Design of coupled finite volume schemes minimizing the grid orientation effect in reservoir simulation

Karine Laurent, Eric Flauraud, Christophe Preux, Quang Huy Tran, Christophe Berthon

Design of coupled finite volume schemes minimizing the grid orientation effect in reservoir simulation

Karine Laurent, Éric Flauraud, Christophe Preux, Quang Huy Tran*
IFP Energies nouvelles
1 et 4 avenue de Bois-Préau, 92852 Rueil-Malmaison Cedex, France

AND

Christophe Berthon
Université de Nantes, Laboratoire Jean Leray
2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 3, France

November 30, 2019

Abstract

In this paper, we propose an analysis method for the so-called *grid orientation effect* (GOE) in the numerical simulation of two-phase flows in porous media. The GOE, which occurs when using coupled finite volume schemes on structured grids, is well known to engineers. Several attempts, most of which are of empirical nature, have been put forward in order to alleviate this undesirable phenomenon. Here, our approach relies on a more rigorous notion of angular error for all directions, which in turn enables us —*via* integration and minimization— to single out the "least anisotropic" scheme within a given family of schemes depending on some tuning parameter(s). Numerical test problems testify to the improvement brought by the new construction. grid orientation effect; reservoir simulation; finite volume schemes; nine-point scheme.

1 Introduction

In oil reservoir simulation, engineers are often faced with a phenomenon called *grid orientation effect* (GOE). This unpleasant effect arises when coupled finite volume schemes are used on structured grids in order to simulate the thrust of a viscous fluid (heavy oil) by a less viscous one (water), which is typical of an injection scenario for enhanced oil recovery. The GOE gives rise to a more or less marked distortion of the computed solution whereas, in particular, the exact solution is radial, as illustrated in Figure 1. As a consequence, the simulation of predicted production of a well also depends on the grid orientation and may not be accurate.

1.1 A simplified model of two-phase flow in porous media

We first present the model under consideration in this paper, which is a simplified version of the *isotherm Dead Oil* [17] corresponding to an oil and water mixture without capillary pressure and gravity. Let $\Omega \subset \mathbb{R}^2$ be a bounded open connected domain with a regular boundary. The two-phase flow is characterized by the common pressure $p(\boldsymbol{x}, t) > 0$ and the water saturation $s(\boldsymbol{x}, t) \in [0, 1]$, where $\boldsymbol{x} = (x, y) \in \Omega$ and $t \geq 0$ are respectively space and time variables. These quantities of

*Corresponding author. Email: quang-huy.tran@ifpen.fr

interest solve

$$\boldsymbol{u} = -\kappa\lambda(s)\boldsymbol{\nabla} p, \tag{1.1a}$$

$$\operatorname{div}(\boldsymbol{u}) = q, \tag{1.1b}$$

$$\phi\partial_t s + \operatorname{div}(f(s)\boldsymbol{u}) = q_w, \tag{1.1c}$$

where the total velocity $\boldsymbol{u}(\boldsymbol{x}, t)$ is given by the Darcy-[24] law (1.1a), and

$$\lambda(s) = \frac{\kappa_{r,w}(s)}{\mu_w} + \frac{\kappa_{r,o}(1-s)}{\mu_o} \tag{1.2}$$

is the total mobility. From now on, equation (1.1b) is referred to as the pressure equation, since it gives $-\operatorname{div}(\kappa\lambda(s)\boldsymbol{\nabla} p) = q$ when combined with (1.1a). The symbol κ stands for the permeability tensor, restricted here to be a scalar. The water relative permeability $\kappa_{r,w}(s)$ is an increasing function of s, while the oil relative permeability $\kappa_{r,o}(1-s)$ is a decreasing function of s. Moreover, the two scalars $\mu_w > 0$ and $\mu_o > 0$ denote the water and oil viscosities. The quantity $\phi(\boldsymbol{x}) \in [0, 1]$ represents the (known) porosity of the medium. Without loss of generality, we impose $\phi \equiv 1$ in the present work.

The water fractional flow $f(s)$ in (1.1c) is defined as

$$f(s) = \frac{\kappa_{r,w}(s)/\mu_w}{\kappa_{r,w}(s)/\mu_w + \kappa_{r,o}(1-s)/\mu_o}, \tag{1.3}$$

where we have set $\kappa_{r,w}(s) = \kappa_{r,w}^\sharp\, \kappa_{r,w}^*(s)$ and $\kappa_{r,o}(1-s) = \kappa_{r,o}^\sharp\, \kappa_{r,o}^*(1-s)$. The normalized relative permeabilities $\kappa_{r,w}^*(s)$ and $\kappa_{r,o}^*(1-s)$ are assumed to be in $[0, 1]$, while $\kappa_{r,w}^\sharp$ and $\kappa_{r,o}^\sharp$ are given dimensionless constants. Examples of explicit values for $\kappa_{r,w}(s)$ and $\kappa_{r,o}(1-s)$ can be found in [5]. The water fractional flow f is a smooth positive and non-decreasing function of s, i.e., $f \geq 0$ and $f' \geq 0$ for $s \in [0, 1]$. It can be put under the reduced form

$$f(s) = \frac{M\kappa_{r,w}^*(s)}{M\kappa_{r,w}^*(s) + \kappa_{r,o}^*(1-s)}, \qquad \text{where} \qquad M = \frac{\mu_o \kappa_{r,w}^\sharp}{\mu_w \kappa_{r,o}^\sharp} \tag{1.4}$$

is the mobility ratio between the displacing water and the displaced oil. It can be shown [7] that M measures, in some sense, the stiffness of the problem. Indeed, as soon as M is larger than some critical threshold, the system (1.1) turns out to be unstable and thus amplifies the numerical errors. In such a context, the errors due to the GOE may become prevailing.

In the right-hand sides in (1.1), the quantities q and q_w are source terms expressing the produced or injected total and water flow in the domain. Equipped with appropriate boundary and initial conditions, model (1.1) is usually discretized in time by the IMPES strategy [1].The pressure p is first solved implicitly by some finite volume discretization in space of the pressure equation (1.1b). Next, the saturation s is then updated explicitly by some finite volume discretization in space of the saturation equation (1.1c).

1.2 Review of literature on the GOE

Over structured grids, the simplest scheme for the pressure equation (1.1b) is the so-called *five-point scheme*, which we abbreviate to 5P. In the finite-volume world [14], the 5P scheme is also known as the TPFA (Two-Point Flux Approximation) scheme. For a Cartesian mesh, [1] and [32] demonstrated that the GOE of the five-point scheme dominates the numerical solution of (1.1) under adverse the mobility ratio, i.e., when M is above some critical threshold. Such failure is displayed in the right panel of Figure 1, where we clearly see that the injected fluid is in advance along the axes of the grid but is late along the diagonals of the grid. Moreover, refining the mesh does not significantly reduce the GOE [4].

In an attempt to alleviate the GOE, [33] advocated a *nine-point* (9P) scheme obtained by superimposing two 5P schemes associated with two square grids rotated by $\pi/4$ relative to each

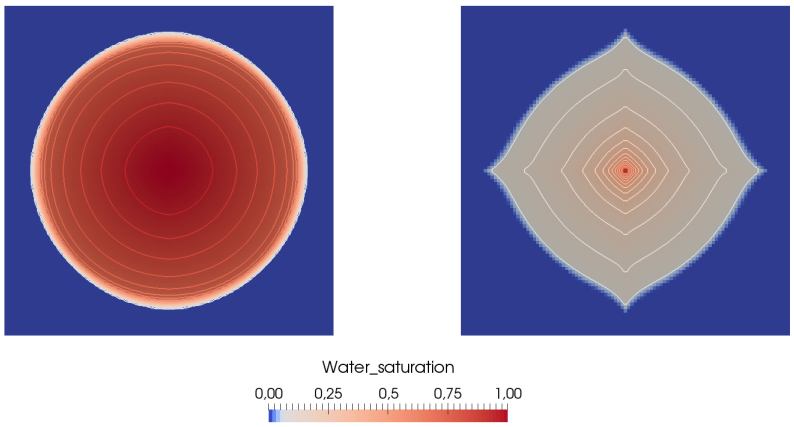


Figure 1: Grid orientation effect. Left: $M = 0.8$; right: $M = 200$.

other. By involving diagonal neighbors into the stencil, the resulting scheme significantly reduces the GOE over square meshes and met an instant success. Two generalizations of the Yanosik-McCracken scheme to rectangular meshes were then proposed by [29] and by [8]. The difference between these two versions lies in the weighting heuristic for the diagonal cells. For this weighting, [13] put forward a more rigorous error analysis leading to a new 9P scheme. Since then, the 9P philosophy has been extended to other porous two-phase models, for example to account for dispersion [19, 30]. The objectionable aspect of these works is that the error analysis —whenever available— is only concerned with the pressure, while the quantity of interest is the saturation. Improving on a previous work by [11] and relying on an analysis of the saturation equation, [15] designed another 9P scheme over square meshes. This methodology is more satisfactory from the theoretical standpoint. However, since the basic idea is to request that the diffusion matrix of the equivalent equation be invariant by a $\pi/4$-rotation, the extension to rectangular meshes does not seem obvious.

In the above-mentioned approaches, the numerical fluxes of the pressure equation (1.1b) are first altered (in structure and values) by taking diagonal cells into the pressure stencil. The modification of the numerical fluxes for the saturation stencil follow suit as an automatic consequence of normal upwinding (see §2 for more details). A natural alternative, investigated by [20, 21, 22] is to focus on more sophisticated discretizations of the saturation equation. This brings out a lot of connections with "genuinely" multidimensional transport schemes for linear advection [9, 28, 12, 2]. Unfortunately, multidimensional advection schemes need exact or highly accurate velocity fields, which cannot be achieved if no effort is dedicated to the pressure equation.

1.3 Objectives and outline of this paper

To our knowledge, the work by [15] —along with [31] for miscible flows— is the first contribution to the GOE issue in which the saturation equation plays a major role and in which the idea of a "good" parameter is highlighted. In the present work, we wish to carry out a mathematical analysis over rectangular meshes of various coupled finite volume schemes for (1.1) where a few degrees of freedom are available. Our ultimate goal is to define the "best" choice that would minimize the GOE in a quantitative sense to be clarified.

In §2, we consider two families of schemes for (1.1) containing tuning parameters. The first one, defined in §2.1 and called 9P1s, has a scalar tuning parameter θ that allows several "historical" schemes such as [33, 29, 8, 13] to appear as special cases of a unified framework. The second one, defined in §2.2 and called 9P2s, has two scalar tuning parameters $\boldsymbol{\theta} = (\theta_x, \theta_y)$, a novelty that we introduce in order to further reduce the GOE.

In §3, we tackle the problem of optimizing these tuning parameters. The same strategy, first laid out in 3.1 for simplicity, is applied to the 9P1s family in 3.2 and to the 9P2s family in 3.3. By

resorting to Taylor expansion and/or Fourier analysis under simplifying assumptions, we succeed in assigning a measure of the angular error to each direction in space. Then, by minimizing the integrated squared difference between this angular error and some ideal behaviour, we are in a position to determine the optimal parameters for each scheme. These optimal values coincide with some formerly proposed values in the literature. Finally, numerical experiments in §4 corroborate our theoretical developments on two test problems.

2 Coupled finite volume schemes with tuning parameters

System (1.1) is usually discretized in time using the IMPES technique [1] where the pressure p is solved implicitly in a first step and the saturation s is solved explicitly (at least for the convection part) in a second step. Using a semi-discrete formulation, the IMPES scheme reads

$$\boldsymbol{u}^{n+1} = -\kappa\lambda(s^n)\boldsymbol{\nabla} p^{n+1}, \tag{2.1a}$$

$$\mathrm{div}(\boldsymbol{u}^{n+1}) = q^{n+1}, \tag{2.1b}$$

$$\Delta t^{-1}(s^{n+1} - s^n) + \mathrm{div}(f(s^n)\boldsymbol{u}^{n+1}) = q_w^{n+1}, \tag{2.1c}$$

where the time-step $\Delta t > 0$ must be restricted by a CFL-like condition [27].

Regarding the discretization in space of the two divergence operators in (2.1), there are two finite volume schemes, one for the pressure equation (2.1b) and another one for the saturation equation (2.1c). The latter is deduced to the former by normal upwinding. In this section, we describe two discretizations in space, namely: (i) in §2.1.2, the 9P1s scheme which makes use of one scalar parameter; (ii) in §2.2.2, the 9P2s scheme which makes use of two scalar parameters. For each method, we first present the discretization of the pressure equation (2.1a)–(2.1b) before exposing the discretization of the saturation equation (2.1c).

The domain Ω is divided into uniform rectangular cells

$$K_{i,j} = (x_{i-1/2}, x_{i+1/2}) \times (y_{j-1/2}, y_{j+1/2})$$

of side lengths $(x_{i+1/2} - x_{i-1/2}, y_{j+1/2} - y_{j-1/2}) = (\Delta x, \Delta y) \in (\mathbb{R}_*^+)^2$. We denote by $\boldsymbol{x}_{i,j} = (x_i, y_j)$ the center of the cell $K_{i,j}$. We restrict ourselves to rectangular meshes since they are widely used in most reservoir simulation software.

2.1 The 9P1s scheme

The 9P1s family includes several classic schemes in a unified formulation.

2.1.1 9P1s for pressure

Let us first assume that the coefficient of $\boldsymbol{\nabla} p$ is uniform in space, that is,

$$\kappa\lambda(s) \equiv 1. \tag{2.2}$$

The semi-discretized pressure equation (2.1a)–(2.1b) then boils down to $-\Delta p = q$, where we have omitted the superscript $n+1$ for the sake of clarity. Our objective is to combine the 1-D discrete Laplace operators per direction

$$(-\Delta_h^x p)_{i,j} = \frac{-p_{i-1,j} + 2p_{i,j} - p_{i+1,j}}{\Delta x^2}, \qquad (-\Delta_h^y p)_{i,j} = \frac{-p_{i,j-1} + 2p_{i,j} - p_{i,j+1}}{\Delta y^2}, \tag{2.3}$$

into a 2-D discrete Laplace operator with a "more isotropic" behavior. The combination we consider is

$$\begin{aligned}(-\Delta_h^\theta p)_{i,j} = {} & \theta(-\Delta_h^x p)_{i,j+1} + (1-2\theta)(-\Delta_h^x p)_{i,j} + \theta(-\Delta_h^x p)_{i,j-1} \\ & + \theta(-\Delta_h^y p)_{i+1,j} + (1-2\theta)(-\Delta_h^y p)_{i,j} + \theta(-\Delta_h^y p)_{i-1,j},\end{aligned} \tag{2.4}$$

where the tuning parameter θ is restricted to $[0, 1/2]$ in order to ensure that each directional combination is convex. The fully expanded stencil of $-\Delta_h^\theta$ reads

$$\begin{aligned}(-\Delta_h^\theta p)_{i,j} = &- \alpha p_{i-1,j+1} - \beta_y p_{i,j+1} - \alpha p_{i+1,j} \\ &- \beta_x p_{i-1,j} + (4\alpha + 2\beta_x + 2\beta_y) p_{i,j} + \beta_x p_{i+1,j} \\ &- \alpha p_{i-1,j-1} - \beta_y p_{i,j-1} - \alpha p_{i+1,j-1},\end{aligned}$$

with

$$\alpha = \theta\left(\frac{1}{\Delta x^2} + \frac{1}{\Delta y^2}\right), \qquad \beta_x = \frac{2\theta}{\Delta y^2} - \frac{1-2\theta}{\Delta x^2}, \qquad \beta_y = \frac{2\theta}{\Delta x^2} - \frac{1-2\theta}{\Delta y^2}.$$

For $\theta = 0$, $-\Delta_h^\theta$ degenerates to the standard 5P scheme, also known as TPFA (Two-Point Flux Approximation) in the finite volume world. For $\theta = 1/6$, $-\Delta_h^\theta$ can be derived from the Q_1 finite element method on the dual rectangular mesh. For $\theta = 1/12$, $-\Delta_h^\theta$ coincides with the Yanosik-Ding 9P scheme [33, 13], although these authors do not present it in this way. The parameter θ is *not* aimed at increasing the order of accuracy for the approximation. Rather, it is aimed at changing the spatial distribution of error, as shown by the forthcoming statement.

Theorem 2.1. *If p is a smooth function of $\boldsymbol{x}$ and if $\Delta x, \Delta y$ are small enough, then*

$$\begin{aligned}(-\Delta_h^\theta p)_{i,j} = (-\Delta p)(\boldsymbol{x}_{i,j}) - \left[\tfrac{1}{12}\Delta x^2\, \partial_{xxxx} p + \tfrac{1}{12}\Delta y^2\, \partial_{yyyy} p(\boldsymbol{x}_{i,j}) + \theta(\Delta x^2 + \Delta y^2)\, \partial_{xxyy} p\right](\boldsymbol{x}_{i,j}) \\ + O(\Delta x^4) + O(\Delta y^4) + O(\Delta x^2 \Delta y^2). \qquad (2.5)\end{aligned}$$

Proof. Starting from the basic 1-D properties

$$\begin{aligned}(-\Delta_h^x p)_{i,j} &= -\partial_{xx}^2 p(\boldsymbol{x}_{i,j}) - \tfrac{1}{12}\Delta x^2\, \partial_{xxxx} p(\boldsymbol{x}_{i,j}) + O(\Delta x^4), \\ (-\Delta_h^y p)_{i,j} &= -\partial_{yy}^2 p(\boldsymbol{x}_{i,j}) - \tfrac{1}{12}\Delta y^2\, \partial_{yyyy} p(\boldsymbol{x}_{i,j}) + O(\Delta y^4),\end{aligned}$$

we carry out Taylor expansions around $\boldsymbol{x}_{i,j}$ by brute force and the proof is completed. □

For a square mesh ($\Delta x = \Delta y = h$), Theorem 2.1 implies

$$(-\Delta_h^\theta p)_{i,j} = (-\Delta p)(\boldsymbol{x}_{i,j}) - \tfrac{1}{12}h^2 \left[\partial_{xxxx} p + \partial_{yyyy} p + 24\theta\, \partial_{xxyy} p\right](\boldsymbol{x}_{i,j}) + O(h^4).$$

Therefore, as soon as $\theta = 1/12$,

$$(-\Delta_h^\theta p)_{i,j} = (-\Delta p)(\boldsymbol{x}_{i,j}) - \tfrac{1}{12}h^2\, \Delta\Delta p(\boldsymbol{x}_{i,j}) + O(h^4). \qquad (2.6)$$

If p is radial, its bi-Laplacian $\Delta\Delta p$ is also radial. It follows from (2.6) that the error between $-\Delta_h^\theta p$ and $-\Delta p$ is then radial, which reflects the desired isotropic behavior. [33] and [13] did not use the same argument but arrived at the same scheme. In §3.1, we will demonstrate that even for a rectangular mesh ($\Delta x \neq \Delta y$), the "optimal" parameter remains $\theta = 1/12$ in a sense that will be made rigorous.

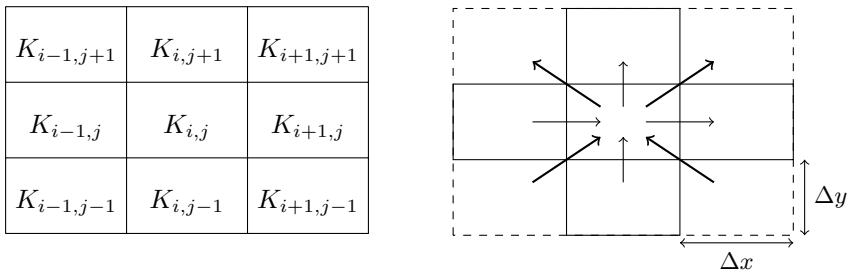


Figure 2: Nine-point stencil (left) and orientation of numerical fluxes (right).

