{K}} \big(\eta_h(s_h^{n,k-1}) \nabla P_h^{n,k,i} + \nabla \varphi_h(s_h^{n,k-1}) \big) \cdot \mathbf{n}_D \, \mathrm{d}\sigma$$

$$= \int_D (q_{\mathrm{n}}^n)_h(s_h^{n,k-1}) \, \mathrm{d}\mathbf{x} - R_{\mathrm{n},D}^{n,k,i}, \tag{A.4a}$$

$$- \int_{\partial D} \underline{\mathbf{K}} M_h(s_h^{n,k-1}) \nabla P_h^{n,k,i} \cdot \mathbf{n}_D \, \mathrm{d}\sigma = \int_D (q_{\mathrm{t}}^n)_h(s_h^{n,k-1}) \, \mathrm{d}\mathbf{x} - R_{\mathrm{t},D}^{n,k,i}. \tag{A.4b}$$

Here, $R_{\mathrm{n}}^{n,k,i}$ and $R_{\mathrm{t}}^{n,k,i}$ are the algebraic residual vectors at the given step i and $R_{\alpha,D}^{n,k,i}$, $\alpha \in \{\mathrm{n}, \mathrm{t}\}$, are the components of these algebraic vectors corresponding to the dual volume $D \in \mathcal{D}_h^{\mathrm{int},n}$.

Altogether, the approximate solution obtained at the time step n, the linearization step k, and the algebraic solver step i is a pair $(s_{h\tau}^{n,k,i}, P_{h\tau}^{n,k,i}) \in V_{h\tau;\overline{s}}|_{I_n} \times V_{h\tau;\overline{P}}|_{I_n}$ given by

$$s_{h\tau}^{n,k,i}(t^n) = s_h^{n,k,i}, \quad s_{h\tau}^{n,k,i}(t^{n-1}) = s_h^{n-1}, \tag{A.5a}$$

$$P_{h\tau}^{n,k,i}(t^n) = P_h^{n,k,i}, \quad P_{h\tau}^{n,k,i}(t^{n-1}) = P_h^{n-1}. \tag{A.5b}$$

A.2 Distinguishing the error components

Computing the pair $(s_{h\tau}^{n,k,i}, P_{h\tau}^{n,k,i})$, defined by (A.5a)–(A.5b), involves a numerical quadrature, an iterative linearization, and an iterative algebraic solver. Therefore this approximate solution does not solve the initial equations (4.2a)–(4.2b) and henceforth the flux reconstructions (4.5a)–(4.5b) do not necessarily satisfy Assumption C. We show below how these fluxes can be reconstructed in such a way that Assumption C is still satisfied, allowing to apply Theorem 1. We further show how to distinguish and estimate separately the additional errors arising from iterative linearizations and algebraic solvers that have not converged completely.

A.2.1 Reconstruction of the fluxes

Let a time step n, a linearization step k, and an algebraic solver step i be given. We construct here the fluxes $\mathbf{u}_{\mathrm{n},h}^{n,k,i}$ and $\mathbf{u}_{\mathrm{t},h}^{n,k,i}$ satisfying Assumption C. To distinguish the different error components, we consider

$$\mathbf{u}_{\mathrm{n},h}^{n,k,i} = \mathbf{d}_{\mathrm{n},h}^{n,k,i} + \mathbf{l}_{\mathrm{n},h}^{n,k,i} + \mathbf{a}_{\mathrm{n},h}^{n,k,i} + \mathbf{q}_{\mathrm{n},h}^{n,k,i}, \tag{A.6a}$$

$$\mathbf{u}_{\mathrm{t},h}^{n,k,i} = \mathbf{d}_{\mathrm{t},h}^{n,k,i} + \mathbf{l}_{\mathrm{t},h}^{n,k,i} + \mathbf{a}_{\mathrm{t},h}^{n,k,i} + \mathbf{q}_{\mathrm{t},h}^{n,k,i}, \tag{A.6b}$$

where all the fluxes above are constructed in the space $\mathbf{RTN}_0(\mathcal{S}_h^n)$, cf. Section 4.1; $\mathbf{d}_{\mathrm{n},h}^{n,k,i}, \mathbf{d}_{\mathrm{t},h}^{n,k,i}$ are called the *discretization fluxes*, $\mathbf{l}_{\mathrm{n},h}^{n,k,i}, \mathbf{l}_{\mathrm{t},h}^{n,k,i}$ the *linearization error fluxes*, $\mathbf{a}_{\mathrm{n},h}^{n,k,i}, \mathbf{a}_{\mathrm{t},h}^{n,k,i}$ the *algebraic error fluxes*, and $\mathbf{q}_{\mathrm{n},h}^{n,k,i}, \mathbf{q}_{\mathrm{t},h}^{n,k,i}$ the *space quadrature–linearization error fluxes*.

