Grid Orientation Effect in coupled Finite Volume Schemes

Robert Eymard, Cindy Guichard, Roland Masson

Grid Orientation Effect in coupled Finite Volume Schemes

R. Eymard[*], C. Guichard[†] and R. Masson[‡]

May 18, 2012

Abstract

The numerical simulation of two-phase flow in a porous medium may lead, when using coupled finite volume schemes on structured grids, to the apparition of the so-called Grid Orientation Effect (GOE). We propose in this paper a procedure to eliminate this phenomenon, based on the use of new fluxes with a new stencil in the discrete version of the convection equation, without changing the discrete scheme for computing the pressure field. Numerical results show that the GOE does not significantly decrease with the size of the discretization using the initial scheme on the coupled problem, but that it is efficiently suppressed by the new procedure, even on coarse meshes. A mathematical study, based on a weak BV inequality using the new fluxes, confirms the convergence of the modified scheme in a particular case.

1 Introduction

In the 1980's, numerous papers have been concerned with the so-called Grid Orientation Effect (GOE), in the framework of oil reservoir simulation. This effect is occurring in the simulation of viscous oil recovery by the injection of a very mobile fluid (water, steam water, miscible gas...). In order to more precisely describe the numerical problem, let us consider the following two-phase flow problem in a bounded open connected domain $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3), with a regular boundary denoted by $\partial\Omega$.

$$\left\{\begin{array}{ll} u_t - \mathrm{div}(k_1(u)\Lambda\nabla p) & = \max(s,0)f(c) + \min(s,0)f(u) \\ (1-u)_t - \mathrm{div}(k_2(u)\Lambda\nabla p) & = \max(s,0)(1-f(c)) + \min(s,0)(1-f(u)), \\ M(u) = k_1(u) + k_2(u) & \text{and } f(u) = \dfrac{k_1(u)}{M(u)}, \end{array}\right. \tag{1}$$

where, for $\boldsymbol{x} \in \Omega$ and $t \geq 0$, $u(\boldsymbol{x}, t) \in [0, 1]$ is the saturation of phase 1, and therefore $1 - u(\boldsymbol{x}, t)$ is the saturation of phase 2, k_1 is the mobility of phase 1 (increasing function such that $k_1(0) = 0$), k_2 is the mobility of phase 2 (decreasing function such that $k_2(1) = 0$), the functions f and M are respectively called the fractional flow and the total mobility, the function s represents a volumic source term, corresponding to injection/pumping fluids into the domain, p is the common pressure of both phases (the capillary pressure is assumed to be negligible in front of the pressure gradients due to injection and production wells) and $\Lambda(\boldsymbol{x})$ denotes the permeability tensor (that is defined by a symmetric positive definite matrix which may depend on the point $\boldsymbol{x} \in \Omega$). The volumic composition of the injected fluid is tuned by the function c, assumed to vary between 0 and 1.

We may then rewrite System (1) as the coupling of an elliptic problem with unknown p and a nonlinear scalar hyperbolic problem with unknown u,

$$\mathrm{div}\ \boldsymbol{v} = s \quad \text{with} \quad \boldsymbol{v} = -M(u)\Lambda\nabla p, \tag{2a}$$

$$u_t + \mathrm{div}(f(u)\boldsymbol{v}) = \max(s,0)f(c) + \min(s,0)f(u) \tag{2b}$$

[*]Université Paris-Est
[†]Université Paris-Est et Université de Nice
[‡]Université de Nice

Let us now consider a coupled finite volume scheme for the approximation of Problem (1), written under the form (2):

$$\sum_{L,(K,L)\in S} F_{K,L}^{n+1} = s_K^{n+1} \text{ and } F_{K,L}^{n+1} + F_{L,K}^{n+1} = 0 \tag{3a}$$

$$|K|\left(u_K^{n+1} - u_K^n\right) + \tau^n \sum_{L,(K,L)\in S} \left(f(u_K^m)(F_{K,L}^{n+1})^{(+)} - f(u_L^m)(F_{L,K}^{n+1})^{(+)} \right) = \tau^n((s_K^{n+1})^{(+)} f(c_K^{n+1}) - (s_K^{n+1})^{(-)} f(u_K^m)). \tag{3b}$$

In the above system, we denote by K, L the control volumes, by $|K|$ the measure of K (volume in 3D, area in 2D), by S the initial stencil of the scheme, defined as the set of pairs (K, L) having a common interface denoted $\sigma_{K,L}$, by n the time index, by τ^n the time step ($\tau^n = t^{n+1} - t^n$), by u_K^n the saturation in control volume K at time t^n, by s_K^{n+1} the quantity $\frac{1}{\tau^n}\int_{t^n}^{t^{n+1}} \int_K s(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t$ and by c_K^{n+1} the quantity $\frac{1}{\tau^n |K|}\int_{t^n}^{t^{n+1}} \int_K c(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t$. The flux $F_{K,L}^{n+1} = (F_{K,L}^{n+1})^{(+)} - (F_{L,K}^{n+1})^{(+)}$ is a generally partially implicit approximation of the flux $-\int_{\sigma_{K,L}} M(u)\Lambda\nabla p \cdot \boldsymbol{n}_{K,L}\mathrm{d}s$ at the interface $\sigma_{K,L}$ at time step n (where $\boldsymbol{n}_{K,L}$ is the unit normal vector to $\sigma_{K,L}$ oriented from K to L), and, for all real a, the values $a^{(+)}$ and $a^{(-)}$ are non-negative and such that $a^{(+)} - a^{(-)} = a$. For example, assuming $\Lambda = \mathrm{Id}$, we may use an admissible mesh in the sense of [8], that is a partition of the domain Ω in control volumes denoted $K \in \mathcal{M}$ such that a particular point $\boldsymbol{x}_K \in K$ is called the "centre" of K. The mesh and the points $\boldsymbol{x}_K$ are assumed to be such that, for a pair (K, L) of neighbouring control volumes, their common interface $\sigma_{K,L}$ is orthogonal to the line $(\boldsymbol{x}_K, \boldsymbol{x}_L)$. We may then define the "Two Point Flux Approximation" $F_{K,L}^{n+1}$ by

$$F_{K,L}^{n+1} = \frac{|\sigma_{K,L}|}{\mathrm{d}(\boldsymbol{x}_K, \boldsymbol{x}_L)} \frac{2M(u_K^m)M(u_L^m)}{M(u_K^m) + M(u_L^m)}(p_K^{n+1} - p_L^{n+1}), \tag{4}$$

denoting by p_K^{n+1} the pressure in the control volume K at time t^{n+1}, by $|\sigma_{K,L}|$ the measure in $\mathbb{R}^{d-1}$ of $\sigma_{K,L}$, and by $\mathrm{d}(\boldsymbol{x}_K, \boldsymbol{x}_L)$ the distance between $\boldsymbol{x}_K$ and $\boldsymbol{x}_L$. The value m is set to n in the case of the "IMPES" scheme (IMplicit in Pressure and Explicit in Saturation), and to $n + 1$ for the implicit scheme. Then numerical evidence shows that Scheme (3)-(4), whose main features are that of most of the industrial codes for oil reservoir simulation, leads to the apparition of the GOE when the function $M(u)$ strongly depends on u and the more mobile fluid is injected.

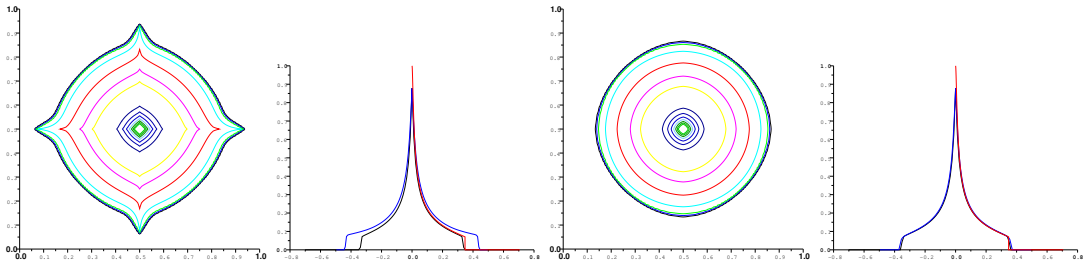


Figure 1: Saturations obtained with Scheme (3): Isovalues of saturation, from boundary to centre: 0.02 0.04 0.06 0.08 0.1 0.125 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55.(extreme left). Saturation profiles: analytical solution (red), profile along median axis (blue), diagonal profile (black). (middle left). Same isovalues (middle right) and saturation profiles (extreme right) for decoupled scheme.

We may observe this effect in the left part of Figure 1, which is resulting from a test case presented in details in the numerical section 3: a mobile fluid is injected at the centre of a square domain gridded by a 121×121 square mesh, and the coupled problem is solved using Scheme (3)-(4). The boundary conditions are prescribed such that there exists an analytical solution with radial symmetry. We see on this figure the advance of the injected fluid along the axes of the grid,

whereas the fluid is late along the diagonals of the grid: this characterises the GOE. Directly setting in Scheme (3b) the analytical values $F_{K,L}^{n+1}$ arising from the radial symmetric solution (this scheme is then called the "decoupled scheme" in Section 3), we obtain the results given in the right part of Figure 1, which no longer show significant GOE on the same grid. We see on this example that the GOE, created by this numerical coupling mechanism, does not seem to easily vanish on fine grids; unfortunately, in the general case, we miss an expression for decoupled fluxes. The solution consisting in using unstructured grids cannot be used in the industrial codes, in which the meshes have to fit the geological layers. Since the approximate fields that are expected in the oil reservoir engineering should be independent of the grid, practical solutions for getting rid of the GOE have been developed in industrial codes using a discretization similar to (4) on structured and regular meshes (mainly based on rectangular parallelepipedic meshes) or Corner Point Geometry [12]. The literature on this problem is huge, and is impossible to exhaustively quote; let us only cite [4, 5, 9, 13, 14, 15] and references therein.

The aim of this paper is to study a new method, consisting in changing the stencil and the fluxes for the approximation of (2b), without modifying the approximation of (2a). The advantage of this method is that it preserves the consistency and convergence properties of the approximation for the second-order space terms. Indeed, the new scheme holds in cases where (4) is no longer used for discretising (2a), and is replaced by a finite volume method adapted to general meshes [1, 3, 6, 11]. Any finite volume approximation of (2a) is then defined by a stencil $S \subset \mathcal{M}^2$ and values $(F_{K,L}^{n+1})_{(K,L)\in S}$ such that (3a) holds. We then consider, for any $(K,L) \in S$, the splitting of the initial fluxes $F_{K,L}^{n+1}$ along given "paths" from K to L, which are defined as finite sequences of control volumes beginning with K and ending with L. The new fluxes $\widehat{F}_{I,J}^{n+1}$ are then obtained by gathering all the partial fluxes along the paths from K to L passing by (I,J). This leads to the definition of a new stencil $\widehat{S}$, including all these pairs (I,J). Then this procedure meets two essential properties: the first one is that the flux continuity holds

$$\widehat{F}_{K,L}^{n+1} + \widehat{F}_{L,K}^{n+1} = 0, \ \forall (K,L) \in \widehat{S},$$

and the second one is that the balance of the new fluxes in the control volumes is the same as that satisfied by the fluxes $(F_{K,L}^{n+1})_{(K,L)\in S}$:

$$\sum_{L,(K,L)\in\widehat{S}} \widehat{F}_{K,L}^{n+1} = \sum_{L,(K,L)\in S} F_{K,L}^{n+1}, \ \forall K \in \mathcal{M}.$$

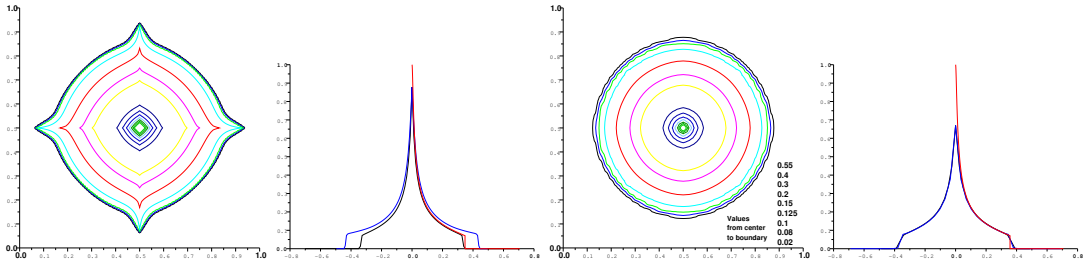


Figure 2: Same isovalues and profiles as in Figure 1, using Scheme (3)-(4) on the fine 121×121 mesh (left) and Scheme (3a)-(4)-(5) on the coarse 41×41 mesh (right).

With these new fluxes and stencil, we replace (3b) by the following new scheme:

$$|K| \ \left(u_K^{n+1} - u_K^n\right) + \tau^n \sum_{L,(K,L)\in\widehat{S}} \left(f(u_K^m)(\widehat{F}_{K,L}^{n+1})^{(+)} - f(u_L^m)(\widehat{F}_{L,K}^{n+1})^{(+)} \right) = \tau^n((s_K^{n+1})^{(+)} f(c_K^{n+1}) - (s_K^{n+1})^{(-)} f(u_K^m)). \tag{5}$$

Figure 2 shows the improvement resulting from the use of the new scheme for the coupled problem on the same problem as the one considered in Figure 1: on a coarse mesh, precise GOE-free results are already obtained.