In order to extend the discretization to the general case $\kappa\lambda(s) \not\equiv 1$, let us reformulate $-\Delta_h^\theta$ in the finite volume language. Multiplying the stencil (2.4) by the measure $\Delta x\Delta y$ of a cell and rearranging the right-hand side, we obtain the discrete flux balance

$$\begin{aligned}\Delta x\Delta y(-\Delta_h^\theta p)_{i,j} &= F^\theta_{i+1/2,j} - F^\theta_{i-1/2,j} + F^\theta_{i,j+1/2} - F^\theta_{i,j-1/2} \\ &\quad + F^{\theta\nearrow}_{i+1/2,j+1/2} - F^{\theta\nearrow}_{i-1/2,j-1/2} + F^{\theta\nwarrow}_{i-1/2,j+1/2} - F^{\theta\nwarrow}_{i+1/2,j-1/2},\end{aligned} \tag{2.7}$$

where we have set

$$F^\theta_{i+1/2,j} = \quad [z - 2\theta(z+z^{-1})](p_{i,j} - p_{i+1,j}), \tag{2.8a}$$

$$F^\theta_{i,j+1/2} = [z^{-1} - 2\theta(z+z^{-1})](p_{i,j} - p_{i,j+1}), \tag{2.8b}$$

$$F^{\theta\nearrow}_{i+1/2,j+1/2} = \quad \theta(z+z^{-1})\,(p_{i,j} - p_{i+1,j+1}), \tag{2.8c}$$

$$F^{\theta\nwarrow}_{i-1/2,j+1/2} = \quad \theta(z+z^{-1})\,(p_{i,j} - p_{i-1,j+1}), \tag{2.8d}$$

using the ratio between the mesh sizes

$$z = \frac{\Delta y}{\Delta x}. \tag{2.9}$$

The selected orientation of the eight numerical fluxes involved in (2.7) is displayed in Figure 2. The arrows $\nearrow$ and $\nwarrow$ indicate the direction in which the flux takes a positive value.

The reformulation (2.7)–(2.8) naturally suggests the scheme

$$\begin{aligned}&F^\theta_{i+1/2,j} - F^\theta_{i-1/2,j} + F^\theta_{i,j+1/2} - F^\theta_{i,j-1/2} \\ &\qquad + F^{\theta\nearrow}_{i+1/2,j+1/2} - F^{\theta\nearrow}_{i-1/2,j-1/2} + F^{\theta\nwarrow}_{i-1/2,j+1/2} - F^{\theta\nwarrow}_{i+1/2,j-1/2} = \Delta x\Delta y\, q_{i,j}\end{aligned} \tag{2.10}$$

for the pressure equation (2.1b) in the general case $\kappa\lambda(s) \not\equiv 1$, where the numerical fluxes

$$F^\theta_{i+1/2,j} = \quad \widetilde{\kappa\lambda}(s^n_{i,j}, s^n_{i+1,j})[z - 2\theta(z+z^{-1})](p^{n+1}_{i,j} - p^{n+1}_{i+1,j}), \tag{2.11a}$$

$$F^\theta_{i,j+1/2} = \widetilde{\kappa\lambda}(s^n_{i,j}, s^n_{i,j+1})[z^{-1} - 2\theta(z+z^{-1})](p^{n+1}_{i,j} - p^{n+1}_{i,j+1}), \tag{2.11b}$$

$$F^{\theta\nearrow}_{i+1/2,j+1/2} = \quad \widetilde{\kappa\lambda}(s^n_{i,j}, s^n_{i+1,j+1})\theta(z+z^{-1})\,(p^{n+1}_{i,j} - p^{n+1}_{i+1,j+1}), \tag{2.11c}$$

$$F^{\theta\nwarrow}_{i-1/2,j+1/2} = \quad \widetilde{\kappa\lambda}(s^n_{i,j}, s^n_{i-1,j+1})\theta(z+z^{-1})\,(p^{n+1}_{i,j} - p^{n+1}_{i-1,j+1}), \tag{2.11d}$$

are now defined by the harmonic mean $\widetilde{\lambda}(s_L, s_R) = 2\lambda(s_L)\lambda(s_R)/[\lambda(s_L) + \lambda(s_R)]$ [14].

Since the factors $z - 2\theta(z+z^{-1})$ and $z^{-1} - 2\theta(z+z^{-1})$ appear in the fluxes (2.8a)–(2.8b), it is advisable to impose the restriction

$$0 \le \theta \le \frac{\min(1, z^2)}{1+z^2} =: \theta_M, \tag{2.12}$$

so that the 9P1s horizontal and vertical fluxes (2.8a)–(2.8b) have the same signs as their 5P counterparts $z(p_{i,j} - p_{i+1,j})$ and $z^{-1}(p_{i,j} - p_{i,j+1})$.

2.1.2 9P1s for saturation

Once the pressure field is computed, the saturation equation (2.1c) can be discretized with a scheme having a similar nine-point and eight-flux structure. More specifically,

$$\begin{aligned}&\Delta x\Delta y\Delta t^{-1}(s^{n+1}_{i,j} - s^n_{i,j}) + G^\theta_{i+1/2,j} - G^\theta_{i-1/2,j} + G^\theta_{i,j+1/2} - G^\theta_{i,j-1/2} \\ &\qquad + G^{\theta\nearrow}_{i+1/2,j+1/2} - G^{\theta\nearrow}_{i-1/2,j-1/2} + G^{\theta\nwarrow}_{i-1/2,j+1/2} - G^{\theta\nwarrow}_{i+1/2,j-1/2} = \Delta x\Delta y\, q_{w;i,j},\end{aligned} \tag{2.13}$$

where the fluxes are upwinded according to [16] as

$$G^\theta_{i+1/2,j} = f(s^n_{i,j})[F^\theta_{i+1/2,j}]^+ + f(s^n_{i+1,j})[F^\theta_{i+1/2,j}]^-, \tag{2.14a}$$

$$G^\theta_{i,j+1/2} = f(s^n_{i,j})[F^\theta_{i,j+1/2}]^+ + f(s^n_{i,j+1})[F^\theta_{i,j+1/2}]^-, \tag{2.14b}$$

$$G^{\theta\nearrow}_{i+1/2,j+1/2} = f(s^n_{i,j})[F^{\theta\nearrow}_{i+1/2,j+1/2}]^+ + f(s^n_{i+1,j+1})[F^{\theta\nearrow}_{i+1/2,j+1/2}]^-, \tag{2.14c}$$

$$G^{\theta\nwarrow}_{i-1/2,j+1/2} = f(s^n_{i,j})[F^{\theta\nwarrow}_{i-1/2,j+1/2}]^+ + f(s^n_{i-1,j+1})[F^{\theta\nwarrow}_{i-1/2,j+1/2}]^-, \tag{2.14d}$$

where $[F]^+ = \max(F,0)$ and $[F]^- = \min(F,0)$ are respectively the positive and negative parts of F. The term $q_{w;i,j}$ expresses the source term which is set to zero from now on.

The scheme (2.13)–(2.14) must be supplemented by a CFL-like condition so as to guarantee the maximum principle for the saturation, at least in regions where both source terms vanish. For this purpose, let us introduce

$$\sigma(s_L,s_R) = \begin{cases} f'(s_L) & \text{if } s_L = s_R, \\ \frac{f(s_R)-f(s_L)}{s_R-s_L} & \text{otherwise;} \end{cases}$$

and the quantities

$$\sigma_{i+1/2,j} = \sigma(s^n_{i,j}, s^n_{i+1,j}), \qquad \sigma_{i+1/2,j+1/2} = \sigma(s^n_{i,j}, s^n_{i+1,j+1}),$$
$$\sigma_{i,j+1/2} = \sigma(s^n_{i,j}, s^n_{i,j+1}), \qquad \sigma_{i-1/2,j+1/2} = \sigma(s^n_{i,j}, s^n_{i-1,j+1}),$$

which are all non-negative since f is an non-decreasing function. For each cell $K_{i,j}$, let

$$\begin{aligned} \rangle\sigma F^\theta\langle_{i,j} = &- \sigma_{i+1/2,j}[F^\theta_{i+1/2,j}]^- + \sigma_{i-1/2,j}[F^\theta_{i-1/2,j}]^+ - \sigma_{i,j+1/2}[F^\theta_{i,j+1/2}]^- + \sigma_{i,j-1/2}[F^\theta_{i,j-1/2}]^+ \\ &- \sigma_{i+1/2,j+1/2}[F^{\theta\nearrow}_{i+1/2,j+1/2}]^- + \sigma_{i-1/2,j-1/2}[F^{\theta\nearrow}_{i-1/2,j-1/2}]^+ \\ &- \sigma_{i-1/2,j+1/2}[F^{\theta\nwarrow}_{i-1/2,j+1/2}]^- + \sigma_{i+1/2,j-1/2}[F^{\theta\nwarrow}_{i+1/2,j-1/2}]^+ \end{aligned}$$

be its total incoming flux.

Proposition 2.1. *If $q_{i,j} = q_{w;i,j} = 0$ at some cell $K_{i,j}$ and*

$$\frac{\Delta t}{\Delta x \Delta y} \rangle\sigma F^\theta\langle_{i,j} \le 1, \tag{2.15}$$

then $s^{n+1}_{i,j}$ is a convex combination of $s^n_{i,j}$ and its eight neighboring saturations at time n.

Proof. Multiplying the pressure balance (2.10) by $f(s^n_{i,j})$, subtracting the product from the saturation balance (2.13), splitting $F = [F]^+ + [F]^-$ for each flux and involving the σ's, we manage to express $s^{n+1}_{i,j}$ as a combination of $s^n_{i,j}$ and its eight neighbours, the coefficients of which depend on the data. We refer the readers to [3] for more details. □

From this Proposition, we deduce the stability condition to be imposed as

$$\frac{\Delta t}{\Delta x \Delta y} \max_{i,j} \rangle\sigma F^\theta\langle_{i,j} \le 1. \tag{2.16}$$

We postpone the error analysis to §3.2, where we will see that the approximation in saturation remains of first-order with respect to $(\Delta x, \Delta y)$. The parameter θ does not improve the order of accuracy. It is simply aimed at reshaping the error distribution in space.

2.2 The 9P2s scheme

We wish to push further the generalization of 9P schemes by considering two tuning parameters instead of one. After all, since we have two privileged directions x, y, two grid-steps $\Delta x, \Delta y$, it seems natural to have θ_x, θ_y in the definition of the scheme. Besides, it is expected that having two degrees of freedom at our disposal will help us fight the GOE more efficiently. The difficulty, however, lies in preserving the finite-volume flux balances when introducing a second parameter.

2.2.1 9P2s for pressure

As in §2.1.1, let us start with the uniform case (2.2). To discretize pressure equation $-\Delta p = q$, we combine the 1-D discrete Laplace operators (2.3) into a 2-D discrete Laplace operator. The combination takes the form

$$\begin{aligned}(-\Delta_h^{\boldsymbol{\theta}} p)_{i,j} = {} & \theta_x(-\Delta_h^x p)_{i,j+1} + (1-2\theta_x)(-\Delta_h^x p)_{i,j} + \theta_x(-\Delta_h^x p)_{i,j-1} \\ & + \theta_y(-\Delta_h^y p)_{i+1,j} + (1-2\theta_y)(-\Delta_h^y p)_{i,j} + \theta_y(-\Delta_h^y p)_{i-1,j},\end{aligned} \tag{2.17}$$

where $\boldsymbol{\theta} = (\theta_x,\ \theta_y)$ is a pair of tuning parameters, ones per direction. The fully expanded stencil of $(-\Delta_h^{\boldsymbol{\theta}} p)$ reads

$$(-\Delta_h^{\boldsymbol{\theta}} p)_{i,j} = -\alpha p_{i-1,j+1} - \beta_y p_{i,j+1} - \alpha p_{i+1,j} \tag{2.18}$$

$$- \beta_x p_{i-1,j} + (4\alpha + 2\beta_x + 2\beta_y) p_{i,j} - \beta_x p_{i+1,j} \tag{2.19}$$

$$- \alpha p_{i-1,j-1} - \beta_y p_{i,j-1} - \alpha p_{i+1,j-1} \tag{2.20}$$

with

$$\alpha = \frac{\theta_x}{\Delta x^2} + \frac{\theta_y}{\Delta y^2}, \qquad \beta_x = \frac{2\theta_y}{\Delta y^2} - \frac{1-2\theta_x}{\Delta x^2}, \qquad \beta_y = \frac{2\theta_x}{\Delta x^2} - \frac{1-2\theta_y}{\Delta y^2}.$$

Theorem 2.2. *If p is a smooth function of $\boldsymbol{x}$ and if Δx, Δy are small enough, then*

$$\begin{aligned}(-\Delta_h^{\boldsymbol{\theta}} p)_{i,j} = (-\Delta p)(\boldsymbol{x}_{i,j}) & - \left[\tfrac{1}{12}\Delta x^2\, \partial_{xxxx} p + \tfrac{1}{12}\Delta y^2\, \partial_{yyyy} p + (\theta_x \Delta x^2 + \theta_y \Delta y^2)\, \partial_{xxyy} p\right](\boldsymbol{x}_{i,j}) \\ & + O(\Delta x^4) + O(\Delta y^4) + O(\Delta x^2 \Delta y^2).\end{aligned} \tag{2.21}$$

Proof. The proof follows along the same lines as in Theorem 2.1. □

At this stage, it appears that only the combination $\theta_x \Delta x^2 + \theta_y \Delta y^2$ matters for the second-order accuracy. Later we will prescribe other rules to determine θ_x and θ_x separately. For the moment, we observe that over a square mesh ($\Delta x = \Delta y = h$), the best choice is $\theta_x + \theta_y = 1/6$. Indeed, as argued in 2.1.1, the error is then $-\frac{1}{12} h^2 \Delta\Delta p$. If p is radial, then the bi-Laplacian $\Delta\Delta p$ is also radial, which ensures isotropy.

To deal with the variable coefficient case $\kappa\lambda(s) \not\equiv 1$, we first need to reformulate $-\Delta_h^{\boldsymbol{\theta}}$ as a finite volume scheme. Multiplying the stencil (2.17) by the measure $\Delta x \Delta y$ of a cell and reorganizing various terms, we end up with the flux balance

$$\begin{aligned}\Delta x \Delta y (-\Delta_h^{\boldsymbol{\theta}} p)_{i,j} = {} & \widetilde{F}^{\boldsymbol{\theta}}_{i+1/2,j} - \widetilde{F}^{\boldsymbol{\theta}}_{i-1/2,j} + \widetilde{F}^{\boldsymbol{\theta}}_{i,j+1/2} - \widetilde{F}^{\boldsymbol{\theta}}_{i,j-1/2} \\ & + \widetilde{F}^{\boldsymbol{\theta}\nearrow}_{i+1/2,j+1/2} - \widetilde{F}^{\boldsymbol{\theta}\nearrow}_{i-1/2,j-1/2} + \widetilde{F}^{\boldsymbol{\theta}\nwarrow}_{i-1/2,j+1/2} - \widetilde{F}^{\boldsymbol{\theta}\nwarrow}_{i+1/2,j-1/2},\end{aligned} \tag{2.22}$$

where

$$\widetilde{F}^{\boldsymbol{\theta}}_{i+1/2,j} = (1-4\theta_x) F_{i+1/2,j}, \qquad \widetilde{F}^{\boldsymbol{\theta}}_{i-1/2,j} = (1-4\theta_x) F_{i-1/2,j}, \tag{2.23a}$$

$$\widetilde{F}^{\boldsymbol{\theta}}_{i,j+1/2} = (1-4\theta_y) F_{i,j+1/2}, \qquad \widetilde{F}^{\boldsymbol{\theta}}_{i,j-1/2} = (1-4\theta_y) F_{i,j-1/2}, \tag{2.23b}$$

$$\widetilde{F}^{\boldsymbol{\theta}\nearrow}_{i+1/2,j+1/2} = \theta_y F_{i,j+1/2} + \theta_x F_{i+1/2,j+1} + \theta_x F_{i+1/2,j} + \theta_y F_{i+1,j+1/2}, \tag{2.23c}$$

$$\widetilde{F}^{\boldsymbol{\theta}\nearrow}_{i-1/2,j-1/2} = \theta_y F_{i-1,j-1/2} + \theta_x F_{i-1/2,j} + \theta_x F_{i-1/2,j-1} + \theta_y F_{i,j-1/2}, \tag{2.23d}$$

$$\widetilde{F}^{\boldsymbol{\theta}\nwarrow}_{i-1/2,j+1/2} = \theta_y F_{i,j+1/2} - \theta_x F_{i-1/2,j+1} - \theta_x F_{i-1/2,j} + \theta_y F_{i-1,j+1/2}, \tag{2.23e}$$

$$\widetilde{F}^{\boldsymbol{\theta}\nwarrow}_{i+1/2,j-1/2} = \theta_y F_{i+1,j-1/2} - \theta_x F_{i+1/2,j} - \theta_x F_{i+1/2,j-1} + \theta_y F_{i,j-1/2}, \tag{2.23f}$$

and

$$F_{i+1/2,j} = z\,(p_{i,j} - p_{i+1,j}), \qquad F_{i,j+1/2} = z^{-1}\,(p_{i,j} - p_{i,j+1}), \tag{2.24}$$

are the 5P fluxes of the uniform case. We recall that $z = \Delta y / \Delta x$ is the ratio between the grid spacings. For a more detailed derivation of (2.23), see [23, §5.1]. In this construction, each diagonal flux is made up of two horizontal fluxes and two vertical fluxes, corresponding to the possible paths

between a cell and any diagonal cell. It is also worth noting that, for $\theta_x = \theta_y = \theta$, although the discrete Laplacian (2.17) is identical to (2.4), the definition of fluxes (2.23)–(2.24) is *not* identical to (2.8). This has a tremendous impact on the discretization of the saturation equation and makes the 9P2s family very different from the 9P1s one.

The reformulation (2.22) naturally suggests the scheme

$$\begin{aligned}\widetilde{F}^{\boldsymbol{\theta}}_{i+1/2,j} - \widetilde{F}^{\boldsymbol{\theta}}_{i-1/2,j} + \widetilde{F}^{\boldsymbol{\theta}}_{i,j+1/2} - \widetilde{F}^{\boldsymbol{\theta}}_{i,j-1/2} \\ + \widetilde{F}^{\boldsymbol{\theta}\nearrow}_{i+1/2,j+1/2} - \widetilde{F}^{\boldsymbol{\theta}\nearrow}_{i-1/2,j-1/2} + \widetilde{F}^{\boldsymbol{\theta}\nwarrow}_{i-1/2,j+1/2} - \widetilde{F}^{\boldsymbol{\theta}\nwarrow}_{i+1/2,j-1/2} = \Delta x \Delta y \, q_{i,j}\end{aligned} \tag{2.25}$$

for the pressure equation (1.1b) in the general case $\kappa\lambda(s) \not\equiv 1$, where the fluxes are defined by relations (2.23) but in which we have plugged the non-uniform 5P fluxes

$$F_{i+1/2,j} = \kappa\widetilde{\lambda}(s^n_{i,j}, s^n_{i+1,j})\, z\, (p^{n+1}_{i,j} - p^{n+1}_{i+1,j}), \qquad F_{i,j+1/2} = \kappa\widetilde{\lambda}(s^n_{i,j}, s^n_{i,j+1})\, z^{-1}\, (p^{n+1}_{i,j} - p^{n+1}_{i,j+1}). \tag{2.26}$$

Since the factors $1 - 4\theta_x$ and $1 - 4\theta_y$ appear in the fluxes (2.23a)–(2.23b), it is advisable to impose the restriction

$$0 \le \theta_x, \, \theta_y \le \tfrac{1}{4} \tag{2.27}$$

so that the 9P2s horizontal and vertical fluxes (2.23a)–(2.23b) have the same sign as their 5P counterparts $F_{i+1/2,j}$ and $F_{i,j+1/2}$.