We first specify $\mathbf{d}_{\mathrm{n},h}^{n,k,i}, \mathbf{d}_{\mathrm{t},h}^{n,k,i}$. For each face F of the mesh $\mathcal{S}_D$ included in ∂D but not in $\partial\Omega$, we define

$$\mathbf{d}_{\mathrm{n},h}^{n,k,i}\cdot\mathbf{n}_F := -\frac{1}{|F|}\int_F \big(\underline{\mathbf{K}}\big(\eta_h(s_h^{n,k,i})\nabla P_h^{n,k,i} + \nabla\varphi_h(s_h^{n,k,i})\big)\cdot\mathbf{n}_F\,\mathrm{d}\sigma, \tag{A.7a}$$

$$\mathbf{d}_{\mathrm{t},h}^{n,k,i}\cdot\mathbf{n}_F := -\frac{1}{|F|}\int_F \big(\underline{\mathbf{K}}M_h(s_h^{n,k,i})\nabla P_h^{n,k,i}\cdot\mathbf{n}_F\,\mathrm{d}\sigma. \tag{A.7b}$$

Next, $\mathbf{l}_{\mathrm{n},h}^{n,k,i}$ and $\mathbf{l}_{\mathrm{t},h}^{n,k,i}$ are specified implicitly by

$$(\mathbf{d}_{\mathrm{n},h}^{n,k,i} + \mathbf{l}_{\mathrm{n},h}^{n,k,i})\cdot\mathbf{n}_F := -\frac{1}{|F|}\int_F \big(\underline{\mathbf{K}}\big(\eta_h(s_h^{n,k-1})\nabla P_h^{n,k,i} + \nabla\varphi_h(s_h^{n,k-1})\big)\cdot\mathbf{n}_F\,\mathrm{d}\sigma, \tag{A.8a}$$

$$(\mathbf{d}_{\mathrm{t},h}^{n,k,i} + \mathbf{l}_{\mathrm{t},h}^{n,k,i})\cdot\mathbf{n}_F := -\frac{1}{|F|}\int_F \big(\underline{\mathbf{K}}M_h(s_h^{n,k-1})\nabla P_h^{n,k,i}\cdot\mathbf{n}_F\,\mathrm{d}\sigma. \tag{A.8b}$$

As discussed in Section 4.2.2, for the remaining degrees of freedom one can proceed as in [50, 21, 23]. As for $\mathbf{a}_{\mathrm{n},h}^{n,k,i}$, $\mathbf{a}_{\mathrm{t},h}^{n,k,i}$ and $\mathbf{q}_{\mathrm{n},h}^{n,k,i}$, $\mathbf{q}_{\mathrm{t},h}^{n,k,i}$, for all $D\in\mathcal{D}_h^{\mathrm{int},n}$ we merely require that

$$\int_D \nabla\cdot\mathbf{a}_{\mathrm{n},h}^{n,k,i}\,\mathrm{d}\mathbf{x} = R_{\mathrm{n},D}^{n,k,i}, \tag{A.9a}$$

$$\int_D \nabla\cdot\mathbf{a}_{\mathrm{t},h}^{n,k,i}\,\mathrm{d}\mathbf{x} = R_{\mathrm{t},D}^{n,k,i} \tag{A.9b}$$

and

$$\int_D \nabla\cdot\mathbf{q}_{\mathrm{n},h}^{n,k,i}\,\mathrm{d}\mathbf{x} = \int_D \big((q_{\mathrm{n}}^n)_h(s_h^{n,k-1}) - q_{\mathrm{n}}^n(s_h^{n,k,i})\big)\,\mathrm{d}\mathbf{x}, \tag{A.10a}$$

$$\int_D \nabla\cdot\mathbf{q}_{\mathrm{t},h}^{n,k,i}\,\mathrm{d}\mathbf{x} = \int_D \big((q_{\mathrm{t}}^n)_h(s_h^{n,k-1}) - q_{\mathrm{t}}^n(s_h^{n,k,i})\big)\,\mathrm{d}\mathbf{x}. \tag{A.10b}$$