This paper is organised as follows. In Section 2, we detail the construction procedure of the new fluxes, with the example of the design of a nine-point scheme, starting from a five-point scheme (this is the procedure used for the production of the results shown in Figure 2). Then detailed numerical results are provided in Section 3 for numerically assessing the efficiency of the method. Three test cases are considered. The first one is the nonlinear radial symmetric case, already considered in the introduction of this paper. Then the results of the implicit and explicit coupled schemes are compared on a linear radial symmetric case. Finally, a simple 3D case with three 2D layers shows the possibility to implement the scheme in industrial reservoir simulators. A short conclusion is then proposed in Section 4. Finally, in an appendix, the new scheme is mathematically analysed in the particular case where $f(u) = u$ and $M(u)$ is constant. The interest of this mathematical study is firstly to show how the mathematical features of the new scheme are used in the convergence proof, secondly to exhibit a simple and general sufficient condition on the initial fluxes for ensuring this convergence property.

2 Mesh, stencils and fluxes

2.1 Construction of the new stencil and fluxes

We consider here a finite nonempty set $\mathcal{M}$ called the mesh, whose elements are the control volumes $K, L \in \mathcal{M}$, defined as nonempty open bounded subsets of $\mathbb{R}^d$, $d \in \mathbb{N}^*$ (more detailed geometric properties are not necessary in this section). For any $K \in \mathcal{M}$, we denote by $h_K > 0$ the diameter of K. This section is devoted to the method of construction of a new stencil and of new fluxes, using the initial ones. We say that $S \subset \mathcal{M}^2$ is an "admissible stencil on $\mathcal{M}$" if, for all $K \in \mathcal{M}$, there exists at least one $L \in \mathcal{M} \setminus \{K\}$ such that $(K, L) \in S$, and such that, for all $(K, L) \in S$, then $(L, K) \in S$.
For any $(K, L) \in S$, we assume that is defined a non empty set $\widehat{\mathcal{P}}_{K,L}$ (called the set of the paths from K to L) such that

1. For all $P \in \widehat{\mathcal{P}}_{K,L}$, there exist $m \in \mathbb{N} \setminus \{0\}$ and a set of m distinct control volumes $\{K_1, \dots, K_m\} \subset \mathcal{M}$ with $K_1 = K$ and $K_m = L$ such that

$$P = \{(K_i, K_{i+1}), i = 1, \dots, m-1\}.$$

   By extension, for any $K \in \mathcal{M}$, we write $K \in P$ if there exists $i = 1, \dots, m$ such that $K = K_i$.

2. For any $P = \{(K_i, K_{i+1}), i = 1, \dots, m-1\} \in \widehat{\mathcal{P}}_{K,L}$, we denote by $P^{\leftarrow}$ the inverse path from L to K, defined by $P^{\leftarrow} = \{(K_{i+1}, K_i), i = 1, \dots, m-1\}$. We assume that, for all $(K, L) \in S$, $\widehat{\mathcal{P}}_{L,K} = \{P^{\leftarrow}, P \in \widehat{\mathcal{P}}_{K,L}\}$.

3. The new stencil $\widehat{S} \subset \mathcal{M}^2$, defined by

$$\widehat{S} = \bigcup_{(K,L) \in S, P \in \widehat{\mathcal{P}}_{K,L}} P \tag{6}$$

   satisfies therefore that for all $(K, L) \in \widehat{S}$, $(L, K) \in \widehat{S}$, and is therefore an admissible stencil on $\mathcal{M}$ in the above sense.

4. We denote by $\theta_{\widehat{\mathcal{P}}}$ the value defined by:

$$\theta_{\widehat{\mathcal{P}}} = \max\{\frac{\max(\sum_{M \in P} h_M, \operatorname{diam}(\bigcup_{M \in P} M))}{\min_{M \in P} h_M}, P \in \widehat{\mathcal{P}}_{K,L}, (K, L) \in S\}. \tag{7}$$

   Note that $\theta_{\widehat{\mathcal{P}}}$ is greater than the maximum number of elements in a path.

For an admissible stencil S on $\mathcal{M}$, we consider a real family $(F_{K,L})_{(K,L)\in S}$, which satisfies the following symmetry property:

$$F_{K,L} + F_{L,K} = 0,\ \forall (K,L) \in S. \tag{8}$$

For all $(K,L) \in S$, let $(F^P_{K,L})_{P\in\widehat{\mathcal{P}}_{K,L}}$ be a family such that

$$\forall (K,L) \in S,\ \ \forall P \in \widehat{\mathcal{P}}_{K,L},\ \ F^P_{K,L}F_{K,L} \geq 0, \tag{9}$$

$$\forall (K,L) \in S,\ \ \sum_{P\in\widehat{\mathcal{P}}_{K,L}} F^P_{K,L} = F_{K,L}, \tag{10}$$

and

$$\forall (K,L) \in S,\ \forall P \in \widehat{\mathcal{P}}_{K,L},\ F^{P^\leftarrow}_{L,K} = -F^P_{K,L}. \tag{11}$$

We define the families $(\widetilde{F}_{K,L}, \widetilde{F}^{(+)}_{K,L})_{(K,L)\in\widehat{S}}$ by

$$\begin{array}{l} \forall (I,J) \in \widehat{S}, \\ \widetilde{F}^{(+)}_{I,J} = \displaystyle\sum_{(K,L)\in S}\ \sum_{P\in\widehat{\mathcal{P}}_{K,L}} \xi^P_{I,J} \max\left(F^P_{K,L}, 0\right), \\ \widetilde{F}_{I,J} = \widetilde{F}^{(+)}_{I,J} + \widetilde{F}^{(+)}_{J,I} = \displaystyle\sum_{(K,L)\in S}\ \sum_{P\in\widehat{\mathcal{P}}_{K,L}} \xi^P_{I,J}|F^P_{K,L}|, \end{array} \tag{12}$$

where $\xi^P_{I,J}$ is such that $\xi^P_{I,J} = 1$ if $(I,J) \in P$ and $\xi^P_{I,J} = 0$ otherwise. We finally define, for a given $\nu \in [0,1]$, the families $(\widehat{F}_{I,J}, \widehat{F}^{(+)}_{I,J})_{(I,J)\in\widehat{S}}$ used in the new convection scheme (5) by

$$\forall (I,J) \in \widehat{S},\ \widehat{F}^{(+)}_{I,J} = G_\nu(\widetilde{F}^{(+)}_{I,J}, \widetilde{F}^{(+)}_{J,I})\ \text{ and } \widehat{F}_{I,J} = \widetilde{F}^{(+)}_{I,J} - \widetilde{F}^{(+)}_{J,I} = \widehat{F}^{(+)}_{I,J} - \widehat{F}^{(+)}_{J,I}, \tag{13}$$

where the function G_ν is defined by

$$\forall \nu \in [0,1], \forall (a,b) \in (\mathbb{R}^+)^2,\ G_\nu(a,b) = \max(a-b, \frac{1}{2}(a-b+\nu(a+b)), 0). \tag{14}$$

The function G_ν is designed in order to minimise $G_\nu(a,b)+G_\nu(b,a)$ (hence introducing the smallest additional numerical diffusion to $G_0(a,b) = \max(a-b,0)$) under the constraints $G_\nu(a,b) \geq 0$ (for monotonicity purposes), $G_\nu(a,b) - G_\nu(b,a) = b-a$ (hence ensuring the conservativity) and $G_\nu(a,b) + G_\nu(b,a) \geq \nu(a+b)$ (this property is used for controlling the fluxes $(\widetilde{F}_{K,L})_{(K,L)\in\widehat{S}}$ from the weak BV inequality in the convergence analysis). Indeed, it is straightforward to check that the continuous function $G_\nu(a,b)$ ensures the following property: if $|a-b| > \nu(a+b)$, we have $G_\nu(a,b) = \max(a-b,0)$ and $G_\nu(b,a) = \max(b-a,0)$. Otherwise, we have $G_\nu(a,b) = \frac{1}{2}(a-b+\nu(a+b))$ and $G_\nu(b,a) = \frac{1}{2}(b-a+\nu(a+b))$. Therefore we get

$$\begin{array}{l} (G_\nu(a,b), G_\nu(b,a)) = \text{argmin}\{c+d,\ (c,d) \in (\mathbb{R}^+)^2,\ c-d = a-b,\ c+d \geq \nu(a+b)\}, \\ \forall (a,b) \in (\mathbb{R}^+)^2, \forall \nu \in [0,1]. \end{array} \tag{15}$$

We can then deduce that

$$\forall (I,J) \in \widehat{S},\ \widehat{F}_{I,J} = \widehat{F}^{(+)}_{I,J} - \widehat{F}^{(+)}_{J,I} = \sum_{(K,L)\in S}\ \sum_{P\in\widehat{\mathcal{P}}_{K,L}} \xi^P_{I,J} F^P_{K,L}. \tag{16}$$

Remark 1 If the fluxes F_{KL} are computed using a Multi-Point Flux Approximation scheme (i.e. there exist coefficients $(a^M_{K,L})_{M\in\mathcal{M}}$ such that $F_{K,L} = \sum_{M\in\mathcal{M}} a^M_{K,L} p_M$ and $\sum_{M\in\mathcal{M}} a^M_{K,L} = 0$), and if $F^P_{K,L} = \omega^P_{K,L} F_{KL}$ with $\omega^P_{K,L} \geq 0$ and $\sum_{P\in\widehat{\mathcal{P}}_{K,L}} \omega^P_{K,L} = 1$, we get, using (16), $\widehat{F}_{I,J} = \sum_{M\in\mathcal{M}} \widehat{a}^M_{I,J} p_M$ with

$$\widehat{a}^M_{I,J} = \sum_{(K,L)\in S}\ \sum_{P\in\widehat{\mathcal{P}}_{K,L}} \xi^P_{I,J} \omega^P_{KL} a^M_{KL},$$

and

$$\sum_{M\in\mathcal{M}} \widehat{a}_{I,J}^M = \sum_{(K,L)\in S} \sum_{P\in\widehat{\mathcal{P}}_{K,L}} \xi_{I,J}^P \omega_{KL}^P \sum_{M\in\mathcal{M}} a_{KL}^M = 0.$$

Besides, if we let $\nu = 0$, the relation

$$\widehat{F}_{I,J}^{(+)} = \max(\widehat{F}_{I,J}, 0) \tag{17}$$

holds, which leads to a standard upstream weighting scheme coupled with a Multi-Point Flux Approximation scheme for the pressure, which may be implemented in standard codes with a simple modification of the stencils and transmissivities. Note that the value $\nu = 0$ is excluded in the mathematical analysis provided in Section 4, but that the numerical tests given in Section 3 show that this value seems to be efficient in practice. On the contrary, for $\nu > 0$, which is assumed in the mathematical analysis, the expression of the new fluxes cannot be obtained from a simple Multi-Point Flux Approximation expression.

Remark 2 If we let $\mathcal{P}_{K,L} = \{P_0\}$ with $P_0 = \{(K,L)\}$ (which leads to $\widehat{S} = S$), the new fluxes are identical to the initial ones, independently of ν chosen in $[0,1]$.

2.2 Properties of the new fluxes

We may now state the following result.

Lemma 2.1 (New stencil and fluxes) *Let $\mathcal{M}$ be a finite nonempty set whose elements are nonempty open bounded subset of $\mathbb{R}^d$, $d \in \mathbb{N}^*$, and let $S \subset \mathcal{M}^2$ be an admissible stencil on $\mathcal{M}$ in the sense defined in this paper. For any $K \in \mathcal{M}$, we denote by $h_K > 0$ the diameter of K. Let $(F_{K,L})_{(K,L)\in S}$ be such that (8) holds. Let $(\widehat{\mathcal{P}}_{K,L})_{(K,L)\in S}$, $\theta_{\widehat{\mathcal{P}}}$, $\widehat{S}$ and $(F_{K,L}^P)_{(K,L)\in S, P\in\widehat{\mathcal{P}}_{K,L}}$ be such that (6)-(11) hold. Let $(\widetilde{F}_{K,L}, \widetilde{F}_{K,L}^{(+)})_{(K,L)\in\widehat{S}}$ be defined by (12), let $\nu \in [0,1]$ be given and let $(\widehat{F}_{I,J}, \widehat{F}_{I,J}^{(+)})_{(I,J)\in\widehat{S}}$ be defined by (13). Then the following properties hold:*

$$\forall (I,J) \in \widehat{S},\ \nu\widetilde{F}_{I,J} \le \widehat{F}_{I,J}^{(+)} + \widehat{F}_{J,I}^{(+)}, \tag{18}$$

$$\sum_{L,(K,L)\in\widehat{S}} \widehat{F}_{K,L} = \sum_{L,(K,L)\in S} F_{K,L},\ \forall K \in \mathcal{M}, \tag{19}$$

and

$$\sum_{(K,L)\in\widehat{S}} \max(h_K, h_L)|\widetilde{F}_{K,L}| \le \theta_{\widehat{\mathcal{P}}}^2 \sum_{(K,L)\in S} \max(h_K, h_L)|F_{K,L}|. \tag{20}$$