2.2.2 9P2s for saturation

Once the pressure field is computed, the saturation equation (1.1c) can be discretized with a scheme having a similar nine-point and eight-flux structure. More specifically,

$$\begin{aligned}\Delta x \Delta y \Delta t^{-1}(s^{n+1}_{i,j} - s^n_{i,j}) + \widetilde{G}^{\boldsymbol{\theta}}_{i+1/2,j} - \widetilde{G}^{\boldsymbol{\theta}}_{i-1/2,j} + \widetilde{G}^{\boldsymbol{\theta}}_{i,j+1/2} - \widetilde{G}^{\boldsymbol{\theta}}_{i,j-1/2} \\ + \widetilde{G}^{\boldsymbol{\theta}\nearrow}_{i+1/2,j+1/2} - \widetilde{G}^{\boldsymbol{\theta}\nearrow}_{i-1/2,j-1/2} + \widetilde{G}^{\boldsymbol{\theta}\nwarrow}_{i-1/2,j+1/2} - \widetilde{G}^{\boldsymbol{\theta}\nwarrow}_{i+1/2,j-1/2} = \Delta x \Delta y \, q_{w;i,j},\end{aligned} \tag{2.28}$$

with the upwinded fluxes

$$\tilde{G}^{\boldsymbol{\theta}}_{i+1/2,j} = f(s^n_{i,j})\,[\widetilde{F}^{\boldsymbol{\theta}}_{i+1/2,j}]^+ + f(s^n_{i+1,j})\,[\widetilde{F}^{\boldsymbol{\theta}}_{i+1/2,j}]^-, \tag{2.29a}$$

$$\tilde{G}^{\boldsymbol{\theta}}_{i,j+1/2} = f(s^n_{i,j})\,[\widetilde{F}^{\boldsymbol{\theta}}_{i,j+1/2}]^+ + f(s^n_{i,j+1})\,[\widetilde{F}^{\boldsymbol{\theta}}_{i,j+1/2}]^-, \tag{2.29b}$$

$$\tilde{G}^{\boldsymbol{\theta}\nearrow}_{i+1/2,j+1/2} = f(s^n_{i,j})\,[\widetilde{F}^{\boldsymbol{\theta}\nearrow}_{i+1/2,j+1/2}]^+ + f(s^n_{i+1,j+1})\,[\widetilde{F}^{\boldsymbol{\theta}\nearrow}_{i+1/2,j+1/2}]^-, \tag{2.29c}$$

$$\tilde{G}^{\boldsymbol{\theta}\nwarrow}_{i-1/2,j+1/2} = f(s^n_{i,j})\,[\widetilde{F}^{\boldsymbol{\theta}\nwarrow}_{i-1/2,j+1/2}]^+ + f(s^n_{i-1,j+1})\,[\widetilde{F}^{\boldsymbol{\theta}\nwarrow}_{i-1/2,j+1/2}]^-. \tag{2.29d}$$

As in §2.1.2, the scheme is also supplemented by a CFL-like condition so as to guarantee the maximum principle for the saturation, at least in regions where both source terms vanish. Let

$$\begin{aligned}\rangle\sigma\widetilde{F}^{\boldsymbol{\theta}}\langle_{i,j} = \sigma_{i-1/2,j}[\widetilde{F}^{\boldsymbol{\theta}}_{i-1/2,j}]^+ - \sigma_{i+1/2,j}[\widetilde{F}^{\boldsymbol{\theta}}_{i+1/2,j}]^- \\ + \sigma_{i,j-1/2}[\widetilde{F}^{\boldsymbol{\theta}}_{i,j-1/2}]^+ - \sigma_{i,j+1/2}[\widetilde{F}^{\boldsymbol{\theta}}_{i,j+1/2}]^- \\ + \sigma_{i-1/2,j-1/2}[\widetilde{F}^{\boldsymbol{\theta}\nearrow}_{i-1/2,j-1/2}]^+ - \sigma_{i+1/2,j+1/2}[\widetilde{F}^{\boldsymbol{\theta}\nearrow}_{i+1/2,j+1/2}]^- \\ + \sigma_{i+1/2,j-1/2}[\widetilde{F}^{\boldsymbol{\theta}\nwarrow}_{i+1/2,j-1/2}]^+ - \sigma_{i-1/2,j+1/2}[\widetilde{F}^{\boldsymbol{\theta}\nwarrow}_{i-1/2,j+1/2}]^-\end{aligned}$$

be the total incoming flux of cell $K_{i,j}$.

Proposition 2.2. *If $q_{i,j} = q_{w;i,j} = 0$ at some cell $K_{i,j}$ and*

$$\frac{\Delta t}{\Delta x \Delta y} \rangle\sigma\widetilde{F}^{\boldsymbol{\theta}}\langle_{i,j} \le 1, \tag{2.30}$$

then $s^{n+1}_{i,j}$ is a convex combination of $s^n_{i,j}$ and its eight neighbouring saturations at time n.

Proof. Similar to that of Proposition 2.1. □

From this we infer a stability condition similar to (2.16). Again, we postpone the error analysis to §3.3, where we will see that the approximation in saturation remains of first-order.

3 Optimization of the parameters

The main issue of this paper is to correctly design the parameters θ in order to decrease as much as possible the anisotropy of the numerical error when the exact solution is radial. Note that this is not equivalent to minimizing the numerical error itself. Once again, we emphasize that the order of the numerical error remains unchanged. In fact, only its distribution in space will change. To this end:

1. Firstly, we need to quantify the anisotropy of the numerical error along each direction. This can be achieved by using Fourier analysis under the simplifying assumption of constant coefficients and velocities.

2. Secondly, we need to introduce an ideal behaviour of the angular error that we declare to be the "least anisotropic" one. There might be some degree of arbitrariness in this choice, but we will try to suggest the most natural one.

3. Finally, we need to minimize to total discrepancy (over all directions) between the angular error corresponding to the scheme and that of the expected ideal one. Most of the time, we will be able to determine the exact solution of this minimization problem.

In §3.1, we are interested in minimizing the anisotropy of the error in pressure by correctly adjusting the parameter θ of the 9P1s scheme of §2.1.1. In §3.2, we also endeavour to adjust the parameter θ, but this time in an attempt to alleviate the anisotropy of the error in saturation when employing the 9P1s scheme of §2.1.2. In §3.3, the same analysis on the saturation error will be achieved on the pair $\boldsymbol{\theta} = (\theta_x, \theta_y)$ of the 9P2s scheme of §2.2.2.

3.1 Optimization of 9P1s based on pressure

We illustrate the above procedure by focusing on the pressure equation $-\mathrm{div}(\kappa\lambda(s)p) = q$. In order to perform the Fourier analysis, we assume an infinite domain and the hypotheses $\kappa\lambda(s) \equiv 1$ and $q \equiv 0$. By inserting into the exact and approximate operators $-\Delta$ and $-\Delta_h^\theta$ the exponential form

$$p_{i,j} = e^{I(ik\Delta x + j\ell\Delta y)}, \tag{3.1}$$

where the imaginary number I satisfies $I^2 = -1$ and $\boldsymbol{k} = (k, \ell) \in \mathbb{R}^2$ is the wave vector, we end up with the multiplicative relations

$$(-\Delta p)_{i,j} = \mathscr{F}[-\Delta](\boldsymbol{k}) p_{i,j}, \qquad (-\Delta_h^\theta p)_{i,j} = \mathscr{F}[-\Delta_h^\theta](\boldsymbol{k}) p_{i,j}. \tag{3.2}$$

The factors $\mathscr{F}[-\Delta](\boldsymbol{k})$ and $\mathscr{F}[-\Delta_h^\theta](\boldsymbol{k})$ do not depend on (i, j) and are called respectively *exact* and *approximate symbols* of the Laplacian. Let

$$\mathscr{E}^\theta_{\Delta x, \Delta y}(\boldsymbol{k}) = \mathscr{F}[-\Delta_h^\theta](\boldsymbol{k}) - \mathscr{F}[-\Delta](\boldsymbol{k}) \tag{3.3}$$

be the error between the two symbols. This error depends not only on $\Delta x, \Delta y, \theta$ but also on the direction of the wave vector $\boldsymbol{k}$. Let

$$\gamma = \arctan \frac{\ell}{k}$$

be the angle between the horizontal axis and the wave vector.

Lemma 3.1. *If $\Delta x, \Delta y$ are small enough, then*

$$\begin{aligned}\mathscr{E}^\theta_{\Delta x, \Delta y}(\boldsymbol{k}) = &- |\boldsymbol{k}|^4 \big\{ [\tfrac{1}{12} - \theta](\Delta x^2 + \Delta y^2) \sin^4\gamma + [-\tfrac{1}{6}\Delta x^2 + \theta(\Delta x^2 + \Delta y^2)] \sin^2\gamma + \tfrac{1}{12}\Delta x^2 \big\} \\ &+ O(\Delta x^4, \Delta y^4, \Delta x^2 \Delta y^2).\end{aligned} \tag{3.4}$$

Proof. It is straightforward to show that the exact symbol is

$$\mathscr{F}[-\Delta](\boldsymbol{k}) = |\boldsymbol{k}|^2 = k^2 + \ell^2. \tag{3.5}$$

By construction of the 9P1s approximation $-\Delta_h^\theta$, the approximate symbol is given by

$$\begin{aligned}\mathscr{F}[-\Delta_h^\theta](\boldsymbol{k}) = &\big[\Delta x^{-2}(-e^{-Ik\Delta x}+2-e^{Ik\Delta x})e^{I\ell\Delta y} &&+\Delta y^{-2}(-e^{I\ell\Delta y}+2-e^{-I\ell\Delta y})e^{-Ik\Delta x}\big]\cdot\theta\\ &+\big[\Delta x^{-2}(-e^{-Ik\Delta x}+2-e^{Ik\Delta x}) &&+\Delta y^{-2}(-e^{I\ell\Delta y}+2-e^{-I\ell\Delta y})\qquad\big]\cdot(1-2\theta)\\ &+\big[\Delta x^{-2}(-e^{-Ik\Delta x}+2-e^{Ik\Delta x})e^{-I\ell\Delta y} &&+\Delta y^{-2}(-e^{I\ell\Delta y}+2-e^{-I\ell\Delta y})e^{Ik\Delta x}\big]\cdot\theta.\end{aligned}$$

Thanks to the trigonometric identity $-e^{-I\varsigma}+2-e^{I\varsigma}=4\sin^2(\varsigma/2)$, we obtain

$$\mathscr{F}[-\Delta_h^\theta](\boldsymbol{k}) = 4\sin^2(k\Delta x/2)\frac{1-4\theta\sin^2(\ell\Delta y/2)}{\Delta x^2}+4\sin^2(\ell\Delta y/2)\frac{1-4\theta\sin^2(k\Delta x/2)}{\Delta y^2}. \tag{3.6}$$

Now, assuming that $|k|\Delta x\ll 1$ and $|\ell|\Delta y\ll 1$, we can use the Taylor expansion

$$\frac{\sin^2\vartheta}{\vartheta^2}=1-\frac{\vartheta^2}{3}+O(\vartheta^4)$$

in order to end up with

$$\mathscr{F}[-\Delta_h^\theta](\boldsymbol{k}) = |\boldsymbol{k}|^2-\big[\tfrac{1}{12}\Delta x^2k^4+\tfrac{1}{12}\Delta y^2\ell^4+\theta(\Delta x^2+\Delta y^2)\ell^2k^2\big]+O(\Delta x^4,\Delta y^4,\Delta x^2\Delta y^2).$$

Since $k=|\boldsymbol{k}|\cos\gamma$ and $\ell=|\boldsymbol{k}|\sin\gamma$, the above equation combined with (3.5) gives (3.4). □

In the right-hand side of (3.4), the bracket in factor of $|\boldsymbol{k}|^4$ depends only on the angle γ. Hence, it is natural to raise it to the status of a definition.

Definition 3.1. The quantity

$$\widetilde{\mathscr{E}}^\theta_{\Delta x,\Delta y}(\gamma) = \big[\tfrac{1}{12}-\theta\big](\Delta x^2+\Delta y^2)\sin^4\gamma+\big[-\tfrac{1}{6}\Delta x^2+\theta(\Delta x^2+\Delta y^2)\big]\sin^2\gamma+\tfrac{1}{12}\Delta x^2 \tag{3.7}$$

is said to be the angular error in pressure along the direction γ associated with the nine-point scheme.

Let us set

$$S=\sin^2\gamma\in[0,1]. \tag{3.8}$$

We observe that $\widetilde{\mathscr{E}}^\theta_{\Delta x,\Delta y}$ is a quadratic polynomial with respect to S. From now on, with a slight abuse of notation, the numerical error $\widetilde{\mathscr{E}}^\theta_{\Delta x,\Delta y}$ is now as a function of S and reads

$$\widetilde{\mathscr{E}}^\theta_{\Delta x,\Delta y}(S) = \big[\tfrac{1}{12}-\theta\big](\Delta x^2+\Delta y^2)S^2+\big[-\tfrac{1}{6}\Delta x^2+\theta(\Delta x^2+\Delta y^2)\big]S+\tfrac{1}{12}\Delta x^2. \tag{3.9}$$

We remark that for all θ,

$$\widetilde{\mathscr{E}}^\theta_{\Delta x,\Delta y}(S=0)=\tfrac{1}{12}\Delta x^2\qquad\text{and}\qquad\widetilde{\mathscr{E}}^\theta_{\Delta x,\Delta y}(S=1)=\tfrac{1}{12}\Delta y^2. \tag{3.10}$$

This implies that the angular errors along the direction of the axes cannot be modified by the tuning parameter θ. As a consequence, with $\Delta x\neq\Delta y$, there is always a residual anisotropy between the x-direction and the y-direction that cannot be removed. However, we are offered the freedom to select an "ideal" transition from $S=0$ to $S=1$. We claim that the straight line

$$\widetilde{\mathscr{E}}^\star_{\Delta x,\Delta y}(S) = \tfrac{1}{12}\left[(\Delta y^2-\Delta x^2)S+\Delta x^2\right]. \tag{3.11}$$

can be regarded as the least anisotropic choice. Indeed, among all functions $\widetilde{\mathscr{E}}:[0,1]\to\mathbb{R}$ with end values $\widetilde{\mathscr{E}}(0)=\Delta x^2/12$ and $\widetilde{\mathscr{E}}(1)=\Delta y^2/12$, the affine function achieves the minimum of the functional $W(\widetilde{\mathscr{E}})=\int_0^1|\widetilde{\mathscr{E}}'(S)|^2\,\mathrm{d}S$ which measures the total squared variations of $\widetilde{\mathscr{E}}$.

Equipped with these preliminary notions, we propose to seek the optimal parameter θ^* to minimize the total anisotropy, defined as the $L^2(0,1)$-distance between $\widetilde{\mathscr{E}}^\theta_{\Delta x,\Delta y}$ and $\widetilde{\mathscr{E}}^\star_{\Delta x,\Delta y}$. In other words,

$$\theta^*=\arg\min_{\theta\in[0,1/2]}\int_0^1|\widetilde{\mathscr{E}}^\theta_{\Delta x,\Delta y}(S)-\widetilde{\mathscr{E}}^\star_{\Delta x,\Delta y}(S)|^2\,\mathrm{d}S. \tag{3.12}$$

Theorem 3.1. *The unique minimizer of* (3.12) *is*

$$\theta^* = \frac{1}{12}. \tag{3.13}$$

Proof. It suffices to note that as soon as $\theta = 1/12$, $\widetilde{\mathscr{E}}^{\theta=1/12}_{\Delta x,\Delta y} \equiv \widetilde{\mathscr{E}}^{\star}_{\Delta x,\Delta y}$. Then, the value of the objective function $\|\widetilde{\mathscr{E}}^{\theta}_{\Delta x,\Delta y} - \widetilde{\mathscr{E}}^{\star}_{\Delta x,\Delta y}\|^2_{L^2(0,1)}$ vanishes. On the other hand, this is the only value of θ such that $\widetilde{\mathscr{E}}^{\theta}_{\Delta x,\Delta y} \equiv \widetilde{\mathscr{E}}^{\star}_{\Delta x,\Delta y}$. □

The value $\theta^* = 1/12$ was already mentioned by [13] but only for square meshes ($\Delta x = \Delta y$). For rectangular meshes ($\Delta x \neq \Delta y$), we expect the anisotropy error to be reasonably small. In Figure 3, a few curves $\widetilde{\mathscr{E}}^{\theta}_{\Delta x,\Delta y}$ are plotted as functions of s for various values of θ. Note that for the 5P scheme ($\theta = 0$), the red curves appear to be very far from the optimal behaviour for both a square mesh (where the ideal error is represented by the green horizontal line) and a rectangular mesh (where the ideal error is represented by the green straight line).

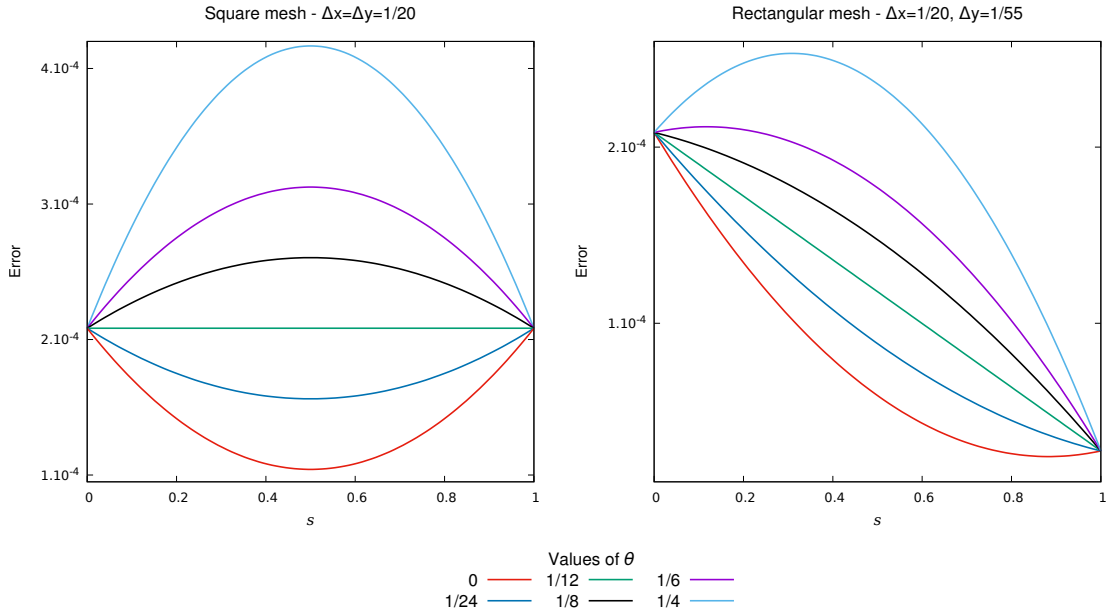


Figure 3: Angular error associated with the nine-point scheme $S \mapsto \widetilde{\mathscr{E}}^{\theta}_{\Delta x,\Delta y}(S)$ for various θ.

3.2 Optimization of 9P1s based on saturation

It could be argued that, despite numerous previous works, the error in pressure considered in §3.1 is not the good quantity to look at. After all, engineers are more interested in the saturation front and therefore it is the error in saturation that should be made more isotropic for a radial solution. Such an analysis was pioneered by [15] for a special scheme in a square mesh. Here, following a different approach, we carry out the analysis for the scheme (2.13)–(2.14) in a rectangular mesh.