From $R_{\alpha,D}^{n,k,i}$ and $(q_\alpha^n)_h(s_h^{n,k-1}) - q_\alpha^n(s_h^{n,k,i})$, $\alpha\in\{\mathrm{n},\mathrm{t}\}$, $\mathbf{a}_{\mathrm{n},h}^{n,k,i}$, $\mathbf{a}_{\mathrm{t},h}^{n,k,i}$ and $\mathbf{q}_{\mathrm{n},h}^{n,k,i}$, $\mathbf{q}_{\mathrm{t},h}^{n,k,i}$ can be constructed using, for instance, the algorithm of [31, Section 7.3] or proceeding as in [24]. The goal is to ensure that $\|\mathbf{q}_{\alpha,h}^{n,k,i}\|_{\underline{\mathbf{K}}^{-\frac12};L^2(\Omega)}$, $\|\mathbf{a}_{\alpha,h}^{n,k,i}\|_{\underline{\mathbf{K}}^{-\frac12};L^2(\Omega)}$, and $\|\mathbf{l}_{\alpha,h}^{n,k,i}\|_{\underline{\mathbf{K}}^{-\frac12};L^2(\Omega)}$, $\alpha\in\{\mathrm{n},\mathrm{t}\}$, go to zero as the quadrature error gets negligible and the algebraic and linearization solver, respectively, converge.

Using (A.4a)–(A.4b), the above definitions lead to:

Lemma A.2 (Assumption C for time step n, linearization step k, and algebraic solver step i)**.** *Let* $\mathbf{u}_{\mathrm{n},h}^{n,k,i}$ *and* $\mathbf{u}_{\mathrm{t},h}^{n,k,i}$ *satisfy* (A.6)–(A.10)*. Then Assumption C holds true for* $\mathbf{u}_{\mathrm{n},h}^{n,k,i}$ *and* $\mathbf{u}_{\mathrm{t},h}^{n,k,i}$*.*

A.2.2 Distinguishing the error components

Given a time step n, a time $t \in I_n$, a volume $D \in \mathcal{D}_h^n$, a linearization step k, and an iterative solver step i, we introduce the following notations, refining the ones in (3.4), (3.5), and (3.8):

$$\eta_{\mathrm{DF,n},D}^{n,k,i}(t) := \|\mathbf{u}_{\mathrm{n},h}^{n,k,i} + \underline{\mathbf{K}}\big(\eta(s_{h\tau}^{n,k,i})\nabla P_{h\tau}^{n,k,i} + \nabla\varphi(s_{h\tau}^{n,k,i})\big)(t)\|_{\underline{\mathbf{K}}^{-\frac{1}{2}};L^2(D)}, \tag{A.11a}$$

$$\eta_{\mathrm{DF,t},D}^{n,k,i}(t) := \|\mathbf{u}_{\mathrm{t},h}^{n,k,i} + \underline{\mathbf{K}}M(s_{h\tau}^{n,k,i})\nabla P_{h\tau}^{n,k,i}(t)\|_{\underline{\mathbf{K}}^{-\frac{1}{2}};L^2(D)}, \tag{A.11b}$$

$$\eta_{\mathrm{R,n},D}^{n,k,i} := m_D\|\partial_t s_{h\tau}^{n,k,i} + \nabla\cdot\mathbf{u}_{\mathrm{n},h}^{n,k,i} - q_{\mathrm{n}}^n(s_h^{n,k,i})\|_{L^2(D)}, \tag{A.11c}$$

$$\eta_{\mathrm{R,t},D}^{n,k,i} := m_D\|\nabla\cdot\mathbf{u}_{\mathrm{t},h}^{n,k,i} - q_{\mathrm{t}}^n(s_h^{n,k,i})\|_{L^2(D)}, \tag{A.11d}$$

$$\eta_{\mathrm{Q,n},D}^{n,k,i}(t) := C_{\mathrm{F},\Omega}h_\Omega c_{\underline{\mathbf{K}},\Omega}^{-\frac{1}{2}}\|q_{\mathrm{n}}^n(s_h^{n,k,i}) - q_{\mathrm{n}}^n(s_{h\tau}^{n,k,i})(t)\|_{L^2(D)}, \tag{A.11e}$$

$$\eta_{\mathrm{Q,t},D}^{n,k,i}(t) := C_{\mathrm{F},\Omega}h_\Omega c_{\underline{\mathbf{K}},\Omega}^{-\frac{1}{2}}\|q_{\mathrm{t}}^n(s_h^{n,k,i}) - q_{\mathrm{t}}^n(s_{h\tau}^{n,k,i})(t)\|_{L^2(D)}. \tag{A.11f}$$