Proof. We get (18), using the properties (15) of the function G_ν defined by (14).
Let us turn to (19). For a given $I \in \mathcal{M}$, by reordering the sums, we can write that

$$\sum_{J,(I,J)\in\widehat{S}} \widehat{F}_{I,J} = \sum_{J,(I,J)\in\widehat{S}} \sum_{(K,L)\in S} \sum_{P\in\widehat{\mathcal{P}}_{K,L}} \xi_{I,J}^P F_{K,L}^P = \sum_{(K,L)\in S} \sum_{P\in\widehat{\mathcal{P}}_{K,L}} \chi_{I,P} F_{K,L}^P$$

where $\chi_{I,P} = \sum_{J,(I,J)\in\widehat{S}} \xi_{I,J}^P$ is equal to 1 if there exists $J \in \mathcal{M}$ such that $(I,J) \in P$ (therefore $I \neq L$), and to 0 otherwise. Note that, for $(K,L) \in S$ with $K \neq I$ and for $P \in \widehat{\mathcal{P}}_{K,L}$ with $\chi_{I,P} = 1$, we have $I \neq L$, $(L,K) \in S$, $P^{\leftarrow} \in \widehat{\mathcal{P}}_{L,K}$ and $\chi_{I,P^{\leftarrow}} = 1$. So, using (11), we obtain

$$\sum_{(K,L)\in S \text{ s.t. } K\neq I} \sum_{P\in\widehat{\mathcal{P}}_{K,L}} \chi_{I,P} F_{K,L}^P = 0.$$

Therefore we can write, using (10),

$$\sum_{J,(I,J)\in\widehat{S}} \widehat{F}_{I,J} = \sum_{L,(I,L)\in S} \sum_{P\in\widehat{\mathcal{P}}_{I,L}} \chi_{I,P} F_{I,L}^P = \sum_{L,(I,L)\in S} \sum_{P\in\widehat{\mathcal{P}}_{I,L}} F_{I,L}^P = \sum_{L,(I,L)\in S} F_{I,L},$$

which proves (19). Finally, let us prove (20). Thanks to (12), reordering the sums and using (7) and (9), we obtain

$$\begin{aligned}\sum_{(I,J)\in\widehat{S}} \max(h_I,h_J)\widetilde{F}_{I,J} &= \sum_{(I,J)\in\widehat{S}} \max(h_I,h_J) \sum_{(K,L)\in S} \sum_{P\in\widehat{\mathcal{P}}_{K,L}} \xi_{I,J}^P |F_{K,L}^P| \\ &\le \theta_{\widehat{\mathcal{P}}} \sum_{(I,J)\in\widehat{S}} \sum_{(K,L)\in S} \max(h_K,h_L) \sum_{P\in\widehat{\mathcal{P}}_{K,L}} \xi_{I,J}^P |F_{K,L}^P| \\ &= \theta_{\widehat{\mathcal{P}}} \sum_{(K,L)\in S} \max(h_K,h_L) \sum_{P\in\widehat{\mathcal{P}}_{K,L}} \#P |F_{K,L}^P| \\ &\le \theta_{\widehat{\mathcal{P}}}^2 \sum_{(K,L)\in S} \max(h_K,h_L) \sum_{P\in\widehat{\mathcal{P}}_{K,L}} |F_{K,L}^P| \\ &= \theta_{\widehat{\mathcal{P}}}^2 \sum_{(K,L)\in S} \max(h_K,h_L) |F_{K,L}|.\end{aligned}$$

□

Let us provide an example of application of this method.

2.3 Example: construction of a 9-point stencil scheme

We apply the method described in Section 2.1 to 2D structured quadrilateral meshes, assuming that the initial stencil S is the natural five-point stencil. For a given pair of neighbouring control volumes (K,L), we define $\widehat{\mathcal{P}}_{K,L}$ by $\widehat{\mathcal{P}}_{K,L} = \{P_i, i = 0,\dots,4\}$ with $P_0 = \{(K,L)\}$ and $P_i = \{(K,M_i),(M_i,L)\}$ for $i = 1,2,3,4$ (see Figure 3). Then we define $(F_{K,L}^P)_{P\in\widehat{\mathcal{P}}_{K,L}}$ as follows. For a given $\omega > 0$ (the value of ω is discussed below), we take

$$\begin{cases} F_{K,L}^{P_0} = (1-4\omega)F_{K,L} \text{ for } P_0 = \{(K,L)\}, \\ \\ F_{K,L}^{P_i} = \omega F_{K,L} \text{ for } P_i = \{(K,M_i),(M_i,L)\},\ \forall i = 1,2,3,4. \end{cases}$$

Then the new stencil $\widehat{S}$ is the classical nine-point stencil (see Figure 3), defined by

$$\widehat{S} = S \cup \{(K,L)\in\mathcal{M}^2,\ \overline{K} \textit{ and } \overline{L} \textit{ have a common point }\}.$$

This method is illustrated by Figure 3, in which the double solid arrows represent the initial connectivity of the five-point stencil S and the double dashed arrows represent the new connectivity of the nine-point stencil $\widehat{S}$.

Assuming that this procedure has been applied to the whole mesh, let us give two examples of computation of $\widetilde{F}_{K,L}^{(+)}$ resulting from (12):

$$\begin{cases} \widetilde{F}_{K,L}^{(+)} = & (1-4\omega)\max(F_{K,L},0) \\ & +\omega\big(\max(F_{K,M_1},0)+\max(F_{M_2,L},0)+\max(F_{K,M_3},0)+\max(F_{M_4,L},0)\big) \\ \widetilde{F}_{K,M_2}^{(+)} = & \omega\big(\max(F_{K,L},0)+\max(F_{L,M_2},0)+\max(F_{K,M_1},0)+\max(F_{M_1,M_2},0)\big). \end{cases} \tag{21}$$

The values $\widehat{F}_{K,L}^{(+)}$ are then obtained using (13).

Following [7], it is then possible to define an optimal value for ω, if the nine-point new fluxes defined by (21) and (13), setting $\nu = 0$, are used in Scheme (5) on a square grid. Let us assume that there exists a constant velocity $\boldsymbol{v}\in\mathbb{R}^2$ such that

$$F_{K,L}^{n+1} = \int_{\sigma_{K,L}} \boldsymbol{v}\cdot\boldsymbol{n}_{K,L}\mathrm{d}s.$$

We replace the notation u_K^n by $u_{i,j}^n$ in a control volume K whose centre has coordinates ih, jh, for $i,j\in\mathbb{Z}$ and for a given space step $h > 0$. Let us assume, without loss of generality, that the

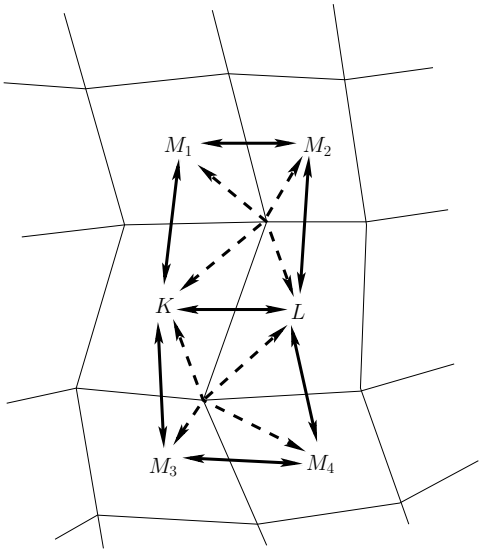


Figure 3: Five and nine point stencils on a structured quadrilateral mesh.

coordinates of $\boldsymbol{v}$ in the axes of the grid are (a, b) with $a \geq b \geq 0$. Then Scheme (5) may be written, using (21) and (17),

$$\begin{array}{l} h^2\left(u_{i,j}^{n+1} - u_{i,j}^{n}\right) \\ +\tau^n h\big((1-4\omega)a(f(u_{i,j}^m) - f(u_{i-1,j}^m)) + (1-4\omega)b(f(u_{i,j}^m) - f(u_{i,j-1}^m)) \\ +2\omega(a+b)(f(u_{i,j}^m) - f(u_{i-1,j-1}^m)) + 2\omega(a-b)(f(u_{i,j}^m) - f(u_{i-1,j+1}^m))\big) \\ = \tau^n((s_{i,j}^{n+1})^{(+)} f(c_{i,j}^{n+1}) - (s_{i,j}^{n+1})^{(-)} f(u_{i,j}^m)). \end{array} \tag{22}$$

Thanks to the following Taylor expansions:

$$\begin{array}{lll} f(u_{i-1,j}^m) = & f(u_{i,j}^m) - h\partial_x f(u_{i,j}^m) & \\ & +\frac{h^2}{2}\partial_{xx}^2 f(u_{i,j}^m) & +\varepsilon_{i-1,j}^m, \\ f(u_{i,j-1}^m) = & f(u_{i,j}^m) - h\partial_y f(u_{i,j}^m) & \\ & +\frac{h^2}{2}\partial_{yy}^2 f(u_{i,j}^m) & +\varepsilon_{i,j-1}^m, \\ f(u_{i-1,j-1}^m) = & f(u_{i,j}^m) - h\partial_x f(u_{i,j}^m) - h\partial_y f(u_{i,j}^m) & \\ & +\frac{h^2}{2}(\partial_{xx}^2 f(u_{i,j}^m) + 2\partial_{xy}^2 f(u_{i,j}^m) + \partial_{yy}^2 f(u_{i,j}^m)) & +\varepsilon_{i-1,j-1}^m, \\ f(u_{i-1,j+1}^m) = & f(u_{i,j}^m) - h\partial_x f(u_{i,j}^m) + h\partial_y f(u_{i,j}^m) & \\ & +\frac{h^2}{2}(\partial_{xx}^2 f(u_{i,j}^m) - 2\partial_{xy}^2 f(u_{i,j}^m) + \partial_{yy}^2 f(u_{i,j}^m)) & +\varepsilon_{i-1,j+1}^m, \end{array}$$

we may express the numerical diffusion term E_c, resulting from the upstream weighting scheme (22) for the approximation of the continuous equation (2b), by

$$E_c = -\frac{h}{2}(a\partial_{xx}^2 f(u) + 8\omega b\partial_{xy}^2 f(u) + (b + 4\omega(a-b))\partial_{yy}^2 f(u)) = -h \text{ div}(\ D(0,\omega,\boldsymbol{v})\ \nabla f(u)),$$

where $D(0,\omega,\boldsymbol{v})$ (the value 0 standing for the initial grid) is the linear mapping, whose matrix in the axes of the grid is given by

$$M = \begin{bmatrix} m_{11} & m_{12} \\ m_{12} & m_{22} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} a & 4\omega b \\ 4\omega b & b + 4\omega(a-b) \end{bmatrix}.$$

The mapping $D(\theta,\omega,\boldsymbol{v})$ is then defined as the numerical diffusion operator of Scheme (22) in a grid whose axes are turned by the angle θ with respect to the initial grid. Then the GOE due to this diffusion term would be theoretically suppressed if we could find a real value ω such that

$D(\theta,\omega,\boldsymbol{v})$ is independent of θ and $\boldsymbol{v}$. Unfortunately, as we show below, this general problem does not seem to have a solution. But we are able to solve a weaker problem: find $\omega \in [0, \frac{1}{4}]$ (in order to ensure condition (9)) such that $D(0,\omega,\boldsymbol{v}) = D(-\frac{\pi}{4},\omega,\boldsymbol{v})$ for all $\boldsymbol{v} \in \mathbb{R}^2$ (note that this rotation leads to the highest discrepancy between the numerical results on the two grids, when using the initial scheme corresponding to $\omega = 0$). Let us express the matrix $\widetilde{M}$ of $D(0,\omega,\boldsymbol{v})$ in the axes turned by $-\frac{\pi}{4}$:

$$\widetilde{M} = \begin{bmatrix} \widetilde{m}_{11} & \widetilde{m}_{12} \\ \widetilde{m}_{12} & \widetilde{m}_{22} \end{bmatrix} = P^{-1}MP, \text{ with } P = \begin{bmatrix} c_\theta & s_\theta \\ -s_\theta & c_\theta \end{bmatrix}$$

denoting $c_\theta = \frac{\sqrt{2}}{2}$ and $s_\theta = \frac{\sqrt{2}}{2}$. It gives

$$\begin{aligned} \widetilde{m}_{11} &= c_\theta^2 m_{11} - 2c_\theta s_\theta m_{12} + s_\theta^2 m_{22}, \\ \widetilde{m}_{12} &= c_\theta s_\theta (m_{11} - m_{22}) + (c_\theta^2 - s_\theta^2) m_{12}, \\ \widetilde{m}_{22} &= s_\theta^2 m_{11} + 2c_\theta s_\theta m_{12} + c_\theta^2 m_{22}. \end{aligned}$$

Let us now express the matrix $\widehat{M}$ of $D(-\frac{\pi}{4},\omega,\boldsymbol{v})$ in the same axes (which are those of the turned grid). The coordinates of $\boldsymbol{v}$ in these axes are given by $(\widehat{a},\widehat{b}) = (c_\theta a - s_\theta b, s_\theta a + c_\theta b)$, which are such that $\widehat{b} \geq \widehat{a} \geq 0$ (since $s_\theta = c_\theta = \frac{\sqrt{2}}{2}$ and $a \geq b \geq 0$), and the matrix $\widehat{M}$ of $D(-\frac{\pi}{4},\omega,\boldsymbol{v})$ is given by the following expression, similar to that of M:

$$\widehat{M} = \begin{bmatrix} \widehat{m}_{11} & \widehat{m}_{12} \\ \widehat{m}_{12} & \widehat{m}_{22} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \widehat{a} + 4\omega(\widehat{b} - \widehat{a}) & 4\omega\widehat{a} \\ 4\omega\widehat{a} & \widehat{b} \end{bmatrix}.$$