Once again, in order to perform Fourier calculations, in (1.1c), we assume an infinite domain and the simplifying hypotheses $f(s) \equiv s$ and $q = q_w \equiv 0$. In addition, we enforce the velocity to be constant, given by $\boldsymbol{u} \equiv {}^t(a, b)$ where $a \geq 0$ and $b \geq 0$ are fixed values. According to the upwinding formulas of s, the scheme (2.13)–(2.14) now writes

$$\begin{aligned} s^{n+1}_{i,j} = s^n_{i,j} - \Delta t(\Delta x\Delta y)^{-1}\big(&s^n_{i,j}\, F^{\theta}_{i+1/2,j} - s^n_{i-1,j}\, F^{\theta}_{i-1/2,j} + s^n_{i,j}\, F^{\theta}_{i,j+1/2} - s^n_{i,j-1}\, F^{\theta}_{i,j-1/2} \\ &+ s^n_{i,j}\, F^{\theta\nearrow}_{i+1/2,j+1/2} - s^n_{i-1,j-1}\, F^{\theta\nearrow}_{i-1/2,j-1/2} + s^n_{i-1/2,j+1/2}\, F^{\theta\nwarrow}_{i-1/2,j+1/2} - s^n_{i+1/2,j-1/2}\, F^{\theta\nwarrow}_{i+1/2,j-1/2}\big), \end{aligned} \tag{3.14}$$

where the fluxes associated with the total velocity $\boldsymbol{u}$ are

$$F^{\theta}_{i\pm1/2,j} = [(1-2\theta)z - 2\theta z^{-1}]a\Delta x, \qquad F^{\theta\nearrow}_{i\pm1/2,j\pm1/2} = \theta[z+z^{-1}](a\Delta x + b\Delta y), \tag{3.15a}$$

$$F^{\theta}_{i,j\pm1/2} = [(1-2\theta)z^{-1} - 2\theta z]b\Delta y, \qquad F^{\theta\nwarrow}_{i\mp1/2,j\pm1/2} = \theta[z+z^{-1}](a\Delta x - b\Delta y), \tag{3.15b}$$

and the interface saturations are

$$(s^n_{i-1/2,j+1/2},\, s^n_{i+1/2,j-1/2}) = \begin{cases} (s^n_{i-1,j+1},\, s^n_{i,j}) & \text{if } a\Delta x - b\Delta y > 0, \\ (s^n_{i,j},\, s^n_{i+1,j-1}) & \text{otherwise.} \end{cases} \tag{3.16}$$

To focus on the discretization in space alone, we study the semi-discrete version of scheme (3.14)

$$\partial_t s_{i,j} + [(\boldsymbol{u}\cdot\boldsymbol{\nabla} s)^{\theta}_h]_{i,j} = 0,$$

where

$$\begin{aligned}[(\boldsymbol{u}\cdot\boldsymbol{\nabla} s)^{\theta}_h]_{i,j} = (\Delta x\Delta y)^{-1}\big(&s_{i,j}\,F^{\theta}_{i+1/2,j} - s_{i-1,j}\,F^{\theta}_{i-1/2,j} + s_{i,j}\,F^{\theta}_{i,j+1/2} - s_{i,j-1}\,F^{\theta}_{i,j-1/2} \\ &+ s_{i,j}\,F^{\theta\nearrow}_{i+1/2,j+1/2} - s_{i-1,j-1}\,F^{\theta\nearrow}_{i-1/2,j-1/2} + s_{i-1/2,j+1/2}\,F^{\theta\nwarrow}_{i-1/2,j+1/2} - s_{i+1/2,j-1/2}\,F^{\theta\nwarrow}_{i+1/2,j-1/2}\big),\end{aligned}$$

with the interface saturations $s_{i\mp1/2,j\pm1/2}$ defined in (3.16). By plugging into the exact and approximate operators $\boldsymbol{u}\cdot\nabla$ and $(\boldsymbol{u}\cdot\nabla)^{\theta}_h$ the exponential form

$$s_{i,j} = e^{I(ik\Delta x + j\ell\Delta y)}, \tag{3.17}$$

where $\boldsymbol{k} = (k,\ell) \in \mathbb{R}^2$ is the wave vector, we arrive at the multiplicative expressions

$$(\boldsymbol{u}\cdot\nabla s)_{i,j} = \mathscr{F}[\boldsymbol{u}\cdot\nabla](\boldsymbol{k})s_{i,j} \qquad \text{and} \qquad ((\boldsymbol{u}\cdot\nabla s)^{\theta}_h)_{i,j} = \mathscr{F}[(\boldsymbol{u}\cdot\nabla)^{\theta}_h](\boldsymbol{k})s_{i,j}.$$

Now, we study the error

$$\mathscr{E}^{\theta}_{\Delta x,\Delta y}(\boldsymbol{u},\boldsymbol{k}) = \mathscr{F}[(\boldsymbol{u}\cdot\nabla)^{\theta}_h](\boldsymbol{k}) - \mathscr{F}[\boldsymbol{u}\cdot\nabla](\boldsymbol{k}).$$

between the exact symbol $\mathscr{F}[\boldsymbol{u}\cdot\nabla]$ and the approximate symbol $\mathscr{F}[(\boldsymbol{u}\cdot\nabla)^{\theta}_h]$. Let

$$\gamma = \arctan\frac{b}{a}, \qquad \varphi = \arctan\frac{\ell}{k}$$

be the angles made by the horizontal axis with respectively the velocity vector and the wave vector. For the sake of simplicity in the notations, we introduce

$$\Omega = \varphi - \gamma, \qquad \gamma^* = \arctan\frac{\Delta x}{\Delta y}.$$

Thus Ω is the angle between $\boldsymbol{u}$ and $\boldsymbol{k}$.

Lemma 3.2. *If $\Delta x, \Delta y$ are small enough, then*

$$\mathscr{E}^{\theta}_{\Delta x,\Delta y}(\boldsymbol{u},\boldsymbol{k}) = |\boldsymbol{k}|^2|\boldsymbol{u}|\,(\cos\Omega,\sin\Omega)\begin{bmatrix}\widetilde{A}^{\theta}_{\Delta x,\Delta y}(\gamma) & \widetilde{B}^{\theta}_{\Delta x,\Delta y}(\gamma)\\ \widetilde{B}^{\theta}_{\Delta x,\Delta y}(\gamma) & \widetilde{C}^{\theta}_{\Delta x,\Delta y}(\gamma)\end{bmatrix}\begin{pmatrix}\cos\Omega\\ \sin\Omega\end{pmatrix} + O(\Delta x^2,\Delta y^2,\Delta x\Delta y), \tag{3.18}$$

where if $\gamma \le \gamma^*$,

$$\widetilde{A}^{\theta}_{\Delta x,\Delta y}(\gamma) = \tfrac{1}{2}\Delta x\cos^3\gamma + \tfrac{1}{2}\Delta y\big[1-2\theta(1+z^2)\big]\sin^3\gamma + 3\Delta y\,\theta(z+z^{-1})\cos\gamma\,\sin^2\gamma, \tag{3.19a}$$

$$\begin{aligned}\widetilde{B}^{\theta}_{\Delta x,\Delta y}(\gamma) = &-\Delta y\,\theta(z+z^{-1})\sin^3\gamma + \tfrac{1}{2}\Delta y\big[1-2\theta(1+z^2)\big]\cos\gamma\,\sin^2\gamma \\ &- \big[\tfrac{1}{2}\Delta x - 2\Delta y\,\theta(z+z^{-1})\big]\cos^2\gamma\,\sin\gamma,\end{aligned} \tag{3.19b}$$

$$\begin{aligned}\widetilde{C}^{\theta}_{\Delta x,\Delta y}(\gamma) = &\Delta y\,\theta(z+z^{-1})\cos^3\gamma + \tfrac{1}{2}\Delta y[1-2\theta(1+z^2)]\cos^2\gamma\,\sin\gamma \\ &+ \big[\tfrac{1}{2}\Delta x - 2\Delta y\,\theta(z+z^{-1})\big]\cos\gamma\,\sin^2\gamma,\end{aligned} \tag{3.19c}$$

while if $\gamma \geq \gamma^*$,

$$\widetilde{A}^{\theta}_{\Delta x,\Delta y}(\gamma) = \tfrac{1}{2}\Delta y \sin^3\gamma + \tfrac{1}{2}\Delta x\big[1-2\theta(1+z^{-2})\big]\cos^3\gamma + 3\Delta x\,\theta(z+z^{-1}))\sin\gamma\,\cos^2\gamma, \quad (3.20a)$$

$$\widetilde{B}^{\theta}_{\Delta x,\Delta y}(\gamma) = \Delta x\,\theta(z+z^{-1})\cos^3\gamma - \tfrac{1}{2}\Delta y\big[1-2\theta(1+z^{-2})\big]\cos^2\gamma\,\sin\gamma + \big[\tfrac{1}{2}\Delta y - 2\Delta x\,\theta(z+z^{-1})\big]\cos\gamma\,\sin^2\gamma, \quad (3.20b)$$

$$\widetilde{C}^{\theta}_{\Delta x,\Delta y}(\gamma) = \Delta x\,\theta(z+z^{-1})\sin^3\gamma + \tfrac{1}{2}\Delta x\big[1-2\theta(1+z^{-2})\big]\cos\gamma\,\sin^2\gamma + \big[\tfrac{1}{2}\Delta y - 2\Delta x\,\theta(z+z^{-1})\big]\cos^2\gamma\,\sin\gamma. \quad (3.20c)$$

Proof. It is plain that, for the exact symbol,

$$\mathscr{F}[\boldsymbol{u}\cdot\nabla](\boldsymbol{k}) = I\,\boldsymbol{k}\cdot\boldsymbol{u} = I(ak+b\ell). \quad (3.21)$$

For the sake of simplicity in the forthcoming developments, we only consider $a\Delta x - b\Delta y \geq 0$ while $a\Delta x - b\Delta y \leq 0$ turns out to be similar and it is let to the reader. Since $a \geq 0$ and $b \geq 0$, this is equivalent to $\gamma \leq \gamma^*$. From the numerical accumulation term (3.14), the approximate symbol can be inferred as

$$\begin{aligned}\mathscr{F}[(\boldsymbol{u}\cdot\boldsymbol{\nabla})^{\theta}_h](\boldsymbol{k}) = (\Delta x\Delta y)^{-1}\big\{\ & \big[(1-2\theta)z - 2\theta z^{-1}\big]a(1-e^{-Ik\Delta x}) \\ & + \big[(1-2\theta)z^{-1} - 2\theta z\big]b\Delta y(1-e^{-I\ell\Delta y}) \\ & + \theta\big[z+z^{-1}\big](a\Delta x + b\Delta y)(1-e^{I(-k\Delta x-\ell\Delta y)}) \\ & + \theta\big[z+z^{-1}\big](a\Delta x - b\Delta y)(1-e^{I(-k\Delta x+\ell\Delta y)})\ \big\}. \end{aligned} \quad (3.22)$$

From various Taylor expansions for $|k|\Delta x \ll 1$ and $|\ell|\Delta y \ll 1$, we get

$$\begin{aligned}\mathscr{F}[(\boldsymbol{u}\cdot\boldsymbol{\nabla})^{\theta}_h](\boldsymbol{k}) = {} & I(ak+b\ell) + \tfrac{1}{2}k^2\Delta x\,a + k\ell\Delta y\,2b\theta(z+z^{-1}) \\ & + \tfrac{1}{2}\ell^2\Delta y\big[b\{1-2\theta(1+z^2)\} + 2a\theta(z+z^{-1})\big] + O(\Delta x^2, \Delta y^2, \Delta x\Delta y). \end{aligned} \quad (3.23)$$

Subtracting (3.21) from this relation, we obtain

$$\begin{aligned}\mathscr{E}^{\theta}_{\Delta x,\Delta y}(\boldsymbol{u},\boldsymbol{k}) = {} & \tfrac{1}{2}k^2\Delta x\,a + k\ell\Delta y\,2b\theta(z+z^{-1}) \\ & + \tfrac{1}{2}\ell^2\Delta y\big\{b\big[1-2\theta(1+z^2)\big] + 2a\theta(z+z^{-1})\big\} + O(\Delta x^2, \Delta y^2, \Delta x\Delta y). \end{aligned} \quad (3.24)$$

Since $k = |\boldsymbol{k}|\cos\varphi$, $\ell = |\boldsymbol{k}|\sin\varphi$, $a = |\boldsymbol{u}|\cos\gamma$, $b = |\boldsymbol{u}|\sin\gamma$, the above equation becomes

$$\mathscr{E}^{\theta}_{\Delta x,\Delta y}(\boldsymbol{u},\boldsymbol{k}) = |\boldsymbol{k}|^2|\boldsymbol{u}|\,(\cos\varphi,\sin\varphi)\begin{bmatrix} A^{\theta}_{\Delta x,\Delta y}(\gamma) & B^{\theta}_{\Delta x,\Delta y}(\gamma) \\ B^{\theta}_{\Delta x,\Delta y}(\gamma) & C^{\theta}_{\Delta x,\Delta y}(\gamma)\end{bmatrix}\begin{pmatrix}\cos\varphi\\ \sin\varphi\end{pmatrix} + O(\Delta x^2, \Delta y^2, \Delta x\Delta y), \quad (3.25)$$

where

$$A^{\theta}_{\Delta x,\Delta y}(\gamma) = \tfrac{1}{2}\Delta x\cos\gamma, \quad (3.26a)$$

$$B^{\theta}_{\Delta x,\Delta y}(\gamma) = \Delta y\,\theta(z+z^{-1})\sin\gamma, \quad (3.26b)$$

$$C^{\theta}_{\Delta x,\Delta y}(\gamma) = \tfrac{1}{2}\Delta y\big\{\big[1-2\theta(1+z^2)\big]\sin\gamma + 2\theta(z+z^{-1})\cos\gamma\big\}. \quad (3.26c)$$

Since $\varphi = \gamma + \Omega$ and because of

$$\begin{pmatrix}\cos\varphi\\ \sin\varphi\end{pmatrix} = \begin{bmatrix}\cos\gamma & -\sin\gamma\\ \sin\gamma & \cos\gamma\end{bmatrix}\begin{pmatrix}\cos\Omega\\ \sin\Omega\end{pmatrix}, \quad (3.27)$$

we easily have

$$\begin{bmatrix} \widetilde{A}^{\theta}_{\Delta x,\Delta y}(\gamma) & \widetilde{B}^{\theta}_{\Delta x,\Delta y}(\gamma) \\ \widetilde{B}^{\theta}_{\Delta x,\Delta y}(\gamma) & \widetilde{C}^{\theta}_{\Delta x,\Delta y}(\gamma)\end{bmatrix} = \begin{bmatrix}\cos\gamma & \sin\gamma\\ -\sin\gamma & \cos\gamma\end{bmatrix}\begin{bmatrix} A^{\theta}_{\Delta x,\Delta y}(\gamma) & B^{\theta}_{\Delta x,\Delta y}(\gamma) \\ B^{\theta}_{\Delta x,\Delta y}(\gamma) & C^{\theta}_{\Delta x,\Delta y}(\gamma)\end{bmatrix}\begin{bmatrix}\cos\gamma & -\sin\gamma\\ \sin\gamma & \cos\gamma\end{bmatrix},$$

to obtain the expected relations. For $a\Delta x - b\Delta y \leq 0$, that is, $\gamma \geq \gamma^*$, the proof is similar, starting from the approximate symbol $\mathscr{F}[(\boldsymbol{u}\cdot\boldsymbol{\nabla})^{\theta}_h]$. Furthermore, it can be checked that the matrix entries (3.19) and (3.20) match with each other when $\gamma = \gamma^*$. □

In the right-hand side of (3.18), the factor of $|\boldsymbol{k}|^2|\boldsymbol{u}|$ depends on two angles. This factor involves a 2×2 *diffusion matrix* whose entries depend on the velocity angle γ and whose action depends on the angle Ω between the velocity and the wave vector. The first diagonal entry $\widetilde{A}^{\theta}_{\Delta x,\Delta y}(\gamma)$ is called *longitudinal error*, as it corresponds to $\Omega=0$. The second diagonal entry $\widetilde{C}^{\theta}_{\Delta x,\Delta y}(\gamma)$ is called *transverse error*, as it corresponds to $\Omega=\pi/2$. The extra diagonal entry is called *cross term error*.

DEFINITION 3.2. The quantity $\widetilde{A}^{\theta}_{\Delta x,\Delta y}(\gamma)$, defined by (3.19a) or (3.20a) according to the sign of $\gamma-\gamma^*$, is said to be the angular error in saturation along the direction γ associated with the 9P1s scheme.

The reason why we opt for the longitudinal error $\widetilde{A}^{\theta}_{\Delta x,\Delta y}(\gamma)$ as a measure of the directional anisotropy is understandable: for a radial solution, the only error that matters for the position of the front is that of the radial diffusion. Equipped with this longitudinal error, we now state the optimal parameter θ. To achieve such an issue, once again, we adopt a suitable comparison with an "ideal error" to be prescribed. First, let us introduce $S=\sin^2\gamma\in[0,1]$ and with some abuse in the notation, let us consider the longitudinal error $\widetilde{A}^{\theta}_{\Delta x,\Delta y}$ as a function of S. To write down this function, let us introduce the transition value

$$S^*=\sin^2\gamma^*=\frac{\tan^2\gamma^*}{1+\tan^2\gamma^*}=\frac{\Delta x^2}{\Delta x^2+\Delta y^2}=\frac{1}{1+z^2}. \tag{3.28}$$

Then, if $S\le S^*$,

$$\widetilde{A}^{\theta}_{\Delta x,\Delta y}(S)=\tfrac{1}{2}\Delta x(1-S)^{3/2}+\tfrac{1}{2}\Delta y\big[1-2\theta(1+z^2)\big]S^{3/2}+3\Delta y\,\theta(z+z^{-1})(1-S)^{1/2}S,$$

while if $S\ge S^*$

$$\widetilde{A}^{\theta}_{\Delta x,\Delta y}(S)=\tfrac{1}{2}\Delta y S^{3/2}+\tfrac{1}{2}\Delta x\big[1-2\theta(1+z^{-2})\big](1-S)^{3/2}+3\Delta x\,\theta(z+z^{-1})(1-S)S^{1/2}.$$

Once again, we point out that

$$\widetilde{A}^{\theta}_{\Delta x,\Delta y}(S=0)=\tfrac{1}{2}\Delta x\qquad\text{and}\qquad\widetilde{A}^{\theta}_{\Delta x,\Delta y}(S=1)=\tfrac{1}{2}\Delta y \tag{3.29}$$

are independent of the parameter θ. Among all functions $\widetilde{A}:[0,1]\to\mathbb{R}$ with end values $\widetilde{A}(0)=\Delta x/2$ and $\widetilde{A}(1)=\Delta y/2$, the affine function

$$\widetilde{A}^{\star}_{\Delta x,\Delta y}(S)=\tfrac{1}{2}\{(\Delta y-\Delta x)S+\Delta x\} \tag{3.30}$$

is supposed to be the "least anisotropic" one, in the sense that it achieves the minimum of the total squared variation $W(\widetilde{A})=\int_0^1|\widetilde{A}'(S)|^2\,\mathrm{d}S$. Therefore, we advocate to look for the optimal parameter θ^* by minimizing the $L^2(0,1)$-distance between $\widetilde{A}^{\theta}_{\Delta x,\Delta y}$ and $\widetilde{A}^{\star}_{\Delta x,\Delta y}$, as

$$\theta^*=\arg\min_{\theta\in[0,\theta_M]}\int_0^1|\widetilde{A}^{\theta}_{\Delta x,\Delta y}(S)-\widetilde{A}^{\star}_{\Delta x,\Delta y}(S)|^2\,\mathrm{d}S, \tag{3.31}$$

where the upperbound θ_M was set in (2.12).