Define the *spatial error estimators* by

$$\begin{aligned}\eta_{\mathrm{sp,n},D}^{n,k,i} :=& \eta_{\mathrm{R,n},D}^{n,k,i} + \|\mathbf{d}_{\mathrm{n},h}^{n,k,i} + \underline{\mathbf{K}}\big(\eta(s_{h\tau}^{n,k,i})\nabla P_{h\tau}^{n,k,i} + \nabla\varphi(s_{h\tau}^{n,k,i})\big)(t^n)\|_{\underline{\mathbf{K}}^{-\frac{1}{2}};L^2(D)},\\ \eta_{\mathrm{sp,t},D}^{n,k,i} :=& \eta_{\mathrm{R,t},D}^{n,k,i} + \|\mathbf{d}_{\mathrm{t},h}^{n,k,i} + \underline{\mathbf{K}}M(s_{h\tau}^{n,k,i})\nabla P_{h\tau}^{n,k,i}(t^n)\|_{\underline{\mathbf{K}}^{-\frac{1}{2}};L^2(D)},\end{aligned}$$

the *temporal error estimators* by

$$\begin{aligned}\eta_{\mathrm{tm,n},D}^{n,k,i}(t) :=& \|\underline{\mathbf{K}}\big(\eta(s_{h\tau}^{n,k,i})\nabla P_{h\tau}^{n,k,i} + \nabla\varphi(s_{h\tau}^{n,k,i})\big)(t) - \underline{\mathbf{K}}\big(\eta(s_{h\tau}^{n,k,i})\nabla P_{h\tau}^{n,k,i} + \nabla\varphi(s_{h\tau}^{n,k,i})\big)(t^n)\|_{\underline{\mathbf{K}}^{-\frac{1}{2}};L^2(D)}\\ &+ \eta_{\mathrm{Q,n},D}^{n,k,i}(t),\\ \eta_{\mathrm{tm,t},D}^{n,k,i}(t) :=& \|\underline{\mathbf{K}}M(s_{h\tau}^{n,k,i})\nabla P_{h\tau}^{n,k,i}(t) - \underline{\mathbf{K}}M(s_{h\tau}^{n,k,i})\nabla P_{h\tau}^{n,k,i}(t^n)\|_{\underline{\mathbf{K}}^{-\frac{1}{2}};L^2(D)} + \eta_{\mathrm{Q,t},D}^{n,k,i}(t),\end{aligned}$$

the *linearization error estimators* by

$$\begin{aligned}\eta_{\mathrm{lin,n},D}^{n,k,i} &:= \|\mathbf{l}_{\mathrm{n},h}^{n,k,i}\|_{\underline{\mathbf{K}}^{-\frac{1}{2}};L^2(D)},\\ \eta_{\mathrm{lin,t},D}^{n,k,i} &:= \|\mathbf{l}_{\mathrm{t},h}^{n,k,i}\|_{\underline{\mathbf{K}}^{-\frac{1}{2}};L^2(D)},\end{aligned}$$

the *algebraic error estimators* by

$$\begin{aligned}\eta_{\mathrm{alg,n},D}^{n,k,i} &:= \|\mathbf{a}_{\mathrm{n},h}^{n,k,i}\|_{\underline{\mathbf{K}}^{-\frac{1}{2}};L^2(D)},\\ \eta_{\mathrm{alg,t},D}^{n,k,i} &:= \|\mathbf{a}_{\mathrm{t},h}^{n,k,i}\|_{\underline{\mathbf{K}}^{-\frac{1}{2}};L^2(D)},\end{aligned}$$

and the *space quadrature–linearization error estimators* by

$$\begin{aligned}\eta_{\mathrm{quad,n},D}^{n,k,i} &:= \|\mathbf{q}_{\mathrm{n},h}^{n,k,i}\|_{\underline{\mathbf{K}}^{-\frac{1}{2}};L^2(D)},\\ \eta_{\mathrm{quad,t},D}^{n,k,i} &:= \|\mathbf{q}_{\mathrm{t},h}^{n,k,i}\|_{\underline{\mathbf{K}}^{-\frac{1}{2}};L^2(D)}.\end{aligned}$$