The identity $D(0,\omega,\boldsymbol{v}) = D(-\frac{\pi}{4},\omega,\boldsymbol{v})$ for all $\boldsymbol{v} \in \mathbb{R}^2$ is obtained under the condition $\widetilde{M} = \widehat{M}$ for all $\boldsymbol{v} \in \mathbb{R}^2$, which may be expressed, for all $a \geq b \geq 0$, by

$$\begin{aligned} c_\theta^2 a - 2c_\theta s_\theta 4\omega b + s_\theta^2 (b + 4\omega(a-b)) &= c_\theta a - s_\theta b + 4\omega(s_\theta a + c_\theta b - c_\theta a + s_\theta b), \\ c_\theta s_\theta (a - (b + 4\omega(a-b))) + (c_\theta^2 - s_\theta^2) 4\omega b &= 4\omega(c_\theta a - s_\theta b), \\ s_\theta^2 a + 2c_\theta s_\theta 4\omega b + c_\theta^2 (b + 4\omega(a-b)) &= s_\theta a + c_\theta b. \end{aligned}$$

The above system then implies

$$\begin{aligned} c_\theta^2 + s_\theta^2 4\omega &= c_\theta + 4\omega(s_\theta - c_\theta), \\ -2c_\theta s_\theta 4\omega + s_\theta^2(1 - 4\omega) &= -s_\theta + 4\omega(c_\theta + s_\theta), \\ c_\theta s_\theta (1 - 4\omega) &= 4\omega c_\theta, \\ -c_\theta s_\theta (1 - 4\omega) + (c_\theta^2 - s_\theta^2) 4\omega &= -4\omega s_\theta, \\ s_\theta^2 + c_\theta^2 4\omega &= s_\theta, \\ 2c_\theta s_\theta 4\omega + c_\theta^2 (1 - 4\omega) &= c_\theta. \end{aligned}$$

It is remarkable that there exists a solution to the above system with 6 equations and 1 unknown in the case $c_\theta = \frac{\sqrt{2}}{2}$ and $s_\theta = \frac{\sqrt{2}}{2}$ (there is no solution for a general rotation angle). This solution is given by

$$\omega = \frac{\sqrt{2} - 1}{4} \simeq 0.1036.$$

This choice is shown to be efficient for suppressing the GOE in the numerical results provided below, as well as in [5].

3 Numerical results

3.1 A 2D nonlinear case with radial symmetry

We consider Problem (1) on $\Omega = (0,1)^2$ in the isotropic case $\Lambda = \mathrm{Id}$, with the following data:

$$k_1(u) = u^2,\ k_2(u) = \frac{1}{\mu}(1-u)^2,\ M(u) = k_1(u) + k_2(u),\ f(u) = \frac{k_1(u)}{M(u)},$$

where $\mu > 0$ corresponds to a viscosity ratio between the two phases. Let $C = (\frac{1}{2}, \frac{1}{2})$ be the centre of the domain Ω (see Figure 4), and consider the polar coordinates (r, θ) with centre C and local basis denoted by $(\mathbf{e}_r, \mathbf{e}_\theta)$. We look for solutions $p(r,t)$, $u(r,t)$ of (1), depending only on time t and on r. For this, we prescribe the following output total flux boundary condition:

$$M(u)\nabla p \cdot \boldsymbol{n}_{\partial\Omega} = -\boldsymbol{v}(r) \cdot \boldsymbol{n}_{\partial\Omega} \text{ on } \partial\Omega,$$

where $\boldsymbol{v}(r) = \frac{1}{2\pi r}\mathbf{e}_r$. Hence, at each linear segment $[A, B]$ of the boundary, one has $\int_{[A,B]} \boldsymbol{v} \cdot \mathbf{n}_{\partial\Omega} \mathrm{d}s = |\widehat{ACB}|/(2\pi)$, where $|\widehat{ACB}|$ denotes the measure of the angle between the segments $[C, A]$ and $[C, B]$ (see Figure 4). A punctual source term, equal to 1, is imposed at the point C

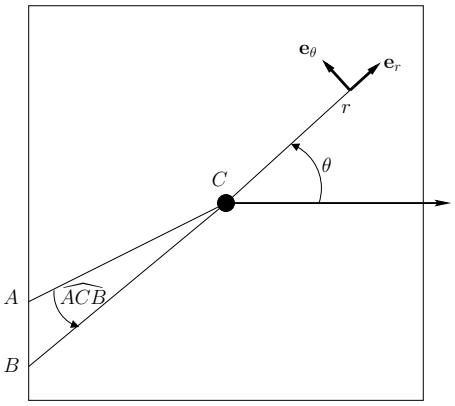


Figure 4: Geometry of the radial circular test.

with input saturation $u(0,t) = 1$. Then, there exists a unique entropy weak analytical solution (p, u) only depending on r and t, called the Buckley-Leverett solution in the framework of oil engineering (recall that, for a nonlinear problem without an entropy criterion, there exists an infinity of weak solutions in the general 1D case; in this 2D case, there may exist weak solutions without radial symmetry):

$$\begin{aligned}
&\bar{u} = 1/\sqrt{\mu+1},\ \bar{v} = (1+\sqrt{\mu+1})/2,\ \bar{r}(t) = \sqrt{\bar{v}t/\pi} \\
&u(r,t) = 0 \text{ for } r > \bar{r}(t), \\
&u(r,t) = (f')^{(-1)}(\pi r^2/t) \text{ for } r < \bar{r}(t), \\
&M(u)\nabla p = -\boldsymbol{v}(r) \text{ which gives } p(r,t) = \int_r^{r_0} \frac{1}{2\pi s M(u(s,t))}\mathrm{d}s + p_0,
\end{aligned}$$

where the value of the pressure is fixed at p_0 at the distance r_0 to point C. The above solution shows a circular discontinuity with height $\bar{u}$, located at the circle with centre C and radius $\bar{r}(t)$. We consider the case where $\mu = 200$ and $t = 0.05$. We then have $\bar{u} \simeq 0.07$ and $\bar{r}(t) \simeq 0.353$ (these are the data used for Figures 1 and 2 in the introduction).

Scheme (3a)-(4) has been implemented in a prototype running under SCILAB environment, together with the method described in Section 2.3 for computing the new fluxes $\widehat{F}_{K,L}^{n+1}$ from $F_{K,L}^{n+1}$. The IMPES scheme $m = n$ is chosen and we compute the new values u_K^{n+1} from Scheme (5). The strategy for determining the time step is based on a desired maximum variation of saturation between two time steps (equal to 0.05).

In order to assess the part of the GOE which is due to the coupling between the two finite volume schemes, we compare the results of Scheme (3a)-(4)-(5), called in the following the *coupled scheme* (recall that, thanks to (21), this scheme is identical, if $\omega = 0$, to the initial scheme (3)-(4)) with the results that are obtained by the scheme consisting in the only resolution of (5) (which is identical to (3b) if $\omega = 0$), in which $F_{K,L}^{n+1}$ is given by the constant radial signed flux

$$F_{K,L}^{n+1} = \frac{|\widehat{ACB}|}{2\pi}, \text{ with } \sigma_{K,L} = [A, B] \text{ and } \boldsymbol{n}_{K,L} \cdot (\frac{A+B}{2} - C) \geq 0, \tag{23}$$

hence defining the *decoupled scheme*. Note that in this decoupled case, the upstream weighted finite volume (5) may be proved, for any value of $\omega \in [0, \frac{1}{4}]$, to converge to the unique entropy weak solution of the problem.

Let us start with a qualitative study of the effect of the parameter ω on the GOE in the case of coarse meshes (which are realistic in practical applications). We respectively plot in Figures 5 and 6 the contours of the saturation and the profiles of the saturation along the median and the diagonal axes, for three values of ω (the grid is the 41×41 one and we set $\nu = 0.1$). Let us remark that the intense GOE in the initial scheme ($\omega = 0$) is completely suppressed with the value $\omega = 0.1$. It might be connected to the analysis of the numerical diffusion in Section 2.3. Then the value $\omega = 0.2$ generates some GOE similar to that which would result from a rotation of the grid with angle $\pi/4$. We see in Figure 6 that the profiles along the median and diagonal axes are nearly not distinguishable, and very close to the analytical solution (this is confirmed by the convergence results given below).

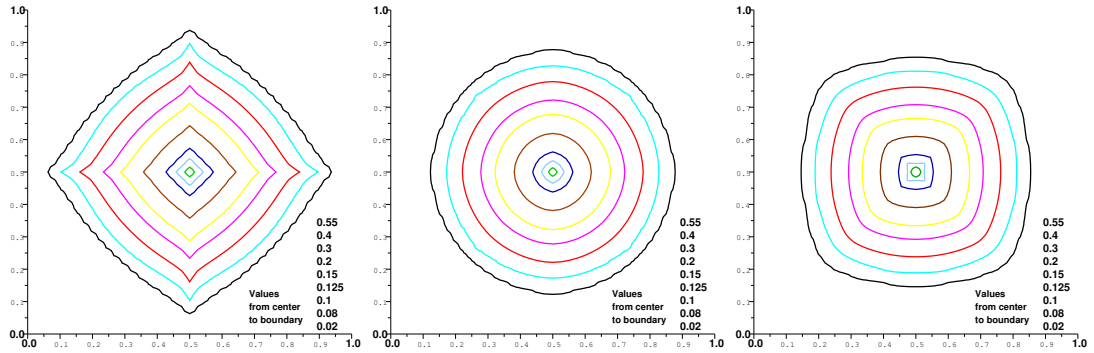


Figure 5: Contours of saturations, using the 41×41 mesh, at $t = 0.05$ ($\mu = 200$) with the initial scheme (identical to modified scheme with $\omega = 0$, left), the modified scheme with $\omega = 0.1$ (middle) and the modified scheme with $\omega = 0.2$ (right).

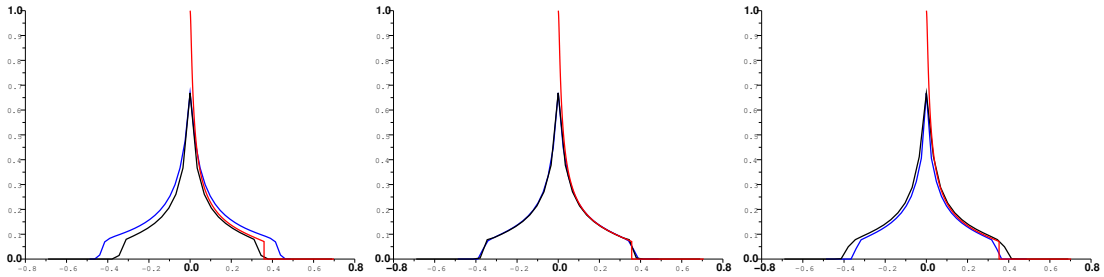


Figure 6: Profiles of saturations, using the 41×41 mesh: analytical solution (red), profile along median axis (blue), diagonal profile (black); initial scheme (left), $\omega = 0.1$ (middle), $\omega = 0.2$ (right).

We now study the behaviour of the approximate pressure with respect to ω, in the coupled scheme (3a)-(4)-(5). Although the analytical pressure tends to $+\infty$ as r tends to 0, we do not draw this infinite branch. In Figure 7, we compare the profiles of the pressure along the median axis, the diagonal axis and the analytical solution, for the initial scheme $\omega = 0$, and the modified scheme $\omega = 0.1$ and $\omega = 0.2$. We observe the confirmation that the pressure is not directly influenced by the GOE, but that increasing values of ω lead to a higher range of the pressure, due to the fact that an increase of the numerical diffusion leads to a decrease of the average values of the saturation near Point C (this can be observed in Figure 6 for the saturation profile along the median axis). We then numerically observe that the parameter $\nu \in [0, 1]$ has only a small influence on the results (recall that $\nu > 0$ is necessary for the convergence proof). Setting $\omega = 0.1$, we see a slight

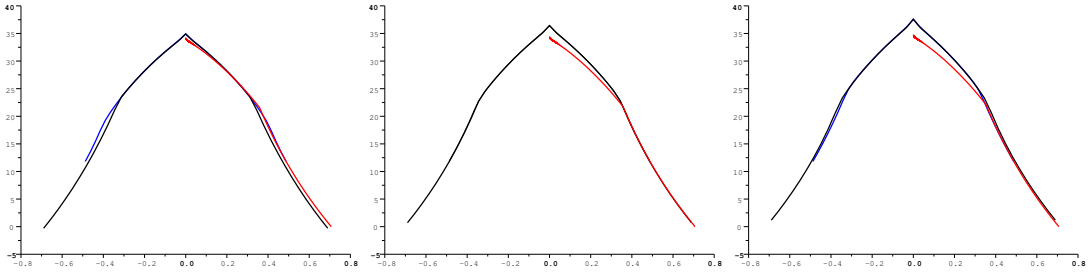


Figure 7: Profiles of pressures, using the 41×41 mesh (p_0 set to 0 at distance $r_0 = \sqrt{2}/2$) : analytical solution (red), profile along median axis (blue), diagonal profile (black); the initial scheme (left) and the modified scheme with $\omega = 0.1$ (middle), with $\omega = 0.2$ (right).

difference on the saturation profiles between the cases $\nu = 0.1$ and $\nu = 1$ (in the latter case, the profile is slightly more diffused for the small values, see Figure 8); there is no difference between the values $\nu = 0.1$ and $\nu = 0$. In the following, all tests are done with $\nu \in [0, 0.1]$.