Theorem 3.2. *The unique minimizer of* (3.31) *is*

$$\theta^*=\min\left(\theta_M,\ \frac{\int_0^1 U_{\Delta x,\Delta y}V_{\Delta x,\Delta y}}{\int_0^1 U^2_{\Delta x,\Delta y}}\right) \tag{3.32}$$

where

$$U_{\Delta x,\Delta y}(S)=\begin{cases}\Delta y\big[3(z+z^{-1})(1-S)^{1/2}S-(1+z^2)S^{3/2}\big] & \text{if } S\le S^*,\\ \Delta x\big[3(z+z^{-1})(1-S)S^{1/2}-(1+z^{-2})(1-S)^{3/2}\big] & \text{if } S\ge S^*,\end{cases} \tag{3.33a}$$

$$V_{\Delta x,\Delta y}(s)=\tfrac{1}{2}\Delta x[(1-S)-(1-S)^{3/2}]+\tfrac{1}{2}\Delta y[S-S^{3/2}]. \tag{3.33b}$$

Proof. By definition of $U_{\Delta x,\Delta y}$ and $V_{\Delta x,\Delta y}$, given by (3.33), we have

$$\widetilde{A}^{\theta}_{\Delta x,\Delta y}(S) - \widetilde{A}^{\star}_{\Delta x,\Delta y}(S) = \theta U_{\Delta x,\Delta y}(S) - V_{\Delta x,\Delta y}(S)$$

to write

$$\|\widetilde{A}^{\theta}_{\Delta x,\Delta y} - \widetilde{A}^{\star}_{\Delta x,\Delta y}\|^2_{L^2(0,1)} = \theta^2 \int_0^1 U^2_{\Delta x,\Delta y} - 2\theta \int_0^1 U_{\Delta x,\Delta y} V_{\Delta x,\Delta y} + \int_0^1 V^2_{\Delta x,\Delta y}. \tag{3.34}$$

To minimize this convex quadratic function in θ over the convex interval $[0, 1/2]$, we can first minimize it over $\mathbb{R}$ and then project the solution obtained on the interval. Over $\mathbb{R}$, the function (3.34) easily gets its minimal value at

$$\theta^\sharp = \frac{\int_0^1 U_{\Delta x,\Delta y} V_{\Delta x,\Delta y}}{\int_0^1 U^2_{\Delta x,\Delta y}}.$$

Moreover, we have $U_{\Delta x,\Delta y}(S) \geq 0$ and $V_{\Delta x,\Delta y}(S) \geq 0$ for all $S \in [0, 1]$. Hence, $\theta^\sharp \geq 0$, and the only projection to be made is $\theta^* = \min(\theta_M, \theta^\sharp)$. This compltes the proof. □

Unfortunately, the exact formulas (3.32)–(3.33) are irrelevant from a practical point of view. Indeed, the involved integrals must be evaluated by numerical quadrature and the resulting optimal parameter θ^* is a highly complicated rational fraction of $\Delta y/\Delta x$. To devise a more effective procedure, we content ourselves with a suboptimal value $\theta^\flat$ such that the curve of $\widetilde{A}^{\theta^\flat}_{\Delta x,\Delta y}$ meets that of $\widetilde{A}^{\star}_{\Delta x,\Delta y}$ at the transition point $S = S^*$, where S^* is defined by (3.28).

Theorem 3.3. *The suboptimal value $\theta^\flat$ defined by*

$$\widetilde{A}^{\theta^\flat}_{\Delta x,\Delta y}(S^*) = \widetilde{A}^{\star}_{\Delta x,\Delta y}(S^*).$$

is given by

$$\theta^\flat = \frac{1}{4}\left(\frac{\Delta x + \Delta y}{\sqrt{\Delta x^2 + \Delta y^2}} - 1\right) = \frac{1}{4}\left(\frac{1+z}{\sqrt{1+z^2}} - 1\right). \tag{3.35}$$

Proof. For $S = S^* = 1/(1+z^2)$, we readily have

$$\widetilde{A}^{\star}_{\Delta x,\Delta y}(S^*) = \frac{\Delta x \Delta y(\Delta x + \Delta y)}{2(\Delta x^2 + \Delta y^2)}, \qquad \widetilde{A}^{\theta}_{\Delta x,\Delta y}(S^*) = \frac{1+4\theta}{2}\frac{\Delta x \Delta y}{\sqrt{\Delta x^2 + \Delta y^2}},$$

for all θ. Equality of these two values for $\theta = \theta^\flat$ implies (3.35). Moreover, it is straightforward to verify that $\theta^\flat \in [0, 1/2]$ and the proof is completed. □

Note that, for a square mesh ($\Delta x = \Delta y = h$), the suboptimal value degenerates to

$$\theta^\flat = \frac{\sqrt{2}-1}{4} \approx 0.103553, \tag{3.36}$$

which coincides with the parameter recommended by [15]. The analysis of [15] is intimately related to a square mesh and does not carry over to a rectangular mesh, contrary to ours. Moreover, direct calculations from (3.19)–(3.20) show that $\theta = \theta^\flat$ is the only value such that

$$\widetilde{A}^{\theta}_{h,h}(\gamma + \pi/4) = \widetilde{A}^{\theta}_{h,h}(\gamma), \quad \widetilde{B}^{\theta}_{h,h}(\gamma + \pi/4) = \widetilde{B}^{\theta}_{h,h}(\gamma), \quad \widetilde{C}^{\theta}_{h,h}(\gamma + \pi/4) = \widetilde{C}^{\theta}_{h,h}(\gamma) \tag{3.37}$$

for all $\gamma \in [0, \pi/4]$. The $\pi/4$-invariant property (3.37) of the diffusion matrix was also known by [15]. However, it emerges from our analysis $\pi/4$-invariance does not guarantee strcit optimality, especially in rectangular meshes. In Figure 4, we display the longitudinal error $\widetilde{A}^{\theta}_{\Delta x,\Delta y}$ as a function of $S \in [0, 1]$ for various values of θ. It can be seen that the transition point S^* moves away from $1/2$ for rectangular meshes, but the longitudinal error remains close to the ideal curve.

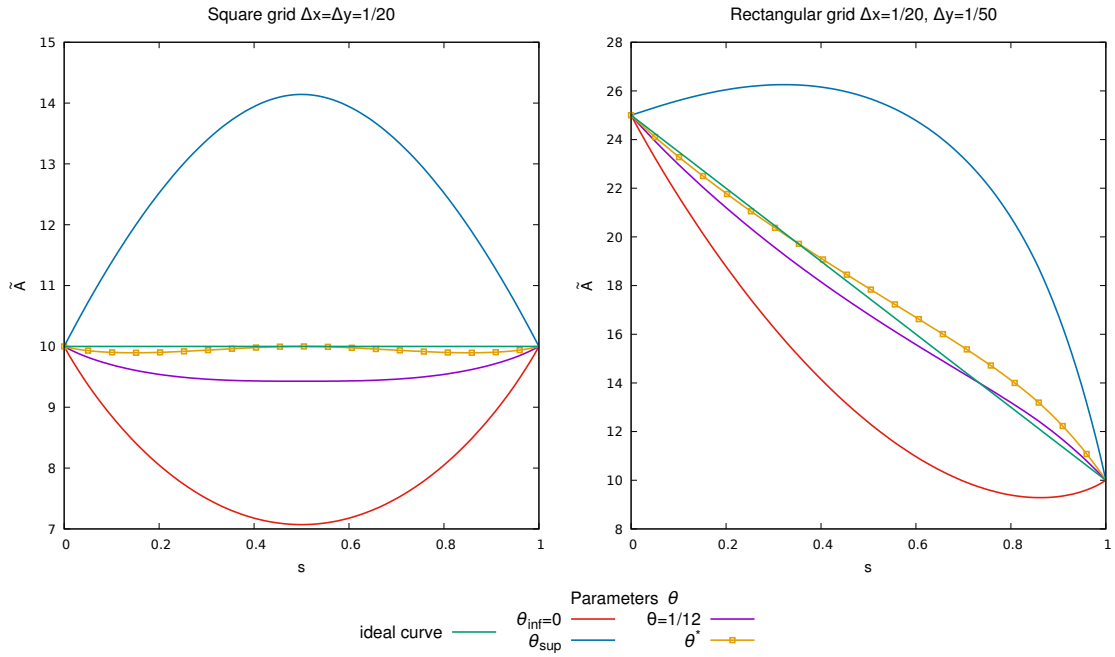


Figure 4: Angular error associated with the 9P1s scheme $s \mapsto \widetilde{A}^{\theta}_{\Delta x,\Delta y}$ for a few values of θ.

3.3 Optimization of 9P2s based on saturation

We now turn to optimizing the 9P2s scheme. Similarly to §3.3, in order to make Fourier analysis possible, we make the linearity assumptions $f(s) = s$, $q = q_w \equiv 0$ and $\boldsymbol{u} \equiv (a, b)$, where $a \geq 0$ and $b \geq 0$ are constants. The saturation transport (2.28) can then be written as

$$\begin{aligned} s^{n+1}_{i,j} = s^n_{i,j} - \Delta t(\Delta x\Delta y)^{-1}\big(&s^n_{i,j}\, F^{\boldsymbol{\theta}}_{i+1/2,j} - s^n_{i-1,j}\, F^{\boldsymbol{\theta}}_{i-1/2,j} + s^n_{i,j}\, F^{\boldsymbol{\theta}}_{i,j+1/2} - s^n_{i,j-1}\, F^{\boldsymbol{\theta}}_{i,j-1/2} \\ &+ s^n_{i,j}\, F^{\boldsymbol{\theta}\nearrow}_{i+1/2,j+1/2} - s^n_{i-1,j-1}\, F^{\boldsymbol{\theta}\nearrow}_{i-1/2,j-1/2} + s^n_{i-1/2,j+1/2}\, F^{\boldsymbol{\theta}\nwarrow}_{i-1/2,j+1/2} - s^n_{i+1/2,j-1/2}\, F^{\boldsymbol{\theta}\nwarrow}_{i+1/2,j-1/2}\big), \end{aligned} \tag{3.38}$$

where the fluxes associated with the total velocity $\boldsymbol{u}$ are

$$F^{\boldsymbol{\theta}}_{i\pm1/2,j} = (1-4\theta_x)a\Delta y, \qquad F^{\boldsymbol{\theta}\nearrow}_{i\pm1/2,j\pm1/2} = 2(\theta_x a\Delta y + \theta_y b\Delta x), \tag{3.39a}$$

$$F^{\boldsymbol{\theta}}_{i,j\pm1/2} = (1-4\theta_y)b\Delta x, \qquad F^{\boldsymbol{\theta}\nwarrow}_{i\mp1/2,j\pm1/2} = 2(\theta_x a\Delta y - \theta_y b\Delta x), \tag{3.39b}$$

the interface saturations are

$$(s^n_{i-1/2,j+1/2}, s^n_{i+1/2,j-1/2}) = \begin{cases} (s^n_{i-1,j+1}, s^n_{i,j}) & \text{if } a\Delta x - b\Delta y > 0, \\ (s^n_{i,j}, s^n_{i+1,j-1}) & \text{otherwise.} \end{cases} \tag{3.40}$$

To focus on the discretization in space alone, we study the semi-discrete version of scheme (3.38)

$$\partial_t s_{i,j} + [(\boldsymbol{u}\cdot\boldsymbol{\nabla} s)^{\boldsymbol{\theta}}_h]_{i,j} = 0,$$

where we have set

$$\begin{aligned} [(\boldsymbol{u}\cdot\boldsymbol{\nabla} s)^{\theta}_h]_{i,j} = (\Delta x\Delta y)^{-1}\big(&s_{i,j}\, F^{\theta}_{i+1/2,j} - s_{i-1,j}\, F^{\theta}_{i-1/2,j} + s_{i,j}\, F^{\theta}_{i,j+1/2} - s_{i,j-1}\, F^{\theta}_{i,j-1/2} \\ &+ s_{i,j}\, F^{\theta\nearrow}_{i+1/2,j+1/2} - s_{i-1,j-1}\, F^{\theta\nearrow}_{i-1/2,j-1/2} + s_{i-1/2,j+1/2}\, F^{\theta\nwarrow}_{i-1/2,j+1/2} - s_{i+1/2,j-1/2}\, F^{\theta\nwarrow}_{i+1/2,j-1/2}\big), \end{aligned}$$

with the interface saturations $s_{i\mp1/2,j\pm1/2}$ defined in (3.40).

Reusing notations from the Fourier setup of §3.2, with $\boldsymbol{k} = (k, \ell) \in \mathbb{R}^2$ the wave vector, we define the exact symbol $\mathscr{F}[\boldsymbol{u}\cdot\nabla]$ and the approximate symbol $\mathscr{F}[(\boldsymbol{u}\cdot\nabla)^{\boldsymbol{\theta}}_h](\boldsymbol{k})$ by plugging the

exponential form (3.17) into the corresponding operators. This allows us to define the Fourier error

$$\mathscr{E}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\boldsymbol{u},\boldsymbol{k}) = \mathscr{F}[(\boldsymbol{u}\cdot\nabla)^{\boldsymbol{\theta}}_h(\boldsymbol{k}) - \mathscr{F}[\boldsymbol{u}\cdot\nabla](\boldsymbol{k}).$$

for which we seek a Taylor expansion in Δx, Δy. This is the purpose of the following statement, in which we have defined the transition angle

$$\gamma^* = \arctan\frac{\theta_x\,\Delta y}{\theta_y\,\Delta x}.$$

Lemma 3.3. *If $\Delta x, \Delta y$ are small enough, then*

$$\mathscr{E}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\boldsymbol{u},\boldsymbol{k}) = |\boldsymbol{k}|^2|\boldsymbol{u}|\,(\cos\Omega,\sin\Omega)\begin{bmatrix}\widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) & \widetilde{B}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma)\\ \widetilde{B}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) & \widetilde{C}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma)\end{bmatrix}\begin{pmatrix}\cos\Omega\\ \sin\Omega\end{pmatrix} + O(\Delta x^2,\Delta y^2,\Delta x\Delta y), \tag{3.41}$$

where if $\gamma \le \gamma^*$,

$$\widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) = \tfrac{1}{2}\Delta x\cos^3\gamma + 2\Delta x(2\theta_y+\theta_x\,z^2)\cos\gamma\sin^2\gamma + \tfrac{1}{2}\Delta y(1-4\theta_y)\sin^3\gamma, \tag{3.42a}$$

$$\widetilde{B}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) = \tfrac{1}{2}\Delta y(1-4\theta_y)\cos\gamma\sin^2\gamma - 2\Delta x\theta_y\sin^3\gamma + \tfrac{1}{2}\Delta x(4\theta_y+4\theta_x z^2-1)\cos^2\gamma\sin\gamma, \tag{3.42b}$$

$$\widetilde{C}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) = 2\Delta x\,\theta_x z^2\cos^3\gamma + \tfrac{1}{2}\Delta y(1-4\theta_y)\cos^2\gamma\sin\gamma + \tfrac{1}{2}\Delta x(1-8\theta_y)\cos\gamma\sin^2\gamma, \tag{3.42c}$$

while if $\gamma \ge \gamma^*$,

$$\widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) = \tfrac{1}{2}\Delta x(1-4\theta_x)\cos^3\gamma + 2\Delta y(2\theta_x+z^{-2})\cos^2\gamma\sin\gamma + \tfrac{1}{2}\Delta y\sin^3\gamma, \tag{3.43a}$$

$$\widetilde{B}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) = -\tfrac{1}{2}\Delta x(1-4\theta_x)\cos^2\gamma\sin\gamma + 2\Delta y\theta_x\cos^3\gamma - \tfrac{1}{2}(4\theta_x+4\theta_y z^{-2}-1)\cos^2\gamma\sin\gamma, \tag{3.43b}$$

$$\widetilde{C}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) = 2\Delta y\theta_y\,z^{-2}\sin^3\gamma + \tfrac{1}{2}\Delta x(1-4\theta_x)\cos\gamma\sin^2\gamma + \tfrac{1}{2}\Delta y(1-8\theta_x)\cos^2\gamma\sin\gamma. \tag{3.43c}$$

Proof. We provide the proof for $a\Delta y\theta_x - b\Delta x\theta_y \ge 0$, the other case $a\Delta y\theta_x - b\Delta x\theta_y \le 0$ being similar. Since $a \ge 0$ and $b \ge 0$, this is equivalent to $\gamma \le \gamma^*$. From the numerical accumulation term (3.38), the approximate symbol can be inferred as

$$\begin{aligned}\mathscr{F}([\boldsymbol{u}\cdot\boldsymbol{\nabla}]^{\boldsymbol{\theta}}_h)(\boldsymbol{k}) = (\Delta x\Delta y)^{-1}\big\{&\,(1-4\theta_x)a\Delta y\,(1-e^{-Ik\Delta x}) + (1-4\theta_y)b\Delta x\,(1-e^{-I\ell\Delta y})\\ &+2\big[\theta_x a\Delta x+\theta_y b\Delta x\big](1-e^{-Ik\Delta x-I\ell\Delta y})\\ &+2\big[\theta_x a\Delta y-\theta_y b\Delta x\big](1-e^{-Ik\Delta x+I\ell\Delta y})\,\big\}.\end{aligned}$$

For $|k|\Delta x \ll 1$ and $|\ell|\Delta y \ll 1$, Taylor expansions yield

$$\begin{aligned}\mathscr{F}([\boldsymbol{u}\cdot\boldsymbol{\nabla}]^{\boldsymbol{\theta}}_h)(\boldsymbol{k}) = I(ak+b\ell) + \tfrac{1}{2}\Delta x\,ak^2 + \tfrac{1}{2}\Delta y\,\big[b(1-4\theta_y)+4a\theta_x z\big]\ell^2 + 4\Delta x\,\theta_y bk\ell\\ + O(\Delta x^2,\Delta y^2,\Delta x\Delta y).\end{aligned}$$

Subtracting (3.21) from the above relation gives

$$\mathscr{E}^{\boldsymbol{\theta}}(\boldsymbol{k},\boldsymbol{u}) = \tfrac{1}{2}\Delta x\,ak^2 + \tfrac{1}{2}\Delta y\,\big[b(1-4\theta_y)+4a\theta_x z\big]\ell^2 + 4\Delta x\,\theta_y bk\ell + O(\Delta x^2,\Delta y^2,\Delta x\Delta y). \tag{3.44}$$

Substituting $k = |\boldsymbol{k}|\cos\varphi$, $\ell = |\boldsymbol{k}|\sin\varphi$, $a = |\boldsymbol{u}|\cos\gamma$, $b = |\boldsymbol{u}|\sin\gamma$ into (3.44) results in

$$\mathscr{E}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\boldsymbol{u},\boldsymbol{k}) = |\boldsymbol{k}|^2|\boldsymbol{u}|\,(\cos\varphi,\sin\varphi)\begin{bmatrix}A^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) & B^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma)\\ B^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) & C^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma)\end{bmatrix}\begin{pmatrix}\cos\varphi\\ \sin\varphi\end{pmatrix} + O(\Delta x^2,\Delta y^2,\Delta x\Delta y), \tag{3.45}$$

where

$$A^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) = \tfrac{1}{2}\Delta x\cos\gamma,\quad B^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) = 2\Delta x\theta_y\sin\gamma,\quad C^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) = \tfrac{1}{2}\Delta y\big[(1-4\theta_y)\sin\gamma+4\theta_x z\cos\gamma\big].$$

By invoking the trigonometric identity (3.27), we are in a position to reformulate equation (3.45) as (3.41), with

$$\begin{bmatrix} \widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) & \widetilde{B}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) \\ \widetilde{B}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) & \widetilde{C}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) \end{bmatrix} = \begin{bmatrix} \cos\gamma & \sin\gamma \\ -\sin\gamma & \cos\gamma \end{bmatrix} \begin{bmatrix} A^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) & B^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) \\ B^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) & C^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma) \end{bmatrix} \begin{bmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma \end{bmatrix}.$$

Formulas (3.42) are recovered thanks to straightforward calculations. Note that continuity holds at the transition angle $\gamma = \gamma^*$ for the matrix entries (3.42) and (3.43). □

Once again, it is worth mentioning that the right-hand side of (3.41) depends on the velocity angle γ and the angle Ω between the velocity and the wave vector. As a consequence, we can regard $\widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma)$ as the *longitudinal error*, $\widetilde{B}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma)$ as the *cross-term error* and $\widetilde{C}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma)$ as the *transverse error*.

