Mathematical complexities in porous media flow

Citation for published version (APA):

Document status and date:
Published: 26/09/2019

Document Version:
Typeset version in publisher’s lay-out, without final page, issue and volume numbers

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl

providing details and we will investigate your claim.
Mathematical Complexities in Porous Media Flow

Koondanibha Mitra

PhD Thesis
Award of degree - Technische Universiteit Eindhoven: Doctor
Award of degree - Universiteit Hasselt: Doctor in Science: Mathematics

Promoters
prof.dr.ir. B. Koren (Technische Universiteit Eindhoven)
prof.dr. I.S. Pop (Universiteit Hasselt and Universitetet i Bergen)
Mathematical Complexities in Porous Media Flow

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven,
op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens,
voor een commissie aangewezen door het College voor Promoties,
in het openbaar te verdedigen op
donderdag 26 september 2019 om 16:00 uur

door

Koondanibha Mitra

geboren te Shibpur, India
Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter: prof.dr. J.J. Lukkien  
1e promotor: prof.dr.ir. B. Koren  
2e promotor: prof.dr. I.S. Pop (Universiteit Hasselt en Universitetet i Bergen)  
externe leden: prof.dr. P. De Maesschalck (Universiteit Hasselt)  
prof.dr.ir. C. Vuik (Technische Universiteit Delft)  
prof.dr. F.A. Radu (Universitetet i Bergen)  
leden TU/e: prof.dr.ir. C.J. van Duijn (Technische Universiteit Eindhoven en Universiteit Utrecht)  
prof.dr.ir. E.H. van Brummelen  
adviseur: dr. J. Schütz (Universiteit Hasselt)

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.
Abstract

Multiphase flow through porous media plays an important role in many practical applications, from groundwater modelling, oil and gas recovery to CO\textsubscript{2} sequestration. In the current work, we address two challenges related to accurate modelling and simulation of such processes. The first is to incorporate non-equilibrium effects such as hysteresis and dynamic capillarity in the models. Experiments have shown that under certain circumstances, phenomena like saturation overshoot and finger formation occur, that cannot be explained by the standard (equilibrium) models. Hence, an extension of these models needs to be considered.

The second is to develop fast, stable and preferably simple numerical techniques that solve the highly nonlinear and possibly degenerate equations governing flow in the extremely heterogeneous porous domains of the real world. Accordingly, this work is divided into two parts:

Non-equilibrium effects

First, a new model is proposed for hysteresis in capillary pressure, which extends and improves the play-type hysteresis model. It is shown that this model is physically consistent and approximates experimentally obtained hysteresis curves. It is then used to solve the problem of horizontal redistribution of water and air, demonstrating that ‘unconventional’ flow, predicted earlier in literature, does indeed occur in certain cases. To follow-up, we show that the model is mathematically well-posed.

Next, gravity-driven infiltration of water into relatively dry soil is considered when the wetting front has the form of a downward propagating travelling wave, i.e., the wetting front moves at a constant speed and shape. We consider various cases with increasing complexity in a number of chapters. In the first, we study the behavior of fronts when either hysteresis or dynamic capillarity is included. In the second chapter, both effects are included simultaneously and both the play-type and the extended play-type models, mentioned in the previous paragraph, are considered. The existence of travelling waves is proved and criteria for the occurrence of
overshoots and the system to reach full saturation are made precise. The techniques developed are further used in the third chapter to describe viscous fingering and to derive the propagation speed of the fingers.

Finally, fronts are analysed for the two-phase case in a very general setting where the relative permeabilities, as well as the capillary pressure, are hysteretic and dynamic capillary effect is included. Existence of all possible travelling wave solutions is shown and a number of qualitative properties are established. The travelling wave solutions are then used to derive admissibility conditions for shocks in the hyperbolic limit. The entropy solutions derived in this way are much broader compared to the standard entropy solutions of the Buckley-Leverett equation since they can be non-monotone and have multiple shocks. These results are used to explain experimental observations such as non-monotone saturation profiles and stable saturation plateaus, which were previously not well-understood.

**Numerical methods**

A linear domain decomposition scheme is proposed for heterogeneous and in particular, layered porous media. Apart from being parallelizable, it is unconditionally convergent for a mild restriction on the time step. Moreover, it is, in general, more stable and better conditioned than standard monolithic schemes such as the Newton or the Picard scheme, while being comparable in speed.

The issue of nonlinearity is handled in the following chapter where a linear iterative scheme is proposed for solving the nonlinear diffusion equations that arise in porous flow problems. Being a modified version of the L-scheme, it converges linearly for a mild restriction on the time step, having convergence rate proportional to an exponent of the time step size. The convergence is also guaranteed for degenerate cases. This makes it faster than both the L-scheme and the Picard scheme and more stable than the Newton and the Picard scheme. Numerical results are provided that support the analytical findings.

Finally, a mixed finite element method is proposed for the two phase flow model with dynamic capillarity effect. Error estimates are derived showing that the scheme is first order in both space and time. The numerical results support our conclusion.
Contents

Abstract v

1 Introduction 1
  1.1 Standard flow model ........................................ 3
    1.1.1 Two-phase flow ....................................... 4
    1.1.2 Unsaturated flow ...................................... 6
    1.1.3 Non-dimensionalization and effective equations ........ 6
  1.2 Non-equilibrium effects .................................... 9
    1.2.1 Hysteresis ............................................ 9
    1.2.2 Dynamic capillarity .................................. 13
    1.2.3 The combined model and some mathematical techniques ... 14
  1.3 Numerical challenges ...................................... 17
    1.3.1 Heterogeneity ....................................... 17
    1.3.2 Nonlinearity and degeneracy .......................... 18
    1.3.3 Numerical methods for non-equilibrium effects ....... 20
  1.4 Overview of the chapters .................................. 21
    1.4.1 Part I: non-equilibrium effects ....................... 21
    1.4.2 Part II: numerical methods ............................ 23

2 Hysteresis and horizontal redistribution in porous media 27
  2.1 Introduction ............................................. 27
    2.1.1 Similarity models ................................... 29
    2.1.2 Play-type hysteresis .................................. 29
    2.1.3 Interfacial area model ................................ 30
  2.2 Extended play-type hysteresis model ....................... 31
  2.3 Redistribution ............................................ 35
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.1</td>
<td>General set up</td>
<td>35</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Possible initial conditions</td>
<td>36</td>
</tr>
<tr>
<td>2.4</td>
<td>Self-similar solutions</td>
<td>38</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Reformulation of the problem</td>
<td>38</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Classification of possible solutions in the half columns $\mathbb{R}^-$ and $\mathbb{R}^+$</td>
<td>39</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Direction of flow</td>
<td>42</td>
</tr>
<tr>
<td>2.4.4</td>
<td>Construction of combined solution for initial conditions on the main drainage and imbibition curves: $E_r \in \partial H^{dr}$, $E_l \in \partial H^{im}$</td>
<td>43</td>
</tr>
<tr>
<td>2.4.5</td>
<td>Construction of combined solution for arbitrary initial conditions: $E_l, E_r \in H \cup \partial H^{im} \cup \partial H^{dr}$</td>
<td>46</td>
</tr>
<tr>
<td>2.5</td>
<td>Numerical study</td>
<td>51</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Numerical scheme</td>
<td>52</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Numerical results</td>
<td>53</td>
</tr>
<tr>
<td>2.6</td>
<td>Conclusion</td>
<td>55</td>
</tr>
<tr>
<td>3</td>
<td>Existence and properties of the solutions of the extended play-type hysteresis model</td>
<td>57</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>57</td>
</tr>
<tr>
<td>3.2</td>
<td>Mathematical formulation</td>
<td>59</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Problem statement and assumptions</td>
<td>60</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Relation between the regularised EPH model and $(\mathcal{P}s)$</td>
<td>61</td>
</tr>
<tr>
<td>3.3</td>
<td>Existence of solutions</td>
<td>65</td>
</tr>
<tr>
<td>3.4</td>
<td>Boundedness and well-posedness of (EPH)</td>
<td>69</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Bounds on $u$ and $v$</td>
<td>69</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Existence of solutions of (EPH)</td>
<td>71</td>
</tr>
<tr>
<td>3.5</td>
<td>Behaviour when $\tau \to 0$</td>
<td>75</td>
</tr>
<tr>
<td>3.6</td>
<td>Conclusion and outlook</td>
<td>76</td>
</tr>
<tr>
<td>4</td>
<td>Travelling wave solutions for the Richards equation incorporating non-equilibrium effects in the capillary pressure</td>
<td>77</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>77</td>
</tr>
<tr>
<td>4.2</td>
<td>Mathematical formulation</td>
<td>79</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Basic equations</td>
<td>79</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Travelling wave formulation</td>
<td>81</td>
</tr>
<tr>
<td>4.3</td>
<td>Capillary hysteresis</td>
<td>83</td>
</tr>
<tr>
<td>4.3.1</td>
<td>The case $S_T = 1$</td>
<td>84</td>
</tr>
<tr>
<td>4.3.2</td>
<td>The case $S_T &lt; 1$</td>
<td>89</td>
</tr>
<tr>
<td>4.4</td>
<td>Dynamic capillarity</td>
<td>96</td>
</tr>
<tr>
<td>4.4.1</td>
<td>General behaviour of the orbits</td>
<td>97</td>
</tr>
<tr>
<td>4.4.2</td>
<td>The case when $f \notin L^1(0,1)$</td>
<td>108</td>
</tr>
<tr>
<td>4.5</td>
<td>Numerical results</td>
<td>112</td>
</tr>
</tbody>
</table>
4.5.1 Numerical scheme ........................................... 112
4.5.2 Iterative scheme ........................................... 113
4.6 Conclusion ................................................... 119

5 Travelling waves: the combined case of capillary hysteresis and dynamic capillarity 121
5.1 Introduction .................................................. 121
5.2 Travelling waves with combined pressure (5.2) .................... 124
  5.2.1 The case $f(S) = 1$ ........................................ 126
  5.2.2 The case $f \in L^1(0,1)$ .................................... 137
  5.2.3 The case $f \notin L^1(0,1)$ .................................... 138
5.3 Travelling waves with combined pressure (5.4) .................... 140
5.4 Numerical results ............................................ 150
5.5 Conclusion ................................................... 153

6 Traveling wave solutions for the Richards equation with hysteresis 155
6.1 Introduction .................................................. 155
6.2 Reformulation of the traveling wave equations ....................... 159
6.3 Existence of hysteretic traveling wave solutions .................... 161
  6.3.1 Construction of normalized orbits for given $c$ ................. 162
  6.3.2 Construction of $c$ to satisfy the flux condition ............... 168
6.4 Discussion .................................................... 170

7 Fronts in two-phase porous media flow problems: the effects of hysteresis and dynamic capillarity 173
7.1 Introduction .................................................. 173
7.2 Mathematical model .......................................... 174
  7.2.1 Governing equations ...................................... 175
  7.2.2 Modelling hysteresis and dynamic capillarity ................. 176
  7.2.3 Effective equations ...................................... 179
  7.2.4 Travelling wave formulation ............................. 181
7.3 No relative permeability hysteresis and small $N_g$ (Scenario A) .... 183
  7.3.1 Preliminaries ............................................. 183
  7.3.2 Problem (TW) with $S_B < S_T \leq \bar{S}$ ....................... 194
  7.3.3 $(S_T, \tau) \notin \mathcal{A}$ ..................................... 197
  7.3.4 Entropy solutions to hyperbolic conservation laws ............ 199
7.4 Hysteretic relative permeabilities and small $N_g$ (Scenario B) .... 201
  7.4.1 Case (i): $S_B < S_T \leq \bar{S}_i$ .......................... 203
  7.4.2 Case (ii): $\bar{S}_d \leq S_T < S_B$ and stability of plateaus .... 204
  7.4.3 Entropy solutions ....................................... 206
7.5 Numerical results ............................................ 207
  7.5.1 Numerical results for Scenario A ........................... 208
7.5.2 Numerical results for Scenario B ........................................... 211
7.6 Conclusion ............................................................................. 214

8 A linear domain decomposition method for partially saturated flow in
layered porous media .................................................................. 219
  8.1 Introduction ................................................................. 219
  8.2 Problem formulation and iterative scheme ......................... 223
    8.2.1 Problem formulation ................................................. 223
    8.2.2 The LDD iterative scheme ....................................... 224
  8.3 Analysis of the scheme ................................................... 227
    8.3.1 Intuitive justification of the L-scheme ....................... 228
    8.3.2 Convergence of the scheme ..................................... 229
  8.4 Numerical experiments ................................................... 233
    8.4.1 Results for a case with known exact solution ............ 235
    8.4.2 Results for a realistic case with van Genuchten parameters 240
    8.4.3 Time performance .................................................. 242
    8.4.4 Parameter dependence and key features .................. 245
    8.4.5 Multi domain and parallelisation study .................... 248
  8.5 Conclusion ........................................................................... 250

9 A linear iterative scheme for nonlinear diffusion problems ........ 251
  9.1 Introduction ................................................................. 251
  9.2 The general problem and linearisation schemes ................. 253
    9.2.1 Time-discrete formulation and properties of functions .... 253
    9.2.2 Standard linearisation techniques ............................. 255
  9.3 The modified L-scheme ................................................... 257
    9.3.1 Quasilinear equations with linear diffusivity and flux .... 257
  9.4 General nonlinear diffusion equation .................................. 261
  9.5 Numerical results .......................................................... 268
  9.6 Conclusion ........................................................................... 275

10 Mixed finite element discretization for the two-phase flow model with
dynamic capillarity: error estimates ..................................... 277
  10.1 Introduction ................................................................. 277
  10.2 Notations and assumptions ............................................. 279
  10.3 Mixed formulation and error estimates ............................. 281
    10.3.1 Existence and uniqueness ....................................... 281
    10.3.2 Error estimates ..................................................... 286
  10.4 Numerical results .......................................................... 292
Chapter 1

Introduction

“Thus the partial differential equation entered theoretical physics as a handmaid, but has gradually become mistress.” ~ A. Einstein

The study of flow through porous materials finds its application in many different fields, including subsurface flow [19]; biological processes such as circulation, transport in tissue, biofilm growth [244]; industrial processes such as catalysis, reactive transport [256], absorption [74]. To elaborate on the societal relevance of such studies, take the flow through the subsurface (underground geological structures) as an example. The challenges associated with it are of paramount importance in the energy and the environmental sector. Extraction of oil and gas, currently responsible for more than 50% of the world’s energy production, is an example of such a process. However, despite decades of advances in enhanced oil recovery, most of the oil in a reservoir cannot be extracted [148] due to economic infeasibility. For groundwater systems, constant monitoring of pollutants, mineral deposition, and water level [19, 20, 204] provide further motivation for research. The issue of climate change too is closely connected with that of geological storage of CO\(_2\) [178] since it is projected that over 100 Giga-tons of CO\(_2\) has to be injected into the subsurface by 2050 to restrict the temperature increase to 2°C [1]. However, the viability of such a process on a large scale is still debated.

From a technical point of view, some of the major hurdles in addressing the problems mentioned above stem from errors introduced in the modelling and the simulation stages. Consider multiphase flow through the subsurface for instance. Examples include oil and gas in case of petroleum extraction, air and water for groundwater systems and supercritical CO\(_2\) and brine for CO\(_2\) sequestration. In this case, the complexity of such systems in the pore scale and the much larger spatial dimension of the domains force one to use an upscaled model [115]. The stan-
The standard model used for this purpose is the Darcy equation coupled with some form of capillary pressure-saturation relationship [19, 115]. Although upscaling techniques like homogenization [240] have been used for many decades in the one phase context [8, 43, 256], a rigorous upscaling of the full-scale multiphase system is yet to be achieved, see [164, 218] for partial results. This implies that the standard model is a useful approximation rather than a model derived based on first principles, and it introduces significant limitations on its predictive capabilities. To illustrate this issue, consider the phenomenon of hysteresis, which plays a major role in viscous fingering [103] (see Figure 1.1(a)), CO$_2$ sequestration [71], trapping of oil between heterogeneous layers [129] and many other important processes [201]. However, it is often excluded from the standard framework due to numerical and analytical challenges involved in its implementation [44, 182].

A similar case can be made for the dynamic capillarity effect, which is known to cause overshoots inside the porous domain if the flow rate is high enough [72], also shown in Figure 1.1(b). This is a clear deviation from the predictions obtained through the standard model [83] which would only predict profiles that are mono-

![Experimental evidence of deviation from the standard model.](image)
1.1 Standard flow model

In the macroscopic description of multiphase flow in porous media, one uses quantities that are averaged over a representative elementary volume \[19, 178\]. For a given phase \( \alpha \), the primary unknowns are saturation \( S_\alpha \), volumetric flux \( Q_\alpha \) [m \cdot s\(^{-1}\)] and pressure \( p_\alpha \) [Pa]. Here, saturation \( S_\alpha \in [0,1] \) is defined as the fraction of the pore space occupied by phase \( \alpha \). Conservation of mass of phase \( \alpha \) demands

\[
\partial_t (\phi \rho_\alpha S_\alpha) + \nabla \cdot (\rho_\alpha Q_\alpha) = R_\alpha.
\]  

(1.1)
Here, $\phi$ is the porosity, defined as the fraction of the volume occupied by pores; $\rho_\alpha$ is the density of the phase and $R_\alpha$ is the source term which is non-zero e.g. in case of reaction, absorption or phase transition. It is assumed implicitly that the porous matrix does not deform.

The Darcy law [19,115,178] relates $\vec{Q}_\alpha$ and $p_\alpha$ with

$$\vec{Q}_\alpha = -\frac{k_{r\alpha}}{\mu_\alpha} \mathbf{K}[\nabla p_\alpha - \rho_\alpha \vec{g}].$$

(1.2)

Here, $\mathbf{K}[\text{m}^2]$ is the absolute permeability. It is a scalar for isotropic domains and a symmetric positive definite tensor otherwise. Generally, $\mathbf{K}$ is considered to be an intrinsic property of the medium that depends only on spatial coordinates. The term $k_{r\alpha} \in [0,1]$ is known as the relative permeability. It represents the fraction of absolute permeability available for phase $\alpha$. The viscosity of the phase $\alpha$ is given in (1.2) by $\mu_\alpha [\text{Pa}\cdot\text{s}]$.

Since (1.1)-(1.2) is not a closed system, one requires more equations. We make this precise for the specific, but important, case of two-phase flow.

### 1.1.1 Two-phase flow

When dealing with two phases, we distinguish between a wetting ($\alpha = w$) and a non-wetting ($\alpha = n$) phase. The wetting phase refers to the phase having more affinity towards the porous matrix. Thus, in water-oil and water-air systems, it represents water. Observe that, by definition

$$S_n + S_w = 1.$$  

(1.3)

However, the relation between $S_\alpha$, $k_{r\alpha}$ and $p_\alpha$ still needs to be prescribed. For this purpose, the standard model assumes the following [19,20,115,178]:

$$p_n - p_w =: p_c = p_c(S_w) \text{ and } k_{r\alpha} =: k_{r\alpha}(S_w).$$

(1.4)

Here, $p_c(\cdot)$ is known as the capillary pressure function. Such relationships are based on experimental observations, which are made under so-called equilibrium conditions. This means that measurements are done only after the system has reached equilibrium. Physical justifications for assuming (1.4) are given in [19,178] and limitations of such assumptions are stated. The existence of weak solutions of the model (1.1)-(1.4) is shown in [145] for strictly positive permeabilities and in [144] for vanishing permeabilities with initial and boundary conditions bounded away from the points where the permeabilities become zero. Uniqueness is shown in [59] and the existence of weak solutions of the model for the simple dual porosity case is shown in [11].
For practical computation of $p_c$ and $k_{r\alpha}$ in (1.4) one generally uses the effective saturation

$$S_e := \frac{S_w - S_{wr}}{1 - S_{wr} - S_{nr}}.$$

(1.5)

Here, $S_{nr} \in [0, 1]$ is the residual saturation of phase $\alpha$, defined as the minimum saturation of $\alpha$ that can be obtained by wetting or drying process alone. Observe that, $S_e$ spans the entire interval $[0, 1]$. Two of the most commonly used models relating $p_c$ and $k_{r\alpha}$ with $S_e$ are the Brooks-Corey model [45],

$$\begin{cases}
    p_c := P_b S_e^{\frac{1}{\lambda}} , \\
    k_{rw} := S_e^{2+\frac{3}{\lambda}} , \\
    k_{rn} := (1 - S_e^2) \left(1 - S_e^{\frac{2}{\lambda}}\right),
\end{cases}$$

(1.6)

and the van Genuchten model [259],

$$\begin{cases}
    p_c := P_r \left(S_e^{-\frac{1}{q}} - 1\right)^{-\frac{1}{q}} , \\
    k_{rw} := S_e^{\gamma_1} \left(1 - \left(1 - S_e^{1+q}\right)^q\right) , \\
    k_{rn} := (1 - S_e)^{\gamma_2} \left(1 - S_e^{1+q}\right)^{2q}.
\end{cases}$$

(1.7)

Figure 1.2: The functions $p_c$ (scaled), $k_{rw}$ and $k_{rn}$ for (left) Brooks-Corey model with $\lambda = .75$, and (right) van Genuchten model with $\ell = 2$. The parameters are taken from [152] and give the closest match with each other. Observe that, for the Brooks-Corey model $p_c(1 - S_{nr}) = P_b > 0$

The positive constants $P_b$, $\lambda$, $\gamma_{1/2}$, $P_r$, $q$, $\ell$ are characteristics of the medium calculated using experimental data. Generally, $\gamma_1 = \frac{1}{2}$, $\gamma_2 = \frac{1}{3}$ and $q = 1 - \frac{1}{\ell}$ are used in (1.7) [115]. The relation between parameters of both models is explained in [152]. Figure 1.2 shows the plots for $p_c$ (scaled), $k_{rw}$ and $k_{rn}$ using expressions (1.6) (left)
and (1.7) (right) for parameters related as in [152]. An important observation is that $p_c(1 - S_{rn}) = 0$ for van Genuchten model, whereas, $p_c(1 - S_{rn}) > 0$ for Brooks-Corey model. This positive end pressure is called the entry pressure. It may cause trapping of the non-wetting (oil) phase in heterogeneous media [165, 220, 252]. In subsequent chapters, the expressions (1.6)-(1.7) and the properties of the curves, as shown in Figure 1.2, will be used multiple times. However, in each chapter we restate the properties of $p_c$ and $k_{rw}$ since each chapter builds on particular assumptions.

1.1.2 Unsaturated flow

A simplification of the model given by (1.1)-(1.4) arises when considering the flow of air and water in the vadose layer of the soil. This is known as unsaturated flow. Since, $\rho_w/\rho_n \approx 1000$ and $\mu_w/\mu_n \approx 100$ for water and air, air is assumed to be much more mobile, and thus, when in contact with the atmosphere, it is assumed to have constant atmospheric pressure throughout the domain. Hence, $p_n = 0$ is substituted into (1.1)-(1.4) which gives the Richards equation, reading

$$\frac{\partial}{\partial t}(\phi \rho_w S_w) = \nabla \cdot \left[ \frac{\rho_w}{\mu_w} k_{rw} K(\nabla p_w - \rho_w \vec{g}) \right] + R_w, \quad S_{wR} \leq S_w \leq 1 - S_{nr},$$

(1.8)

closed by the expressions

$$p_w = -p_c(S_w), \quad k_{rw} = k_{rw}(S_w).$$

(1.9)

For constant $\phi$, $\rho_w$ and $\mu_w$, (1.8)-(1.9) reduce to a nonlinear diffusion problem. Existence of solutions for this problem for the one dimensional case has been studied in [245, 253]. Weak solutions for the general case are shown to exist in [9, 10] and uniqueness is proved in [181].

1.1.3 Non-dimensionalization and effective equations

In this part we use dimensional scaling to derive effective equations that will be analysed in the subsequent chapters. For simplicity, we assume the following:

(A1) The flow is incompressible, i.e. $\rho_{\alpha}$ are constants.

(A2) Viscosity of the fluids remain constant, i.e. $\mu_{\alpha}$ are constants.

(A3) The porosity does not change with time and is bounded away from 0, i.e. $0 < \phi_m \leq \phi < 1$ for some constant $\phi_m > 0$.

(A4) The absolute permeability is a positive and bounded function, i.e. we assume $K = K$ with two constant $K_m, K_M > 0$ existing such that $0 < K_m \leq K \leq K_M$. This
1.1 Standard flow model

condition can be generalised to \( K = \{ K_{ij} \} \) being a \( d \times d \) matrix-valued function satisfying

\[
K_m \xi^2 \leq \sum_{i,j} \xi_i K_{ij} \xi_j \leq K_M \xi^2,
\]

for any vector \( \xi = \{ \xi_i \} \in \mathbb{R}^d \). We choose a scalar \( K \), however, for simplicity.

To non-dimensionalize we use reference quantities \( L_{\text{ref}}, t_{\text{ref}}, p_{\text{ref}} \) that depend on the problem, and define

\[
\tilde{x}^* := \frac{\tilde{x}}{L_{\text{ref}}}, \quad t^* := \frac{t}{t_{\text{ref}}}, \quad p_a^* := \frac{p_a}{p_{\text{ref}}}, \quad S_w^* := S_e.
\]

The two-phase flow model

Let the domain of interest have a characteristic length \( L_{\text{ref}} \) and the total flow (see below) at the influx boundary have a reference value \( v_{\text{ref}} \). If the two fluids have similar properties, i.e. \( \mu_n \sim \mu_w \) and \( \rho_n \sim \rho_w \), and the flow is advection dominated, then the following choice of reference quantities is usually made,

\[
t_{\text{ref}} = (1 - S_{wr} - S_{nr}) \frac{\phi_{\text{ref}} L_{\text{ref}}}{v_{\text{ref}}}, \quad p_{\text{ref}} = \sigma \sqrt{\phi_{\text{ref}} K_{\text{ref}}}. \tag{1.10}
\]

Here, \( \sigma \) is the surface tension between the two-fluids. The scaling for \( p_{\text{ref}} \) is known as the J-Leverett scaling [153] and generally \( P_b, P_r \sim p_{\text{ref}} \) in (1.6)-(1.7). Defining the non-dimensional numbers

\[
N_c := \frac{K_{\text{ref}} p_{\text{ref}}}{\mu_n v_{\text{ref}}}, \quad N_g := \frac{K_{\text{ref}} (\rho_w - \rho_n)}{|\hat{g}|}, \tag{1.11}
\]

and quantities

\[
\begin{align*}
&k_n := k_r n, \quad k_w := \frac{\mu_n}{\mu_w} k_r w, \quad \kappa = \frac{K}{K_{\text{ref}}}, \\
&\hat{g} = \frac{\hat{g}}{|\hat{g}|}, \quad \hat{g}_n := \frac{\rho_n \hat{g}}{\rho_w - \rho_n}, \quad \hat{g}_w := \frac{\rho_w \hat{g}}{\rho_w - \rho_n},
\end{align*}
\]  

we get the dimensionless equations

\[
(\text{TP}) \left\{ \begin{array}{l}
\phi \partial_t S_w = \nabla \cdot [k_w \kappa (N_c \nabla p_w - N_g \hat{g}_w)] + R_w, \\
-\phi \partial_t S_w = \nabla \cdot [k_n \kappa (N_c \nabla p_n - N_g \hat{g}_n)] + R_n.
\end{array} \right. \tag{1.13a}\tag{1.13b}
\]

The superscript * of the dimensionless quantities is omitted for ease of presentation. The numbers \( N_c \) and \( N_g \) are called capillary number and gravity number respectively. The terms \( R_n, R_w \) are scaled with the rest of the parameters. Realistic values of dimensional and scaled quantities are given in [161]. Equation (1.13) will be used multiple times in our study.
Adding (1.13a) and (1.13b) we further write

\[ \nabla \cdot \vec{Q} = R_n + R_w, \tag{1.14} \]

where

\[
\vec{Q} = -k_w \kappa (N_c \nabla p_n - N_g \vec{g} w) - k_n \kappa (N_c \nabla p_n - N_g \vec{g} n) \\
= -N_c (k_n + k_w) \kappa \nabla p_n - N_c k_n \kappa \nabla p_c + N_g \kappa (k_n \vec{g} n + k_w \vec{g} w), \tag{1.15}
\]

is the total flow. Moreover, \( p_c = p_n - p_w \) as in (1.4). If \( \vec{Q} \) is known, which is often the case when the problem is in one spatial dimension, equations (1.13a) and (1.13b) can be combined into

\[
(BL) \left\{ \begin{array}{l}
\phi \partial_t S_w + \nabla \cdot \left[ f_w \vec{Q} + N_g h_w \kappa \vec{g} + N_c h_w \kappa \nabla p_c \right] = R_w,
\end{array} \right. \tag{1.16}
\]

with

\[
f_w := \frac{k_w}{k_w + k_n}, \quad h_w := \frac{k_n k_w}{k_w + k_n} = k_n f_w. \tag{1.17}
\]

The quantities \( f_w, h_w \) are the so-called fractional flow functions. Equation (1.16) is the well-known Buckley-Leverett equation [20, 115, 178]. In particular, the limit \( N_c \to 0 \) is relevant as it gives the hyperbolic Buckley-Leverett equation

\[
(HBL) \left\{ \begin{array}{l}
\phi \partial_t S_w + \nabla \cdot \left[ f_w \vec{Q} + N_g h_w \kappa \vec{g} \right] = R_w,
\end{array} \right. \tag{1.18}
\]

where \( \vec{Q} \) satisfies (1.14). From the scaling, it is evident that (1.18) approximates (1.16) for large \( L_{\text{ref}} \). For a given \( \vec{Q} \), the solution to (1.18) in one spatial dimension is referred to as the Oleinik entropy solution [180] in the sense of Remark 1.1. The solution is composed of shocks, rarefaction waves and constant states [180]. This will be discussed in detail in a later chapter.

Remark 1.1 (Entropy solutions). If a pair \( (S_w^N, p_c^N) \) solves (BL) for some \( N_c > 0 \) and \( S_w^N \to S_w \) in some sense, for \( N_c \to 0 \), then in our discussion \( S_w \) will be regarded as the entropy solution of the (HBL) (equation (1.18)) [150]. We use this limiting procedure to select the physically relevant solution to (HBL), which might otherwise admit multiple weak solutions.

The unsaturated flow model

In the case of unsaturated flow, see Section 1.1.2, we consider a slightly different scaling. Assuming that the gravity is the dominating force, we take

\[
u_{\text{ref}} := \frac{K_{\text{ref}}}{{\mu_w v_{\text{ref}}}}, \quad N_c := \frac{K_{\text{ref}} \rho_{\text{w}} \mu_{\text{w}}}{\mu_w \nu_{\text{ref}} L_{\text{ref}}}, \quad k := k_{\text{rw}}, \tag{1.19}
\]
1.2 Non-equilibrium effects

the scales $p_{\text{ref}}$, $t_{\text{ref}}$, $L_{\text{ref}}$ and $\kappa$ being as before. This gives the dimensionless Richards equation

$$
(\text{Ri}) \left\{ \begin{array}{l}
\phi \frac{\partial S_w}{\partial t} = \nabla \cdot \left[ k\kappa \left( N_c \nabla p_w - \hat{g} \right) \right] + R_w, \\
0 \leq S_w \leq 1.
\end{array} \right.
$$

(1.20)

Equation (1.20) is widely used in the following chapters.

Recalling (1.3), in the standard model it is assumed that

$$
p_c, k_n, k_w, k \text{ are functions of } S_w \text{ only.}
$$

(1.21)

However, we will see that this assumption holds only under certain conditions. Thus a revision/extension needs to be made.

1.2 Non-equilibrium effects

The assumption that $p_c$, $k_n$, $k_w$, $k$ depend only on $S_w$, gives accurate predictions when the system is close to equilibrium since they are determined from experiments only after the system has reached equilibrium. However, in many practical situations the saturation is changing rapidly or flow-reversal takes place. Then the behaviour of the flow cannot be described accurately by the assumption (1.21) as will be seen shortly after. The deviations are attributed to two separate effects, hysteresis and dynamic capillarity.

1.2.1 Hysteresis

A system is said to be hysteretic when its state is described not only by the state variables but also by its evolution [261]. Hysteresis plays an important role in fields like magnetism, electronics and wettability. In flow through porous media, hysteresis appears in two separate terms that we have introduced, capillary pressure and relative permeabilities.

Capillary hysteresis

It was first observed by Haines in 1930 [107] that the capillary pressure $p_c$ exhibits hysteresis, the observation thereafter being verified by Richards in 1931 [212]. Since then, there has been an abundance of experimental results outlining the nature of capillary hysteresis, [170, 197, 273] being some examples. It is observed that, when measured by a pure imbibition/infiltration/wetting process, the capillary pressure-saturation relationship follows a curve $p_c^{(i)}$ in the $p_c$-$S_w$ plane, which differs significantly from the curve $p_c^{(d)}$ obtained from pure drainage/drying experiments. Mathematically, this is expressed as

$$
\partial_t S_w \begin{cases} 
\geq 0 & \text{if } p_c = p_c^{(i)}(S_w), \\
\leq 0 & \text{if } p_c = p_c^{(d)}(S_w).
\end{cases}
$$

(1.22)
Intermediate curves, hereafter referred to as the *scanning curves*, are followed if the process switches from imbibition to drainage or from drainage to imbibition. Figure 1.3 illustrates this based on experimental results from [170]. The macroscopic effect of hysteresis is generally attributed to the complex fluid-structure interactions and hysteresis between advancing and receding contact angles of the fluid phases in the microscopic level [20, 107, 178]. A mathematical justification can be found in [217] and a pore network study is given in [210]. As mentioned earlier, hysteresis is very important in modelling processes such as CO$_2$ sequestration [71] since the injected CO$_2$ tends to get displaced once the injection ends. Similarly, in oil and gas recovery, important effects such as trapping, redistribution and finger formation, depend crucially on hysteresis [103, 129, 189, 201, 209].

A multitude of models have been proposed to incorporate hysteresis. An extensive study of different kind of models can be found in Chapter 2. Here, we give a brief overview. The independent and dependent domain models [172, 242] aim to describe hysteresis based on certain assumptions on the pore distribution functions [188, 196]. The Lenhard-Parker model [184] uses scaled versions of $p_c^{(i)}$ and $p_c^{(d)}$ as scanning curves. The models explained above are complicated to handle analytically and numerically because the scanning curves are calculated from the evolution history. However, the scanning curves obtained this way are close approximations to the realistic ones. Interfacial area models take a different approach and assume that hysteresis stems from a hidden state variable, i.e. interfacial area. Interfacial area is assumed to depend on saturation and pressure and satisfies an evolution equation [111, 113, 176]. However, the nature of hysteresis, as seen from
1.2 Non-equilibrium effects

experiments, cannot be fully described by this reformulation only [193].

For the most part in our study, we choose the play-type hysteresis model for describing hysteresis. It was proposed in [24] based on thermodynamic considerations. The scanning curves for this model are vertical, having constant saturations. This makes it simple and local as it is described by the expression,

$$ p_c = p_n - p_w \in p_c^+ (S_w) - p_c^- (S_w) \cdot \text{sign}(\partial_t S_w), \quad (1.23) $$

where

$$ p_c^+ = \frac{1}{2} (p_c^{(d)} + p_c^{(i)}), \quad p_c^- = \frac{1}{2} (p_c^{(d)} - p_c^{(i)}), \quad (1.24) $$

and \( \text{sign}(\cdot) \) is the multi-valued mapping

$$ \text{sign}(\zeta) = \begin{cases} 
1 & \text{for } \zeta > 0, \\
[-1,1] & \text{for } \zeta = 0, \\
-1 & \text{for } \zeta < 0. 
\end{cases} \quad (1.25) $$

Observe that (1.23) is consistent with (1.22) since \( p_c = p_c^{(i)} (S_w) \) forces \( \text{sign}(\partial_t S_w) = 1 \), implying \( \partial_t S_w \geq 0 \). The same is true for \( p_c^{(d)} \). If \( p_c \) is between \( p_c^{(i)} (S_w) \) and \( p_c^{(d)} (S_w) \), then \( \partial_t S_w = 0 \) is forced, meaning that the scanning curves are vertical.

The play-type hysteresis model has garnered a lot of interest recently due to the expression (1.23) being simple and closed-form. Existence [141, 149, 222] and uniqueness [53] results are available for the model when (1.23) is made invertible with respect to \( \partial_t S_w \). This is done either by inclusion of dynamic effects or by regularising \( \text{sign}(\cdot) \) as an increasing function. In [221] it is shown that play-type hysteresis does not ensure an \( L^1 \)-contraction property. An upscaling of the model using homogenization was performed in [218] to get more realistic scanning curves.

However, a criticism of the play-type hysteresis model is that in real life, the scanning curves are generally inclined and not vertical, see Figure 1.3. This curtails the applicability of the play-type model in practical situations, making an extension necessary. This will be a recurrent topic in the chapters that follow.

The numerical challenges associated with modelling capillary hysteresis are mentioned later in this chapter. We turn now to the other term which contains hysteresis.

Relative permeability hysteresis

The relative permeabilities \( k, k_n \) and \( k_w \) are known to be hysteretic too. The effect is less pronounced compared to capillary hysteresis. For the two-phase case, the hysteresis in relative permeability is reported in [38, 102, 137]. Generally, changes in the relative permeability of the non-wetting phase are greater compared to the wetting phase. Hysteresis is also observed for unsaturated flows [197, 243]. Figure 1.4 plots
experimental measurements for $k$ and $k_n$ for an unsaturated [243] and a two-phase flow experiment [102]. It holds that

$$
\partial_t S_w \begin{cases}
\geq 0 & \text{if } \zeta = \zeta^{(i)}(S_w), \\
\leq 0 & \text{if } \zeta = \zeta^{(d)}(S_w),
\end{cases}
$$

for $\zeta \in \{k, k_n, k_w\}$. (1.26)

The scanning curves for $k$ are also shown in Figure 1.4.

Figure 1.4: Hysteresis in relative permeability. The $k_n$ plots use data from two-phase experiments in [102] and the $k$ plots are using data from unsaturated flow experiments in [243]. Scanning curves for the unsaturated case are shown.

From a modelling point of view, the discussions on the relative permeability hysteresis are generally limited to the analysis of hyperbolic Buckley-Leverett equation (1.18). In [22, 190, 215] the authors propose the scanning hysteresis model (SHM), which introduces an artificial variable $\pi$ to keep track of the location in the $k$-$S_w$ plane. An evolution equation of $\pi$ is provided and numerical methods and entropy conditions are derived based on that. Further, in [2, 190] a relaxation term is added in the evolution equation of $\pi$, similar to dynamic capillarity that will be introduced later in this chapter for capillary pressure. A jump hysteresis model is considered in [76, 119]. The models mentioned above are constrained by the fact that the non-equilibrium effects in the pressure are not taken into account. The role of the variable $\pi$ for SHM is played in reality by the physical variable $p_c$, and thus the question of hysteresis in $k$ and $p_c$ are intertwined. This is explained in detail in Chapter 7. Hence, by including hysteresis in $p_c$, the model can become more physical. The jump hysteresis model, on the other hand, is limited in application because of its discontinuous coefficients.

Results that incorporate hysteresis in both terms, are not very common to our knowledge. In [216] a numerical method is proposed for doing so. In [272, Chapter
1.2 Non-equilibrium effects

5] a play-type model is proposed for modelling hysteresis in both capillary pressure and relative permeability. However, we show in Chapter 7 that such a model is ill-posed since a separate equation for evolution of \( k \) is not provided.

Based on these observations, we propose in Chapter 7 a new model that establishes a relation between hysteresis in \( p_c \) and in \( k \). However, this only works when the second non-equilibrium effect, mentioned briefly before, is also present in the model.

1.2.2 Dynamic capillarity

The existence of the dynamic capillarity effect was first postulated from experiments showing the occurrence of overshoots. To elaborate, in experiments like [72, 229] it is seen that if water is injected in a long porous column at a high enough rate, then a local maximum known as overshoot develops inside the domain, see Figure 1.1(b). This cannot be explained by the standard model as it violates the maximum principle [83, 221, 255]. The dynamic effect is evident from drainage and redistribution experiments too [35, 272], see Figure 1.5 (left). In [112], by thermodynamic arguments, the authors proposed the following extension of the standard model:

\[
p_n - p_w = p_c = p_c^+(S_w) - \tau f(S_w) \partial_t S_w.
\] (1.27)

Observe that, in the above, no hysteresis effects are taken into account. The constant \( \tau > 0 \) and the function \( f(\cdot) > 0 \) are known as the dynamic capillarity coefficient and function respectively. This law is consistent with the second law of thermodynamics [112] since, in the absence of hysteresis, entropy generation requires

\[
-\partial_t S_w[(p_n - p_w) - p_c^+(S_w)] \geq 0.
\] (1.28)

The term \( \tau f(S_w) \) is determined experimentally in [47, 70, 134, 275]. Figure 1.3 (right) is a reproduction of [47, Figure 7] showing how \( \tau f(S_w) \) typically depends on \( S_w \) for different soils. An overview of experiments, modelling and applications of the dynamic capillary effect can be found in [110, 267]. In recent years, a generalisation of (1.27) using thermodynamically constrained averaging theory (TCAT) has been proposed [106] which takes into account the evolution of interfaces. Some analysis of the model has been done [235] in relation to (BL) equation (1.16). However, in this study, we mainly focus on expression (1.27) for simplicity.

The dynamic capillarity model has been extensively studied both mathematically and numerically. The importance of dynamic capillarity in trapping was shown in [246]. Stability was studied in [255]. When (1.27) is substituted into (1.20) it gives a third-order nonlinear pseudo-parabolic system. Existence results for such equations can be found in [31, 55, 94, 163], and uniqueness in [54, 94]. The pseudo-parabolic equations have several interesting properties. For example, the solutions
Figure 1.5: Dynamic capillarity effect. The (lower left) plot shows the non-monotone variation of $p_n - p_w$ with respect to saturation $S_w$ shown in (upper left) plot, during the drainage experiments in [35]. This implies the presence of dynamic capillarity effect. In the (right) plot, the dependence of the term $\tau f(S_w)$ on $S_w$ is shown for different soils. The figure is taken from [47].

of linear pseudo-parabolic equations inherit the spatial regularity of the initial condition [69], which means that if the initial condition is discontinuous then the discontinuity vanishes as time goes to infinity. Another important property is that the solutions do not satisfy the maximum principle. This has been used in [68, 209, 222, 248, 254] to explain overshoots. Travelling wave solutions for the dynamic capillarity model have been investigated both for the unsaturated [68] and the two-phase case [236, 248, 254]. Among these, [254] is particularly important as it shows that dynamic capillarity has significant effect on solutions of (HBL) (equation (1.18)) despite capillary pressure not being present explicitly in the expression. This is because the entropy solutions become vastly different and possibly non-monotone if the viscous term includes dynamic capillarity.

As before, we leave the discussions on numerical methods for the next section. Before that, we formulate the problem when both hysteresis and dynamic effects are present.

1.2.3 The combined model and some mathematical techniques

When both play-type hysteresis and dynamic capillarity are considered in the capillary pressure, the $p_c$-$S_w$ relationship becomes [24, 25],

$$p_n - p_w = p_c \in p_c^+ (S_w) - p_c^- (S_w) \cdot \text{sign} (\partial_t S_w) - \tau f(S_w) \partial_t S_w. \quad (1.29)$$
1.2 Non-equilibrium effects

This extension is consistent with (1.28). For a fixed $S_w$, observe that, $p_c(S_w)\zeta + \tau f(S_w)\zeta$ is an increasing function of $\zeta$, which allows rewriting (1.29) as

$$\partial_t S_w = \frac{1}{\tau f(S_w)} \mathcal{F}(S_w, p_c) := \frac{1}{\tau f(S_w)} \begin{cases} p_c^{(i)}(S_w) - p_c & \text{when } p_c \leq p_c^{(i)}(S_w), \\ 0 & \text{when } p_c \in [p_c^{(i)}(S_w), p_c^{(d)}(S_w)], \\ p_c^{(d)}(S_w) - p_c & \text{when } p_c \geq p_c^{(d)}(S_w). \end{cases}$$

(1.30)

The combined model has been studied extensively. Some mathematical results on well-posedness are found in [53, 141, 149, 222]. Apart from it being useful for including both the non-equilibrium effects, it is shown in [209] that the combination of dynamic capillary effect and hysteresis is required to explain the development of viscous fingers in a homogeneous media, see Figure 1.6. This observation concurs with the experiments in [103].

![Figure 1.6: Numerical experiments on viscous fingering in a homogeneous porous media, taken from [209, Figure 5]. The top row shows the development of fingers with time when both hysteresis and dynamic capillarity are present. The bottom row has only dynamic capillarity. The distinction between the two cases is clear. For the standard model and for model including only dynamic capillarity, the development of fingers is ruled out by [83, 221, 255].](image)

The model (1.30), coupled with (1.13) or (1.20), is difficult to analyse in its full generality, even when the permeabilities are assumed to be functions of $S_w$ only. To circumvent this issue, often problems are considered in lower spatial dimensions, 1-D for example. This allows one to verify hypotheses or make predictions using techniques in the mathematical glossary. Here, we apply two such techniques, i.e. travelling waves and self-similar solutions. In travelling wave approximation, a one-dimensional infinite domain is considered, and the solutions are assumed to be of
the form
\[ \eta = \eta(\xi) \text{ for } \eta \in \{S_w, p_n, p_w, p_c\}, \] (1.31)
where the travelling wave coordinate for space coordinate \( x \) and time \( t \) is given by,
\[ \xi = ct - x. \]

Here, \( c \) is the Rankine-Hugoniot wave speed. Such approximations are used in diverse fields [84, 228, 249] to study the qualitative behaviour of solutions. In relation to the non-equilibrium effects, travelling waves have been rigorously studied in [68, 236, 248]. The stability of the travelling waves have been studied in [175]. The admissible shocks, in case of hyperbolic problems, are also derived using such analysis [150, 254].

Self-similar solutions, on the other hand, are useful in porous media flow models for purely diffusive problems in a semi-infinite domain. The substitution used is:
\[ \eta = \eta(\xi) \text{ for } \eta \in \{S_w, p_n, p_w, p_c\} \text{ and } \xi = \frac{x}{\sqrt{t}}. \] (1.32)

This has been widely used to study redistribution in horizontal porous media [110, 189, 193, 247, 252, 253]. In Chapter 2 we use it for the same purpose. The self-similar solutions are also helpful in studying Stefan problems [127].

<table>
<thead>
<tr>
<th>Mathematical results for non-equilibrium effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Existence</td>
</tr>
<tr>
<td>1P: ( k = 1 ) [218, 221]</td>
</tr>
<tr>
<td>2P: [31, 94, 163]</td>
</tr>
<tr>
<td>2P: [55]</td>
</tr>
<tr>
<td>1P: [149, 222]</td>
</tr>
<tr>
<td>2P: [141]</td>
</tr>
<tr>
<td>1P: Ch. 3</td>
</tr>
<tr>
<td>1P: Ch. 3</td>
</tr>
<tr>
<td>Uniqueness</td>
</tr>
<tr>
<td>1P: [54, 94]</td>
</tr>
<tr>
<td>2P: [53]</td>
</tr>
<tr>
<td>Special solutions</td>
</tr>
<tr>
<td>(TW)</td>
</tr>
<tr>
<td>1P: Ch. 4</td>
</tr>
<tr>
<td>2P: Ch. 7</td>
</tr>
<tr>
<td>(Sim) Ch. 3</td>
</tr>
<tr>
<td>1P: Ch. 4</td>
</tr>
<tr>
<td>2P: [236, 248, 254]</td>
</tr>
<tr>
<td>(TW)</td>
</tr>
<tr>
<td>1P: Ch. 4</td>
</tr>
<tr>
<td>2P: [236, 248, 254]</td>
</tr>
<tr>
<td>(TW)</td>
</tr>
<tr>
<td>1P: Ch. 5</td>
</tr>
<tr>
<td>(Sim) Ch. 2</td>
</tr>
<tr>
<td>(TW)</td>
</tr>
<tr>
<td>1P: Ch. 7</td>
</tr>
<tr>
<td>2P: Ch. 7</td>
</tr>
<tr>
<td>Stability</td>
</tr>
<tr>
<td>[221]</td>
</tr>
<tr>
<td>[175, 255]</td>
</tr>
<tr>
<td>[149, 209]</td>
</tr>
<tr>
<td>Other results</td>
</tr>
<tr>
<td>derivation: [24], upscaling: [218]</td>
</tr>
<tr>
<td>interface condition: [246]</td>
</tr>
<tr>
<td>finger speed: Ch. 6</td>
</tr>
<tr>
<td>derivation: [24]</td>
</tr>
<tr>
<td>derivation: Ch. 7</td>
</tr>
</tbody>
</table>

Table 1.1: Summary of mathematical results concerning non-equilibrium effect in unsaturated/two-phase flow in porous media. The abbreviations used are 1P: unsaturated, 2P: two-phase, TW: travelling waves, Sim: self-similar solutions and Ch.: Chapter.
To summarize this section, a table of known results and open questions is given. This illustrates where the current work fits in. The extended play-type model mentioned in one of the columns of Table 1.1 refers to the model proposed in Chapter 2.

1.3 Numerical challenges

In this section we would look at some of the major challenges in numerical implementation of the models mentioned above. In particular, we would focus on heterogeneity, nonlinearity, degeneracy and non-equilibrium effects along with the combinations of these four. Some of the other important challenges that will not be addressed here are the existence and growth of fractures, thermal and reactive transport and poromechanics.

1.3.1 Heterogeneity

![Figure 1.7: Heterogeneity in the top layer of the SPE10 benchmark reservoir [63]. The (left) figure shows the logarithmic variation of the horizontal diagonal component of the absolute permeability tensor, $K_{11}$, and the (right) figure shows the variation of porosity, $\phi$. The layered structure of the subsurface is evident from the figure.](image)

The subsurface is extremely heterogeneous as the material properties fluctuate up to orders of magnitude over small distances. Moreover, compositionally, the subsurface is made of layers of rocks having very different properties. In oil and gas recovery and CO$_2$-sequestration, it is extremely important to model conduction through these layers accurately, since the oil/CO$_2$ is often confined between less permeable layers of bedrock and caprock. Table 1.2 shows values of absolute permeability for different types of rocks, whereas Figure 1.7 shows typical variation of permeability and porosity inside the benchmark reservoir SPE10 [63].
## 1.2 Absolute permeability of different rock types

Table 1.2: Absolute permeability of different rock types. The permeability can vary over 10 orders of magnitude depending on the rock.

<table>
<thead>
<tr>
<th>Rock Type</th>
<th>Permeability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fractured rocks</td>
<td>$10^{-3} - 10^{-6}$</td>
</tr>
<tr>
<td>Oil reservoir rocks</td>
<td>$10^{-7} - 10^{-9}$</td>
</tr>
<tr>
<td>Sandstone</td>
<td>$10^{-10} - 10^{-11}$</td>
</tr>
<tr>
<td>Limestone</td>
<td>$10^{-12} - 10^{-13}$</td>
</tr>
<tr>
<td>Granite</td>
<td>$10^{-14} - 10^{-15}$</td>
</tr>
</tbody>
</table>

For the standard model, the effects of heterogeneity have been studied extensively both analytically and numerically. Here we restrict ourselves to few examples. Effective equations and transmission conditions for highly heterogeneous media were derived analytically in [61,165,218,220,252]. Different spatial discretizations have been used for the numerical treatment of the porous flow problem including finite differences [42], finite volumes [92,93,162], finite elements [58,177], mixed finite elements [13,81,179,208,268], mortar finite elements [186], discontinuous Galerkin [7,15,16,86,87] and multi-point flux approximation methods [140]. In particular, the issue of heterogeneity is addressed in the papers [81,86,87,140,143,186]. Finite difference and finite volume methods are proposed for highly anisotropic diffusion problems in [257] and [258]. Domain decomposition methods (DD) [75,199] are often used when layered structures are involved. Some examples in this regard are [232,233] where nonlinear DD solvers and preconditioners are proposed, [174] where DD preconditioners for layered media are analysed and [28,29] where a non-overlapping DD scheme for the Richards equation is considered. DD techniques are discussed in detail in Chapter 8. A different approach to deal with the heterogeneities is the multiscale approach. The underlying principle in this case is to divide the domain into smaller subdomains, find basis functions that capture the heterogeneities at the smaller scale and reconstruct the global solution using these functions. Examples include multiscale finite volume [128], multiscale finite elements [62,121] and iterative multiscale methods [108]. Although the multiscale methods are known to be comparatively faster and parallelizable, in the multiphase case it becomes difficult to prove the convergence of the scheme analytically, see [118,265] for error estimates of some nonlinear cases. In the direction of adaptive refinements, a-posteriori estimates have been derived in [5,50,88,231,262] and refinement strategies are investigated in [17,61,64,266].

### 1.3.2 Nonlinearity and degeneracy

Another issue to be addressed is the nonlinearity and degeneracy of the flow models. For the standard model, the nonlinearity stems from the functions $k_n$, $k_w$ and $p_c$ being nonlinear, as seen from (1.6) and (1.7). Degeneracy, on the other hand, refers to the loss of parabolicity of (TP) and (Ri) (equations (1.13) and (1.20)) due to the associated functions blowing up or vanishing at $S_w = 0,1$, i.e.

$$
\lim_{{S \to 0}} p_c^c(S) = \infty, \quad k_w(0) = k(0) = 0, \quad k_n(1) = 0.
$$

(1.33)
Stability of a numerical scheme depends inherently on the degeneracy of the problem.

To solve nonlinear parabolic equations like the ones rising from the standard model, backward Euler time-discretization is generally used due to its stability. This results in a sequence of nonlinear elliptic problems which is solved using a set of linear iterations. One of the most common choices in this regard is the Newton iteration [26, 151]. The Newton iteration, also known as the Newton scheme, shows quadratic convergence in the best case scenario. However, convergence is achieved only if the initial guess is close enough to the time-discrete solution. To improve the convergence of the Newton solver, many alterations have been proposed over time, including multigrid pre-conditioning [130], line-search methods, different parametrizations [39], implicit pressure explicit saturation methods (IMPES) [143] etc. The Newton scheme calculates the Jacobian at every iteration, which is computationally expensive. The alterations mentioned above, add to that cost in some cases reducing the order of convergence. Moreover, the stiffness matrices are relatively ill-conditioned [155]. To circumvent these problems the modified Picard scheme was proposed in [57]. However, it is also converging only when the initial guess is close enough to the solution and the convergence is linear [151]. In [183] it is shown for non-degenerate elliptic problems that for mixed finite element discretization, in order to ensure convergence of the Newton method, the difference between initial guess and the solution should be of the order of $\frac{1}{h^d}$, $h$ being the mesh-size and $d$ the spatial dimension. The analysis is generalised for parabolic problems in [206] where it is shown that the Newton and the modified Picard schemes converge if

$$\Delta t \leq C \varepsilon^r h^d,$$

where $\Delta t > 0$ is the time step size, $C, \varepsilon, r > 0$ are constants and $p^{\beta}(S_w) \leq \varepsilon^{-1}$. Not only is it a restrictive condition for higher dimensional computations but it also shows that Newton and Picard schemes become unstable as $S_w \to 0$, see (1.33).

An alternative scheme that does not have the aforementioned stability problem, is the L-scheme developed in [192, 194, 234]. It is globally convergent, meaning it converges irrespective of the initial guess and the convergence is guaranteed even for degenerate cases, provided that the flux is bounded. Further, no derivatives have to be calculated at each iteration and the matrices are well-conditioned. The L-scheme analysis has now been extended to cover two-phase flows [203], Hölder continuous functions [202], domain-decomposition approaches [227] and geomechanics problems [33, 34]. However, the L-scheme is only linearly convergent and the convergence is considerably slower compared to the Newton or the Picard scheme [155]. A detailed analysis of different linearisation schemes is found in Chapter 9. Regarding spatial discretization, convergence of gradient schemes for nonlinear elliptic/parabolic problems is studied in [78].

To deal with degeneracy, the global pressure formulation is often used where Kirchhoff transformation is applied to absorb the nonlinearity of the diffusion coeffi-
cient [9, 13]. Error estimates are derived in [191] for the time discrete formulation that uses this transformation. However, calculating the inverse transformation is computationally expensive. Furthermore, the method does not generalise to cover all non-equilibrium effects. In [39] degeneracy is handled by assuming the pressure and the saturation both to be functions of a new primary variable. An energy stable method for degenerate parabolic equations is analysed in [49] that preserves the gradient flow structure at a discrete level. Other approaches to resolve this issue include regularisation of the associated functions [160] and shifting of initial/boundary data [195]. In Chapter 9 we propose a fast linearisation technique that converges for degenerate problems and does not require the Kirchhoff transform.

1.3.3 Numerical methods for non-equilibrium effects

The effects of dynamic capillarity, specially in relation to development of overshoots, have been studied extensively using computations, see [267] for an overview. For one and two dimensional infiltration experiments, non-monotone saturation profiles, resulting from dynamic effects, have been investigated in [214]. A similar study but focusing more on the effect of different $\tau f(S_w)$ terms, has been reported in [276]. Based on travelling wave analysis, an adaptive moving mesh method was proposed in [270] which substantially reduces the computational costs of the numerical experiments mentioned above.

For heterogeneous media, transmission conditions are derived in [246] incorporating dynamic capillarity effect. Numerical schemes including and excluding entry pressures have been proposed in [117] and [116] respectively, to deal with heterogeneity. The importance of dynamic effects for flow through layered media was shown numerically in [96]. Regarding spatial discretization, a cell-centered finite difference method and a conservative Eulerian-Lagrangian approach are used in [187], whereas a mimetic finite difference method is developed in [271]. With regard to finite volume methods, a two point flux approximation scheme was proposed in [85], whereas, a multipoint flux approximation scheme was proposed in [52]. Furthermore, a Galerkin time-stepping method for dynamic effects is derived in [90] and discontinuous Galerkin method is derived in [135, 136]. A heterogeneous multiscale method is considered in [138, 139] where dynamic capillarity plays a role in the micro-scale. The domain decomposition approach proposed in Chapter 8 is extended to cover hysteresis and dynamic capillarity in [159]. A scheme for Buckley-Leverett equation, taking into consideration dynamic viscous terms, is derived in [264]. For a similar scenario, in [69] a scheme is proposed that takes into account possible discontinuities in saturation. Operator splitting schemes are discussed in [3] also in this context. In Chapter 10 we further discuss numerical schemes for the dynamic capillarity case.