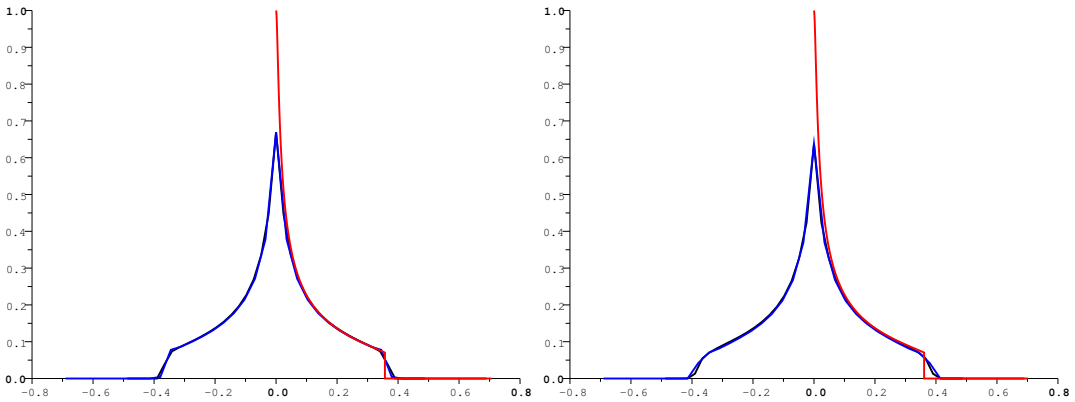


Figure 8: Profiles of saturation, using the 41×41 mesh: analytical solution (red), profile along median axis (blue), diagonal profile (black); modified scheme with $\nu = 0.1$ and $\omega = 0.1$ (left), with $\nu = 1$ and $\omega = 0.1$ (right).

Let us now turn to the convergence orders as the size of the mesh decreases. We give in Table 1 the $L^1(\Omega)$-error of the saturation for six $k \times k$ square meshes, selecting $k = 21, 41, 61, 81, 101, 121$, and for the same three values of ω as above (0 for the initial scheme, 0.1 for the best correction of the GOE, 0.2 for the purpose of the comparison). We remark that there is no clear indication that the scheme is converging to the analytical solution in the case $\omega = 0$, in confirmation with Figure 1. On the contrary, the value $\omega = 0.1$ provides a significantly converging behaviour. The contents of Table 1 is plotted in Figure 9. It is particularly clear in this figure that the convergence properties of the coupled scheme are completely different from that of the decoupled one.
In order to observe the interaction between the convergence of the saturation and that of the fluxes in this coupled case, we explore in Table 2 the E_1 error of the fluxes in the following sense (the hypothesis that this error tends to 0 is done in the convergence theorem 4.4 in the appendix):

$$E_1 = \sum_{(K,L)\in S} \max(h_K, h_L) \left| F_{K,L}^{n+1} - \int_{\sigma_{K,L}} \boldsymbol{v} \cdot \boldsymbol{n}_{K,L} \mathrm{d}s \right|, \tag{24}$$

where n corresponds to the final time. We observe that the convergence of the fluxes is again much stronger in the case $\omega = 0.1$ than in the two other cases. It is worth noticing that the error

size	$\omega = 0$	conv. ord.	$\omega = 0.1$	conv. ord.	$\omega = 0.2$	conv. ord.
21 cpl	0.00912	-	0.00472	-	0.00762	-
41 cpl	0.00677	0.445	0.00262	0.879	0.00444	0.807
61 cpl	0.00573	0.419	0.00169	1.10	0.00382	0.378
81 cpl	0.00531	0.268	0.00132	0.871	0.00334	0.473
101 cpl	0.00505	0.227	0.00116	0.585	0.00315	0.265
121 cpl	0.00478	0.304	0.000969	0.996	0.00293	0.401
21 dec	0.00513	-	0.00467	-	0.00646	-
41 dec	0.00330	0.659	0.00249	0.939	0.00360	0.873
61 dec	0.00241	0.791	0.00163	1.06	0.00258	0.838
81 dec	0.00188	0.876	0.00126	0.908	0.00205	0.811
101 dec	0.00155	0.874	0.00111	0.574	0.00171	0.822
121 dec	0.00140	0.563	0.000899	1.17	0.00148	0.799

Table 1: $L^1(\Omega)$-errors of the saturation at time 0.05, coupled scheme (denoted by "cpl") and decoupled scheme (denoted by "dec")

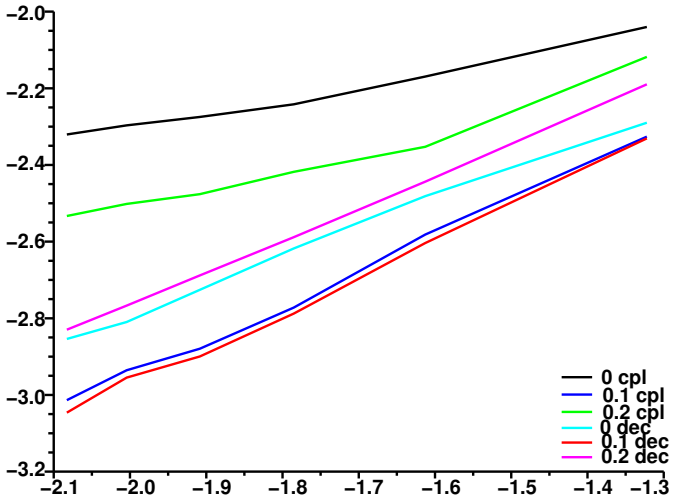


Figure 9: $\log_{10}(\|u_{\mathcal{D}}(\cdot,t) - u(\cdot,t)\|_{L^1(\Omega)})$, where $u_{\mathcal{D}}$ (resp. u) denotes the approximate (resp. analytical) solution, as a function of $-\log_{10}(k)$ for $\omega = 0, 0.1, 0.2$, for the coupled ("cpl") and decoupled ("dec") schemes.

E_2, defined by

$$(E_2)^2 = \sum_{(K,L)\in S} \frac{\max(h_K, h_L)}{|\sigma_{K,L}|} \left(F_{K,L}^{n+1} - \int_{\sigma_{K,L}} \boldsymbol{v} \cdot \boldsymbol{n}_{K,L} \mathrm{d}s \right)^2$$

does not tend to zero in this case where the elliptic problem has a measure in the right hand side.

These numerical results show the efficiency of the new scheme with $\omega = 0.1$, since the error is already reduced for the small values of k. Let us then mention that these convergence results are improved by the nonlinearity of the problem, as shown by comparison with the linear case down below.

size	$\omega = 0$	conv. ord.	$\omega = 0.1$	conv. ord.	$\omega = 0.2$	conv. ord.
21	0.0717	-	0.0158	-	0.0285	-
41	0.0601	0.263	0.0103	0.639	0.0259	0.142
61	0.0534	0.297	0.00786	0.680	0.025202	0.068
81	0.0498289	0.244	0.0064550	0.694	0.0243511	0.121
101	0.0476573	0.201	0.0059253	0.388	0.0236399	0.134
121	0.0459696	0.199	0.0055491	0.363	0.0226995	0.224

Table 2: E_1-error of the fluxes at time 0.05, coupled scheme (L_1 norm of the fluxes 0.7164889).

3.2 A 2D linear case with radial symmetry

We now again consider Problem (1) on $\Omega = (0,1)^2$ in the isotropic case $\Lambda = \mathrm{Id}$, with the following data:

$$k_1(u) = u,\ k_2(u) = 1 - u,\ M(u) = 1,\ f(u) = u.$$

The same radial symmetric conditions as above are imposed, and the solution is given by

$$\begin{aligned}
&\bar{r}(t) = \sqrt{t/\pi} \\
&u(r,t) = 0 \text{ for } r > \bar{r}(t), \\
&u(r,t) = 1 \text{ for } r < \bar{r}(t), \\
&\nabla p = -\boldsymbol{v}(r) \text{ which gives } p(r,t) = \frac{1}{2\pi}\log(\frac{r_0}{r}) + p_0,
\end{aligned}$$

where the value of the pressure is again fixed at p_0 at the distance r_0 to point C. The above solution shows a circular discontinuity with height 1, located at the circle with centre C and radius $\bar{r}(t)$. The final time is taken equal to $t = 0.1$. For the linear runs, explicit scheme $m = n$, we take for the time step the constant $1/k^2$, equal to the measure of the central grid block in which the input flux is equal to 1. In the case of the implicit scheme $m = n + 1$, we prescribe a desired change of saturation between two time steps equal to 10^{-6}, in order to make negligible the error due to the time discretisation. Although we require a such small desired variation in the implicit scheme, we all the same observe that the implicit scheme provides smaller CPU times than the explicit one in this SCILAB implementation of the schemes, due to much larger time steps at the end of the simulation. Note that, in both implicit and explicit cases, the approximate fluxes $F_{K,L}^{n+1}$ and the approximate pressure p_K^{n+1} do no longer depend on n.
We show in Table 3 the $L^1(\Omega)$-error of the saturation for the explicit and implicit coupled schemes respectively. Since the relative difference in the $L^1(\Omega)$-error of the saturation between the coupled scheme (3a)-(4)-(5) and the decoupled scheme (5)-(23) is lower than 2 percent (for both the implicit and the explicit schemes), we don't provide the results for the decoupled scheme (the greatest difference is the case 21×21 with $\omega = 0$: we observe an error equal to 0.0432 for the coupled explicit scheme, 0.0425 for the decoupled explicit scheme, 0.0458 for the coupled implicit scheme and 0.0452 for the decoupled implicit scheme). This is partly resulting from the very good convergence of the fluxes as shown in Table 4: an order 1 is observed although the elliptic problem is singular.

It is interesting to notice that the observed orders of convergence for the $L^1(\Omega)$-error of the saturation are lower than that obtained in the nonlinear case: they remain about 1/2, which is the expected value in the linear case. Although some GOE is visible on the contours of the saturation (see Figure 10), the errors obtained with $\omega = 0.1$ are greater than that of the initial scheme. Note that the results obtained using the explicit and the implicit schemes are very similar. This is due to the fact that the time step is regulated in the explicit case by the measure of one control volume $1/k^2$, instead of behaving as $1/k$ (classical order in a less singular case). Hence the compensation between the time and space errors, which classically occurs for explicit schemes, does not lead to a significant diminution of the error, compared to the implicit scheme, regulated in such a way that the time error remains very small.

size	$\omega = 0$	conv. ord.	$\omega = 0.1$	conv. ord.	$\omega = 0.2$	conv. ord.
21 exp	0.0432	-	0.0460	-	0.0478	-
41 exp	0.0308	0.503	0.0339	0.455	0.0362	0.418
61 exp	0.0255	0.481	0.0275	0.530	0.0290	0.556
81 exp	0.0220	0.516	0.0240	0.482	0.0255	0.455
101 exp	0.0197	0.502	0.0214	0.519	0.0227	0.527
121 exp	0.0180	0.493	0.0196	0.473	0.0209	0.451
21 imp	0.0458	-	0.0484	-	0.0501	-
41 imp	0.0321	0.532	0.0350	0.483	0.0372	0.445
61 imp	0.0262	0.512	0.0281	0.553	0.0296	0.575
81 imp	0.0225	0.525	0.0245	0.490	0.0260	0.462
101 imp	0.0201	0.511	0.0218	0.528	0.0231	0.535
121 imp	0.0184	0.487	0.0200	0.464	0.0213	0.444

Table 3: $L^1(\Omega)$-error of the saturation at time 0.1, explicit coupled scheme ("exp") and implicit coupled scheme ("imp").

k	21	41	61	81	101	121
E_1	0.01341	0.00722	0.00493	0.00375	0.00302	0.00253

Table 4: Linear case, flux error in the coupled scheme, providing a numerical convergence order close to 1.

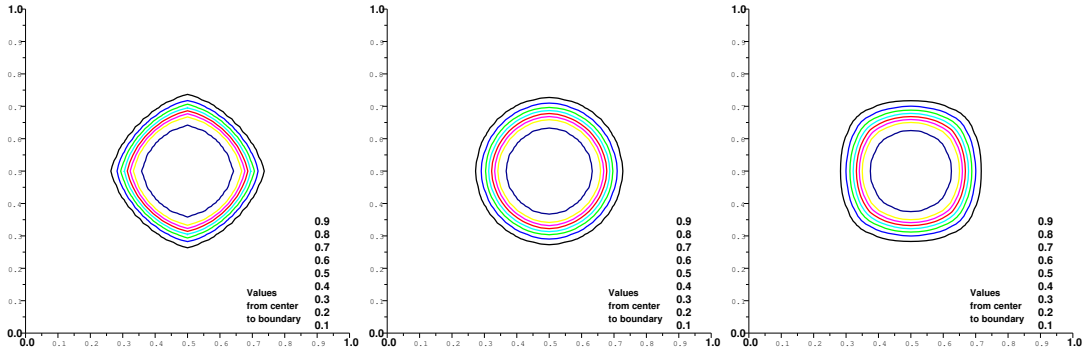


Figure 10: Contours of saturation for the explicit coupled scheme. From left to right: $\omega = 0$ (initial scheme), $\omega = 0.1$ and $\omega = 0.2$ with the 41×41 mesh.