The estimators introduced above have global counterparts, defined as

$$\big(\eta_{\mathrm{sp}}^{n,k,i}\big)^2 := 2\tau^n \sum_{\alpha\in\{\mathrm{n,t}\}}\sum_{D\in\mathcal{D}_h^n}\big(\eta_{\mathrm{sp},\alpha,D}^{n,k,i}\big)^2, \tag{A.12a}$$

$$\big(\eta_{\mathrm{tm}}^{n,k,i}\big)^2 := 4\sum_{\alpha\in\{\mathrm{n,t}\}}\int_{I_n}\sum_{D\in\mathcal{D}_h^n}\big(\eta_{\mathrm{tm},\alpha,D}^{n,k,i}(t)\big)^2\,\mathrm{d}t, \tag{A.12b}$$

$$\big(\eta_{\mathrm{lin}}^{n,k,i}\big)^2 := 2\tau^n \sum_{\alpha\in\{\mathrm{n,t}\}}\sum_{D\in\mathcal{D}_h^n}\big(\eta_{\mathrm{lin},\alpha,D}^{n,k,i}\big)^2, \tag{A.12c}$$

$$\big(\eta_{\mathrm{alg}}^{n,k,i}\big)^2 := 2\tau^n \sum_{\alpha\in\{\mathrm{n,t}\}}\sum_{D\in\mathcal{D}_h^n}\big(\eta_{\mathrm{alg},\alpha,D}^{n,k,i}\big)^2, \tag{A.12d}$$

$$\big(\eta_{\mathrm{quad}}^{n,k,i}\big)^2 := 2\tau^n \sum_{\alpha\in\{\mathrm{n,t}\}}\sum_{D\in\mathcal{D}_h^n}\big(\eta_{\mathrm{quad},\alpha,D}^{n,k,i}\big)^2. \tag{A.12e}$$

With the above notations, we have the following

Corollary A.3 (Distinguishing the space, time, linearization, and algebraic solver errors)**.** *Let the time step n, the linearization step k, and the iterative solver step i be fixed. The estimators of* (A.11) *can be bounded locally as*

$$\eta^{n,k,i}_{\mathrm{DF},\alpha,D}(t)+\eta^{n,k,i}_{\mathrm{R},\alpha,D}+\eta^{n,k,i}_{\mathrm{Q},\alpha,D}(t)\le\eta^{n,k,i}_{\mathrm{sp},\alpha,D}+\eta^{n,k,i}_{\mathrm{tm},\alpha,D}(t)+\eta^{n,k,i}_{\mathrm{lin},\alpha,D}+\eta^{n,k,i}_{\mathrm{alg},\alpha,D}+\eta^{n,k,i}_{\mathrm{quad},\alpha,D},\qquad t\in I_n,\ \alpha\in\{\mathrm{n},\mathrm{t}\},$$

and globally as

$$\begin{aligned}&\left\{\sum_{\alpha\in\{\mathrm{n},\mathrm{t}\}}\int_{I_n}\left(\left\{\sum_{D\in\mathcal{D}_h^n}(\eta^{n,k,i}_{\mathrm{DF},\alpha,D}(t)+\eta^{n,k,i}_{\mathrm{R},\alpha,D})^2\right\}^{\frac12}+\left\{\sum_{D\in\mathcal{D}_h^n}(\eta^{n,k,i}_{\mathrm{Q},\alpha,D}(t))^2\right\}^{\frac12}\right)^2\mathrm{d}t\right\}^{\frac12}\\&\le\eta^{n,k,i}_{\mathrm{sp}}+\eta^{n,k,i}_{\mathrm{tm}}+\eta^{n,k,i}_{\mathrm{lin}}+\eta^{n,k,i}_{\mathrm{alg}}+\eta^{n,k,i}_{\mathrm{quad}}.\end{aligned}$$

A.3 Stopping criteria and adaptivity

Based on the estimate of Corollary A.3, we now give criteria for stopping the iterative linearization and the iterative algebraic solver, for controlling the space quadrature error, and for adaptive space–time mesh refinement.

A.3.1 A stopping criterion for iterative algebraic solvers

Let $0<\gamma_{\mathrm{alg}}$ be a user-given weight, typically close to 1. Following [31, 24], the iterative algebraic solver can be stopped whenever

$$\eta^{n,k,i}_{\mathrm{alg}}\le\gamma_{\mathrm{alg}}(\eta^{n,k,i}_{\mathrm{sp}}+\eta^{n,k,i}_{\mathrm{tm}}+\eta^{n,k,i}_{\mathrm{lin}}+\eta^{n,k,i}_{\mathrm{quad}}).\tag{A.13}$$

Essentially (A.13) indicates when the error due to the iterative algebraic solver starts to be dominated by other terms in the global error. After meeting this criterion, one should focus on reducing the other components in the overall error.