The role of hysteresis in CO$_2$ storage is shown numerically in [71] and its role in formation of fingers is demonstrated in [131, 209]. For hysteresis in capillary
pressure, a play-type model is known to cause instability of the numerical algorithm [44, 216] due to a vertical switch between the imbibition and the drainage curves. To resolve this, a regularisation is often applied for computations. This can either be done by introducing dynamic capillarity [141, 149] or by smoothening the scanning curves [216]. For hysteresis in permeability, a Gudunov method is developed in [215] for solving Riemann problems using scanning hysteresis model. For similar objectives, operator splitting strategies are considered in [2] for a wide range of parameter values.

To solve the nonlinear equations associated with the problems mentioned above, standard methods such as the Newton or the Picard scheme are often used. However, we show in Chapters 2, 4 and 7 that in many situations these methods fail to yield a solution. In [136] the L-scheme is used to resolve this issue. The computations shown in Chapters 2, 4, 5 and 7 are also based on the L-scheme and its modified variant proposed in Chapter 9. The formulation is partly inspired by the semi-implicit discretization discussed in [95].

1.4 Overview of the chapters

So far we discussed some of the major challenges in modelling, analysis and simulation of flow through porous media. In this work we attempt to address these challenges. The work is divided into two parts. The first part is dedicated towards improving the existing models and predictions by inclusion of the non-equilibrium effects. The second part deals with the numerical challenges that were mentioned earlier.

1.4.1 Part I: non-equilibrium effects

In this part, we derive models that include the non-equilibrium effects and use mathematical techniques to understand and quantify their effect. In particular, we use phase plane techniques to analyse the behaviour of the travelling waves and functional analytic and numerical techniques to tackle the partial differential equations. We proceed step by step, starting with the unsaturated flow case with capillary hysteresis and dynamic capillarity considered separately, and ending with the two-phase case with both types of hysteresis and dynamic capillarity included simultaneously in the model. The chapter-wise description is as follows:

Chapter 2

Different hysteresis models are examined and a new thermodynamically consistent model, referred to as the extended play-type hysteresis model is proposed. For this model the scanning curves are non-vertical and can be constructed to approximate
experimentally observed scanning curves. Furthermore it does not require any bookkeeping when the flux reverses at some point in space. The problem of horizontal redistribution is considered to illustrate the strength of the new model. By self-similar transformation it is shown that all cases of redistribution can be handled, including the unconventional flow cases. Finally numerical experiments are presented that recover all the redistribution cases.

Chapter 3

The well-posedness of the extended play-type hysteresis model is investigated for unsaturated flows. The resulting equations are split into a parabolic equation coupled with an ordinary differential equation. First, the existence of solutions is proved for the regularised model. The result extends to the degenerate cases. For uniqueness, a non-standard approach is taken, where short time boundedness of the gradient is proved using Rothe’s method which finally yields uniqueness at all times. Further, the solutions are shown to satisfy a version of the maximum principle. The consequences of passing the regularisation parameter to zero are explored.

Chapter 4

In this chapter we investigate how the gravity driven wetting fronts behave while moving through long vertical homogeneous porous columns. The focus is on travelling wave solutions of the Richards equation, for which the existence is proved first for the model including play-type hysteresis and subsequently for the model including dynamic capillarity effect. A threshold value for \( \tau \) is found over which the travelling waves become non-monotone, resulting in overshoots. Similar thresholds of \( \tau \) are found, that dictate whether the overshoot will have regions of positive pressure or whether the overshoot will reach a maximum corresponding to the full saturation. Moreover, it is shown that the saturation stays within physical bounds when \( \int_0^1 f(S_w) = \infty \). The paper ends with numerical experiments confirming the theoretical results.

Chapter 5

The analysis of Chapter 4 is extended to include combined capillary hysteresis and dynamic effects. Both the play-type and the extended play-type models are considered for this purpose. It is shown that the overshoots will develop if certain parametric conditions are satisfied and the number of overshoots will be finite for the extended model. The characteristics of the profiles are explained in detail for all the cases. Moreover, parametric conditions that inhibit the fronts from reaching full saturation are laid out. The analysis agrees well with experimental observations. Finally, numerical results are shown that confirm all the theoretical predictions.
1.4 Overview of the chapters

Chapter 6

The travelling wave analysis of Chapter 5 is repeated with a special emphasis on finding monotone solutions. Existence of such solutions is proved for appropriate boundary conditions. This, in theory, gives the speed of the fingers in homogeneous isotropic media.

Chapter 7

This chapter accumulates work in previous chapters, as we study the behaviour of saturation fronts for two-phase flow through a long homogeneous porous column. In particular, the model includes hysteresis and dynamic effects in the capillary pressure, and hysteresis in the permeabilities. The analysis uses travelling wave approximation. Entropy solutions are derived for Riemann problems that are arising in this context. These solutions belong to a much broader class compared to the standard Oleinik solutions, where hysteresis and dynamic effects are neglected. The relevant cases are examined and the corresponding solutions are categorized. They include non-monotone profiles, multiple shocks and self-developing stable saturation plateaus. Numerical results illustrate the mathematical analysis. Finally, experimental results are compared with the theoretical findings.

1.4.2 Part II: numerical methods

The chapters are roughly divided into discussions of heterogeneity, nonlinearity and non-equilibrium effects. The chapter-wise description goes as follows:

Chapter 8

We propose a linear iterative domain decomposition scheme for heterogeneous and layered porous media, which uses the linearisation (L-type) iterations to also decouple the subdomains. The convergence of the scheme is proved rigorously under mild restrictions on the time step size. The scheme is numerically compared with other monolithic schemes, such as the Newton and the Picard scheme. It is seen that the proposed scheme is more stable than the Newton scheme while remaining comparable in computational time, even when no parallelisation is adopted. After presenting a parametric study that can be used to optimise the proposed scheme, we briefly discuss the effect of parallelisation and give an example of a four-domain implementation.

Chapter 9

A linearisation technique is proposed for solving nonlinear elliptic partial differential equations that are obtained from the time-discretization of a wide variety of
nonlinear parabolic problems. The scheme is inspired by the L-scheme, which gives unconditional convergence of the linear iterations. Here we take advantage of the fact that at a particular time step, the initial guess for the iterations can be taken as the solution of the previous time step. First it is shown for quasilinear equations that have linear diffusivity that the scheme always converges, irrespective of the time step size, the spatial discretization and the degeneracy of the associated functions. Moreover, it is shown that the convergence is linear with convergence rate proportional to the time step size. Next, for the general case it is shown that the scheme converges linearly if the time step size is smaller than a certain threshold which does not depend on the mesh size, and the convergence rate is proportional to the square root of the time step size. Finally numerical results are presented that show that the scheme is at least as fast as the modified Picard scheme, faster than the L-scheme and is more stable than the Newton or the Picard scheme.

Chapter 10

A fully discrete numerical scheme for the model describing two-phase immiscible flow in porous media with dynamic effects in the capillary pressure is proposed. For time discretization, the Euler implicit method is used. The spatial discretization is based on the mixed finite element method (MFEM). Specifically, the lowest order Raviart–Thomas elements are applied. The error estimates for the saturation, fluxes and phase pressures are derived for the temporal and spatial discretization to show the convergence of the scheme. Finally, numerical results are presented that support the theoretical findings.

Finally we conclude this work by summarizing our findings and reiterating the important open questions. Based on the results, a recommendation is made for a model that incorporates the non-equilibrium effects without significantly increasing the computational expense.
Part I:

Non-equilibrium effects
Chapter 2

Hysteresis and horizontal redistribution in porous media

2.1 Introduction

In this chapter we consider the flow of two immiscible and incompressible fluids through a homogeneous and isotropic porous medium. It is assumed that the pores of the medium are fully occupied by these fluids. One fluid is the wetting phase and the other one is the non-wetting phase. They are denoted by the subscripts \( w \) and \( n \). We disregard the influence of gravity, as we are interested in a horizontal physical system. Further we assume that there are no internal sinks or sources. The corresponding (i.e. macroscopic) equations are well-known \([19, 65, 115]\) and in the dimensionless form they read

\[
\begin{align*}
\partial_t S_\alpha + \nabla \cdot \vec{F}_\alpha &= 0, \\
\vec{F}_\alpha &= -k_\alpha(S_\alpha)\nabla p_\alpha \\
S_n + S_w &= 1, \\
S_n, S_w &\geq 0.
\end{align*}
\]  
(2.1a)

For the stated assumptions, (2.1) directly follows from (TP) (equation (1.13)), since the homogeneity implies \( \kappa, \phi \equiv 1 \), and the absence of source terms and gravity implies \( R_n, R_w \equiv 0 \) and \( \hat{g} \equiv 0 \) respectively. Further, since in this chapter, we do not prescribe flux at the boundary, \( v_{\text{ref}} \) in (1.11) is chosen so that \( N_c = 1 \). Similar to before, in (2.1a), \( S_\alpha \), \( p_\alpha \) and \( \vec{F}_\alpha \) denote saturation (assumed to be scaled such that \( 0 \leq S_\alpha \leq 1 \)), pressure and volumetric flux of phase \( \alpha \), and \( k_\alpha \) is the relative permeability of the

---

porous medium with respect to phase $\alpha$. If the system contains water as the wetting phase and air as the non-wetting phase, then it is generally assumed that $p_n = p_{\text{air}}$ is constant, see Section 1.1.2. The equations (2.1a) reduce to Richards equation (Ri) which is the standard equation that models flow in unsaturated porous domains [19, 65, 115]:

$$ \partial_t S_w + \nabla \cdot \vec{F}_w = 0, \quad \vec{F}_w = -k_w(S_w)\nabla p_w \text{ with } 0 \leq S_w \leq 1. $$

(2.2)

In general, the closure relation between $p$ and $S_w$ is given in form of an algebraic relationship determined from experiments:

$$ p_n - p_w = p_c(S_w), $$

(2.3)

where $p_c : (0,1) \rightarrow \mathbb{R}$ is the capillary pressure, see Section 1.1. There is a vast amount of literature available on the capillary pressure and its interpretation [19, 65, 115]. For properties and closed-form relations we refer to the well-known references [45, 259].

![Figure 2.1: $p_c$-$S_w$ relationships from experiments and existing models. (left) Schematic diagram of main drainage and imbibition curves and scanning curves [272]. (right) Schematic diagram of scanning curves for play-type hysteresis and similarity models [24, 171].](image)

It is known that multiphase flow in porous media displays hysteretic effects [212]. The capillary pressure-saturation relationship (henceforth called $p_c$-$S_w$ relationship) traces one path while going through an infiltration/imbibition/wetting process and another path while going through a drainage/drying process. This effect can be incorporated into the classical model by replacing $p_c(S_w)$ function in (2.3) by two different functions: $p_{c(i)}(S_w)$ for imbibition and $p_{c(d)}(S_w)$ for drainage. This works quite well if the porous medium is going through only an imbibition or a drainage
process. But if there is a switch between the two, then the $p_c-S_w$ curves span the region between the $p_c^{(i)}$ and $p_c^{(d)}$ curves in form of scanning curves [170]. Typical behaviour is shown in Figure 2.1. Some of the most common models for porous media hysteresis are listed below.

### 2.1.1 Similarity models

The main idea behind this class of models is to express the scanning curves by a closed-form relationship that allows the scanning curve to be similar in shape to the imbibition or drainage curves. With the similarity hypothesis, Philip [188] was able to obtain reasonable approximations for scanning curves. A formally equivalent model was derived by Poulouvassilis [196] based on the independent domain theory. Concepts of domain theory of capillary hysteresis were further used in the models proposed by Mualem [171,172], Mualem & Dagan [173], Lenhard-Parker [184] and others (see [142,260]). The scanning curves obtained in this way are close to the experimental scanning curves. Moreover, this class of models can describe “secondary hysteresis”. This refers to the phenomenon that the scanning curve through a point, switching from imbibition to drainage, is different from the scanning curve through the same point switching from drainage to imbibition. All other models discussed in this paper do not include this secondary behaviour, because for those models an intermediate point can move back and forth on the same scanning curve. Similarity models can also be used to explain many cases of the horizontal redistribution problem [114], which is an important benchmark problem for flows in porous media.

However similarity models are not straightforward to apply because in these models the saturation at a point is a function of all previous reversal points (when it switches from imbibition to drainage or vice versa) [142,260]. The closed-form expressions of the scanning curves actually take in these reversal points as parameters. The order at which the processes (imbibition/drainage) have gone through plays also a vital role. This leads to book-keeping for each point in space, making the models difficult to handle in practice and in any numerical or analytical approach.

### 2.1.2 Play-type hysteresis

In this approach one models scanning curves as vertical lines between $p_c^{(i)}$ and $p_c^{(d)}$ [24,222,261]. To close equation (2.1a) one replaces (2.3) by an expression of the form

$$p_n - p_w := p_c(S_w, \partial_t S_w) \in p_c^+(S_w) - p_c^-(S_w) \cdot \text{sign}(\partial_t S_w),$$

(2.4)

where sign(*) is defined in (1.25) and $p_c^\pm$ are defined in (1.24). This model implies that if $\partial_t S_w > 0$, then $p_n - p_w = p_c^+(S_w) - p_c^-(S_w) = p_c^{(i)}(S_w)$, and if $\partial_t S_w < 0$, then $p_n - p_w = p_c^-(S_w) + p_c^+(S_w) = p_c^{(d)}(S_w)$. If $p_c^{(i)}(S_w) < p_n - p_w < p_c^{(d)}(S_w)$, then $\partial_t S_w = 0$. 
The model based on (2.2) and (2.4) is well posed in the mathematical sense [53, 141, 149, 222] and can be physically justified by pore-scale [217] or thermodynamic [24] arguments. Furthermore, upscaling of the model using homogenization has been done in [218] and a stability analysis has been performed in [221]. Although this model has the advantage of being simple and local in time (no information on the history of a point is required) the resulting vertical scanning curves do not really resemble the ones from experiments. Moreover, as we show later, the play-type hysteresis model cannot describe all cases of horizontal redistribution.

2.1.3 Interfacial area model

Pore-scale simulations have shown [210] that interfacial areas play an important role in the $p_c-S_w$ relationship. Motivated by this, a model was proposed in [111, 113, 176] based on thermodynamic considerations. The main idea is to introduce the volumetric interfacial area ($a_{wn}$) as a new variable, in addition to saturation and pressure, and to assume that $a_{wn}$ is a unique function of saturation and capillary pressure:

$$a_{wn} = a_{wn}(S_w, p_n - p_w).$$

(2.5)

A transport equation for $a_{wn}$ was proposed leading to a new formulation for multiphase flow. The original idea was that this new formulation including interfacial area could account for the full hysteretic nature of the water retention characteristic. However, when analysing the model [193] it was shown that for any fixed point $x_0$ in space, there exists a unique $p_c-S_w$ curve which satisfies

$$\frac{dp_c}{dS}(S_w) = g(p_c, S_w), \quad \text{with } p_c(S_w(x_0, 0)) = p_c(x_0, 0),$$

(2.6)

where $g$ is a given smooth function of $p_c$ and $S_w$. Thus if $S_w(x, 0)$ and $p_c(x, 0)$ have only two values, as they do for horizontal redistribution, two $P_c-S_w$ curves arise. Clearly for general initial conditions, infinitely many $P_c-S_w$ curves may arise. Hence the concept of primary imbibition and drainage is not described by the interfacial area model. Hysteresis, in the sense of switching between two curves at a fixed $x_0$, could only be included by introducing rate dependent terms in the coefficients [274].

The purpose of this chapter is to introduce a hysteresis model that is based on the play-type approach, having primary drainage/imbibition curves and in between non-vertical scanning curves. These scanning curves can be chosen in such a way that they are close to experimental data. This model is presented and discussed in Section 2.2. In Section 2.3 the equations and conditions for (horizontal) redistribution are given and in Section 2.4 self-similar solutions are discussed describing all possible redistribution cases using the new model. Then, in Section 2.5 a numerical
scheme (L-scheme) for the partial differential equation is presented and computational results are compared with the redistribution cases. Conclusion are given in Section 2.6.

2.2 Extended play-type hysteresis model

In this section we will introduce the extended play-type hysteresis model, and discuss its background and some of its properties. For ease of presentation, the capillary pressure is denoted by \( u \), i.e.

\[
  u := p_c = p_n - p_w. \tag{2.7}
\]

We extend the play-type hysteresis model by introducing non-vertical scanning curves which become vertical near the saturation end points \( S_w = 0 \) and \( S_w = 1 \). To this end we replace \( \partial_t S_w \) in the sign function in (2.4) by \( \partial_t (H(S_w) + u) \), where \( H : (0, 1) \to (0, \infty) \) is a function chosen in such a way that the corresponding scanning curves have the desired properties. Thus instead of (2.4) we propose

\[
  u \in p_c^+ (S_w) - p_c^- (S_w) \cdot \text{sign}(\partial_t H(S_w) + \partial_t u), \tag{2.8}
\]

where \( p_c^+ (S_w) \) and \( \text{sign}(\cdot) \) are defined as in (1.24) and (1.25), respectively. To better understand what relation (2.8) implies, let us introduce the following sets:

\[
  \mathcal{H} := \{(S_w, u) : S_w \in (0, 1], p_c^{(i)}(S_w) < u < p_c^{(d)}(S_w)\},
  \partial \mathcal{H}^{dr} := \{(S_w, u) : S_w \in (0, 1], u = p_c^{(d)}(S_w)\},
  \partial \mathcal{H}^{im} := \{(S_w, u) : S_w \in (0, 1], u = p_c^{(i)}(S_w)\}.
\]

Figure 2.2: Sets in \((S_w, u)\) plane according to Definition 2.1.

Definition 2.1. In the \((S_w, u)\) plane we consider the sets

- \( \mathcal{H} := \{(S_w, u) : S_w \in (0, 1], p_c^{(i)}(S_w) < u < p_c^{(d)}(S_w)\} \)
- \( \partial \mathcal{H}^{dr} := \{(S_w, u) : S_w \in (0, 1], u = p_c^{(d)}(S_w)\} \)
- \( \partial \mathcal{H}^{im} := \{(S_w, u) : S_w \in (0, 1], u = p_c^{(i)}(S_w)\} \)
In these definitions we considered $S_w \in (0, 1]$, meaning $0 < S_w \leq 1$ to avoid $S_w = 0$, as $p_c^{(i)}(S_w)$ and $p_c^{(d)}(S_w)$ become singular at that point.

If $(S, u) \in \mathcal{H}$, then by (1.24)

$$-1 = \frac{p_c^+(S_w) - p_c^{(i)}(S_w)}{p_c^-(S_w)} < \frac{p_c^+(S_w) - u}{p_c^-(S_w)} < \frac{p_c^+(S_w) - p_c^{(d)}(S_w)}{p_c^-(S_w)} = 1.$$  

Hence from (2.8)

$$-1 < \text{sign}(\partial_t H(S_w) + \partial_t u) < 1.$$  

Definition (1.25) of sign(·) then gives

$$\partial_t H(S_w) + \partial_t u = 0.$$  

(2.9)

This means that we have

$$\frac{d u}{d S_w} = -\frac{d H}{d S_w}(S_w) \text{ for } (S_w, u) \in \mathcal{H},$$  

(2.10)

implying scanning curves with slope $-\frac{d H}{d S_w}(S_w)$. Note that a point can move back and forth along the same scanning curve, see also Figure 2.3. This property holds along the vertical scanning curves of the play-type model as well.

If $(S, u) \in \partial \mathcal{H}^{im}$ then,

$$\frac{p_c^+(S_w) - u}{p_c^-(S_w)} = \frac{p_c^+(S_w) - p_c^{(i)}(S_w)}{p_c^-(S_w)} = 1,$$

and hence sign($\partial_t H(S_w) + \partial_t u$) = 1 implying

$$\partial_t H(S_w) + \partial_t p_c^{(i)}(S_w) = \left(\frac{d H}{d S_w} + \frac{d p_c^{(i)}}{d S_w}\right) \partial_t S_w \geq 0.$$  

(2.13)

Thus if $\frac{d H}{d S_w} > -\frac{d p_c^{(i)}}{d S_w}$ we have

$$\partial_t S_w \geq 0 \text{ for } (S_w, u) \in \partial \mathcal{H}^{im}.$$  

(2.11)

Similarly if $\frac{d H}{d S_w} > \frac{d p_c^{(d)}}{d S_w}$, then

$$\partial_t S_w \leq 0 \text{ for } (S_w, u) \in \partial \mathcal{H}^{dr}.$$  

(2.12)

Now choosing $H$ such that for each $0 < S_w \leq 1$

$$\frac{d H}{d S_w}(S_w) \geq \max \left\{-\frac{d p_c^{(i)}}{d S_w}(S_w), -\frac{d p_c^{(d)}}{d S_w}(S_w)\right\},$$  

(2.13)
2.2 Extended play-type hysteresis model

both the lower bound conditions are satisfied and so are inequalities (2.11) and (2.12). Moreover, condition (2.13) ensures that scanning curves originating from arbitrary points on \( \partial H_{\text{im}} \) or \( \partial H_{\text{dr}} \) remain in \( H \).

The question arises how to choose and construct a function \( H \) that gives scanning curves close to experimental data and satisfies (2.13) for mathematical consistency. We present a construction that is based on the experiments of Morrow and Harris [170]. Their results for drainage and imbibition are shown in Figure 1.3. Here the variables are as in [170]: the saturation is unscaled \( 0 < S_{wr} < S_w < 1 - S_{nr} \) and capillary pressure \( u \) is in cm of water. In the construction of \( H \) we use the same variables to get a meaningful comparison. We propose for \( H \) the form:

\[
H(S_w) = \alpha(S_w) \int_{S_{\text{ref}}}^{S_w} \max \left\{ -\frac{dp_c^{(i)}}{ds_w}(S_w), -\frac{dp_c^{(d)}}{ds_w}(S_w) \right\} dS,
\]

(2.14)

for \( S_w > S_{\text{ref}} \), where \( S_{\text{ref}} \) is chosen close to \( S_{wr} \) (in fact \( p_c^{(i)}(S_{\text{ref}}) = 35 \)) and \( p_c^{(i)} \) and \( p_c^{(d)} \) are taken from the experiment. Clearly if \( \alpha(S_w) \geq 1 \) and \( \frac{d\alpha}{ds_w}(S_w) \geq 0 \), then (2.13) is satisfied. With trial and error we found that

\[
\alpha(S_w) = 2 \left( 1 + \frac{1}{5} S_w^5 \right),
\]

(2.15)

gives a good approximation to both drainage and imbibition. The scanning curves shown in Figure 2.3 result from \( u = \text{constant} - H(S_w) \), where the constant is defined by the intersection point with \( \partial H_{\text{im}} \) or \( \partial H_{\text{dr}} \).

![Figure 2.3: Scanning curves based on (2.14) and (2.15).](image)

The hysteretic function \( H \) obtained this way gives a close match to the experiments of Morrow and Harris [170]. Having explicit expressions for \( p_c^{(i)} \) and \( p_c^{(d)} \), for
instance the van Genuchten expressions, would result in a (semi) explicit expression for $H$.

**Remark 2.1.** For simplicity one could consider $\frac{dH}{dS_w}(S_w) := \frac{1}{\delta}$ (is constant) for some $\delta > 0$. Then the scanning curves are described by the straight lines:

$$\frac{du}{dS_w} = -\frac{1}{\delta}.$$  \hspace{1cm} (2.16)

By the same argument as above, one would need for each $0 < S_w \leq 1$,

$$\frac{1}{\delta} \geq \max \left\{ -\frac{dp_{c}^{(i)}}{dS_w}(S_w), -\frac{dp_{c}^{(d)}}{dS_w}(S_w) \right\},$$  \hspace{1cm} (2.17)

to ensure inequality (2.13). However this would unrealistically restrict the range of saturations for which the scanning curves remain in $H$.

Beliaev and Hassanizadeh [24] were the first to derive the extended play-type hysteresis closure relationship. Using thermodynamical arguments they obtained an expression in the inverse form of (2.8): see [24, eq. 35]. In our notation it translates into:

$$u \in P^{+}(S_w) - P^{-}(S_w) \cdot \text{sign} \left\{ 1 - C \frac{dp_{c}^{0}}{dS_w}(S_w) \right\} \partial_1 S_w + C \partial_1 u \right\},$$

where $C > 0$ is a constant and $P_{c}^{0}$ is a reference curve between $p_{c}^{(i)}$ and $p_{c}^{(d)}$ that intersects all scanning curves. For example $P_{c}^{0}$ could be $P^{+}$. The authors argued, using experimental data, that the term $CP_{c}^{0}(S_w)$ has roughly 10% contribution in expression (2.18). Also they pointed out that $C \to 0$ corresponds to the play-type hysteresis model. As $C^{-1}(1 - C \frac{dp_{c}^{0}}{dS_w}(S_w)) > 0$ for each $0 < S_w \leq 1$, one could write,

$$\frac{dH}{dS_w}(S_w) := C^{-1} \frac{dp_{c}^{0}}{dS_w}(S_w).$$ \hspace{1cm} (2.19)

to put (2.18) in the form of (2.8). Note however that the reference pressure $P_{c}^{0}$ does not necessarily satisfy condition (2.13).

Extension of play-type hysteresis model by inclusion of non-vertical scanning curves has been hypothesized also in the context of numerical analysis. One of the earliest example of this is found in [109] where a numerical method has been proposed for one dimensional hysteretic columns. In [44], authors use non-vertical approximations to vertical scanning curves as the original play-type hysteresis model poses difficulties for convergence. Similarly in [149, 216] and Chapter 4, non-vertical scanning curves have been introduced to regularise the sign function for mathematical analysis and numerical experiments.
Remark 2.2. Because of the appearance of the term $\partial_t u$ in (2.8), the extended play-type hysteresis model requires initial conditions both in capillary pressure and saturation. This is in contrast to play-type hysteresis where only an initial condition in saturation is required.

2.3 Redistribution

Following Philip [189], Raats & van Duijn [201] and Pop et al [193] we make the effect of hysteresis explicit in cases of horizontal redistribution. We show that the extended play-type hysteresis as described by (2.8), with $H(S_w)$ satisfying (2.13), covers all these cases. This is one of the main purposes of the paper.

2.3.1 General set up

Consider a horizontal porous column of infinite extent, directed along the $x$-axis. In the column the flow is one dimensional and describes the redistribution of fluids. We shall restrict ourselves to the case of flow in an unsaturated porous media, with water as the wetting phase and air as the non-wetting phase. Setting $p_n = p_{air} = 0$, the variables to be determined are the water saturation $S_w$ and the capillary pressure $u = -p_w$. For brevity we drop the subscript $w$ from the notation. Let $\mathbb{R}^- = \{x < 0\}$ and $\mathbb{R}^+ = \{x > 0\}$ denote respectively the left and right half of the column. The governing equations are Richards equation (2.2) along with the closure relation (2.8):

\begin{align}
\partial_t S + \partial_x F &= 0, \quad \text{for } x \in \mathbb{R}, t > 0, \quad (2.20a) \\
F &= k(S)\partial_x u \quad \text{for } x \in \mathbb{R}, t > 0, \quad (2.20b) \\
u &\in p_c^+ (S) - p_c^- (S) \cdot \text{sign}(\partial_t H(S) + \partial_t u). \quad (2.20c)
\end{align}

The halves $\mathbb{R}^-$ and $\mathbb{R}^+$ have constant, but different, initial saturation and pressure at $t = 0$. We impose

\begin{align}
S(x,0) &= \begin{cases} 
S_l & \text{for } x \in \mathbb{R}^- \\
S_r & \text{for } x \in \mathbb{R}^+ 
\end{cases} \quad (2.21) \\
u(x,0) &= \begin{cases} 
u_l & \text{for } x \in \mathbb{R}^- \\
u_r & \text{for } x \in \mathbb{R}^+ \quad (2.22)
\end{cases}
\end{align}

Throughout this paper we assume that the functions $k$, $p_c^{(i)}$, $p_c^{(d)}$ and $H$ are smooth and satisfy the structural properties:

(A1) $k(0) = 0$; $k(S)$, $\frac{dk}{dS}(S) > 0$ for $0 < S \leq 1$ with $k(1) < \infty$. 

Hysteresis and horizontal redistribution in porous media

(A2) \( p_c^{(d)}(S) > p_c^{(i)}(S) > 0 \) if \( 0 < S < 1 \) with \( p_c^{(d)}(1) \geq p_c^{(i)}(1) = 0 \).

(A3) \( \frac{dp_c^{(i)}}{dS}(S) \cdot \frac{dp_c^{(d)}}{dS}(S) < 0 \) for \( 0 < S < 1 \).

(A4) \( H \) satisfies (2.13).

For the initial conditions (2.21) and (2.22) to be consistent with expression (2.20c) we must impose,

\[
\begin{align*}
\rho_c^{(i)}(S_j) \leq u_j \leq \rho_c^{(d)}(S_j) & \quad \text{for } j \in \{l, r\}.
\end{align*}
\]

(2.23)

In this and later sections we use the notation,

**Definition 2.2.** \( E_l := (S_l, u_l) \) and \( E_r = (S_r, u_r) \).

Thus in terms of Definitions 2.1 and 2.2, condition (2.23) reads: \( E_l, E_r \in \mathcal{H} \cup \partial \mathcal{H}^{im} \cup \partial \mathcal{H}^{dr} \).

Mass conservation and momentum conservation requires the flux \( F \) and the suction \( u \) to be continuous. However, the saturation maybe discontinuous across \( x = 0 \). Thus the strategy is to solve (2.20) separately for \( \mathbb{R}^- \) and for \( \mathbb{R}^+ \), subject to (2.21) and (2.22) and then to match possible solutions so that flux and suction are continuous at \( x = 0 \). Such solutions will be either in the imbibition state or in the drainage state in \( \mathbb{R}^- \) and \( \mathbb{R}^+ \). Thus in addition to (2.20)-(2.22) we shall explicitly use

\[
\begin{align*}
\rho^i(S_j) & \leq u_j \leq \rho^d(S_j) & \quad \text{for } j \in \{l, r\}.
\end{align*}
\]

(2.24)

for each \( t > 0 \). Here we use the notation \( f(0^\pm) = \lim_{x \to 0^\pm} f(x) \).

### 2.3.2 Possible initial conditions

Let \( E_r \in \partial \mathcal{H}^{dr}, E_l \in \partial \mathcal{H}^{im} \) and let \( u_l > u_r \). Since this implies \( S_l < S_r \) one expects that water flows from the wet half column to the dry half column. This is called “conventional flow”. It is described by Philip in his classical paper [189].

He found that in this case the right half column \( (\mathbb{R}^+) \) is in the drainage state, with \( (S, u) \) following a trajectory on \( \partial \mathcal{H}^{dr} \), and the left half column \( (\mathbb{R}^-) \) is in the imbibition state, with \( (S, u) \) following a trajectory on \( \partial \mathcal{H}^{im} \). These trajectories are connected at \( x = 0 \) by a horizontal jump in the \( (S, u) \) plane where \( u = u_{0,1} \in (u_r, u_l) \). This value is uniquely chosen so that the flux is continuous. The behavior is sketched in the \( (S, u) \) plane in Figure 2.4.

However it was later realized and pointed out by Raats & van Duijn in [201] that this construction fails for \( u_l < u_r \). Then one has to use the scanning curves
emerging from the points $E_l$ and $E_r$ as in Figure 2.4. With these scanning curves one follows the same procedure. But now the left half is in the drainage state and becomes drier, while the right half is in the imbibition state and becomes wetter. This is called “unconventional flow” because additionally if $S_l < S_r$ then counterintuitively water flows from the dry half to the wet half. Using the Mualem model, Heinen & Raats [114] demonstrated numerically that this type of redistribution does indeed occur.

Going one step further one can ask what happens when $E_j \in \mathcal{H}$ ($j \in \{l, r\}$). Although redistribution results arising from interfacial area models [193] are available for general pressure and saturation initial conditions, they do not specify any directions in the induced $p_c - S_w$ curves. Thus, they can not describe hysteretic redistribution in the broadest sense. This is the same reason why redistribution results for heterogeneous semi-infinite blocks [247] cannot be extended to cover hysteretic domains.

Remark 2.3. In the extended play-type model, the primary scanning curves are described by equation (2.9). On such curves in $\mathcal{H}$ one can, in principle, go back and forth: i.e. the same scanning curves are used for drainage and imbibition. Hence no secondary scanning curves are generated. However, if there were secondary scanning curves, they would not play a role in the horizontal redistribution cases we consider, since the right and left halves of the column can only be in one state: either drainage or imbibition. This applies for example to the Mualem model used by Heinen & Raats [114].

Figure 2.4: (green) Redistribution according to Philip [189] yielding “conventional” flow; (cyan) redistribution according to Raats & van Duijn yielding “unconventional” flow. The arrows indicate the direction of increasing $x$. 
2.4 Self-similar solutions

2.4.1 Refomulation of the problem

One can reduce the system of partial differential equations (2.20) to a system of
ordinary differential equations by introducing the similarity transformation:

\[
\begin{align*}
S(x, t) &= S(\eta), \quad u(x, t) = u(\eta), \quad F(x, t) = F(\eta) \\
& \text{where } \eta = \frac{x}{\sqrt{t}}, \quad -\infty < \eta < \infty.
\end{align*}
\]

(2.25)

Since \( \eta \to \infty \) as \( t \to 0 \) and \( x > 0 \), and \( \eta \to -\infty \) as \( t \to 0 \) and \( x < 0 \), the piece-wise constant initial conditions (2.21),(2.22) become boundary conditions at \( \eta = \pm \infty \). Hence
substituting (2.25) into (2.20), using initial conditions (2.21)-(2.22) and matching
conditions (2.24) one obtains the boundary value problem \((\mathcal{P})\): (we will use primes
to denote differentiation with respect to \( \eta \))

\[
\begin{align*}
\eta^2 S' &= F', \quad F = k(S)u' \quad \text{for } \eta < 0 \text{ and } \eta > 0, \\
u \in p_c^+(S) - p_c^-(S) \cdot \text{sign} \left[ -\eta(H(S) + u) \right], \\
(S(\eta), u(\eta)) &\to E_l \quad \text{as } \eta \to -\infty, \\
(S(\eta), u(\eta)) &\to E_r \quad \text{as } \eta \to \infty, \\
u(0^-) &= u(0^+), \quad F(0^-) = F(0^+).
\end{align*}
\]

(2.26)-(2.30)

Here flux \( F \) has been redefined: it is \( \sqrt{t} \) times the original flux. To obtain (2.27) we used
\[
\text{sign} \left[ \partial_t (H(S) + u) \right] = \text{sign} \left[ -\frac{\eta}{2t} (H(S) + u)' \right] = \text{sign} \left[ -\eta(H(S) + u)' \right].
\]

In the formulation of Problem \((\mathcal{P})\) the capillary pressure \( u \) and the flux \( F \) are continuous. Integrating by parts the first equation in (2.26), the flux continuity implies that \( \eta S(\eta) \) is continuous as well. Hence the saturation \( S \) can only be discontinuous at \( \eta = 0 \).

When discussing the solutions of Problem \((\mathcal{P})\) we shall represent them as trajectories \( (S(\eta), u(\eta)) : -\infty < \eta < \infty \) in the \((S, u)\) plane. The trajectories run from \( E_l \) (as \( \eta \to -\infty \)) to \( E_r \) (as \( \eta \to \infty \)). In the figures below the arrows indicate the direction of increasing \( \eta \) (or increasing \( x \)).

Remark 2.4. The smoothness of the coefficients in (2.26)-(2.29) implies that the
functions \( S,F,u \) are smooth when \( \eta \neq 0 \) and the equations are satisfied in the classical sense except at points where \((S,u)\) moves from one of the sets \( \mathcal{H}, \partial \mathcal{H}^{\text{im}}, \partial \mathcal{H}^{\text{dr}} \) to another.
2.4.2 Classification of possible solutions in the half columns $\mathbb{R}^-$ and $\mathbb{R}^+$

Figure 2.5: (a) Possible solutions of Problem $(\mathcal{P})$ emerging from $E_l \in \mathcal{H}^{(i)}$. The arrows indicate increasing $\eta$. (b) Switch from imbibition to scanning curve.

We first consider all possible solutions in $\mathbb{R}^-$, that satisfy boundary condition (2.28). Depending upon the location of the initial condition $E_l$ in the $(S,u)$ plane, we distinguish three cases:

Case 1. Initial condition on main wetting curve: $E_l \in \partial \mathcal{H}^{im}$ (i.e. $u_l = p_c^{(i)}(S_l)$)

Then there are four possibilities, indicated in Figure 2.5a.

1.1 Uniform and constant $(S,u)$ in $\mathbb{R}^-$: $(S(\eta), u(\eta)) = E_l$ for all $\eta < 0$;

1.2 Imbibition on main imbibition curve: $(S(\eta), u(\eta)) \in \partial \mathcal{H}^{im}$ and $S'(\eta) > 0$ for all $\eta < 0$;

1.3 Drainage on scanning curve: $(S(\eta), u(\eta)) \in \mathcal{H}$ and $S'(\eta) < 0$ for all $\eta < 0$. This implies (see equation (2.9)),

$$H(S(\eta)) + u(\eta) = H(S_l) + u_l$$

for all $\eta < 0$;

1.4 Drainage on scanning and main drainage curves: There exists $\eta_0 < 0$ so that

$(S(\eta), u(\eta)) \in \mathcal{H}$ and $S'(\eta) < 0$ for $-\infty < \eta < \eta_0$,

$(S(\eta), u(\eta)) \in \partial \mathcal{H}^{dr}$ and $S'(\eta) < 0$ for $\eta_0 \leq \eta < 0$. 


One may wonder why in Case 1.2 the trajectory stays on $\partial H^{\text{im}}$ for all $\eta < 0$. Suppose it does not. Then there exists a reversal point $\eta^* < 0$ at which the trajectory switches from the main wetting curve to a drying scanning curve, as in Figure 2.5b. Since $S'(\eta) > 0$ for $-\infty < \eta < \eta^*$, it follows from (2.26) that $F'(\eta) < 0$ for $-\infty < \eta < \eta^*$. Using $F(-\infty) = 0$, we have $F(\eta^*) < 0$. But this would imply $u'(\eta^*) = \frac{F(\eta^*)}{k(S(\eta^*))} < 0$, contradicting the reversal of direction at $\eta^*$.

This argument can be used repeatedly to show that $S'(\eta)$ cannot change its sign in $\mathbb{R}^-$ which eliminates all other possibilities except the ones presented in Case 1.

**Case 2. Initial condition on scanning curve:** $E_l \in \mathcal{H}$ (i.e. $p^{(i)}_c(S_l) < u_l < p^{(d)}_c(S_l)$)

Now there are three possibilities, see Figure 2.6.

2.1 Uniform and constant $(S, u)$ in $\mathbb{R}^-$: $(S(\eta), u(\eta)) = E_l$ for all $\eta < 0$;

2.2 Imbibition or drainage on scanning curve: $(S(\eta), u(\eta)) \in \mathcal{H}$ and $S'(\eta) \leq 0$ for all $\eta < 0$. Again this implies

$$H(S(\eta)) + u(\eta) = H(S_l) + u_l$$

for all $\eta < 0$;

2.3 Imbibition on scanning and main imbibition curves or drainage on scanning and main drainage curves: There exists $\eta_0 < 0$ so that

$$(S(\eta), u(\eta)) \in \mathcal{H}$$

for $-\infty < \eta < \eta_0$,

and either $(S(\eta), u(\eta)) \in \partial \mathcal{H}^{\text{im}}$ for $\eta_0 \leq \eta < 0$ and $S'(\eta) > 0$ for all $-\infty < \eta < 0$, or $(S(\eta), u(\eta)) \in \partial \mathcal{H}^{\text{dr}}$ for $\eta_0 \leq \eta < 0$ and $S'(\eta) < 0$ for all $-\infty < \eta < 0$.

![Figure 2.6: Possible solutions of Problem (P) emerging from $E_l \in \mathcal{H}$.](image-url)
Case 3. Initial condition on main drying curve: $E_l \in \partial \mathcal{H}^{\text{dr}}$ (i.e. $u_l = p^{(d)}_c(S_l)$)

As in Case 1 there are four possibilities, see Figure 2.7.

3.1 Uniform and constant $(S, u)$ in $\mathbb{R}^-$: $(S(\eta), u(\eta)) = E_l$ for all $\eta < 0$;

3.2 Drainage on main drainage curve: $(S(\eta), u(\eta)) \in \partial \mathcal{H}^{\text{dr}}$ and $S'(\eta) < 0$ for all $\eta < 0$;

3.3 Imbibition on scanning curve: $(S(\eta), u(\eta)) \in \mathcal{H}$ and $S'(\eta) > 0$ for all $\eta < 0$, again giving

$$H(S(\eta)) + u(\eta) = H(S_l) + u_l$$

for all $\eta < 0$;

3.4 Imbibition on scanning and main imbibition curves: There exists $\eta_0 < 0$ so that

$$(S(\eta), u(\eta)) \in \mathcal{H}$$
$$S'(\eta) > 0 \text{ for } -\infty < \eta < \eta_0,$$

$$(S(\eta), u(\eta)) \in \partial \mathcal{H}^{\text{im}} \text{ and } S'(\eta) > 0 \text{ for } \eta_0 \leq \eta < 0.$$

A similar distinction of possible solutions can be made with respect to $E_r$. We omit the details. So far we have classified all possible piece-wise solutions in $\mathbb{R}^-$ and $\mathbb{R}^+$. To combine the solutions in the two half columns we present one final observation.
2.4.3 Direction of flow

Except for the trivial case $S(\eta) = S_l$ for $\eta < 0$ and $S(\eta) = S_r$ for $\eta > 0$ (implying that $F(\eta) = 0$ for all $-\infty < \eta < \infty$), the sign of $S'(\eta)$ is either strictly positive or strictly negative in the half columns $R^-$ and $R^+$. This was demonstrated in Section 2.4.2, see also Figure 2.5b. We now show that this implies that the flux $F$ cannot change its sign in the whole column: either $F(\eta) > 0$ or $F(\eta) < 0$ for all $-\infty < \eta < \infty$.

Suppose $S'(\eta) < 0$ for $\eta < 0$. Then (2.26) implies $F'(\eta) > 0$ for $\eta < 0$. Using this and the flux $F(-\infty) = 0$ we find $F(\eta) > 0$ for all $\eta < 0$. In particular, by continuity of the flux, $F(0) > 0$. This observation, the fact that $F'(\eta)$ has a fixed sign for $\eta > 0$ and $F(\infty) = 0$ imply $F(\eta) > 0$ for all $\eta > 0$. A similar argument is used when $S'(\eta) > 0$ for $\eta < 0$. Since $u_r - u_l = \int_{-\infty}^{\infty} u' = \int_{-\infty}^{\infty} \frac{F(\eta)}{K(S(\eta))}$, we conclude

4.1 If $u_r > u_l$ then $F(\eta) > 0$ and $u'(\eta) > 0$ for all $\eta \in \mathbb{R}$.

4.2 If $u_r < u_l$ then $F(\eta) < 0$ and $u'(\eta) < 0$ for all $\eta \in \mathbb{R}$.

Remark 2.5. **Redistribution and play-type hysteresis:** With play-type hysteresis the expression (2.4) is used. Since no time derivative of capillary pressure is involved, the redistribution problem only requires an initial saturation given by (2.21). In case of ‘unconventional’ flow, as suggested in [201], the $(S, u)$ profile will lie on scanning curves for $x \neq 0$ which means that the saturation is constant for $x < 0$ and $x > 0$. Using this result in (2.20a) we get that, $F(x, t)$ is constant for all $x \in \mathbb{R}$ and $t > 0$. Since $F(\pm \infty, t) = 0$ (any other value would give an unbounded pressure at $x = \pm \infty$ from (2.20b)) we have $F(x, t) = 0$ for all $x \in \mathbb{R}$ and $t > 0$, meaning that the capillary pressure $u = u_0 = constant$ for all $x \in \mathbb{R}$. This observation has the following consequence:

![Figure 2.8: Possible solutions with play-type hysteresis.](image-url)
If \( p_c^{(i)}(S_l) \leq p_c^{(d)}(S_r) \), then any horizontal connection is possible as indicated in Figure 2.8. The corresponding saturation profile is ‘frozen’ in the sense that:

\[
S(x, t) = \begin{cases} 
S_l & \text{for } x < 0, \\
S_r & \text{for } x > 0, 
\end{cases} 
\tag{2.31}
\]

for all \( t > 0 \). The capillary pressure is a undetermined constant \( u_0 \) as long as \( p_c^{(i)}(S_l) \leq u_0 \leq p_c^{(d)}(S_r) \). If however \( p_c^{(i)}(S_l) \geq p_c^{(d)}(S_r) \) then the solution is given by classical Philip construction.

Interestingly, the saturation profile remains frozen even when the domain is finite. If the domain is \([-1, 1]\) and Neumann conditions \( F(\pm 1, t) = 0 \) (no flow) are imposed at the boundaries for \( t > 0 \), then by the same argument one obtains saturations as in (2.31).

The cases discussed above for \( E_l \), and likewise for \( E_r \), are the building blocks in the construction of the full solution. In Section 2.4.4 we consider \( E_l \in \partial H_{im} \) and \( E_r \in \partial H_{dr} \). In Section 2.4.5 \( E_l \) and \( E_r \) are arbitrarily chosen: i.e. inside \( \mathcal{H} \) and/or on the boundaries \( \partial H_{im} \) and \( \partial H_{dr} \).

### 2.4.4 Construction of combined solution for initial conditions on the main drainage and imbibition curves: \( E_r \in \partial H_{dr}, E_l \in \partial H_{im} \)

With reference to Figure 2.9 we fix a point \( E_r \in \partial H_{dr} \) and let \( S_* \) be such that \( p_c^{(i)}(S_*) = u_r \). Further the curve \( H(S) + u = H(S_r) + u_r \) intersects \( \partial H_{im} \) at \( S = S_* \). Clearly \( 0 < S_* < S^* < 1 \). We consider five typical positions for \( E_l \) with respect to the given \( E_r \).

(i) \( 0 < S_l < S_* \). This is the classical Philip redistribution. In terms of Section 2.4.1 we have Case 1.2 for \( E_l \) and a similar case for \( E_r \). The value of \( u(0) = u_{0,1} \), where the saturation jumps, is determined from the flux continuity.

(ii) \( S_l = S_* \). Special case where no flow occurs and the system is in equilibrium. Here the flux \( F(\eta) = 0 \) for all \( \eta \in \mathbb{R} \). It is Case 1.1 for \( E_l \) and Case 3.1 for \( E_r \).

(iii) \( S_* < S_l < S^* \). This is unconventional flow since the dry (left) half column becomes drier and the wet (right) half column wetter if \( S_l < S_r \). Here \( (S(\eta), u(\eta)) \in \mathcal{H} \) for all \( \eta \in \mathbb{R} \). It is Case 1.3 for \( E_l \) and similar for \( E_r \). As before, the value of \( u(0) = u_{0,3} \) follows from flux continuity.

(iv) \( S_l = S^* \). As in (iii), but now \( \{(S(\eta), u(\eta)) : \eta \in \mathbb{R}\} \) belongs to a single scanning curve. Here the saturation is continuous at \( \eta = 0 \).
(v) $S^* < S_l < 1$. This is a rather complicated case. Depending on the position of $E_l$ with respect to $E_r$ three connections (i.e. solutions) are possible. One is shown in Figure 2.9 where $(S(\eta), u(\eta)) \in \mathcal{H}$ for all $\eta \in \mathbb{R}$. This situation relates to the Case 1.3 for $E_l$ and similar for $E_r$. But there are also connections possible using part of $\partial \mathcal{H}^{dr}$ (Case 1.4 for $E_l$) or part of $\mathcal{H}^{(i)}$ (case similar to 3.4 for $E_r$). These constructions will be discussed in detail in Section 2.4.5.

The saturation profiles corresponding to (i)-(v) are sketched in Figure 2.10. It is relatively straightforward to check that the constructions (i)-(v) are the only ones allowed based on our discussions in Section 2.4.2.

In [193] and [247] mathematical aspects of the Philip redistribution (Case (i))
are considered. Before analysing the other cases in detail, observe that for the extended model \((S(\eta), u(\eta)) \in \mathcal{H}\) implies that \(H(S(\eta)) + u(\eta)\) is constant which means that \(u' = -\frac{dH}{dS}(S)S'\). Substituting this in (2.26) we get:

\[
\frac{\eta}{2} S' + (D_{S'} S')' = 0,
\]

(2.32)

if \((S(\eta), u(\eta)) \in \mathcal{H}\), where

\[
D_{S'} := k(S) \frac{dH}{dS}(S).
\]

(2.33)

Case (iii) and (iv) are covered by (2.32) and (2.33) for \(\eta \in \mathbb{R}\), subject to boundary conditions \(S(-\infty) = S_l\) and \(S(+\infty) = S_r\). Case (iv) is governed by the general theory in [253] whereas for Case (iii) we need to recall some of the arguments from [193, 247, 253]. To show how to construct the solutions in general, let us briefly consider the unconventional Case (iii). Then \((S(\eta), u(\eta)) \in \mathcal{H}\) for \(\eta \in \mathbb{R}\) and so for the capillary pressure we have:

\[
u = (H(S_l) + u_l) - H(S)\quad \text{in } \mathbb{R}^-,
\]

(2.34)

\[
u = (H(S_r) + u_r) - H(S)\quad \text{in } \mathbb{R}^+.
\]

(2.35)

Now consider the sub-problems

\[
(\mathcal{P}^-) \quad \begin{cases} 
\frac{\eta}{2} S' + (D_{S'} S')' = 0 \text{ on } \mathbb{R}^- \\
S(-\infty) = S_l, \quad S(0^-) = S^- < S_l, \\
u \text{ satisfies (2.34);}
\end{cases}
\]

(2.36)

(2.37)

(2.38)

and

\[
(\mathcal{P}^+) \quad \begin{cases} 
\frac{\eta}{2} S' + (D_{S'} S')' = 0 \text{ on } \mathbb{R}^+, \\
S(0^+) = S^+ > S_r, \quad S(\infty) = S_r, \\
u \text{ satisfies (2.35).}
\end{cases}
\]

(2.39)

(2.40)

(2.41)

In these problems the saturation \(S^-\) and \(S^+\) will be chosen so that the capillary pressure \(u\) as well as the flux \(F = -D_{S'} S'\) are continuous across \(\eta = 0\). From (2.34) and (2.35) it follows that

\[
u(0) = H(S_l) + u_l - H(S^-) = H(S_r) + u_r - H(S^+).
\]

(2.42)

The algebraic conditions imply that if \(u(0)\) ranges from \(u_l\) to \(u_r\), then \(S^-\) ranges from \(S_l\) to \(S_{\min} < S_l\) and \(S^+\) from \(S_{\max} > S_r\) to \(S_r\), see Figure 2.11.
Boundary value problems like \((P^-)\) and \((P^+)\) have been studied in detail in [253]. There it is shown that the flux at \(\eta = 0^\pm\) depends continuously and monotonically on the boundary saturation \(S^\pm\). Denoting the flux at \(\eta = 0^\pm\) by,

\[ F^+(S^+) := F(\eta = 0^+; S^+), \]
\[ F^-(S^-) := F(\eta = 0^-; S^-), \]  

then \(F^+(S^+):=\) is continuous and strictly increasing for \(S_r \leq S^+ \leq S_{\text{max}}\) with \(F^+(S_r) = 0\), and \(F^-(S^-)\) is continuous and strictly decreasing for \(S_{\text{min}} \leq S^- \leq S_l\) with \(F^-(S_l) = 0\).

Writing the fluxes as functions of \(u(0)\) we have:

- \(F^+(u_r) = 0\) and \(F^+(u(0))\) is non-negative, continuous and strictly decreasing for \(u_l \leq u(0) \leq u_r\).
- \(F^-(u_l) = 0\) and \(F^-(u(0))\) is non-negative, continuous and strictly increasing for \(u_l \leq u(0) \leq u_r\).

Hence there exists a unique capillary pressure \(u(0) = u_0\) where the fluxes intersect, as in Figure 2.12. This pressure uniquely determines saturations \(S^+\) and \(S^-\) by (2.42). Taking the composite function of solutions of \((P^-)\) and \((P^+)\) completes the construction of Case (iii).

### 2.4.5 Construction of combined solution for arbitrary initial conditions: \(E_l,E_r \in \mathcal{H} \cup \partial \mathcal{H}^{\text{im}} \cup \partial \mathcal{H}^{\text{dr}}\)

For redistribution positive and negative directions are interchangeable. Therefore, without loss of generality we may assume

\[ u_l < u_r. \]  

(2.45)
Again we fix $E_r$, but this time $E_r \in \mathcal{H} \cup \partial \mathcal{H}^{im} \cup \partial \mathcal{H}^{dr}$. With assumption (2.45), we can sort out all typical cases based on the location of $E_l$. For this purpose, and with reference to Figure 2.13, the sets $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$ are introduced. Let $S^*$ and $S_*$ be defined as before and let the curve $H(s) + u = H(p_c^{(d)}(u_r)) + u_r$ intersect $u = p_c^{(i)}(S^*)$ at $S = S^\dagger$. The formal definitions are:

\[ u_{0,1}, u_{0,2}, u_{0,3} \]

\[ E_{l,1} \in \mathcal{H}_1 \text{ (Case A)} \]
\[ E_{l,2} \in \mathcal{H}_2 \text{ (Case B(b))} \]
\[ E_{l,3} \in \mathcal{H}_3 \text{ (Case C(c))} \]

Figure 2.13: The redistribution scenario for general cases.

Figure 2.12: Intersection of fluxes yielding the unique capillary pressure $u(0) = u_0$. 
Definition 2.3.

\[ \mathcal{H}_1 = \{(S, u) : 0 < S < 1, \max\{p_c^{(i)}(S), p_c^{(i)}(S^*)\} \leq u \leq \min\{u_r, H(p_c^{(d)^{-1}}(u_r)) + u_r - H(S)\}\}, \]
\[ \mathcal{H}_2 = \{(S, u) : 0 < S < 1, \max\{p_c^{(i)}(S^*), H(p_c^{(d)^{-1}}(u_r)) + u_r - H(S)\} \leq u \leq p_c^{(d)}(S)\}, \]
\[ \mathcal{H}_3 = \{(S, u) : 0 < S < 1, p_c^{(i)}(S) \leq u \leq \min\{p_c^{(i)}(S^*), p_c^{(d)}(S)\}\}. \]

Accordingly we distinguish

**Case A:** \( E_l \in \mathcal{H}_1 \)

This situation is similar to Case (iii) in Section 2.4.4. Here \( (S(\eta), u(\eta)) \in \mathcal{H}_1 \cap \mathcal{H} \) for all \( \eta \in \mathbb{R} \). Hence one uses Problems \((\mathcal{P}^-)\) and \((\mathcal{P}^+)\) to determine the saturations \( S^- \) and \( S^+ \) and the pressure \( u(0) \) for which the flux is continuous at \( \eta = 0 \). Figure 2.13 shows the trajectory running from \( E_l = E_{l,1} \) to \( E_r \). It is comprised of part of the scanning curve through \( E_{l,1} \) and part of the scanning curve through \( E_r \). At \( \eta = 0 \) is the horizontal switch where \( u(0) = u_{0,1} \).

**Case B:** \( E_l \in \mathcal{H}_2 \)

For left states in this set there are two possibilities. One is as described above for \( E_l \in \mathcal{H}_1 \), that is \( (S(\eta), u(\eta)) \in \mathcal{H} \) for all \( \eta \in \mathbb{R} \). We call this possibility ‘Case B(a)’. The other possibility is more involved and needs further attention. It occurs when \( E_l \) is close to \( \partial \mathcal{H}^{dr} \). Let \( \hat{S} \) denote the saturation at which the scanning curve through \( E_l \) intersects \( \partial \mathcal{H}^{dr} \), and let \( \hat{u} := p_c^{(d)}(\hat{S}) \) be the corresponding suction. Then for \( \hat{S} \leq S \leq S_l \),

\[ u = H(S_l) + u_l - H(S). \quad (2.46) \]

Now suppose, as in Figure 2.14(a), that there is a switch from the scanning curve through \( E_l \) to the scanning curve through \( E_r \). Then (2.34) holds with \( \hat{S} < S^- < S_l \) and in terms of the fluxes one would have a unique intersection at \( u(0) = u_0 \) satisfying \( u_l < u_0 < \hat{u} \) (Figure 2.14(b)). But what if \( F^+(\hat{u}) > F^-(\hat{u}) \) as in Figure 2.14(c)? Then the construction fails. We call this case: Case B(b).

To resolve it one needs to follow \( u = p_c^{(d)}(S) \) for \( S < \hat{S} \). This gives in the left column the suction saturation relation,

\[ u = \begin{cases} 
H(S_l) + u_l - H(S) & \text{for } \hat{S} < S \leq S_l, \\
p_c^{(d)}(S) & \text{for } S \leq \hat{S}.
\end{cases} \quad (2.47) \]
In the right hand column (\(\mathbb{R}^+\)) relation (2.35) still holds. This leads to the same problem (\(\mathcal{P}^+\)) in (\(\mathbb{R}^+\)), but to a modified problem in \(\mathbb{R}^-\):

\[
(\mathcal{P}^-) \begin{cases}
\frac{\eta}{2} S' + (\hat{D}(S) S)' = 0 \text{ in } \mathbb{R}^-,
S(-\infty) = S_l, \quad S(0^-) = S^- < S_l,
u \text{ satisfies (2.47) in } \mathbb{R}^-;
\end{cases}
\]

where

\[
\hat{D}(S) = \begin{cases}
 k(S) \frac{dH(S)}{dS} & \text{for } S < S_l
-k(S) \frac{dp_c(d)}{dS}(S) & \text{for } S \leq \hat{S}
\end{cases}.
\]

Here the diffusivity \(\hat{D}\) is in general discontinuous at \(\hat{S}\). But the results in [253] only require boundedness of the diffusivity. Thus with the notation of Section 2.4.4 (where we use \(\hat{F}^-\) to denote the flux in problem (\(\mathcal{P}^-\))), we have that \(\hat{F}^-(S^-)\) is strictly decreasing and continuous for \(S^- < S_l\) with \(\hat{F}^-(S_l) = 0\). In terms of the pressure \(u(0)\), the flux is strictly increasing and continuous for \(u_l \leq u(0) \leq u_r\) with \(\hat{F}^-(u_l) = 0\). Thus, as in Figure 2.15, the fluxes \(\hat{F}^-\) and \(F^+\) intersect at a unique suction \(u(0) = u_0\). The composite function of the solution of (\(\mathcal{P}^-\)) and (\(\mathcal{P}^+\)) describe the case of redistribution.