3.3 A 3D test case with three layers

The numerical tests presented here are inspired by [10]. The domain is defined by

$$\Omega = (-0.5, 0.5) \times (-0.5, 0.5) \times (-0.15, 0.15).$$

The permeability $\Lambda(\boldsymbol{x}), \boldsymbol{x} \in \Omega$ is equal to 1 if the distance from $\boldsymbol{x}$ to the vertical axis $0z$ is lower than 0.48, and to 10^{-3} otherwise (see Figure 11), which ensures the confinement of the flow in the cylinder with axis $0z$ and radius 0.48. The density ratio is equal to 0.8. We use Corey-type relative permeability, $k_1(u) = u^4$ and $k_2 = (1-u)^2/100$. At the initial state, the reservoir is assumed to be saturated by the oil phase. Water is injected at the origin by an injection well. Two production wells, denoted by P_1 and P_2, are respectively located at the points $(-0.3\cos\frac{\pi}{3}, -0.3\sin\frac{\pi}{3}, 0)$ and $(0.3\cos\frac{\pi}{3}, -0.3\sin\frac{\pi}{3}, 0)$

A prototype of an industrial code written in FORTRAN, based on an implicit scheme, is used for obtaining numerical results with two Cartesian grids, the second one deduced from the first one by a rotation of angle $\theta = \frac{\pi}{6}$ with axis Oz. The number of cells in each direction (x, y, z) are $N_x = N_y = 51$ and $N_z = 3$ (which means that the three wells are numerically taken into account as source terms in the middle layer of the mesh).

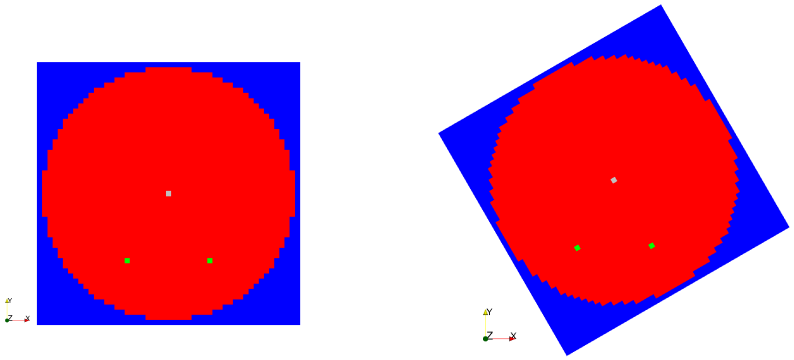
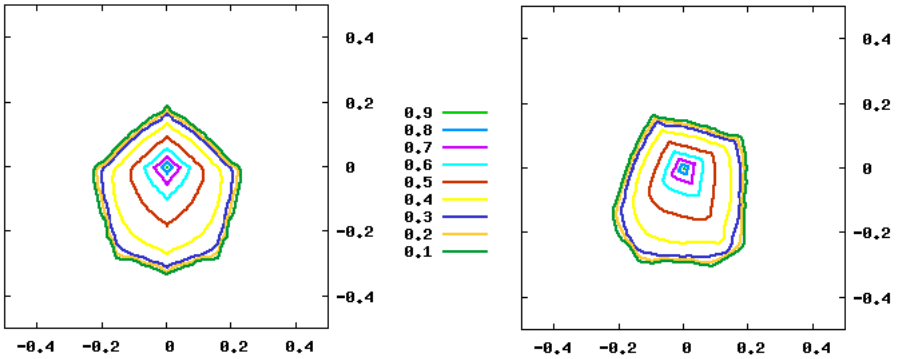


Figure 11: The two meshes used. In red, the highest permeability zone, in blue the lower permeability zone. Squares indicate wells.

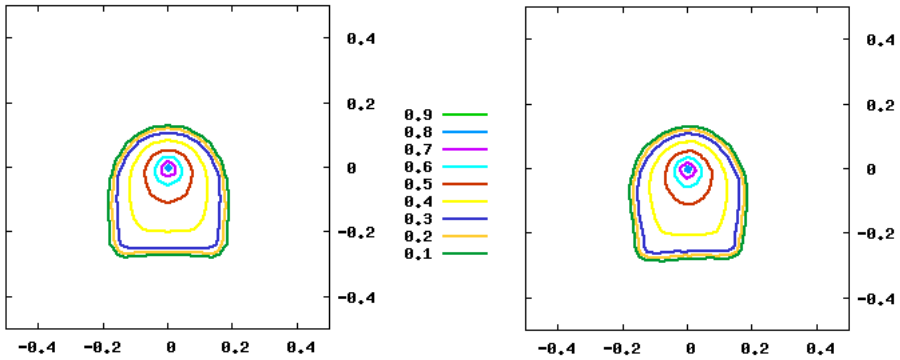
At each time step, we use the Multi-Point Flux Approximation L-scheme [2] for solving the pressure equation, providing the values $F_{K,L}$. Then the method described in section 2.3 is used for the definition of new stencils, selecting $\omega = 0.1$ for all faces which are inscribed in the cylinder. The parameter ν is taken equal to 0, allowing to implement the scheme in standard industrial codes by only modifying the stencil of the Multi-Point Flux Approximation scheme (see Remark 1).

The same value for the time step is used for all the computations, which are stopped once a given quantity of water has been injected. Note that, in the mesh depicted on the right part of Figure 11, the line (P_2, O) becomes the $0y$ axis of the mesh.

We see on Figure 12 the resulting contours of the saturation. We observe that the results obtained using the method described in this paper look very similar in the two grids, whereas the ones obtained using the initial five-point stencil are strongly distorted by the GOE.



(a) $\theta = 0$, stencil S (c) $\theta = \frac{\pi}{6}$, stencil S



(d) $\theta = 0$, stencil $\widehat{S}$ (f) $\theta = \frac{\pi}{6}$, stencil $\widehat{S}$

Figure 12: Water saturation contours u at the same time.

4 Conclusion

In this paper we have considered the nonlinear system of PDE's resulting from the conservation equations of two incompressible immiscible phases flowing within a porous medium. This system, which may also be seen as the coupling of a diffusion equation with respect to the pressure and a convection equation with respect to the saturation, is shown in practical cases of mobility contrast, to lead to the apparition of the so-called Grid Orientation Effect (GOE). We propose a new procedure to overcome this phenomenon, based on the modification of the stencil of the discrete version of the convection equation, without modifying the pressure equation. This procedure preserves the scheme used for the coupled diffusion equation.

Some numerical results, including the comparison with an analytical solution, show the efficiency and the accuracy of the method in nonlinear coupled cases, for which there is no indication that the initial scheme should converge to a GOE-free solution, as the size of the mesh tends to zero. This is different with the linear case, where the GOE can be suppressed by decreasing the size of the mesh, in the same way as in the nonlinear decoupled case.

For some values of the parameters of the method, we obtain a natural version of the nine-point schemes defined some decades ago on regular grids, whose advantage is to apply on the structured but not regular grids used in reservoir simulation, in association with Multi-Point Flux Approximation finite volume schemes. In this case, it may be immediately implemented in standard industrial codes by a simple modification of the stencils.

Appendix: Convergence analysis in a simplified case

For the sake of the mathematical analysis, we only consider Problem (1) in the case where $f(u) = u$ and where the function $k_1(u) + k_2(u)$ is constant. Indeed, the analysis of Problem (1) in the case $k_1(u)+k_2(u)$ not constant is an open problem, and the case of a general function f may be studied using the methods of [8]. Hence the mathematical study is focused on the convergence of the new approximate scheme for the following problem on $\Omega \times (0,T)$:

$$\mathrm{div}\boldsymbol{v} = s, \tag{25}$$

$$u_t + \mathrm{div}(u\boldsymbol{v}) = \max(s,0)c + \min(s,0)u \text{ in } \Omega \times (0,T), \tag{26}$$

together with the initial condition

$$u = u_{\mathrm{ini}} \text{ in } \Omega, \tag{27}$$

under the following hypotheses, denoted (H) in this section:

Definition 4.1 (Hypotheses (H))

1. *Ω is a bounded open connected subset of $\mathbb{R}^d$, $T > 0$ is the period of observation.*
2. *We assume that $\boldsymbol{v} \in C^1(\overline{\Omega})$ is such that $\boldsymbol{v} \cdot \boldsymbol{n}_{\partial\Omega} = 0$ on $\partial\Omega$. We denote by $s = \mathrm{div}\boldsymbol{v}$.*
3. *We assume that $c \in L^\infty(\Omega \times (0,+\infty))$ and $u_{\mathrm{ini}} \in L^\infty(\Omega)$, where the functions c and u_{ini} are essentially bounded by 0 and 1.*

Then Problem (26)-(27) is considered in the following weak sense:

$$\begin{array}{l}\displaystyle\int_0^{+\infty}\int_\Omega (u\varphi_t + u\boldsymbol{v}\cdot\nabla\varphi + (\max(s,0)c + \min(s,0)u)\varphi)\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})\varphi(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} = 0, \\ \forall \varphi \in C_c^\infty(\mathbb{R}^d \times \mathbb{R}).\end{array} \tag{28}$$

4.1 Approximation by an upstream weighting scheme

Let us first precise the definition for the admissible space-time discretizations which will be considered here.

Definition 4.2 *Let $\Omega \subset \mathbb{R}^d$, with $d \in \mathbb{N} \setminus \{0\}$ be a bounded open connected domain and let $T > 0$. We say that $\mathcal{D} = (\mathcal{M}, \mathcal{F}, N, (t^n)_{n=0,\ldots,N})$ is an admissible space-time discretization of $\Omega \times (0,T)$ if:*

1. *The set $\mathcal{M}$ of the control volumes is such that all elements of $\mathcal{M}$ are disjoint open connected subsets of Ω with regular boundary, and such that $\overline{\Omega} = \bigcup_{K\in\mathcal{M}} \overline{K}$. The d-dimensional measure of K (resp. Ω) is denoted by $|K|$ (resp. $|\Omega|$) and the diameter of K is denoted h_K. We denote by $h_\mathcal{D}$ the maximum value of $(h_K)_{K\in\mathcal{M}}$.*
2. *The interior faces of the mesh $\sigma \in \mathcal{F}_{\mathrm{int}}$ are obtained by $\overline{K} \cap \overline{L} := \sigma_{K,L}$, for all pairs of neighbouring control volumes $K \in \mathcal{M}$ and $L \in \mathcal{M}$. They are assumed to be planar, with constant unit normal vector $\boldsymbol{n}_{K,L}$ oriented from K to L. The exterior faces of the mesh $\sigma \in \mathcal{F}_{\mathrm{ext}}$ are obtained by $\sigma = \overline{K} \cap \overline{\partial\Omega}$, for all control volumes $K \in \mathcal{M}$. The set of all the faces of the mesh $\mathcal{F}$ is defined by $\mathcal{F} = \mathcal{F}_{\mathrm{int}} \cup \mathcal{F}_{\mathrm{ext}}$. The $d-1$-dimensional measure of $\sigma \in \mathcal{F}$ is denoted by $|\sigma|$, assumed to be strictly positive. For all $K \in \mathcal{M}$, it is assumed that there exists a subset of $\mathcal{F}$, denoted by $\mathcal{F}_K$, such that $\partial K = \bigcup_{\sigma\in\mathcal{F}_K} \sigma$.*
3. *$N \in \mathbb{N} \setminus \{0\}$ and $(t^n)_{n=0,\ldots,N}$ is a real family such that $t_0 = 0 < t^1 \ldots < t^N = T$.*

We then define

$$\theta_\mathcal{D} = \max_{K\in\mathcal{M}} \frac{h_K \sum_{\sigma\in\mathcal{F}_K} |\sigma|}{|K|}. \tag{29}$$

and we denote $\tau^n = t^{n+1} - t^n$ for $n = 0, \ldots, N-1$.

Assuming Hypotheses (H), let $\mathcal{D} = (\mathcal{M}, \mathcal{F}, N, (t^n)_{n=0,\ldots,N})$ be an admissible space-time discretization of $\Omega \times (0,T)$ in the sense of Definition 4.2. We define an admissible stencil S on $\mathcal{M}$ in the sense precised above by the set of all pairs (K,L) such that $K \in \mathcal{M}$, $L \in \mathcal{M} \setminus \{K\}$ and $|\sigma_{K,L}| > 0$. Let $(F_{K,L})_{(K,L)\in S}$ be such that (8) hold. We then denote, for $\sigma \in \mathcal{F}_K$ such that $\sigma = \sigma_{K,L}$, $F_{K,\sigma} = F_{K,L}$, and for $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\rm ext}$, $F_{K,\sigma} = 0$. We assume that $(F_{K,L})_{(K,L)\in S}$ satisfies the following discrete conservation property

$$\sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma} = \sum_{L,(K,L)\in S} F_{K,L} = s_K,\ \forall K \in \mathcal{M}, \tag{30}$$

where we denote

$$s_K = \int_K s(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\ \forall K \in \mathcal{M}, \tag{31}$$

Let $\widehat{S}$, $(\widehat{\mathcal{P}}_{K,L})_{(K,L)\in S}$, $\theta_{\widehat{\mathcal{P}}}$ and $(F^P_{K,L})_{(K,L)\in S, P\in\widehat{\mathcal{P}}_{K,L}}$ such that (6)-(11) hold. Let $(\widetilde{F}_{K,L}, \widetilde{F}^{(+)}_{K,L})_{(K,L)\in\widehat{S}}$ be defined by (12), let $\nu \in (0,1]$ be given (the value $\nu = 0$ is excluded, since some bounds in Lemma 4.3 and Theorem 4.4 are obtained with respect to $1/\nu$, see also Remark 1) and let $(\widehat{F}_{I,J}, \widehat{F}^{(+)}_{I,J})_{(I,J)\in\widehat{S}}$ be defined by (13).
The implicit version of the upstream weighting scheme devoted for approximating (28) on $[0,T]$ may be written

$$\begin{aligned}
&|K|\,\frac{u_K^{n+1} - u_K^n}{\tau^n} \\
&+ \sum_{L\in\mathcal{M}} \left(\widehat{F}^{(+)}_{K,L} u_K^{n+1} - \widehat{F}^{(+)}_{L,K} u_L^{n+1}\right) + s_K^{(-)} u_K^{n+1} - s_K^{(+)} c_K^{n+1} = 0, \\
&\forall n = 0,\ldots,N-1,\ \forall K \in \mathcal{M},
\end{aligned} \tag{32}$$

letting $\widehat{F}^{(+)}_{I,J} = 0$ for pairs of control volumes $(I,J) \notin \widehat{S}$, and where

$$c_K^{n+1} = \frac{1}{|K|\,\tau^n} \int_{t^n}^{t^{n+1}} \int_K c(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t,\ \forall n = 0,\ldots,N-1,\ \forall K \in \mathcal{M}, \tag{33}$$

$$s_K^{(+)} = \int_K \max(s(\boldsymbol{x}),0)\mathrm{d}\boldsymbol{x},\ s_K^{(-)} = \int_K \max(-s(\boldsymbol{x}),0)\mathrm{d}\boldsymbol{x},\ \forall K \in \mathcal{M}, \tag{34}$$

and

$$u_K^0 = \frac{1}{|K|} \int_K u_{\rm ini}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\ \forall K \in \mathcal{M}. \tag{35}$$

4.2 Estimates

The following lemma, which may be proved in the spirit of [8, Proposition 26.1 p. 918] and [8, Proposition 25.2 p. 913], provides the existence, the uniqueness of the discrete solution and a weak BV-inequality.