A local version of (A.13) can also be stated. Given $D\in\mathcal{D}_h^n$, let $0<\gamma_{\mathrm{alg},D}$. The iterative algebraic solver should be stopped whenever

$$\eta^{n,k,i}_{\mathrm{alg},\alpha,D}\le\gamma_{\mathrm{alg},D}(\eta^{n,k,i}_{\mathrm{sp},\alpha,D}+\eta^{n,k,i}_{\mathrm{tm},\alpha,D}+\eta^{n,k,i}_{\mathrm{lin},\alpha,D}+\eta^{n,k,i}_{\mathrm{quad},\alpha,D}),\qquad\alpha\in\{\mathrm{n},\mathrm{t}\}.\tag{A.14}$$

A.3.2 A stopping criterion for iterative linearizations

Let $0<\gamma_{\mathrm{lin}}$ be a user-given weight, typically close to 1. Following [21, 24], the iterative linearization solver can be stopped whenever

$$\eta^{n,k,i}_{\mathrm{lin}}\le\gamma_{\mathrm{lin}}(\eta^{n,k,i}_{\mathrm{sp}}+\eta^{n,k,i}_{\mathrm{tm}}+\eta^{n,k,i}_{\mathrm{quad}}).\tag{A.15}$$

Whenever (A.15) holds, the overall error is dominated by the other components than the linearization error, therefore no improvement in the approximation can be expected by when continuing iterating in the linearization step.

A local version of (A.15) can be defined as in the previous paragraph. Given $D\in\mathcal{D}_h^n$, let $0<\gamma_{\mathrm{lin},D}$. Then the iterative linearization solver should be stopped whenever

$$\eta^{n,k,i}_{\mathrm{lin},\alpha,D}\le\gamma_{\mathrm{lin},D}(\eta^{n,k,i}_{\mathrm{sp},\alpha,D}+\eta^{n,k,i}_{\mathrm{tm},\alpha,D}+\eta^{n,k,i}_{\mathrm{quad},\alpha,D}),\qquad\alpha\in\{\mathrm{n},\mathrm{t}\}.\tag{A.16}$$

A.3.3 Controlling the quadrature errors

Whenever the nonlinear source functions q_α are present, the fluxes $\mathbf{q}^{n,k,i}_{\mathrm{n},h}$, $\mathbf{q}^{n,k,i}_{\mathrm{t},h}$ and the corresponding estimators $\eta^{n,k,i}_{\mathrm{quad,n},D}$, $\eta^{n,k,i}_{\mathrm{quad,t},D}$ are nonzero. As above, let $0<\gamma_{\mathrm{quad}}$ be a user-given weight, typically close to 1. We require

$$\eta^{n,k,i}_{\mathrm{quad}}\le\gamma_{\mathrm{quad}}(\eta^{n,k,i}_{\mathrm{sp}}+\eta^{n,k,i}_{\mathrm{tm}}),\tag{A.17}$$

i.e., that the error stemming from the approximation of the nonlinear source functions is controlled by the spatial and temporal ones. Locally, given $D\in\mathcal{D}_h^n$, let $0<\gamma_{\mathrm{quad},D}$. Then the quadrature errors are to be controlled by

$$\eta^{n,k,i}_{\mathrm{quad},\alpha,D}\le\gamma_{\mathrm{quad},D}(\eta^{n,k,i}_{\mathrm{sp},\alpha,D}+\eta^{n,k,i}_{\mathrm{tm},\alpha,D}),\qquad\alpha\in\{\mathrm{n},\mathrm{t}\}.\tag{A.18}$$

A.3.4 Adaptive space–time mesh refinement

Once the conditions (A.13), (A.15), and (A.17) or (A.14), (A.16), and (A.18) are verified, we are left with balancing the spatial and temporal errors. First, we require that

$$\eta_{\mathrm{tm}}^{n,k,i} \approx \eta_{\mathrm{sp}}^{n,k,i}. \tag{A.19}$$

In contrast to (A.13), (A.15), and (A.17), the goal here is to have $\eta_{\mathrm{tm}}^{n,k,i}$ and $\eta_{\mathrm{sp}}^{n,k,i}$ of comparable size instead of getting either $\eta_{\mathrm{tm}}^{n,k,i}$ much smaller than $\eta_{\mathrm{sp}}^{n,k,i}$ or vice versa. Further, a local version of (A.19) can also be stated: for all $D \in \mathcal{D}_h^n$,

$$\eta_{\mathrm{tm},\alpha,D}^{n,k,i} \approx \eta_{\mathrm{sp},\alpha,D}^{n,k,i}, \qquad \alpha \in \{\mathrm{n},\mathrm{t}\}. \tag{A.20}$$

Achieving (A.19) or (A.20) is to be done by changing the time step τ^n or the spatial mesh $\mathcal{T}_h^n$. In the latter, the goal is to ensure that all $\eta_{\mathrm{sp},\alpha,D}^{n,k,i}$, $\alpha \in \{\mathrm{n},\mathrm{t}\}$, $D \in \mathcal{D}_h^n$, are of comparable size, i.e., that error is equally distributed in space.