![Figure 2.14: (a) Possible switch with \(u_l < u_0 < \hat{u}\); (b) Intersection of fluxes; (c) no intersection with \(F^+(\hat{u}) > F^-(\hat{u})\).](image-url)
Figure 2.15: Intersection of the fluxes with $F^-$ and $F^+$

Figure 2.16: Behaviour of saturation $S$ and pressure $u$ between $E_{l,2}$ and $E_r$. Figure 2.16(a) shows the trajectory running from $E_l = E_{l,2}$ to $E_r$. It is composed of part of the scanning curve through $E_{l,2}$, part of $\partial H^{dr}$ and then part of scanning curve through $E_r$. The horizontal segment or switch is at $\eta = 0$ when $u(0) = u_{0,2}$. The corresponding saturation and suction are sketched in Figure 2.16(b) as functions of $\eta$. Due to the discontinuity in $D(S)$ at $S = \hat{S}$, the saturation has a kink at $S = \hat{S}$ corresponding to $\eta = \eta_0$ ($\eta_0$ is defined in Case 2.3, Section 2.4.2). The suction has a kink at $\eta = 0$ due to the jump in saturation.
2.5 Numerical study

Case C: $E_l \in \mathcal{H}_3$

There are two arrangements, denoted by Case C(a) and Case C(b), that are similar to the Cases B(a) and B(b) respectively. But in this set one has $u_l < p_c^{(i)}(S^*)$, implying that a third type of construction is possible where part of the two scanning curves, $\partial H^\text{dr}$ as well as $\partial H^\text{im}$ are being used. We call this ‘Case C(c)’.

Let $u^* := p_c^{(i)}(S^*)$. Case C(c) arises if

$$F^+(u = u^*) < \hat{F}^-(u = u^*),$$

(2.48)

where the fluxes $F^+$ and $\hat{F}^-$ are defined in the description of previous cases. If (2.48) holds, then the fluxes cannot intersect when $u^* < u(0) < u_r$, see Figure 2.17. To resolve this situation one needs to modify Problem $\mathcal{P}^+$ by including $p_c^{(i)}$, similar to the definition of Problem $\mathcal{P}^-$. The remaining argument is omitted since it is almost identical to the argument used for $E_l \in \mathcal{H}_2$. Figure 2.13 shows the trajectory running from $E_l = E_{l,3}$ to $E_r$.

![Figure 2.17: Non-intersection of the fluxes when $E_l \in \mathcal{H}_3$.](image)

2.5 Numerical study

In this section we present a numerical approach for the redistribution problem and use our theoretical findings to classify and validate the computational results. For sufficiently large dimensionless column half-length $W > 0$ we consider the initial-boundary value problem (compare (2.20)-(2.22))

$$\partial_t S + \partial_x F = 0 \text{ for } |x| < W, t > 0,$$

(2.49a)

$$F = k(S) \partial_x u \text{ for } |x| < W, t > 0,$$

(2.49b)

$$u \in p_c^+ (S) - p_c^- (S) \cdot \text{sign}(\partial_t H(S) + \partial_t u);$$

(2.49c)
with

\[
S(x,0) = \begin{cases} S_l & \text{for } -W < x < 0, \\ S_r & \text{for } 0 < x < W, \end{cases}
\]

(2.50)

\[
u(x,0) = \begin{cases} u_l & \text{for } -W < x < 0, \\ u_r & \text{for } 0 < x < W, \end{cases}
\]

(2.51)

and

\[F(\pm W, t) = 0 \text{ for } t > 0.\]

(2.52)

For sufficiently small time \(t\), the zero-flux boundary conditions (2.52) have a negligible influence on the redistribution process. Hence, for small \(t\), a solution of (2.49)-(2.52) behaves as if the domain is unbounded and is close to the self-similar solutions discussed in this paper. We will reveal all typical cases.

### 2.5.1 Numerical scheme

We start with replacing the sign(\(\cdot\)) graph in (2.49c) by a strictly increasing smooth function \(G_\varepsilon\) satisfying:

- \(dG_\varepsilon/d\zeta > 0\) in \(\mathbb{R}\) such that \(dG_\varepsilon(0)/\varepsilon = \frac{1}{\varepsilon}\) and \(\lim_{\varepsilon \to 0} G_\varepsilon(\zeta) = \text{sign}(\zeta)\) for \(\zeta \neq 0\).

- \(\lim_{\zeta \to \pm \infty} dG_\varepsilon(\zeta) = \gamma \varepsilon\) for some constant \(\gamma > 0\) and for each \(\varepsilon > 0\).

These properties allow us to define the inverse \(\Phi_\varepsilon = G_\varepsilon^{-1}\), so that (2.49c) can be written as:

\[
\partial_t H(S) + \partial_t u = \Phi_\varepsilon \left( \frac{p_c^+(S) - u}{p_c^-(S)} \right).
\]

(2.53)

Next we introduce the variable:

\[
v := H(S) + u, \text{ with } S = H^{-1}(v - u),
\]

(2.54)

and the function

\[
\Psi_\varepsilon(u, v) := \Phi_\varepsilon \left( \frac{p_c^+(H^{-1}(v - u)) - u}{p_c^{-}(H^{-1}(v - u))} \right).
\]

(2.55)

In terms of \(u, v\) and \(\Psi_\varepsilon\), equations (2.49) are transformed into

\[
\partial_t u = \frac{dH}{dS}(H^{-1}(v - u))\partial_x \left( k(H^{-1}(v - u))\partial_x u \right) + \Psi_\varepsilon(u, v),
\]

(2.56a)

\[
\partial_t v = \Psi_\varepsilon(u, v),
\]

(2.56b)
2.5 Numerical study

for \(|x| < W\) and \(t > 0\). Hence we have written (2.49) as a coupled system consisting of a parabolic equation for the capillary pressure \(u\) and an abstract ordinary differential equation for \(v\). This type of splitting is well-known in the mathematical literature, for instance see [222] and Chapters 4, 5 and 7.

Let the time interval \([0, T]\) be divided into \(N\) intervals of width \(\Delta t\) \((T = N\Delta t)\) and let \(w_n\) be the variable \(w\) at \(t = n\Delta t\), with \(1 \leq n \leq N\). We calculate \(v_n\) from the explicit time-discrete form of (2.56b) \((v_0(x) = v(x, 0) = H(S(x, 0)) + u(x, 0)):\)

\[
v_n = v_{n-1} + \Delta t \Psi^c(u_{n-1}, v_{n-1}), \quad \text{for} \quad 1 \leq n \leq N.
\]

(2.57)

We want to solve (2.56a) implicitly for stability. For this, the equation (2.56a) needs to be linearized. We use a L-scheme [192] type linearisation technique along with inner iterations to solve for \(u_n\). This scheme is defined by (for \(i = 1, 2, \ldots\))

\[
(1 + L)u_n^i - \Delta t \frac{dH}{dS}(H^{-1}(v_n - u_n^{i-1}))\partial_x \left(k(H^{-1}(v_n - u_n^{i-1}))\partial_x u_n^i\right)
= Lu_n^{i-1} + \Delta t \Psi^c(u_n^{i-1}, v_n) + u_n^{i-1},
\]

(2.58)

with \(u_0^0 = u_{n-1}\). Here \(u_n^i\) is the \(i\)th iteration of the \(n\)th time step. If assumptions (A1)-(A4) are satisfied and if \(\partial_x u, \partial_u \Psi^c\) and \(\partial_x \Psi^c\) are bounded, then the scheme converges for \(L\) sufficiently large and for \(\Delta t\) sufficiently small [192]. In (2.58) we use finite differences for spatial discretization and the whole scheme was implemented in Matlab. In the computations we use

\(W = 100, \quad \Delta x = 0.1, \quad \Delta t = 0.001\) and \(L = 1\).

2.5.2 Numerical results

The numerical results are obtained for

\(k(S) = S^2\) and (for simplicity) \(H(S) = \frac{S}{\delta}\) with \(\delta = \frac{1}{40}\).

For \(p_c^{(d)}\) and \(p_c^{(i)}\) we took somewhat artificial expressions in order to visualize all possible cases:

\[
p_c^{(i)}(S) = \frac{1}{5} \left(\frac{1}{S} - 1\right) + \frac{2}{5} (1 - S)^2 \quad \text{and} \quad p_c^{(d)}(S) = \left(\frac{1}{S} - 1\right) + 6(1 - S)^2.
\]

Taking realistic (van Genuchten) expressions would make some of the cases hard to distinguish. Finally we use for \(\Phi^c\) the expression
Φε(ζ) = \begin{cases} 
\sqrt[3]{\varepsilon} + \frac{1}{\gamma \varepsilon}(\zeta - 1) & \text{for } \zeta > 1 \\
\varepsilon \zeta \sqrt{1 - (1 - (\varepsilon^2 \gamma)^{\frac{3}{2}})\zeta^2} & \text{for } \zeta \in [-1, 1], \\
-\sqrt[3]{\varepsilon} - \frac{1}{\gamma \varepsilon}(\zeta + 1) & \text{for } \zeta < -1 
\end{cases}

where \(\epsilon = 10^{-4}\) (fixed). Observe that \(\Phi_\varepsilon\) is continuously differentiable in \(\mathbb{R}\). The value of \(\gamma\) can be chosen small as long as \((S, u) \in \mathcal{H}\) (then \(\gamma = 1\)), but needs a large value when \((S, u)\) is on \(\partial \mathcal{H}^{\text{im}}\) or on \(\partial \mathcal{H}^{\text{dr}}\).

Given the size of the domain \((W = 100)\) we show the computational results at \(t = 1\). This is sufficiently small so that the zero-flux boundary conditions have no influence on the redistribution process. We visualize the results as trajectories in the \((S, u)\) plane (i.e. \((S(x, t), u(x, t))\) where \(x\) runs from \(-W\) to \(W\)) and as \(S\) vs. \(\eta\) profiles.

![Figure 2.18: Numerical results for \(E_r \in \partial \mathcal{H}^{\text{dr}}\) fixed and \(E_l \in \partial \mathcal{H}^{\text{im}}\) variable, at \(t = 1\). Here \(S_r = 0.45\) and \(S_{l,1} = 0.30\) (unconventional flow); \(S_{l,2} = 0.52 = S^*\) (saturation is continuous) and \(S_{l,3} = 0.60\) (conventional flow corresponding to Case C(a)).](image)

Figure 2.18 shows results where \(E_r \in \partial \mathcal{H}^{\text{dr}}\) is fixed and where \(E_l \in \partial \mathcal{H}^{\text{im}}\) is varied. Note that the green trajectory represents unconventional flow because the trajectory moves to the left when leaving \(E_l\) and when entering \(E_r\). This means that the dry half column becomes drier and the wet half column becomes wetter.

Figure 2.19 shows results for \(E_l, E_r \in \mathcal{H} \cup \partial \mathcal{H}^{\text{im}} \cup \partial \mathcal{H}^{\text{dr}}\). Here we took \(\gamma = 100\) to ensure convergence of the iterations in (2.58). The cases discussed in Section 2.4 are accurately recovered by the computations. This validates the analysis and ex-
2.6 Conclusion

explains the complex behaviour of the computed saturation and pressure. In fact, the agreement is excellent.

![Figure 2.19: Numerical results for $E_l, E_r \in H \cup \partial H \text{Im} \cup \partial H \text{dr}$, where $u_r = 3$ is fixed. With reference to Section 2.4 we have Case A: $S_r = 0.4$, $E_l = (0.35, 2)$; Case B(b): $S_r = 0.25$, $E_l = (0.65, 1)$; Case C(c): $S_r = 0.15$, $E_l = (0.8, 0.3)$; Philip: $S_r = 0.075$, $E_l = (0.9, 0.17)$.](image)

2.6 Conclusion

In this paper we discussed different hysteresis models for multiphase flow through porous media. To incorporate the effect of non-vertical scanning curves in a simple closed form, we proposed an extension to the play-type hysteresis model and showed that this model resembles the experimental scanning curves accurately. We outlined different properties of the model and discussed available physical and numerical justifications for the model.

After this we investigated horizontal redistribution in an infinite coloumn in the context of hysteresis. In this problem, the two halves of the infinite horizontal porous column have different but constant initial saturation and pressure conditions that causes redistribution to occur. It was pointed out that existing models cannot give a complete description of the redistribution phenomenon. The extended hysteresis model was used to analyse the problem and the resulting system of equations was simplified using a similarity transformation. By distinguishing all possible cases and then using the flux and pressure continuity criterion repeatedly, we constructed unique solutions for the redistribution problem. In fact, we showed that redistribution will always take place, even for unconventional cases, if the initial pressure condition is different in the two halves. Moreover, we categorized all possible scenarios of redistribution into different cases.
Finally, a numerical scheme was proposed for the regularised nonlinear system of equations arising from the extended play-type hysteresis model that converges irrespective of initial guesses. Numerical results from the scheme corroborated our analytical findings.
Chapter 3

Existence and properties of the solutions of the extended play-type hysteresis model

3.1 Introduction

This chapter discusses the well-posedness of the extended play-type hysteresis (henceforth called EPH) model, proposed in Chapter 2 to incorporate the effects of capillary hysteresis in porous flow problems. In the context of flow of water through soil, the model reads in the dimensionless form,

\[
\frac{\partial}{\partial t} S + \nabla \cdot [k(S)(\nabla p - \hat{g})] = 0, \tag{3.1a}
\]

\[
p \in \frac{1}{2}(p_c^{(d)}(S) + p_c^{(i)}(S)) - \frac{1}{2}(p_c^{(d)}(S) - p_c^{(i)}(S)) \cdot \text{sign}(\partial_t H(S) + \partial_t p). \tag{3.1b}
\]

Equation (3.1a) is the Richards equation (1.20) written with water saturation \(S\) and capillary pressure \(p\) (which in this case is simply the additive inverse of water pressure) as primary unknowns. The absolute permeability \(\kappa\), porosity \(\phi\), and capillary number \(N_c\) in (1.20) are assumed to be 1 in this chapter, for the sake of simplicity. Equation (3.1b) represents the EPH model. The functions \(p_c^{(i)}, p_c^{(d)}, H: (0, 1) \rightarrow \mathbb{R}\) and \(k: (0, 1) \rightarrow \mathbb{R}^+\) are determined based on experiments [20, 45, 115, 259] and \(\text{sign}(\cdot)\) is the multivalued signum graph given in (1.25).

To analyse the model (3.1) above, we will study another class of problems, i.e.

\[
\frac{\partial}{\partial t} u + \nabla \cdot \mathbf{F}(u, v) = \nabla \cdot [\mathbf{G}(u, v) \nabla u] + \Psi_1(u, v), \tag{3.2a}
\]

\[
\frac{\partial}{\partial t} v = \Psi_2(u, v), \tag{3.2b}
\]

completed with suitable boundary and initial conditions. The equivalence of (3.2) with the regularised version of (3.1) will be shown in Section 3.2. System (3.2) has
an interesting mathematical structure as it consists of a nonlinear parabolic partial differential equation in a two-way coupling with an ordinary differential equation. This implies that standard techniques, such as the $L^1$ contraction [181] principle or the Kirchhoff transform [10] can not directly be applied to this system. Moreover, the maximum principle can be violated for the system (3.2) which gives rise to the overshoot phenomenon that will be discussed in Chapters 4 and 5.

Further motivation for our analysis comes from establishing a physically consistent, easily implementable, numerically robust, and for the main focus of the current chapter, well-posed model for hysteresis. The phenomenon of capillary hysteresis has been discussed in detail in Sections 1.2.1 and 2.1. In the standard model (see Section 1.1), hysteresis is often disregarded due to its complex nature and capillary pressure $p$ is assumed to be a function of $S$, i.e.

$$p = p_c(S).$$  \hfill (3.3)

Equation (3.1a), along with (3.3), constitutes a nonlinear diffusion problem with diffusivity $-k(S)p_c(S)$. The existence of weak solutions for such problems are studied in [9, 10] and uniqueness results are found in [56, 181]. However, as shown in Chapters 1 and 2, disregarding hysteresis introduces significant modelling errors for the problem. To account for hysteresis in a comparatively simple way, in [24], the play-type hysteresis model was proposed based on thermodynamic considerations. It reads,

$$p \in \frac{1}{2}(p_c^{(d)}(S) + p_c^{(i)}(S)) - \frac{1}{2}(p_c^{(d)}(S) - p_c^{(i)}(S)) \cdot \text{sign}(\partial_t S).$$  \hfill (3.4)

Observe that, this gives $\partial_t S \geq 0$ when $p = p_c^{(i)}(S)$ and $\partial_t S \leq 0$ when $p = p_c^{(d)}(S)$. If $p_c^{(i)}(S) < p < p_c^{(d)}(S)$ then $\partial_t S = 0$, which implies that the scanning curves for this model are vertical lines having constant saturation, see Figure 2.1 (right).

The play-type hysteresis model has been studied extensively analytically due to its local and closed-form structure in contrast to other more complicated models such as the Lenhard-Parker model [184]. If the sign graph in (3.4) is regularised, then together with (3.1a), it constitutes a nonlinear pseudo-parabolic equation for $S$ [53, 149, 163, 222]. Existence results for such pseudo-parabolic equations can be found in [30–32]. The existence of solutions of the play-type hysteresis model for degenerate capillary pressure and permeability functions is shown in [222]. Existence for the two-phase case is discussed in [141]. For the constant relative permeability case, the existence of weak solutions for the unregularised play-type hysteresis model is shown in [218]. In the same paper, an upscaled version of the play-type model has also been derived. Uniqueness is shown in [53] for the two-phase case with dynamic capillarity present. In [221] it is shown that the play-type hysteresis model does not define an $L^1$-contraction. This leads to unstable planar fronts. Travelling wave solutions are investigated in Chapters 4 to 7. For mathematical analysis, in many cases the pseudo-parabolic system is split into an elliptic equation coupled with an ordinary differential equation [31, 32, 54, 149, 222].
However, it was pointed out in Chapter 2 that due to the approximation of scanning curves by vertical segments, play-type hysteresis model makes certain physically inaccurate predictions. For example, it predicts that in many cases water will not redistribute when two columns having constant but different saturations are joined together. In Chapter 5, it will be shown that the model predicts infinitely many interior maxima of saturation (called overshoots in this context) for high enough injection rates through a long column. This contradicts observations from experiments [72]. Moreover, it is well documented that the numerical methods incorporating the play-type model become unstable if the regularisation parameter is sent to zero [44, 216, 251, 271]. This motivated an extension of the play-type hysteresis model (EPH) presented in Chapter 2 by (2.8). The model is thermodynamically consistent, as evidenced by [24], and was used in Chapter 2 to cover all cases of horizontal redistribution.

In this chapter, we investigate the existence of weak solutions of the regularised EPH model and analyse the properties of the solutions. An alternative form of (3.1) is proposed in Section 3.2 and mathematical preliminaries are stated. In Section 3.3, existence of solutions for the model given by (3.2) is proved using Rothe’s method. In Section 3.4, it is shown that a maximum principle holds for the solution under certain assumptions. This also gives the existence of solutions for the degenerate EPH model. Section 3.5 is dedicated to investigating the behavior of the solutions when the regularisation parameter is passed to 0. It is shown that in the limit the solution is contained within physically consistent bounds.

### 3.2 Mathematical formulation

Let \( \Omega \in \mathbb{R}^d \) be a bounded open domain with \( \partial \Omega \in C^1 \). The \( L^2(\Omega) \) inner product in this domain is denoted by \( \langle \cdot, \cdot \rangle \), whereas \( \| \cdot \| \) and \( \| \cdot \|_p \) denotes the \( L^2(\Omega) \) and \( L^p(\Omega) \) norms respectively for \( 1 \leq p \leq \infty \). For any other space \( V(\Omega) \), the norm is denoted by \( \| \cdot \|_V \). Further, \( W^{k,p} \) denotes the Sobolev space containing functions that have up to \( k \)th order derivatives in \( L^p(\Omega) \). In particular, \( H^k(\Omega) = W^{k,2}(\Omega) \) and \( H^k_0 = \{ u \in H^k : u = 0 \text{ on } \partial \Omega \text{ in a trace sense} \} \). Let \( H^{-1}(\Omega) \) denote the dual of \( H^1_0 \) and \( \langle \cdot, \cdot \rangle \) the duality pairing of \( H^1_0(\Omega) \) with \( H^{-1}(\Omega) \). The space of function having up to \( \ell \)th order space derivatives \( \alpha \)-Hölder continuous, will be referred to as \( C^{\ell,\alpha}(\Omega) \) with \( C^\ell(\Omega) = C^{\ell,0}(\Omega) \).

Let \( T > 0 \) represent a maximum time with \( Q = \Omega \times (0,T) \). The Bochner space \( L^p(0,T;X(\Omega)) \) represents the space of functions \( u : Q \to \mathbb{R} \) having norm \( \| u \|_{L^p(0,T;X(\Omega))} := \left( \int_0^T \| u(t) \|_X^p \, dt \right)^{1/p} < \infty \). Finally we introduce the space

\[
\mathcal{W} = \{ u \in L^2(0,T;H^1_0(\Omega)) : \partial_t u \in L^2(0,T;H^{-1}(\Omega)) \}.
\]

Following [230], \( \mathcal{W} \) is compactly embedded in \( L^2(0,T;L^2(\Omega)) \), i.e. \( \mathcal{W} \hookrightarrow L^2(0,T;L^2(\Omega)) \).
The inequalities that will be used repeatedly in our analysis include Cauchy-Schwarz inequality; Poincaré inequality, with $C_{\Omega}$ denoting the constant appearing in it, i.e. $\|u\|_{H^1_0} > C_{\Omega} \|u\|$; Young’s inequality which states that for $a, b \in \mathbb{R}$ and $\sigma > 0$ one has
\[
abla \leq \frac{1}{2\sigma} a^2 + \frac{\sigma}{2} b^2,
\]
and finally the discrete Gronwall’s lemma, which states that if $\{y_n\}, \{f_n\}$ and $\{g_n\}$ are non-negative sequences and
\[
y_n \leq f_n + \sum_{0 \leq k < n} g_k y_k, \quad \text{for all } n \geq 0,
\]
then
\[
y_n \leq f_n + \sum_{0 \leq k < n} f_k g_k \exp \left( \sum_{k < j < n} g_j \right), \quad \text{for all } n \geq 0.
\]
In our notation $\{\cdot\}$ represents the positive part function, i.e. $|u| = \max\{u, 0\}$. Further, $C > 0$ will denote a generic constant throughout the paper.

### 3.2.1 Problem statement and assumptions

For the most part, in this chapter we study the well-posedness of the system (3.2) for suitable boundary and initial conditions. In completeness, it is written as

\[
\begin{aligned}
\mathcal{P}_S \quad & \begin{aligned}
\partial_t u + \nabla \cdot \tilde{F}(u, v) &= \nabla \cdot [\mathcal{D}(u, v) \nabla u] + \Psi_1(u, v) \quad \text{in } Q, \\
\partial_t v &= \Psi_2(u, v) \quad \text{in } \tilde{Q}, \\
u(0, \cdot) &= u_0(\cdot) \quad \forall \cdot \in \Omega, \\
v(\cdot, 0) &= v_0(\cdot) \quad \forall \cdot \in \Omega, \\
u &= 0 \quad \text{on } \partial \Omega \times [0, T].
\end{aligned}
\end{aligned}
\]

The relation between $\mathcal{P}_S$ and EPH is established in Section 3.2.2.

The properties of the functions $\mathcal{D}$, $\Psi_{1,2}$, $u_0$ and $v_0$ are as follows:

(A1) $\mathcal{D} \in C^1(\mathbb{R}^2)$; $0 < \mathcal{D} \leq \mathcal{D}(u, v) \leq \mathcal{D}_M < \infty$ for $u, v \in \mathbb{R}$.

(A2) $\tilde{F}: \mathbb{R}^2 \rightarrow \mathbb{R}^d$ with the $j$-th component $F_j \in C^1(\mathbb{R}^2)$ satisfying $|\tilde{F}| \leq F_M$ for some $F_M > 0$.

(A3) $|\Psi_j(u_1, v_1) - \Psi_j(u_2, v_2)| \leq \Psi_u|u_1 - u_2| + \Psi_v|v_1 - v_2|$ for $j \in \{1, 2\}$ and constants $\Psi_u, \Psi_v > 0$. Moreover, $\frac{\Psi_1(u, v) - \Psi_1(u_1, v_1)}{v - v_1}, \frac{\Psi_2(u, v) - \Psi_2(u_1, v_1)}{u - u_1} \leq 0$ for all $u, u_1, v, v_1 \in \mathbb{R}$.

(A4) $u_0 \in H^1_0(\Omega) \cap L^\infty(\Omega)$ and $v_0 \in H^1(\Omega) \cap L^\infty(\Omega)$, such that $\Psi_2(0, v_0) = 0$ at $\partial \Omega$.

Observe that, (3.7) combined with (A3) and (A4) imply that $v(t)$ restricted to $\partial \Omega$ remains unchanged for all $t > 0$. The weak solution of $\mathcal{P}_S$ is defined as
Definition 3.1. The pair \((u, v)\) with \(u \in \mathcal{W}\) and \(v \in H^1(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega))\) is a weak solution of \((\mathcal{P}s)\) if it solves for all \(\phi \in L^2(0, T; H^1_0(\Omega))\) and \(\xi \in L^2(Q)\)

\[
\begin{align*}
(P_2) & \quad \int_0^T (\partial_t u, \phi) + \int_0^T (\mathcal{P}(u, v) \nabla u, \nabla \phi) = \int_0^T (F(u, v), \nabla \phi) + \int_0^T (\Psi_1(u, v), \phi); \quad (3.10a) \\
& \quad (\partial_t v, \xi) = (\Psi_2(u, v), \xi); \quad (3.10b)
\end{align*}
\]

and \(u(0) = u_0\) and \(v(0) = v_0\).

Remark 3.1 (Boundary Conditions). For simplicity, a zero Dirichlet condition has been assumed at the boundary for our current analysis. Nevertheless, Definition 3.1 can be generalised to include Dirichlet, Neumann, and mixed type boundary conditions.

Remark 3.2 (Assumptions). The condition in (A3) that \(\Psi_j\) is decreasing with respect to both the variables \(u\) and \(v\) is not required for proving the existence of the weak solutions. It is used in Section 3.4 to prove that the solutions are bounded in \(L^\infty\). Similarly \(u_0, v_0 \in L^\infty(\Omega)\) is only used for proving the \(L^\infty\) bound.

### 3.2.2 Relation between the regularised EPH model and \((\mathcal{P}s)\)

Although (3.1) is closer to the expressions of the hysteresis models used in practice, \((\mathcal{P}s)\) is more convenient to analyse mathematically. We show below that the EPH model with the \(\text{sign}()\) graph regularised is a particular case of \((\mathcal{P}s)\). To be more precise, we start with assuming the following properties of the functions \(p_c^{(i)}\), \(p_c^{(d)}\) and \(k\) used in (3.1), (3.3) and (3.4), which are consistent with the data obtained from experiments [115, 170, 197].

\[\begin{align*}
& (P1) \quad p_c^{(i)}(1) = p_c^{(d)(1)} = p_c^{(d)}(1) \text{ and } \lim_{S \to 0} p_c^{(i)}(S) = \lim_{S \to 0} p_c^{(d)}(S) = \lim_{S \to 1} p_c^{(i)}(S) = \lim_{S \to 1} p_c^{(d)}(S) = \infty. \\
& (P2) \quad k \in C^1(\mathbb{R}); k(S) = k(0) \geq 0 \text{ for } S \leq 0, k(S) = k(1) \text{ for } S \geq 1 \text{ and } k'(S) > 0 \text{ for } 0 < S < 1.
\end{align*}\]

As seen from (2.56), the set of equations (3.1) cannot be reduced to the standard weak formulation used for partial differential equations. Thus, we consider an alternative expression to (3.1b) representing the EPH model. Completed with suitable boundary and initial conditions, the model reads

\[
\begin{align*}
(\text{EPH}) & \quad \begin{cases}
\partial_t S = \nabla \cdot [k(S)(\nabla p - \mathbf{g})] & \text{in } Q, \quad (3.11) \\
p \in \frac{1}{2}(p_c^{(d)}(S) + p_c^{(i)}(S)) - \frac{1}{2}(p_c^{(d)}(S) - p_c^{(i)}(S)) \cdot \text{sign} \left[ \partial_t (S + b(p)) \right] & \text{in } Q, \quad (3.12) \\
S(\cdot, 0) = S_0(\cdot), \quad p(\cdot, 0) = p_0(\cdot) & \text{in } \Omega, \quad (3.13) \\
p = 0 & \text{on } \partial \Omega \times [0, T]. \quad (3.14)
\end{cases}
\end{align*}
\]
Observe that, for relation (3.12) the scanning curves are given by \( S + b(p) = \text{constant} \), instead of \( H(S) + p = \text{constant} \) as used in Chapter 2. Moreover, if \( p = p_c^{(i)}(S) \) in some open subset of \( \Omega \), then from (3.12), \( \partial_t S + \partial_t b(p_c^{(i)}(S)) \geq 0 \) or \( (1 + b'(p_c^{(i)}(S))p_c^{(i)'}(S))\partial_t S \geq 0 \). Since (1.22) demands that \( \partial_t S \geq 0 \) in this case, for consistency \( 1 + b'(p_c^{(i)}(S))p_c^{(i)'}(S) > 0 \) has to be satisfied. Similar result holds if \( p = p_c^{(d)}(S) \) in some open subset of \( \Omega \). Combining these consistency criteria, similar to the bound (2.13) assumed for \( H(S) \) in Chapter 2, here we assume

(P3) \( b \in C^1(\mathbb{R}) \) with \( b(0) = 0 \) and

\[
0 < b'(p_c^{(j)}(S)) < -\frac{1}{p_c^{(j)'}(S)} \quad \text{for all } 0 < S < 1 \text{ and } j \in \{i, d\}. \quad (3.15)
\]

**Remark 3.3.** The inequality (3.15) also guarantees that any scanning curve passing through \((S_0, p_0)\) for \( S_0 \in (0, 1) \) and \( p_0 \in (p_c^{(i)}(S_0), p_c^{(d)}(S_0)) \) intersects \( p_c^{(i)} \) at some \( S_1 < 1 \) and \( p_c^{(d)} \) at some \( S_d \in (0, S_1) \).

For the initial and boundary conditions we assume:

(P4) \( S_0 \in H^1(\Omega) \) and \( p_0 \in H^1_0(\Omega) \). Moreover, an \( \epsilon > 0 \) exists such that \( \epsilon \leq S_0 \leq 1 - \epsilon \) and \( p_c^{(i)}(S_0) \leq p_0 \leq p_c^{(d)}(S_0) \) a.e. in \( \Omega \).

The condition \( p_c^{(i)}(S_0) \leq p_0 \leq p_c^{(d)}(S_0) \) comes from the physical constraint that \( p_0 \) stays intermediate to \( p_c^{(i)} \) and \( p_c^{(d)} \) when only hysteretic effects are considered.

**Remark 3.4.** If \( \lim_{S \to 0} p_c^{(j)}(S) = \infty \) for \( j \in \{i, d\} \) or \( k(0) = 0 \), then \( S = 0 \) at any interior point in the domain makes the problem degenerate, see Section 1.3.2 also. Similarly, the problem becomes degenerate at \( S = 1 \) since \( p_c^{(i)'}(1) = p_c^{(d)'}(1) = -\infty \). Moreover, \( S \) must satisfy the physical bound \( 0 \leq S \leq 1 \), otherwise \( p_c^{(i)}(S) \) and \( p_c^{(d)}(S) \) become ill defined. Treating the degeneracy and proving the physical bounds pose extra challenges in considering the problem (EPH) compared to (EPs).

Having stated the properties of the associated functions, we now show the equivalence of the regularised (EPH) model and (EPs). For this purpose, the following transformations are introduced:

\[
u := b(p) = \int_0^p b'(\varphi) d\varphi \quad \text{and} \quad v := S + b(p) = S + u. \quad (3.16)
\]

We recast (EPH) in terms of \( u \) and \( v \). Since \( \lim_{S \to 0} p_c^{(d)}(S) = \infty \) we get \( b(\infty) - b(p_c^{(d)}(1)) = \int_{p_c^{(d)}(1)}^\infty b'(p)p \leq -\int_{p_c^{(d)}(1)}^\infty \frac{1}{p_c^{(d)'}(p)p_c^{(i)'(p)}} dp \leq 0 \). Note that, if \( S \to 0 \) then \( v \to u \). From these observations, we define the terms \( U_M, V_M, U_m, V_m \) that will become important later;

\[
U_M = V_M = b(\infty), \quad U_m = b(p_c^{(i)}(1)) = b(p_c^{(d)}(1)), \quad V_M = 1 + U_m > V_m. \quad (3.17)
\]
Next, we express \( p = p_c^{(i)}(S) \) in terms of a relation between \( u \) and \( v \). From (3.16), \( p = p_c^{(i)}(S) \) implies \( v = v_i(u) := (p_c^{(i)})^{-1}(b^{-1}(u)) + u \). Since, \( v_i'(u) = \frac{1}{b'(p_c^{(i)}(S)p_c^{(d)}(S)} + 1 < 0 \) from (P3) for \( S = (p_c^{(i)})^{-1}(b^{-1}(u)) \), an inverse of \( v_i(u) \) is defined, denoted by \( \rho_i(u) \). In a similar way \( \rho_d(u) \) is defined. This is summarized into

\[
\rho_i = ((p_c^{(i)})^{-1} \circ b^{-1} + 1)^{-1}, \quad \rho_d = ((p_c^{(d)})^{-1} \circ b^{-1} + 1)^{-1}.
\]

From (P1) and (P3), one immediately obtains

(P5) There exists a constant \( M_\rho > 0 \) such that \( -M_\rho < \rho_i', \rho_d' < 0 \) for \( v \in [V_m, V_M] \). Further, \( \rho_d(v) > \rho_i(v) \) for all \( v \in (V_m, V_M) \); \( \rho_i(V_m) = \rho_d(V_m) = U_M \) and \( \rho_i(V_M) = \rho_d(V_M) = U_m \).

Figure 3.1 (right) plots \( \rho_i \) and \( \rho_d \) curves calculated from realistic \( p_c^{(i)} \) and \( p_c^{(d)} \) curves shown in Figure 3.1 (left). By definition, \( p = p_c^{(i)}(S) \) iff \( u = \rho_i(v) \) and \( p = p_c^{(d)}(S) \) iff \( u = \rho_d(v) \). Furthermore, \( \partial_t v = 0 \) implies \( p_c^{(i)}(S) \leq p \leq p_c^{(d)}(S) \) which is same as having \( \rho_i(v) \leq u \leq \rho_d(v) \). Thus, an equivalent relation to (3.12) is

\[
u \in \frac{1}{2}(\rho_d(v) + \rho_i(v)) - \frac{1}{2}(\rho_d(v) - \rho_i(v)) \cdot \text{sign}(\partial_t v).
\]

---

**Figure 3.1:** (left) Realistic \( p_c^{(i)} \) and \( p_c^{(d)} \) curves in the \( S-p \) plane calculated using the van Genuchten model (1.7); (right) corresponding \( \rho_i \) and \( \rho_d \) curves in the \( v-u \) plane.

The \( b(p) \) used here is such that \( b'(p) = -\frac{1}{2} \max \left( \frac{1}{(p_c^{(i)}(p)c^{(i)}(S)}(p)) \right. \), The black dotted lines correspond to the scanning curves with respect to this choice. In particular, curves for \( S + b(p) = 0.5, 0.7, 0.9 \) are shown. The values \( U_m, U_M, V_m \) and \( V_M \) are marked in the (right) figure.
Since \( \text{sign}(\cdot) \) is not single-valued, we regularise expression (3.19). In Chapter 4, \( \text{sign}(\cdot) \) is regularised by a smooth and increasing function. Here, we use the expression

\[
u \in \frac{1}{2}(\rho_d(v) + \rho_\dagger(v)) - \frac{1}{2}(\rho_d(v) - \rho_\dagger(v)) \cdot \text{sign}(\partial_t v) - \tau \partial_t v, \tag{3.20}\]

where \( \tau > 0 \) is the regularisation parameter. This approach has been used in [54, 149, 209] and in Chapter 5. The right most term in (3.20) also has physical significance, since, it gives rise to the dynamic capillarity phenomenon in porous media [53, 112, 222], see Section 1.2.2. Moreover, expression (3.20) is thermodynamically consistent as it leads to entropy generation as follows from [24], see specifically equation (35). Since the function \( \frac{1}{2}(\rho_d(v) - \rho_\dagger(v))\text{sign}(\cdot) + \tau \xi \) is increasing with respect to \( \xi \), from [24, 25, 54], expression (3.20) can be inverted, which gives

\[
\partial_t v = \Phi_T(u, v) := \begin{cases} \rho_d(v) - u & \text{when } u > \rho_d(v), \\ 0 & \text{when } u \in [\rho_\dagger(v), \rho_d(v)], \\ \rho_\dagger(v) - u & \text{when } u < \rho_\dagger(v). \end{cases} \tag{3.21}\]

Setting in (EPH)

\[
\mathcal{D}(u, v) = \frac{k(v - u)}{b'(b^{-1}(u))}, \quad \tilde{F}(u, v) = \tilde{g}k(v - u), \quad \Psi_1(u, v) = \Psi_2(u, v) = \Phi_T(u, v), \tag{3.22}\]

we recover (\( \mathcal{D}s \)). It is straightforward to verify from (P5) that \( \Psi_1 \) and \( \Psi_2 \), defined as in (3.22), satisfy assumption (A3). Similarly \( u_0 \) and \( v_0 \) are consistent with (A4) when defined from a \( (S_0, p_0) \) pair satisfying (P4). Furthermore, \( U_m < u_0 < U_M \) and \( V_m < v_0 < V_M \). However, to show that \( \mathcal{D}(u, v) \) defined in (3.22), satisfies assumption (A1) we need to show that \( S = v - u \) is bounded away from zero. This is done in Section 3.4.

Based on our discussion so far, we define the weak solution of the (EPH) model for \( \tau > 0 \) as

**Definition 3.2.** The pair \( (S, p) \) with \( p \in \mathcal{W}', S - S_0 \in \mathcal{W} \) and \( S \in [0, 1] \) a.e. in \( Q \) is a weak solution of (EPH) for \( \tau > 0 \) if for any \( \phi \in L^1(0, T; H^1_0(\Omega)) \) and \( \xi \in L^2(\Omega) \) the following is satisfied

\[
(\mathcal{P}_{\text{EPH}}) \begin{cases} \int_0^T \langle \partial_t S, \phi \rangle = \int_0^T (k(S)(\nabla p - \tilde{g})) \cdot \nabla \phi; \quad \tag{3.23a} \\
\langle \partial_t (S + b(p)), \xi \rangle = (\Phi_T(b(p), S + b(p)), \xi); \quad \tag{3.23b}\end{cases}
\]

with \( p(0) = p_0, S(0) = S_0 \) and \( \Phi_T \) defined in (3.21).

Observe that, according to Definition 3.2, \( S \) has a trace on \( \partial \Omega \) that does not change with time, i.e., it is fixed by \( S_0 \).
3.3 Existence of solutions

The main existence result of this section is as follows:

**Theorem 3.1.** Assume (A1)-(A4). Then (Pw) has a weak solution (u, v). Moreover, \( u \in W \cap L^\infty(0, T; L^2(\Omega)) \), \( v \in L^\infty(0, T; H^1(\Omega)) \) and \( \partial_t v \in L^\infty(0, T; L^2(\Omega)) \).

To prove this, we apply Rothe’s method [132]. Let the time \( T \) be divided into \( N \) time steps of width \( \Delta t \) \((T = N\Delta t)\) and for any \( n \in \{1, 2, ..., N\} \) let \( \partial_t w \) be approximated by \( (w_n - w_{n-1})/\Delta t \) for \( w \in (u, v) \). Here \( w_n \) stands for the value of the function \( w \) at time \( t_n = n\Delta t \). The time-discrete solution is defined as

**Definition 3.3.** For a given \( n \in \{1, 2, ..., N\} \) and \( (u_n, v_n) \in (L^2(\Omega))^2 \), the time discrete weak solution of (Pw) at \( t = t_n \) is a pair \((u_n, v_n) \in H^1_0(\Omega) \times L^2(\Omega) \) such that for all \( \phi \in H^1_0(\Omega) \) and \( \xi \in L^2(\Omega) \),

\[
\begin{align*}
(\mathcal{D}_r) \quad & (u_n - u_{n-1}, \phi) + \Delta t(\mathcal{F}(u_{n-1}, v_{n-1}) \nabla u_n, \nabla \phi) = \Delta t(\Psi_1(u_n, v_n), \phi) + (u_{n-1}, \phi), \\
& (v_n, \xi) = (v_{n-1}, \xi) + \Delta t(\Psi_2(u_n, v_n), \xi).
\end{align*}
\]

(3.24a, 3.24b)

For the rest of the section the shorthand \( \mathcal{D}_n := \mathcal{D}(u_n, v_n) \), \( \tilde{F}_n := \tilde{F}(u_n, v_n) \) and \( \Psi_{jn} := \Psi_j(u_n, v_n) \) will be used extensively for \( n \in \{1, 2, ..., N\} \) and \( j \in \{1, 2\} \). We show first that the pair \((u_n, v_n)\) exists.

**Lemma 3.1.** For small enough \( \Delta t \), a pair \((u_n, v_n)\) solving (P_r) exists.

**Proof.** We define the operator \( \mathcal{D} : (L^2(\Omega))^2 \to H^1_0(\Omega) \times L^2(\Omega) \) as follows: \( \mathcal{D}(\tilde{u}, \tilde{v}) = (u^*, v^*) \) solving

\[
\begin{align*}
(u^*, \phi) + \Delta t(\mathcal{D}_{n-1} \nabla u^* - \tilde{F}_{n-1}, \nabla \phi) = & \Delta t(\Psi_1(\tilde{u}, \tilde{v}), \phi) + (u_{n-1}, \phi), \\
(v^*, \xi) = & (v_{n-1}, \xi) + \Delta t(\Psi_2(\tilde{u}, \tilde{v}), \xi).
\end{align*}
\]

(3.25a, 3.25b)

for any \( \phi \in H^1_0(\Omega) \) and \( \xi \in L^2(\Omega) \). It follows from Lax-Milgram theorem [269, Chapter 2] that \( \mathcal{D} \) is well-defined. Next, we apply the Banach fixed point theorem. For two pairs \((\tilde{u}_1, \tilde{v}_1)\) and \((\tilde{u}_2, \tilde{v}_2)\) in \( H^1_0(\Omega) \times L^2(\Omega) \) let the outputs of \( \mathcal{D} \) be \((u^*_1, v^*_1)\) and \((u^*_2, v^*_2)\). Defining \( e^*_u = u^*_1 - u^*_2 \), putting \( \phi = e^*_u \) in (3.25a) and using Young’s inequality we get

\[
\|e^*_u\|^2 + \Delta t \mathcal{D}_m \|\nabla e^*_u\|^2 \leq (e^*_u, e^*_u) + \Delta t(\mathcal{D}_{n-1} \nabla e^*_u, \nabla e^*_u) = \Delta t(\Psi_1(\tilde{u}_1, \tilde{v}_1) - \Psi_1(\tilde{u}_2, \tilde{v}_2), e^*_u) \\
\leq \frac{\Delta t^2}{2} \|\Psi_1(\tilde{u}_1, \tilde{v}_1) - \Psi_1(\tilde{u}_2, \tilde{v}_2)\|^2 + \frac{1}{2} \|e^*_u\|^2 \leq C \Delta t^2 \|\tilde{u}_1 - \tilde{u}_2\|^2 + \|\tilde{v}_1 - \tilde{v}_2\|^2 + \frac{1}{2} \|e^*_u\|^2,
\]

for some \( C > 0 \), the last inequality following from (A3). This implies,

\[
\frac{1}{2} \|e^*_u\|^2 + \Delta t \mathcal{D}_m \|\nabla e^*_u\|^2 < C \Delta t^2 \|\tilde{u}_1 - \tilde{u}_2\|^2 + \|\tilde{v}_1 - \tilde{v}_2\|^2.
\]

(3.26)
Similarly defining \( e^*_{\nu} = v^*_1 - v^*_2 \) we get from (3.25b)
\[
\| e^*_{\nu} \|^2 < C \Delta t^2 \{ \| \bar{u}_1 - \bar{u}_2 \|^2 + \| \bar{v}_1 - \bar{v}_2 \|^2 \}. \tag{3.27}
\]
This clearly shows that \( \mathcal{B} \) defines a contraction in \( H^1_0(\Omega) \times L^2(\Omega) \) for \( \Delta t \) small enough. More precisely, observing that \( \| u \|_t = \sqrt{\| u \|^2 + 2 \Delta t D_{\text{eff}} \| \nabla u \|^2} \) defines a norm on \( H^1_0(\Omega) \) which is equivalent to the standard one, \( \mathcal{B} \) is contractive with respect to the norm \( \sqrt{\| u \|_t^2 + \| v \|^2} \) for small \( \Delta t \), following from (3.26) and (3.27). Hence, a fixed point \((u_n, v_n)\) of \( \mathcal{B} \) exists in \( H^1_0(\Omega) \times L^2(\Omega) \) so that \( \mathcal{B}(u_n, v_n) = (u_n, v_n) \). This proves the lemma. We remark that the condition on \( \Delta t \) is moderate, i.e. \((u_n, v_n)\) exists if \( 0 < \Delta t \leq C \) where \( C > 0 \) does not depend on \( n \) or \((u_{n-1}, v_{n-1})\).

From now on we assume that \( \Delta t \) is small enough which guarantees the existence of solution pairs to the time discrete problems \((\mathcal{B} r)\). Our goal will be to construct the solution \((u, v)\) from the time-discrete solutions. For this purpose, we introduce the following interpolation functions: for \( w \in \mathcal{W} \) and \( t \in (0, T) \) the piece-wise constant interpolations \( \bar{w}_{\Delta t}, \tilde{w}_{\Delta t} \) and the linear interpolation \( \hat{w}_{\Delta t} \) are defined in \( \mathcal{Q} \) so that for \( t \in (t_{n-1}, t_n) \) (here \( t_n = n \Delta t \)),
\[
\bar{w}_{\Delta t}(t) = w_n, \quad \tilde{w}_{\Delta t}(t) = w_{n-1}, \quad \hat{w}_{\Delta t}(t) = w_{n-1} + \frac{t - t_{n-1}}{\Delta t} (w_n - w_{n-1}). \tag{3.28}
\]

As a first step we show that \( \bar{u}_{\Delta t} \) and \( \tilde{v}_{\Delta t} - v_0 \) are bounded uniformly in \( \mathcal{W} \) and then we would use embedding theorems to construct the weak solution.

**Lemma 3.2.** Given \( \{(u_n, v_n) : n \in \{1, \ldots, N\}\} \) solving \((\mathcal{B} r)\), \( \bar{u}_{\Delta t}, \tilde{u}_{\Delta t}, \hat{u}_{\Delta t} \in L^2(0, T; H^1_0(\Omega)) \) and \( \partial_t \bar{u}_{\Delta t} \in L^2(0, T; H^{-1}(\Omega)) \) and the corresponding norms are bounded uniformly with respect to \( \Delta t \). Similarly, the bounds of \( \bar{v}_{\Delta t}, \tilde{v}_{\Delta t}, \hat{v}_{\Delta t} \in L^\infty(0, T; H^1(\Omega)) \) and \( \partial_t \bar{v}_{\Delta t} \in L^\infty(0, T; L^2(\Omega)) \) are uniform.

**Proof.** (Part 1) The membership of the functions \( \bar{u}_{\Delta t}, \tilde{u}_{\Delta t}, \hat{u}_{\Delta t} \) in \( L^2(0, T; H^1_0(\Omega)) \) is straightforward. We proceed by showing their uniform boundedness. Putting \( \phi = u_n \) in (3.24a) and multiplying (3.24b) by \( v_n \) and integrating over \( \Omega \) we get,
\[
\begin{align*}
(\| u_n \|^2 - & \| u_{n-1} \|^2 + \| u_n - u_{n-1} \|^2 + \Delta t D_{\text{eff}} \| \nabla u_n \|^2) \leq C \Delta t [1 + \| u_n \|^2 + \| v_n \|^2], \\
(\| v_n \|^2 - & \| v_{n-1} \|^2 + \| v_n - v_{n-1} \|^2) \leq C \Delta t [1 + \| u_n \|^2 + \| v_n \|^2],
\end{align*}
\]
for some \( C > 0 \). Here, the identity \( (a - b, b) = \frac{1}{2} (\| a \|^2 - \| b \|^2) + \frac{1}{2} \| a - b \|^2 \) has been invoked and \( (\bar{F}_{n-1}, \nabla u_n) \leq \frac{1}{2} \| \bar{F}_{n-1} \|^2 + \frac{2\eta}{2} \| \nabla u_n \|^2 \) is used. Combining both in-
3.3 Existence of solutions

equalities and summing the results up from \( n = 1 \) to \( P \leq N \) yields

\[
\|u_P\|^2 + \|v_P\|^2 + \sum_{k=1}^{P} (\|u_k - u_{k-1}\|^2 + \|v_k - v_{k-1}\|^2) + \mathcal{D}_m \sum_{k=1}^{P} \|\nabla u_k\|^2 \Delta t \\
\leq (\|u_0\|^2 + \|v_0\|^2) + 2CP\Delta t + 2C' \Delta t \sum_{k=0}^{P-1} (\|u_k\|^2 + \|v_k\|^2).
\]

With \( C_0 = \|u_0\|^2 + \|v_0\|^2 \) the discrete Gronwall’s lemma is applied to get:

\[
\|u_P\|^2 + \|v_P\|^2 \leq (C_0 + 2CP\Delta t) \exp(2CP\Delta t) \leq (C_0 + 2CT) \exp(2C'T).
\]  (3.29)

Further, it gives two other important bounds both of which are used later, i.e.

\[
\sum_{k=1}^{N} \|\nabla u_k\|^2 \Delta t < C_1(T), \quad \sum_{k=1}^{N} (\|u_k - u_{k-1}\|^2 + \|v_k - v_{k-1}\|^2) < C_2(T),
\]  (3.30)

with \( C_1(T) > 0 \) being independent of \( N \) for \( i \in \{1,2\} \). This directly gives the bounds \( L^2(0,T;H^1_0(\Omega)) \) since, for example,

\[
\int_0^T \|\partial_{\Delta t} u\|^2_{H^1_0(\Omega)} dt = \sum_{k=1}^{N} \|\nabla u_k\|^2 \Delta t \leq C_1(T),
\]

with the rest of the bounds following accordingly.

(Part 2) We need to show that \( \partial_{\Delta t}, \bar{v}_{\Delta t}, \tilde{v}_{\Delta t} \) \( \in L^\infty(0,T; H^1(\Omega)) \). So far we have from (3.29) that \( \bar{v}_{\Delta t}, \tilde{v}_{\Delta t} \in L^\infty(0,T; L^2(\Omega)) \). Since (3.3) does not explicitly involve any spatial derivatives of \( v_n \), in order to prove its spatial regularity, we use directional derivatives: \( D^h w = (w(\bar{x} + \delta h) - w(\bar{x}))/h \) for \( \bar{x} \in \Omega \) and an arbitrary unit vector \( \delta \in \mathbb{R}^d \). Choose an open subset \( V \subset \Omega \) such that \( \text{dist}(V, \partial \Omega) > 2h > 0 \). Then from (3.24b) after multiplying with \( D^h v_n \) and integrating over \( V \) we get:

\[
\int_V (D^h v_n - D^h v_{n-1}) D^h v_n = \Delta t \int_V D^h \Psi_{2,n} D^h v_n \leq \Delta t C \int_V |D^h u_n|^2 + \int_V |D^h v_n|^2,
\]

where \( C > 0 \) does not depend on \( n \) or \( \Delta t \). After summing from \( n = 1 \) to \( P \), where \( P \leq N \) is chosen arbitrarily, and using \( \sum_{k=1}^{P} \int_V |D^h u_k|^2 \Delta t \leq \tilde{C}_3 \sum_{k=1}^{P} \|\nabla u_k\|^2 \Delta t \leq \tilde{C}_3 C_1(T) =: C_3 \) for some \( \tilde{C}_3 > 0 \) (see Theorem 3, Chapter 5.8 of [89]) we have

\[
or \int_V |D^h v_P|^2 - \int_V |D^h v_0|^2 + \sum_{k=1}^{P} \int_V |D^h v_k - D^h v_{k-1}|^2 \leq C_3 + 2C \sum_{k=1}^{P} \int_V |D^h u_k|^2 \Delta t.
\]

With the application of discrete Gronwall’s lemma one obtains that \( \int_V |D^h v_n|^2 \) is bounded independent of \( V \) and \( h \). The smoothness of the boundary \( \partial \Omega \) further implies that we can extend \( V \) to \( \Omega \) and hence by applying the theorem mentioned above it is proved that \( \|\nabla v_n\| \) is bounded. Consequently, \( \partial_{\Delta t}, \bar{v}_{\Delta t}, \tilde{v}_{\Delta t} \in L^\infty(0,T; H^1(\Omega)) \).
(Part 3) Finally we prove the regularity of time derivatives of \( \tilde{u}_{\Delta t} \) and \( \tilde{v}_{\Delta t} \). Observe that, as \( \tilde{u}_{\Delta t}, \tilde{v}_{\Delta t} \in L^\infty(0,T;L^2(\Omega)) \), \( \Psi_2(\tilde{u}_{\Delta t}, \tilde{v}_{\Delta t}) \in L^\infty(0,T;L^2(\Omega)) \) which gives from (3.24b),

\[
\|\partial_t \tilde{v}_{\Delta t}\| = \frac{1}{\Delta t} \|v_n - v_{n-1}\| = \|\Psi_2\| = \|\Psi_2(\tilde{u}_{\Delta t}, \tilde{v}_{\Delta t})\| \leq C_4,
\]

for some \( C_4 > 0 \) and \( t \in (t_{n-1}, t_n) \). As for \( \partial_t \tilde{u}_{\Delta t} \), from (3.24a)

\[
\|\partial_t \tilde{u}_{\Delta t}\|_{H^{-1}(\Omega)} = \sup_{\|\phi\|_{H^1(\Omega)} = 1} \langle \partial_t \tilde{u}_{\Delta t}, \phi \rangle = \sup_{\|\phi\|_{H^1(\Omega)} = 1} \left( \frac{u_n - u_{n-1}}{\Delta t}, \phi \right) 
\]

\[
\leq D_M \|u_n\| + \|\tilde{F}_{n-1}\| + \|\phi\|_{H^1(\Omega)} = D_M \|u_{\Delta t}\| + \|F_{\Delta t}\| + \frac{1}{C_\Omega} \|\Psi_1(\tilde{u}_{\Delta t}, \tilde{v}_{\Delta t})\|.
\]

Since \( \tilde{u}_{\Delta t} \in L^2(0,T;H^1_0(\Omega)) \) and \( \Psi_1(\tilde{u}_{\Delta t}, \tilde{v}_{\Delta t}) \in L^2(Q) \), we have that \( \partial_t \tilde{u}_{\Delta t} \in L^2(0,T;H^{-1}(\Omega)) \).