Lemma 4.3 **L^∞ estimate, existence, uniqueness of the discrete solution and weak BV-inequality.**
Under Hypotheses (H), let $\mathcal{D} = (\mathcal{M}, \mathcal{F}, N, (t^n)_{n=0,\ldots,N})$ be an admissible space-time discretization of $\Omega \times (0,T)$ in the sense of Definition 4.2. Let S be the set of all pairs (K,L) such that $K \in \mathcal{M}$, $L \in \mathcal{M} \setminus \{K\}$ and $|\sigma_{K,L}| > 0$, and let $(F_{K,L})_{(K,L)\in S}$ be such that (8), (30) and (31) hold. Let $(\widehat{\mathcal{P}}_{K,L})_{(K,L)\in S}$, $\widehat{S}$, $\theta_{\widehat{\mathcal{P}}}$ and $(F^P_{K,L})_{(K,L)\in S, P\in\widehat{\mathcal{P}}_{K,L}}$ such that (6)-(11) hold. Let $(\widetilde{F}_{K,L}, \widetilde{F}^{(+)}_{K,L})_{(K,L)\in\widehat{S}}$ be defined by (12), let $\nu \in (0,1]$ be given and let $(\widehat{F}_{I,J}, \widehat{F}^{(+)}_{I,J})_{(I,J)\in\widehat{S}}$ be defined by (13). Let $(u_K^n)_{K\in\mathcal{M},n=0,\ldots,N}$ be such that (32)-(35) hold. Then

$$0 \le u_K^n \le 1,\ \forall n = 0,\ldots,N,\ \forall K \in \mathcal{M}. \tag{36}$$

Therefore, there exists one and only one $(u_K^n)_{K\in\mathcal{M},n=0,\dots,N}$ *such that* (32)-(35) *hold, which is moreover such that there exists* $C_{\rm BV}>0$*, only depending on* Ω*,* s *and* T *with:*

$$\sum_{n=0}^{N-1}\tau^n\sum_{(K,L)\in\widehat{S}}\widetilde{F}_{K,L}(u_K^{n+1}-u_L^{n+1})^2\le\frac{C_{\rm BV}}{\nu}. \tag{37}$$

4.3 Convergence study

It is now possible to give a convergence proof for the scheme in the linear case. This proof could be extended to the nonlinear scalar hyperbolic case by following the methods proposed in [8], based on the convergence to the unique entropy process solution. Let us also note that, referring to Remark 2, the present mathematical analysis applies (with $\nu>0$) to an upstream weighting scheme written with the initial fluxes.

Theorem 4.4 *Under Hypotheses (H), let* $\mathcal{D}=(\mathcal{M},\mathcal{F},N,(t^n)_{n=0,\dots,N})$ *be an admissible space-time discretization of* $\Omega\times(0,T)$ *in the sense of Definition 4.2. Let* S *be the set of all pairs* (K,L) *such that* $K\in\mathcal{M}$*,* $L\in\mathcal{M}\setminus\{K\}$ *and* $|\sigma_{K,L}|>0$*, and let* $(F_{K,L})_{(K,L)\in S}$ *be such that* (8)*,* (30) *and* (31) *hold. Let* $(\widehat{\mathcal{P}}_{K,L})_{(K,L)\in S}$*,* $\widehat{S}$*,* $\theta_{\widehat{\mathcal{P}}}$ *and* $(F^P_{K,L})_{(K,L)\in S,P\in\widehat{\mathcal{P}}_{K,L}}$ *such that* (6)-(11) *hold. Let* $(\widetilde{F}_{K,L},\widetilde{F}^{(+)}_{K,L})_{(K,L)\in\widehat{S}}$ *be defined by* (12)*, let* $\nu\in(0,1]$ *be given and let* $(\widehat{F}_{I,J},\widehat{F}^{(+)}_{I,J})_{(I,J)\in\widehat{S}}$ *be defined by* (13)*. Let* $(u_K^n)_{K\in\mathcal{M},n=0,\dots,N}$ *be such that* (32)-(35) *hold and let* $u_{\mathcal{D}}$ *be the function defined by*

$$u_{\mathcal{D}}(x,t)=u_K^{n+1},\ \text{for a.e. } (x,t)\in K\times(t^n,t^{n+1}),\ \forall n=0,\dots,N-1,\ \forall K\in\mathcal{M}. \tag{38}$$

We assume that

$$\lim_{h_{\mathcal{D}}\to0}\sum_{(K,L)\in S}\max(h_K,h_L)\left|F_{K,L}-\int_{\sigma_{K,L}}\boldsymbol{v}(\boldsymbol{x})\cdot\boldsymbol{n}_{K,L}\mathrm{d}s(\boldsymbol{x})\right|=0. \tag{39}$$

Then, as $h_{\mathcal{D}}\to0$ *and* $\max\tau^n\to0$ *while* ν *remains fixed,* $\theta_{\mathcal{D}}$ *and* $\theta_{\widehat{\mathcal{P}}}$ *remain bounded,* $u_{\mathcal{D}}$ *converges for the weak-*$\star$ *topology of* $L^\infty(\Omega\times(0,T))$ *to the unique function* $u\in L^\infty(\Omega\times(0,T))$ *satisfying* (28)*.*

Remark 3 Condition (39) *is naturally satisfied if* $F_{K,L}=\int_{\sigma_{K,L}}\boldsymbol{v}(\boldsymbol{x})\cdot\boldsymbol{n}_{K,L}\mathrm{d}s(\boldsymbol{x})$*. More interestingly, it also holds if* $F_{K,L}$ *is obtained using a finite volume scheme for the approximation of Problem* $\mathrm{div}\boldsymbol{v}=s$ *with* $\boldsymbol{v}=-\Lambda\nabla p$ *and Neumann boundary conditions (see [8], pp. 996–1012).*

Proof. In order to prove Theorem 4.4, we consider a sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of admissible space-time discretizations, such that $h_{\mathcal{D}_m}$ (denoted by h_m in the following) and $\max_n(\tau_m^n)$ tend to zero as $m\to\infty$. We assume that, for each m, the families implicitly indexed by m: $(F_{K,L})_{(K,L)\in S}$, $\widehat{S}$, $(\widehat{\mathcal{P}}_{K,L})_{(K,L)\in S}$ and $(F^P_{K,L})_{(K,L)\in S,P\in\widehat{\mathcal{P}}_{K,L}}$ satisfy the hypotheses of the theorem with the same value $\nu\in(0,1]$, while $\theta_{\mathcal{D}_m}$ and $\theta_{\widehat{\mathcal{P}}}$ remain bounded as m tends to ∞. We denote $u_m=u_{\mathcal{D}_m}$ for all $m\in\mathbb{N}$.
Let us prove the convergence of the sequence $(u_m)_{m\in\mathbb{N}}$ to the weak solution u of Problem (28) for the weak-$\star$ topology of $L^\infty(\Omega\times(0,T))$, for all $T>0$. The classical argument of the uniqueness of this limit suffices for concluding the proof of the theorem.
We first notice that, thanks to Lemma 4.3, we get the existence of a subsequence, again noted $(u_m)_{m\in\mathbb{N}}$, which converges to some function $u\in L^\infty(\Omega\times(0,T))$ for the weak-$\star$ topology of $L^\infty(\Omega\times(0,T))$ as $m\to+\infty$. The aim of this proof is to show that u satisfies (28).

Let $\varphi\in C_c^\infty(\mathbb{R}^d\times\mathbb{R})$ be such that $\varphi=0$ in $\mathbb{R}^d\times[T,+\infty)$. In this proof, we denote by C_φ an L^∞ bound of first and second derivatives of φ. Let $m\in\mathbb{N}$. In the following, we drop some indices m, using the notations $\mathcal{D}=\mathcal{D}_m$. We define φ_K^n by

$$\varphi_K^{n+1}=\frac{1}{|K|}\int_K\varphi(\boldsymbol{x},t^n)\mathrm{d}\boldsymbol{x}\mathrm{d}t,\ \forall K\in\mathcal{M},\ \forall n=0,\dots,N.$$

We get the following equation from $\sum_{L\in\mathcal{M}}(\widehat{F}^{(+)}_{K,L}-\widehat{F}^{(+)}_{L,K})+s^{(-)}_K-s^{(+)}_K=0$ and (32):

$$|K|\,(u_K^{n+1}-u_K^n)+\tau^n\sum_{L\in\mathcal{M}}\widehat{F}^{(+)}_{L,K}(u_K^{n+1}-u_L^{n+1})+\tau^n\ s_K^{(+)}(u_K^{n+1}-c_K^{n+1})=0,\quad\forall K\in\mathcal{M}. \tag{40}$$

We then multiply (40) by φ_K^{n+1}, sum over $K\in\mathcal{M}$ and $n=0,\ldots,N-1$. We obtain $T_1^{(m)}+T_2^{(m)}=0$, with

$$T_1^{(m)}=\sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}}|K|\,(u_K^{n+1}-u_K^n)\varphi_K^{n+1}$$

and

$$T_2^{(m)}=\sum_{n=0}^{N-1}\tau^n\sum_{K\in\mathcal{M}}\left(\sum_{L\in\mathcal{M}}\widehat{F}^{(+)}_{L,K}(u_K^{n+1}-u_L^{n+1})\varphi_K^{n+1}+s_K^{(+)}(u_K^{n+1}-c_K^{n+1})\varphi_K^{n+1}\right).$$

We classically show, using the weak-$\star$ convergence of $(u_m)_{m\in\mathbb{N}}$ to u, that

$$\lim_{m\to+\infty}T_1^{(m)}=-\int_0^{+\infty}\int_\Omega u(\boldsymbol{x},t)\varphi_t(\boldsymbol{x},t)\ \mathrm{d}\boldsymbol{x}\mathrm{d}t-\int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})\varphi(\boldsymbol{x},0)\ \mathrm{d}\boldsymbol{x}=0.$$

Let us now prove the convergence of the sequence $(T_2^{(m)})_{m\in\mathbb{N}}$ to T_3 defined by

$$T_3=-\int_0^{+\infty}\int_\Omega(u\boldsymbol{v}\cdot\nabla\varphi+(\max(s,0)c+\min(s,0)u)\varphi)\ \mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

To this purpose, let us now define $T_4^{(m)}$ by

$$T_4^{(m)}=-\sum_{n=0}^{N-1}\tau^n\sum_{K\in\mathcal{M}}\left(u_K^{n+1}\sum_{\sigma\in\mathcal{F}_K}\varphi_\sigma^{n+1}\int_\sigma\boldsymbol{v}(\boldsymbol{x})\cdot\boldsymbol{n}_{K,\sigma}\mathrm{d}s(\boldsymbol{x})\mathrm{d}t+s_K^{(+)}(c_K^{n+1}-u_K^{n+1})\varphi_K^{n+1}\right),$$

with

$$\varphi_\sigma^{n+1}=\frac{1}{|\sigma|}\int_\sigma\varphi(\boldsymbol{x},t^n)\mathrm{d}s(\boldsymbol{x}).$$

The proof that

$$\lim_{m\to+\infty}T_4^{(m)}=T_3$$

can be done, using the weak-$\star$ convergence of $(u_m)_{m\in\mathbb{N}}$ to u and similar techniques to [8, Theorem 35.1 p. 1006]. We now consider $T_5^{(m)}$ (recalling that $F_{K,\sigma}=F_{K,L}$ for $\sigma=\sigma_{K,L}$ else $F_{K,\sigma}=0$ for $\sigma\in\mathcal{F}_{\mathrm{ext}}$), defined by

$$T_5^{(m)}=-\sum_{n=0}^{N-1}\tau^n\sum_{K\in\mathcal{M}}\left(u_K^{n+1}\sum_{\sigma\in\mathcal{F}_K}\varphi_\sigma^{n+1}F_{K,\sigma}+s_K^{(+)}(c_K^{n+1}-u_K^{n+1})\varphi_K^{n+1}\right).$$