DEFINITION 3.3. The quantity $\widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma)$, defined by (3.42a) and (3.43a), is said to be the angular error in saturation along the direction γ associated with the 9P2s scheme.

The choice of $\widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(\gamma)$ is justified on the same grounds as in §3.2. Following the same procedure as in §3.2 and slightly abusing notations, we now consider $\widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}$ as a function of $S = \sin^2\gamma$. Let us introduce the transition value and the transition value

$$S^* = \sin^2\gamma^* = \frac{\omega^2}{1+\omega^2}, \qquad \text{where} \qquad \omega = \tan\gamma^* = \frac{z\,\theta_x}{\theta_y} \tag{3.46}$$

Then, if $S \leq S*$,

$$\widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(S) = \tfrac{1}{2}\Delta x(1-S)^{3/2} + 2\Delta x(2\theta_y + \theta_x z^2)(1-S)^{1/2}S + \tfrac{1}{2}\Delta y(1-4\theta_y)S^{3/2}, \tag{3.47a}$$

while if $S \geq S^*$

$$\widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(S) = \tfrac{1}{2}\Delta x(1-4\theta_x)(1-S)^{3/2} + 2\Delta y(2\theta_x + \theta_y z^{-2})(1-S)S^{1/2} + \tfrac{1}{2}\Delta y S^{3/2}. \tag{3.47b}$$

Its values at $S = 0$ and $S = 1$ do not depend on $\boldsymbol{\theta}$ but only on Δx, Δy. Indeed,

$$\widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(S=0) = \tfrac{1}{2}\Delta x, \qquad \widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(S=1) = \tfrac{1}{2}\Delta y.$$

As a consequence, it is still possible to keep the function $\widetilde{A}^{\star}_{\Delta x,\Delta y}$ defined in (3.30) as the "ideal" least anisotropic reference. As before, the expensive exact optimal

$$\boldsymbol{\theta}^* = \arg \min_{0\leq\theta_x,\theta_y\leq 1/4} \int_0^1 |\widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(S) - \widetilde{A}^{\star}_{\Delta x,\Delta y}(S)|^2 \,\mathrm{d}S$$

can be replaced by the suboptimal value $\boldsymbol{\theta}^\flat$ such that the curve of $\widetilde{A}^{\boldsymbol{\theta}^\flat}_{\Delta x,\Delta y}$ meets that of $\widetilde{A}^{\star}_{\Delta x,\Delta y}$ at the transition point $S = S^*$, i.e.,

$$\widetilde{A}^{\boldsymbol{\theta}^\flat}_{\Delta x,\Delta y}(S^*) = \widetilde{A}^{\star}_{\Delta x,\Delta y}(S^*). \tag{3.48}$$

This time, contrary to §3.2, the transition value S^* depends itself on the parameters $\boldsymbol{\theta}$. We can take advantage of this dependency to move S^* as much as possible to $1/2$. The reason for this is that the closer S^* is to $1/2$, the better the whole curve $\widetilde{A}^{\boldsymbol{\theta}^\flat}$ matches that of $\widetilde{A}^{\star}_{\Delta x,\Delta y}$. Let us work out a solution to this minimization problem in two stages.

Theorem 3.4. *If $\omega = z\theta^\flat_x/\theta^\flat_y$ is prescribed at a given value, then the solution of* (3.48) *is given by*

$$\theta^\flat_x(z,\omega) = \frac{\sqrt{1+\omega^2}(z\omega^2+1) - (1+z\omega^3)}{8z\omega}, \qquad \theta^\flat_y(z,\omega) = \theta^\flat_x(z,\omega)\,z/\omega. \tag{3.49}$$

Proof. At the transition value $S = S^* = \omega^2/(1+\omega^2)$, straightforward calculations show that

$$\widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}(S^*) = \tfrac{1}{2}\Delta x(1+\omega^2)^{-3/2}\big[1+4(2\theta_y+\theta_x z^2)\omega^2+(1-4\theta_y)z\omega^3\big],$$
$$\widetilde{A}^{\star}_{\Delta x,\Delta y}(S^*) = \tfrac{1}{2}\Delta x(1+\omega^2)^{-1}\big[1+z\omega^2\big].$$

Since $\theta_x = \omega\theta_y/z$, equality of these two values for $(\theta_x,\theta_y) = (\theta_x^\flat,\theta_y^\flat)$ implies (3.49). □

We wish to require $\omega = 1$, so that $S^* = 1/2$. Unfortunately, $\theta_x^\flat(z,1)$ and $\theta_y^\flat(z,1)$ may exceed $1/4$ for some z. To comply with (2.27), the idea is to specify $\omega = \omega^*(z)$ in such a way that $\omega = 1$ for "reasonable" values of z and $\max(\theta_x^\flat,\theta_y^\flat) = 1/4$ otherwise.

Proposition 3.1. *The suboptimal pair* $\boldsymbol{\theta}^\flat = (\theta_x^\flat(z,\omega^*(z)),\theta_y^\flat(z,\omega^*(z)))$ *satisfies* (2.27) *for*

$$\omega^*(z) = \begin{cases} \frac{7}{2}z & \text{if } 0 \le z \le 2/7, \\ 1 & \text{if } 2/7 \le z \le 7/2, \\ \frac{2}{7}z & \text{otherwise.} \end{cases}$$

Proof. See [23, §5.3.2]. Calculations rely on the symmetry properties $\theta_x^\flat(z^{-1},\omega^{-1}) = \theta_y^\flat(z,\omega)$, $\theta_y^\flat(z^{-1},\omega^{-1}) = \theta_x^\flat(z,\omega)$ and $\omega^*(z^{-1}) = [\omega^*(z)]^{-1}$. □

For a square mesh ($\Delta x = \Delta y = h$), we recover (3.36). In Figure 5, we display $\widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}$ as a function of $S \in [0,1]$ for various values of (θ_x,θ_y), with $z = 1$ in the left panel and $z = 0.4$ in the right panel.

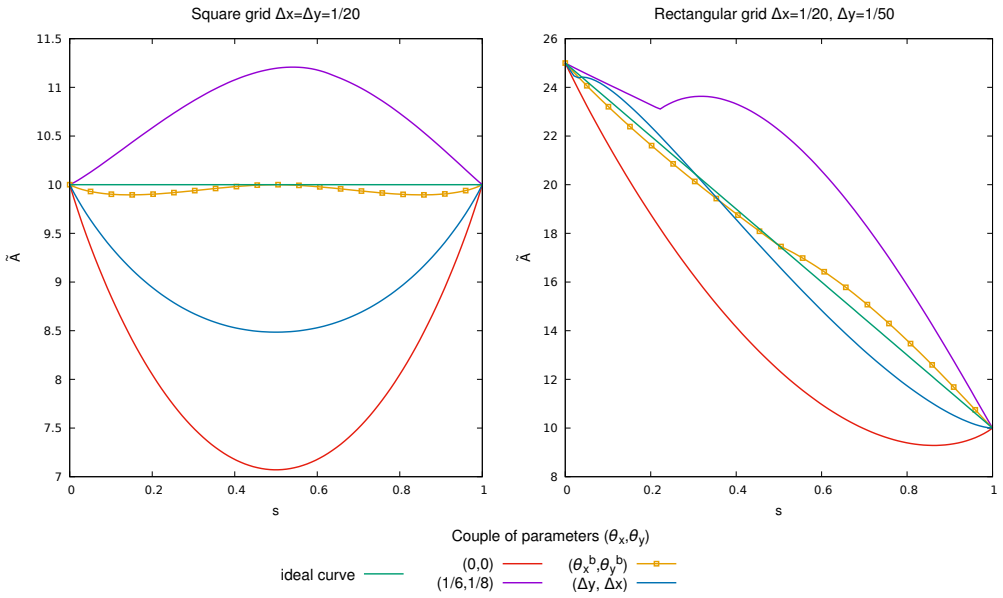


Figure 5: Angular error associated with the 9P2s scheme $s \mapsto \widetilde{A}^{\boldsymbol{\theta}}_{\Delta x,\Delta y}$ for a few pairs (θ_x,θ_y).

4 Numerical results

Two test problems are now supplied in order to demonstrate the effectiveness of the methods designed in §3 for reducing the GOE. problems, the exact solution exhibits a radial symmetry.

4.1 Radial test case

The first problem models an injector well in a homogeneous infinite domain. Consider the system

$$\boldsymbol{u} = -\lambda(s)\boldsymbol{\nabla} p, \tag{4.1a}$$
$$\partial_t s + \operatorname{div}(f(s)\boldsymbol{u}) = \delta_0, \tag{4.1b}$$
$$\operatorname{div}(\boldsymbol{u}) = \delta_0, \tag{4.1c}$$

in $\mathbb{R}^2 \times [0,T]$, $T = 0.05$, with the initial data $s(\boldsymbol{x}, t=0) = 0$ in $\mathbb{R}^2$. In (4.1), $q = q_w = \delta_0$ are Dirac sources expressing liquid injection at $\boldsymbol{x} = \boldsymbol{0}$. The absolute permeability has been assigned the constant value $\kappa = 1$, while the relative permeabilities correspond to the model of [10], that is,

$$\kappa_{r,w}(s) = s^2 \qquad \text{and} \qquad \kappa_{r,o}(1-s) = (1-s)^2. \tag{4.2}$$

As a consequence, the water fractional flux is

$$f(s) = \frac{Ms^2}{Ms^2 + (1-s)^2}, \qquad \text{with} \qquad M = \frac{\mu_o}{\mu_w}. \tag{4.3}$$

Setting $\mu_o = 200$ and $\mu_w = 1$ results in $M = 200$, which is a highly unfavourable mobility ratio.

Lemma 4.1. *Let $r = |\boldsymbol{x}|$ be the distance from the origin and $\boldsymbol{e}_r = \boldsymbol{x}/|\boldsymbol{x}|$ be the unit radial vector. The exact solution of (4.1)–(4.3) is given by*

$$\boldsymbol{u}(r,t) = \boldsymbol{e}_r / 2\pi r \tag{4.4a}$$
$$s(r,t) = \begin{cases} (f')^{-1}_{[s^*,1]}(\pi r^2/t) & \text{if } 0 < r^2 < f'(s^*)\, t/\pi, \\ 0 & \text{otherwise,} \end{cases} \tag{4.4b}$$
$$p(r,t) = p_0 + \frac{1}{2\pi} \int_{r_0}^{r} \frac{\mathrm{d}\varsigma}{\lambda(s(\varsigma,t))\varsigma}, \tag{4.4c}$$

where $s^* = (1+M)^{-1/2}$ and $(r_0, p_0) \in \mathbb{R}_*^+ \times \mathbb{R}$ are some arbitrary constants.

Proof. See [23, §2.4.1] or [18]. □

Let us now switch to the finite computational domain $\Omega = [-0.5, 0.5]^2$, over which all of the equations (4.1)–(4.3) are considered. To mimic the infinite problem, we further prescribe the inhomogeneous Neumann boundary condition

$$-\lambda(s)\boldsymbol{\nabla} p \cdot \boldsymbol{n} = \frac{1}{2\pi r}\, \boldsymbol{e}_r \cdot \boldsymbol{n}, \tag{4.5}$$

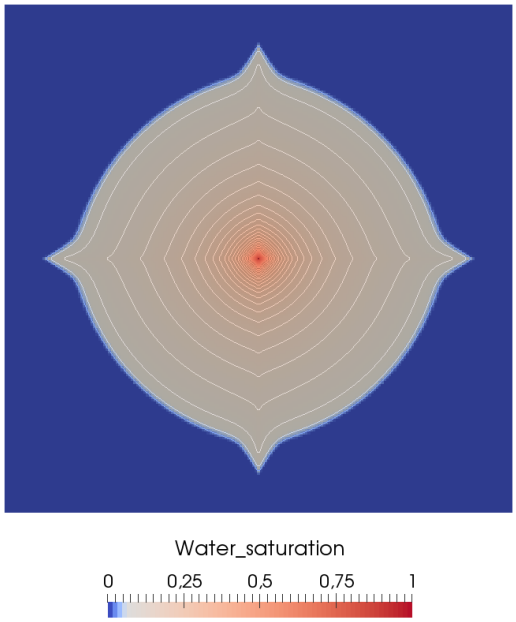
where $\boldsymbol{n}$ denotes the unit outward normal vector of $\partial\Omega$. In other words, the value of the Neumann condition is computed from the exact velocity (4.4a). The following geometrical property greatly helps implementing (4.5), in that it enables one to integrate the outgoing flux over a boundary edge.

Proposition 4.1. *Let A and B be two distinct points in the plane such that the origin $\mathsf{O} = (0,0)$ does not lie on the segment $[\mathsf{AB}]$. Let $\boldsymbol{n}$ be the unit vector such that $(\overrightarrow{\mathsf{AB}}, \boldsymbol{n}) = -\pi/2$. Then,*

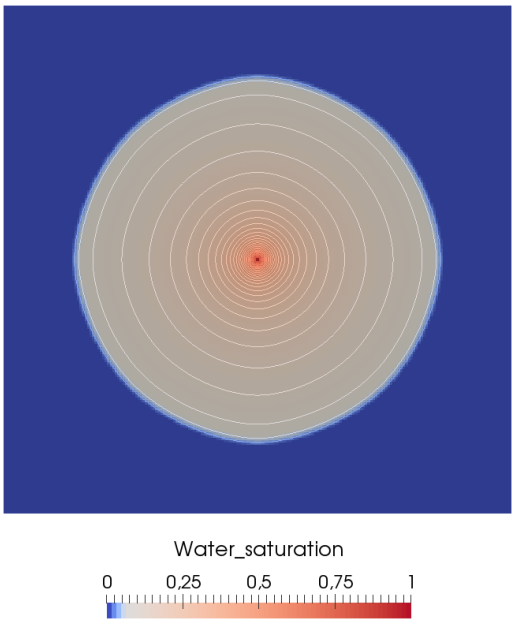
$$\int_{[\mathsf{AB}]} \frac{1}{2\pi r}\, \boldsymbol{e}_r \cdot \boldsymbol{n} = \frac{1}{2\pi}(\overrightarrow{\mathsf{OA}}, \overrightarrow{\mathsf{OB}}),$$

where angles are oriented and measured in radian.

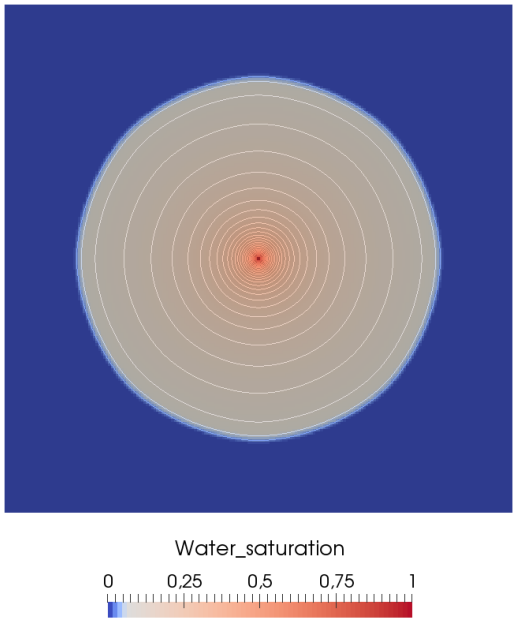
Proof. See [23, §2.4.1] or [18]. □



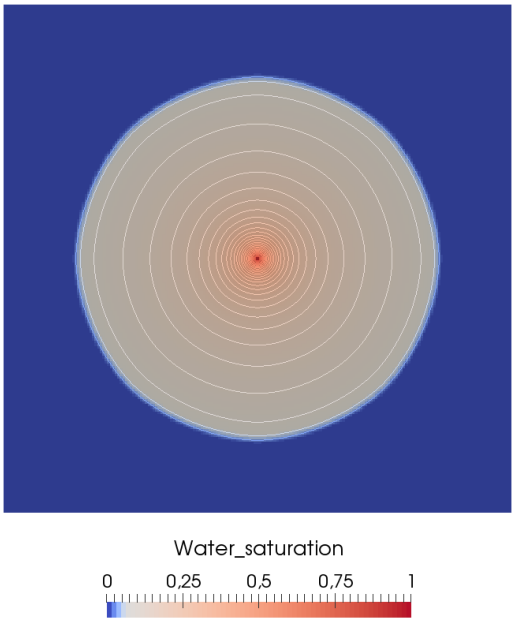
(a) 5P scheme.



(b) 9P1s scheme with $\theta = 1/12$.

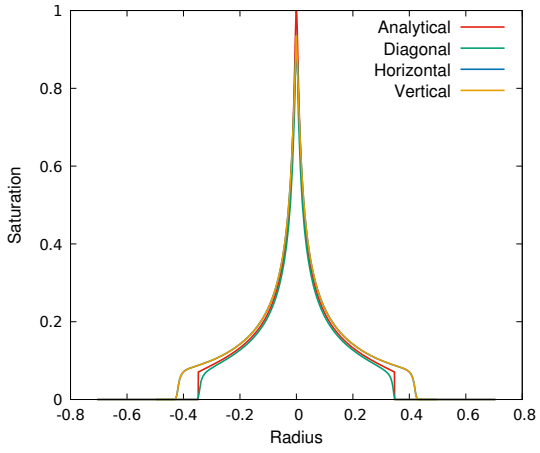


(c) 9P1s scheme with $\theta^\flat$.

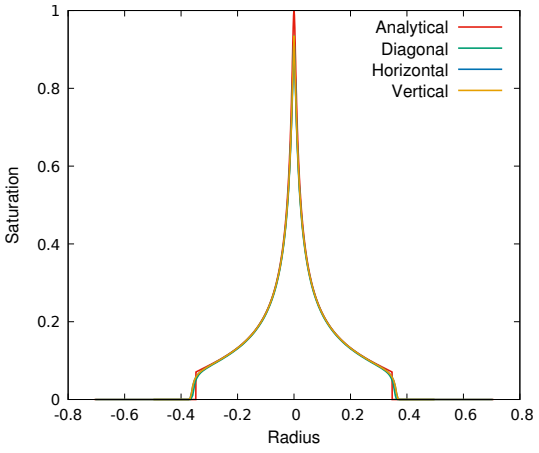


(d) 9P2s scheme with $(\theta_x^\flat, \theta_y^\flat)$.

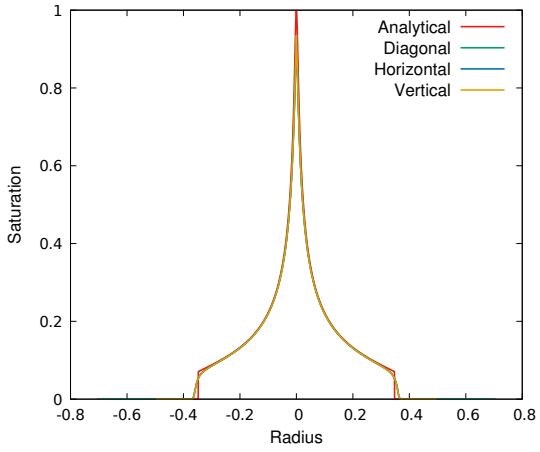
Figure 6: Water saturation fields at $T = 0.05$ for the radial problem on square mesh using four different schemes.