**Proof of Theorem 3.1.** Lemma 3.2 shows that \( \tilde{u}_{\Delta t} \) is bounded in \( \mathcal{W} \) irrespective of \( \Delta t \). Hence, there exists a sequence of time-steps \( \{\Delta t_p\} \in \mathbb{N} \) with \( \lim \Delta t_p = 0 \) such that

\[
\tilde{u}_{\Delta t_p} \rightharpoonup u \text{ and } (\tilde{v}_{\Delta t_p} - v_0) \rightharpoonup (v - v_0) \text{ weakly in } \mathcal{W}.
\]  

Due to the compact embedding of \( \mathcal{W} \) in \( L^2(Q) \) this further implies that,

\[
\tilde{w}_{\Delta t_p} \rightharpoonup w, \text{ strongly in } L^2(Q) \text{ for } w \in \{u, v\}.
\]

From (3.30) the following inequalities are obtained,

\[
\int_0^T \|\tilde{u}_{\Delta t} - \tilde{u}_{\Delta t_p}\|^2 \, dt = \sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \left\|\frac{t_k - t}{\Delta t}(u_k - u_{k-1})\right\|^2 \, dt = \frac{\Delta t}{3} \sum_{k=1}^{N} \|u_k - u_{k-1}\|^2 \leq \frac{C_2}{3} \Delta t,
\]

\[
\int_0^T \|\tilde{u}_{\Delta t} - \tilde{u}_{\Delta t_p}\|^2 \, dt = \sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \|(u_k - u_{k-1})\|^2 \, dt = \Delta t \sum_{k=1}^{N} \|u_k - u_{k-1}\|^2 \leq C_2 \Delta t.
\]

This shows that \( \tilde{u}_{\Delta t_p} \rightarrow u \text{ as } \Delta t_p \rightarrow 0 \text{ since } \|u - \tilde{u}_{\Delta t_p}\| \leq \|u - \tilde{u}_{\Delta t_p} + \tilde{u}_{\Delta t_p} - \tilde{u}_{\Delta t_p}\|. \) By repeating this argument, one shows that the same holds for \( \tilde{u}_{\Delta t_p} \). Hence,

\[
\tilde{u}_{\Delta t_p}, \tilde{u}_{\Delta t_p}, \tilde{u}_{\Delta t_p} \rightarrow u, \text{ strongly in } L^2(Q).
\]

In an identical way, for \( \tilde{v}_{\Delta t} \) there exists a subsequence \( \Delta t_q \rightarrow 0 \) such that

\[
\tilde{v}_{\Delta t_q}, \tilde{v}_{\Delta t_q}, \tilde{v}_{\Delta t_q} \rightarrow v, \text{ strongly in } L^2(Q).
\]

Finally, due to the bounds of \( \partial_t \tilde{u}_{\Delta t} \) and \( \partial_t \tilde{v}_{\Delta t} \) as given in Lemma 3.2, for a \( \Delta t_m \rightarrow 0 \) we also have,

\[
\partial_t \tilde{w}_{\Delta t_m} \rightarrow \partial_t w, \text{ weakly in } L^2(0,T;H^{-1}(\Omega)) \text{ for } w \in \{u, v\}.
\]


We claim that \((u, v)\) solves \((\mathcal{P}w)\). Let \(\Delta t_r \to 0\) be a sequence that satisfies the limits (3.31), (3.32), (3.33) and (3.34). From (3.24) we have for \(\phi \in L^2(0, T; H^1_0(\Omega))\) and \(\xi \in L^2(Q),\)

\[
T \int_0^T (\partial_t \tilde{u}_{\Delta t_r}, \phi) + \int_0^T \langle \mathcal{D}(\tilde{u}_{\Delta t_r}, \tilde{v}_{\Delta t_r}) \nabla \tilde{u}_{\Delta t_r}, \nabla \phi \rangle = \int_0^T \langle \Psi_1(\tilde{u}_{\Delta t_r}, \tilde{v}_{\Delta t_r}), \phi \rangle + \int_0^T \langle \tilde{F}(\tilde{u}_{\Delta t_r}, \tilde{v}_{\Delta t_r}), \nabla \phi \rangle,
\]

\[
T \int_0^T (\partial_t \tilde{v}_{\Delta t_r}, \xi) = \int_0^T \langle \Psi_2(\tilde{u}_{\Delta t_r}, \tilde{v}_{\Delta t_r}), \xi \rangle.
\]

Since \(\partial_t \tilde{v}_{\Delta t_r} - \partial_t v \in L^2(\Omega)\), the second equation directly gives (3.10b) in the limit. From (3.34), \(\int_0^T (\partial_t \tilde{u}_{\Delta t_r}, \phi) - \int_0^T (\partial_t u, \phi)\) and (3.32),(3.33) gives \(\int_0^T \langle \Psi_1(\tilde{u}_{\Delta t_r}, \tilde{v}_{\Delta t_r}), \phi \rangle - \int_0^T \langle \tilde{F}(\tilde{u}_{\Delta t_r}, \tilde{v}_{\Delta t_r}), \nabla \phi \rangle\). Convergence of the term with \(\nabla \tilde{u}_{\Delta t_r}\) remains to be shown. For this we first observe that \(\mathcal{D}(\tilde{u}_{\Delta t_r}, \tilde{v}_{\Delta t_r}) \nabla \tilde{u}_{\Delta t_r}\) is bounded in \(L^2(Q)\) uniformly with respect to \(\Delta t\). This means that a \(\tilde{\xi} \in L^2(0, T; L^2(\Omega)^d)\) exists such that \(\mathcal{D}(\tilde{u}_{\Delta t_r}, \tilde{v}_{\Delta t_r}) \nabla \tilde{u}_{\Delta t_r} - \tilde{\xi}\) weakly. To identify \(\tilde{\xi} = \mathcal{D}(u, v) \nabla u\), we restrict the test function to \(\phi \in C^\infty_0(Q)\). Using the strong convergence of \(\mathcal{D}(\tilde{u}_{\Delta t_r}, \tilde{v}_{\Delta t_r})\) and the weak convergence of \(\nabla \tilde{u}_{\Delta t_r}\) one gets with \(C_\phi = \|\phi\|_{C^1}\) that

\[
\left| \int_0^T \langle \mathcal{D}(\tilde{u}_{\Delta t_r}, \tilde{v}_{\Delta t_r}) \nabla \tilde{u}_{\Delta t_r} - \mathcal{D}(u, v) \nabla u, \nabla \phi \rangle \right| \leq \left| \int_0^T \langle (\mathcal{D}(\tilde{u}_{\Delta t_r}, \tilde{v}_{\Delta t_r}) - \mathcal{D}(u, v)) \nabla \tilde{u}_{\Delta t_r}, \nabla \phi \rangle \right| + \left| \int_0^T \langle \mathcal{D}(u, v) \nabla (\tilde{u}_{\Delta t_r} - u), \nabla \phi \rangle \right| \leq C_\phi \left| \int_0^T \langle (\mathcal{D}(\tilde{u}_{\Delta t_r}, \tilde{v}_{\Delta t_r}) - \mathcal{D}(u, v)) \nabla \tilde{u}_{\Delta t_r}, \nabla \phi \rangle \right| + \left| \int_0^T \langle \mathcal{D}(u, v) \nabla (\tilde{u}_{\Delta t_r} - u), \mathcal{D}(u, v) \nabla \phi \rangle \right| \to 0,
\]

as \(\Delta t_r \to 0\). Since the weak limit is unique, we have \(\mathcal{D}(\tilde{u}_{\Delta t_r}, \tilde{v}_{\Delta t_r}) \nabla \tilde{u}_{\Delta t_r} - \tilde{\xi} = \mathcal{D}(u, v) \nabla u\). This shows that \((u, v)\) is a weak solution of \((\mathcal{P}w)\).

\[\Box\]

### 3.4 Boundedness and well-posedness of (EPH)

#### 3.4.1 Bounds on \(u\) and \(v\)

Next, we investigate whether a solution of \((\mathcal{P}w)\) satisfies the maximum principle or not. This is an interesting question primarily due to two reasons. Firstly, the maximum principle is used to prove the existence of solutions of \((\mathcal{P}_{\mathcal{EPH}})\) in the case when \(k(0) = 0\). This is discussed in details later. Secondly, for pseudo-parabolic equations arising from the dynamic capillarity effect, it is shown in [67] that the maximum principle does not hold. Moreover, in Chapter 5 we prove that travelling
wave solutions of the EPH model, when combined with dynamic capillarity, do not abide by the maximum principle. Having similar structure to the systems mentioned above, one might wonder if a maximum principle holds for \( (\mathcal{P}w) \). As it turns out, a maximum principle does hold in this case if the convective term behaves in a certain way. This is preferred as a property of the EPH since hysteresis alone is known not to cause deviation from the maximum principle (see the pure hysteresis limits shown in Chapters 4 and 7 for unsaturated and two-phase flow cases). However, it is to be noted that the maximum principle does not necessarily generalise to other convective terms, as is expected in the case of pseudo-parabolic equations.

**Proposition 3.1.** Assume that \( \tilde{F}(u, v) = \tilde{F}(u) \). For the pair \((v_l, v_r)\), \( v_l = \inf \{v_0\} \) and \( v_r = \sup \{v_0\} \) let there exist a pair \((u_l, u_r)\) such that \( u_0 \in [u_l, u_r] \) a.e. and

\[
\Psi_1(u_l, v_l) \leq 0 \leq \Psi_2(u_r, v_l), \quad \Psi_2(u_l, v_r) \leq 0 \leq \Psi_1(u_l, v_r).
\]

Then the weak solution \((u, v)\) of \( (\mathcal{P}w) \) satisfies \( v_l \leq v \leq v_r \) and \( u_l \leq u \leq u_r \) a.e.

The rationals behind the assumptions used in the proposition are explained later in the context of \((\mathcal{E}P)\).

**Proof.** We only show the proof that \( u < u_r \) and \( v > v_l \) a.e. and omit the other half of the proof since the arguments are identical. For any \( t \in [0, T] \), with \( \chi_{[0, t]} \) denoting the characteristic function of \([0, t]\), taking \( \phi = \chi_{[0, t]} \{u-u_r\} \) in (3.10a) yields:

\[
\frac{1}{2}\|u(t) - u_r\|^2 + D_m \int_0^t \|\nabla (u - u_r)\|^2 \leq \int (\partial_t u, [u - u_r]_+) + \int (\partial_r u, [u - u_r]_+) + \int (\tilde{F}(u, v) \nabla u, \nabla [u - u_r]_+)
\]

\[
= \int (\Psi_1(u, v), [u - u_r]_+) + \int (\tilde{F}(u), \nabla [u - u_r]_+)
\]

\[
\leq \int (\Psi_1(u, v) - \Psi_1(u_l, v_l), [u - u_r]_+) + \int (\tilde{F}(u) - \tilde{F}(u_r), \nabla [u - u_r]_+)
\]

\[
= \int (\Psi_1(u, v) - \Psi_1(u_l, v_l), [u - u_r]_+) + \int (\tilde{F}(u) - \tilde{F}(u_r), \nabla [u - u_r]_+) + \int (\partial_t \tilde{F} u, [u - u_r]_+, \nabla [u - u_r]_+)
\]

\[
\leq \Psi_u \int \|u - u_r\|^2 + \frac{t}{2} (\Psi_1(u, v) - \Psi_1(u_l, v_l), [u - u_r]_+) + \frac{C_2}{2} \int \|\nabla [u - u_r]_+\|^2 + \frac{D_m}{2} \int \|\nabla [u - u_r]_+\|^2
\]

\[
\leq C_m \int \|u - u_r\|^2 + \|v_l - v_r\|^2 + \frac{D_m}{2} \int \|\nabla [u - u_r]_+\|^2.
\]

Here \( C, C_m > 0 \) are constants. The inequality \( 0 \leq -\frac{\Psi_1(u, v) - \Psi_1(u_l, v_l)}{v_l - v_l} \leq \Psi_v \) follows from \((A3)\). Moreover, \( \tilde{F} \in C^1(\mathbb{R}) \) and \( \int_{\Omega} \tilde{F}(u_r) \cdot \nabla [u - u_r]_+ = 0 \) are used.
Similarly, testing (3.10b) with \( \chi_{[0,t]}(v_l - v)_+ \) one gets
\[
\frac{1}{2} \| u(t) - v(t) \|_+^2 = \int_0^t (-\Psi_2(u,v),(v_l - v)_+) \leq \int_0^t (\Psi_2(u_r,v_l) - \Psi_2(u,v), (v_l - v)_+) \\
\leq C_m \int_0^t \| u - u_r \|_+^2 + \| v_l - v \|_+^2.
\]

Finally adding them yields
\[
\| u(t) - u_r \|_+^2 + \| v_l - v(t) \|_+^2 \leq 4C_m \int_0^t \| u - u_r \|_+^2 + \| v_l - v \|_+^2.
\] (3.35)

Since \( \| v_l - v_0 \|_+ = 0 \) and \( \| u_0 - u_r \|_+^2 = 0 \), we conclude from Gronwall’s lemma that \( \| u(t) - u_r \|_+ = 0 \) and \( \| v_l - v(t) \|_+ = 0 \) for all \( t > 0 \). This proves the proposition.  

### 3.4.2 Existence of solutions of (EPH)

The main result of this section is the existence of a solution to (EPH). This is obtained first for the case when \( k(0) > 0 \), and then for \( k(0) = 0 \) in the absence of convective terms.

**Theorem 3.2.** Assume (P1)-(P5).

(a) If \( k(0) > 0 \) then a solution of \((\mathcal{P}_{\text{EPH}})\) exists.

(b) If \( |g| = 0 \), then a solution \((S, p)\) of \((\mathcal{P}_{\text{EPH}})\) exists for \( k(0) = 0 \). Moreover, there exists saturations \( 0 < S_l < S_r < 1 \) such that \( S \in [S_l, S_r] \) and \( p \in [p_c^{(i)}(S_r), p_c^{(d)}(S_l)] \) a.e. for all \( t > 0 \).

We start by observing that (P3) implies \( b'(p) \to 0 \) for \( p \to \pm\infty \) which might cause the model to become degenerate. Therefore, we consider a non-degenerate system that approximates \((\mathcal{P}_{\text{EPH}})\) first. Let \( \rho_i(v) \) and \( \rho_d(v) \), defined in (3.18), be extended to \( \mathbb{R} \) such that \( \rho_i(v) = \rho_d(v) = U_M \) for \( v \leq V_m \), and \( \rho_i(v) = \rho_d(v) = U_m \) for \( v \geq V_M \). For \( \Phi_r \) defined in (3.21) this implies that \( \| \Phi_r(u,v) \|^2 < C[1 + \| u \|^2] \), since \( \rho_i, \rho_d \) are bounded.

For some \( \delta > 0 \) small enough, let \( p_c^{(i)}(1) < p_l^\delta < p_r^\delta < \infty \) be such that \( b'(p_l^\delta) = b'(p_r^\delta) = \delta \). Assume \( b'(p) > \delta \) for \( p_l^\delta < p < p_r^\delta \). Define \( b_\delta \in C^1(\mathbb{R}) \) to be a regularised version of \( b(p) \) such that (see Figure 3.2 (left))

(P8) \( b_\delta'(p) \geq \delta > 0 \) for all \( p \in \mathbb{R} \) with \( b_\delta'(p) = \delta \) for \( p \leq p_l^\delta \) or \( p \geq p_r^\delta \) and \( b_\delta(p) = b(p) \) for \( p_l^\delta < p < p_r^\delta \). Clearly, \( |b_\delta(p) - b(p)| < C\delta \| p \| + 1 \) for some constant \( C > 0 \).
Consequently, as \( p_0 \) satisfies (P4), \( p_0^\delta < p_0 < p_r^\delta \) for \( \delta \) small enough, making \( b_\delta(p_0) = b(p_0) \).

We look for solutions \((p_\delta, u_\delta, v_\delta)\), with \( p_\delta, u_\delta \in \mathcal{W} \) and \( v_\delta - S_0 - b(p_0) \in \mathcal{W} \), such that for any \( \phi \in L^2(0, T; H^1_0(\Omega)) \) and \( \xi \in L^2(Q) \) the following is satisfied

\[
\begin{align*}
(\mathcal{D}_\delta) \quad & \left\{ \begin{array}{ll}
T \int \langle \partial_t u_\delta, \phi \rangle + T \int k(v_\delta - u_\delta) |\nabla p_\delta - \hat{g}| \nabla \phi = T \int (\Phi_T(u_\delta, v_\delta), \phi); \\
\int (\partial_t v_\delta, \xi) = (\Phi_T(u_\delta, v_\delta), \xi), & u_\delta = b_\delta(p_\delta) \quad \text{in } \hat{Q};
\end{array} \right.
\end{align*}
\]

with \( u_\delta(0) = b(p_0), \ v_\delta(0) = S_0 + b(p_0) \). The existence of such a triplet \((p_\delta, u_\delta, v_\delta)\) follows from Theorem 3.1 by setting \( \tilde{F}, \Psi_1, \Psi_2 \) as in (3.22) and \( \mathcal{G}(u, v) = \frac{k(v - u)}{b_\delta^{-1}(u)} \).

For \( k(0) > 0 \) it follows directly that all the Assumptions (A1)-(A4) are satisfied. Hence \((p_\delta, u_\delta, v_\delta)\) exists. Next, we show uniform bounds of \( p_\delta, u_\delta \) and \( v_\delta \) with respect to \( \delta \).

**Lemma 3.3.** Under the assumptions of Theorem 3.2, \( p_\delta \) and \( u_\delta \) are uniformly bounded in \( L^2(0, T; H^1_0(\Omega)) \) and \( \mathcal{W} \) respectively, whereas, \( v_\delta - S_0 - b(p_0) \) is uniformly bounded in \( \mathcal{W} \).
3.4 Boundedness and well-posedness of (EPH)

Proof. To show this, we first test (3.36a) with \(u_\delta\), which gives

\[
\|u_\delta\|^2 - \|b(p_0)\|^2 + \frac{k(0)}{b_M} \int_0^T \|\nabla u_\delta\|^2 \\
\leq \frac{T}{0} \langle \partial_t u_\delta, u_\delta \rangle + \int_0^T (k(v_\delta - u_\delta) \nabla p_\delta, b_\delta'(p_\delta) \nabla p_\delta) \\
\leq \frac{T}{0} \int_0^T (k(v_\delta - u_\delta) \hat{g}, \nabla u_\delta) + \frac{T}{0} \|\Phi_T, u_\delta\| \leq \frac{b_M}{2k(0)} \|k(1)\| p_\delta^2 + \frac{k(0)}{2b_M} \int_0^T \|\nabla u_\delta\|^2 + C \frac{T}{0} [1 + \|u_\delta\|^2].
\]

Here \(b_M = \max_{\varphi \in \mathbb{R}} b_\delta'(\varphi) = \max_{\varphi \in \mathbb{R}} b_\delta'\). This clearly shows the boundedness of \(u_\delta\) in \(L^2(0, T; H^1_0(\Omega))\). Testing (3.36a) with \(p_\delta\) gives

\[
\left\| \frac{T}{0} \int_0^T \varphi \cdot b_\delta'(\varphi) d\varphi \right\| - \left\| \frac{T}{0} \int_0^T \varphi \cdot b_\delta'(\varphi) d\varphi \right\|_1 + k(0) \|\nabla p_\delta\|^2 \\
\leq \frac{T}{0} \int_0^T \varphi \cdot b_\delta'(p_\delta) \partial_t p_\delta + \frac{T}{0} \int_0^T (k(v_\delta - u_\delta) \nabla p_\delta, \nabla p_\delta) \\
= \frac{T}{0} \int_0^T (k(v_\delta - u_\delta) \hat{g}, \nabla p_\delta) + \frac{T}{0} \|\Phi_T, p_\delta\| \\
\leq \frac{\|k(1)\|^2}{k(0)} \|p_\delta\|^2 + \frac{k(0)}{4} \int_0^T \|\nabla p_\delta\|^2 + \frac{1}{k(0)C_\Omega^2} \frac{T}{0} \|\Phi_T\|^2 + \frac{k(0)}{4} \int_0^T \|\nabla p_\delta\|^2.
\]

We have used Poincaré inequality and the fact that \(\int_0^T \varphi \cdot b_\delta'(\varphi) d\varphi > 0\) for all \(\varphi \in \mathbb{R}\) here.

This gives that \(p_\delta\) is bounded uniformly in \(L^2(0, T; H^1_0(\Omega))\). Finally, by following the steps of Lemma 3.2 (Part 2), one concludes that \(v_\delta \in L^\infty(0, T; H^1(\Omega))\) while \(\partial_t v_\delta \in L^\infty(0, T; L^2(\Omega))\), the bounds being uniform in both cases. Moreover, following the steps of (Part 3) of Lemma 3.2

\[
\|\partial_t u_\delta\|_{H^{-1}} \leq k(1) \|\hat{g}\| + k(1) \|\nabla p_\delta\| + \frac{1}{C_\Omega} \|\Phi_T\|.
\]

which shows that \(\partial_t u_\delta \in L^2(0, T; H^{-1}(\Omega))\), concluding the proof. \[\square\]

Proof of Theorem 3.2(a). Define \(S_\delta := v_\delta - u_\delta\). From Lemma 3.3 it follows that there exists a sequence \((\delta_r)_{r \in \mathbb{N}}\) with \(\lim \delta_r = 0\) such that

\[
u_\delta_r \to u, v_\delta_r \to v, S_\delta_r \to S = v - u \text{ in } L^2(Q); \\
p_\delta_r \to p \text{ in } L^2(0, T; H^1_0(\Omega)), \partial_t u_\delta_r \to \partial_t u \text{ in } L^2(0, T; H^{-1}(\Omega)) \text{ and } \partial_t v_\delta_r \to \partial_t v \text{ in } L^2(Q).
\]

Further, following the proof of Theorem 3.1 we conclude that \(u, v, p\) and \(S\) satisfy

\[
\frac{T}{0} \int \partial_t S, \phi \right\|_1 \right\|_0 + k(0) \int_0^T \|\nabla u_\delta\|^2 + \frac{1}{k(0)C_\Omega^2} \frac{T}{0} \|\Phi_T\|^2 + \frac{k(0)}{4} \int_0^T \|\nabla p_\delta\|^2.
\]

Finally, by following the steps of (Part 3) of Lemma 3.2

\[
\|\partial_t u_\delta\|_{H^{-1}} \leq k(1) \|\hat{g}\| + k(1) \|\nabla p_\delta\| + \frac{1}{C_\Omega} \|\Phi_T\|.
\]
for \( \phi \in L^2(0,T;H^1_0(\Omega)) \) and \( \xi \in L^2(Q) \). For completeness, we still need to show that \( u = b(p) \) and \( S \in [0,1] \) a.e. To show the former, observe that \( b(p_{\delta_{r}}) \to u \) in \( L^2(Q) \), since,

\[
\|u - b(p_{\delta_{r}})\| \leq \|u - u_{\delta_{r}}\| + \|b_{\delta_{r}}(p_{\delta_{r}}) - b(p_{\delta_{r}})\| \to 0. \tag{3.37}
\]

The first term on the right vanishes as \( u_{\delta_{r}} \to u \) strongly. For the second we use (P6), giving \( \|b_{\delta_{r}}(p_{\delta_{r}}) - b(p_{\delta_{r}})\| \leq C\delta[1 + \|p_{\delta_{r}}\|] \) for some \( C > 0 \), which approaches 0 as \( \delta_{r} \to 0 \). Now, from (P3) one gets

\[
\|b(p) - b(p_{\delta_{r}})\|^2 \leq b_M(b(p) - b(p_{\delta_{r}}), p - p_{\delta_{r}}) = b_M(b(p) - u, p - p_{\delta_{r}}) + b_M(u - b(p_{\delta_{r}}), p - p_{\delta_{r}}).
\]

As \( \delta_{r} \to 0 \), both terms vanish due to (3.37) and the weak convergence of \( p_{\delta_{r}} \), thus proving \( u = b(p) \).

Finally, \( u = b(p) \in [U_m,U_M] \) a.e. implies that \( v \in [V_m,V_M] \) a.e. To see this, we test with \( \xi = \{V_m - v\}^+_\) to get

\[
\frac{1}{2} \partial_t \{\|V_m - v\|^2 = -\Phi_T(u,v),[V_m - V]_+ = \frac{1}{r}(u - \rho_i(v),[V_m - V]_+) \leq 0.
\]

Here, as \( v < V_m \) and \( u < U_M \), \( \Phi_T = \frac{\rho_i(v) - u}{r} \geq \frac{\rho_i(V_m) - U_M}{r} = 0 \), see also Figure 3.2 (right). This proves that \( \|V_m - v\| = 0 \) meaning \( v \geq V_m \) a.e. implying that \( S = v - u \geq V_m - U_M \). Similarly one shows that \( S \leq 1 \) which concludes our proof.

**Proof of Theorem 3.2(b).** Since \( k(0) = 0 \) we approximate \( k \) by \( k_{\mu}, \mu > 0 \), satisfying \( k_{\mu}(S) = k(S) > 0 \) for \( S > \mu \) and \( k_{\mu}(S) = k(S) \) otherwise. Then the solution \( (S_{\mu},p_{\mu}) \) of (\( \mathcal{P}_{EPH} \)) exists. Define \( u_{\mu} = b(p_{\mu}) \) and \( v_{\mu} = S_{\mu} + b(p_{\mu}) \), as before. The pair \( (v_{\mu}, v_{r}) \) is defined as in Proposition 3.1, i.e. \( v_{\mu} = \inf(S_{\mu} + b(p_{\mu})) \), \( v_r = \sup(S_{\mu} + b(p_{\mu})) \). Note that, \( v_{\mu}, v_r \in (V_m,V_M) \) due to (P4). Define \( u_\mu, u_r \in (U_m,U_M) \) as

\[
u = \rho_d(v_\mu), \quad u_\mu = \rho_i(v_r) \quad \text{giving} \quad \Phi_T(v_\mu, u_r) = \Phi_T(v_r, u_r) = 0,
\]

see Figure 3.2 (right) for clarification. Moreover, since \( p_{\mu}(S_{\mu}) \leq p_\mu \leq p_{\mu}(S_0) \), it implies that \( u_0 = b(p_0) \in [u_\mu, u_r] \). Thus, recalling \( |g| = 0 \), we have from Proposition 3.1 that \( u_\mu \leq u_r \) and \( v_\mu \leq v_r \) a.e. Consequently, \( S_{\mu} = v_\mu - u_\mu \geq v_{\mu} - u_r > 0 \). Defining

\[
S_{\mu} = v_{\mu} - u_r \quad \text{and} \quad S_r = v_r - u_\mu \quad \text{one has} \quad 0 < S_{\mu} \leq S_r < 1 \quad \text{a.e.} \tag{3.38}
\]

By taking \( \mu < S_{\mu} \), one actually gets \( k_{\mu}(S_{\mu}) \geq k(S_{\mu}) > 0 \) a.e. in \( Q \). Thus, passing \( \mu \to 0 \), we obtain the solution. \( \square \)
3.5 Behaviour when $\tau \to 0$

In this section we address one other important question: what is the behavior of $(S, p)$ for the limit $\tau \to 0$ which corresponds to the unregularised EPH model. Ideally $p_c(S) \leq p \leq p_c(S)$ should be satisfied if $\tau \to 0$ since the dynamic capillarity contribution is vanishing. One might also wonder whether a limiting solution exists in this case or not. In this section, we take a step forward towards understanding this limiting situation. To this aim, we make the parametric dependence of the solutions $(S, p)$ of $(\mathcal{P}_{\text{EPH}})$ explicit. More precisely, given $\tau > 0$, let $(S_\tau, p_\tau)$ denote the solution of $(\mathcal{P}_{\text{EPH}})$ and further define $u_\tau = b(p_\tau)$ and $v_\tau = S_\tau + b(p_\tau)$. Then, the following holds:

**Proposition 3.2.** Let the assumptions of Theorem 3.2 hold. Then for some $C > 0$,  

$$ 0 \leq T_0 \int \|u_\tau - \rho_d(v_\tau)\|_+^2 + T_0 \int \|\rho_i(v_\tau) - u_\tau\|_+^2 < C. $$

**Proof.** Observe that $\Phi_\tau$ can be rewritten simply as 

$$ \Phi_\tau(u, v) = -\frac{1}{\tau}|u - \rho_d(v)|_+ - \frac{1}{\tau}|u - \rho_i(v)|_- . $$

(3.39)

We consider here the case when $|\bar{g}| = 0$ and $k(0) = 0$, the case $|\bar{g}| \neq 0$ and $k(0) > 0$ being simpler. From Theorem 3.2(b), $p_\tau$ and $S_\tau$ are bounded independent of $\tau$. For $D(u, v)$, defined as in (3.22), this implies that there exists a $D_m = k(S)/\max_{p \in \mathbb{R}} (b'(p)) > 0$ such that (A1) is satisfied. The pair $(u_\tau, v_\tau)$ satisfies 

$$ 0 \leq T_0 \int \partial_\tau u_\tau, \phi + T_0 \int \langle \rho_d(u_\tau, v_\tau) \nabla u_\tau, \nabla \phi \rangle = T_0 \int \Phi_\tau(u_\tau, v_\tau, \phi) \text{ and } (\partial_\tau v_\tau, \xi) = (\Phi_\tau(u_\tau, v_\tau), \xi), $$

(3.40)

for all $\phi \in L^2(0, T; H^1_0(\Omega))$ and $\xi \in L^2(Q)$. We first test with $\phi = u_\tau$ which gives 

$$ \frac{1}{2} \int \|u_\tau\|^2 + D_m \int T_0 \|\nabla u_\tau\|^2 \leq \frac{1}{2} \int \|u_0\|^2 + T_0 \int \Phi_\tau(u_\tau) . $$

(3.41)

Testing the $v$-equation with $\rho_d(v_\tau)$ and integrating over Q one gets 

$$ 0 \leq T_0 \int \int_\Omega \left( v_\tau \int_0^T v_m \rho_d(q) dq \right) = T_0 \int \Phi_\tau, u_\tau - \rho_d(v_\tau). $$

(3.42)

Subtracting (3.42) from (3.41), one further obtains 

$$ \frac{1}{2} \int \|u_\tau\|^2 + D_m \int T_0 \|\nabla u_\tau\|^2 \leq \frac{1}{2} \int \|u_0\|^2 + \int_0^T \int \rho_d(q) dq \leq T_0 \int \Phi_\tau, u_\tau - \rho_d(v_\tau). $$
The term \( \int_0^T \rho_d(\varphi) \, d\varphi \) is bounded in \( L^\infty(\Omega) \) since both \( \tau \) and \( \rho_d(\varphi) \) are bounded. The last term can be estimated as

\[
\begin{align*}
\Phi_T, u_T - \rho_d(v_T) &= -\frac{1}{T} \| [u_T - \rho_d(v_T)]_+ \|^2 - \frac{1}{T} \| [u_T - \rho_i(v_T)]_-, u_T - \rho_d(v_T) \| \\
& \leq \frac{1}{T} \| [u_T - \rho_d(v_T)]_+ \|^2 - \frac{1}{T} \| [u_T - \rho_i(v_T)]_-, u_T - \rho_i(v_T) \| \\
& \leq \frac{1}{T} \| [u_T - \rho_d(v_T)]_+ \|^2 - \frac{1}{T} \| [u_T - \rho_i(v_T)]_- \|^2.
\end{align*}
\]

Combining everything, we get

\[
\begin{align*}
\frac{1}{2} \| u_T \|^2 + \mathcal{D}_m \int_0^T \| \nabla u_T \|^2 + \frac{1}{T} \int_0^T \| [u_T - \rho_d(v_T)]_+ \|^2 + \frac{1}{T} \int_0^T \| [u_T - \rho_i(v_T)]_- \|^2 \\
& \leq \frac{1}{2} \| u_0 \|^2 + \left\| \int_0^T \rho_d(\varphi) \, d\varphi \right\|_1.
\end{align*}
\]

(3.43)

This proves the assertion. The proof for the case \( k(0) > 0 \) is similar.

An immediate consequence of (3.43), specifically the boundedness of \( \int_0^T \| \nabla u_T \|^2 \), is the following corollary.

**Corollary 3.1.** Under assumptions of Proposition 3.2, there exists a \( u \in L^2(0, T; H_0^1(\Omega)) \cap L^\infty(0, T; L^2(\Omega)) \) such that \( u_T \rightharpoonup u \) weakly.

### 3.6 Conclusion and outlook

In this chapter, we analysed the extended play-type hysteresis (EPH) model, proposed in Chapter 2, for the unsaturated flow case. A modification of the model is proposed that conforms with the standard weak formulation used in analysis. A regularised version of the problem is considered and it is shown to be equivalent to a nonlinear parabolic equation coupled with an ordinary differential equation completed with suitable boundary and initial conditions. Using Rothe's method, the existence of weak solutions is first proven for the equivalent system and then for the EPH model. Solutions are shown to exist even when the capillary pressure functions degenerate and when relative permeability vanishes for zero saturation. To cover the latter case, the maximum principle needs to hold, which is proven to be true in the absence of an advective term. The consequences of passing the regularisation parameter to zero are explored which show that the solutions are bounded in the physically relevant regime in the limit.

The existence result proven in this chapter, also holds for the model proposed for hysteresis in relative permeabilities in Chapter 7 when combined with extended play-type hysteresis model. We leave the proof of uniqueness for a future study.
Chapter 4

Travelling wave solutions for the Richards equation incorporating non-equilibrium effects in the capillary pressure

4.1 Introduction

In this chapter we consider unsaturated flow through porous media which is encountered in many applications of societal and engineering relevance. Some examples are given in Chapter 1. A commonly used mathematical model for unsaturated flows is the Richards equation (Ri) (see (1.20)), which is obtained after inserting the Darcy law into the water mass balance equation. The two main unknowns in this equation are the water saturation and the water pressure, written in this chapter as $S$ and $p$ respectively. In standard porous media flow models, these two unknowns are related through the strictly decreasing capillary pressure function $p_c(\cdot)$, namely $p = -p_c(S)$, which is determined experimentally under equilibrium conditions. Commonly used expressions for $p_c$ are given in Section 1.1.

However, as seen from Chapters 1 and 2, under non-equilibrium conditions such relations no longer give accurate results. One of the main reasons of deviation from the predictions of the equilibrium model, is the presence of hysteresis in capillary pressure. For a detailed explanation of the nature of hysteresis we refer to Section 1.2.1 and a case study of different hysteresis models available is given in Section 1.2.2.

tion 2.1. In the current chapter, we consider the play-type hysteresis model for the pressure-saturation dependence. Subsequently, in Chapter 5, the extended play-type hysteresis model is considered under similar circumstances.

Secondly, the non-monotone profiles of saturation, experimentally observed for high enough injection rates in a long homogenous column [72], also plead for the inclusion of dynamic effects in the pressure-saturation relationship, as suggested in [112]. The interior saturation maxima, as observed from these experiments, are known as 'overshoots' and are shown in Figure 1.1(b).

In mathematical terms, models like those mentioned above are evolution equations of pseudo-parabolic type, or involve differential inclusions. Such models will be called below "non-equilibrium type models". In this chapter, we investigate how the solution profiles for unsaturated flow through a long, homogeneous porous column are affected by such non-equilibrium effects. The analysis is based on travelling waves (TW), allowing to reduce the model first to a nonlinear ordinary differential equation, and then to a dynamical system. This provides insight in the structure and behaviour of the solutions, and in particular how the non-equilibrium regime affects the profiles. The present analysis follows the ideas in [250], which studies the existence of TW solutions for reactive flow and transport models in porous media. In [68] TW solutions are analysed for nonlinear models that are similar to the Richards equation, but where higher order effects are included inspired by the ones describing dynamic capillarity. The nonlinear functions taken in [68] are of power-like type, in particular the flux function is convex. The existence of TW solutions is analysed, and in particular it is shown that oscillations behind the infiltration front may occur, depending on the magnitude of the dynamic effect. A similar analysis, but for two-phase flow models implying convex-concave flux functions is carried out in [236, 248, 254]. Also related are the diffusive-dispersive equations appearing as models for the phase transition dynamics, but in which the higher order terms are in terms of the spatial derivatives only [21, 66]. Though having a different motivation, the associated TW equation is similar to the one for the dynamic capillarity models, in particular since both involve a non-convex nonlinearity in the lower order terms. In this context, in [254] it is proved that the saturation profile may have overshoot in form of a plateau separated by two fronts (infiltration-drainage), similar to the ones obtained in [72]. The dependence of the saturation value at such plateaus on the magnitude of the dynamic effect is proved rigorously in [254], and non-standard entropy conditions are defined for the shock solutions of the limiting hyperbolic case when the capillary effects are neglected. This analysis is extended to the case of degenerate models in [236, 248]. Due to the degeneracy in the model, the saturation remains between the physically relevant values, but the TW solutions may have discontinuous derivatives. The possibility of encountering non-monotonic TW profiles for various extensions of the Richards equation, including dynamic capillarity models, is evidenced numerically in [3, 73, 83]. Finally, we mention [271] for a numerical study of the saturation and capillary pressure profiles for several of the
hysteresis concepts discussed above, combined with dynamic capillarity.

The present analysis consists of three parts. First the existence of TW solutions is
analysed for the models involving hysteresis. The TW profiles are obtained by regularising the multi-valued function involved in the hysteretic term. In particular, we analyse the orbits associated with the TW solution in the saturation-pressure plane. We prove that in the initial and the final stages these orbits follow scanning curves that become vertical when the regularisation parameter vanishes, and in between they follow the corresponding primary curves (imbibition/drainage).

Next, the case where dynamic effects are present in the pressure-saturation relationship is discussed. The existence of TW solutions is obtained and criteria ensuring their non-monotonicity are provided. These include also situations where full-saturation is achieved.

In the last part we discuss a numerical scheme for approximating the solution of the nonlinear, pseudo-parabolic partial differential equations modelling the processes described above. The scheme is implicit, so at each time step one has to solve a nonlinear problem. In this context we propose an iterative method which is unconditionally convergent. Finally, numerical results validating the theoretical findings are provided. As will be seen below, the numerical solutions to the original model are reproducing nicely analytically predicted structures and properties of the TW solutions.

4.2 Mathematical formulation

4.2.1 Basic equations

We consider the unsaturated water flow in a one-dimensional, homogeneous porous medium. Let \( t \) and \( x \) denote the time and space variable respectively. Assuming that the medium is vertical so that gravity effects are playing a role, we use Richards equation (Ri) which reads

\[
\partial_t S = \partial_x \left[ k(S) \left( \partial_x p - 1 \right) \right],
\]

for this case. To obtain (4.1) from (Ri), \( L_{\text{ref}} \) in (1.19) is chosen such that \( N_c = 1 \). This is a valid choice since the domains considered here, are unbounded. Further, \( \kappa, \phi \equiv 1 \) due to homogeneity. The relative permeability \( k(\cdot) \) is a given, positive and increasing function that characterizes the medium and can be determined experimentally.

The model is completed by an equation describing the dependence between \( p \) and \( S \). For the standard model, as stated in Section 1.1, this dependence is algebraic,

\[-p = p_c(S),\]

where \( p_c(\cdot) \) is the capillary pressure function. Its specific form is determined experimentally. As mentioned, the results available in the literature assume a local equilibrium and disregard the history of the system.
Here we consider the non-equilibrium model proposed in [24], which combines dynamic effects in the $p$-$S$ relationship with a simple play-type hysteresis model, see Chapters 1 and 2. The closure relationship is given by

$$- p \in p^+_c(S) - p^-_c(S) \cdot \text{sign}(\partial_t S) - \tau f(S)\partial_t S,$$

where $\text{sign}(\cdot)$ is defined in (1.25) and $p^\pm_c$ are defined in (1.24). As introduced in Section 1.2, $\tau$ is the dynamic capillary coefficient and $f(\cdot)$ is the dynamic capillary function. Both work as damping terms and are non-negative. In [112], a thermodynamic justification of such models has been given. Also, homogenisation techniques are employed in [37] for justifying the dynamic terms. For experimental studies concerning the value of $\tau$ and the shape of the function $f$ we refer to [35].

Next we state the assumptions on the nonlinear functions involved in the model. The assumptions are justified from physical point of view. In what follows, the superscript $'$ denotes differentiation with respect to the argument of the function.

(A1) $k \in C^1([0,1])$, $k'(S) > 0$ for $0 < S \leq 1$, $k(0) = 0$, $k(1) = 1$ and $k$ is convex.

(A2) The damping parameter is positive, $\tau > 0$. The damping function $f \in C([0,1])$ and $f(S) > 0$ for $0 < S < 1$.

(A3) The capillary pressure functions $p^{(j)}_c$, $j \in \{i, d\}$ satisfy $p^{(j)}_c : (0,1) \rightarrow [0,\infty)$, $p^{(j)}_c \in C^1((0,1))$, $p^{(j)}_c(1) = 0$, $p^{(j)}_c(S) < 0$ and $p^{(i)}_c(S) < p^{(d)}_c(S)$ for $S \in (0,1)$.

An immediate consequence is that the functions $p^\pm_c(\cdot)$ defined in (1.24) are in $C^1((0,1))$, satisfying $p^\pm_c(1) = 0$ and $p^\pm_c(S) > 0$ for all $S \in (0,1)$. Figure 1.3 displays an example of primary drainage and imbibition curves.

To analyse the effect of hysteresis, which is modelled by means of a multivalued function, we consider a regularisation approach. With $\varepsilon > 0$ being a small regularisation parameter, one can approximate the sign function by another function $H_\varepsilon : \mathbb{R} \rightarrow \mathbb{R}$ satisfying the following

(A4) For each $\varepsilon > 0$, $H_\varepsilon$ is smooth and satisfies

$$H_\varepsilon(-s) = -H_\varepsilon(s)$$

and

$$\lim_{s \rightarrow -\infty} H_\varepsilon(s) = \pm 1, \quad \lim_{\varepsilon \rightarrow 0} H_\varepsilon(s) = \begin{cases} -1 & \text{if } s < 0, \\ 1 & \text{if } s > 0. \end{cases}$$

Further, $H_\varepsilon$ depends smoothly and monotonically on $\varepsilon$: if $\varepsilon_1 > \varepsilon_2 > 0$ then $|H_{\varepsilon_1}(s)| < |H_{\varepsilon_2}(s)|$ for all $s \neq 0$. 


When sign is replaced by $H_\varepsilon$ in (4.2), the regularised model for the pressure-saturation relationship becomes

$$-p = p_c^+(S) - p_c^-(S) \cdot H_\varepsilon(\partial_t S) - \tau f(S) \partial_t S. \quad (4.4)$$

Such regularisation has been used in [209, 221] for proving the existence of weak solutions to such models, and for developing appropriate numerical schemes.

One may wonder if the regularisation (4.4) has a physical interpretation. In the play-type limit as $\varepsilon \to 0$, a switch from drainage to imbibition is through a vertical scanning curve, whereas, when considering the regularised model (4.4), scanning curves have a steep but finite slope as observed in experiments [170].

Another motivation for considering regularised models can be found in [261], where the play-type hysteresis is viewed as a 'friction-controlled backslash' process. This means that dissipative forces, which are mostly continuous in porous media, are responsible for it. At the pore scale, hysteresis occurs because of the difference in the advancing and receding contact angles of the wetting phase, which is a continuous phenomenon and hence jump phenomena should not be expected. These evidences in favour of non-vertical scanning curves have motivated us to propose the extended play-type model in Chapter 2.

Based on the above, $\varepsilon$ can be seen as a physical parameter, or at least can be used to fit more realistic $p_c$-$S$ scanning curves. Having this in mind, in the subsequent discussions we will analyse first the case $\varepsilon > 0$ and then the limiting case of $\varepsilon \to 0$.

Before doing so we mention that (4.1), combined with the constitutive relationship (4.2) or its regularised counterpart (4.4), becomes a nonlinear, pseudo-parabolic equation. In general, one cannot expect that solutions exist in a classical sense. We refer to [30,32,48,53–55,141,163,166,221,222] for results concerning the existence and uniqueness of weak solutions for hysteresis models, dynamic capillarity models, or for models including both effects. In particular we refer to [54,221,222] where, as suggested in [24,25], (4.4) is used to express $\partial_t S$ as a function of $S$ and $p$. We rely on the same idea for the TW analysis below.

### 4.2.2 Travelling wave formulation

To simplify the analysis and to understand the profile of the solutions to the regularised mathematical model (4.1), (4.4) we look for TW solutions. We assume that the solutions have profiles that do not change in time, but travel with a velocity $c$ that will be determined later. Specifically, we extend the domain (the porous medium) to the entire real axis $\mathbb{R}$ and assume that the saturation and the pressure depend on the TW variable $\zeta = ct - x$. Note that this choice is the opposite of $x - ct$ which is commonly used in literature. But our choice is convenient for the analysis below. Moreover, for the ease of presentation, we introduce the negative pressure

$$u = -p.$$
In groundwater terms \( u \) is called suction. In this paper, however, we still refer to \( u \) as pressure. Thus we set

\[
S(x, t) = S(\xi) \quad \text{and} \quad u(x, t) = u(\xi), \quad \text{with} \quad \xi = ct - x.
\]

(4.5)

The wave velocity \( c \in \mathbb{R} \) will be determined later. In terms of \( \xi \), equations (4.1) and (4.4) become

\[
cS' = (k(S)(-u' + 1))',
\]

(4.6)

\[
u = p^+(S) - p^-(S)H_\varepsilon(cS') - c\tau f(S)S',
\]

(4.7)

where \(-\infty < \xi < \infty\). Replacing \( H_\varepsilon \) by \( \text{sign} \) and “\( = \)” by “\( \in \)” we get the travelling wave system corresponding to (4.1) and (4.2).

We consider the case where the saturation and the pressure admit horizontal asymptotes at \( \pm \infty \), i.e.

\[
\lim_{\xi \to -\infty} S(\xi) = S_B, \quad \lim_{\xi \to \infty} S(\xi) = S_T,
\]

(4.8)

\[
\lim_{\xi \to -\infty} u(\xi) = u_B, \quad \lim_{\xi \to \infty} u(\xi) = u_T.
\]

(4.9)

for given saturations \( S_T, S_B \) satisfying \( 0 < S_B < S_T \leq 1 \) and for given pressures \( u_B \) and \( u_T \). We restrict ourselves to the case \( S_B < S_T \) for two reasons:

(i) If a travelling wave exists with \( c > 0 \), then \( S_B < S_T \) describes a wetting (infiltration) front moving from top to bottom through the porous column. This is precisely the physical setting that we aim to describe.

(ii) The convexity of \( k \) implies that travelling waves can only exist if \( S_B < S_T \). This follows directly from the sign of \( u' \) in the \((S, u)\) phase plane.

Integrating (4.6) gives

\[
cS + A = k(S)(-u' + 1) \quad \text{in} \quad \mathbb{R},
\]

(4.10)

where \( A \) is a constant of integration. Using now (4.8) and (4.9) in (4.10) and (4.7) yields

\[
\lim_{\xi \to -\infty} S'(\xi) = \lim_{\xi \to -\infty} u'(\xi) = 0.
\]

Hence (4.7) implies

\[
\lim_{\xi \to -\infty} u(\xi) = p^+(S_B) \quad \text{and} \quad \lim_{\xi \to \infty} u(\xi) = p^+(S_T),
\]

which provides a necessary condition for the existence of TW solutions. We have

**Proposition 4.1.** A necessary condition for the existence of TW solutions is that the components in the left and right states are compatible, namely \( u_\alpha = p^+(S_\alpha) \quad (\alpha \in \{B, T\}) \).
4.3 Capillary hysteresis

In what follows, this compatibility condition is always assumed.

Applying boundary conditions (4.8) and (4.9) to (4.10) we get

\[ c = \frac{k(S_T) - k(S_B)}{S_T - S_B}, \]  

(4.11)

and \( u' = g(S; S_B, S_T) \),

(4.12)

where

\[ g(S; S_B, S_T) = 1 - \frac{c(S - S_B) + k(S_B)}{k(S)} = 1 - \frac{c(S - S_T) + k(S_T)}{k(S)}. \]  

(4.13)

The last equality uses the wave speed expression (4.11).

In this chapter we investigate the effect of hysteresis and dynamic capillarity separately. The combined case will be considered in Chapter 5.

From (4.8), (4.9) and Proposition 4.1 it follows that the points \( E_\alpha = (S_\alpha, u_\alpha) \) with \( u_\alpha = p + c(S_\alpha) \) and \( \alpha \in \{B, T\} \) are equilibria for the dynamical system (4.6), (4.7). Much of the TW analysis will be in terms of orbits in the \((S, u)\) plane, connecting \( E_B \) and \( E_T \).

Clearly, the corresponding waves are translation invariant (invariant to shift in \( \xi \)). To fix the orbit we impose the normalization

\[ S(0) = \frac{1}{2} (S_B + S_T) \text{ and } S(\xi) < \frac{1}{2} (S_B + S_T) \text{ for all } \xi < 0. \]  

(4.14)

We will see later that the inequality in (4.14) is needed as \( S \) has oscillating behaviour near \( S_T \) when \( E_T \) becomes a stable spiral sink. From now on, while discussing travelling waves or orbits, we implicitly assume that (4.14) is satisfied.

4.3 Capillary hysteresis

Dropping the dynamic terms in (4.7) we have

\[ u = p^+_c(S) - p^+_c(S)H_\varepsilon(cS'). \]  

(4.15)

For a given regularisation \( H_\varepsilon \) satisfying (A4) we introduce

\[ \Phi_\varepsilon(r) = \frac{1}{\varepsilon} H_\varepsilon^{-1}(r) \text{ for } -1 < r < 1. \]  

(4.16)

Then \( \Phi_\varepsilon \) satisfies

**Proposition 4.2.** \( \Phi_\varepsilon : (-1, 1) \to \mathbb{R} \) is a smooth, odd and increasing function satisfying \( \Phi_\varepsilon'(0) = \frac{r}{\varepsilon} \) for all \( \varepsilon > 0 \). Also, given two regularisation parameters \( \varepsilon_{1,2} \) such that \( \varepsilon_2 > \varepsilon_1 > 0 \) one has \( |\Phi_{\varepsilon_1}(r)| < |\Phi_{\varepsilon_2}(r)| \) for all \( r \in (-1, 1) \). Finally, \( \lim_{\varepsilon \to 0} \Phi_\varepsilon(r) = 0 \) for all \( r \in (-1, 1) \).
The proof is straightforward and is omitted. Figure 4.1 shows a sketch of $\Phi_\varepsilon$ for different values of $\varepsilon$.

![Sketch of $\Phi_\varepsilon$. The actual plots are for $\Phi_\varepsilon(r) = \frac{\varepsilon r}{\sqrt{1 - r^2}}$ and for the indicated values of $\varepsilon$.](image)

Rewriting (4.15) in terms of $\Phi_\varepsilon$ we obtain for $S$ and $u$ the dynamical system

\[
S' = \Phi_\varepsilon \left( \frac{p^*_c(S) - u}{p^*_c(S)} \right),
\]

\[
u' = \mathcal{G}(S; S_B, S_T).
\]

We consider the cases $S_T = 1$ and $S_T < 1$ separately.

### 4.3.1 The case $S_T = 1$

Since $-1 < H_\varepsilon < 1$, equation (4.15) implies for $0 < S < 1$,

\[
p_c^{(i)}(S) < u < p_c^{(d)}(S)
\]

and so

\[
-1 < \frac{p_c^+(S) - u}{p_c^-(S)} < 1.
\]

The main result of this section is

**Theorem 4.1.** Let $0 < S_B < S_T = 1$ and $E_B = (S_B, p_c^+(S_B)), E_T = (1, 0)$. 
4.3 Capillary hysteresis

(a) Let $\varepsilon > 0$ be fixed. The system (4.17) has a unique orbit $(S_\varepsilon, u_\varepsilon)$ connecting the points $E_B$ and $E_T$. Along the orbit $S$ is increasing and $u$ is decreasing. Consequently, for any $S \in (S_B, 1)$ there exists a unique $\xi(S)$ such that $S_\varepsilon(\xi(S)) = S$. A similar result holds for $u \in (0, p_c^d(S_B))$.

(b) The orbits $(S_\varepsilon, u_\varepsilon)$ are well ordered with respect to $\varepsilon$ and do not intersect except at the equilibrium points $E_B$ and $E_T$. Specifically, if $\varepsilon_2 > \varepsilon_1 > 0$ and $S_{\varepsilon_1}(\xi_1) = S_{\varepsilon_2}(\xi_2) = S$ for some $S \in (S_B, 1)$ and $\xi_{1,2} \in \mathbb{R}$, then $u_{\varepsilon_2}(\xi_1) > u_{\varepsilon_1}(\xi_2)$.

(c) Let $S \in (S_B, 1)$ be fixed. For arbitrary $\varepsilon > 0$, let $w_\varepsilon(S) := u_\varepsilon(\xi(S))$. Then $\lim_{\varepsilon \to 0} w_\varepsilon = p_c^{(i)}(S)$, uniformly on compact subsets of $(S_B, 1)$.

The monotone behaviour of the orbits imply that the TW solutions are monotone in both components. In particular, no overshoot occurs in either pressure or saturation. Moreover the functions $S_\varepsilon : \mathbb{R} \to (S_B, 1)$ and $u_\varepsilon : \mathbb{R} \to (0, p_c^d(S_B))$ are one to one. This is used in (c) of Theorem 4.1: given $S \in (S_B, 1)$, there exists a unique $\xi(S) \in \mathbb{R}$, where $S_\varepsilon(\xi(S)) = S$, which defines the corresponding pressure $w_\varepsilon(S) = u_\varepsilon(\xi(S))$. The function $w_\varepsilon(S)$, with $S_B < S < 1$ and arbitrary $\varepsilon > 0$, describes the orbits as a function of $S$. Observe that, the definition of $\xi(S)$ and $w_\varepsilon$ makes sense only if $S_\varepsilon$ is monotone. If $S_\varepsilon$ is not monotone globally, the functions $\xi(S)$ and $w_\varepsilon$ can still be defined but restricted to intervals where the monotonicity of the saturation holds. This generalisation will be used to describe the case $S_T < 1$ and for the analysis of the dynamic capillarity case.

Differentiation of $w_\varepsilon(S)$ with respect to $S$ gives

$$
w_\varepsilon'(S) = \frac{dw_\varepsilon(S)}{dS} = \frac{dS}{d\xi(S)} \frac{d\xi(S)}{ds} = \frac{\Theta(S; S_B, 1)}{\Phi_\varepsilon \left( \frac{p_c^{(i)}(S) - w_\varepsilon}{p_c^d(S)} \right)}.
$$

(4.20)

To prove Theorem 4.1 we first need some intermediate results. We start with

**Proposition 4.3.** The region $\mathcal{H}^- = \{(S,u) : S_B < S < 1 \text{ and } p_c^{(i)}(S) < u < p_c^{(d)}(S)\}$ is positive invariant for the dynamical system (4.17).

**Proof.** Since $k(\cdot)$ is a convex function it follows that $\Theta(S; S_B, 1) < 0$ for any $S \in (S_B, 1)$. Also $S'_c > 0$ whenever $p_c^{(i)}(S) < u < p_c^{(d)}(S)$. Therefore any orbit $(S_\varepsilon, u_\varepsilon)$ will be monotone in both components as long as it remains in $\mathcal{H}^-$, and the function $w_\varepsilon$ introduced above is well defined.

Referring to Figure 4.2, since $S'_c = 0$ and $u'_c < 0$ along the graph of $p_c^d$, the orbit cannot leave $\mathcal{H}^-$ through the upper boundary. The same holds for the vertical boundary $S = S_B$, since along it one has $S'_c > 0$. Finally, as the orbit approaches the primary imbibition curve $p_c^{(i)}$, one has $S'_c \to +\infty$ and therefore $\frac{dw_\varepsilon}{ds} \to 0$. Since
$p_c^{(i)}' < 0$, this implies that the orbit cannot leave $\mathcal{H}^-$ through the lower boundary as well. Hence $\mathcal{H}^-$ is invariant.

The next proposition characterises the equilibrium point $E_B$.

**Proposition 4.4.** $E_B$ is a saddle type equilibrium.

**Proof.** Linearising (4.17) around any equilibrium point $E_\alpha = (S_\alpha, p_c^+(S_\alpha))$ ($\alpha \in \{B, T\}$) yields the characteristic equation

$$
\lambda^2 - \Phi'_c(0) \frac{p_c^+(S_\alpha)}{p_c^+(S_\alpha)} \lambda + \Phi'_c(0) \frac{(k'(S_\alpha) - c)}{k(S_\alpha)p_c^+(S_\alpha)} = 0.
$$

(4.21)

Since $k$ is convex one has $k'(S_B) < c < k'(1)$. Hence, at $E_B$, the last term on the left is negative, which proves the result.

**Remark 4.1.** Since $\Phi'_c(0) = \frac{\xi}{c}$ the positive eigenvalue in 4.21 at $E_B$, $\lambda = \lambda_{+,B,\varepsilon}$, satisfies

$$
\lambda_{+,B,\varepsilon} = \sqrt{C_1^2 \varepsilon^2 + C_2 \varepsilon - C_1 \varepsilon^2} = \theta(\sqrt{\varepsilon}), \quad \text{as } \varepsilon \to 0
$$

for appropriately chosen $C_{1,2} > 0$.

We now turn to the proof of Theorem 4.1.

**Proof.** (a) Consider the situation near the saddle at $E_B$. A direct calculation shows that the eigenvector corresponding to the unstable eigenvalue $\lambda_{+,B,\varepsilon} > 0$ points into the region $\mathcal{H}^-$ for increasing $S$. Let $(S_\varepsilon, u_\varepsilon)$ be the unique orbit leaving $E_B$ in this direction. By the invariance of $\mathcal{H}^-$ and the sign of the right hand sides in (4.17), the
orbit remains in $\mathcal{H}^-$ with increasing $S_\varepsilon$ and decreasing $u_\varepsilon$. As $\xi \to +\infty$ it can only end up in the boundary point $E_T = (1,0)$.

(b) Letting $S \to S_B$ in (4.20) we obtain

$$w_\varepsilon'(S_B) = \frac{p_c^+(S_B)}{2} \left( 1 + \sqrt{1 + \frac{4(\varepsilon - k'(S_B))p_c^-(S_B)}{k(S_B)\Phi_c'(0)(p_c^+(S_B))^2}} \right).$$

Since $\Phi_c'(0) = \frac{\varepsilon}{2}$, it follows that $w_{\varepsilon_1}'(S_B) < w_{\varepsilon_2}'(S_B) < 0$ for any $0 < \varepsilon_1 < \varepsilon_2$. Using $w_{\varepsilon_1}(S_B) = w_{\varepsilon_2}(S_B) = p_c^+(S_B)$ we have $w_{\varepsilon_1}(S) < w_{\varepsilon_2}(S)$ in a right neighbourhood of $S_B$. Now suppose there exists $S^* \in (S_B,1)$ such that $w_{\varepsilon_1}(S^*) < w_{\varepsilon_2}(S^*)$ for $S_B < S < S^*$ and $w_{\varepsilon_2}(S^*) = w_{\varepsilon_1}(S^*)$. Then $w_{\varepsilon_1}'(S^*) > w_{\varepsilon_2}'(S^*)$. This contradicts (4.20) at $S^*$.

(c) At this point we know that for all $\varepsilon > 0$,

- $w_\varepsilon(S_B) = p_c^+(S_B), w_\varepsilon(1) = 0$ and $w_\varepsilon > p_c^{(j)}$ in $(S_B,1)$.
- For any pair $0 < \varepsilon_1 < \varepsilon_2$, $w_{\varepsilon_1} < w_{\varepsilon_2}$ in $(S_B,1)$.

As a consequence

$$\lim_{\varepsilon \to 0} w_\varepsilon(S) = \tilde{w}(S) \text{ for each } S_B \leq S \leq 1,$$

where $\tilde{w} : [S_B,1) \to [0,p_c^+(S_B)]$ satisfies $\tilde{w}(S_B) = p_c^+(S_B), \tilde{w}(1) = 0$ and $\tilde{w}(S) \geq p_c^{(j)}(S)$ for $S_B < S < 1$. Moreover $\tilde{w}(S)$ is non-increasing in $[S_B,1)$, which is inherited from the monotonicity of $w_\varepsilon$ in $[S_B,1]$.