We have

$$T_5^{(m)}-T_4^{(m)}=-\sum_{n=0}^{N-1}\tau^n\sum_K u_K^{n+1}\sum_{\sigma\in\mathcal{F}_K}\varphi_\sigma^{n+1}(F_{K,\sigma}-\int_\sigma\boldsymbol{v}(\boldsymbol{x})\cdot\boldsymbol{n}_{K,\sigma}\mathrm{d}s(\boldsymbol{x})).$$

Using (30) (which implies $\sum_{\sigma\in\mathcal{F}_K}(F_{K,\sigma}-\int_\sigma\boldsymbol{v}(\boldsymbol{x})\cdot\boldsymbol{n}_{K,\sigma}\mathrm{d}s(\boldsymbol{x}))=0$), we get

$$T_5^{(m)}-T_4^{(m)}=-\sum_{n=0}^{N-1}\tau^n\sum_K u_K^{n+1}\sum_{\sigma\in\mathcal{F}_K}(\varphi_\sigma^{n+1}-\varphi_K^{n+1})(F_{K,\sigma}-\int_\sigma\boldsymbol{v}(\boldsymbol{x})\cdot\boldsymbol{n}_{K,\sigma}\mathrm{d}s(\boldsymbol{x})),$$

which leads to

$$|T_5^{(m)} - T_4^{(m)}| \leq C_\varphi \sum_{n=0}^{N-1} \tau^n \sum_K \sum_{\sigma \in \mathcal{F}_K} h_K |F_{K,\sigma} - \int_\sigma \boldsymbol{v}(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}s(\boldsymbol{x})|,$$

which tends to zero thanks to (39).
Gathering by pairs of control volumes (each one appears once in the summation), we have

$$T_2^{(m)} - T_5^{(m)} = \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} (u_K^{n+1} - u_L^{n+1})(\widehat{F}_{L,K}^{(+)} \varphi_K^{n+1} - \widehat{F}_{K,L}^{(+)} \varphi_L^{n+1} + F_{K,L} \varphi_{K,L}^{n+1}),$$

setting $\varphi_{K,L}^{n+1} = \varphi_{\sigma_{K,L}}^{n+1}$ if $(K,L) \in S$ else $\varphi_{K,L}^{n+1} = \frac{1}{2}(\varphi_K^{n+1} + \varphi_L^{n+1})$ (recall that $F_{K,L} = 0$ if $(K,L) \notin S$).
Let us prove that $\lim_{m \to \infty} |T_2^{(m)} - T_5^{(m)}| = 0$, result which completes our proof.
Since $\widehat{F}_{K,L}^{(+)} - \widehat{F}_{L,K}^{(+)} = \widehat{F}_{K,L}$, we get $T_2^{(m)} - T_5^{(m)} = T_6^{(m)} + T_7^{(m)}$ with

$$T_6^{(m)} = \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} (u_K^{n+1} - u_L^{n+1}) \big(\widehat{F}_{L,K}^{(+)} (\varphi_K^{n+1} - \varphi_{K,L}^{n+1}) - \widehat{F}_{K,L}^{(+)} (\varphi_L^{n+1} - \varphi_{K,L}^{n+1}) \big),$$

and

$$T_7^{(m)} = \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} (u_K^{n+1} - u_L^{n+1})(F_{K,L} - \widehat{F}_{K,L}) \varphi_{K,L}^{n+1}.$$

We may write

$$\begin{aligned} |T_6^{(m)}| &\leq C_\varphi \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} |u_K^{n+1} - u_L^{n+1}| \ \max(h_K, h_L)(\widehat{F}_{K,L}^{(+)} + \widehat{F}_{L,K}^{(+)}) \\ &\leq C_\varphi \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} |u_K^{n+1} - u_L^{n+1}| \ \max(h_K, h_L) \widetilde{F}_{K,L}. \end{aligned}$$

Turning to the study of $T_7^{(m)}$, we get $T_7^{(m)} = T_8^{(m)} - T_9^{(m)}$ with

$$T_8^{(m)} = \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} (u_K^{n+1} - u_L^{n+1}) F_{K,L} \varphi_{K,L}^{n+1},$$

and

$$T_9^{(m)} = \sum_{n=0}^{N-1} \tau^n \sum_{\{K,L\} \subset \mathcal{M}} (u_K^{n+1} - u_L^{n+1}) \widehat{F}_{K,L} \varphi_{K,L}^{n+1}.$$

Remarking that, for $P \in \widehat{\mathcal{P}}_{K,L}$, we have

$$\sum_{(I,J) \in P} (u_I^{n+1} - u_J^{n+1}) = (u_K^{n+1} - u_L^{n+1}),$$

and that

$$\sum_{P \in \widehat{\mathcal{P}}_{K,L}} F_{K,L}^P = F_{K,L},$$

we get that

$$T_8^{(m)} = \frac{1}{2} \sum_{n=0}^{N-1} \tau^n \sum_{(K,L) \in S} \sum_{P \in \widehat{\mathcal{P}}_{K,L}} \sum_{(I,J) \in P} (u_I^{n+1} - u_J^{n+1}) \varphi_{K,L}^{n+1} \ F_{K,L}^P.$$

Besides, we have

$$\begin{aligned} T_9^{(m)} &= \frac{1}{2}\sum_{n=0}^{N-1}\tau^n \sum_{(I,J)\in\widehat{S}} (u_I^{n+1}-u_J^{n+1})\widehat{F}_{I,J}\varphi_{I,J}^{n+1} \\ &= \frac{1}{2}\sum_{n=0}^{N-1}\tau^n \sum_{(I,J)\in\widehat{S}} (u_I^{n+1}-u_J^{n+1}) \sum_{(K,L)\in S}\ \sum_{P\in\widehat{\mathcal{P}}_{K,L}} \xi_{I,J}^P F_{K,L}^P \varphi_{I,J}^{n+1}, \end{aligned}$$

which leads, thanks to $\xi_{I,J}^P = 1$ if $(I,J)\in P$ else $\xi_{I,J}^P = 0$, to

$$T_9^{(m)} = \frac{1}{2}\sum_{n=0}^{N-1}\tau^n \sum_{(K,L)\in S}\ \sum_{P\in\widehat{\mathcal{P}}_{K,L}}\ \sum_{(I,J)\in P} (u_I^{n+1}-u_J^{n+1})\varphi_{I,J}^{n+1}\ F_{K,L}^P.$$

Hence

$$T_7^{(m)} = \frac{1}{2}\sum_{n=0}^{N-1}\tau^n \sum_{(K,L)\in S}\ \sum_{P\in\widehat{\mathcal{P}}_{K,L}}\ \sum_{(I,J)\in P} (u_I^{n+1}-u_J^{n+1})(\varphi_{I,J}^{n+1}-\varphi_{K,L}^{n+1})\ F_{K,L}^P.$$

We have $|\varphi_{I,J}^{n+1}-\varphi_{K,L}^{n+1}| \le C_\varphi \theta_{\widehat{\mathcal{P}}} \max(h_I,h_J)$ thanks to the definition (7) of $\theta_{\widehat{\mathcal{P}}}$. Therefore we get

$$|T_7^{(m)}| \le \frac{C_\varphi}{2}\theta_{\widehat{\mathcal{P}}}\sum_{n=0}^{N-1}\tau^n \sum_{(K,L)\in S}\ \sum_{P\in\widehat{\mathcal{P}}_{K,L}}\ \sum_{(I,J)\in P} |u_I^{n+1}-u_J^{n+1}|\ \max(h_I,h_J)|F_{K,L}^P|,$$

which may also be rewritten as

$$|T_7^{(m)}| \le \frac{C_\varphi}{2}\theta_{\widehat{\mathcal{P}}}\sum_{n=0}^{N-1}\tau^n \sum_{(K,L)\in\widehat{S}} |u_K^{n+1}-u_L^{n+1}|\max(h_K,h_L)\widetilde{F}_{K,L}.$$

Hence we get, setting $C_1 = C_\varphi + C_\varphi\theta_{\widehat{\mathcal{P}}}$

$$|T_2^{(m)} - T_5^{(m)}| \le C_1\sum_{n=0}^{N-1}\tau^n \sum_{\{K,L\}\subset\mathcal{M}} |u_K^{n+1}-u_L^{n+1}|\max(h_K,h_L)\widetilde{F}_{K,L}.$$

Thanks to the Cauchy-Schwarz inequality and defining T_{10} by

$$T_{10}^{(m)} = \sum_{n=0}^{N-1}\tau^n \sum_{\{K,L\}\subset\mathcal{M}} \max(h_K,h_L)\widetilde{F}_{K,L},$$

we have, thanks to Lemma 4.3,

$$(T_2^{(m)}-T_5^{(m)})^2 \le C_1^2 T_{10}^{(m)}\left(h_m\sum_{n=0}^{N-1}\tau^n \sum_{\{K,L\}\subset\mathcal{M}} (u_K^{n+1}-u_L^{n+1})^2\widetilde{F}_{K,L}\right) \le C_1^2 T_{10}^{(m)} h_m \frac{C_{\rm BV}}{\nu}.$$

It now suffices to show that $T_{10}^{(m)}$ remains bounded. Using (20), we have

$$T_{10}^{(m)} \le \theta_{\widehat{\mathcal{P}}}^2 \sum_{\{K,L\}\subset\mathcal{M}} \max(h_K,h_L)|F_{K,L}|.$$

We then remark that the term $\sum_{(K,L)\in S}\max(h_K,h_L)|F_{K,L}|$ remains bounded thanks to (39) and to the bound $\theta_{\mathcal{D}}|\Omega|\|\boldsymbol{v}\|_\infty$ on $\sum_{(K,L)\in S}\max(h_K,h_L)|\int_{\sigma_{K,L}}\boldsymbol{v}\cdot\boldsymbol{n}_{K,L}\mathrm{d}s|$. This completes the proof that

$$\lim_{m\to\infty}(T_2^{(m)}-T_5^{(m)}) = 0.$$

and therefore that

$$\lim_{m\to+\infty} T_2^{(m)} = T_3.$$

We have then proved that u satisfies (28), which concludes the proof of convergence of the scheme. □

References

[1] Aavatsmark, I., Eigestad, G.T.: Numerical Convergence of the MPFA O-method and U-method for General Quadrilateral Grids. *Int. J. Numer. Meth. Fluids* **51**, 939–961 (2006)

[2] Aavatsmark, I., Eigestad, G.T., Heimsund, B.-O., Mallison, B.T., Nordbotten, J.M., Oian, E.: A New Finite-Volume Approach to Efficient Discretization on Challenging Grids. *SPE J.* **15(3)**, 658–669 (2010)

[3] Agelas, L., Guichard, C., Masson, R.: Convergence of Finite Volume MPFA O-type Schemes for Heterogeneous Anisotropic Diffusion Problems. *IJVF* **7(2)**, (2010)

[4] Aziz, K., Ramesh, A.B., Woo, P.T.: Fourth SPE Comparative Solution Project: Comparison of Steam Injection Simulators. *J. Pet. Tech.* **39**, 1576–1584 (1987)

[5] Corre, B., Eymard, R., Quettier, L.: Applications of a Thermal Simulator to Field Cases, *SPE ATCE*, (1984)

[6] Dawson, C., Sun, S., Wheeler, M.F.: Compatible Algorithms for Coupled Flow and Transport. *Comput. Meth. Appl. Mech. Eng.* **193**, 2565–2580 (2004)

[7] Eymard, R., Gallouët, T.: Schémas à neuf points pour équations de transport. *in R. Eymard, Techniques numériques de simulation d'écoulements polyphasiques en milieu poreux : applications à des cas industriels, thèse de l'université de Chambéry, France* (1987).

[8] Eymard, R., Gallouët, T., Herbin, R.: The Finite Volume Method. *Handbook of Numerical Analysis, Ph. Ciarlet J.L. Lions eds*, **7**, 715–1022 (2000).

[9] Eymard, R., Sonier, F.: Mathematical and Numerical Properties of Control-Volume Finite-Element Scheme for Reservoir Simulation. *SPE Reservoir Eng.* **9**, 283–289 (1994)

[10] Keilegavlen, E., Kozdon, J., Mallison, B.T.: Monotone Multi-dimensional Upstream Weighting on General Grids. *Proc. ECMOR XII*, Oxford, (2010)

[11] Lipnikov, K., Moulton, J.D., Svyatskiy, D.: A Multilevel Multiscale Mimetic (M3) Method for Two-phase Flows in Porous Media. *J. Comput. Phys.* **14**, 6727–6753 (2008)

[12] D.K. Ponting. Corner Point Geometry in Reservoir Simulation. *In Clarendon Press, editor*, *Proc. ECMOR I*, 45–65, Cambridge, (1989)

[13] Shubin, G.R., Bell, J.B.: An analysis of the grid orientation effect in numerical simulation of miscible displacement. *Comput. Methods Appl. Mech. Eng.* **47**, 47–71 (1984).

[14] Vinsome, P., Au, A.: One Approach to the Grid Orientation Problem in Reservoir Simulation. *Old SPE J.* **21**, 160–161 (1981)

[15] Yanosik, J.L., McCracken, T.A.: A Nine-point, Finite-Difference Reservoir Simulator for Realistic Prediction of Adverse Mobility Ratio Displacements. *Old SPE J.* **19**, 253–262 (1979)