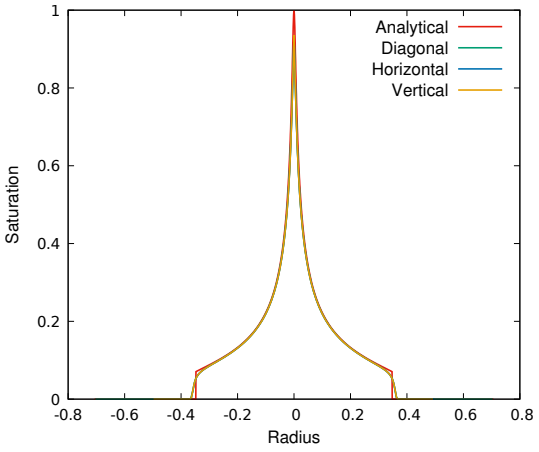
(a) 5P scheme.



(b) 9P1s scheme with $\theta = \frac{1}{12}$.

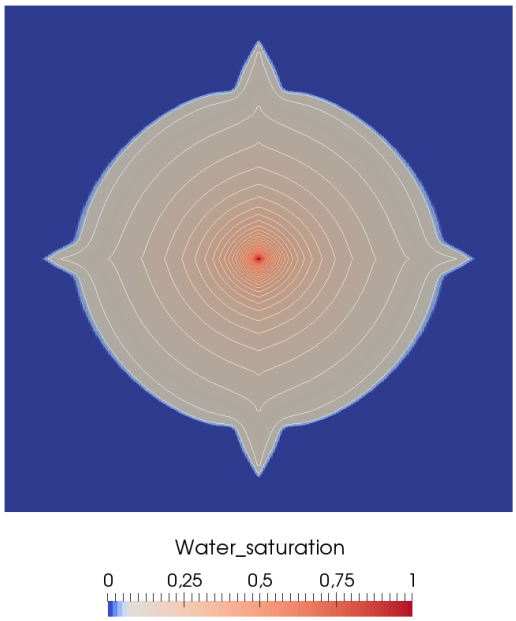


(c) 9P1s scheme with $\theta^\flat$.

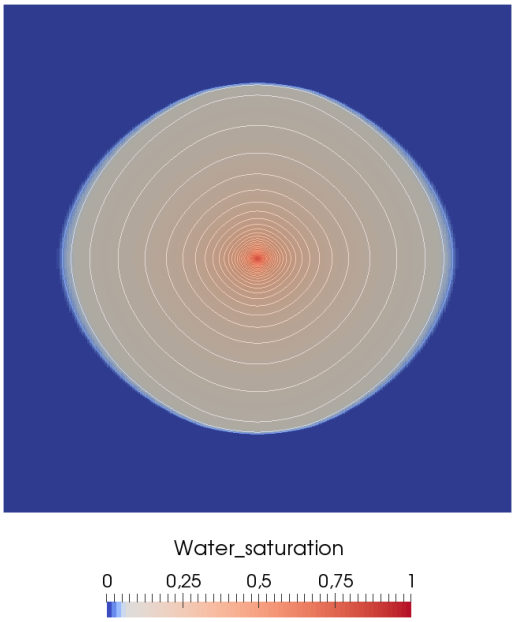


(d) 9P2s scheme with $(\theta_x^\flat, \theta_y^\flat)$.

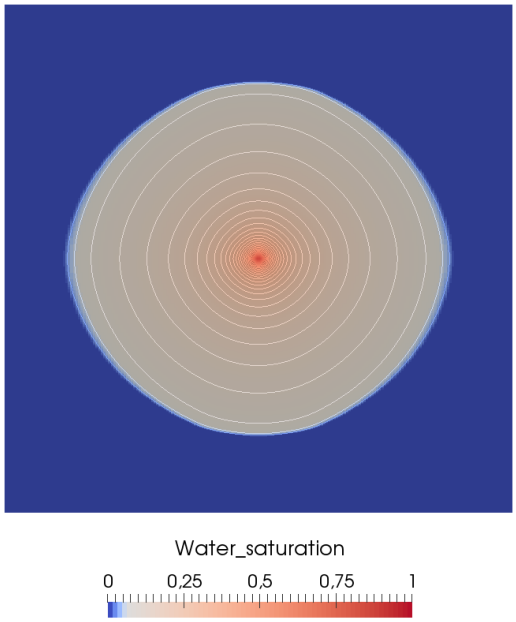
Figure 7: Water saturation profiles for the radial problem on square mesh along the diagonal (green), horizontal (blue) and vertical (yellow) axes.



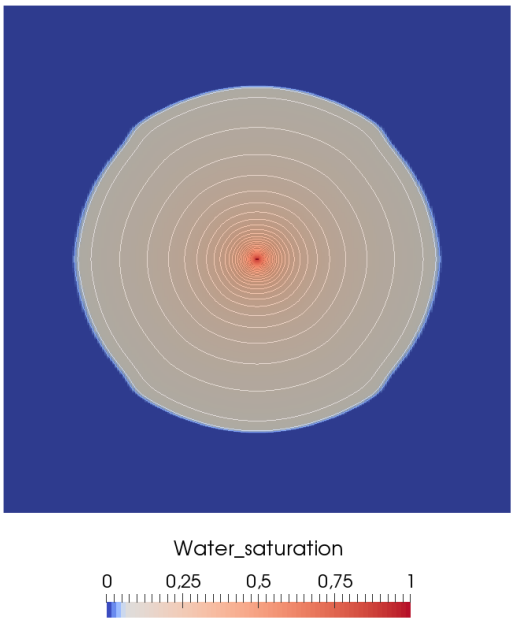
(a) 5P scheme.



(b) 9P1s scheme with $\theta = \frac{1}{12}$.

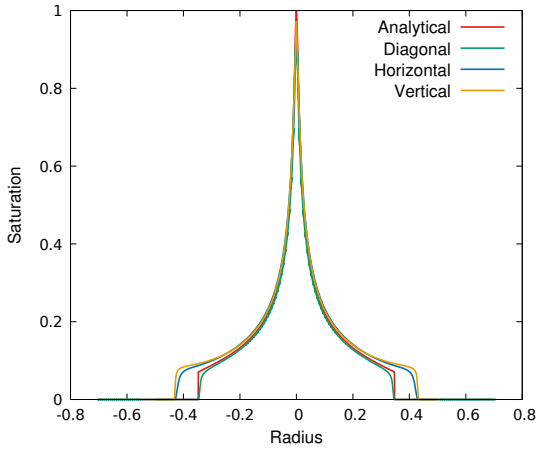


(c) 9P1s scheme with $\theta^\flat$.

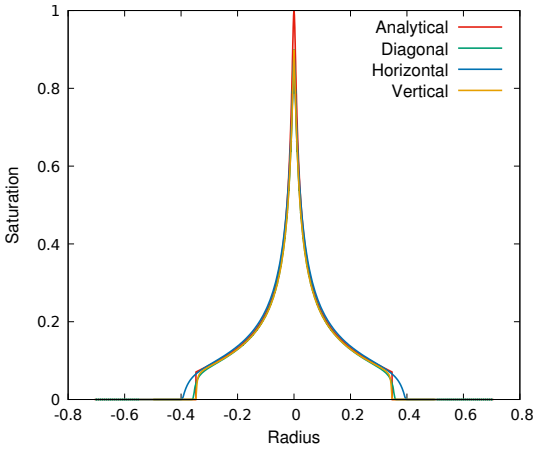


(d) 9P2s scheme with $(\theta_x^\flat, \theta_y^\flat)$.

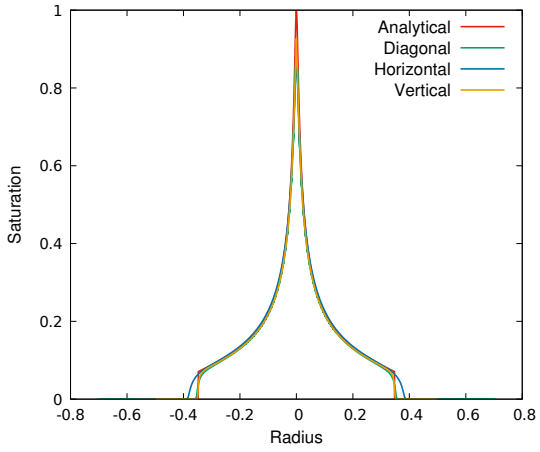
Figure 8: Water saturation fields at $T = 0.05$ for the radial problem on rectangular mesh ($\Delta x = 3\Delta y$) using four different schemes.



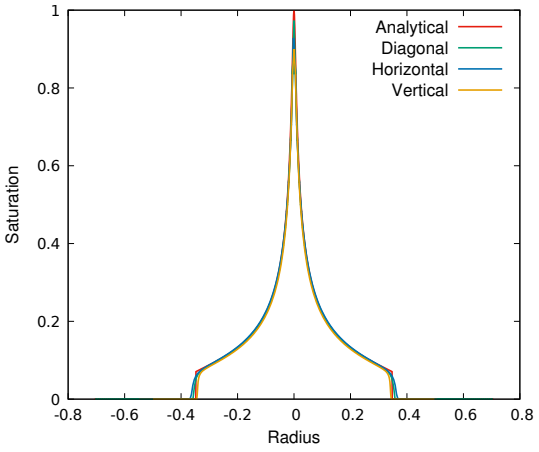
(a) 5P scheme.



(b) 9P1s scheme with $\theta = \frac{1}{12}$.



(c) 9P1s scheme with $\theta^\flat$.



(d) 9P2s scheme with $(\theta_x^\flat, \theta_y^\flat)$.

Figure 9: Water saturation profiles for the radial problem on rectangular mesh ($\Delta x = 3\Delta y$) along the diagonal (green), horizontal (blue) and vertical (yellow) axes.

For the Dirac mass in (4.1) to be correctly discretized, its location $\boldsymbol{x} = \mathbf{0}$ should lie at the center of a cell. Consequently, because of symmetry, the number of cells in each direction should be odd. The simulations are run on two uniform grids: a 201×201 square mesh (Figures 6 and 7) and a 201×601 rectangular mesh (Figures 8 and 9). For each grid, we first display snapshots of $s(\cdot, T)$ computed by 4 methods: (a) the five-point scheme; (b) the 9P1s scheme with $\theta = 1/12$; (c) the 9P1s scheme with $\theta^\flat$ given by (3.35); (d) the 9P2s scheme with (θ_x, θ_y) given by (3.49). Then, we extract 1-D cross-sections along various directions.

The results with the 9P1s scheme and the 9P2s scheme are at the bottom of Figure 6 for the snapshot and of Figure 7 for the saturation profile. The red color is for the analytical solution, the green line is the numerical solution on the diagonal of the domain, the blue one represents the numerical solution on the x-direction and the y-direction is coloured in yellow. Notice that the analytical solution is only represented once because it is invariant per rotation so it is the same in each direction of the mesh. Remark that when the nine-point scheme is used on the saturation equation, a perfect radial solution is obtained on square meshes, that means that the solution is quasi invariant per rotation. On rectangular mesh, the invariant of the solution is not obtained neither for the 9P1s scheme (Figure 8b or c) nor for the 9P2s scheme (Figure 8d) although that the quality of the solution looks better. Compared to the five-point scheme results (Figure 8a), the solution is more radial with our two schemes because of the absence of spikes along the axis of the mesh. Those spikes are visible on the profiles of the saturation too (Figures 7 and 9). On square mesh, it can only be seen the diffusion of the shock whereas the solution is no more radial in rectangular meshes; the x- and y-directions not being together.

4.2 Five-well test case

The second test case is inspired from [20] where we use five wells rather than two. As in the previous test case, we assume that the reservoir is initially saturated with oil and water enters the center of the domain by an injection well. To compare the solutions, two square domains are used, namely,

$$\Omega_1 = (-L/2, L/2)^2, \qquad \Omega_2 = \big\{(x, y) \in \mathbb{R}^2 \mid |x| + |y| \le L/\sqrt{2}\big\},$$

which are deduced from each other by a rotation of angle $\pi/4$. In both domains, the injector well is located at $\boldsymbol{X} = \mathbf{0}$, while the producer wells are located at

$$\boldsymbol{X}_{1,2,3,4} = (\pm d/\sqrt{2}, \pm d/\sqrt{2}),$$

for $0 < d < L/2$, as shown in Figures 10–11. A simulation is performed for each $\Omega \in \{\Omega_1, \Omega_2\}$ in order to approximate the solution of the system

$$\boldsymbol{u} = -\kappa(\boldsymbol{x})\lambda(s)\boldsymbol{\nabla} p, \qquad \text{in } \Omega \times (0, T), \tag{4.6a}$$

$$\operatorname{div}(\boldsymbol{u}) = Q\delta_{\mathbf{0}} - \sum_{\mathrm{L}=1}^{4} Q_{\mathrm{L}}(s, p)\delta_{\boldsymbol{X}_{\mathrm{L}}}, \qquad \text{in } \Omega \times (0, T), \tag{4.6b}$$

$$\partial_t s + \operatorname{div}(f(s)\boldsymbol{u}) = Q\delta_{\mathbf{0}} - \sum_{\mathrm{L}=1} Q_{\mathrm{L}}(s, p) f(s)\delta_{\boldsymbol{X}_{\mathrm{L}}}, \qquad \text{in } \Omega \times (0, T), \tag{4.6c}$$

with the Neumann boundary condition $\boldsymbol{u} \cdot \boldsymbol{n} = 0$ on $\partial\Omega$ and the initial data $s(\boldsymbol{x}, 0) = 0$. As producers work with imposed pressure p_W, the outflows Q_{L} are modeled by [25, 26] as

$$Q_{\mathrm{L}}(s, p) = \lambda(s)\, \frac{2\pi\kappa(\boldsymbol{X}_{\mathrm{L}})}{\ln(r_e/r_{p,\mathrm{L}})}\, (p - p_W(\boldsymbol{X}_{\mathrm{L}})), \tag{4.7}$$

where $p_W(\boldsymbol{X}_{\mathrm{L}})$ is the pressure at the bottom of the well, $r_{p,\mathrm{L}}$ is the radius and $r_e \approx 0.14\sqrt{\Delta x^2 + \Delta y^2}$ is the equivalent radius of the cell. The ratio

$$\mathrm{WP}_{\mathrm{L}} = \frac{2\pi\kappa(\boldsymbol{X}_{\mathrm{L}})}{\ln(r_e/r_{p,\mathrm{L}})}$$

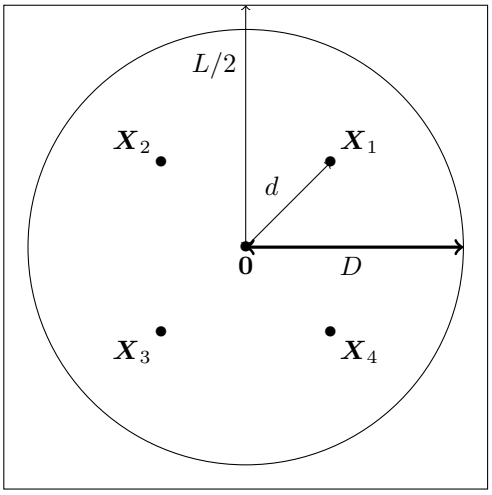


Figure 10: Domain Ω_1 : Diagonal grid.

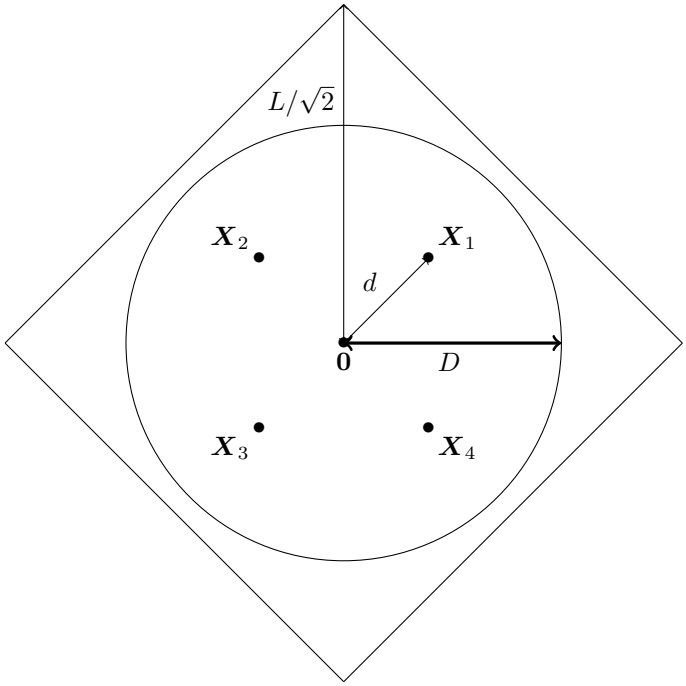


Figure 11: Domain Ω_2 : Parallel grid.

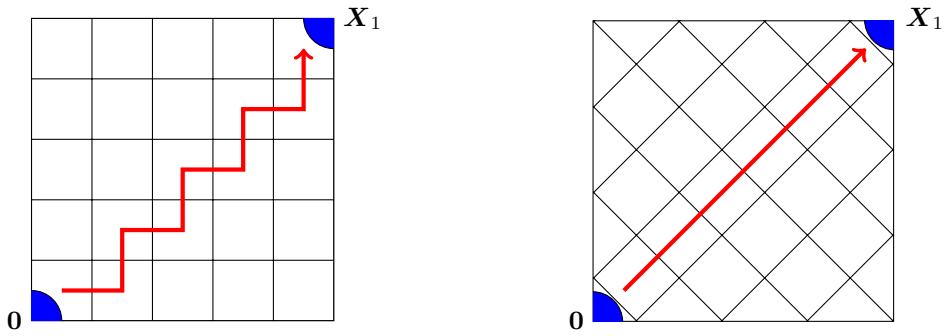


Figure 12: Location of injector well $\mathbf{0}$ and producer well $\boldsymbol{X}_1$. Paths of the numerical flux for a diagonal mesh (on the left) and a parallel mesh (on the right).

is called *Peaceman well index*. Normally, $r_{p,\mathrm{L}} \ll r_e$, so as this index is positive.

The permeability κ is now a function of $\boldsymbol{x}$ that takes two constant values, i.e.,

$$\kappa(\boldsymbol{x}) = \begin{cases} \kappa_M & \text{if } r = \|\boldsymbol{x}\| < D, \\ \kappa_m & \text{otherwise,} \end{cases} \tag{4.8}$$

with $0 < \kappa_m \ll \kappa_M$ and $0 < d < D < L/2$. Having a low permeability where $r > D$ is aimed at preventing the fluid from flowing outside the circle of radius D.

The simulations are run with

$$L = 101\,\mathrm{m}, \qquad N_x = N_y = 101, \qquad \Delta x = \Delta y = 1\,\mathrm{m}.$$

In other words, the two domains Ω_1 and Ω_2 are discretized with squares (see Figure 12). It is important to see that, at the discrete level, the relative position of the producer with resepct to the injector is different for the two domains. Domain Ω_1 is called *diagonal mesh*, because the line connecting the injector to each producer goes diagonally through the mesh. Domain Ω_2 is called *parallel mesh*, to the extent that the same line coincides with the main direction of the mesh.