Now suppose there exists $S_0 \in (S_B,1)$ such that $\tilde{w}(S_0) > p_c^{(j)}(S_0)$. Then there exists $\delta > 0, \delta$ small enough, so that $S_0 - \delta > S_B$ and $w_\varepsilon(S) > w_\varepsilon(S_0) \geq \tilde{w}(S_0) > p_c^{(j)}(S_0 - \delta)$ for $S \in (S_0 - \delta, S_0)$. In this situation, all orbits pass through the region (see Figure 4.3)

$$\mathcal{R} = \{(S,u) : S_0 - \delta/2 < S < S_0 \text{ and } p_c^{(j)}(S_0 - \delta) < u < p_c^+(S)\}.$$  \hspace{1cm} (4.23)

In $\mathcal{R}$ we have

$$\mathcal{G}(S;S_B,S_T) < -C \text{ for some } C > 0,$$  \hspace{1cm} (4.24)

and since $\mathcal{R}$ does not touch $p_c^{(j)}$,

$$m = \sup_{(S,u) \in \mathcal{R}} \left( \frac{p_c^+(S) - u}{p_c^+(S)} \right), \text{ with } 0 < m < 1.$$  \hspace{1cm} (4.25)

Using this in (4.20) we find, using the monotonicity of $\Phi_\varepsilon$

$$-w_\varepsilon'(S) \geq \frac{C}{\Phi_\varepsilon(m)},$$  \hspace{1cm} (4.25)

for all $\varepsilon > 0$ and for $S_0 - \delta/2 < S < S_0$. Integration from $S = S_0 - \delta/2$ to $S = S_0$ gives

$$\frac{C\delta}{2\Phi_\varepsilon(m)} \leq w_\varepsilon \left( S_0 - \frac{\delta}{2} \right) - w_\varepsilon(S_0) < p_c^+(S_B) - p_c^{(j)}(S_0).$$

Letting $\varepsilon \to 0$ we reach a contradiction. Hence $\tilde{w}(S) = p_c^{(j)}(S)$ for $S_B < S \leq 1$. By Dini’s Theorem the convergence is uniform on any closed interval $[S_B + \mu,1]$ with $\mu > 0$. \qed
Passing the limit $\varepsilon \to 0$ gives the TW solutions that corresponds to the play-type hysteresis. In terms of saturation $S = S(\xi)$ it runs from $S = S_B$ as $\xi \to -\infty$ to $S = S_T = 1$ as $\xi \to \infty$, while $u = p_c ^{(i)}(S)$. We make this precise in the following corollary

**Corollary 4.1.** Let $\xi^*_\varepsilon \in \mathbb{R}$ be such that $u_\varepsilon(\xi^*_\varepsilon) = p_c ^{(i)}(S_B)$ and let $S^*_\varepsilon = S_c(\xi^*_\varepsilon)$. Then

$$\lim_{\varepsilon \to 0} S^*_\varepsilon = S_B$$

and

$$\lim_{\varepsilon \to 0} \xi^*_\varepsilon = -\infty.$$

Before giving the proof we observe that in view of the convergence result in Theorem 4.1 (c), this corollary shows that for $\varepsilon \searrow 0$ the orbits become vertical when approaching $E_B$.

**Proof.** Since the orbits $(S_\varepsilon, u_\varepsilon)$ are ordered, $S^*_\varepsilon$ decreases with $\varepsilon$. Moreover, by construction $S^*_\varepsilon > S_B$. Hence $\lim S^*_\varepsilon = S^*$ exists. Assuming $S^* > S_B$ leads to a contradiction as in the proof of Theorem 4.1 (c). Thus $S^* = S_B$.

To prove the second statement we first write equation 4.17a in terms of $\xi_\varepsilon$, $\xi_\varepsilon$ being defined in Theorem 4.1 (a):

$$\frac{d\xi_\varepsilon}{dS}(S) = \frac{1}{\Phi_\varepsilon \left( \frac{p_c ^{(i)}(S) - w_\varepsilon(S)}{p_c(S)} \right)}.$$  \hspace{1cm} (4.26)

As $S^*_\varepsilon = S(\xi^*_\varepsilon)$ and $S_\varepsilon(0) = (S_B + S_T)/2$, integrating (4.26) and using (4.20) we get

$$-\xi^*_\varepsilon = \int_{S^*_\varepsilon}^{S_\varepsilon(0)} \frac{dS}{\Phi_\varepsilon \left( \frac{p_c ^{(i)}(S) - w_\varepsilon(S)}{p_c(S)} \right)} = \int_{S^*_\varepsilon}^{S_\varepsilon(0)} \frac{w'_\varepsilon(S)}{\mathcal{G}(S; S_B, 1)} dS.$$  

![Figure 4.3: The saturation $S_0$ and the region $\mathcal{R}$ for $\varepsilon > 0$.](image)
Since $S^*_e \to S_B$, for any $\delta > 0$ there exists a $\bar{\mu} = \bar{\mu}(\delta)$ such that $S_B < S^*_e < S_B + \delta$ for all $0 < \varepsilon < \bar{\mu}(\delta)$. Using $\mathcal{G} \in C^1([S_B, 1])$ and $M_{\mathcal{G}} = \max\{|\mathcal{G}'(S)| : S \in [S_B, \frac{1}{2}(S_T + S_B)]\}$ we estimate

\[ -\xi^*_e \geq \frac{1}{M_{\mathcal{G}}} \int_{S^*_e} S^*_e - w'_e(S) dS \geq \frac{1}{M_{\mathcal{G}}} \int_{S_B + \delta} \frac{1}{2} (S_T + S_B) - w'_e(S) dS =: \frac{1}{M_{\mathcal{G}}} h_{\varepsilon}. \tag{4.27} \]

Evaluating $h_{\varepsilon}$ gives

\[ h_{\varepsilon} = \frac{w_e(S_B + \delta)}{\delta} = 2 \int_{S_B + \delta} \frac{1}{2} (S_T + S_B) - \frac{1}{2} (S_B + S_B) dS. \]

Since $w_e$ converges uniformly in $|S_B + \delta, \frac{1}{2}(S_B + S_T)|$ and since $p_c^{(i)} \in C^1$ we have

\[ \lim_{\varepsilon \to 0} h_{\varepsilon} = \frac{p_c^{(i)}(S_B + \delta)}{\delta} - 2 \int_{S_B + \delta} \frac{1}{2} (S_T + S_B) dS = \frac{1}{2} (S_T + S_B) =: h_0. \]

Therefore for any $\nu > 0$, there exists a $\mu^*(\nu) > 0$ such that $h_{\varepsilon} > h_0 - \nu$ for all $\varepsilon \in (0, \mu^*(\nu))$. Thus for $0 < \varepsilon < \min\{\bar{\mu}(\delta), \mu^*(\nu)\}$

\[ -\xi^*_e \geq \frac{1}{M_{\mathcal{G}}} (h_0 - \nu) \geq \frac{M_p}{M_{\mathcal{G}}} \ln \left( \frac{S_T - S_B}{2\delta} \right) - \frac{\nu}{M_{\mathcal{G}}}, \tag{4.28} \]

where $M_p = \min\{-p_c^{(i)}(S) : S_B < S < \frac{1}{2}(S_B + S_T)\}$. Since $\delta$ can be chosen arbitrarily small, this concludes the proof.

### 4.3.2 The case $S_T < 1$

We consider now the case when the top and bottom saturations satisfy $0 < S_B < S_T < 1$. In the analysis we use the region

\[ \mathcal{H} = \{(S, u) : S_B \leq S \leq 1, p_c^{(i)}(S) \leq u \leq p_c^{(d)}(S)\}, \]

and its subregions (see Figure 4.4)

\[ \mathcal{H}_1 = \{(S, u) : S_B \leq S \leq S_T, p_c^{(i)}(S) \leq u \leq p_c^{+}(S)\}, \]
\[ \mathcal{H}_2 = \{(S, u) : S_T \leq S \leq 1, p_c^{(i)}(S) \leq u \leq p_c^{+}(S)\}, \]
\[ \mathcal{H}_3 = \{(S, u) : S_T \leq S \leq 1, p_c^{+} \leq u \leq p_c^{(d)}(S)\}, \]
\[ \mathcal{H}_4 = \{(S, u) : S_B \leq S \leq S_T, p_c^{+} \leq u \leq p_c^{(d)}(S)\}. \]

We first analyse the case when $\varepsilon > 0$. 

Properties for fixed $\varepsilon > 0$

The key properties of the orbits are stated in

Theorem 4.2. Let $0 < S_B < S_T < 1$ and $\varepsilon > 0$. Then the following holds

(a) There exists a unique orbit $(S_\varepsilon, u_\varepsilon)$ satisfying (4.17), (4.14) and connecting $E_B$ and $E_T$.

(b) There exists a $\varepsilon_m > 0$ such that $E_T$ is a stable spiral sink whenever $0 < \varepsilon < \varepsilon_m$.

Proof. (a) Repeating the proof for the $S_T = 1$ case, we observe that the equilibrium $E_B$ is a saddle and all orbits leaving $E_B$ along the unstable direction and for increasing $S_\varepsilon$ enter the region $\mathcal{H}^-$ introduced in Proposition 4.3. Also, no orbit can leave the region $\mathcal{H}$ defined above through the primary curves $p_i^{(i)}(S)$ and $p_d^{(d)}(S)$. Now we let $\bar{S}$ be such that $p_d^{(d)}(\bar{S}) = p_i^{(i)}(S_B)$. Then two cases can be identified, $S_T < \bar{S}$ and $S_T \geq \bar{S}$.

The case $S_T < \bar{S}$: With respect to Figure 4.4, the orbit leaving $E_B$ for increasing $S_\varepsilon$ enters first the region $\mathcal{H}_1$. Then there are four possibilities (see Figure 4.4)

1. The orbit goes through $\mathcal{H}_2, \mathcal{H}_3, \mathcal{H}_4$ and returns to $E_B$.
2. The orbit goes through $\mathcal{H}_2, \mathcal{H}_3, \mathcal{H}_4$ and leaves $\mathcal{H}_4$ through the segment $(S_B, u_B)$, $(S_B, p_d^{(d)}(S_B))$.
3. The orbit goes through $\mathcal{H}_2, \mathcal{H}_3, \mathcal{H}_4$ and then leaves $\mathcal{H}_4$ through the arc $(S, p_c^{(c)}(S))$ between $E_B$ and $E_T$. This in turn gives rise to two possibilities:
4.3 Capillary hysteresis

The case $S_T \geq \hat{S}$: In this case, if the orbit enters from $\mathcal{H}_3$ to $\mathcal{H}_4$ at some $\xi = \xi_{3-4}$, $u_e(\xi_{3-4}) < p_c^{(d)}(S_T) < p_c^{(d)}(\hat{S}) = u_B$. But in $\mathcal{H}_4$, $u_\xi$ is decreasing, hence $u_e < u_B$ for all arguments $\xi > \xi_{3-4}$, which rules out the first two possibilities (possibility 1 and 2) above.

To show that actually 3.B is the only possibility in both cases, we follow an argument from [68], based on Divergence Theorem. We define the vector-valued function

$$F : \mathcal{H} \to \mathbb{R}^2, \quad F(S, u) = (\Phi_e(S, u), G(S)),$$

and denote its components by $F_s$ and $F_u$ respectively. A direct calculation gives

$$\nabla \cdot F = \frac{\Phi'_e}{(p_c^{-})^2} \cdot (p_c^+ p_c^- - p_c^+ p_c^- + u p_c^-),$$

where the arguments $S$ and $u$ are disregarded. Hence, for $(S, u) \in \mathcal{H}$ one has

$$\nabla \cdot F(S, u) < \begin{cases} 
\frac{\Phi'_e}{(p_c^{-})^2} \cdot (p_c^+ p_c^- - p_c^+ p_c^- + p_c^- p_c^{(i)}) & \text{if } p_c^- > 0 \\
\frac{\Phi'_e}{(p_c^{-})^2} \cdot (p_c^+ p_c^- - p_c^+ p_c^- + p_c^- p_c^{(d)}) & \text{if } p_c^- < 0 
\end{cases}$$

The last factor in the first inequality gives

$$\frac{1}{4} (p_c^{(d)} + p_c^{(i)}) (p_c^{(d)} - p_c^{(i)}) - \frac{1}{4} (p_c^{(d)} + p_c^{(i)}) (p_c^{(d)} - p_c^{(i)})' + \frac{p_c^{(i)}}{2} (p_c^{(d)} - p_c^{(i)})' < 0,$$

Similarly, in the second inequality one gets $\frac{1}{2} (p_c^{(d)} - p_c^{(i)}) p_c^{(d)}' < 0$. Thus we have shown that

$$\nabla \cdot F(S, u) < 0 \quad \text{for all } (S, u) \in H.$$

We can now investigate the possibilities mentioned above. To rule out the first two possibilities in the case $S_T < \hat{S}$ we define the domain $\Omega$ bounded by the closed orbit, or by the orbit and the segment $(S_B, u_B), (S_B, p_c^{(d)}(S_B))$ (see Figure 4.5). Let the orbit intersect the segment $(S_B, u_B), (S_B, p_c^{(d)}(S_B))$ at the point $T$. So for possibility 1, $T$ is simply $E_T$. By (4.30) one has

$$0 > \int_{\Omega} \nabla \cdot F = \int_{E_B}^{T} F \cdot \dot{r} + \int_{T}^{E_B} F \cdot \dot{r} = 0 - \int_{T}^{E_B} F_s,$$
with the last integral on the right appearing only in the second possibility listed above. Since $F_S \leq 0$ in the region $\mathcal{H} \setminus \mathcal{H}^-$, this gives a contradiction.

Finally, to eliminate 3.A we observe that by the Poincaré-Bendixson Theorem, if the orbit does not end up in $E_T$ then it must approach a limit cycle around $E_T$. However, one can use again the argument above, to show that limit cycles do not exist. So, the only possible behaviour of the orbits is as stated in possibility 3.B. This is displayed in the left plot of Figure 4.6. Also this orbit is unique if condition (4.14) is taken into account as this clearly fixes $\xi = 0$.

(b) Having proved the existence of an orbit connecting $E_B$ and $E_T$, showing that the orbit forms a stable spiral around $E_T$ for small enough $\epsilon$ is a matter of calculation. Using the properties of $\Phi_{\epsilon}$, $p_c^+$ and the convexity of $k$ in (4.21) it is easy to show that for small values of $\epsilon$, the eigenvalues corresponding to equilibrium point $E_T$ will be complex with negative real part. This completes the proof.

The (left) plot in Figure 4.6 shows the phase portrait in the $S$-$u$ plane. In the (right) plot one has orbit component $S$ as function of $-\xi$, in the case when $E_T$ is a stable spiral. Note the usage of $-\xi = x - ct$ instead of $\xi$, which is because in the original problem (with $x$ and $t$ as independent variables) the left state ($x = -\infty$) corresponds to $S_T$ and right state ($x = \infty$) corresponds to $S_B$. This convention is used when comparing with numerical solutions to (4.1), (4.4).

**Properties for the limit case $\epsilon \to 0$**

Knowing now the structure of the orbits for fixed $\epsilon > 0$, we study their behaviour as $\epsilon \to 0$. In certain aspects, the results obtained for $S_T = 1$ and for $S_T < 1$ are quite similar. The major difference is in the fact that the orbits are not monotone anymore.
Consequently, the function $w_\varepsilon$ introduced Theorem 4.1 can only be defined as long as $S_\varepsilon$ remains monotone. Clearly, when starting from $E_B$ the monotonicity is lost for the first argument $\xi$ where $S_\varepsilon(\xi) = S_T$. We define $\xi^T_\varepsilon$ as

$$\xi^T_\varepsilon = \min\{\xi \in \mathbb{R} : S_\varepsilon(\xi) = S_T\}. \quad (4.31)$$

From now on we refer to the function $w_\varepsilon$ as the one obtained for $\xi \in (-\infty, \xi^T_\varepsilon]$. With this, one has

**Proposition 4.5.** (a) As long as $S \leq S_T$ the orbits $(S_\varepsilon, u_\varepsilon)$ are well ordered with respect to $\varepsilon > 0$, and do not intersect.

(b) For any $S \in (S_B, S_T]$, $w_\varepsilon(S) \to p_c^{(i)}(S)$ as $\varepsilon \to 0$, uniformly on compact subsets of $(S_B, S_T]$.

The proof is the same as for Theorem 4.1 and is therefore omitted here.

For the case $S_T = 1$, Corollary 4.1 is stating the limit behaviour of the orbits when $\varepsilon \searrow 0$. The nature of the equilibrium $E_B$ remains unchanged when $S_T < 1$. Therefore similar results hold as before: if $\xi^*_\varepsilon \in \mathbb{R}$ is such that $u_\varepsilon(\xi^*_\varepsilon) = p_c^{(i)}(S_B)$, for $S^*_\varepsilon = S_\varepsilon(\xi^*_\varepsilon)$ one has

$$\lim_{\varepsilon \to 0} S^*_\varepsilon = S_B \quad \text{and} \quad \lim_{\varepsilon \to 0} \xi^*_\varepsilon = -\infty,$$

and the corresponding orbits become vertical when approaching $E_B$.

The situation changes for $E_T$ since the orbits $(S_\varepsilon, u_\varepsilon)$ form stable spirals for small $\varepsilon$. To understand this behaviour we let $\tilde{\xi}_\varepsilon = \min\{\xi \in \mathbb{R} : S_\varepsilon(\xi) = S_T\}$ and define (see Figure 4.6 (left))

$$R_\varepsilon = \sup\{S_\varepsilon(\xi) : u_\varepsilon(\xi) = p_c^{+}(S_\varepsilon(\xi))\} \quad \text{and} \quad L_\varepsilon = \inf\{S_\varepsilon(\xi) : u_\varepsilon(\xi) = p_c^{+}(S_\varepsilon(\xi)), \xi > \tilde{\xi}_\varepsilon\}.$$

The following statement is proved
Proposition 4.6. For $\tilde{\xi}_\varepsilon$, $L_\varepsilon$ and $R_\varepsilon$ introduced above, one has

$$\lim_{\varepsilon \to 0} \tilde{\xi}_\varepsilon = \infty, \quad \lim_{\varepsilon \to 0} L_\varepsilon = S_T \quad \text{and} \quad \lim_{\varepsilon \to 0} R_\varepsilon = S_T.$$ 

Proof. The proof for $\tilde{\xi}_\varepsilon$ is almost identical to the proof of Corollary 4.1. For the remaining part we only consider $R_\varepsilon$, the proof for $L_\varepsilon$ being similar.

Clearly, $R_\varepsilon \geq S_T$. Assuming that a $\delta > 0$ and a sequence $\varepsilon_k \to 0$ exist such that $R_{\varepsilon_k} > S_T + \delta$ for all $k \in \mathbb{N}$. Let

$$R = \{(S, u) : S_T + \frac{\delta}{2} < S < S_T + \delta \text{ and } p_c^{(i)}(S_T) \leq u \leq p_c^{(i)}(S)\}.$$ 

Clearly, all orbits pass through $R$. Letting

$$M = \sup_{(S, u) \in R} \left(\frac{p_c^{(i)}(S) - u}{p_c^{(i)}(S)}\right)$$

one has $0 \leq M < 1$ and $0 \leq \Phi_\varepsilon\left(\frac{p_c^{(i)}(S) - u}{p_c^{(i)}(S)}\right) < \Phi_\varepsilon(M)$ for all $(S, u) \in R$. From (4.20) and recalling that $k$ is convex, for any $S \in (S_T + \frac{\delta}{2}, S_T + \delta)$ one has

$$w_{\varepsilon_k}'(S) > \frac{\Phi(S; S_B, S_T)}{\Phi_\varepsilon(M)} > \frac{(k(S) - k(S_T)) - c(S - S_T)}{k(S_T + \delta) \cdot \Phi_\varepsilon(M)}.$$ 

Integrating the above over $(S_T + \frac{\delta}{2}, S_T + \delta)$ and using the properties of $k$, a constant $C_0 > 0$ depending on $\delta$ but not on $\varepsilon$ exists such that

$$w_{\varepsilon_k}(S_T + \delta) - w_{\varepsilon_k}\left(S_T + \frac{\delta}{2}\right) > \frac{C_0}{\Phi_\varepsilon(M)}.$$ 

In the above, the difference on the left is bounded by $p_c^{(i)}(S_T) - p_c^{(i)}(S_T + \delta)$. However, by Proposition 4.2, the ratio on the right goes to $\infty$ when $\varepsilon \to 0$, which gives a contradiction. This implies that $R_\varepsilon \to S_T$ for $\varepsilon \to 0$.

Proposition 4.7. The limit solution $(S(\xi), p_c^{(i)}(S(\xi)))$ solves the following boundary value problem

$$cS' = [k(S)((p_c^{(i)}(S))' + 1)], \quad S(-\infty) = S_B, \quad S(\infty) = S_T.$$
4.3 Capillary hysteresis

Proof. From (4.26) we get that
\[
\xi_\varepsilon(S) = -\int_{\frac{1}{2}(S_B+S_T)}^{S} \frac{w_\varepsilon'(S)}{\mathcal{G}(S)} \, dS.
\] (4.32)
Following the exact steps produced in Corollary 4.1, i.e. integrating by parts (4.32) and then using \(\lim_{\varepsilon \to 0} w_\varepsilon(S) = p_c^{(i)}(S)\), one obtains that the limit
\[
\xi_0(S) = \lim_{\varepsilon \to 0} \xi_\varepsilon(S),
\]
exists. Moreover, \(\xi_0(S)\) depends continuously and strictly monotonically on \(S\) for all \(S \in (S_B, S_T)\) and \(\xi_0(S_B) = -\infty, \xi_0(S_T) = \infty\), the last one following from Proposition 4.6. This means for any \(\xi \in \mathbb{R}\) one can find a \(S = \xi^{-1}_0(\xi)\). This makes the solution \(S(\xi)\) well defined. Differentiating the limit version of (4.32) with respect to \(S\) and inverting we get
\[
\frac{dS}{d\xi}(\xi) = -\frac{\mathcal{G}(S(\xi))}{p_c^{(i)}(S(\xi))}.
\] (4.33)
Rearranging terms completes the proof.

Remark 4.2. Propositions 4.5, 4.6 and 4.7 characterise the behaviour of the orbits in the limiting case \(\varepsilon \to 0\). These orbits are approaching vertical segments at \(S = S_B\) and \(S = S_T\), and in between the primary imbibition curve (see Figure 4.7). Possible oscillations can appear around \(E_T\) when \(S_T < 1\). As \(\varepsilon \to 0\), these oscillations are damped in the \(S\) component, but we are unable to show a similar behaviour for

\[\text{Figure 4.7: Orbit for limiting case } \varepsilon \to 0 \text{ in } S-u \text{ plane (left); and saturation and pressure profiles for the limiting orbit as a function of } -\xi = x - ct \text{ (right).}\]
the pressure. Computational results shown in Figure 4.17 and 4.18 indicate that pressure oscillations do not decay as \( \varepsilon \) decreases. However, these oscillations cannot be observed in reality for \( \varepsilon \to 0 \) as they are pushed towards infinity. Proceeding as in Corollary 4.1, one can show that \( \xi^* \to -\infty \) as \( \varepsilon \to 0 \) and a similar result holds for the other side, determined by \( S_T \). In other words, the oscillations move to infinity and at any finite point the limiting waves are monotone in both saturation and pressure and they lie on the primary imbibition curve.

### 4.4 Dynamic capillarity

Now we discuss the case without hysteresis, but include dynamic effects in the \( p_c - S \) relationship. More precisely, we assume that the primary curves in (1.22) are the same, \( \max(p_c^i) = \min(p_c^d) \), giving \( p_c^-(S) = 0 \) and \( p_c^+(S) = \max(p_c^i(S)) \) for all \( S \). For the ease of presentation, as many results in this case are similar to the ones for the hysteresis model, we still use the notations \( p_c^\pm \).

At the same time we now take \( \tau > 0 \) and thus (4.2) and (4.4) become

\[

u = p_c^+(S) - \tau f(S) \partial_t S.

\]

With the TW velocity \( c \) given in (4.11), the dynamical system (4.6)-(4.7) associated to the TW solutions become,

\[

S' = \frac{p_c^+(S)}{c} - u,

u' = \mathcal{G}(S; S_B, S_T).
\]

As before, we seek orbits that connect the equilibria \( E_B = (S_B, p_c^+(S_B)) \) and \( E_T = (S_T, p_c^+(S_T)) \), where \( 0 < S_B < S_T \leq 1 \). To fix the orbits we normalize the orbits by assuming that \( S(0) = (S_B + S_T)/2 \). We remark that this section borrows ideas and extends results from [68, 236, 248, 254]. Before investigating the existence of the TW solutions to the system (4.34) we observe that in certain cases the analysis can be reduced to the simpler case \( f \equiv 1 \). To see this we introduce the transformed variable

\[

Y = Y(S) = \int_0^S f(\varrho) d\varrho.
\]

Following from Assumption (A2), this transformation has a unique inverse which we write as \( S = S(Y) \). Also, \( Y(1) < \infty \) if and only if \( f \in L^1(0, 1) \). In terms of \( (Y, u) \), the system (4.34) becomes

\[

Y' = \frac{\tilde{p}_c^+(Y) - u}{c},

u' = \tilde{\mathcal{G}}(Y; Y_B, Y_T),
\]

with \( Y_\alpha = \int_0^{S_\alpha} f(\varrho) d\varrho \) for \( \alpha \in \{ B, T \} \) and the functions \( \tilde{p}_c^+, \tilde{\mathcal{G}} \) defined as

\[

\tilde{p}_c^+(Y) = p_c^+(S(Y)) \quad \text{and} \quad \tilde{\mathcal{G}}(Y; Y_B, Y_T) = \mathcal{G}(S(Y); S_B, S_T).
\]
Observe that the system (4.36) is qualitatively similar to (4.34) for the constant damping function, \( f \equiv 1 \). The difference is in a reinterpretation of the nonlinearities \( p_c^+ \) and \( \psi \). In view of this we start analysing the existence of TW solutions and their properties by replacing (4.34a) with the simpler equation

\[
S' = \frac{p_c^+(S) - u}{c\tau}.
\]

The more general case when \( f \in L^1(0, 1) \) is discussed briefly at the end of Section 4.4.1. Moreover, the analysis also extends to cases when \( f \not\in L^1(0, 1) \). As will be seen in Section 4.4.2, the case \( f \not\in L^1 \) gives a natural framework in which the saturation remains within the physically relevant range, \( S \in [0,1] \).

4.4.1 General behaviour of the orbits

![Figure 4.8: The directions followed by the orbits in the S-u plane for dynamic capillary case.](image)

As for the hysteresis case, in this part we analyse the existence of orbits of the system (4.34b), (4.38) connecting the equilibrium points \( E_B \) and \( E_T \). Clearly, these orbits will depend on \( \tau \), motivating the notation \( (S, u) \). Below we use the regions

\[
\mathcal{H}_1 = \{(S, u) : S_B \leq S \leq S_T, u \leq p_c^+(S)\}, \quad \mathcal{H}_2 = \{(S, u) : S_T \leq S \leq 1, u \leq p_c^+(S)\},
\]

\[
\mathcal{H}_3 = \{(S, u) : S_T \leq S \leq 1, p_c^+(S) \leq u\}, \quad \mathcal{H}_4 = \{(S, u) : S_B \leq S \leq S_T, p_c^+(S) \leq u\}.
\]

Figure 4.8 shows the directions followed by the orbits of the system (4.38), (4.34b). Note that if an orbit goes through \( \mathcal{H}_1 \), there it is monotone in both components, namely \( u'_\tau < 0 \) and \( S'_\tau > 0 \). Hence an orbit can only exit \( \mathcal{H}_1 \) through the line \( S = S_T \).
A straightforward calculation shows that the eigenvalues for the linearisation of (4.34b), (4.38) around $E = (S_B, p_c^k(S_B))$ ($a \in (B, T)$) are

$$
\lambda^\pm = \frac{(p_c^k)'(S_B)}{2cT} \left( 1 \pm \sqrt{1 - \frac{4cT(k'(S_B) - c)}{k(S_B)(p_c^k(S_B))^2}} \right). 
$$

(4.39)

Since $k$ is convex, one has $k'(S_T) > c > k'(S_B)$, which shows that $E_B$ is a saddle point. Further, the unstable orbit leaving $E_B$ to the right enters the region $H_1$. To understand its behaviour as $\xi \to \infty$ we begin with

**Proposition 4.8.** Given $\tau > 0$, the orbit $(S_T, u_\tau)$ leaving $E_B$ into $H_1$ either approaches $E_T$ from $H_1$ as $\xi \to \infty$, or leaves $H_1$ through the vertical line $S = S_T$.

**Proof.** In view of the monotonicity inside $H_1$, if $(S_T, u_\tau)$ does not leave $H_1$ through its right boundary it will approach an equilibrium contained in $H_1$ and at the right of $E_B$. Since $k$ is a convex function, the only such point is $E_T$. \qed

As for the hysteresis model, all orbits $(S_T, u_\tau)$ are monotone between $(S_B, S_T)$. So, similar to Theorem 4.1, with the stated normalization $S_T(0) = (S_B + S_T)/2$, it is possible to define the functions $\xi_\tau$, $w_\tau : (S_B, S_T) \to \mathbb{R}$ for the dynamic capillarity model as well. More precisely, for any $S \in (S_B, S_T)$, a unique $\xi_\tau(S)$ exists such that $S_T(\xi_\tau(S)) = S$ and $S_T(\xi) < S$ for all $\xi < \xi_\tau(S)$. With this, $w_\tau(S) = u_\tau(\xi_\tau(S))$. Also one can extend $w_\tau$ to the closed interval $[S_B, S_T]$.

We emphasize on the fact that the functions are defined as long as $S_T$ remains increasing. In particular, this holds until the orbit leaves $H_1 \cup H_2$. Similar to (4.20), $w_\tau$ satisfies the equation

$$
w_\tau'(S) = \frac{\tau c \theta(S; S_B, S_T)}{p_c^k(S) - w_\tau}. 
$$

(4.40)

The propositions below explain how the orbits $(S_T, u_\tau)$ depend on $\tau$, before they leave $H_1$.

**Proposition 4.9.** For the family of functions $w_\tau$ introduced above one has

(a) $w_\tau \to p_c^k$ uniformly in $[S_B, S_T]$ as $\tau \to 0$.

(b) For any $S \in (S_B, S_T)$, $w_\tau(S) \to -\infty$ as $\tau \to \infty$.

**Proof.** We define the family of functions $v_\tau : [S_B, S_T] \to [0, \infty)$, $v_\tau(S) = p_c^k(S) - w_\tau(S)$. Note that since $(S_T, u_\tau) \in H_1$, $v_\tau$ is always positive. By (4.40) we get

$$
\frac{1}{2}(v_\tau^2)'(S) = v_\tau v_\tau'(S) = -cT \theta(S; S_B, S_T) + v_\tau p_c^k \leq -cT \theta(S; S_B, S_T). 
$$

(4.41)
Integration from \( S = S_B \) to an arbitrary \( S \in (S_B, S_T) \) gives

\[
\nu_1^2(S) \leq -2c \int_{S_B}^{S} g(q; S_B, S_T) dq \leq -2c \int_{S_B}^{S_T} g(q; S_B, S_T) dq = 2\tau \tilde{K},
\]

with \( \tilde{K} = -c \int_{S_B}^{S_T} \varphi(q) dq \geq 0 \). This implies

\[
0 \leq p_c^+(S) - \nu_1(S) \leq \sqrt{2c\tilde{K}}. \tag{4.42}
\]

Observing that \( \tilde{K} \) does not depend on \( S \), the conclusion follows immediately.

For the second part, assume there exists \( L > 0 \) and \( S^* \in (S_B, S_T) \) such that \( w_{\tau_k}(S^*) \geq p_c^+(S^*) - L \) for a sequence \( \{\tau_k\}_{k \in \mathbb{N}} \) going to infinity. Since \( w_{\tau_k} \) is strictly decreasing in \([S_B, S_T]\) we have \( p_c^+(S) - w_{\tau_k}(S) < p_c^+(S) - p_c^+(S^*) + L \) if \( S_B < S < S^* \). Since \( g(S; S_B, S_T) < 0 \) in \( \mathcal{H}_1 \) integration of (4.40) gives

\[
w_{\tau_k}(S^*) = w_{\tau_k}(S_B) + c\tau_k \int_{S_B}^{S^*} \frac{g(q)}{p_c^+(q) - w_{\tau_k}(q)} dq
\]

\[
< p_c^+(S_B) + c\tau_k \int_{S_B}^{S^*} \frac{g(q; S_B, S_T)}{p_c^+(q) - p_c^+(S^*) + L} dq = p_c^+(S_B) - c\tau_k K_s, \tag{4.43}
\]

with \( K_s = \int_{S_B}^{S^*} \frac{g(q; S_B, S_T)}{p_c^+(q) - p_c^+(S^*) + L} dq \). Clearly, \( K_s > 0 \). Since \( \lim_{k \to \infty} \tau_k = \infty \), this contradicts the assumed boundedness of \( w_{\tau_k} \) and the proposition is proved.

The orbits depend continuously and monotonically on \( \tau \), as follows from

**Proposition 4.10.** For all \( S \in [S_B, S_T] \), \( \nu_1(S) \) is continuously decreasing with respect to \( \tau \).

**Proof.** The proof for the monotonicity follows the arguments in the proof of Theorem 4.1 (b) and is omitted.

For the continuity we take \( S \in [S_B, S_T] \) and \( 0 < \tau_1 < \tau_2 \), and use again the functions \( v_i = p_c^+ - \nu_1 \), \( i \in \{1, 2\} \). From (4.41) and using the monotonicity of \( \nu_1 \) with respect to \( \tau \) one obtains

\[
\frac{1}{2}(v_2^2 - v_1^2)'(S) = -c(\tau_2 - \tau_1)g(S; S_B, S_T) + (v_2 - v_1)p_c^+(S) < -c(\tau_2 - \tau_1)g(S; S_B, S_T).
\]

With \( \tilde{K} \) defined above, integration gives

\[
0 < v_2^2(S) - v_1^2(S) < 2(\tau_2 - \tau_1)\tilde{K},
\]

which implies the continuity with respect to \( \tau \) of \( v_1 \) and consequently of \( \nu_1 \). 
\( \Box \)
From the discussion so far we conclude that the orbits \((S_t, u_t)\) are close to the graph of \(p_c^+\) for small values of \(t\), but move away from it as \(t\) increases, and for \(S \in (S_B, S_T)\). This situation is presented in Figure 4.9. In the remaining part of this subsection we focus on the behaviour of the system beyond the point \(S = S_T\). The main goal is to show that orbits connecting \(E_B\) and \(E_T\) exist for all values of \(t > 0\). In Theorem 4.3 we show this for small values of \(t\) and for larger \(t\) values the existence is shown in Theorems 4.4 and 4.5.

**Theorem 4.3.** Let \(\{ (S_t, u_t) \}_{t \geq 0}\) be the family of orbits of (4.38), (4.34b), originating from \(E_B\) and entering \(\mathcal{H}\). Then there exists a \(\tau_* > 0\) such that \(w_{\tau_*}(S_T) = 0\). For all \(\tau \in (0, \tau_*)\) the system (4.38), (4.34b) has a unique orbit \((S_t, u_t)\) satisfying \(S_T(0) = (S_B + S_T)/2\) and connecting \(E_B\) and \(E_T\).

**Proof.** The existence of a \(\tau_*\) for which \(w_{\tau_*}(S_T) = 0\) follows directly from Propositions 4.9 and 4.10. Also, \(w_T(S_T) < 0\) for \(\tau > \tau_*\) and \(w_T(S_T) > 0\) for \(\tau < \tau_*\).

To understand the behaviour of \((S_t, u_t)\) for \(\tau < \tau_*\) we recall Proposition 4.8, which states that the orbit either approaches \(E_T\) or enters \(\mathcal{H}_2\) through \(S = S_T\) at a finite \(\xi_T\). In the latter case, which is displayed in Figure 4.10, \(u_t\) becomes increasing for \(\xi > \xi_T\). With \(\tau < \tau_*\), since \(p_c^+(1) = 0 < u_T(\xi_T) < p_c^+(S_T)\) the orbit must intersect the graph of \(p_c^+\) at some \(\xi = \xi_P\) and enter \(\mathcal{H}_3\), where \(S_T\) becomes decreasing whereas \(u_T\) is still increasing. We claim that the orbit either approaches \(E_T\), or enters \(\mathcal{H}_4\) for some \(\xi = \xi_R\).

To see this, assume that a \(\delta > 0\) exists such that \(S_T(\xi) \geq S_T + \delta\) for all \(\xi > \xi_P\). As \(S_T\) is bounded and decreasing, the limit \(\lim_{\xi \to \infty} S_T(\xi)\) exists and is finite. Denoting it by \(\tilde{S}_T\) we have \(\tilde{S}_T \geq S_T + \delta\). Further, since \(u_T\) is only bounded from below, a similar reasoning shows that either \(\lim_{\xi \to \infty} u_T(\xi) = \tilde{u}_T \in [p_c^+(\tilde{S}_T), \infty)\) or \(u_T \to \infty\).

Since \(S_T\) is decreasing with \(\xi\) and bounded from below, \(\lim_{\xi \to \infty} S_T' = 0\). From (4.38) one gets \(\tilde{u}_T = p_c^+(\tilde{S}_T)\). Therefore \(u_T\) has a (finite) limit as \(\xi \to \infty\) and from

![Figure 4.9: The dependence of the orbits \((S_t, u_t)\) on \(t\) for \(S_B < S < S_T\).](image)
(4.34b) we get \( \lim_{\xi \to \infty} u'_\tau = 0 \). In other words, \((\hat{S}_\tau, \hat{u}_\tau)\) is an equilibrium point, which is not possible since \( k \) is a convex function and therefore \( \mathcal{W} \) has only two zeros. This rules out the possibility that \( S_\tau \) is bounded away from \( S_T \), so either \( \lim_{\xi \to \infty} S_\tau(\xi) = S_T \), or the orbit enters \( \mathcal{H}_4 \) at some finite argument \( \xi_R \).

In the former case it follows as before that the orbit ends up in \( E_T \). In the latter case we follow the arguments in Theorem 4.2 to prove that \((S_\tau, u_\tau)\) cannot end up back in \( E_B \), or leave \( H_4 \) through the line \( S = S_B \). This means that it enters \( \mathcal{H}_1 \) again at some \( \xi = \xi_Z \). However, in this case the incoming part of the orbit is above the part emerging from \( E_B \), and therefore the set bounded by \( \{(S_T(\xi), u_T(\xi)) : \xi < \xi_Z\} \) and the graph of \( p_c^+ \) from \( E_B \) to \((S_T(\xi_Z), u_T(\xi_Z))\) is positive invariant. With this, the proof continues as in Theorem 4.2.

Theorem 4.3 states that the orbits go to \( E_T \) for all \( \tau \in (0, \tau_*) \) but it does not state how the orbits behave close to \( E_T \). This is given in

**Proposition 4.11.** There exists a \( \tau_m > 0 \) such that for \( \tau \in (0, \tau_m] \) any orbit going to \( E_T \) goes either directly or after a finite number of turns around \( E_T \), and for \( \tau > \tau_m \) the orbit is a stable spiral around \( E_T \).

**Proof.** To prove this part we use the eigenvalues of the linearisation around \( E_T \), computed in (4.39). Let \( \tau_m = \frac{k(S_T)(p_c^+(S_T))^2}{4c(k(S_T)-c)} \). Note that \( E_T \) is a stable sink for \( 0 < \tau \leq \tau_m \) and a stable spiral for \( \tau > \tau_m \). This proves the statement of Proposition 4.11. □

Having explained the behaviour of orbits close to \( E_T \) we again turn to existence, this time for \( \tau > \tau_* \). As will be seen below, the orbits connecting \( E_B \) and \( E_T \) exist for \( \tau > \tau_* \) too, but to prove this we have to introduce

\[
\alpha(S_B, S_T) = \int_{S_B}^{1} \mathcal{W}(\rho; S_B, S_T) d\rho. \tag{4.44}
\]

![Figure 4.10: Behaviour of the orbit \((S_\tau, u_\tau)\) for \( \tau \leq \tau_* \).](image)
By the convexity of \( k \), as stated in Assumption (A1), and the definition \( \mathcal{G}(S;S_B,S_B) = 1 - k'(S_B)(S - S_B) + k(S_B) \), for any fixed \( S \in (S_B,1) \) the function \( \mathcal{G} \) is decreasing with respect to \( S_T \in (S_B,1) \). Also, one has \( \mathcal{G}(S;S_B,S_T) < 0 \) if \( S \in (S_B,S_T) \) and \( \mathcal{G}(S;S_B,S_T) > 0 \) if \( S \in (S_T,1) \). Moreover,

\[
\alpha(S_B,1) < 0 < \alpha(S_B,S_T), \tag{4.45}
\]

and \( \alpha(S_B,\cdot) \) is decreasing in \( [S_B,S_T] \). Observe that \( \alpha(S_B,S_T) \) does not depend on \( \tau \).

Figure 4.11 shows how the functions \( k(S) \), \( \mathcal{G}(S) \) and \( \int_{S_B}^{S} \mathcal{G}(\varrho;S_B,S_T)d\varrho \) vary with \( S \).

With this we can now state the following

**Theorem 4.4.** Let \( S_B,S_T \in (0,1) \), \( S_B < S_T \) and \( \alpha(S_B,S_T) \) be defined as above. If \( \alpha(S_B,S_T) \geq 0 \) then for all \( \tau > \tau^* \) the orbit \((S_\tau,u_\tau)\) reaches \( E_T \) as \( \xi \to \infty \).

**Proof.** Since \( \alpha(S_B,S_T) \geq 0 \), by the properties of \( \mathcal{G} \) an \( S_\alpha \in [S_T,1) \) exists such that

\[
\int_{S_\alpha}^{1} \mathcal{G}(S;S_B,S_T)dS = \alpha(S_B,S_T). \tag{4.46}
\]

Clearly, \( S_\alpha < 1 \) if \( \alpha(S_B,S_T) > 0 \) and \( S_\alpha = 1 \) if \( \alpha(S_B,S_T) = 0 \). Figure 4.12 (left) shows the location of \( S_\alpha \) as the point where the hashed areas, below and above the \( S \)-axis, are equal. We rewrite (4.40) as

\[
\frac{d}{dS} \left( p_\tau^+(S)w_\tau - \frac{1}{2}w_\tau^2 \right) = c\tau\mathcal{G}(S) + w_\tau dp_\tau^+/dS. \tag{4.47}
\]
4.4 Dynamic capillarity

Figure 4.12: (left) $S_\alpha$ is the saturation at which the hashed area above the $S$-axis equals the one below the $S$-axis. (right) The $(S_\tau, u_\tau)$ orbits for $\tau > \tau_*$ and $\alpha(S_B, S_T) > 0$. $S_1(\tau)$, $S_2(\tau)$ and $S_\alpha$ are shown in the image for this particular $S_T$ value.

Since $\tau > \tau_*$, $w_\tau(S_T) < 0$. Let $S_1(\tau) \in (S_B, S_T)$ be such that $w_\tau(S) > 0$ for all $S \in [S_B, S_1(\tau)]$, i.e. the first point where the orbit $(S_\tau, u_\tau)$ enters the region $u < 0$. Observe that, $w_\tau$ is increasing for $S > S_T$. Further, let $S_2(\tau) \in (S_T, 1]$ be such that $w_\tau(S_2(\tau)) = 0$ and $w_\tau(S) < 0$ for all $S \in (S_1(\tau), S_2(\tau))$. We prove that $S_2(\tau) < 1$, thus the orbit returns in the upper half plane (see also Figure 4.12 (right)). More precisely, since $\alpha(S_B, S_T) \geq 0$, we prove in Proposition 4.13 that $S_2(\tau) < S_\alpha$ for all $\tau > \tau_*$. Assume that $S_2(\tau) = 1$ for some $\tau > \tau_*$, then the domain of definition of $w_\tau$ can be extended to $[S_B, 1]$. Integrating (4.47) from $S_1(\tau)$ to 1 gives

$$-\frac{1}{2} w_\tau^2(1) = \frac{1}{cT} \int_{S_1}^{S_B} \mathcal{G} dS + \int_{S_1}^{1} w_\tau \frac{dp^+_C}{dS}. $$

Moreover, for $S \in (S_1(\tau), 1)$ one has $w_\tau(S) < 0$ and since $\mathcal{G}(S; S_B, S_T) < 0$ for $S \in (S_B, S_T)$ one has

$$\alpha = \int_{S_B}^{1} \mathcal{G}(S) dS = \int_{S_B}^{S_1} \mathcal{G} + \int_{S_1}^{1} \mathcal{G} = \int_{S_B}^{S_1} \mathcal{G} - \frac{1}{2cT} w_\tau^2(1) - \frac{1}{cT} \int_{S_1}^{1} w_\tau \frac{dp^+_C}{dS} < 0,$$

which contradicts the assumption $\alpha(S_B, S_T) \geq 0$. Therefore, if $\tau > \tau_*$, a $S_2(\tau) \leq 1$ exists such that $w_\tau(S_2(\tau)) = 0$, meaning that the orbit $(S_\tau, u_\tau)$ intersects the axis $u = 0$ for the second time. Following the reasoning in the proof of Theorem 4.3 one obtains that $(S_\tau, u_\tau)$ ends up in $E_T$. \[\square\]
The proof of Theorem 4.4 introduces three important values for the saturation, $S_\alpha$ given by (4.46), and $S_1(\tau)$, $S_2(\tau)$, the abscissas where the orbit intersects the axis $u = 0$. Below we give some results on the boundedness of $w_\tau$, $S_\alpha$ and $S_2(\tau)$. We start with

**Proposition 4.12.** Let $\tau > \tau_*$ be such that $S_2(\tau) \in (S_T, 1)$ exists. Then

$$w_\tau(S_T) > -K\sqrt{\tau},$$

where $K^2 = 2c \int_{S_T}^1 \vartheta(S) dS$

**Proof.** Equation (4.40) gives $(p_c^+(S) - w_\tau) w_\tau' = c \tau \vartheta(S)$. As $w_\tau'(S) > 0$ for $S \in (S_T, S_2(\tau)]$, this gives $-w_\tau^2(S) < 2c \tau \vartheta(S)$. The proof follows by integrating this inequality over $(S_T, S_2(\tau)]$.

Observe that the estimate in Proposition 4.12 gives a lower bound for $w_\tau$ since $w_\tau(S_T)$ is a minimum for $w_\tau$. Also, the result does not require that $\alpha(S_B, S_T) \geq 0$. The behaviour of $S_\alpha$ and $S_2(\tau)$ is stated in

**Proposition 4.13.** Under the assumptions of Theorem 4.4, one has $S_2(\tau) < S_\alpha$ and

$$\lim_{\tau \to \infty} S_2(\tau) = S_\alpha.$$

**Proof.** To estimate $S_2(\tau)$ we integrate (4.47) from $S_1(\tau)$ to $S_2(\tau)$ and obtain

$$c \tau \int_{S_1(\tau)}^{S_2(\tau)} \vartheta(S) dS + \int_{S_1(\tau)}^{S_2(\tau)} w_\tau \frac{dp_c^+}{dS} = 0.$$

Using this, one can split the integrals in (4.44) to obtain

$$\int_{S_2(\tau)}^{1} \vartheta = \alpha - \int_{S_B}^{S_1(\tau)} \vartheta + \frac{1}{c \tau} \int_{S_1(\tau)}^{S_2(\tau)} w_\tau \frac{dp_c^+}{dS}.$$

Denoting by $I_1(\tau)$ and $I_2(\tau)$ the two integrals on the right, since $\vartheta < 0$ for $S \in (S_B, S_T)$ and $w_\tau(S) < 0$ for $S \in (S_1(\tau), S_2(\tau))$ one gets $I_1(\tau) < 0$ and $I_2(\tau) > 0$. This gives $\int_{S_2(\tau)}^{S_1(\tau)} \vartheta > 0$. As $S_2(\tau) > S_T$, $\vartheta > 0$ for $S \in (S_2(\tau), 1)$ and therefore $S_2(\tau) < S_\alpha$ for all $\tau > \tau_*$.

To obtain the limit we start by proving that $S_1(\tau) \to S_B$ as $\tau \to \infty$. Clearly $S_1(\tau)$ decreases with increasing $\tau$ and remains bounded from below by $S_B$. Now suppose $S_1(\tau) \geq S_B + \delta$ for some $\delta > 0$ and for all $\tau > \tau_*$. Since $w_\tau(S) > 0$ and $\vartheta(S) < 0$ for $S \in (S_B, S_1(\tau))$, integrating (4.40) from $S_B$ to $S_1(\tau)$ gives

$$p_c^+(S_B) = c \tau \int_{S_B}^{S_1} \frac{-\vartheta(S)}{p_c^+(S) - w_\tau(S)} dS > -\frac{c \tau}{p_c^+(S_B)} \int_{S_B}^{S_B + \delta} \vartheta(S) dS.$$

This gives a contradiction for large $\tau$ as $c$ and $\vartheta$ do not depend on $\tau$. Hence $\lim_{\tau \to \infty} I_1(\tau) = 0$. 

To estimate $I_2$ we use Proposition 4.12 and the properties of $w_T$

\[ 0 < I_2(\tau) = \frac{1}{ct} \int_{S_1}^{S_2} u_T \frac{d P^+_c}{dS} < \frac{1}{c \sqrt{T}} P^+_c(S_B) K. \tag{4.49} \]

Hence \( \lim_{\tau \to \infty} \int_{S_2}^1 \mathcal{G}(S) dS = \alpha = \int_{S_\alpha}^1 \mathcal{G}(S) dS \). This proves that \( S_2 \to S_\alpha \) for \( \tau \to \infty \).

Having understood the behaviour of the orbits for the case \( \alpha(S_B, S_T) \geq 0 \) we proceed by analysing the case \( \alpha(S_B, S_T) < 0 \). In particular this situation occurs when \( S_T \) is close enough or equal to 1.

**Lemma 4.1.** Let \( S_B, S_T \in (0, 1) \), \( S_B < S_T \) and \( \alpha(S_B, S_T) \) introduced in (4.44). If \( \alpha(S_B, S_T) < 0 \) then a \( \tau^* \) exists such that for all \( \tau > \tau^* \), the orbit \((S_T, u_T)\) passes through a point \((1, w_T(1))\) with \( w_T(1) < 0 \).

**Proof.** We use ideas that are similar to the ones in the proof of Theorem 4.4. Assume that \( S_2(\tau) \leq 1 \) for all \( \tau > \tau^* \). Integrating (4.47) from \( S = S_B \) to \( S = S_2(\tau) \) gives

\[
-\frac{1}{2} p^+_c(S_B)^2 = cT \int_{S_B}^{S_2(\tau)} \mathcal{G}(S) + \int_{S_B}^{S_2(\tau)} \frac{d P^+_c}{dS} \frac{w_T}{\tau} < c\tau a + w_T(S_T)(p^+_c(S_2) - p^+_c(S_B))
\]

\[
< c\tau a - w_T(S_T)p^+_c(S_B) < c\tau a + p^+_c(S_B)K\sqrt{T}. \]

Since \( \alpha < 0 \) this gives a contradiction for \( \tau \) exceeding a \( \tau^* \geq \tau^* \), where \( \tau^* \) is determined such that the term on the right in the equation above becomes equal to \( -\frac{1}{2} p^+_c(S_B)^2 \). From this it follows that for \( \tau > \tau^* \) the orbit \((S_T, u_T)\) has no second intersection point with the \( u \)-axis before passing through the vertical line \( S = 1 \), therefore \( w_T(1) < 0 \).

From Lemma 4.1 we see that, if \( \alpha(S_B, S_T) < 0 \) and \( \tau \) is large enough, the orbit \((S_T, u_T)\) wants to exit the strip \([0,1] \times \mathbb{R}\) through the half-line \((1, u) : u < 0\). However, the functions \( p^+_c \) and \( k \) are only defined inside the physically relevant regime \( S \in [0,1] \) which makes the continuation of the orbits impossible and non-physical. Below we propose an extension of the model which allows continuation of the orbit within the physically relevant strip. It is based on the multi-valued extension of the \( p^+_c \) curve,

\[ P_e(S) = \begin{cases} p^+_c(S), & \text{for } 0 < S < 1, \\ (-\infty, 0] & \text{for } S = 1. \end{cases} \tag{4.50} \]

Such an approach is also used for defining extended pressure conditions in the case of porous media with block-type heterogeneities when models involving an entry pressure are adopted (see e.g. [40, 220]). With the extension the equations read

\[ S' \in \frac{P_e(S) - u}{ct}, \tag{4.51a} \]

\[ u' = \mathcal{G}(S; S_B, S_T). \tag{4.51b} \]
This formulation implies that if \( S = 1 \) in a set \( I \) of positive measure, then \( S' = 0 \) in \( I \) and \( u \in P_e(1) = (-\infty, 0) \) in \( I \). Moreover, from the \( u \)-equation,

\[
u' = \vartheta(1; S_B, S_T) \text{ in } I.
\]

When \( S_T = 1 \) we have \( \vartheta(1; S_B, S_T) = 0 \). Then any point on the half-line \( \{ S = 1 \} \times \{-\infty < u \leq 0 \} \) is an equilibrium point, and the compatibility condition \( u = p^+_e(S_T) \) should be interpreted as \( u_T \in P_e(1) = (-\infty, 0) \). We exploit this observation in the following construction.

\( S_T < 1 \)

From Lemma 4.1 if \( \tau > \tau^* \), the orbit \(( S_T, u_T )\) starting from \( E_B \) reaches \( S = 1 \) at finite \( \xi = \hat{\xi} \) where \( u_T(\hat{\xi}) = \hat{u} < 0 \). Then at \((1, \hat{u})\) we continue the orbit by the vertical upwards segment \( \{ 1 \} \times [\hat{u}, 0) \). Observe that along the segment, \((4.51)\) are still satisfied. Since now \( u' = \vartheta(1; S_B, S_T) = \) constant, we have \( u' = \Delta u = -\frac{\Delta \hat{u}}{\Delta \xi} = \vartheta(1; S_B, S_T) > 0 \), yielding the length of the \( \xi \) interval when \( S = 1 \).

Now taking \((1, 0)\) as the starting point of \((4.51)\) with \( \xi > \hat{\xi} + \Delta \xi \), we continue the construction as before. Again one uses the divergence argument from Theorem 4.2 to show that the orbit spirals into \( E_T \). In particular the orbit cannot reach \( S = 1 \) for a second time as this would lead to a limit cycle which is ruled out from previous arguments.

\( S_T = 1 \)

In this case the entire half-line \( \{ 1 \} \times (-\infty, 0] \) consists of equilibrium points. As before the orbit reaches \( S = 1 \) at finite \( \xi = \hat{\xi} \), with \( u = \hat{u} < 0 \). But it stays at this point for all \( \xi \geq \hat{\xi} \).

Summarising we have

**Theorem 4.5.** Let \( \tau^* \) be as in Lemma 4.1 and for any \( \tau > \tau^* \) let \(( S_T, u_T )\) be the orbit satisfying \((4.51)\) emerging from \( E_B \). Then

(a) For \( S_T < 1 \) the orbit reaches \( S = 1 \) at finite \( \xi = \hat{\xi} \) with \( u(\hat{\xi}) = \hat{u} < 0 \). It continues along the segment \( \{ S = 1 \} \times (-\infty < u \leq 0) \). At the point \((1, 0)\) it re-enters the set \( \{ S < 1 \} \times \mathbb{R} \) and connects to \( E_T \) as \( \xi \to -\infty \).

(b) For \( S_T = 1 \), again the orbit reaches \( S = 1 \) at finite \( \xi = \hat{\xi} \) with \( u(\hat{\xi}) = \hat{u} < 0 \). Since \((1, \hat{u})\) is an equilibrium, the orbit remains in this point for all \( \xi \geq \hat{\xi} \).

**Remark 4.3.** To avoid non-physical saturation regimes, we have considered a multi-valued extension of the \( p^+_e \)-\( S \) curve. Whenever \( S = 1 \), the specific value of \( P_e \) is taken such that \( P_e \) and \( u \) are in equilibrium yielding \( S' = 0 \). For analysing the orbits in this case, one can also consider a regularised approximation of \( P_e \). More precisely, with \( \delta > 0 \) being a small regularisation parameter, define

\[
p^*_e(\delta)(S) = \begin{cases} p^+_e(S) & \text{if } S < 1, \\ \frac{1}{\delta} & \text{if } S \geq 1. \end{cases} \quad \vartheta_e(S; S_B, S_T) = \begin{cases} \vartheta(S; S_B, S_T) & \text{if } S < 1, \\ \vartheta(1; S_B, S_T) & \text{if } S \geq 1. \end{cases}
\]

(4.53)
Letting now \((S^\delta_t, u^\delta_t)\) be the orbits satisfying
\[
\begin{align*}
S' &= \frac{1}{c^\tau} (P^\delta(S) - u), \\
u' &= \mathcal{G}(S; S_B, S_T),
\end{align*}
\]
and starting from \(E_B\), one can analyse the behaviour of these orbits when \(\delta \to 0\). In fact, this regularisation approach is being used for the numerical solutions presented in Section 4.5.

Having understood the above we can now distinguish the following situations which are shown in Figure 4.13. If \(\alpha > 0\) the orbits stay away from \(S = 1\) and approach \(E_T\) either directly or after spiraling (see Figure 4.13(a)). The situation is similar if \(\alpha < 0\) and \(\tau < \tau^*\). Whenever \(\alpha < 0\) and \(\tau > \tau^*\) then the orbit \((S^\tau_t, u^\tau_t)\) has a vertical section at \(S = 1\). The orbits \((S^\tau_t, u^\tau_t)\) for \(\alpha < 0\) are shown in Figure 4.13(b).