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] G. Akrivis, C. Makridakis, and R. H. Nochetto, *A posteriori error estimates for the Crank-Nicolson method for parabolic equations*, Math. Comp., 75 (2006), pp. 511–531.

[3] H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311–341.

[4] L. Angermann, P. Knabner, and K. Thiele, *An error estimator for a finite volume discretization of density driven flow in porous media*, Appl. Numer. Math., 26 (1998), pp. 179–191. Proceedings of the International Centre for Mathematical Sciences Conference on Grid Adaptation in Computational PDEs: Theory and Applications (Edinburgh, 1996).

[5] S. N. Antontsev, A. V. Kazhikhov, and V. N. Monakhov, *Boundary value problems in mechanics of nonhomogeneous fluids*, North-Holland, Amsterdam, 1990. Studies in Mathematics and Its Applications, Vol. 22.

[6] T. Arbogast, *The existence of weak solutions to single porosity and simple dual-porosity models of two-phase incompressible flow*, Nonlinear Anal., 19 (1992), pp. 1009–1031.

[7] J. Bear, *Dynamics of Fluids in Porous Media*, American Elsevier, New York, 1972.

[8] J. Bear and Y. Bachmat, *Introduction to Modeling of Transport Phenomena in Porous Media*, vol. 4 of Theory and Applications of Transport in Porous Media, Kluwer Academic Publishers, Dordrecht, Holland, 1990.

[9] F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*, vol. 15 of Springer Series in Computational Mathematics, Springer-Verlag, New York, 1991.

[10] C. Cancès and T. Gallouët, *On the time continuity of entropy solutions*, J. Evol. Equ., 11 (2011).

[11] J. Carrillo, *Entropy solutions for nonlinear degenerate problems*, Arch. Ration. Mech. Anal., 147 (1999), pp. 269–361.

[12] G. Chavent and J. Jaffré, *Mathematical models and finite elements for reservoir simulation*, North-Holland, Amsterdam, 1986. Studies in Mathematics and Its Applications, Vol. 17.

[13] Y. Chen and W. Liu, *A posteriori error estimates of mixed methods for miscible displacement problems*, Internat. J. Numer. Methods Engrg., 73 (2008), pp. 331–343.

[14] Z. Chen, *Degenerate two-phase incompressible flow. I. Existence, uniqueness and regularity of a weak solution*, J. Differential Equations, 171 (2001), pp. 203–232.

[15] ——, *Degenerate two-phase incompressible flow. II. Regularity, stability and stabilization*, J. Differential Equations, 186 (2002), pp. 345–376.

[16] Z. Chen and R. E. Ewing, *Degenerate two-phase incompressible flow. III. Sharp error estimates*, Numer. Math., 90 (2001), pp. 215–240.

[17] ——, *Degenerate two-phase incompressible flow. IV. Local refinement and domain decomposition*, J. Sci. Comput., 18 (2003), pp. 329–360.

[18] Z. Chen and G. Ji, *Sharp L^1 a posteriori error analysis for nonlinear convection-diffusion problems*, Math. Comp., 75 (2006), pp. 43–71.

[19] J. de Frutos, B. García-Archilla, and J. Novo, *A posteriori error estimates for fully discrete nonlinear parabolic problems*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 3462–3474.

[20] D. A. Di Pietro, M. Vohralík, and C. Widmer, *An a posteriori error estimator for a finite volume discretization of the two-phase flow*, in Finite Volumes for Complex Applications VI, J. Fořt, J. Fürst, J. Halama, R. Herbin, and F. Hubert, eds., Berlin, Heidelberg, 2011, Springer-Verlag, pp. 341–349.

[21] L. El Alaoui, A. Ern, and M. Vohralík, *Guaranteed and robust a posteriori error estimates and balancing discretization and linearization errors for monotone nonlinear problems*, Comput. Methods Appl. Mech. Engrg., 200 (2011), pp. 597–613.

[22] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems. IV. Nonlinear problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1729–1749.

[23] A. Ern and M. Vohralík, *A posteriori error estimation based on potential and flux reconstruction for the heat equation*, SIAM J. Numer. Anal., 48 (2010), pp. 198–223.

[24] ——, *Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs*. In preparation, 2011.

[25] R. Eymard and T. Gallouët, *Convergence d'un schéma de type éléments finis–volumes finis pour un système formé d'une équation elliptique et d'une équation hyperbolique*, RAIRO Modél. Math. Anal. Numér., 27 (1993), pp. 843–861.