The remaining lengths of the problem are

$$d = 29.7\,\mathrm{m}, \qquad D = 48.5\,\mathrm{m}.$$

The well parameters are

$$Q = 5\,\mathrm{m}^3 \cdot \mathrm{d}^{-1}, \qquad p_{\mathrm{W}}(\boldsymbol{X}_{\mathrm{L}}) = 50 \cdot 10^5\,\mathrm{Pa}, \qquad r_{p,\mathrm{L}} = 0.05\,\mathrm{m}.$$

Permeabilities and fluid viscosities are

$$\kappa_M = 100\,\mathrm{mD}, \qquad \kappa_m = 10^{-4}\,\mathrm{mD}, \qquad \mu_w = 1\,\mathrm{cP}, \qquad \mu_o = 100\,\mathrm{cP}.$$

The relative permeabilities are taken from [6], that is,

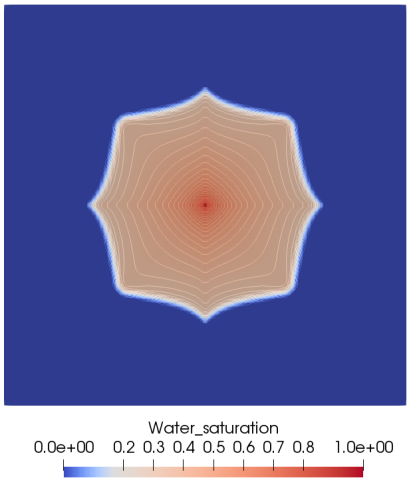
$$\kappa_{r,w}(s) = s^4, \qquad \kappa_{r,o}(1-s) = (1-s)^2, \tag{4.9}$$

from which it follows that

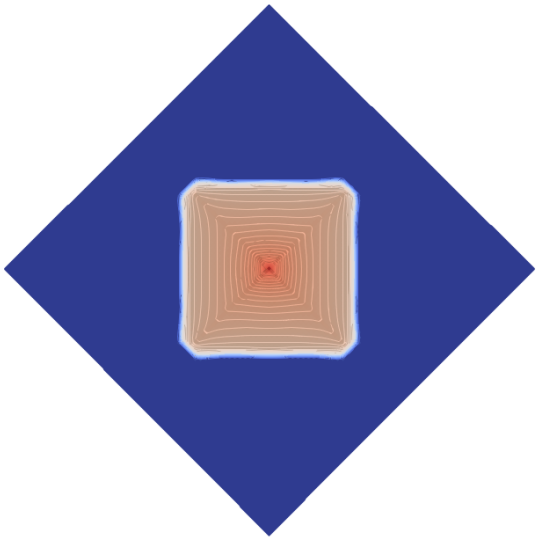
$$f(s) = \frac{Ms^4}{Ms^4 + (1-s)^2}.$$

We simulate a period of $T = 200$ days.

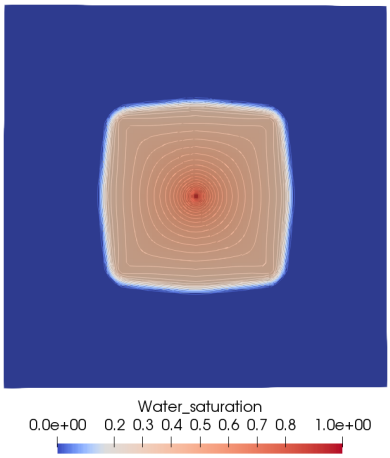
The numerical results obtained with the different schemes and on the two meshes are shown in Figures 13 and 14. We observe that the water saturation profiles obtained using the five-point scheme are very different in the two meshes due to the GOE whereas the ones obained using the nine-point schemes, presented in this paper, are very similar. These observations can also be done on the water production rates at the producers that are presented in Figure 15. Indeed, with the 5P scheme, the curves are not identical between the parallel and the diagonal meshes and in particular, the breakthrough times do not occur at the same time. However, because of the symmetry of the problem, we should obtain the same curves between the two meshes and it is what we observe with the 9P schemes.



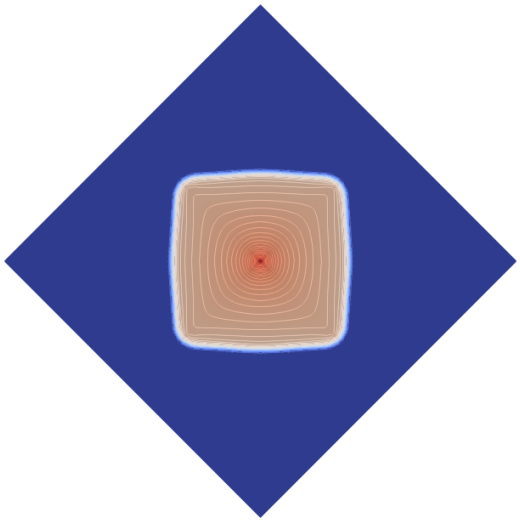
(a) 5P scheme and diagonal mesh.



(b) 5P scheme and parallel mesh.

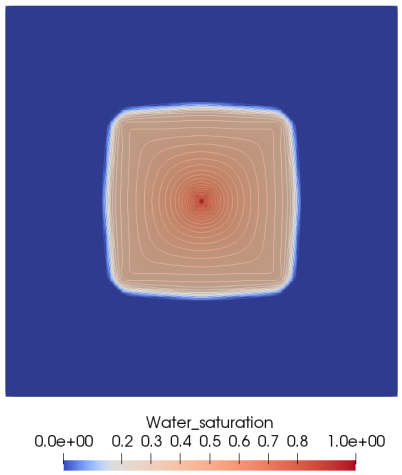


(c) 9P1s scheme with $\theta = 1/12$ and diagonal mesh.

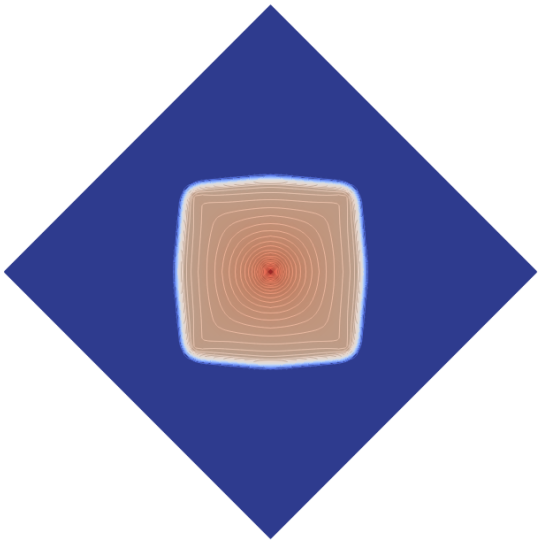


(d) 9P1s scheme with $\theta = 1/12$ and parallel mesh.

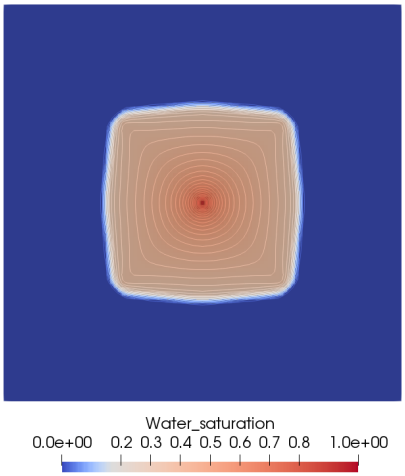
Figure 13: Water saturation fields at $T = 200$ days for the five-well problem using the 5P scheme (panels a–b) and 9P1s scheme (panels c–d) with $\theta = \frac{1}{12}$.



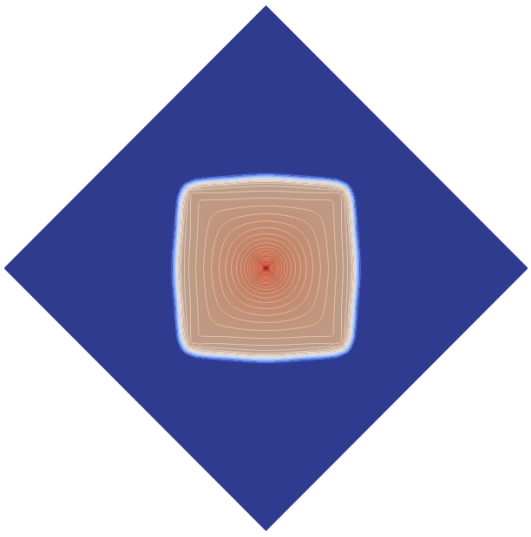
(a) 9P1s scheme with $\theta^\flat$ and diagonal mesh.



(b) 9P1s scheme with $\theta^\flat$ and parallel mesh.

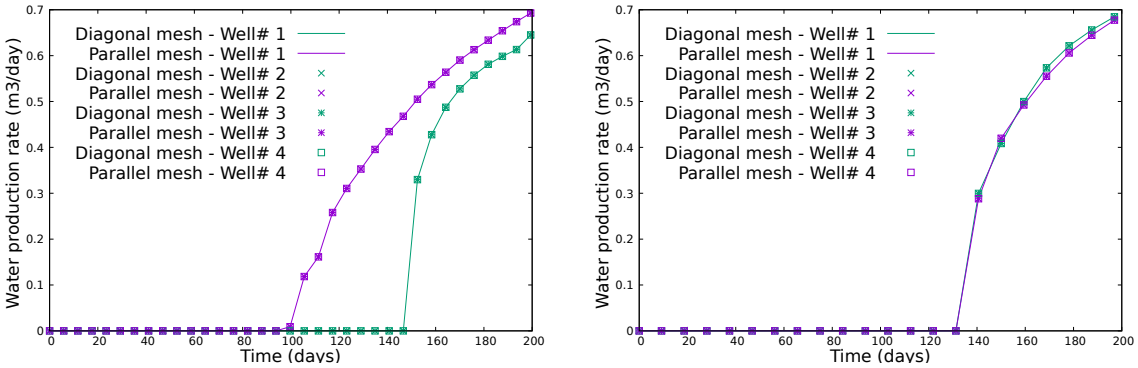


(c) 9P2s scheme with $(\theta_x^\flat, \theta_y^\flat)$ and diagonal mesh.



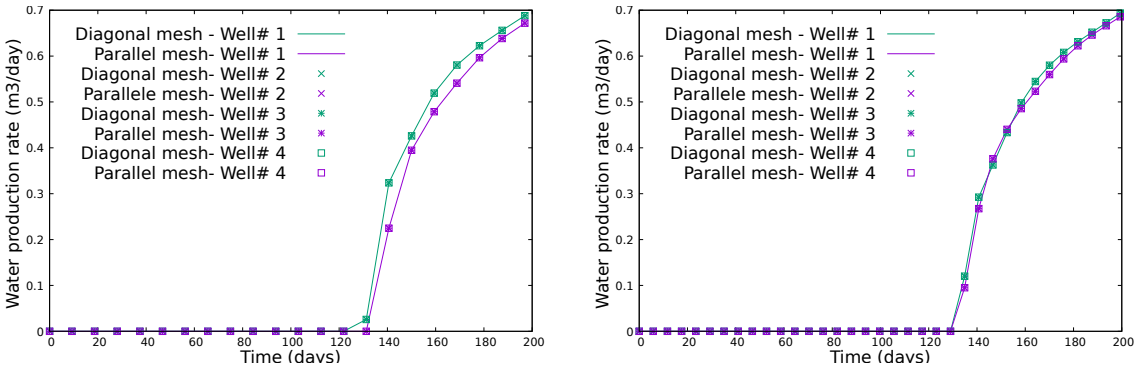
(d) 9P2s scheme with $(\theta_x^\flat, \theta_y^\flat)$ and parallel mesh.

Figure 14: Water saturation fields at $T = 200$ days for the five-well problem using the 9P1s scheme with $\theta^\flat$ (panels a–b) and the 9P2s scheme with $(\theta_x^\flat, \theta_y^\flat)$ (panels c–d).



(a) 5P scheme.

(b) 9P1s scheme with $\theta = \frac{1}{12}$.



(c) 9P1s scheme with $\theta^\flat$.

(d) 9P2s scheme with $(\theta_x^\flat, \theta_y^\flat)$.

Figure 15: Water production curves for the four producers on the two meshes using the 5P scheme (a), the 9P1s scheme with $\theta = 1/12$ (b), the 9P1s scheme with $\theta^\flat$ (c) and the 9P2s scheme with $(\theta_x^\flat, \theta_y^\flat)$ (d).

5 Conclusion

The GOE is an unavoidable consequence of discretization on Cartesian grids. However, under adverse mobility ratios, it is so much amplified that the numerical results produced by the classical 5P scheme become unacceptable for reservoir engineers. In this paper, we have designed a mathematical formalism based on Fourier error analysis in order to define a notion of directional error and to minimize the anisotropy of the computed solutions. Applied to two families of numerical schemes depending on tuning parameters, our paradigm has given rise to two schemes —9P1s and 9P2s— to remedy the GOE.

The first family 9P1s depends on one scalar parameter and provides a unified framework that includes several well-known schemes. Depending on whether the optimization is carried out with respect to pressure or saturation, the optimal and suboptimal values for the tuning parameter happen to be those formerly suggested by various authors in a more or less heuristic way. In this respect, our approach has brought a rigorous justification to these previous works. The second family 9P2s depends on two scalar parameters and is, to our knowledge, a novel construction. The introduction of a second parameter enables us to further reduce the GOE, as testified by the good results of two numerical tests.

To be of practical interest to real simulations, our approach must of course be broadened to take into account more sophisticated physics, such as capillary pressure, anisotropic permeability tensor, gravity effect and polymer injection. Another direction for future research would be to extend the promising ideas of this paper to more complex, non-orthogonal but structured meshes such as CPG (Corner Point Geometry), where it still makes sense to talk about the GOE.

References

[1] K. Aziz and A. Settari, *Petroleum Reservoir Simulation*, Applied Science Publishers, London, 1979.

[2] J. Bohbot, Q. H. Tran, A. Velghe, and N. Gillet, *A multi-dimensional spatial scheme for massively parallel compressible turbulent combustion simulation*, in 5th European Conference on Computational Fluid Dynamics, 14-17 June, J. C. F. Pereira and A. Sequeira, eds., Lisbon, Portugal, 2010, ECCOMAS CFD.

[3] B. Braconnier, C. Preux, É. Flauraud, Q.-H. Tran, and C. Berthon, *An analysis of physical models and numerical schemes for polymer flooding simulations*, Comput. Geosci., 21 (2017), pp. 1267–1279.

[4] C. W. Brand, J. E. Heinemann, and K. Aziz, *The grid orientation effect in reservoir simulation*, in SPE Reservoir Simulation Symposium, 17-20 February, Anaheim, California, Anaheim, California, February 1991, Society of Petroleum Engineers, Society of Petroleum Engineers.

[5] R. H. Brooks and A. T. Corey, *Hydraulic properties of porous media and their relation to drainage design*, Trans. ASAE, 7 (1964), pp. 26–28.

[6] S. E. Buckley and M. C. Leverett, *Mechanism of fluid displacement in sands*, Trans. AIME, 146 (1942), pp. 107–116.

[7] G. Chavent and J. Jaffré, *Mathematical Models and Finite Elements for Reservoir Simulation: Single Phase, Multiphase and Multicomponent Flows through Porous Media*, vol. 17 of Studies in Mathematics and its Applications, North-Holland, Amsterdam, 1986.

[8] K. H. Coats and A. D. Modine, *A consistent method for calculating transmissibilities in nine-point difference equations*, in SPE Reservoir Simulation Symposium, 15-18 November, San Francisco, California, 1983, Society of Petroleum Engineers. SPE-12248-MS.

[9] P. Colella, *Multidimensional upwind methods for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 171–200.

[10] A. T. Corey, *Mechanics of heterogenous fluids in porous media*, Water Resources Publications, Fort Collins, Colorado, 1977.

[11] B. Corre, R. Eymard, and L. Quettier, *Applications of a thermal simulator to field cases*, in SPE Annual Technical Conference and Exhibition, 16-19 September, Houston, Texas, 1984, Society of Petroleum Engineers. SPE-13221-MS.

[12] B. Després and F. Lagoutière, *Genuinely multi-dimensional non-dissipative finite-volume schemes for transport*, Int. J. Appl. Math. Comput. Sci., 17 (2007), pp. 321–328.

[13] D. Y. Ding, *Étude des effets d'orientation de maillage en simulation de réservoir*, Technical report 38232, IFP, Rueil-Malmaison, August 1990.

[14] J. Droniou, *Finite volume schemes for diffusion equations: introduction to and review of modern methods*, Math. Models Methods Appl. Sci., 24 (2014), pp. 1575–1619.

[15] R. Eymard, C. Guichard, and R. Masson, *Grid orientation effect in coupled finite volume schemes*, IMA J. Numer. Anal., 33 (2013), pp. 582–608.

[16] J. C. Frauenthal, R. B. di Franco, and B. F. Towler, *Reduction of grid-orientation effects in reservoir simulation with generalized upstream weighting*, SPE Journal, 25 (1985), pp. 902–908. SPE-11593-PA.

[17] G. Gagneux and M. Madaune-Tort, *Analyse mathématique de modèles non linéaires de l'ingénierie pétrolière*, vol. 22 of Mathématiques et Applications, Springer, Berlin, 1995.

[18] C. Guichard, *Schémas volumes finis sur maillages généraux en milieux hétérogènes anisotropes pour les écoulements polyphasiques en milieux poreux*, PhD thesis, Université Paris-Est, 2011.

[19] F. S. V. Hurtado, C. R. Maliska, and A. F. C. da Silva, *On the factors influencing the grid orientation effect in reservoir simulation*, in 19th International Congress of Mechanical Engineering, November 5-7, Brasilia, 2007, ABCM, pp. 5–9.

[20] E. Keilegavlen, J. E. Kozdon, and B. T. Mallison, *Multidimensional upstream weighting for multiphase transport on general grids*, Comput. Geosci., 16 (2012), pp. 1021–1042.

[21] J. E. Kozdon, B. T. Mallison, and M. G. Gerritsen, *Multidimensional upstream weighting for multiphase transport in porous media*, Comput. Geosci., 15 (2011), pp. 399–419.

[22] J. E. Kozdon, B. T. Mallison, M. G. Gerritsen, and W. H. Chen, *Multi-D upwinding for multi phase transport in porous media*, in SPE Reservoir Simulation Symposium, 2-4 February, The Woodlands, Texas, February 2009, Society of Petroleum Engineers. SPE-119190-MS.

[23] K. Laurent, *Étude de nouveaux schémas numériques pour la simulation des écoulements à rapport de mobilités défavorable dans un contexte EOR*, PhD thesis, Université Paris-Saclay, 2019.

[24] M. Muskat, *The flow of homogeneous fluids through porous media*, International Series in Physics, McGraw-Hill, New York, 1937.

[25] D. W. Peaceman, *Interpretation of well-block pressures in numerical reservoir simulation*, SPE Journal, 18 (1978), pp. 183–199. SPE-6893-PA.

[26] ——, *Interpretation of well-block pressures in numerical reservoir simulation with nonsquare grid blocks and anisotropic permeability*, SPE Journal, 23 (1983), pp. 531–543. SPE-10528-PA.

[27] C. Preux and F. McKee, *Study and approximation of IMPES stability: the CFL criteria*, in Finite Volumes for Complex Applications VI: Problems & Perspectives, J. Fořt, J. Fürst, J. Halama, R. Herbin, and F. Hubert, eds., vol. 4 of Springer Proceedings in Mathematics, Springer, Prague, June 2011, pp. 713–721.

[28] P. L. Roe and D. Sidilkover, *Optimum positive linear schemes for advection in two and three dimensions*, SIAM J. Numer. Anal., 29 (1992), pp. 1542–1568.

[29] P. C. Shah, *A nine-point finite difference operator for reduction of the grid orientation effect*, in SPE Reservoir Simulation Symposium, 15-18 November, San Francisco, California, San Francisco, 15-18 November 1983, Society of Petroleum Engineers, pp. 171–174. SPE-12251-MS.

[30] G. S. Shiralkar and R. E. Stephenson, *A general formulation for simulating physical dispersion and a new nine-point scheme*, SPE Reserv. Eng., 6 (1991), pp. 115–120. SPE-16975-PA.

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

[32] M. R. Todd, P. M. O'dell, and G. J. Hirasaki, *Methods for increased accuracy in numerical reservoir simulators*, SPE Journal, 12 (1972), pp. 515–530. SPE-3516-PA.

[33] J. L. Yanosik and T. A. McCracken, *A nine-point, finite-difference reservoir simulator for realistic prediction of adverse mobility ratio displacements*, SPE Journal, 19 (1979), pp. 253–262. SPE-5734-PA.