We conclude this subsection by comparing the case \(f \in L^1(0, 1)\), \(f \not\equiv 1\) to the case \(f \equiv 1\). This means that one has to repeat the previous arguments for the system (4.36), now in terms of \(Y\). With the functions introduced in (4.37), the direction of the orbits in the \((Y, u)\) plane remains unaffected since \(\bar{\mathcal{G}}(Y; Y_B, Y_T) < 0\) for \(Y_B < Y < Y_T\) and \(\bar{\mathcal{G}} \geq 0\) elsewhere in \((0, \infty)\) due to the convexity of \(k\). The eigenvalues at \((Y_B, u_B)\) and \((Y_T, u_T)\) behave similarly: now the critical \(\tau = \tilde{\tau}_m\) value at \((Y_T, u_T)\) becomes
\[
\tilde{\tau}_m = \frac{(p_c^+)'(Y_T)^2}{4c\tilde{\mathcal{G}}'(Y_T)} = \frac{(p_c^+)'(S_T)^2}{4c\mathcal{G}'(S_T)f(S_T)} = \frac{\tau_m}{f(S_T)}.
\]
Propositions 4.8 and 4.10 remain valid, with the redefinition \(\tilde{K} = -c \int_{Y_B}^{Y_T} \bar{\mathcal{G}}(Y) dY\). The existence Theorem 4.3 works also for this case, but now the divergence argument
uses the function \( F = \left( \frac{1}{\varepsilon T} (\tilde{\rho}_c^* (Y) - u), \tilde{\mathcal{G}}(Y) \right) \), for which \( \nabla \cdot F = \frac{1}{\varepsilon T} \left( \tilde{\rho}_c^* (Y) \right)' = \frac{\left( \rho_c^* (Y) \right)}{\varepsilon T f(S)} < 0 \). Also, the parameter \( \alpha \) given in (4.44), needs to be redefined as
\[
\tilde{\alpha}(Y_B, Y_T) = \int_{Y_B}^{Y(1)} \tilde{\mathcal{G}}(Y; Y_B, Y_T) dY = \int_{S_B}^{1} f(S) \tilde{\mathcal{G}}(S; S_B, S_T) dS. \tag{4.55}
\]
With this, the statement of Theorem 4.4 remains unchanged. \( Y_\alpha \), which corresponds to \( S_\alpha \) defined in (4.46), can now be defined as
\[
\int_{Y(1)}^{Y(1)} \tilde{\mathcal{G}}(Y; Y_B, Y_T) dY = \tilde{\alpha}(Y_B, Y_T). \tag{4.56}
\]
Consequently, the new \( S_\alpha \) satisfies \( Y_\alpha = \int_0^{S_\alpha} f(S) dS \), or \( \int_{S_\alpha} f(S) dS = 0 \). The constant \( K \) used in Propositions 4.12 and 4.13 becomes \( K = 2c \int_{Y(1)}^{Y(1)} \tilde{\mathcal{G}}(Y) dY \). Lemma 4.1 remains the same. Finally, for proving Theorem 4.5 we now use the extension
\[
\tilde{p}_e(Y) = \begin{cases} \tilde{p}_c^*(Y), & \text{for } 0 < Y < Y(1), \\ (-\infty,0] & \text{for } Y = Y(1). \end{cases} \tag{4.57}
\]
Following the arguments of Theorem 4.5 we get that \( Y \) reaches \( Y(1) \) for a finite \( \xi = \tilde{\xi} \) with \( u(\xi) = \tilde{u} < 0 \) and depending upon whether \( S_T < 1 \) or \( S_T = 1 \) either the orbit reaches \( (Y_T,u_T) \) or stays at \( (Y(1),\tilde{u}) \).

### 4.4.2 The case when \( f \notin L^1(0,1) \)

The TW analysis of the dynamic capillarity model up to now is restricted to the case when \( f \in L^1(0,1) \). This might not always be true. Since \( f \) is assumed continuous and positive on \([0,1]\), \( f \notin L^1(0,1) \) implies that it becomes unbounded at \( S = 1 \). As will be proved below, in this case \( S = 1 \) is an upper bound for the saturation and the orbits remain inside the physically relevant regime \( 0 \leq S \leq 1 \). This is like in the case \( \alpha(S_B,S_T) < 0 \) discussed before, but now extending the capillary pressure is not needed anymore.

Let \( \delta > 0 \) be arbitrarily small. Whenever \( S \leq 1 - \delta \), one can apply the transformation (4.35) to reduce the model (4.34) to the case analysed in Section 4.4.1 and most of the results there still remain valid. In particular, the orbits remain monotone if \( S \in (S_B,S_T) \). The main difference appears close to \( S = 1 \), whenever this value is approached. We have

**Theorem 4.6.** Assume \( f \notin L^1(0,1) \) and let \( \tau > 0 \), \( S_B \in (0,1) \), \( S_T \in (S_B,1] \) be given. For the orbits \((S_T,u_T)\) leaving \( E_B \) one has
(a) If \( S_T < 1 \), then \( S_T(\xi) < 1 \) for all \( \xi \in \mathbb{R} \).
(b) If \( S_T = 1 \), then two cases can occur.
(b.1) If \( f \not\in L^1(0,1) \) then as \( \xi \to \infty \), \( S_T \to 1 \) and \( u_T \to -\infty \).

(b.2) If \( f \in L^1(0,1) \) then there exists a \( u^* \in (-\infty, p_c^+(1)] \) such that \( \lim_{\xi \to \infty} (S_T, u_T) \to (1, u^*) \).

Proof. (a) Assume first that \( S_T < 1 \). Compared to the situation analysed in Theorem 4.3, the differences appear whenever \( S_T \) approaches 1. We therefore focus on part of the orbit satisfying \( S_T > S_T \). In this case, \( u'_T > 0 \) whereas \( S''_T > 0 \) as long as the orbit \( (S_T, u_T) \) stays below the \( p_c^+ \)-curve. Two situations are possible: the orbit either intersects the \( p_c^+ \)-curve for some argument \( \xi_3 \), or it reaches the line \( S = 1 \).

In the former situation, let \( S_{3,T} = S_T(\xi_3) \). We know that \( S_T(\xi) \leq S_{3,T} \) for all \( \xi \in \mathbb{R} \), so if \( S_{3,T} < 1 \) then the proof is completed. Assuming the contrary, namely that a \( \tau_0 > 0 \) exists such that \( S_{3,T0} = 1 \), one has \( u_{\tau_0}(\xi_3) \leq p_c^+(1) \) and (4.40) gives

\[
\frac{du_{\tau_0}}{dS} = \frac{\tau c f(S) \mathcal{G}(S)}{p_c^+(S) - u_{\tau_0}}.
\]

(4.58)

As \( p_c^+(S) \) and \( \mathcal{G}(S) > 0 \) for \( S \in [S_T,1) \) one uses (4.42) to see that \( p_c^+ - u_{\tau_0} \geq 0 \) decreases for \( S \in [S_T,1) \). Further, integration of (4.41) gives (with redefinition \( \bar{k} = -c \int_{S_B}^{S_T} f(\rho) \mathcal{G}(\rho) d\rho \))

\[
\sqrt{2\tau_0 \bar{k}} > p_c^+(S_T) - u_{\tau_0}(S_T) > \int_{S_T}^{1} f(S) \mathcal{G}(S) dS \geq \int_{S_T}^{1} cT_0 m_\mathcal{G} f(S) \mathcal{G}(S) dS
\]

with \( m_\mathcal{G} = \min\{\mathcal{G}(S), \frac{1}{2}(S_T + 1) \leq S \leq 1 \} \). Since \( m_\mathcal{G} > 0 \) and \( f \not\in L^1(S_B,1) \), the integral on the right is unbounded, which gives a contradiction.

The second case, when the orbit reaches the line \( S = 1 \), can be ruled out by similar arguments. We omit the details here.

(b) For \( S_T = 1 \), observe that \( S'_T(\xi) > 0 \) for \( (S_T, u_T) \in \mathcal{H}_1 \) and \( S_T \) is bounded above by 1 following the arguments used for proving Corollary 4.1. Consequently \( S_T \) has a limit \( S_\infty \) for \( \xi \to \infty \). Assume \( S_\infty < 1 \). We know that \( u'_T(\xi) \) decreases monotonically for \( \xi \in \mathbb{R} \) so that there are two possibilities. If \( \lim_{\xi \to \infty} u_T(\xi) = u_\infty > -\infty \) then from (4.34) it follows that \( S'_T \) and \( u'_T \) both have a limit as \( \xi \to \infty \). Moreover, since \( S_T \) and \( u_T \) have horizontal asymptotes, it means that \( \lim_{\xi \to \infty} S'_T(\xi) = \lim_{\xi \to \infty} u'_T(\xi) = 0 \). From (4.34b) we then get \( \mathcal{G}(S_\infty) = 0 \), contradicting \( S_\infty < 1 \). On the contrary, if \( \lim_{\xi \to \infty} u_T(\xi) = -\infty \) then from (4.34b) we get

\[
S'_T(\xi) = \frac{p_c^+(S) - u_T(\xi)}{cT f(S)} \geq \frac{p_c^+(S)}{cT f(S)} \geq \inf_{S \in [S_B, S_\infty]} \left\{ \frac{p_c^+(S)}{cT f(S)} \right\} > 0,
\]
for all $\xi > M_\xi$ with some large enough $M_\xi$. This means that $S_T$ cannot have a limit $S_\infty < 1$. Therefore the only possibility remaining is $\lim_{\xi \to \infty} S_T = S_\infty = 1$.

Now let us consider the case $f \not\in L^1(0,1)$. Observe that since $G < 0$ for $S \in (S_B,1)$ one has $\int_{S_B}^1 f(-G) = \infty$. If $w_T$ tends to $u^* > -\infty$ then integrating (4.58) from $S_B$ to 1 and multiplying by $-1$ we get

$$p_\xi^+(S_B) - u^* = \int_{S_B}^1 \frac{-\tau c f(S)G(S)}{p_\xi^+(S) - w_T(S)} dS > \frac{\tau c}{p_\xi^+(S_B) - u^*} \int_{S_B}^1 f(S)(-G(S))dS,$$

which is a contradiction since the term on the left is bounded whereas the integral on the right is not. Hence $\lim_{\xi \to \infty} w_T = -\infty$.

Next, for $f \not\in L^1(0,1)$ after redefining $\bar{K}$ as $\bar{K} = -c \int_{S_B}^1 G f$, Proposition 4.9 gives a lower bound for $w_T(S_T)$ that is uniform for all $S_B < S_T \leq 1$. Also observe that for a fixed $S_B$, $w_T(S_B;S_B,S_T)$ are well ordered with respect to $S_T$ meaning that for $S_B < S_{T,1} < S_{T,2} < 1$, $w_T(S_B;S_B,S_{T,1}) > w_T(S_B;S_B,S_{T,2})$ in their common domain of definition. To see why this holds observe that for $S \in (S_B,S_{T,1})$ and $u < p_\xi^+(S)$,

$$\frac{\tau c f(S)G(S;S_B,S_{T,1})}{p_\xi^+(S) - u} > \frac{\tau c f(S)G(S;S_B,S_{T,2})}{p_\xi^+(S) - u},$$

with $G(S;S_B,S_{T,1}) > G(S;S_B,S_{T,2})$ following from the convexity of $k$. Using (4.58) and proceeding as in the proof of Theorem 4.1 we conclude that the orbits are well-ordered in $S \in (S_B,S_{T,1})$ with respect to $S_T$. As $w_T(S_B;S_B,S_{T,1}) > w_T(S_B;S_B,S_{T,1})$ for $S > S_{T,1}$, the well ordering holds throughout the common domain of definition. In view of the boundedness of $w_T(S_T)$ mentioned before, $\lim_{S_T \to 1} w_T(S_T;S_B,S_T) = u^* > -\infty$. Finally proceeding like proof of Corollary 4.1 one proves that this value can be only attained as $\xi \to \infty$. \hfill \Box

From the proof above we see that in the case $S_T < 1$, for any $\tau > 0$ the orbit may turn around the equilibrium $E_T$ without reaching the line $S = 1$. In particular, an $S_{3,T} \in (S_T,1)$ exists such that the orbit intersects the graph of $p_\xi^+$ for the first time after $E_B$ in the point $(S_{3,T},p_\xi^+(S_{3,T}))$, see Figure 4.14. Moreover, since $f \in C(0,1)\backslash L^1(0,1)$ whereas $G \in C[0,1]$ with $G(1) > 0$ one has $\lim_{S_t \to 1} \int_{S_B}^S f(S)G(S)dS = \infty$. Since $G < 0$ on $(S_B,S_T)$, a unique $S_{a} \in (S_T,1)$ exists such that

$$\int_{S_B}^{S_a} f(S)G(S)dS = 0.$$

Observe that this simply extends the definition of $S_{a}$ in (4.44), given for the case $f \equiv 1$ to $f \in L^1(0,1)$ and $f \not\in L^1(0,1)$. Having introduced the above, as in the case $f \equiv 1$, it is interesting to see what happens if $\tau$ becomes very large. We have:

**Corollary 4.2.** Let $S_T < 1$ and $S_{\alpha}$, $S_{3,T}$ be as introduced above. Then $\lim_{\tau \to \infty} S_{3,T} = S_{\alpha}$. 
4.4 Dynamic capillarity

Proof. As in the proof of Proposition 4.10, the orbits are ordered with respect to \( \tau \). Therefore \( S_{3,\tau} \) is increasing with respect to \( \tau \) and bounded from above, \( S_{3,\tau} < 1 \). Hence there exists the limit \( \lim_{\tau \to \infty} S_{3,\tau} = S_3^* \). As in Theorem 4.4, for \( \tau > \tau_\ast \) and \( S > S_T \) let \( S_{2,\tau} \) be the abscissa where the orbit intersects the axis \( u = 0 \). Following the argument in Theorem 4.4, one proves that \( \lim_{\tau \to \infty} S_{2,\tau} = S_\alpha \). Also as \( S_{3,\tau} > S_{2,\tau} \) it is easy to see that \( S_3^* \geq S_\alpha \). Now integrating (4.58) from \( S_{2,\tau} \) to \( S_{3,\tau} \) gives

\[
p_c^+(S_T) > w_\tau(S_{3,\tau}) = \int_{S_{2,\tau}}^{S_{3,\tau}} \frac{\tau c f(S) \vartheta(S)}{p_c^+(S) - w_\tau(S)} dS \geq \frac{\tau c}{p_c^+(S_T)} \int_{S_{2,\tau}}^{S_{3,\tau}} f(S) \vartheta(S) dS.
\]

Observe that, if \( S_3^* > S_\alpha \), since \( \lim_{\tau \to \infty} S_{3,\tau} = S_3^* \) the integral on the right becomes positive for \( \tau \) large enough. On the other hand, since the term on the left is bounded, as \( \tau \to \infty \) this integral must approach 0. For \( S_3^* > S_\alpha \) this gives a contradiction, so the only possibility is that \( S_3^* = S_\alpha \). \( \square \)

Remark 4.4. For \( S_T < 1 \), Corollary 4.2 shows that for all \( \tau > 0 \) the orbits remain at the left of \( S = S_\alpha < 1 \). This means that the travelling waves exist without needing to extend the capillary pressure in the non-physical domain \( S > 1 \).

Remark 4.5. Observe that \( \tau \) and \( f \) have different effects. Specifically, changing \( \tau \) affects the orbit for all values of \( \zeta \), whereas \( f \) plays a major role only in the vicinity of \( S = 1 \).

![Figure 4.14: Typical \((S_\tau, u_\tau)\) orbit for the case \( f \notin L^1(0,1), S_T < 1 \).](image-url)
4.5 Numerical results

The numerical results presented in this section complement the theoretical findings in the previous sections. Specifically, after solving numerically the system (4.1), (4.4) we verify the predictions made in previous sections for sufficiently large times. In the numerical calculations we take a simple relative permeability function, \( k(S) = S^2 \). The other nonlinear functions are specified later.

4.5.1 Numerical scheme

We start by presenting the numerical scheme. Below \( \varepsilon > 0 \) and \( \tau \geq 0 \) are fixed. With \( S_B < S_T \) we consider the system given by (4.1), (4.4) for \( t > 0 \) and \( x \in (-\ell, \ell) \). The space interval is taken large enough to allow the saturation and pressure to develop profiles resembling the travelling wave profiles. In all cases, \( \ell \geq 50 \).

For the numerical solution we first discretize in time (4.1) and (4.4). Let \( \Delta t > 0 \) be the time step and let \( t_n = n\Delta t \) for \( n \in \mathbb{N} \). The time discrete unknowns \( S_n, u_n \) approximate the saturation and pressure at \( t_n \). We introduce the function \( \Theta \) which gives the discretization of \( \partial_t S \). One gets from (4.4) that

\[
\Theta(S, u) = \Phi_\varepsilon \left( \frac{p^*_c(S) - u}{p_c(S)} \right)
\]

for the hysteresis case and \( \Theta(S, u) = \frac{1}{\tau} (p^*_c(S) - u) \) for the dynamic capillarity case. With the \( \Theta \)-notation, the explicit discretization of (4.4) reads

\[
S_n = S_{n-1} + \Delta t \Theta(S_{n-1}, u_{n-1}).
\]

(4.59)

For stability we solve the time discrete version of (4.1) implicitly,

\[
\partial_x (k(S_n) \partial_x u_n) = -\Theta(S_n, u_n) - \partial_x k(S_n),
\]

(4.60)

together with the pressure boundary conditions at \( x = \pm \ell \),

\[
u_n(-\ell) = u_T, \quad u_n(\ell) = u_B. \]

(4.61)

For the spatial discretization we use standard finite differences.

Observe that for \( n = 0 \), \( u_0 \) is obtained from (4.60) by using the initial condition for saturation, \( S_0 \). The saturation initial condition is \( C^1 \) approximation of the Riemann data and it is consistent with the boundary conditions. Specifically \( S_0 : [-\ell, \ell] \rightarrow (0, 1) \) satisfies

\[
S_0(x) = S_T, \text{ if } x \leq -\ell_1, \quad \text{respectively } S_0(x) = S_B \text{ if } x \geq \ell_1.
\]

(4.62)

Here \( \ell_1 \ll \ell \) is a positive number, and \( S_T \) and \( S_B \) are compatible with the corresponding pressure values, i.e. \( u_j = p^*_c(S_j) \) (\( j \in \{B, T\} \)). Here \( \ell_1 = 5 \). Inside \( (-\ell_1, \ell_1) \) we take \( S_0(x) = \frac{(S_B + S_T)}{2} + \frac{(S_T - S_B)}{4\ell_1^3} x(x^2 - 3\ell_1^2) \). It is to be noted that the choice \( S_0 \) does
not have considerable impact on the end results as long as the necessary assumptions are satisfied.

As a validation of the numerical results, we compare the propagation speed of the numerical profile with the Rankine-Hugoniot speed given in (6.14). The profile speed is calculated as the speed of the point \( x_f(t) \) at which \( S(x_f(t), t) = \frac{1}{2}(S_B + S_T) \). By (4.14), \( \xi = 0 \) at this point, meaning that \( x_f(t) = ct \). Figure 4.15a presents the results for the hysteresis case, which agrees with the TW speed up to the discretization error. The results for the dynamic capillarity model are similar. We emphasise the fact that all results presented below are numerical approximations of the solutions of the original system of partial differential equations and do not assume any TW structure. Nevertheless, for \( t \) long enough the numerical solutions obtained for the specified initial and boundary conditions develop profiles resembling closely the TW solutions.

### 4.5.2 Iterative scheme

Observe that the equation (4.60) is nonlinear in \( u_n \) for the capillary hysteresis case and so an linear iterative scheme has to be used to solve for \( u_n \). For the \( \Phi_\varepsilon \) function defined in Proposition 4.2, the iterative schemes show poor convergence properties particularly because \( |\partial_u \Theta(S, u)| \) can become unbounded in this case in two different ways. Firstly, if \( u \rightarrow p^{(i)}(S) \) or \( u \rightarrow p^{(d)}(S) \), then \( |\partial_u \Theta(S, u)| \rightarrow \infty \). To resolve this we define \( \Phi_\varepsilon \) on \( \mathbb{R} \) in a way such that \( \Phi_\varepsilon'(r) = \frac{1}{\varepsilon} \) for \( |r| > 1 \). Note that this \( \Phi_\varepsilon \) is different from the function \( \Phi_\varepsilon \) given in Proposition 4.2, but satisfies Assumption (A4). This particular choice guarantees the numerical convergence of the nonlinear problem (4.60).

Secondly, \( |\partial_u \Theta(S, u)| \rightarrow \infty \) if \( S \rightarrow 0 \) or \( S \rightarrow 1 \). This problem is avoided by taking \( S_B > 0 \) and \( S_T < 1 \). So when studying the case \( S_T = 1 \), we actually show the result for the limit \( S_T \nearrow 1 \). With these modifications, \( \Theta \) becomes locally Lipschitz in both variables \( S \) and \( u \) for a fixed \( \varepsilon > 0 \).

Because \( \Phi_\varepsilon \) becomes unbounded as \( \varepsilon \rightarrow 0 \), iterative schemes like Newton’s method fail to converge because of the requirement of having good initial guesses. Therefore to solve (4.60) we use a linear iteration scheme inspired by the L-scheme discussed in [155, 192]. Specifically, for a sufficiently large \( L \) that will be specified later and with \( i \) as the iteration index, we solve the linear elliptic equation

\[
Lu^i_n - \partial_x \left[ k(S_n) \partial_x u^i_n \right] = Lu^{i-1}_n + \Theta(S_n, u^{i-1}_n) + \partial_x k(S_n). \tag{4.63}
\]

Following the arguments from [155, 192] one can show that if \( L \geq L_{\text{min}} > 0 \), the scheme in (4.63) becomes a contraction and converges irrespective of the initial guess. However, a natural choice is to start with \( u^0_n = u_{n-1} \). The lower bound \( L_{\text{min}} \) is the Lipschitz constant of \( \Theta \) with respect to the variable \( u \).
The iterative process is stopped if the $L^2$ norm of the difference between two iterates decreases below $10^{-10}$. Following [155, 192], the convergence is linear regardless of the mesh size. This can be seen in Figure 4.15, where the convergence of the iterative process is shown for the hysteresis model.

**Capillary Hysteresis**

We start by presenting the results for the capillary hysteresis case. The primary drainage and imbibition curves are taken such that

\[
p^+_c(S) = \left(1 - \frac{S}{S} \right), \quad \text{and} \quad p^-_c(S) = 2(1 - S)^2.
\]  (4.64)
4.5 Numerical results

With $b = \sqrt[3]{\epsilon}$ and $a = (1 - (\epsilon^2)^{\frac{2}{3}})$, the function $\Phi_\epsilon$ used in the numerical scheme is

$$\Phi_\epsilon(r) = \begin{cases} 
  b + \frac{1}{\epsilon}(r - 1) & \text{for } r > 1, \\
  \epsilon r (1 - a r^2)^{-1/2} & \text{for } r \in [-1, 1], \\
  -b + \frac{1}{\epsilon}(r + 1) & \text{for } r < -1.
\end{cases}$$

The case $S_T = 1$ was studied first. Observe that, in this case the model degenerates whenever $S$ approaches 1, where $p_c^\pm$ vanishes. To avoid this degeneracy, the calculations were performed for an $S_T$ slightly less than 1. This yields monotone profiles of $S$ and $u$, as shown in Figure 4.16, and is in good agreement with the TW profiles for $S_T = 1$. The right plot presents the pair $(S_\epsilon(x, t), u_\epsilon(x, t))$ in the $S$-$u$ plane, for a fixed $t$ and $x \in (-\ell, \ell)$. In analogy with the dynamical system analysis for the TW solutions, we call this an “orbit”. We use this term to refer to all similar plots that will be presented below.

![Figure 4.16: the profiles of $S_\epsilon$ and $u_\epsilon$ in the transformed coordinate $x - ct$ (left); (right) the orbit $(S_\epsilon, u_\epsilon)$ for the hysteresis model in the limit case $S_T \uparrow 1$. The figures are obtained for $S_T = .97$, $S_B = 0.2$, $\Delta x = .1$, $\Delta t = 10^{-3}$.](image)

Next we consider the $S_T < 1$ case. We fix $S_T = 0.6$ and $S_B = 0.2$ and vary $\epsilon$. Figure 4.17 shows the results for $\epsilon = 10^{-1}$. Observe that the orbit is monotone and $E_T$ is a stable sink. According to Theorem 4.2(b), $E_T$ becomes a spiral sink as $\epsilon$ becomes small enough. This is indeed the situation displayed in Figure 4.18, obtained for $\epsilon = 10^{-2}$. We clearly see that the (numerical) orbit spirals toward $E_T$. Consequently, for $t$ sufficiently large, the profiles of $u_\epsilon$ and $S_\epsilon$ are non-monotone.

The results for $\epsilon = 10^{-3}$ are similar, as shown in Figure 4.19. However, when compared to the case $\epsilon = 10^{-2}$ a longer time is required until the numerical solutions develop a profile resembling the travelling waves. To explain this, we observe that whenever $p_c^{(i)}(S) < u < p_c^{(d)}(S)$ one has $\frac{\partial_t S}{S} = H_\epsilon \left( \frac{p_c^{(i)}(S) - u}{p_c^{(d)}(S)} \right) \approx \epsilon H_1 \left( \frac{p_c^{(i)}(S) - u}{p_c^{(d)}(S)} \right)$. Therefore the time required for a profile to develop to a travelling wave profile scales with $\frac{1}{\epsilon}$. 
Also note that close to $S = S_B$, the numerical orbit for $\varepsilon = 10^{-3}$ has a steeper profile than the one for $\varepsilon = 10^{-2}$. This is in agreement with Corollary 4.1.

![Figure 4.17](image)

**Figure 4.17:** The orbit $(S_{\varepsilon}, u_{\varepsilon})$ for $S_B = 0.2$, $S_T = 0.6$, $\Delta x = .1$, $\Delta t = 10^{-3}$ and $\varepsilon = 10^{-1}$.

Another observation is that the oscillations in the $S$ and $u$ profiles are wider for $\varepsilon = 10^{-3}$ than for $\varepsilon = 10^{-2}$. This also follows from Remark 4.1, stating that the period of oscillation scales with $O(\varepsilon^{-1/2})$. Lastly one can see that the amplitude of oscillations in saturation for $\varepsilon = 10^{-3}$ is less than that of $\varepsilon = 10^{-2}$. This follows from Proposition 4.6. This can be seen in the $S$-$u$ phase plane as well: the $S$-range of the spirals decreases with $\varepsilon$. Therefore we conclude that the numerical results are in good agreement with the TW analysis for the hysteresis model.

![Figure 4.18](image)

**Figure 4.18:** (left) The profiles of $S_{\varepsilon}$ and $u_{\varepsilon}$ in the transformed coordinate $x - ct$; (right) the orbit $(S_{\varepsilon}, u_{\varepsilon})$ for $\varepsilon = 10^{-2}$. The other parameters are $S_B = 0.2$, $S_T = 0.6$, $\Delta x = .1$, $\Delta t = 10^{-3}$.
4.5 Numerical results

Dynamic Capillarity

The numerical results for the dynamic capillarity model are obtained for the quadratic function \( k \) and the \( p_c^+ \) function given in (4.64). Recall that in this case the two primary curves, drainage and imbibition, are equal. This means that \( p_c^- \) is vanishing.

We first take \( f(S) = 1 \). Figure 4.20 displays results in the case \( S_B = 0.2 \) and \( S_T = 0.5 \), when \( \alpha(S_B, S_T) > 0 \). By Proposition 4.11, a \( \tau_m \) exists such that for \( \tau < \tau_m \) the profiles of \( S \) and \( u \) are monotone, and for \( \tau > \tau_m \) they are non-monotone as \( E_T \) becomes a spiral sink. The value \( \tau_3 \) in Figure 4.20 is taken so that \( w_{\tau_3}(S_T) \approx 0 \). Also, a case with \( \tau \sim \tau_m \) is shown. In this case no oscillations are observed and the orbit goes directly to \( E_T \). These behaviours agree with the results given in Propositions 4.9 to 4.11.

Next we take the case \( S_B = 0.2 \) and \( S_T = 0.8 \), in which case \( \alpha(S_B, S_T) < 0 \). To avoid the unphysical saturation regimes we have considered extended \( P_{e,S} \) model given in (4.51). However, for the numerical solution this multi-valued extension is replaced by the regularised \( P_{e,S}^\delta \) curve given in (4.53) with \( \delta = 10^{-3} \). Figure 4.21 shows the profiles and orbits for \( \tau > \tau^* \). Observe that, due to the regularisation, \( S \) is still exceeding 1 and for \( S \geq 1 \) the orbit is not vertical but has a steep slope. As \( \delta \to 0 \) the possibility of having \( S > 1 \) is eliminated and the orbit goes vertically along \( S = 1 \). Moreover, in this case pressure remains continuously differentiable, but a kink can be observed at the transition from \( S = 1 \) to \( S < 1 \) which is as one expects from extension (4.54) given in Remark 4.3. Therefore we claim that the results are in good agreement with the theory.
Finally, we investigate the case when \( f \not\in L^1(0,1) \). We choose \( f(S) = \frac{1}{1-S} \) with \( S_B = 0.2, S_T = 0.8 \). The results are given in Figure 4.22. The profile takes considerably more time to develop and hence a kink is still visible in Figure 4.22 as a remnant of the initial condition. Compared to the case \( f \in L^1(0,1) \), we observe that the saturation stays below \( S = 1 \), but as \( \tau \) increases the saturation approaches \( S_\alpha \), which for Figure 4.22 is \( S_\alpha = 0.9903 \). This is in good agreement with the results in Theorem 4.6 and Corollary 4.2.

---

Figure 4.20: (left) The profiles of \( S_\tau \) and \( u_\tau \) in the transformed coordinate \( x-\tau t \); (right) the orbit \((S_\tau, u_\tau)\) for different \( \tau \) values. The \( \tau \) values used are \( \tau_1 = 1, \tau_2 = 10 \) and \( \tau_3 = 100 \). The other parameters are \( S_B = 0.2, S_T = 0.5, \Delta x = 0.1, \Delta t = 10^{-3} \). In this case \( \alpha(S_B, S_T) > 0 \).

Figure 4.21: (left) The profiles of \( S_\tau \) and \( u_\tau \) in the transformed coordinate \( x-\tau t \) with \( \tau = 100 > \tau^* \) and \( \alpha(S_B, S_T) < 0 \); (right) the orbit \((S_\tau, u_\tau)\) in the \((S, u)\) plane. Here \( S_B = 0.2, S_T = 0.8 \). The regularised extended model (4.54) given in Remark 4.3 has been used for this computation with a \( \delta \) value of \( 10^{-3} \) used in \( P_\delta \) (see (4.53)).
4.6 Conclusion

In this chapter we discussed the implications of including non-equilibrium effects in unsaturated porous flow models. Specifically, the play-type hysteresis and dynamic capillarity effects are considered in the saturation-pressure relationship. One focus was on analysing the occurrence of non-monotonic saturation or pressure profiles (overshoots) arising due to the non-equilibrium effects mentioned above. To this end, the traveling wave analysis is considered to understand the flow in a long, homogeneous vertical porous column.

The analysis is done first for hysteresis models. In this case, the existence of travelling wave solutions was shown first for the regularised case and then for the limiting case, leading to a play-type hysteresis model. It was proved that oscillations may appear in the regularised hysteresis models, which correspond to non-vertical scanning curves. However, in the limit situation these oscillations disappear and the saturation-pressure orbits lie on the imbibition curve.

Next we have investigated the dynamic capillarity effects, for which the existence of TW solutions is proved. Furthermore, the existence of a threshold value for the dynamic capillary parameter is shown so that for values less than this the travelling waves are monotonic, and become non-monotonic for values above the threshold. Moreover, similar thresholds are found for the dynamic capillary parameter that dictates whether the overshoot will have regions of positive pressure or whether the overshoot will reach a maximum corresponding to the full saturation. Also mechanisms to restrict the saturation to physically relevant values are analysed.

Finally, a semi-implicit numerical scheme to solve the nonlinear, pseudo parabolic equations corresponding to the non-equilibrium model was proposed. For solving the
emerging time discrete, nonlinear equations, the L-scheme was used. This scheme is used for solving the original partial differential equation in a large, but finite domain. For sufficiently large times the numerical solutions show a good resemblance with the travelling wave profiles predicted theoretically.
Chapter 5

Travelling waves: the combined case of capillary hysteresis and dynamic capillarity

5.1 Introduction

In the last chapter we considered wetting fronts in the form of travelling wave solutions of the one-dimensional Richards equation with the presence of either capillary hysteresis or dynamic capillarity. In this chapter we study how the results change when the combination of capillary hysteresis and dynamic capillarity is considered. Similar to previous chapters, Richards equation (1.20) is written as

\[ \partial_t S + \partial_x (k(S)(\partial_x u + 1)) = 0, \]

(5.1)

with the assumption that the one-dimensional infinite domain is homogeneous and points along gravity. In (5.1), \( S \in [0,1] \) is the water saturation, \( u := p_c = -p_w \) is the capillary pressure and the function \( k(S) \) denotes the relative permeability. The non-equilibrium effects, with play-type hysteresis, are represented by

\[ u \in p^+_c(S) - p^-_c(S) \cdot \text{sign}(\partial_t S) - \tau f(S) \partial_t S, \]

(5.2)

see (1.30). Here, \( p^+_c(S) \) and \( p^-_c(S) \) are calculated from the primary drainage \((p^{(d)}_c(S))\) and infiltration \((p^{(i)}_c(S))\) curves using relation (1.24). Further, \( \tau \) is the dynamic capillary coefficient and \( f(S) \) the dynamic capillary function. If hysteresis is absent then \( p^-_c(S) \equiv 0 \), whereas, if dynamic capillarity is absent then \( \tau = 0 \).

We assume, as before, 

(A2) \( k \in C^1([0,1]), k'(S) > 0 \) for \( 0 < S \leq 1 \), \( k(0) = 0 \), \( k(1) = 1 \) and \( k \) is strictly convex in \([0,1] \).

(A3) The capillary coefficient is positive, \( \tau > 0 \). The capillary function \( f \in C^1([0,1]) \) and \( f(S) > 0 \) for \( 0 < S < 1 \).

(A4) The capillary pressure functions \( p_c^{(k)} \) \((k \in \{i,d\})\) satisfy \( p_c^{(k)} : (0,1) \to [0,\infty) \), \( p_c^{(k)} \in C^1((0,1)), p_c^{(k)}(1) = 0 \), \( p_c^{(k)(0)}(S) < 0 \) and \( p_c^{(i)}(S) < p_c^{(d)}(S) \) for \( 0 < S < 1 \).

In (A2)-(A4), and elsewhere in this chapter, a prime denotes differentiation with respect to the argument.

As explained in the paragraph preceding (1.30), (5.2) can be rewritten as

\[
\partial_t S = \frac{1}{\tau f(S)} \mathcal{F}(S, p) = \frac{1}{\tau f(S)} \begin{cases} 
    p_c^{(d)}(S) - u & \text{when } u \geq p_c^{(d)}(S), \\
    0 & \text{when } u \in [p_c^{(i)}(S), p_c^{(d)}(S)], \\
    p_c^{(i)}(S) - u & \text{when } u \leq p_c^{(i)}(S),
\end{cases}
\tag{5.3}
\]

see also [24, 25]. Hence, the model based on (5.2) gives vertical scanning curves \((\partial_t S = 0)\) when \( p_c^{(i)}(S) < u < p_c^{(d)}(S) \).

In this chapter we consider, in addition to (5.2), the extended play-type hysteresis model introduced in Chapter 2. Combined with the dynamic capillary effect, it is expressed here as

\[
u \in p_c^{(d)}(S) - p_c^{(i)}(S) \cdot \text{sign}(\partial_t S + \tau \partial_t u) - \tau f(S)\partial_t S.
\tag{5.4}
\]

This corresponds to \( H(S) = \frac{S}{\tau} \) in Chapter 2 and \( b(u) = \epsilon u \) in Chapter 3. The advantages of this model are stated in Chapter 2. Similar models were studied in [24, 271].

In Chapter 2 it is shown that the model is consistent if \( \epsilon \) satisfies certain upper bounds. The specific bounds are stated explicitly in Section 5.3.

As in (5.3), \( \partial_t S \) can be made explicit from (5.4). The calculations are given in the Appendix. The result is

\[
\partial_t S = \mathcal{A}(S,u,\partial_t u) := \begin{cases} 
    \frac{p_c^{(d)}(S) - u}{\tau f(S)} & \text{when } u \geq p_c^{(d)}(S) + \epsilon \tau f(S)\partial_t u, \\
    -\epsilon \partial_t u & \text{when } p_c^{(i)}(S) + \epsilon \tau f(S)\partial_t u < u < p_c^{(d)}(S) + \epsilon \tau f(S)\partial_t u, \\
    \frac{p_c^{(i)}(S) - u}{\tau f(S)} & \text{when } u \leq p_c^{(i)}(S) + \epsilon \tau f(S)\partial_t u.
\end{cases}
\tag{5.5}
\]

Letting \( \epsilon \to 0 \) in this expression yields the play-type expression (5.3) and letting \( \tau \to 0 \) gives

\[
\partial_t (S + \epsilon u) = 0 \text{ when } p_c^{(i)}(S) < u < p_c^{(d)}(S),
\]

We are now in a position to describe the combined case of capillary hysteresis and dynamic capillarity.
5.1 Introduction

i.e. scanning curves in the form of straight lines with inclination $-\frac{1}{\varepsilon}$ [80].

Using (5.3) to eliminate $\partial_t S$ from (5.1), yields an elliptic equation for $u$. Likewise, using (5.5) in (5.1) gives an elliptic-parabolic equation for which an initial pressure is required. This is natural for extended play-type problems, as was shown in Chapter 2 where we studied the case of horizontal redistribution.

Assumption (A3) includes the possibility of the capillary function $f(S)$ having a singular behaviour at $S = 1$. The behaviour of $f(S)$, as observed from experiments [47], suggests that this should indeed be taken into consideration. Therefore we distinguish, as in Chapter 4, between the classes

A. $f \in L^1(0,1)$

B. $f \not\in L^1(0,1)$

We show in Section 5.2 that $f \in L^1(0,1)$ is in essence equivalent to the case where $f(S) = \text{constant} = 1$.

It is known that the dynamic term in (5.2) may cause overshoot or oscillations in solutions of (5.1), (5.2). This has been observed both analytically (e.g. [68, 141, 175, 251]) and numerically (e.g. [209, 214, 271]). Moreover, we saw from the previous chapters that, the stronger the dynamic term $\tau$, the larger the overshoot/oscillations. To prevent the saturation from exceeding its physical maximal value $S = 1$, the capillary pressure was extended as a graph in Chapter 4, [251]. For the hysteretic case, this means that both $p_c^{(i)}$ and $p_c^{(d)}$ are extended to be set-valued at $S = 1$:

$$p_c^{(k)}(S) = \begin{cases} p_c^{(k)}(S) \text{ (as in (A4))} & \text{when } 0 < S < 1, \\ (-\infty, 0] & \text{when } S = 1, \end{cases} \quad (5.6)$$

for $k \in \{i, d\}$. Definition (5.6) allows us to deal with large values of $\tau$, causing the porous medium to be saturated ($S = 1$) in parts of the domain. The set-valued pressure at $S = 1$ acts as a natural saturation cut-off.

This chapter is organized as follows. In Section 5.2 travelling waves describing wetting fronts are introduced for the model based on (5.1), (5.2) and (5.6). The main emphasis, Section 5.2.1, is on the case $f(S) = 1$. Special attention is given to the role of the parameter $\tau$ in relation to extension (5.6). Generalisations to non-constant $f$ are given in Section 5.2.2 ($f$ from class A) and in Section 5.2.3 ($f$ from class B). In Section 5.3 travelling waves based on the extended expression (5.4) are treated, making the role of the hysteretic parameter $\varepsilon$ explicit. Section 5.4 shows computational results for the models (5.2) and (5.4). We present the conclusions in Section 5.5.
5.2 Travelling waves with combined pressure (5.2)

As in Chapter 4, we seek travelling wave solutions of the form
\[ S(x, t) = S(\xi), \quad u(x, t) = u(\xi) \text{ and } \xi = ct - x, \]  
where \(-\infty < \xi < \infty\) and
\[ \lim_{\xi \to -\infty} S(\xi) = S_B, \quad \lim_{\xi \to \infty} S(\xi) = S_T. \]  
Here \(S_B\) corresponds to the ‘initial’ saturation and \(S_T\) to the ‘injected’ saturation. In the physical context of wetting fronts we choose
\[ 0 < S_B < S_T \leq 1. \]  
Due to the convexity of \(k(S)\), this choice is necessary for mathematical reasons as well [248, 251].

The boundary conditions for the pressure need special attention. For the moment we impose
\[ u'(\pm\infty) = 0. \]  
Substituting (5.7) into (5.1) and (5.3) and applying conditions (5.8) and (5.10), details are given in Chapter 4, one finds for the wave speed \(c\) the Rankine-Hugoniot expression
\[ c = \frac{k(S_T) - k(S_B)}{S_T - S_B}, \]  
and for \(S\) and \(u\) the first order equations
\[ S' = \frac{1}{ctf(S)} \mathcal{F}(S, u), \]  
\[ u' = \mathcal{G}(S; S_B, S_T). \]  
Here, \(\mathcal{F}(S, p)\) is as given in (1.30) and (5.3), whereas,
\[ \mathcal{G}(S; S_B, S_T) = 1 - \frac{k(S_B) + c(S - S_B)}{k(S)}, \]  
as in Chapter 4. These functions are sketched in Figure 5.1. We are interested in solutions of (5.12) belonging to the strip \(\{0 < S \leq 1, u \in \mathbb{R}\}\) signifying the physically relevant regime. In this strip we introduce, for future reference, the sets
\[ \mathcal{H}^{(d)} = \{u > p^{(d)}_c(S)\}, \]  
\[ \mathcal{H} = \{p^{(i)}_c(S) < u < p^{(d)}_c(S)\}, \]  
\[ \mathcal{H}^{(i)} = \{u < p^{(i)}_c(S)\}, \]
5.2 Travelling waves with combined pressure (5.2)

\[ \mathcal{F}(S, u) \]

\[ \mathcal{G}(S; S_B, S_T) \]

Figure 5.1: (left) Graph of \( \mathcal{F}(S, u) \) for given \( S \in (0, 1) \), (right) Graph of \( \mathcal{G}(S; S_B, S_T) \) for given \( 0 < S_B < S_T < 1 \), with \( k(S) = S^2 \). Here \( S_m \) is the unique saturation value for which \( \int_{S_B}^{S_m} q(S; S_B, S_T) dS = 0 \) and \( S_l \in (S_B, S_T) \) is such that \( \mathcal{G}(S_l) = \min_{S \in (0,1)} \mathcal{G}(S; S_B, S_T) \).

and the points \( (k \in \{i, d\}) \)

\[ E_{B,k} = (S_B, p_c^{(k)}(S_B)), \quad E_{T,k} = (S_T, p_c^{(k)}(S_T)). \]  

(5.15)

They are indicated in Figure 5.2, together with the direction of the solution orbits of (5.12).

The right hand side of (5.12) vanishes iff

\[ S \in \{S_B, S_T\} \text{ and } (S, u) \in \mathcal{H}. \]  

(5.16)

Hence, all points of the closed segments \( E_{B,i}E_{B,d} \) and \( E_{T,i}E_{T,d} \) are equilibrium points of (5.12). Since all the trajectories are vertical in \( \mathcal{H} \) and all points on the vertical

Figure 5.2: The sets (5.14) and points (5.15) in the strip \( 0 < S \leq 1 \). The arrows indicate the direction of the solution orbits (5.12).
lines $E_{B,i}E_{B,d}$ and $E_{T,i}E_{T,d}$ are equilibrium points, no interior points can serve as a start and/or end point of any trajectory. Only the end points $E_{B,k}$ and $E_{T,k}$, $k \in \{i, d\}$, can play that role. We investigate the nature of these points later in this section.

With the capillary function $f$ satisfying (A3), the function

$$Y(S) = \int_0^S f(\varrho) d\varrho$$

is well defined for $0 \leq S < 1$ and $Y(1) < \infty$ iff $f \in L^1(0, 1)$. Since $f > 0$ in $(0, 1)$, the inverse $S = S(Y)$ is given by

$$S = S(Y) \text{ for } 0 \leq Y \leq Y(1).$$

Rewriting (5.12) in terms of $Y$ and $u$ gives

$$Y' = \frac{1}{c\tau} \tilde{F}(Y, u),$$

$$u' = \tilde{G}(Y; Y_B, Y_T),$$

(5.19)

where $\tilde{F}(Y, u) = F(S(Y), u)$ and $\tilde{G}(Y; Y_B, Y_T) = G(S(Y); S_B, S_T)$ with $Y_{B,T} = \int_0^{S_{B,T}} f(\varrho) d\varrho$. Thus the transformed system only involves a stretching of the $S$-coordinate.

If $f \in L^1(0, 1)$, solutions belong to the strip $\{0 < Y \leq Y(1)\}$ and the transformed system (5.19) is qualitatively similar to (5.12) with $f(S) = 1$. Therefore the results for (5.12) with $f(S) = 1$ can be translated directly to the case $f(S) \neq 1$ and $f \in L^1(0, 1)$. If $f \not\in L^1(0, 1)$, solutions belong to the half space $\{Y > 0\}$. This situation needs separate attention.

Remark 5.1. Since (5.12) is autonomous, its solutions are unique modulo a shift in $\xi$. One usually fixes this by setting $S(0) = \frac{1}{2} (S_B + S_T)$.

5.2.1 The case $f(S) = 1$

Since $F(S, p)$ is not smooth across $p_c^{(i)}$ and $p_c^{(d)}$, linearising (5.12) when approaching $E_{B,k}$ from $\mathcal{H}^{(k)}$, $k \in \{i, d\}$, gives different results compared to when approached from $\mathcal{H}$. When $E_{B,i}$ is approached from $\mathcal{H}^{(i)}$ and $E_{B,d}$ from $\mathcal{H}^{(d)}$, the linearisation gives a positive and a negative eigenvalue at these points. Hence $E_{B,i}$ and $E_{B,d}$ are saddles of which the direction of the eigenvectors is sketched in Figure 5.3.

Similarly at the points $E_{T,i}$ and $E_{T,d}$ one finds for the eigenvalues (see also references [68, 175, 214, 251])

$$\lambda_k = \frac{p_c^{(i)}(S_T)}{2c\tau} \left[ 1 \pm \sqrt{1 - \frac{\tau}{\tau_{c,k}}} \right]$$

(5.20)
where
\[ \tau_{c,k} = \frac{(P_k'(S_T))^2}{4c\varphi'(S_T;S_B,S_T)} > 0 \quad (k = i, d). \] (5.21)

Without loss of generality we may assume
\[ p_c^{(d)'}(S_T) < p_c^{(i)'}(S_T) < 0 \]
implying
\[ \tau_{c,i} < \tau_{c,d}. \] (5.22)

Then we have
- \( \tau < \tau_{c,i} \Rightarrow E_{T,i} \) and \( E_{T,d} \) are sinks;
- \( \tau_{c,i} < \tau < \tau_{c,d} \Rightarrow E_{T,i} \) is stable spiral and \( E_{T,d} \) is sink;
- \( \tau > \tau_{c,d} \Rightarrow E_{T,i} \) and \( E_{T,d} \) are stable spirals.

As a consequence, if \( \tau > \tau_{c,i} \), no orbit can reach \( E_{T,i} \) along a fixed direction. This can only happen when \( \tau < \tau_{c,i} \). The same holds at \( E_{T,d} \) for \( \tau > \tau_{c,d} \).

For a given \( \tau > 0 \), the goal is to construct an orbit that runs from \( E_{B,i}E_{B,d} \) as \( \xi \to -\infty \) to \( E_{T,i}E_{T,d} \) as \( \xi \to +\infty \). This ensures that (5.8) and (5.10) are satisfied. Inspection of the flow direction in Figure 5.2 gives that this can only be achieved by considering the orbit leaving \( E_{B,i} \) in the direction of \( \mathcal{H}^{(i)} \) for increasing \( S \). As long as the orbit is in \( \mathcal{H}^{(i)} \), we can use the results of Chapter 4 to describe its behaviour. Here we stay below \( p_c^{(i)} \), whereas in Chapter 4 we stay below \( p_c = p_c^{(i)} = p_c^{(d)} \). For small \( \tau \), the behaviour of the orbit leaving \( E_{B,i} \) is straightforward as it connects to \( E_{T,i} \) when \( \xi \to \infty \). This follows from the following proposition.

![Figure 5.3: Eigenvectors of the saddles \( E_{B,i} \) and \( E_{B,d} \). The arrows indicate the direction of the corresponding trajectories for increasing \( \xi \).](image-url)
There exists a $\tau_m \in (0, \tau_{c,i})$ such that for all $0 < \tau \leq \tau_m$ the orbit satisfying (5.12) and leaving $E_{B,i}$ moves monotonically through $\mathcal{H}^{(i)}$, implying that $S$ increases and $u$ decreases with $\xi$. Moreover, $(S, u) \rightarrow E_{T,i}$ as $\xi \rightarrow \infty$.

Proof. The orbit $(S, u)$ leaving $E_{B,i}$ and moving through $\mathcal{H}^{(i)} \cap \{S_B < S < S_T\}$ is monotone since $S' > 0$ and $u' < 0$ in this region. This follows from (5.12) and the characteristics of the functions $F$ and $G$. Next, we introduce the function $\mathcal{L}_r(S) = p_c^{(i)}(S) + r(S - S_T)$ for a constant $r > 0$. Our claim is that there exists a $\tau_m \in (0, \tau_{c,i})$ and a $r > 0$ such that

$$\mathcal{L}_r(S(\xi)) < u(\xi) < p_c^{(i)}(S(\xi)) \quad \text{for} \quad 0 < \tau \leq \tau_m, \quad r > r \quad \text{and} \quad \xi \in \mathbb{R}. \quad (5.23)$$

Since $u = \mathcal{L}_r(S)$ and $u = p_c^{(i)}(S)$ intersect only at the equilibrium point $E_{T,i}$ and the orbit $(S, u)$ is monotone, the statement of the proposition follows.

To prove (5.23) assume that $(S, u)$ intersects $\mathcal{L}_r(S)$ for some $r > 0$. Let the first intersection be at $\xi = \xi^*$ with $S = S(\xi^*) < S_T$ and $u = u(\xi^*)$, see Figure 5.4. This implies that $u = p_c^{(i)}(S) + r(S - S_T)$ and

$$\frac{u'(\xi)}{S'(\xi)} \leq \mathcal{L}_r'(S) \quad \text{or} \quad \frac{c \tau \mathcal{G}(S; S_B, S_T)}{p_c^{(i)}(S) - u} = \frac{c \tau \mathcal{G}(S; S_B, S_T)}{r(S_T - S)} \leq p_c^{(i)}(S) + r. \quad (5.24)$$

Since $\mathcal{G}(S_T; S_B, S_T) = 0$ we have $\frac{\mathcal{G}(S; S_B, S_T)}{S - S_T} = \mathcal{G}'(S; S_B, S_T)$ for some $\xi \in (S_B, S_T)$. Defining

$$p := \min_{S \in (S_B, S_T)} [-p_c^{(i)}(S)] > 0, \quad m := \max_{S \in (S_B, S_T)} \{\mathcal{G}'(S; S_B, S_T)\} > 0 \quad \text{and} \quad \tau_m := \frac{p^2}{4cm},$$

we get from (5.24) that the intersection occurs if

$$r^2 - pr + c \tau m \geq 0 \quad \text{which is satisfied for} \quad r \not\in \left[\frac{p}{2} \left(1 - \sqrt{1 - \frac{r}{\tau_m}}\right), \frac{p}{2} \left(1 + \sqrt{1 - \frac{r}{\tau_m}}\right)\right].$$
Observe from (5.21) that $\tau_m \leq \tau c, i$. Choosing $r^- = \frac{p}{2} \left[ 1 - \sqrt{1 - \frac{\tau}{\tau_m}} \right]$ and $\tau \leq \tau_m$, one concludes that the orbit $(S, u)$ can not intersect $\mathcal{L}_r(S)$ for $r > r^-$, hence, proving (5.23) and consequently, the proposition.

However, when $\tau$ is larger, the behaviour of the orbit depends crucially on the value of $S_T$ with respect to $S_B$. To make this precise, we introduce as in Chapter 4 the expression

$$\alpha(S_B, S_T) = \int_{S_B}^{1} \varrho(S; S_B, S_T) dS,$$

where $\varrho$ is given by (5.13).

Clearly, $\alpha(S_B, S_T = S_B) > 0$, $\alpha(S_B, S_T = 1) < 0$ and $\alpha(S_B, \cdot)$ strictly decreases in $(S_B, 1)$; see Figure 5.5 for an impression. Hence for each $0 < S_B < 1$, there exists a unique $S_T = S_T^*(S_B)$ such that

$$\alpha(S_B, S_T^*(S_B)) = 0$$

and

$$\begin{cases} 
\alpha(S_B, S_T) > 0 & \text{if } S_T < S_T^*(S_B), \\
\alpha(S_B, S_T) < 0 & \text{if } S_T > S_T^*(S_B).
\end{cases}$$

Further, for $S_T < S_T^*(S_B)$, we define the saturation $S_m \in [S_T, 1]$ by

$$\int_{S_m}^{1} \varrho(S; S_B, S_T) dS = \alpha(S_B, S_T)$$

or, see Figure 5.1 (right),

$$\int_{S_B}^{S_m} \varrho(S; S_B, S_T) dS = 0.$$  

Note that if $S_T \not\sim S_T^*(S_B)$, then $S_m \not\sim 1$.

The behaviour of the orbits, as demonstrated in Chapter 4, is summarized in the following theorem.
Theorem 5.1. Let $0 < S_B < 1$ be fixed and let $S_B < S_T \leq 1$. Then

(i) If $\tau > \tau_{c,i}$, the orbit leaving $E_{B,i}$ cannot enter $E_{T,i}$ in a monotone way through $\mathcal{H}_i$;

(ii) Let $S_T < S_T^*(S_B)$ (or $\alpha(S_B, S_T) > 0$). Then for any $\tau > 0$, the orbit leaving $E_{B,i}$ remains to the left of the line $S = S_m$;

(iii) Let $S_T > S_T^*(S_B)$ (or $\alpha(S_B, S_T) < 0$). Then there exist a $\tau^* > 0$ such that for $\tau < \tau^*$ the orbit leaving $E_{B,i}$ remains to the left of $S = 1$. If $\tau > \tau^*$, the orbit intersects the line $S = 1$ at a negative pressure.

Remark 5.2. (i) The dynamical system (5.12) has three characteristic values for $\tau$ where the behaviour of its solutions changes: $\tau_{c,k}$ for $k \in \{i,d\}$ defined by (5.21) and $\tau^*$ from Theorem 5.1 (iii). By assumption $\tau_{c,i} < \tau_{c,d}$. It is not clear how or if $\tau^*$ is ordered with respect to $\tau_{c,i}$ and $\tau_{c,d}$. Therefore we include in the discussion the possibility $\tau^* < \tau_{c,i}$ as well as $\tau^* \in (\tau_{c,i},\tau_{c,d})$ and $\tau^* > \tau_{c,d}$.

(ii) If $\tau < \tau_m$, the orbit moves in a monotone way through $\mathcal{H}_i$ from the saddle $E_{B,i}$ to the sink $E_{T,i}$. It is possible for this behaviour to continue up to $\tau = \tau_{c,i}$. However, for $\tau > \tau_{c,i}$, the nature of the point $E_{T,i}$ allows us only to make statement (i) of Theorem 5.1.

(iii) No matter how large $\tau$, the orbit stays to the left of $S = S_m$ for $S_T < S_T^*(S_B)$.

(iv) For $S_T > S_T^*(S_B)$ and $\tau$ sufficiently large, the orbit intersects $S = 1$ yielding a solution with a plateau at $S = 1$ where $u' = \mathcal{G}(1; S_B, S_T)$. This solution is allowed by the extended capillary pressure (5.4).

Figure 5.6: (left) Behaviour for $\tau < \tau_m < \tau_{c,i}$; (center) Here $S_T < S_T^*(S_B)$ and orbit remains to the left of $S = S_m$. (right) Now $S_T > S_T^*(S_B)$ and $\tau < \tau^*$ in the top orbit and $\tau > \tau^*$ in the bottom orbit.

The results described by Theorem 5.1 and Remark 5.2 are shown in Figure 5.6. The orbit in Figure 5.6 (center) and the top orbit in Figure 5.6 (right) exit the region.
\( \mathcal{H}^{(i)} \) at the points \( A_i = (S_A, p_c^{(i)}(S_A)) \). The bottom orbit of Figure 5.6 (right) exits at the point \( B = (S = 1, u < 0) \).

Since \( \mathcal{F} = 0 \) in \( \mathcal{H} \), the orbits at \( A_i \) move vertically upwards until they reach the drainage curve \( p_c^{(d)}(S_A) \) at \( A_d = (S_A, p_c^{(d)}(S_A)) \). Using again results from Chapter 4, we are left with four possibilities that are shown in Figure 5.7.

![Figure 5.7](image_url)

Figure 5.7: Four possibilities for orbits starting at the point \( A_d \): (a) Orbit moves directly and monotonically to \( E_T, d \) as \( \xi \to \infty \); (b) orbit passes \( S = S_T \) and returns to the hysteretic region \( \mathcal{H} \); (c) orbit passes \( S = S_T \) and ends up at \( E_B, d \) as \( \xi \to \infty \); (d) Orbit passes first \( S = S_T \) and then \( S = S_B \) above \( E_B, d \).

Cases (c) and (d) are ruled out by the following divergence argument (see also [68, 251]). Write system (5.12) as

\[
\ddot{u} = \tilde{R}(\tilde{u}),
\]
where
\[
\vec{u} = \begin{pmatrix} S \\ u \end{pmatrix} \quad \text{and} \quad \vec{R}(\vec{u}) = \begin{pmatrix} \frac{1}{c_T} \mathcal{F}(S, u) \\ \mathcal{G}(S, S_B, S_T) \end{pmatrix},
\]
and observe that
\[
\text{div} \vec{R} = \frac{1}{c_T} \frac{\partial \mathcal{F}}{\partial S} \leq 0,
\]
as
\[
\frac{\partial \mathcal{F}}{\partial S} = \begin{cases} p^i_c(S) & \text{when } u \in \mathcal{H}_i, \\ 0 & \text{when } u \in \mathcal{H}, \\ p^d_c(S) & \text{when } u \in \mathcal{H}_d. \end{cases}
\]

Let \( \Omega \) in Figure 5.7(c) denote the region enclosed by the orbit running from \( E_{B,i} \) to \( E_{B,d} \) and the vertical segment at \( S = S_B \). Then the divergence theorem applied to \( \vec{R} \) in \( \Omega \) gives a contradiction. The same argument applies to case (d) (see Chapter 4). Hence we are left with cases (a) and (b) in which the orbit is captured in a region around the segment \( E_{T,i}E_{T,d} \). In case (a) the solution moves monotonically from \( A_d \) to the equilibrium point \( E_{T,d} \). Clearly this is only possible when \( \tau < \tau_{c,d} \) (and \( E_{T,d} \) is a sink). In case (b) the orbit goes at least once around \( E_{T,i}E_{T,d} \) and it may reach \( E_{T,i} \) (if \( \tau < \tau_{c,i} \)) or \( E_{T,d} \) (if \( \tau < \tau_{c,d} \)) after finitely many rotations. If \( \tau > \tau_{c,d} > \tau_{c,i} \), the nature of the points \( E_{T,k} \), \( (k \in \{i, d\}) \) prevents this behaviour.

When the orbit exits at point \( B \) as in Figure 5.6 (right), it moves vertically upwards with \( S = 1 \) and \( u' = \mathcal{G}(1; S_B, S_T) > 0 \) until it reaches \( S = 1, u = 0 \). Then we repeat the previous argument.