[26] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis, Vol. VII, North-Holland, Amsterdam, 2000, pp. 713–1020.

[27] R. Eymard, R. Herbin, and A. Michel, *Mathematical study of a petroleum-engineering scheme*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 937–972.

[28] L. Gallimard, P. Ladevèze, and J. P. Pelle, *Error estimation and time–space parameters optimization for FEM non-linear computation*, Computers & Structures, 64 (1997), pp. 145–156.

[29] R. Huber and R. Helmig, *Node-centered finite volume discretizations for the numerical simulation of multiphase flow in heterogeneous porous media*, Comput. Geosci., 4 (2000), pp. 141–164.

[30] W. Jäger and J. Kačur, *Solution of doubly nonlinear and degenerate parabolic problems by relaxation schemes*, RAIRO Modél. Math. Anal. Numér., 29 (1995), pp. 605–627.

[31] P. Jiránek, Z. Strakoš, and M. Vohralík, *A posteriori error estimates including algebraic error and stopping criteria for iterative solvers*, SIAM J. Sci. Comput., 32 (2010), pp. 1567–1590.

[32] D. Kröner and S. Luckhaus, *Flow of oil and water in a porous medium*, J. Differential Equations, 55 (1984), pp. 276–288.

[33] S. N. Kružkov, *First order quasilinear equations with several independent variables.*, Mat. Sb. (N.S.), 81 (123) (1970), pp. 228–255.

[34] A. Michel, *A finite volume scheme for two-phase immiscible flow in porous media*, SIAM J. Numer. Anal., 41 (2003), pp. 1301–1317.

[35] P. Neittaanmäki and S. Repin, *Reliable methods for computer simulation*, vol. 33 of Studies in Mathematics and its Applications, Elsevier Science B.V., Amsterdam, 2004. Error control and a posteriori estimates.

[36] R. H. Nochetto, G. Savaré, and C. Verdi, *A posteriori error estimates for variable time-step discretizations of nonlinear evolution equations*, Comm. Pure Appl. Math., 53 (2000), pp. 525–589.

[37] R. H. Nochetto and C. Verdi, *Approximation of degenerate parabolic problems using numerical integration*, SIAM J. Numer. Anal., 25 (1988), pp. 784–814.

[38] M. Ohlberger, *A posteriori error estimate for finite volume approximations to singularly perturbed nonlinear convection–diffusion equations*, Numer. Math., 87 (2001), pp. 737–761.

[39] ——, *A posteriori error estimates for vertex centered finite volume approximations of convection–diffusion–reaction equations*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 355–387.

[40] F. Otto, *L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations*, J. Differential Equations, 131 (1996), pp. 20–38.

[41] I. S. Pop, *Error estimates for a time discretization method for the Richards' equation*, Comput. Geosci., 6 (2002), pp. 141–160.

[42] I. S. Pop, F. Radu, and P. Knabner, *Mixed finite elements for the Richards' equation: linearization procedure*, J. Comput. Appl. Math., 168 (2004), pp. 365–373.

[43] W. Prager and J. L. Synge, *Approximations in elasticity based on the concept of function space*, Quart. Appl. Math., 5 (1947), pp. 241–269.

[44] F. A. Radu, I. S. Pop, and P. Knabner, *Newton-type methods for the mixed finite element discretization of some degenerate parabolic equations*, in Numerical mathematics and advanced applications, Springer, Berlin, 2006, pp. 1192–1200.

[45] ——, *Error estimates for a mixed finite element discretization of some degenerate parabolic equations*, Numer. Math., 109 (2008), pp. 285–311.

[46] S. I. Repin, *A posteriori estimates for partial differential equations*, vol. 4 of Radon Series on Computational and Applied Mathematics, Walter de Gruyter GmbH & Co. KG, Berlin, 2008.

[47] J. E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II, North-Holland, Amsterdam, 1991, pp. 523–639.

[48] R. Verfürth, *A review of a posteriori error estimation and adaptive mesh-refinement techniques*, Teubner-Wiley, Stuttgart, 1996.

[49] 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*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 487–518.

[50] M. Vohralík, *Guaranteed and fully robust a posteriori error estimates for conforming discretizations of diffusion problems with discontinuous coefficients*, J. Sci. Comput., 46 (2011), pp. 397–438.

[51] M. Vohralík and M. F. Wheeler, *A posteriori error estimates, stopping criteria, and adaptivity for two-phase flows.* In preparation, 2011.