With these observations in mind, we now investigate the behaviour of the orbits near \( S = S_T \). We first consider

\[ \tau > \tau_{c,d} (\geq \tau_{c,i}) \]

In this case the endpoints \( E_{T,i} \) and \( E_{T,d} \) are stable spirals, implying that the behaviour, as sketched in Figure 5.7(a), cannot occur. Hence we are left with Figure 5.7(b) as the only option. Since an orbit cannot intersect itself, it will be captured in a neighborhood of the segment \( E_{T,i}E_{T,d} \). Moreover, since orbits move vertically through \( \mathcal{H} \), they cannot reach an interior point of \( E_{T,i}E_{T,d} \) and the corresponding saturation has plateaus below and above \( S_T \). Clearly there are infinitely many of these plateaus, otherwise the orbits would end up in \( E_{T,i} \) or \( E_{T,d} \) after finitely many rotations.

We are now in a position to prove
Lemma 5.1. Let $0 < S_B < S_T < 1$ and $\tau > \tau_{c,d}$. Further, for a given orbit satisfying (5.12), let $\{S_{n+1}^\pm\}_{n=1}^\infty$ denote the saturation plateaus above ($\pm$) and below ($\mp$) $S_T$. Then

(i) $S_n^- < S_{n+1}^- < S_T$ for $n \geq 1$ and $\lim_{n \to \infty} S_n^- / S_T$;

(ii) $S_n^+ > S_{n+1}^+ > S_T$ for $n \geq 1$ and $\lim_{n \to \infty} S_n^+ \downarrow S_T$.

Proof. We only show (i). The ordering of $S^\pm_n$ is a consequence of the fact that orbits cannot intersect themselves. Now suppose that a $\delta > 0$ exists such that

$$S_n^- \leq S_T - \delta$$

for all $n \geq 1$. (5.27)

Let,

$$\mathcal{R}_\delta = \{(S, u) : S_T - \delta < S < S_T, p^{(i)}_c(S) < u < p^{(d)}_c(S_T - \delta)\},$$

see Figure 5.8. Since the orbit is confined to a neighborhood of $E_T_i E_T_d$ and rotates around $\mathcal{R}_\delta$, it must converge, as $\xi \to \infty$, to a limit cycle $C$. Let $\Omega$ denote the region enclosed by $C$. Clearly $\Omega \supseteq \mathcal{R}_\delta$. Since $\text{div} \vec{R} \leq 0$ in $\Omega$, in particular $\text{div} \vec{R} = \frac{1}{\tau T} p^{(d)}_c(S_T - \delta)$ in $\{S, u) : S_T - \delta < S < S_T, p^{(d)}_c(S) < u < p^{(d)}_c(S_T - \delta)\}$, the divergence theorem applied to $\vec{R}$ in $\Omega$ yields a contradiction. Hence $\delta = 0$. \qed

![Figure 5.8: Rotation of orbit around $\mathcal{R}_\delta$.](image)

When the orbit crosses the hysteresis domain $\mathcal{H}$ at some $S = \hat{S} \in (0, 1)$, the pressure equation satisfies

$$u' = \mathcal{G}(\hat{S}; S_B, S_T).$$
Hence, the $\xi$-interval ($\ell$) in which this takes place has length

$$
\mu(\ell) = \frac{p_c^{(d)}(S) - p_c^{(i)}(S)}{q(S; S_B, S_T)}.
$$

Thus as a consequence of Lemma 5.1, we have

Figure 5.9: Sketch of saturation and pressure behaviour. Here $S_T < S^*_T(S_B)$ implying $S < S_m < 1$. If $S_T > S^*_T(S_B)$ and $\tau > \tau^*$, the first saturation plateau occurs at $S = 1$.

**Corollary 5.1.** Corresponding to the saturation plateaus $\{S^\pm_n\}_{n=1}^\infty$, there exist $\xi$-intervals $\{\ell^\pm_n\}_{n=1}^\infty$, $\ell^\pm_n \subset \mathbb{R}$, satisfying

(i) $\ell^\pm_n$ is between $\ell^+_n$ and $\ell^+_n$. 

5.2 Travelling waves with combined pressure (5.2)

\[ (ii) \quad \mu(\ell_n^\pm) = \frac{p_c^{(d)}(S_n^\pm) - p_c^{(i)}(S_n^\pm)}{\mathcal{G}(S_n^\pm; S_B, S_T)}. \]

\[ (iii) \quad \lim_{n \to \infty} \mu(\ell_n^\pm) = \infty. \]

Remark 5.3. Lemma 5.1 implies that the pressure \( u(\xi) \) oscillates between \( p_c^{(i)}(S_T) \) and \( p_c^{(d)}(S_T) \) as \( \xi \to \infty \). But Corollary 5.1 shows that \( \lim_{\xi \to \infty} u'(\xi) = 0 \).

Proof. (i) This is immediate from the orbit rotating around \( \overline{E_T,iE_T,d} \).

(ii) The expression follows directly from (5.29).

(iii) This is a consequence of Lemma 5.1, (ii), \( p_c^{(d)}(S) > p_c^{(i)}(S) \) for \( 0 < S < 1 \) and \( \mathcal{G}(S_T; S_B, S_T) = 0 \).

The behaviour described by Lemma 5.1 and Corollary 5.1 is sketched in Figure 5.9.

\[ \tau \leq \tau_{c,d} \]

In this case either \( E_{T,d} \) (if \( \tau_{c,i} < \tau \leq \tau_{c,d} \)) or both \( E_{T,i} \) and \( E_{T,d} \) (if \( \tau \leq \tau_{c,i} \)) are a sink. This gives the following result.

Lemma 5.2. Let \( 0 < S_B < S_T < 1 \) and \( \tau \leq \tau_{c,d} \). Then for any orbit satisfying (5.12), we have:

(i) If \( 0 < \tau \leq \tau_{c,i} \), it moves directly and monotonically through \( \mathcal{H}_i \) from \( E_{B,i} \) to \( E_{T,i} \) (a sufficient condition for this behaviour is given in Proposition 5.1) or it rotates finitely many times around \( \overline{E_{T,i}E_{T,d}} \) and ends up in \( E_{T,i} \) or \( E_{T,d} \);

(ii) If \( \tau_{c,i} < \tau \leq \tau_{c,d} \), it ends up in \( E_{T,d} \) after a finite number of passages through \( \mathcal{H} \).

Proof. (i) If the orbit does not move directly from \( E_{B,i} \) to \( E_{T,d} \) through \( \mathcal{H}_i \), it has to cross the hysteretic domain \( \mathcal{H} \) or it has a saturation plateau at \( S = 1 \). Then we argue as in Lemma 5.1 to show that the orbit must approach \( \overline{E_{T,i}E_{T,d}} \) as \( \xi \to \infty \). The fact that both \( E_{T,i} \) to \( E_{T,d} \) are sinks implies that the orbit will be captured by one of these points after finitely many rotations.

(ii) Now only \( E_{T,d} \) is a sink. Arguing as in (i) gives the result. \( \square \)
In the discussion and figures so far $S_T < 1$ has been considered. If $S_T = 1$, the orbit cannot intersect the imbibition pressure curve at $S_T < 1$. This follows from the flow direction as shown in Figure 5.10. Furthermore, each point of the half line $(S = 1, u \leq 0)$ is an equilibrium point of (5.12) with the extended capillary pressure (5.4). In this case the existence result is given by

\[ u \leq 0 \]

**Lemma 5.3.** Let $0 < S_B < S_T = 1$.

(i) If $p_c^{(i)}(1) = 0$, then for any $\tau > 0$ the solution of (5.12) is in $\mathcal{H}_i$ and runs from $E_{B,i}$ as $\xi \to -\infty$ to a point $(S = 1, u = u_\tau < 0)$ at a finite value of $\xi = \xi_\tau$. For $\xi \geq \xi_\tau$ the solution is given by $S(\xi) = 1$ and $u(\xi) = u_\tau$. The pressure value $u_\tau$ decreases in $\tau$ and has the lower bound $u_\tau \geq -\frac{2cK}{c^2\tau}$, where $K = -\alpha(S_B,1) > 0$. Furthermore, $\lim_{\xi \to \xi_\tau} S'(\xi) = -\frac{u_\tau}{c\tau}$ and $\lim_{\xi \to \xi_\tau} u'(\xi) = 0$.

(ii) If $p_c^{(i)}(1) < 0$, then the solution behaves as in (i) when $\tau > \tau_{c,i}$. For $\tau \leq \tau_{c,i}$ it may end up in $(1,0)$ as $\xi \to -\infty$.

**Proof.** Since the orbit belongs to $\mathcal{H}_i$, equations (5.12) reduce to

\[
\begin{align*}
S' & = \frac{1}{c\tau} (p_c^{(i)}(S) - u), \\
u' & = g(S;S_B,1).
\end{align*}
\]

(i) With $p_c^{(i)}(1) = 0$, the eigenvalues at the point $(1,0)$ are given by

\[
\lambda_k = \pm i\sqrt{c\tau g'(1;S_B,1)},
\]

where $i^2 = -1$. This means that for any $\tau > 0$, the orbits cannot reach $(1,0)$ as $\xi \to -\infty$. Hence they must exit $\mathcal{H}_i$ at some point $(S = 1, u = u_\tau < 0)$ for some finite $\xi = \xi_\tau$. For
\(5.2\) Travelling waves with combined pressure (5.2)

\[\xi \geq \xi_\tau\] we extend the solution by the equilibrium \((S = 1, u = u_\tau)\). The properties of the pressure value \(u_\tau\) are demonstrated along the lines of the proof of Theorem 4.1 (b) and Proposition 4.9. Finally, letting \(\xi \to \xi_\tau\) in equations (5.30) gives

\[
\lim_{\xi \to \xi_\tau} S'(\xi) = \frac{1}{c_T} \lim_{\xi \to \xi_\tau} \left( p_c^{(i)}(S(\xi)) - u(\xi) \right) = -\frac{u_\tau}{c_T}
\]

and

\[
\lim_{\xi \to \xi_\tau} u'(\xi) = \lim_{\xi \to \xi_\tau} \mathcal{G}(S(\xi); S_B, 1) = 0.
\]

(ii) If \(p_c^{(i)'}(1) < 0\), the orbit behaves as in (i) when \(\tau > \tau_{c, i}\). For \(\tau \leq \tau_{c, i}\) the point \((1, 0)\) is a sink and thus the orbit may end up at that point as \(\xi \to \infty\). In particular this happens for \(0 < \tau \leq \tau_m\).

\[\square\]

5.2.2 The case \(f \in L^1(0, 1)\)

When the capillary functions in (5.2) is non-constant, one uses (5.17) to recast (5.12) in the equivalent form (5.19), where \(f\) is now removed from the equations. If \(f \in L^1(0, 1)\), then \(0 < Y < Y(1) = \int_0^1 f(\varrho) d\varrho\). The behaviour in terms of \((Y, u)\) in the strip \(\{(Y, u) : 0 < Y \leq Y(1), u \in \mathbb{R}\}\) is similar to the behaviour of \((S, u)\) with \(f(S) = 1\), as discussed in Section 5.2.1.

In terms of the transformed nonlinearities \(\tilde{\mathcal{F}}, \tilde{\mathcal{G}}\) and \(\tilde{p}_c^{(k)} (k \in \{i, d\})\), see definitions below (5.19), the characteristic \(\tau\)-values in terms of \(Y\) are

\[
\bar{\tau}_{c, k} = \frac{(\tilde{p}_c^{(k)'}(Y_T))^2}{4c_T \tilde{\mathcal{G}}'(Y_T; Y_B, Y_T)} \quad (k \in \{i, d\}).
\]

Likewise,

\[
\bar{\tau}(Y_B, Y_T) = \int_{Y_B}^{Y(1)} \tilde{\mathcal{G}}(\varrho; Y_B, Y_T) d\varrho
\]

and

\[
\int_{Y_B}^{Y_m} \tilde{\mathcal{G}}(\varrho; Y_B, Y_T) d\varrho = 0.
\]

Transforming these expressions back to the original saturation \(S\) now gives

\[
\tau_{c, k} = \frac{(p_c^{(k)'}(S_T))^2}{4c f(S_T) \mathcal{G}'(S_T; S_B, S_T)} \quad (k \in \{i, d\}),
\]

\[
\alpha(S_B, S_T) = \int_{S_B}^{1} \mathcal{G}(\varrho; S_B, S_T) f(\varrho) d\varrho,
\]

\[
\int_{S_B}^{S_m} \mathcal{G}(\varrho; S_B, S_T) f(\varrho) d\varrho = 0.
\]
Given these new definitions one can repeat the arguments of Section 5.2.1 to obtain similar results. Clearly \( S_T^*(S_B) \), see Figure 5.5, depends on \( f \) through the expression (5.35). Also the value \( \tau^* \) from Theorem 5.1 will change with \( f \).

Remark 5.4. Let \( f_1, f_2 \in L^1(0, 1) \) and \( f_1 < f_2 \) in \((0, 1)\). Then the critical \( \tau \)-values (5.34) are ordered as well:

\[
\tau_{c,k,1} > \tau_{c,k,2} \quad (k \in \{i, d\})
\]

Hence in case 2, overshoot is expected to occur for smaller values of \( \tau \).

### 5.2.3 The case \( f \not\in L^1(0, 1) \)

Again we use transformation (5.17) and the corresponding redefinitions of the non-linearities. Borrowing ideas from Section 4.4.2, we distinguish the cases \( S_T < 1 \) and \( S_T = 1 \).

\( S_T < 1 \). For this case, there always exists \( S_m \in (S_T, 1) \) satisfying (5.36). To see this, observe that for any \( \delta \in (0, \frac{1-S_T}{2}) \)

\[
\int_{S_B}^{1-\delta} f\theta dS = \int_{S_B}^{(S_T+1)/2} f\theta dS + \int_{(S_T+1)/2}^{1-\delta} f\theta dS \geq \int_{S_B}^{(S_T+1)/2} f\theta dS + \theta((S_T+1)/2; S_B, S_T) \int_{(S_T+1)/2}^{1-\delta} f dS.
\]

Since \( \lim_{\delta \to 0} \int_{S_B}^{1-\delta} f\theta dS = \infty \) and as \( \int_{S_B}^{S_T} f\theta dS < 0 \) the assertion follows. As \( S_m < 1 \), Theorem 5.1 further implies that \( S(\xi) \leq S_m < 1 \) for all \( \tau > 0 \) and \( \xi \in \mathbb{R} \). Hence, \( f(S(\xi)) \) is bounded away from \( \infty \) for \( \xi \in \mathbb{R} \). Consequently, the results presented in Lemma 5.1-5.2 and Corollary 5.1 remain valid and depending upon the value of \( \tau \), the orbits can either end up in \( E_{T,i} \) or \( E_{T,d} \) or rotate infinitely many times around the segment \( E_{T,i}E_{T,d} \) while approaching it.

\( S_T = 1 \). Recall that each point on the half-line \( \{S = 1, u \leq 0\} \) is an equilibrium point. In fact the extension of \( p_c^{(k)}(k \in \{i, d\}) \) as in (5.6), is not required in this case, as \( (P_k(S) - u)/cT f \to 0 \) for \( S \not\to 1, k \in \{i, d\} \). This implies that an orbit \((S, u)\) leaving \( E_{B,i} \) cannot cross the line \( \{S = 1, u \leq 0\} \). Moreover, \( u'(<\xi) = \theta(S(\xi); S_B, 1) < 0 \) and \( S'(\xi) > 0 \) for all \( \xi \in \mathbb{R} \). This along with the direction of orbits shown in Figure 5.10 implies that there are only three options for the orbit \((S, u)\) as \( \xi \to \infty \): (i) \((S, u) \to (1, 0)\), (ii) \((S, u) \to (1, u_T)\) with \(-\infty < u_T < 0\) or (iii) \((S, u) \to (1, -\infty)\).

From Theorem 4.6 it follows that (iii) occurs if and only if

\[
\int_{S_B}^{1} \theta(\varrho; S_B, 1) f(\varrho) d\varrho = -\infty.
\]
However, since \( \mathcal{G}(1; S_B, 1) = 0 \) this integral may be finite. As \( \mathcal{G}(S; S_B, 1) \) is Lipschitz and negative for \( S > S_B \), the integral is finite if

\[
\int_0^1 (1 - \varrho) f(\varrho) d\varrho < \infty.
\]

In this case, one defines \( \alpha(S_B, 1) \) by putting \( S_T = 1 \) in (5.35). As \( \alpha(S_B, 1) > -\infty \), \( (S, u) \to (1, u_T) \) as \( \xi \to \infty \) for some \( -\infty < u_T \leq 0 \). Moreover, \( u_T \) is bounded below by the estimate given in Lemma 5.3. Since \( \mathcal{G}'(1; S_B, 1) > 0 \) and \( \lim_{S \to 1} f(S) = \infty \) we have \( \tau_{c,i} = 0 \) from (5.34). Thus, the orbit cannot reach \((0, 1)\) for any \( \tau > 0 \). This leaves (ii) as the only possibility for this particular case.

We summarize these findings in the following lemma

**Lemma 5.4.** Let \( f \not\in L^1(0,1) \) and assume that \( \tau > 0, S_B > 0 \) and \( S_T \in (S_B, 1) \). Further, if \( (1 - \varrho) f(\varrho) \in L^1(0,1) \), let \( \alpha(S_B, 1) := \int_{S_B}^1 \mathcal{G}(\varrho; S_B, 1) f(\varrho) d\varrho \). Then for orbit \((S(\xi), u(\xi))\) satisfying (5.12) we have:

(i) If \( S_T < 1 \), results presented in Lemmas 5.1-5.2 and Corollary 5.1 remain valid and \( S(\xi) \leq S_m < 1 \) for all \( \xi \in \mathbb{R} \), where \( S_m \) is defined by (5.36).

(ii) If \( S_T = 1 \), we distinguish two cases:

(ii.1) If \( (1 - \varrho) f(\varrho) \not\in L^1(0,1) \), then \( S(\xi) \not\to 1 \) and \( u(\xi) \to -\infty \) as \( \xi \to \infty \).

(ii.2) If \( (1 - \varrho) f(\varrho) \in L^1(0,1) \), then \( (S, u) \to (1, u_T) \) as \( \xi \to \infty \) for some \( u_T \in (-\infty, 0] \).

Here \( u_T \) satisfies \( u_T \geq -\sqrt{-2ca(S_B, 1)\tau} > -\infty \).

Figure 5.11 shows a schematic diagram outlining the cases.

![Figure 5.11: Behaviour of orbits for \( f \not\in L^1(0,1) \). The dotted (orange) line represents the orbit for \( S_T = S_{T,1} < 1 \). The dashdotted line (green) represents the orbit for \( S_T = S_{T,2} = 1 \), when \( (1 - \varrho) f(\varrho) \in L^1(0,1) \) and \( \tau > \hat{\tau} \).](image-url)
5.3 Travelling waves with combined pressure (5.4)

In the previous section we discussed travelling waves arising from the combined pressure model with play-type hysteresis (5.2). The scanning curves for the play-type hysteresis model are vertical, whereas scanning curves obtained from experiments are inclined [24, 170] (see Figure 1.3). As pointed out in Chapters 1 and 2, this may lead to inaccurate predictions. The combined model with extended play-type hysteresis given by (5.4), incorporates non-vertical scanning curves that can be fitted to match experimental observations, see Chapter 2. Hence, it is interesting to study how the behaviour of the infiltration fronts changes if one uses the extended model to describe hysteresis.

As seen from Section 5.2, for sufficiently large values of \( \tau \), the orbit \((S(\xi), u(\xi))\) revolves around the line segment \(ET_{Te}E_{Td}\) infinitely many times. From a physical point of view, this means that for an infinitely long column, one observes infinitely many decaying oscillations trailing behind the first overshoot. This behaviour is indicated in Figure 5.9. However, few to no oscillations are observed behind the first overshoot in experimentally obtained breakthrough curves [35, 72] even for long columns and large injection rates, see also Figure 1.1(b). In this section we explain this by considering non-vertical scanning curves.

In the subsequent discussions, we assume that the inclination of the scanning curves satisfies

\[
\min\{p_c^{(i)}(S), p_c^{(d)}(S)\} > -\frac{1}{c}, \quad \text{for all } S_B \leq S \leq 1.
\]

As shown in [80], consistency of the extended model requires that (5.40) should be satisfied for all \( S \in (0, 1] \). Hence, for more general problems a non-constant \( \varepsilon \) needs to be considered. However, since travelling waves satisfy \( S \geq S_B \), and since \( p_c^{(i)}, p_c^{(d)} \) are bounded in \( S \in [S_B, 1] \) it suffices to consider (5.40) for \( \varepsilon \) sufficiently small.

Substituting the travelling wave form (5.7) into (5.1) and (5.5), and applying conditions (5.8) and (5.10) we obtain the same Rankine-Hugoniot wave speed \( c \) and the same function \( \vartheta(S; S_B, S_T) \), see (5.11) and (5.13). The extended system now reads:

\[
\begin{align*}
S' &= \frac{1}{c\varrho_f(S)} E(S, u), \\
0' &= \vartheta(S; S_B, S_T),
\end{align*}
\]

where

\[
E(S, u) = \begin{cases} 
  p_c^{(d)}(S) - u & \text{when } u > p_c^{(d)}(S) + \varepsilon \tau c f(S) \vartheta(S; S_B, S_T), \\
  -c\varrho_f(S) \vartheta(S; S_B, S_T) & \text{when } p_c^{(i)}(S) + \varepsilon \tau c f(S) \vartheta(S; S_B, S_T) \\
  \leq u \leq p_c^{(d)}(S) + \varepsilon \tau c f(S) \vartheta(S; S_B, S_T), \\
  p_c^{(i)}(S) - u & \text{when } u < p_c^{(i)}(S) + \varepsilon \tau c f(S) \vartheta(S; S_B, S_T).
\end{cases}
\]

(5.42)
Equations (5.41), with expression (5.42), shows that \( f(S) \) cannot be eliminated by transformation (5.17). Therefore, for simplicity we restrict ourselves here to the case \( f(S) = 1 \) only. Then the system becomes

\[
\begin{align*}
S' &= \frac{1}{cT} \mathcal{E}(S, u), \\
u' &= \mathcal{G}(S; S_B, S_T),
\end{align*}
\]

(5.43a, 5.43b)

Here

\[
\mathcal{E}(S, u) = \begin{cases} 
p_c^{(d)}(S) - u & \text{when } u > \hat{u}(S), \\
-\varepsilon \tau \mathcal{G}(S; S_B, S_T) & \text{when } \hat{u}(S) \leq u \leq \check{u}(S), \\
p_c^{(i)}(S) - u & \text{when } u < \check{u}(S),
\end{cases}
\]

where \( \hat{u}, \check{u} : (0, 1) \rightarrow \mathbb{R} \) are given by

\[
\hat{u}(S) = p_c^{(i)}(S) + \varepsilon \tau \mathcal{G}(S; S_B, S_T) \quad \text{and} \quad \check{u}(S) = p_c^{(d)}(S) + \varepsilon \tau \mathcal{G}(S; S_B, S_T).
\]

(5.44)

Note that for each \( \delta > 0, \mathcal{E}(S, u) \) is uniformly Lipschitz continuous in the strip \( \{ \delta \leq S \leq 1 \} \). Similar to the sets \( \mathcal{H}, \mathcal{H}_1, \mathcal{H}_d \) in (5.14), we now divide the strip \( \{ 0 < S < 1 \} \) into

\[
\begin{align*}
\mathcal{H} &= \{ u > \hat{u}(S) \}, \\
\mathcal{H}_1 &= \{ \hat{u}(S) < u < \check{u}(S) \}, \\
\mathcal{H}_d &= \{ u < \check{u}(S) \}.
\end{align*}
\]

(5.45a, 5.45b, 5.45c)

For given \( 0 < S_B < S_T \leq 1 \), the parameters of the system are \( \varepsilon \) and \( \tau \). In the region \( \mathcal{H} \) one has

\[
S' = -\varepsilon \mathcal{G}(S; S_B, S_T) = -\varepsilon u',
\]

(5.46)

implying indeed scanning curves with inclination \( -1/\varepsilon \) in \( \mathcal{H} \). As in play-type hysteresis, each point of the segments \( E_{B,i}E_{B,d} \) and \( E_{T,i}E_{T,d} \) is an equilibrium point. From equation (5.46) and the sign of \( \mathcal{G}(S; S_B, S_T) \) it follows that interior points of \( E_{B,i}E_{B,d} \) are unstable, whereas interior points of \( E_{T,i}E_{T,d} \) are stable. Further, the nature of the end points \( E_{B,k}E_{T,k} \), \( k \in \{ i, d \} \) as in Section 2: \( E_{B,k} \) are saddles and \( E_{T,k} \) are stable nodes and/or stable spirals depending on the value of \( \tau \). This classifies all equilibrium points of (5.43).

The main result of this section is the existence of solutions of (5.43), subject to conditions (5.8) and (5.10), in the form of orbits that connect arbitrarily chosen points \( E_B \in E_{B,i}E_{B,d} \) (as \( \xi \rightarrow -\infty \)) to corresponding points \( E_T \in E_{T,i}E_{T,d} \) (as \( \xi \rightarrow +\infty \)). We shall give the proof in context of Figure 5.12. This figure shows the regions \( \mathcal{H}, \mathcal{H}_1, \mathcal{H}_d \), the functions \( \hat{u}, \check{u} \) and the general flow directions of solution orbits \( (S, u) \). To support the proof we added Figure 5.13 which focuses on the region between the segments \( E_{B,i}E_{B,d} \) and \( E_{T,i}E_{T,d} \).
Figure 5.12: The sets (5.45) and the curves (5.44) in the strip $\{0 < S \leq 1\}$. The arrows indicate the direction for increasing $\xi$ of solutions of (5.43).

Figure 5.13: Possible orbits starting, as $\xi \to -\infty$, from points $E_B \in E_{B,i}, E_{B,d}$. Here $E_B = (S_B, u_B)$ and $\tan \beta = -\frac{1}{\varepsilon}$.

Let $u_B \in [p_c^{(i)}(S_B), p_c^{(d)}(S_B)]$ and let

$$\ell(S; u_B) = u_B - \frac{1}{\varepsilon}(S - S_B) \text{ for } S > S_B.$$  \hspace{1cm} (5.47)
5.3 Travelling waves with combined pressure (5.4)

With reference to Figure 5.12 and Figure 5.13 we assume that \( \epsilon \) and \( \tau \) are such that

\[
\ell(S; p_c^{(d)}(S)) \leq \hat{u}(S) \quad \text{for} \quad S_B \leq S \leq 1,
\]
and

\[
\hat{u}'(S) \geq -\frac{1}{\epsilon}.
\]

As \( \ell(S_B, p_c^{(d)}(S_B)) = p_c^{(d)}(S_B) \), a sufficient condition for (5.48) to be true is

\[
1 + \epsilon \hat{u}'(S) = 1 + \epsilon p_c^{(d)'}(S) + c_t \epsilon^2 \varphi'(S; S_B, S_T) \geq 0,
\]
for \( S_B \leq S \leq 1 \). Similarly for (5.49) the condition is

\[
1 + \epsilon \hat{u}'(S) = 1 + \epsilon p_c^{(i)'}(S) + c_t \epsilon^2 \varphi'(S; S_B, S_T) \geq 0
\]
for \( S_B \leq S \leq 1 \). From (5.40), as \( 1 + \epsilon p_c^{(i)}(S) > 0 \) \((k \in \{i, d\})\), Conditions (5.50) are satisfied if \( \epsilon^2 \tau \) is small enough. Moreover, because \( \varphi' > 0 \) for \( S \geq S_I \) (see Figure 5.1) it follows from (5.40) that Conditions (5.50) are automatically satisfied for \( S > S_I \). Hence, Conditions (5.50) are only to be assumed for \( S \in [S_B, S_I] \).

Observe that from (5.40),

\[
\ell(1; p_c^{(d)}(S_B)) < 0 = \lim_{S \to 1} p_c^{(i)}(S).
\]

Using (5.40) and (5.51) it follows that there exists a unique saturation value \( \hat{S} \in (S_B, 1) \) such that

\[
\ell(\hat{S}; p_c^{(d)}(S_B)) = p_c^{(i)}(\hat{S}).
\]

Figure 5.14 shows that if \( S_T < \hat{S} \), then there exists a unique \( u_B^* \in (p_c^{(i)}(S_B), p_c^{(d)}(S_B)) \) for which the following holds: All orbits starting at \( (S_B, u = u_B \geq u_B^*) \) remain in \( \hat{\mathcal{H}} \) and follow the corresponding scanning curves until \( E_{T,i}E_{T,d} \) is reached. The underlying travelling waves are solutions of (5.46). They result in monotone saturation and pressure profiles for any \( \tau > 0 \). Orbits starting at \( (S_B, u_B < u_B^*) \) intersect \( \hat{u} \) at some \( S < S_T \) and enter the region \( \hat{\mathcal{H}} \). When \( S_T > \hat{S} \) any orbit leaving from \( E_{B,i}E_{B,d} \) intersects \( \hat{u} \) and enters the region \( \hat{\mathcal{H}} \). The saddle point \( E_{B,i} \) needs special attention. If the orbit leaves \( E_{B,i} \) into region \( \hat{\mathcal{H}} \), its local direction \( \frac{du}{dS}(S_B) \), see Section 4.3, must satisfy

\[
\frac{du}{dS}(S_B) = \frac{p_c^{(i)'}(S_B)}{2} \left[ 1 + \sqrt{1 - \frac{4c_t \varphi'(S; S_B, S_T)}{(p_c^{(i)'}(S_B))^2}} \right] \leq \frac{d\hat{u}}{dS}(S_B).
\]

We have
Proposition 5.2. Condition (5.50b) at $S_B$ $\iff$ inequality (5.53).

Proof. We use the short hand notation $p^{(i)}_c := p^{(i)}_c(S_B)$ and $\gamma' = \gamma'(S_B;S_B,S_T)$. Evaluation of (5.50b) at $S_B$ gives

$$1 + \varepsilon p^{(i)}_c + \varepsilon^2 c \gamma' \geq 0$$

or $-\frac{1}{p^{(i)}_c} \geq \varepsilon + \frac{\varepsilon^2 c \gamma'}{p^{(i)}_c}$. Hence, as $\gamma', p^{(i)}_c < 0$ we get

$$1 - \frac{4c \gamma'}{(p^{(i)}_c)^2} \geq 1 + \frac{4c \gamma'}{(p^{(i)}_c)^2} + \left(1 + \frac{2c \gamma'}{(p^{(i)}_c)^2} \right)^2,$$

implying

$$\sqrt{1 - \frac{4c \gamma'}{(p^{(i)}_c)^2}} \leq 1 + 2c \gamma' p^{(i)}_c.$$

This gives

$$\frac{p^{(i)}_c}{2} + \frac{p^{(i)}_c}{2} \sqrt{1 - \frac{4c \gamma'}{(p^{(i)}_c)^2}} \leq p^{(i)}_c + c \gamma' p^{(i)}_c$$

which is inequality (5.53).

Summarizing the above gives the following picture. When conditions (5.50) are fulfilled, orbits that leave from points $E_B = (S_B,u_B) \in \mathcal{E}_{B,i} E_{B,d}$, with $u_B > p^{(i)}_c(S_B)$, either stay in $\mathcal{R}$ and reach corresponding points of $\mathcal{E}_{T,i} \mathcal{E}_{T,d}$ directly, or cross the

![Figure 5.14: Sketch of orbits when $S_T < \bar{S}$. Here, $\tan(\beta) = -\frac{1}{\varepsilon}$.](image-url)
lower bound \( \tilde{u} \) and enter the region \( \tilde{\mathcal{H}} \). The orbit that leaves \( E_{B,i} \) enters directly \( \tilde{\mathcal{H}} \). Once the orbits are in \( \tilde{\mathcal{H}} \), we are in a situation that compares to Theorem 5.1 and Remark 5.2 since in \( \tilde{\mathcal{H}} \) one has

\[
S' = \frac{1}{cT} (p^{(i)}_c(S) - u).
\]

What remains is to show that any such orbit is captured by the segment \( E_{T,i}E_{T,d} \). We make this the content of the following theorem.

**Theorem 5.2.** Let \( 0 < S_B < S_T < 1 \), \( \tau > 0 \) and let conditions (5.50) and (5.40) be satisfied. Further, let \( (S,u) \) be the solution orbit that originates from a point \( E_B = (S_B,u_B) \) on \( E_{B,i}E_{B,d} \), where \( u_B < u^*_B \) if \( S_T \leq \tilde{S} \) or \( u_B \in [p^{(i)}_c(S_B), p^{(d)}_c(S_B)] \) if \( S_T > \tilde{S} \) (see Figure 5.14). Then the following holds

(i) If \( \tau > \tau_{c,d} \), the orbit exceeds \( S_T \) and reaches a point \( E_T \in E_{T,i}E_{T,d} \). This happens after, at most, a finite number of rotations around \( E_{T,i}E_{T,d} \);

(ii) If \( \tau_{c,i} < \tau \leq \tau_{c,d} \), then in addition, the orbit may reach the boundary point \( E_{T,d} \) from \( \tilde{\mathcal{H}} \);

(iii) If \( \tau \leq \tau_{c,d} \), then in addition the orbit may reach the boundary point \( E_{T,i} \) from \( \tilde{\mathcal{H}} \).

For small enough \( \tau \) this will happen in a direct way, without exceeding \( S_T \).

**Proof.** The proof has two parts. In the first we use again the divergence argument to demonstrate the existence of orbits connecting \( E_B \) and corresponding points of \( E_{T,i}E_{T,d} \). In the second we show that the orbits can go around \( E_{T,i}E_{T,d} \) only finitely many times.

To use the divergence argument we write \( \vec{R}(\vec{u}) \) as

\[
\vec{R}(\vec{u}) = \left( \frac{1}{cT} \mathcal{G}(S,u), \mathcal{G}(S;S_B,S_T) \right).
\]

Hence

\[
\text{div} \vec{R} = \frac{1}{cT} \frac{\partial \mathcal{G}}{\partial S} + \frac{\partial \mathcal{G}}{\partial S} = \begin{cases} 
p_c^{(i)}(S) & \text{when } u \in \tilde{\mathcal{H}}, \\
p_c^{(d)}' + \mathcal{G}'(S;S_B,S_T) & \text{when } u \in \mathcal{H}, \\
-p_c^{(d)}(S) & \text{when } u \in \mathcal{H}_d. \end{cases}
\]

Note that \( \text{div} \vec{R} \) does not have a fixed sign as the sign of \( \mathcal{G}' \) changes in \( \tilde{\mathcal{H}} \). With reference to Figure 5.12, we see that there are three possibilities for orbits that originate from \( E_{B,i}E_{B,d} \):

Case (a) They go around \( E_{T,i}E_{T,d} \) and end up in \( E_{B,d} \) or they leave the region \( S > S_B \) through the segment \( \{S = S_B, u > p^{(d)}_c(S_B)\} \).
Case (b) They go into a limit cycle around $E_{T,i} E_{T,d}$.

Case (c) They approach the line $E_{T,i} E_{T,d}$.

Figure 5.15: If Case (a) holds, then the orbit connects to the segment \{\(S = S_B, u \geq p_c^{(d)}(S_B)\)\}. The regions $\Omega$, $\Omega_1$, $\Omega_2$ used in the proof are marked by the colours.

We first consider the orbit leaving $E_{B,i}$. Case (a) for this orbit is shown in Figure 5.15. Let $\Omega$ be the region enclosed by the orbit and the line $S = S_B$. As shown in Chapter 4, the orbit can only exit through $S = S_B$ if

$$\int_{\Omega} \text{div}\mathbf{R} = \int_{\partial\Omega} \hat{n} \cdot \mathbf{u} \geq 0, \quad (5.55)$$

$\hat{n}$ being the unit vector that is outward normal to $\partial\Omega$. Define the regions $\Omega_1, \Omega_2 \subset \Omega$ by

$$\Omega_1 = \{(S, u): S_B < S < S_T, \hat{u}(S) < u < p_c^{(d)}(S_B)\},$$

$$\Omega_2 = \{(S, u): S_B < S < S_T, \hat{u}(S) < u < \hat{u}(S)\}, \quad (5.56)$$

and observe that

$$\int_{\Omega \setminus (\Omega_1 \cup \Omega_2)} \text{div}\mathbf{R} \leq 0,$$

since $\frac{\partial \xi}{\partial S} \leq 0$ in $\Omega/\Omega_2$. Further note that

$$\int_{\Omega_1} \text{div}\mathbf{R} = \int_{S_B}^{S_T} \int_{\hat{u}(S)}^{p_c^{(d)}(S_B)} \frac{p_c^{(d)}(S)}{cT} du dS = \frac{1}{cT} \int_{S_B}^{S_T} p_c^{(d)}(S)p_c^{(d)}(S_B) - p_c^{(d)}(S) - c\varepsilon G(S)dS$$

$$= -\frac{1}{2cT}(p_c^{(d)}(S_T) - p_c^{(d)}(S_B))^2 - \varepsilon \int_{S_B}^{S_T} p_c^{(d)} G dS,$$
and
\[
\int_{\Omega_2} \text{div} \vec{R} = \int_{S_B}^{S_T} \int_{p_{c}^{(d)}(S)}^{p_{c}^{(d)}(S)} (-\varepsilon g'(S))dudS = -\varepsilon \int_{S_B}^{S_T} g'(p_{c}^{d}(S)) - p_{c}^{(i)}(S))dS \\
= -\varepsilon \int_{S_B}^{S_T} g'(p_{c}^{d}(S)) - p_{c}^{(i)}(S))dS + \varepsilon \int_{S_B}^{S_T} (p_{c}^{d}(S)) - p_{c}^{(i)}(S))dS \\
= \varepsilon \int_{S_B}^{S_T} (p_{c}^{d}(S)) - p_{c}^{(i)}(S))dS.
\]

Summing up gives
\[
\int_{\Omega} \text{div} \vec{R} < -\frac{1}{2\varepsilon \tau} (p_{c}^{d}(S_T) - p_{c}^{d}(S_B))^2 - \varepsilon \int_{S_B}^{S_T} (p_{c}^{d}(S)) dS < 0.
\]

This contradicts (5.55) and hence the orbit cannot exist through the segment \( \{S = S_B, u \geq p_{c}^{(d)}(S_B)\} \). As no two orbits can intersect, the orbit originating from any point on \( E_{B,1} \) cannot leave the region \( \{S > S_B\} \).

![Figure 5.16: Case (b): If the orbits go to a limit cycle then the limit cycle will be as the black dashdotted closed loop. The regions \( \Omega^0, \Omega^1, \Omega^2 \) and \( \{S_{in}, S_{out}\} \) are used in the proof.](image)

We need to further eliminate Case (b). As can be seen from Figure 5.14 a limit cycle clearly cannot exist for \( S_T \leq \bar{S} \). For \( S_T > \bar{S} \), if the limit cycle exists, then it will be as shown in Figure 5.16. Let \( \Omega^0 \) be the region bounded by the closed loop and let \( \{S_{in}, S_{out}\} \) be saturation values at which the limit cycle enters and leaves the region \( \mathcal{H} \) such that \( S_{in} < S_{out} < S_T \). Further, let \( \Omega^1 := \{(S, u) \in \Omega^0 : S_{in} < S < S_T, \bar{u}(S) < u < p_{c}^{(d)}(S_{in})\} \) and let \( \Omega^2 \subset \Omega^0 \) be the open subregion in \( S < S_T \) bounded by \( S = S_T \),
\( u = \hat{u}(S), \quad u = \check{u}(S) \) and the closed loop. Again, we have \( \text{div} \vec{R} < 0 \) in \( \Omega^0 / \Omega^2 \). The integral of \( \text{div} \vec{R} \) over \( \Omega^1 \) is

\[
\int_{\Omega^1} \text{div} \vec{R} = \frac{1}{ct} \int_{S_{in}}^{S_T} p_c^{(d)}(S)(p_c^{(d)}(S_{in}) - p_c^{(d)}(S) - ct\varepsilon \mathcal{G}(S))dS
\]

\[
= -\frac{1}{2ct}(p_c^{(d)}(S_T) - p_c^{(d)}(S_{in}))^2 - \varepsilon \int_{S_{in}}^{S_T} p_c^{(d)}GdS,
\]

and over \( \Omega^2 \) is

\[
\int_{\Omega^2} \text{div} \vec{R} = -\varepsilon \int_{S_{in}}^{S_{out}} \mathcal{G}'(\int_{S_{in}}^{(d)}(S_{in})^{-\frac{1}{2}}(S-S_{in}) du) dS - \varepsilon \int_{S_{out}}^{S_{in}} \mathcal{G}'(\int_{S_{in}}^{\hat{u}(S)} du) dS.
\]

Integrating by parts yields

\[
\int_{\Omega^2} \text{div} \vec{R} = \varepsilon \int_{S_{in}}^{S_T} p_c^{(d)}G + \int_{S_{in}}^{S_{out}} (1 + ct\varepsilon^2 \mathcal{G}'(\mathcal{G} - \varepsilon \int_{S_{in}}^{S_T} p_c^{(d)}GdS. \quad (5.62)
\]

In the last step, condition (5.50b) was used. Summing up the integrals gives

\[
\int_{\Omega^0} \text{div} \vec{R} < -\frac{1}{2ct}(p_c^{(d)}(S_T) - p_c^{(d)}(S_{in}))^2 - \varepsilon \int_{S_{in}}^{S_T} p_c^{(d)}GdS < 0.
\]

As before, this gives a contradiction. Hence we are left with Case (c).

Figure 5.17: The points \( S^\pm \) and \( S_T^n \) and the final turn of the orbit.
What remains is to show that the orbit goes at most finitely many times around $E_T$ before it reaches one of its points. Define, see Figure 5.17, $S^+ > S_T$ and $S^- < S_T$ such that

$$\hat{u}(S^-) + \frac{1}{\varepsilon} S^- = p_c^{(i)}(S_T) + \frac{1}{\varepsilon} S_T$$ and $$\hat{u}(S^+) + \frac{1}{\varepsilon} S^+ = p_c^{(d)}(S_T) + \frac{1}{\varepsilon} S_T,$$

If the orbit enters $\hat{H}$ in the $n$th turn at $S = S_n^-$ with $S_n^- \in (S^-,S_T)$, then this is the final turn of the orbit. Similarly, if the orbit enters $\hat{H}$ at the $n$th turn at $S = S_n^+$ with $S_n^+ \in (S_T,S^+)$, then this is the final turn. If no such $n$ exists, then the orbit must go into a limit cycle whose closed loop contains the shaded region in Figure 5.17. This clearly gives a contradiction.

Statements (ii) and (iii) of Theorem 5.2 follow from the nature of the eigenvalues, see also the proof of Lemma 5.2.

Conditions (5.50) are sufficient but not necessary to prove existence as described in Theorem 5.2. Below we give an (almost) explicit construction of a solution when (5.50) is violated.

**Example 3.1** Let $S_B \in (0,1)$ and $\varepsilon > 0$ be given, and let $p_c^{(i)}$ and $p_c^{(d)}$, $S_T \in (S_B,1)$ and $u_B^*$ be such that

$$\ell(S_T,u_B^*) = p_c^{(i)}(S_T),$$

where $\ell(S,u_B)$ is as given in (5.47), see Figure 5.12 and Figure 5.18.

![Figure 5.18: Construction of orbit when (5.50) is violated. Here the function $\hat{u}$ is convex in $(S_B, S_T)$ and for $\tau = \tau^*$ it touches $\ell(S,u^*)$ at $S = S^*$.](image)

Further assume $p_c^{(d)} \in C^2(0,1)$ and $p_c^{(d)}(S) > 0$ for $S_B < S < S_T$. For the relative permeability we consider the explicit case $k(S) = S^2$. A straightforward calculation
gives from (5.13) that \( \mathcal{G} \) satisfies
\[ \mathcal{G}'' > 0 \text{ in } [S_B, S_T], \]
provided \( S_T < 2S_B \). As a consequence,
\[ \dot{u}'' = p_c^{(d)''} + c\epsilon\tau\mathcal{G}'' > 0 \text{ in } (S_B, S_T), \tag{5.65} \]
if \( S_T < 2S_B \). Next we increase \( \tau \) from \( \tau = 0 \), when \( \dot{u} = p_c^{(d)} \), until \( \tau = \tau^* \) when \( \dot{u} \) touches the line \( \ell(S; u_B^*) \) at some \( S = S^* \). By the convexity we have
\[ \dot{u}' < -\frac{1}{\epsilon} \text{ for } S_B \leq S < S^*. \]

The monotonicity of the orbits in the shaded region in Figure 5.18 ensures the existence of orbit running from \( E_B = (S_B, u_B) \), with \( u_B \in [u_B^*, p_c^{(d)}(S_B)] \), to a corresponding point \( E_T \in \bar{E}_{T,i} \). The saturation and pressure profiles are monotone in this example.

If (5.50b) is satisfied, the orbit from \( E_B \) directly enters \( \mathcal{H} \), and does not feel the influence of the parameter \( \epsilon \). Hence, for this orbit, the condition of whether it remains to the left of \( S = 1 \) is the same as for play-type hysteresis, see Theorem 5.1. Moreover, as the orbits are ordered and cannot intersect when \( S < 1 \), the same applies to any orbit leaving \( E_B \). Only if \( S_T > S^*_T(S_B) \) (or \( a(S_B, S_T) < 0 \) and \( \tau \) is sufficiently large) can orbits reach \( S = 1 \). Here one has to be careful with condition (5.50) which requires a certain smallness of the product \( \epsilon^2 \tau \). Large values of \( \tau \) are allowed provided \( \epsilon \) is sufficiently small.

### 5.4 Numerical results

In this section we discuss a number of numerical PDE-solutions and compare the results to the theoretical travelling wave results. Clearly, in the numerical approach one works in a truncated domain: \(-W < x < W\) for some \( W > 0 \), chosen sufficiently large so that well-developed saturation profiles appear.

For the play-type model, the equations are (5.1) and (5.3), where we write the latter as \( \partial_t S = \frac{1}{\tau f(S)} \mathcal{F}(S, u) \), see expression (1.30). Combining the equations and using the semi-implicit time discretization scheme described in Chapters 2 and 4, we have
\[ S_n = S_{n-1} + \frac{\Delta t}{\tau f(S_{n-1})} \mathcal{F}(S_{n-1}, u_{n-1}), \tag{5.66a} \]
\[ \frac{1}{\tau f(S_n)} \mathcal{F}(S_n, u_n) + \partial_x (k(S_n)(\partial_x u_n + 1)) = 0. \tag{5.66b} \]
Here $\Delta t > 0$ is the time step and $n \in \{1, \ldots, N\}$ for some $N \in \mathbb{N}$. The pair $(S_n, u_n)$ approximates $(S, u)$ at the $n$th time step, i.e. at $t = n\Delta t$. The initial condition for the saturation at $n=0$ is a smooth approximation of the Riemann data

$$S_0 := \begin{cases} S_T & \text{if } x < 0, \\ S_B & \text{if } x > 0. \end{cases} \quad (5.67)$$

The boundary conditions are

$$k(S_n)(\partial_x u_n + 1) = k(S_T) \quad \text{at } x = -W, \quad (5.68a)$$
$$u_n = u_B \in [p_c^{(i)}(S_B), p_c^{(d)}(S_B)] \quad \text{at } x = W, \quad (5.68b)$$

for each $n \geq 1$. The first boundary condition implies that we prescribe a constant water flux $k(S_T)$ at $x = -W$. For well-developed profiles this means that approximately $S_n(-W) = S_T$ for $n$ sufficiently large. As for the second condition, we will see that independent of the choice of $u_B$, the profile develops into a travelling wave as predicted. Taking $u_B = p_c^{(i)}(S_B)$ gives a compatible boundary condition. Taking any $u_B \neq p_c^{(i)}(S_B)$, gives a boundary layer at $x = W$.

For the extended model, one needs to discretize the combination (5.1) and (5.5) in the domain $(-W, W)$. Here we propose the semi-implicit scheme

$$S_n = S_{n-1} + \Delta t \partial_x (k(S_{n-1})(\partial_x u_{n-1} + 1)), \quad (5.69a)$$
$$\mathcal{A}(S_n, u_n, \frac{u_n - u_{n-1}}{\Delta t}) + \partial_x (k(S_n)(\partial_x u_n + 1)) = 0, \quad (5.69b)$$

for $n \geq 1$. The function $\mathcal{A} : (0, 1) \times \mathbb{R}^2 \to \mathbb{R}$ is defined in (5.5). Initial and boundary conditions are again given by (5.67) and (5.68). To solve equations (5.69), an initial condition for the pressure is required as well. In the computations we have taken

$$u(x, 0) = \frac{1}{2} (p_c^{(i)}(S(x, 0)) + p_c^{(d)}(S(x, 0))) \text{ for } -W < x < W. \quad (5.70a)$$

To avoid a boundary layer at $x = W$, we take in (5.68b)

$$u_B = \frac{1}{2} (p_c^{(i)}(S_B) + p_c^{(d)}(S_B)). \quad (5.70b)$$

Note that for $x < 0$, the initial condition need not be compatible with the unique pressure $u_T \in E_{T,\widetilde{d}}$ that results from the traveling wave analysis with $u_B$ given by (5.70b). This pressure is approximately attained after a sufficiently large number of time steps.

Distinguishing the cases $\frac{f(S)}{\Delta t} \lesssim 1$, one straightforwardly verifies that $\mathcal{A}(S_n, u_n, u_n - u_{n-1}/\Delta t)$ strictly decreases with respect to $u_n$. Hence, for given $S_n, u_{n-1} \in L^2((-W, W))$, such that $k(S_n) \geq \delta$ for some $\delta > 0$, there exists a unique $u_n \in H^1((-W, W)$ satisfying
(5.69b) and (5.68), e.g. see [269, Chapter 2]. Equations (5.66b) and (5.69b) are solved numerically by means of the L-scheme [192, 251].

Finite differences are used to solve equations (5.66) and (5.69), with $\Delta x = 0.2$ and $\Delta t = 0.02$. The length of the domain $(2W)$ is chosen according to the problem. The nonlinearities $k$, $p_c^{(i)}$ and $p_c^{(d)}$ are as in Chapter 2 and we restrict ourselves to $f(S) = 1$ only. The extension of pressures $P_k$ ($k \in \{i, d\}$) is achieved by setting $P_k = 10^3 \cdot (1 - S)$ for $S > 1$.

The results for play-type hysteresis are shown in Figure 5.19. They agree well with the predictions. Fronts quickly develop into travelling waves with wave-speed $c$ predicted by (5.11). For $S_T < S_T^*(S_B)$ the profiles stay away from $S = 1$ and the behaviour with respect to $\tau$ is as to be expected. For $\tau < \tau_{c,i}$ the travelling wave orbit goes directly to $E_{T,i}$; for $\tau \in (\tau_{c,i}, \tau_{c,d})$ it goes around $E_{T,i}$ to reach $E_{T,d}$ and for $\tau > \tau_{c,d}$ it oscillates around $E_{T,i}E_{T,d}$, though infinitely many oscillations cannot be seen due to the finite domain size. For $S_T > S_T^*(S_B)$ and $\tau$ large enough, the profile indeed reaches $S = 1$ and a plateau is observed.

Figure 5.20 shows results for extended play-type hysteresis. The scanning curves within the hysteretic region are inclined by the fixed angle $\tan^{-1}(-\frac{1}{2})$, where $\varepsilon = \frac{1}{10}$ is used. Observe that, for all the $(\varepsilon, \tau)$ pairs used in Figure 5.20, conditions (5.50) are satisfied. In all cases, small irregularities are observed at the locations where orbits meet the segment $E_{T,i}E_{T,d}$. This reflects the behaviour in $\{x < 0\}$ as observed earlier in this section. These irregularities disappear when $n$ becomes large.
5.5 Conclusion

In this paper we studied the behaviour of infiltration fronts in a long homogeneous vertical unsaturated porous column. The main focus was on describing the fronts when non-equilibrium effects, such as capillary hysteresis and dynamic capillarity are considered. For this purpose, travelling wave approximation was used. Two models for hysteresis were considered, the play-type hysteresis and the extended play-type hysteresis model.

In case of play-type hysteresis model, it was shown that for small dynamic capillary constant \( \tau \), the fronts will be monotone. If \( \tau \) is larger than thresholds calculated from known parameters, the fronts will first develop finitely many and then infinitely many overshoots. Moreover, saturation plateaus will form at the location of undershoots and overshoots and the plateaus will become wider as one moves away from the wave-front. Criteria for the saturation overshoots to reach full saturation was also investigated. If the integral (5.35), which depends solely on the parameters of the system, is positive then the saturation cannot reach the maximum possible saturation value even if \( \tau \) is large. Similarly, if \( f \not\in L^1(0, 1) \), then the saturation cannot reach the maximum value.

To explain in more detail behaviour of wetting fronts observed from experiments, we also considered extended play-type hysteresis model which incorporates non-vertical scanning curves. The existence of travelling wave solutions were proved under conditions (5.50). One of the key differences in this case is that, if the initial saturation and the injection saturation are close enough then there will be no
overshoots irrespective of the value of \( \tau \). Another important distinction is that, in this case only finitely many oscillations can occur. This is also observed from experiments [35, 72], see also Figure 1.1(b).

Finally, a set of numerical tests were performed showing that the fronts indeed develop into travelling wave profiles for long columns. The predictions from our analysis were reproduced by the numerical results.
Chapter 6

Traveling wave solutions for the Richards equation with hysteresis

6.1 Introduction

When water is infiltrating a porous medium, the process can show a homogeneous saturation front or the formation of fingers; in the latter case, water enters preferentially along thin channels of high water saturation. In this chapter, we investigate a non-equilibrium Richards equation with hysteresis, a model that allows the formation of fingers. We are interested in the propagation of saturation profiles and restrict ourselves to the one-dimensional case. The main result of this contribution is that there exist, for appropriate parameter ranges, monotone traveling wave solutions for the system. In the case of a homogeneous front, the traveling wave describes the propagation of the front. In the case of fingering, the traveling wave describes the saturation profile in one finger. Our results predict the speed of the finger, but they also yield that, possibly after a transition time, no overshoot occurs in the finger tip, even in the case of a positive dynamic capillary term $\tau$.

Before we compare our results with the existing literature, we formulate the model problem. Spatial points of the one-dimensional system are denoted by $x \in \mathbb{R}$, time instances are denoted by $t \in [0, \infty)$. The physical state in $(x, t)$ is described by two variables: saturation $S$ and negative pressure $u$, i.e. $S, u : \mathbb{R} \times [0, \infty) \rightarrow \mathbb{R}$. Combining

---

Darcy law with mass conservation yields the unsaturated media flow equation
\[ \partial_t S + \partial_x [k(S)(\partial_x u + 1)] = 0, \]  
commonly referred to as the Richards equation. This specific normalized form has been studied in Chapters 2, 4 and 5. Gravity points in the positive \( x \)-direction and is also normalized. The hydraulic conductivity \( k \) is a given coefficient function \( S \mapsto k(S) \); it is a non-negative and non-decreasing function.

Equation (6.1) is complemented with a constitutional law that relates pressure and saturation. Models for slow processes without hysteresis use an algebraic relation: With a given capillary pressure function \( p_c \) one demands \( u = p_c(S) \). We can include non-equilibrium effects (\( \tau > 0 \)) and play-type hysteresis in the form
\[ u \in p_c^{(i)}(S) - \gamma(S) \cdot H(\partial_t S) - \tau \partial_t S. \]  
In this relation, \( (0, 1] \ni S \mapsto p_c^{(i)}(S) \in \mathbb{R} \) is a given decreasing function, the capillary pressure curve for imbibition. The function \( \gamma : (0, 1] \to [0, \infty) \) is a measure for the hysteretic width;
\[ \gamma(S) = p_c^{(d)}(S) - p_c^{(i)}(S), \]  
and \( H \) is the multivalued step function
\[ H(\zeta) := \begin{cases} 0 & \text{for } \zeta > 0, \\ [-1, 0] & \text{for } \zeta = 0, \\ -1 & \text{for } \zeta < 0. \end{cases} \]

The dynamic capillarity coefficient \( \tau \geq 0 \) is a measure for the redistribution time in the microscopic geometry. Typical plots of \( p_c^{(i)} \) and \( p_c^{(d)} \) are shown in Figure 1.3 and the scanning curves for the play-type hysteresis model are shown in Figure 2.1 (right). In the equilibrium case \( \tau = 0 \), the model imposes that imbibition (\( \partial_t S > 0 \)) occurs always with the imbibition capillary pressure \( u = p_c^{(i)}(S) \), while drainage (\( \partial_t S < 0 \)) occurs always with the drainage capillary pressure \( u = p_c^{(i)}(S) + (p_c^{(d)}(S) - p_c^{(i)}(S)) = p_c^{(d)}(S) \). Observe that, (6.2) is mathematically equivalent to (1.29) except that we have assumed a constant dynamic capillarity function, \( f \equiv 1 \), for the current analysis. The reason for using a different expression for (1.29) is stated later.

The above model (6.1)–(6.2) has received considerable attention in recent years. It was suggested and studied in [24, 112] as a natural extension of the Richards equation. It is thermodynamically consistent and includes the hysteresis effect with a play-type model. We note that Richards emphasizes in his famous article of 1931 the importance of hysteresis [212]. For an overview regarding hysteresis modelling and the development of the mathematical theory see [223] and Chapter 2.

Further research on (6.2)–(6.2) was mainly performed along three different lines:
A. Analysis of the hysteresis system in arbitrary dimension and fingering effect. It was shown that the above system is well-posed for constant $\gamma$ and positive permeability $k$ [149], that solutions are unique [54], that the system does not define an $L^1$-contraction (and hence allows for the fingering effect) [221], and that it indeed can produce the fingering effect, see e.g. [209]. Two-phase flow equations are treated in [53] and [141]. For a stability analysis for the system without hysteresis see e.g. [83,255]. In the broader context we mention [30] and [32] for uniqueness results and [163,166,224] for existence results.

B. Numerical methods. The numerical investigations of [119] and [237] are performed in order to detect the occurrence of saturation overshoots (a phenomenon that is described below in part C) and their dependence on the data of the problem (occurrence is meant in the sense that a travelling wave profile with overshoot exists). Saturation plateaus are recovered in [237]. In [216], a numerical scheme is suggested for a version of the above equations with non-vertical scanning curves. Indeed, introducing third order splines as scanning curves seems to facilitate the Newton iterations in a numerical scheme. A fixed point scheme was analysed in [136]. A numerical analysis of a discontinuous Galerkin scheme is performed in [135]. For a splitting scheme we mention [3] and for rigorous convergence proofs of finite volume schemes [36,51,52].

C. One-dimensional analysis and saturation overshoots. Much of the analysis in this field (including this contribution) is motivated by the effect of saturation overshoots: Experiments show that an imbibition front can be non-monotone. In this case, the saturation directly behind the front is higher than at larger distances behind the front [72]. Even though the experimental results come with large errors, one may interpret that the saturation forms plateaus. This is consistent with the shape of solutions to Riemann problems and is well understood in the case of two-phase flow equations [254]. It is also well understood that a positive parameter $\tau$ in (6.2) can produce non-monotone profiles, see [68,248] for systems without static hysteresis, and [169, 251] for the full system. The methods are based on a phase-space analysis of ordinary differential equations. For other rigorous proofs for the existence of traveling wave solutions in dynamic capillarity models see [67,228,236].

**Traveling waves**

In order to discuss traveling waves in further detail, we specify initial and boundary conditions for the above systems. Let $S_* \in (0,1)$ denote the initial saturation and $0 < F_0 \in \mathbb{R}$ prescribe an inflow at the top boundary $(x = -\infty)$. We then complement
(6.1)–(6.2) by
\[ S(x,0) = S_* \quad \text{for every } x \in \mathbb{R}, \]  
\[ S(x,t) \to S_* \quad \text{as } x \to +\infty \text{ for every } t \in (0,\infty), \]  
\[ k(S(x,t))(|\partial_x u(x,t) + 1| - F_0) \to 0 \quad \text{as } x \to -\infty \text{ for every } t \in (0,\infty). \]

Much of the one-dimensional analysis is concerned with traveling wave solutions. In the traveling wave ansatz, one seeks profile functions \( S, u : \mathbb{R} \to \mathbb{R} \) and a scalar velocity \( c > 0 \) such that the time-dependent functions have the special form
\[ S(x,t) := S(\xi), \quad u(x,t) := u(\xi) \text{ with } \xi = ct - x. \]  

Equations (6.1)–(6.2) provide the following system for \( S, u : \mathbb{R} \to \mathbb{R} \) and \( c > 0 \):
\[ cS' + [k(S)(u' - 1)]' = 0, \]  
\[ u \in p_c(S) - \gamma(S) \cdot H(S') - ctS', \]

where we used \( \text{sign}(cS') = \text{sign}(S') \) for positive \( c \in \mathbb{R} \) and referred \( p_c^{(i)}(S) \) simply as \( p_c \) for the sake of brevity. In our notation \( ' \) denotes the differentiation with respect to the argument. The boundary conditions are
\[ S(\xi) \to S_* \quad \text{as } \xi \to -\infty, \]  
\[ k(S(\xi))(-u'(\xi) + 1) \to F_0 \quad \text{as } \xi \to +\infty. \]

We are concerned with the case that drainage does not occur. This happens in the physical experiment when \( \gamma \) is large compared to the front width. Mathematically, it means that the difference \( u - p_c(S) + \tau cS' \) that appears in (6.6b) is everywhere smaller than \( \gamma(S) \). We may interpret this situation formally as \( \gamma \equiv +\infty \).

The mathematical model for this physical situation is simple and can be read off from (6.6b): The argument of \( H \) can never be negative. We replace (6.6b) by the two conditions:
\[ p_c(S(\xi)) - u(\xi) = \tau cS'(\xi) \quad \text{if } S'(\xi) > 0, \]  
\[ p_c(S(\xi)) - u(\xi) \leq \tau cS'(\xi) \quad \text{if } S'(\xi) = 0, \]

and we demand \( S'(\xi) \geq 0 \) for all \( \xi \in \mathbb{R} \). We emphasize that the case \( \gamma = +\infty \) is of physical relevance in the effect of gravity fingering, see the discussion in Section 6.4.

**Main result.** Our main result on the existence of traveling waves is derived under natural assumptions on the functions \( k \) and \( p_c \).

**Assumption 6.1.** Let the coefficient function \( k : [0,1] \to [0,\infty) \) be of class \( C^1 \), non-decreasing and strictly convex, positive on \( (0,1) \). Let the coefficient function \( p_c : (0,1) \to \mathbb{R} \) be of class \( C^1 \) with \( p'_c(S) < 0 \) for all \( S \in (0,1) \). We furthermore assume \( p_c(S) \to -\infty \) for \( S \nearrow 1 \).
6.2 Reformulation of the traveling wave equations

Remark 6.1. The assumption $\lim_{S \to 1} p_c(S) = -\infty$ is different from the assumption $p_c(1) = 0$ we have used in Chapters 2, 4 and 5. This is imposed to restrict the travelling wave orbits to the physically relevant regime $S \leq 1$. Alternatively, one can introduce the dynamic capillary function $f \in C((0,1))$ in (6.2) and (6.6b), see (1.29), such that $f > 0$ and $\int_{s}^{1} f(S) dS = \infty$. In Section 4.4.2 and Section 5.2.3 it is shown that this keeps the saturation bounded away from 1. However, for simplicity, in this chapter we assume $\lim_{S \to 1} p_c(S) = -\infty$ instead.

Essentially, our main result states that for every initial saturation $S* \in (0, 1)$ and for every small influx parameter $F_0 > 0$, there exists a monotone traveling wave solution $S, u: \mathbb{R} \to \mathbb{R}$ and $c > 0$. However, we must demand that the parameter $\tau$ exceeds a critical value.

Theorem 6.2. Let $k$ and $p_c$ be coefficient functions as in Assumption 6.1. Let $S_\ast \in (0, 1)$ and $F_0 > 0$ be boundary data of the problem. With $c_- = k'(S_\ast)$ we demand that the flux satisfies the bounds $k(S_\ast) + c_-(1 - S_\ast) < F_0 < k(1)$. We use $\tilde{S} \in (S_\ast, 1)$ defined by $k(\tilde{S}) = F_0$ and $\tilde{c} \in (c_-, c_+)$ (here $c_+ = (k(1) - k(S_\ast))/(1 - S_\ast)$) defined by $\tilde{c} = (k(\tilde{S}) - k(S_\ast))/(\tilde{S} - S_\ast)$. There exists a critical value $\tau_* = \tau_* (k, p_c, S_\ast, F_0) \geq 0$ such that, for every $\tau > \tau_*$, there exists a velocity $c \in (c_-, \tilde{c})$ and a traveling wave solution $(S, u)$ to the system (6.6a), (6.7), and (6.8).

6.2 Reformulation of the traveling wave equations

We will construct solutions to (6.6a), (6.7), and (6.8) in a special form: We obtain saturation profiles $S$ that are constant for $\xi > 0$ and pressure profiles $u$ that are affine for $\xi > 0$. The following lemma provides a simplified set of equations. Solutions to the simplified equations can be extended as indicated above to find solutions to the original problem.

Lemma 6.1 (Reduced traveling wave system). Let $F_0, S_\ast > 0$ be boundary data and let $S_0 > 0$ satisfy $k(S_0) \geq F_0$. For $c > 0$ let $(S, u): (-\infty, 0) \to \mathbb{R} \times \mathbb{R}$ be a classical solution of the ordinary differential equation

\begin{align*}
\frac{c(S - S_\ast)}{S} - k(S) u' &= k(S) - k(S_\ast) - k(S) u' \\
\frac{u}{c(S - S_\ast)} - \tau c S' &= p_c(S) - \tau \frac{c S'}{c(S - S_\ast)}
\end{align*}

on $(-\infty, 0)$, satisfying the conditions

\begin{align*}
S(\xi) &\to S_\ast \quad \text{as} \quad \xi \to -\infty, \\
S(0) &= S_0, \\
S'(0) &= 0, \\
c(S_0 - S_\ast) + k(S_\ast) &= F_0.
\end{align*}
160 Traveling wave solutions for the Richards equation with hysteresis

Figure 6.1: A numerically computed traveling wave solution. The figure shows the saturation profile as a solid line and the pressure profile ($-u$) as a dashed line. The solution is computed for $p_c(S) = S$, $k(S) = S^2$, $\tau = 5$, $S_* = 0.1$, $F_0 = 0.25$. We obtain numerically $c = 0.2139$ and $S_0 = 0.8013$.

We furthermore assume that the function $\xi \mapsto S(\xi)$ is monotonically increasing on $(-\infty, 0)$. If we extend $S$ and $u$ for $\xi > 0$ by setting

\[
S(\xi) = S_0, \quad \tag{6.11a}
\]

\[
u(\xi) = p_c(S_0) + \left(1 - \frac{F_0}{k(S_0)} \right) \xi, \quad \tag{6.11b}
\]

then the extended functions $(S, u): \mathbb{R} \to \mathbb{R} \times \mathbb{R}$ solve the system (6.6a), (6.7)–(6.8).

A numerically computed solution as in Lemma 6.1 is depicted in Figure 6.1.

Proof. (a) Equations on $(-\infty, 0)$. Differentiating (6.9a) yields

\[
c S' + [k(S)(u' - 1)]' = 0,
\]

on $(-\infty, 0)$, which is (6.6a). We demanded that $S$ is monotonically increasing on $(-\infty, 0)$, hence $S' \geq 0$ holds. The case $S'(\xi) = 0$ for some $\xi < 0$ can easily be excluded. Indeed, an increasing function $S$ with $S'(\xi) = 0$ satisfies also $S''(\xi) = 0$; in this case, (6.9) implies that $S$ and $u$ are constant to the right of $\xi$, in contradiction to the assumption that $S$ is monotonically increasing. We conclude that $S' > 0$ holds such that we are always in the first case of (6.8). Therefore (6.9b) implies (6.8) on $(-\infty, 0)$.

(b) Boundary conditions. Condition (6.10a) is equivalent to (6.7a). In order to verify (6.7b), we differentiate (6.11b). We obtain for any $\xi > 0$

\[
k(S(\xi))(-u'(\xi) + 1) = F_0,
\]
6.3 Existence of hysteretic traveling wave solutions

In view of Lemma 6.1, our aim in the following is to construct solutions to system (6.9)–(6.10). The equations (6.9) can be written as

\[
\begin{pmatrix}
S' \\
u'
\end{pmatrix} = \begin{pmatrix}
(p_c(S) - u)/(ct) \\
1 - (c(S - S_*) + k(S_*))/k(S)
\end{pmatrix}.
\]  

(6.13)

With the function

\[
\mathcal{G}(S) := \mathcal{G}(S; c, S_*) := 1 - \frac{c(S - S_*) + k(S_*)}{k(S)},
\]  

(6.14)

we can write the unknowns and the right hand side as

\[
y := \begin{pmatrix} S \\ p \end{pmatrix} : (-\infty, 0] \to \mathbb{R}^2, \quad G(y) := \begin{pmatrix} (p_c(S) - u)/(ct) \\ \mathcal{G}(S) \end{pmatrix}.
\]  

(6.15)

and the dynamical system (6.13) in the plane reads

\[
y' = G(y).
\]  

(6.16)

This is precisely the system analysed also in Section 4.4. We recall that the right hand side depends on \( k, c, S_* \) and \( \tau \). An example for the function \( \mathcal{G} \) is sketched.

since \( S(\xi) = S_0 \) by (6.11a). This yields (6.7b).

(c) Equations on \((0, \infty)\). The extensions of \( S \) and \( u \) in (6.11) are defined in such a way that \( S' \equiv 0 \) and \( -u' + 1 = F_0/k(S_0) \) holds on \((0, \infty)\). Since \( k(S) \) is constant, we see that both sides of equation (6.6a) vanish. We check that (6.8) holds on \((0, \infty)\):

\[
\text{There holds } S' \equiv 0, \text{ hence we are in the second case of (6.8). The pressure satisfies } u \geq p_c(S_0) \text{ by definition in (6.11b) and the assumption } F_0/k(S_0) \leq 1. \text{ By } S = S_0 \text{ we have verified (6.8).}
\]

(d) Equations in \(0\). We evaluate equation (6.9a) in \( \xi = 0 \) and insert (6.10b) to obtain

\[
c(S_0 - S_*) = k(S_0) - k(S_*) - k(S_0)u'(0).
\]  

Condition (6.10d) provides the value of the flux in \(0\),

\[
k(S_0)(-u'(0) + 1) = F_0.
\]  

(6.12)

Both, the saturation \( S \) and the flux quantity \( [k(S)(-u' + 1)] \) are continuous in \( \xi = 0 \).

In the following section we obtain our main result. We will show that, given \( F_0, S_* > 0 \), we find \( c > 0, S_0 > 0 \) with \( k(S_0) \geq F_0 \), and \( S \) solving (6.9)–(6.10).
Traveling wave solutions for the Richards equation with hysteresis

Figure 6.2: The graph of the function $G(S)$ for the choice $k(S) = S^2$, $S_* = 0.1$ and $c = 0.4$. We find a maximum with value $\min_{S \in (S_*, 1)} G(S) = -1/3$ and a second root $S^* = 0.3$.

in Figure 6.2, see also Figure 5.1 (right). The function always vanishes in $S_*$. Our assumptions will make sure that $G$ has a negative slope $G'(S_*)$ and is positive in $S = 1$.

Our aim is to find an orbit $y = (S, u): (-\infty, 0] \to \mathbb{R}^2$ satisfying (6.16) and the boundary conditions

$$y(\xi) \to (S_*, p_c(S_*)) \quad \text{for} \quad \xi \to -\infty,$$

$$u(0) = p_c(S(0)),$$

$$c(S(0) - S_*) + k(S_*) = F_0. \quad (6.17c)$$

We note that (6.17a) implies (6.10a), (6.17b) implies (6.10b) for $S_0 := S(0)$ and, by the first equations of (6.13), also (6.10c). Finally, (6.17c) is identical to (6.10d).

6.3.1 Construction of normalized orbits for given $c$

We introduce two bounds for the wave speed. A minimal speed $c_- > 0$ is given by the slope of $k$ in $S_*$. A maximal speed $c_+ > c_-$ is given by the constraint $S \leq 1$. More precisely, we set

$$c_- := k'(S_*), \quad c_+ := \frac{k(1) - k(S_*)}{1 - S_*}. \quad (6.18)$$

For every $c \in (c_-, c_+)$ we denote by $S^* = S^*(c) > S_*$ the second root of $G$: The real number $S^*$ is the solution of (see Figures 6.2 and 6.3)

$$c(S^* - S_*) + k(S_*) = k(S^*). \quad (6.19)$$

By Assumption 6.1 on $k$, there exists a unique solution $S^* \in (S_*, 1)$ to this equation. The number $S^*$ is the point at which the line with slope $c$ intersects the graph of
k, compare Figure 6.3. We emphasize that the value $S^* = S^*(c) \in [0,1]$ depends on $c \in (c_-, c_+)$. 

![Figure 6.3: The coefficient function $S \mapsto k(S)$ and the construction of the upper saturation value $S^* \in (S_*, 1)$.](image)

In the next result, we prescribe a velocity $c > 0$ and determine the flux $F_0$ in dependence of $c$. It will turn out that we must distinguish qualitatively different cases with a critical parameter $\tau_0$. A similar distinction appears also in [68, 248, 254] and in Chapters 4 and 5.

**Proposition 6.1** (Construction of orbits for given $c$). Let coefficient functions $k$ and $p_c$ be as in Assumption 6.1 and let $c \in (c_-, c_+)$ be fixed. Then there exists a critical value $\tau_0 = \tau_0(k, p_c, S_*, c) \geq 0$ such that the following holds:

1. For every $\tau > \tau_0$, there exists a solution $y = (S, u)$ to problem (6.13) that connects the stationary point $(S_*, p_c(S_*))$ with a point $(S_0, p_c(S_0))$, $S_0 \in (S_*, 1)$, which is reached for $x = 0$. In other words: The solution satisfies the boundary conditions (6.17a) and (6.17b).

2. For every $0 < \tau < \tau_0$, there exists a heteroclinic orbit $y : \mathbb{R} \to \mathbb{R}^2$ that connects the stationary point $(S_*, p_c(S_*))$ with the stationary point $(S^*, p_c(S^*))$.

Solutions of Item 1 satisfy $u(\xi) < p_c(S(\xi))$ and $S'(\xi) > 0$ for all $\xi \in (-\infty, 0)$, and $u'(0) > 0$.

With $S^* = S^*(c)$ of (6.19) and $S(0)$ the value of the solution of Item 1, we define the function

$$S_0 : \mathbb{R}_+ \times (c_-, c_+) \to [0,1], \quad S_0(\tau, c) := \begin{cases} S(0) & \text{for } \tau > \tau_0, \\ S^*(c) & \text{for } \tau \leq \tau_0. \end{cases} \quad (6.20)$$
The function $S_0$ is continuous. There holds

$$
\tau_0(k, p_c, S_*, c) \leq \tau_{\text{crit}}(c) := \frac{(p_c'(S_*(c)))^2}{4c G'(S_*(c))}.
$$

(6.21)

Proof. Step 1: The stationary point $(S_*, p_c(S_*))$. We observe that the point $y_* := (S_*, p_c(S_*))$ is a stationary point for (6.16): The property $G(y_*) = 0$ holds by $p_c(S_*) - p_c(S_* = 0$ and by $G(S_*) = 0$. In the following, we seek orbits that approach the stationary point $y_*$ for $\xi \to -\infty$.

The linearisation of the right-hand side of (6.16) in the point $y_* = (S_*, p_c(S_*))$ is given by the matrix

$$
\mathcal{A} := DG(y_*) = \begin{pmatrix}
  p_c'(S_*) / (c \tau) & -1 / (c \tau) \\
  G'(S_*) & 0
\end{pmatrix}.
$$

(6.22)

Because of $c > c_-$, there holds $G'(S_*) = k'(S_*) / k(S_*) - c / k(S_*) < 0$. This yields $\det \mathcal{A} < 0$, which implies that there is a stable and an unstable direction. The characteristic polynomial of $\mathcal{A}$ is

$$
\det(\mathcal{A} - \lambda \mathcal{I}) = \lambda^2 - \frac{p_c'(S_*)}{c \tau} \lambda + \frac{G'(S_*)}{c \tau},
$$

and the eigenvalues of $\mathcal{A}$ are

$$
\lambda_{\pm} = -\frac{p_c'(S_*)}{2c \tau} \left[ -1 \pm \sqrt{1 - \frac{4c \tau G'(S_*)}{p_c'(S_*)^2}} \right].
$$

(6.23)

The eigenvalues of $\mathcal{A}$ are real and of opposite sign, $\lambda_+ > 0$ and $\lambda_- < 0$. The Stable Manifold Theorem [185, Chapter 2.7] implies that there is a one-dimensional unstable manifold $U \subset \mathbb{R}^2$ for the point $y_* = (S_*, p_c(S_*))$ (we recall that the set $U$ is the collection of points $\begin{pmatrix} y_{1,0} \\ y_{2,0} \end{pmatrix}$ such that solutions $y$ of (6.16) with $y(0) = \begin{pmatrix} y_{1,0} \\ y_{2,0} \end{pmatrix}$ satisfy $y(\xi) \to y_*$ as $\xi \to -\infty$). The tangential space to $U$ at $y_* = (S_*, p_c(S_*))$ is

$$
T_{y_*} U = \text{Ker} \begin{pmatrix} p_c'(S_*) / (c \tau) - \lambda_+ & -1 / (c \tau) \\
G'(S_*) & -\lambda_+
\end{pmatrix} = \mathbb{R} \begin{pmatrix} 1 \\
-p_c'(S_*) - c \tau \lambda_+
\end{pmatrix} =: E_{\lambda_+}.
$$

(6.24)

The calculation provides that the slope of the tangential space is lesser than $p_c'(S_*)$. 

In the following we consider the branch $U_+$ of the unstable manifold $U$ that approaches the point $(S_*, u_*)$ $(u_*) = p_c(S_*)$ from the right; for some $\epsilon > 0$ holds:\n
$(S, u) \in U_+$ with $\| (S, u) - (S_*, u_*) \| < \epsilon$ implies $S > S_*$. We can parametrize the branch $U_+$ with an orbit $y : \xi \mapsto y(\xi) \in U_+$ that satisfies $y(\xi) \rightarrow (S_*, u_*)$ for $\xi \rightarrow -\infty$. One of our aims is to show for large $\tau > 0$: There is a parametrization of the orbit (a shift in $\xi$) such that the re-parametrized orbit ends on a point $(y_1, y_2)(0) = (S, u)(0)$ with $u(0) = p_c(S(0))$.

*Step 2: General properties of orbits.* Given $\tau$ and $c$, we fix an orbit $y$ that parametrizes $U_+$. We want to collect properties of this orbit.
In the phase space $\mathbb{R}^2$, we consider the unbounded domain

$$\Sigma := \{(S,u) \in \mathbb{R}^2 | S_0 < S < 1, u < p_c(S)\}. \quad (6.25)$$

For sufficiently small $\xi$ (negative and large in absolute value), there holds $y(\xi) \in \Sigma$. This is a consequence of the fact that the slope of the orbit is given by $E_{\lambda,}$, and it is smaller than $p'_c(S_0)$.

The orbit $y$ can never leave $\Sigma$ through the left boundary since $S'$ is positive on $\partial \Sigma \cap \{S = S_0\}$, see the first equation of (6.13).

We claim that the orbit is bounded, as long as it is contained in $\Sigma$. To prove the claim, we consider the quantity $\nu(\xi) := p_c(S(\xi)) - u(\xi)$, for which (6.13) provides the relation

$$\nu'(\xi) = p_c'(S(\xi)) \frac{\nu(\xi)}{c} - \varphi(S(\xi)). \quad (6.26)$$

We note that $\nu(\xi) \to 0$ holds as $\xi \to -\infty$. The ordinary differential equation (6.26) implies

$$\nu(\xi) \leq v_0 := \frac{\min_{S \in [S_0,1]} \varphi(S)}{\max_{S \in [S_0,1]} p_c'(S)} \quad (6.27)$$

for all $\xi \in \mathbb{R}$. Indeed, $\nu$ satisfies the relation for small $\xi$ (negative and large in absolute value) by positivity of the right hand side. Whenever $\nu$ exceeds the value $v_0$, there holds $\nu'(\xi) \leq 0$ by (6.26). This implies that the value cannot be exceeded.

We finally observe that the orbit moves to the right at all times: $S'(\xi) \geq 0$ for all $\xi$ with $y(\xi) \in \Sigma$. We collect our results on the orbit $y$: The orbit starts in $(S_0, u_0)$, or, more precisely: $y(\xi) \to (S_0, u_0)$ for $\xi \to -\infty$. The orbit travels in $\Sigma$, it does not diverge to infinity and it travels always to the right. We recall the property of $p_c$ that $p_c(S) \to -\infty$ for $S \to 1$. We conclude that there are only two possibilities:

1.) The orbit reaches the $p_c$-curve after finite time; in this case we can re-parametrize such that this time instance is zero. We find that $(S(0), u(0)) = (S_0, p_c(S_0))$ for some $S_0 \in (S_0, 1)$.

2.) The orbit approaches a stationary point. Since all stationary points lie on the $p_c$-curve, we find that $(S(\xi), u(\xi)) \to (S^*, p_c(S^*))$ as $\xi \to -\infty$ for some $S^* \in (S_0, 1)$.

Step 3: Investigation of case 2.) and definition of $\tau_0$. Let us further investigate case 2.). Since $(S^*, p_c(S^*))$ is a stationary point, $S^*$ is indeed the second zero of $G$. In this point holds $\varphi'(S^*) > 0$. The eigenvalues of the linearisation of the dynamical system can be calculated as above; we find

$$\lambda^*_\pm = \frac{p_c'(S^*)}{2c \tau} \left[ \pm \sqrt{1 - \frac{4c\varphi'(S^*)}{p_c'(S^*)^2}} \right].$$

We see that, for $\tau > \tau_{\text{crit}}(c) = (p_c'(S^*))^2 / (4c\varphi'(S^*))$, the eigenvalues are non-real. This implies that the point $(S^*, p_c(S^*))$ is a spiral sink for $\tau > \tau_{\text{crit}}(c)$: orbits are going in...
spiral into the stationary point. Since we defined the orbit \( y \) such that it is contained in \( \Sigma \), we conclude that \( \tau > \tau_{\text{crit}}(c) \) excludes case 2.)

We define
\[
\tau_0(c) := \inf \{ \tau > 0 : \text{the orbit is as in case 1.} \}, \tag{6.28}
\]
and note that \( 0 \leq \tau_0(c) \leq \tau_{\text{crit}}(c) \) holds. In fact, from Proposition 5.1 one further concludes that \( \tau_0(c) > 0 \).

**Step 4: Investigation of case 1.** In case 1.), there holds \( S(0) > S^* \). Indeed, for all \( \xi \) with \( S(\xi) \in S^* \), there holds \( u'(\xi) = G(S(\xi)) \leq 0 \). On the curve \( u = p_c(S) \) holds \( S'(\xi) = 0 \) and, for \( S < S^* \), \( u'(\xi) < 0 \). This implies that the orbit cannot hit the lower boundary of \( \Sigma \) for \( S < S^* \).

Let us assume that \( \tau > 0 \) is such that the orbit \( y \) is as in case 1.). Then, for every value \( \hat{\tau} > \tau \), the corresponding orbit \( \hat{y} \) is also as in case 1.). This can be concluded as follows: For a larger value of \( \tau \), the slope of \( E_{\hat{1},c} \) is more negative, see (6.24) and (6.23). Furthermore, the orbit for \( \tau \) (denoted with \( (S,u) \)) and the orbit for \( \hat{\tau} \) (denoted with \( (\hat{S},\hat{u}) \)) can never intersect in \( S \in (S_*,S^*) \): For two points \( \xi \) and \( \hat{\xi} \) with \( S(\xi) = \hat{S}(\hat{\xi}) < S^* \) and \( u(\xi) = \hat{u}(\hat{\xi}) \), there holds \( u'(\xi) = \hat{u}'(\hat{\xi}) = 0 \). This implies that the orbit marked with hats remains always below the other orbit for \( S \in (S_*,S^*) \). In particular, it cannot approach the point \( (S^*,p_c(S^*)) \), it is also an orbit as described in case 1.).

We conclude from these considerations: For every \( \tau > \tau_0(c) \), orbits are as described in case 1.), for every \( \tau < \tau_0(c) \), orbits are as described in case 2.).

**Step 5: Properties of solutions and continuity of \( S_0 \).** The properties of solutions as in case 1.) are clear from the construction: \( u(\xi) < p_c(S(\xi)) \) and \( S'(\xi) > 0 \) for all \( \xi \in (-\infty,0) \). The inequality \( S(0) > S^* \) implies \( u'(0) = G(S(0)) > G(S^*) = 0 \). The upper bound for \( \tau_0 \) in (6.21) was already observed after (6.28).

It remains to verify the continuity of \( S_0 \). In points \( (\tau,c) \) with \( \tau \neq \tau_0(c) \) the continuity is clear by continuous dependence of solutions to ordinary differential equations. Let us consider a sequence of parameters \( (\tau_p,c_p) \to (\tau_0(c),c) \) as \( N \to \infty \), \( c \in (c_-,c_+) \), \( \tau_0(c) > 0 \). Without loss of generality, we can assume \( \tau_p > \tau_0(c_p) \), since otherwise \( S_0(\tau_p,c_p) = \{ 0 \} \), and the function \( S^* \) is continuous.

For every \( (\tau_p,c_p) \), the corresponding orbit satisfies \( (S_{(p)}(0),u_{(p)}(0)) = (S_0^p,p_c(S_0^p)) \). It remains to show \( S_0^p \to S^* \). Let us assume the contrary. Then we find a subsequence (not relabeled) with \( S_0^p \to S^\infty > S^* \). In this situation, the orbits \( (S_{(p)}(0),u_{(p)}(0)) \) converge to a limiting orbit that reaches \( S^\infty \) for \( \xi = 0 \). We conclude that, for \( \tau = \tau_0(c) \), there is an orbit as in Item 1.), reaching \( S^\infty > S^* \) after finite time. Since the stable manifold (and hence the orbit) depends continuously on the parameter \( \tau \) (compare [185, Chapter 2.7]), small perturbations of \( \tau \) can be performed and we find orbits as in Item 1.) for \( \tau \) slightly smaller than \( \tau_0(c) \). This contradicts Item 2.)
6.3.2 Construction of $c$ to satisfy the flux condition

We now change the unknowns of the system, since the wave-speed $c$ is a-priori unknown. The task of physical relevance is the following: Given an initial saturation $S_*$ and a flux boundary value $F_0$, we seek for the appropriate traveling wave speed $c$ to satisfy (6.17c).

The setting of the physical experiment makes it reasonable to assume $F_0 > k(S_*)$: The inflow rate at the top boundary must exceed the flux at $x = \infty$ which is induced by gravity. On the other hand, inflow is lower than that for full saturation, hence we will use $F_0 < k(1)$. We will actually strengthen these assumptions below.

For the subsequent construction it is useful to define further quantities: We define $\bar{S} \in (S_*, 1)$ and $\check{c} \in (c_-, c_+)$ by

$$k(\bar{S}) = F_0, \quad \check{c} = \frac{k(\bar{S}) - k(S_*)}{\bar{S} - S_*}. \quad (6.29)$$

The number $\check{c}$ is defined such that $S^*(\check{c}) = \bar{S}$ and hence $k(S^*(\check{c})) = k(\bar{S}) = F_0$.

**Lemma 6.2** (Construction of the wave speed). Let $k$ and $p_c$ be as in Assumption 6.1 and $S_* \in (0, 1)$. We use $c_-$ from (6.18) and define $F_- := k(S_*) + c_-(1 - S_*)$. We use $\check{c}$ from (6.29). Then, for every $F_0 \in (F_-, k(1))$, there exists a critical value $\tau_* = \tau_*(S_*, F_0) > 0$ such that the following holds: For every $\tau > \tau_*$, there exists a wave speed $c \in (c_-, \check{c})$ such that the orbit $(S, u)$ of Proposition 6.1 satisfies all boundary conditions (6.17).

**Proof. Step 1: The function $Q$.** We introduce the function $Q : \mathbb{R}_+ \times (c_-, c_+) \to \mathbb{R}$,

$$Q(\tau, c) := c (S_0(\tau, c) - S_*) + k(S_*) - F_0. \quad (6.30)$$

Loosely speaking, zeroes of $Q$ are parameters $(\tau, c)$ with the following property: the corresponding orbit satisfies the boundary condition (6.17c) with the prescribed value of $F_0$. In order to find roots of $Q$, we will use the intermediate value theorem for continuous functions.

The argument is slightly more involved since we have to make sure that the orbit is ending in a point $(S_0, p_c(S_0))$ at $\xi = 0$. To be precise, we look for $\tau > 0$ and $c \in (c_-, \check{c})$ such that

$$Q(\tau, c) = 0 \quad \text{and} \quad S_0(\tau, c) > S^*(c). \quad (6.31)$$

Let us assume that $(\tau, c)$ satisfies (6.31). Because of $S_0(\tau, c) > S^*(c)$, the parameters are such that we are in the first case of Proposition 6.1. Therefore, there exists an orbit $(S, u)$ with $(S, u)(\xi) \to (S_*, u_*)$ for $\xi \to -\infty$ and $(S, u)(0) = (S_0, p_c(S_0))$ for some $S_0 \in (S_*, 1)$. In particular, the orbit satisfies (6.17a) and (6.17b). In order to check (6.17c), we calculate, using first the definition of $Q$ and then (6.31),

$$c(S(0) - S_*) + k(S_*) = Q(\tau, c) + F_0 = F_0.$$
6.3 Existence of hysteretic traveling wave solutions

This verifies (6.17c).

Step 2: Solution of problem (6.31).

Step 2a: Sign of $Q$ for $c = \tilde{c}$. We calculate the sign of $Q$ for parameters $(\tau, \tilde{c})$ with $\tau > \tau_0(\tilde{c})$. The defining relation (6.29) implies $S^*(\tilde{c}) = \tilde{S}$. Since the orbit ends at a point $S_0(\tau, \tilde{c}) > S^*(\tilde{c}) = \tilde{S}$, we find

$$Q(\tau, \tilde{c}) = \tilde{c}(S_0(\tau, \tilde{c}) - S_*) + k(S_*) - F_0$$
$$> \tilde{c}(\tilde{S} - S_*) + k(S_*) - F_0$$
$$= k(\tilde{S}) - k(S_*) + k(S_*) - F_0 = k(\tilde{S}) - F_0 = 0.$$

We see that, for large $\tau > 0$, the zeroes of $Q$ are not at the boundary $|c = \tilde{c}|$.

Step 2b: Sign of $Q$ for $c \approx c_\pm$. For a small number $\epsilon > 0$, we now calculate the sign of $Q$ for parameters $(\tau, c)$ with $c \leq c_- + \epsilon$. We calculate

$$Q(\tau, c) = c(S_0(\tau, c) - S_*) + k(S_*) - F_0$$
$$< c(1 - S_*) + k(S_*) - F_0$$
$$\leq (c_- + \epsilon)(1 - S_*) + k(S_*) - F_0$$
$$\leq F_- + \epsilon(1 - S_*) - F_0 < 0,$$

where the last inequality is satisfied for all sufficiently small values of $\epsilon > 0$ by the choice of $F_-$ and the assumption $F_0 > F_-$. We see that zeroes of $Q$ are not near the boundary $|c = c_-|$.

For $0 < \epsilon < (F_0 - F_-)/(1 - S_*)$ and with $\tau_{\text{crit}}(c)$ of (6.21), we set

$$\tau_* := \sup[\tau_{\text{crit}}(c) | c \in [c_- + \epsilon, \tilde{c}]].$$

(6.32)

Note that $\tau_{\text{crit}}(c)$ tends to infinity for $c$ tending to $c_-$. It is therefore important that we define $\tau_*$ as above with $\epsilon > 0$. With our definition we achieve, by continuity of $\tau_{\text{crit}}$, the finiteness $\tau_* < \infty$.

Let $\tau > \tau_*$ be arbitrary. Since $Q(\tau, .) : (c_-, \tilde{c}] \rightarrow \mathbb{R}$ depends continuously on $c$, we can use the intermediate value theorem. The function is negative for small values of $c$ and positive at the right boundary. We conclude the existence of $c \in (c_-, \tilde{c})$ with $Q(\tau, c) = 0$. Because of $\tau > \tau_*$, the second condition of (6.31) is also satisfied.

Lemma 6.2 imposes for the flux $F_0$ the lower bound $F_0 > F_- := k(S_*) + c_- (1 - S_*)$. With a more involved argument, the lower bound can be improved to $F_0 > F_- := k(S_*)$.

The existence result of Lemma 6.2 for a wave speed $c > 0$ allows to show our main result, formulated in Theorem 6.2.
Proof of Theorem 6.2. The assumptions of the theorem imply that Proposition 6.1 and Lemma 6.2 can be applied. Thus, for any $F_0$ as in the theorem, there exist functions $S,u$ satisfying the dynamical system (6.13) with boundary conditions (6.17). The functions $S,u$ solve also the equivalent problem (6.9)–(6.10).

We want to apply Lemma 6.1. The monotonicity assumption on $S$ is satisfied since $S' > 0$ holds on $(-\infty,0)$ by Proposition 6.1. Proposition 6.1 also provides $u'(0) > 0$ and we can calculate, using (6.13) and inserting (6.17c) into (6.14): $0 < u'(0) = \theta(S(0)) = 1 - (F_0/k(S(0))).$ Hence, $F_0 < k(S_0)$ is satisfied and Lemma 6.1 can be applied. The lemma implies that $S,u$ can be extended to a traveling wave solution.

6.4 Discussion

We have found monotone traveling wave solutions for the Richards equation with play-type hysteresis for a positive dynamic term ($\tau > 0$). This result is surprising since models without hysteresis (and also regularised hysteresis models) usually show saturation overshoots for $\tau > 0$.

Figure 6.5: A solution of the two-dimensional system corresponding to (6.1)–(6.2) with play-type hysteresis and $\tau > 0$, for the setting see [209]. In this result, the fingers are of moderate length and the drainage process in the fingers has not set in. This justifies the assumption that only imbibition and scanning curves are relevant.

Previous results of related systems [68,169,237,248,251,254] provide traveling wave solutions with saturation overshoot. The overshoot is obtained when either the play-type hysteresis is regularised or when the pressure oscillations are large enough to see both, the imbibition and the drainage part of the hysteresis loop. In the work at hand we studied a situation in which the drainage curve is not reached by the traveling wave profile. Instead, the profile only explores the imbibition curve and a scanning curve.
In our setting, monotone saturation profiles can occur — at least for sufficiently large \( \tau \). Our results show the existence of a critical value \( \tau_* \geq 0 \) such that, for \( \tau > \tau_* \), monotone traveling waves occur. We may interpret these solutions as the known “plateau-shaped” solutions that have been observed before; in our setting, these solutions are truncated in the plateau and extended with a constant saturation profile.

We have obtained traveling waves with the help of the simplified system (6.9). This system can be used to understand the formation of fingers in two space dimensions, see Figure 6.1. We interpret \( x \) as the vertical direction and consider a slice through one finger. The saturation profile along the slice can be expected to be a one-dimensional traveling wave solution to the system studied here. With this interpretation, our system allows to predict the speed \( c \) of the finger and the constant saturation \( S_0 \) in the finger, away from the finger tip.
Chapter 7

Fronts in two-phase porous media flow problems: the effects of hysteresis and dynamic capillarity

7.1 Introduction

This chapter generalises the results of Chapters 4 to 6 to the two-phase case and incorporates dynamic capillarity and hysteresis in both capillary pressure and relative permeability. The premise of the chapters mentioned above was the following setting: water is injected at a constant rate from the top of a homogeneous vertical porous column, whose length along the vertical axis is much greater than its horizontal width, see Figure 7.1. Using travelling wave (TW) analysis we studied the propagation of the imbibition fronts in the unsaturated flow regime. The Richards equation (Ri), as given in (5.1), was used for the modelling and non-equilibrium effects were included in the capillary pressure relationship, see Section 1.2. This includes capillary hysteresis and dynamic capillary effects, a detailed description of which can be found in Section 1.2.1 and Section 1.2.2 respectively. In particular, Chapter 4 analysed the influence of hysteresis and dynamic capillarity separately, whereas, Chapter 5 investigated the combined effect of hysteresis and dynamic capillarity both for the play-type hysteresis model (5.2) and the extended play-type hys-

---

The travelling wave solutions approximate the saturation and pressure profiles in infiltration experiments that have similar setting to Figure 7.1 [35, 72, 229]. Thus, they can be used to explain phenomenon such as occurrence of overshoots and development of saturation plateaus [72, 229], see Figure 1.1(b). For the unsaturated case, TW solutions with dynamic effect were first analysed in [68]. The results were improved in Chapters 4 and 5 where apart from proving the existence of TW solutions, criteria for non-monotonicity and reaching full saturation were stated. For the two-phase case it was shown rigorously in [236, 248, 254] that non-monotone travelling waves and non-standard entropy solutions are existing if one includes dynamic capillary effect. Similar analysis but for higher order viscous terms containing spatial derivatives were performed in [21, 84]. It is evidenced in [216] that hysteresis can explain stable saturation plateaus but it cannot initiate them. In [23, 209] it is shown how both hysteresis and dynamic capillarity are required to explain the growth of viscous fingers. The entropy conditions for Buckley-Leverett equation considering hysteresis in only permeability were derived in [22, 190, 215]. However, hysteresis and nonlinearities were not included in the viscous term. This is taken into consideration in [2] where the authors add a dynamic term to model permeability hysteresis, while disregarding hysteresis and dynamic effects in capillary pressure. The behaviour of TWs for a non-monotone flux function in the presence of a third order term was described in [228].

In our current work we build upon [68, 169, 248, 251, 254] to describe the behaviour of fronts in the two-phase case when dynamic capillarity and both type of hysteresis are included. The models that are used in our analysis are introduced in Section 7.2. Section 7.3 discusses the existence of TWs when the flux function is monotone and hysteresis and dynamic effects are included in the capillary pressure but not in the permeabilities. Entropy conditions are derived and they reveal that there can be non-classical shocks. In Section 7.4, the analysis is extended to include hysteresis in the permeabilities for small contributions of the dynamic capillary term. This makes self developing saturation plateaus and a broader class of entropy solutions possible. Section 7.5 presents numerical results that support our analytical findings. Finally, we make some concluding remarks in Section 7.6.

### 7.2 Mathematical model

This section is dedicated to the formulation of a mathematical model that can be used to describe an imbibition process of a fluid into a homogeneous porous column.
7.2 Mathematical model

7.2.1 Governing equations

Here we consider the one-dimensional situation where the flow problem is defined on an interval \((0, H)\). This simplification is justified by the fact that the walls of the porous column, in which the fluids are injected are impermeable and that saturation is in general almost constant across the section area of the column. The axis is pointing in the direction of gravity. The medium is assumed to be homogeneous and the fluids are incompressible. Under these constraint, for \(t\) and \(x\) denoting the time and space variables respectively, we rewrite the Buckley-Leverett equation (BL) from Section 1.1.3 as

\[
\partial_t S = \partial_x \left( f Q + N_g h + N_c h \partial_x p_c \right). \tag{7.1}
\]

Here, \(S \in [0,1]\) is the effective water saturation; \(p_c\) is the capillary pressure \(p_c := p_n - p_w\); \(Q\) is the total flux defined in (1.15) and \(f, h\) are fractional flow functions.

![Figure 7.1: Setup of an infiltration experiment. At the inlet of a column having the height \(H\) water is injected by a constant rate. The main axis of the column is orientated such that it is aligned with the gravity vector.](image)
defined as
\[ f := \frac{k_{rw} - \mu_w k_{rn}}{k_{rw} + \mu_w k_{rn}}, \quad h := \frac{k_{rw} k_{rn}}{k_{rw} + \mu_w k_{rn}} = k_{rn} f, \] (7.2)

see (1.17). The subscripts \( \alpha \in \{n, w\} \) represent non-wetting and wetting phases. Moreover, \( \mu_\alpha \) is the viscosity and \( k_{r\alpha} \) is the relative permeability of phase \( \alpha \).

As the medium is homogeneous, we have taken the scaled porosity \( \phi \) and the absolute permeability \( \kappa \) to be 1 in (BL) (see (1.16)). The capillary number (\( N_c \)) and the gravity number (\( N_g \)) are calculated from (1.11) by putting \( L_{ref} = H \). Furthermore, as there are no source terms, from (1.14) we get that \( Q \) is constant in space for a fixed \( t > 0 \). In addition to that, we assume that \( Q \) is also constant in time, which occurs, e.g. if a constant influx (injection rate) is prescribed at the inlet \( x = 0 \). This implies that \( Q \) is a dimensionless constant and thus, by choosing \( v_{ref} \) properly in (1.11) the following holds
\[ Q(x, t) \equiv 1 \text{ for } x \in (0, H) \text{ and } t > 0. \] (7.3)

From here, dropping the subscript \( c \) in \( p_c \) we write (7.1) as
\[ \partial_t S = \partial_x (f + N_g h + N_c h \partial_x p). \] (7.4)

Note that, \( f, h \), and \( F \) are functions of \( S \) and possibly of \( p \), as shown below.

### 7.2.2 Modelling hysteresis and dynamic capillarity

To incorporate hysteresis and dynamic capillarity in the model, one needs to extend capillary pressure and relative permeability given in the closure relationship (1.4).

**Capillary pressure**

For a detailed description of capillary hysteresis, we refer to Section 1.2.1 and Chapter 2, particularly to Figure 1.3. Dynamic capillarity is described in Section 1.2.2. To include play-type hysteresis and dynamic capillarity in the capillary pressure, we use (1.29), which in this context translates to
\[ p \in \frac{1}{2} p_c^{(i)} (S) + p_c^{(d)} (S) - \frac{1}{2} (p_c^{(d)} (S) - p_c^{(i)} (S)) \cdot \text{sign} (\partial_t S) - \tau \partial_t S, \] (7.5)

where \( p_c^{(i)}, p_c^{(d)} \) are respectively the imbibition and drainage capillary pressures and \( \text{sign}(\cdot) \) denotes the multi-valued signum graph defined in (1.25). The second and third term in the right hand side of (7.5) describe, respectively, capillary hysteresis [24] and dynamic capillarity [112]. Further, \( \tau \geq 0 \) denotes the dynamic capillary coefficient. It models relaxation or damping in the capillary pressure. Although in practice \( \tau \) may depend on \( S \) [35, 47], here we assume it to be constant. The case of non-constant \( \tau \) is considered in Chapters 4 and 5. The capillary pressure functions \( p_c^{(j)}, j \in \{i, d\} \), fulfill [20, 115, 170]:

Here, and later in this paper, a prime denotes differentiation with respect to the argument. In the absence of dynamic effects, i.e. \( \tau = 0 \), expression (7.5) implies

\[
p = \begin{cases} 
p_c^{(i)}(S) & \text{when } \partial_t S > 0, \\
p_c^{(d)}(S) & \text{when } \partial_t S < 0.
\end{cases}
\]

When \( \partial_t S = 0 \), \( p \) is between \( p_c^{(i)}(S) \) and \( p_c^{(d)}(S) \). For this reason, the hysteresis described by (7.5) is called play-type hysteresis: i.e. the scanning curves between \( p_c^{(i)}(S) \) and \( p_c^{(d)}(S) \) are vertical.

Before discussing the case \( \tau > 0 \), we introduce for convenience the sets

\[
\mathcal{H}^{(i)} := \{(S, p) : S \in (0,1), \ p < p_c^{(i)}(S)\},
\]

(7.6)

\[
\mathcal{H}^{(d)} := \{(S, p) : S \in (0,1), \ p > p_c^{(d)}(S)\},
\]

(7.7)

\[
\mathcal{H} := \{(S, p) : S \in (0,1), \ p_c^{(i)}(S) \leq p \leq p_c^{(d)}(S)\},
\]

(7.8)

and the strip \( \mathcal{W} = \mathcal{H}^{(i)} \cup \mathcal{H} \cup \mathcal{H}^{(d)} = [0 < S \leq 1] \). In (5) it is shown that pressure expression (7.5) can be written as

\[
\frac{\partial S}{\partial t} = \frac{1}{\tau} \mathcal{F}(S, p) := \frac{1}{\tau} \begin{cases} 
p_c^{(d)}(S) - p & \text{if } (S, p) \in \mathcal{H}^{(d)}, \\
0 & \text{if } (S, p) \in \mathcal{H}, \\
p_c^{(i)}(S) - p & \text{if } (S, p) \in \mathcal{H}^{(i)}.
\end{cases}
\]

(7.9)

Relative permeability

An overview of relative permeability hysteresis can be found in Section 1.2.1, see Figure 1.4 in particular. To make the effect of hysteresis explicit we need to incorporate a dependence of both \( S \) and \( \partial_t S \) in the relative permeabilities. This dependence should satisfy

\[
k_{\alpha \alpha}(S, \partial_t S) = \begin{cases} 
k_c^{(i)}(S) & \text{if } \partial_t S > 0, \\
k_c^{(d)}(S) & \text{if } \partial_t S < 0,
\end{cases}
\]

(7.10)

Here \( k_{\alpha \alpha}, k_c^{(i)}, k_c^{(d)} : [0,1] \rightarrow \mathbb{R} \) are the imbibition and drainage relative permeabilities obtained from experiments [38, 102, 137, 197, 243]. In line with the experimental outcomes, we assume here for \( j \in \{i,d\} \),

(P2) \( k_r^{(j)} \in C^2([0,1]), k_r^{(j)}(S) > 0 \) for \( 0 < S \leq 1 \), \( k_r^{(j)}(0) = 0 \) and \( k_r^{(j)} \) is strictly convex. Moreover, for \( 0 < S < 1 \), \( k_r^{(i)}(S) < k_r^{(d)}(S) \).
(P3) \( k^{(j)}_{\alpha} \in C^2([0,1]) \), \( k^{(j)}_{\alpha}(S) < 0 \) for \( 0 \leq S < 1 \), \( k^{(j)}_{\alpha}(1) = 0 \) and \( k_{\alpha} \) is strictly convex.

Moreover, for \( 0 < S < 1 \), \( k^{(d)}_{\alpha}(S) < k^{(i)}_{\alpha}(S) \).

Note the reverse ordering in \( k_{\alpha} \) and \( k_{\alpha} \) when switching from imbibition to drainage. This is demonstrated experimentally in [38,102,243], see also Figure 1.4.

In [272], a play-type approach has been proposed to model \( k_{\alpha} \) where

\[
\bar{k}_{\alpha} = \begin{cases} 
  k^{(d)}_{\alpha}(S) & \text{if } (S, p) \in \mathcal{H}^{(d)}, \\
  \bar{k}_{\alpha}^{(i)}(S) & \text{if } (S, p) \in \mathcal{H}^{(i)}, \\
  \bar{k}_{\alpha}(S) & \text{if } (S, p) \in \mathcal{H}^{(i)}. 
\end{cases}
\]

Here \( \bar{k}_{\alpha} : \mathcal{H} \to [0,\infty) \) is a given function that satisfies

(P4) \( \bar{k}_{\alpha} \in C^2(\mathcal{H}) \) such that \( k_{\alpha} \in C(\mathcal{H}) \) for \( \alpha \in \{w,n\} \) and \( \partial_p \bar{k}_{\alpha} > 0 \), \( \partial_p \bar{k}_{\alpha} < 0 \) in \( \mathcal{H} \).

Observe that, this implies \( \bar{k}_{\alpha}(S, p^{(j)}_C(S)) = k^{(j)}_{\alpha}(S) \) for \( j \in \{i,d\} \). For the moment we leave the choice of \( \bar{k}_{\alpha} \) unspecified, except for properties (P4), as it neither influences the entropy conditions nor the critical \( \tau \) values introduced afterwards.

**Remark 7.1.** In the computations one needs to specify an expression for \( \bar{k}_{\alpha} \). In Section 7.5 we use

\[
\bar{k}_{\alpha}(S, p) = k^{(i)}_{\alpha}(S) + (k^{(d)}_{\alpha}(S) - k^{(i)}_{\alpha}(S)) \left( \frac{p - p^{(i)}_C(S)}{p^{(d)}_C(S) - p^{(i)}_C(S)} \right).
\]

This expression is obtained by considering sign in (7.5) and (7.11) as a function that can be eliminated. With \( \tau = 0 \) in (7.5), this results in (7.13). Since the fraction (7.13) is bounded by 0 and 1 for \( (S,p) \in \mathcal{H} \), we have \( \lim_{S \to 0} \bar{k}_{\alpha}(S,p) = k^{(i)}_{\alpha}(0) = k^{(d)}_{\alpha}(0) \) and similar for \( S \to 1 \).

Observe that (7.12) is consistent with (7.10) as from (7.9), \( \partial_t S > 0 \iff p < p^{(i)}_C(S) \) and \( \partial_t S < 0 \iff p > p^{(i)}_C(S) \). Moreover, the scanning curves for \( k_{\alpha} \) have constant \( S \).

Although not true in general, see for instance Figure 1.4, we restrict ourselves to play-type for both \( p \) and \( k_{\alpha} \). An extension describing non-vertical scanning curves is discussed in Chapter 5.
7.2 Mathematical model

Using (7.12) and (7.2), the nonlinearities \( f \) and \( h \) are expressed in terms of \( S \) and \( p \) as well:

\[
\xi(S, p) = \begin{cases} 
\xi^{(d)}(S) & \text{if } (S, p) \in \mathcal{H}^{(d)}, \\
\xi(S, p) & \text{if } (S, p) \in \mathcal{H}, \\
\xi^{(i)}(S) & \text{if } (S, p) \in \mathcal{H}^{(i)}, 
\end{cases}
\]

From (P2)-(P4) we deduce for \( f \) and \( h \):

(P5)  \( f \in C(\mathcal{W}), \ f \in C^2(\mathcal{H}) \) and \( \partial_p \hat{f} > 0 \) in \( \mathcal{H} \). For \( j \in \{i, d\} \), \( f^{(j)} \in C^2([0,1]), f^{(j)'}(S) > 0 \) for \( 0 < S < 1 \), \( f^{(j)}(0) = 0 \), \( f^{(j)}(1) = 1 \). Moreover, for \( 0 < S < 1 \), \( f^{(i)}(S) < f^{(d)}(S) \).

(P6) \( h \in C(\mathcal{W}), h \in C^2(\mathcal{H}), h^{(j)} \in C^2([0,1]) \) and \( h^{(j)}(0) = h^{(j)}(1) = 0 \) for \( j \in \{i, d\} \).

Observe that, in general no ordering holds between \( h^{(i)} \) and \( h^{(d)} \). Typical curves for \( f^{(i)} \) and \( h^{(i)} \) are shown in Figure 7.2. The equations (7.4), (7.5) and (7.14) are the complete set of equations for our model.

7.2.3 Effective equations

Observe that, in Section 1.1.3 we did not prescribe a scaling for \( \tau \). Assuming (7.5) to be relating dimensional quantities and proceeding as in Section 1.1.3 one obtains

\[
p \in \left[ \frac{1}{2} (p^{(i)}_c(S) + p^{(d)}_c(S)) - \frac{1}{2} (p^{(d)}_c(S) - p^{(i)}_c(S)) \right] \cdot \text{sign}(\partial_t S) - \frac{\tau_{\text{ref}}}{p_{\text{ref}} t_{\text{ref}}} \partial_t S. \tag{7.15}
\]

Here \( p_{\text{ref}}, t_{\text{ref}} \) are the reference values of pressure and time as introduced in Section 1.1.3; \( t_{\text{ref}} \) is the reference value of dynamic capillary coefficient \( \tau \) which is rescaled as \( \tau := \tau / t_{\text{ref}} \). Now choosing \( t_{\text{ref}} = N_c p_{\text{ref}} t_{\text{ref}} = p_{\text{ref}}^2 \phi_{\text{ref}} K_{\text{ref}} (1 - S_{wr} - S_{nr}) \), the Leverett scaling [153] for \( p_{\text{ref}} \) (see (1.10)) gives

\[
\tau_{\text{ref}} = \frac{\sigma^2 \phi_{\text{ref}}^2}{\mu_n v_{\text{ref}}^2} (1 - S_{wr} - S_{nr}). \tag{7.16}
\]

This choice leaves us with a characteristic dynamic coefficient \( \tau \) that is independent of the length scale of the problem. This is precisely the scaling used in [105,248,254] that is consistent with the hyperbolic limit. Realistic values of dimensional and scaled quantities are given in [161].

Thus, we are left with the dimensionless system

\[
(\mathcal{P}) \quad \begin{cases} 
\partial_t S + \partial_x (F(S, p) + N_c h(S, p) \partial_x p) = 0, \\
\partial_t S = \frac{1}{N_c \tau} \mathcal{F}(S, p),
\end{cases} \tag{7.17}
\]
where \( F = f + N_g h \).

(7.17c)

This system can be seen as a regularisation of the hyperbolic Buckley-Leverett equation with gravity. Here the regularisation involves hysteresis and dynamic capillarity. Compared to the usual second order parabolic regularisation, yielding shocks that satisfy the standard Oleinik conditions [180], different (non-parabolic) regularisations may yield shocks that violate these conditions, see e.g. [150,254]. Such shocks are called non-classical.

One of the main issues of this paper is to show the existence of non-classical shocks originating from System \((\mathcal{P})\). To this end we proceed as in [254] and study the existence of travelling wave (TW) solutions of \((\mathcal{P})\) that connect a left state \( S_B \) to a right state \( S_T \) in the presence of both hysteresis and dynamic capillarity. Travelling waves for the model with only dynamic capillarity are analysed in [236, 248]. For the case of unsaturated flow, i.e. Richards equation with a convex flux function, existence and qualitative properties of travelling waves are considered in detail in [68,169,251].

For the purpose of travelling waves we consider System \((\mathcal{P})\) in the domain \(-\infty < x < \infty\). Then the capillary number \( N_c \) can be removed from the problem by the scaling

\[
x := x/N_c \quad \text{and} \quad t := t/N_c.
\]

This yields the \( N_c \) independent formulation

\[
(\tilde{\mathcal{P}}) \begin{cases} 
\partial_t S + \partial_x \left( F(S,p) + h(S,p) \partial_x p \right) = 0, \\
\partial_t S = \frac{1}{\tau} F(S,p),
\end{cases}
\]

(7.18a)

with \(-\infty < x < \infty \) and \( t > 0 \). This is the starting point for the TW analysis.

**Remark 7.2.** Using the Brooks-Corey type expression, e.g. see [45],

\[ k_{rw}(S) = S^q \quad \text{and} \quad k_{rn}(S) = (1-S)^q, \]

(7.19)

with \( q = 2 \), the nonlinearities (7.2) and (7.17c) become

\[
f(S) = \frac{S^2}{S^2 + M(1-S)^2}, \quad h(S) = (1-S)^2 f(S), \quad F(S) = S^2 \frac{1 + N_g(1-S)^2}{S^2 + M(1-S)^2},
\]

where \( M = \frac{\mu_w}{\mu_n} \) denotes the viscosity ratio. A plot is shown in Figure 7.2. Some elementary calculations give

(a) **Monotonicity:** If \( N_g \leq M \) then \( F'(S) > 0 \) for all \( 0 < S < 1 \) and if \( N_g > M \) then there exists a unique \( S_F \in (0,1) \) such that \( F'(S) > 0 \) for all \( 0 < S < S_F \) and \( F'(S) < 0 \) for \( S_F < S < 1 \). Since \( F(1) = 1 \), clearly \( F(S_F) > 1 \).
(b) **Inflection points:** $f(S)$ has only one inflection point in $(0, 1)$ whereas, $F(S)$ has at most two. To see this for $f(S)$, note that $f''(S) = P(S)Q(S)$ with $Q(S)$ being a positive function and $P(S) = M - (3M + 3)S^2 + (2M + 2)S^3$. Since $P(0) = M$, $P(1) = -1$ and $P'(S) < 0$ for $S \in (0, 1)$, the result follows.

These properties of $f$ and $F$ will be used when discussing the different cases of travelling waves.

![Figure 7.2: The functions $f(S)$, $h(S)$ and $F(S)$ as given in Remark 7.2. Here $M = 2$ and $N_g = 4.$](image)

### 7.2.4 Travelling wave formulation

Having derived the non-dimensional hysteretic two-phase flow System $(\hat{\mathcal{S}})$, we investigate under which conditions travelling wave solutions exist. These are solutions of the form

$$S(x, t) = S(\xi), \quad p(x, t) = p(\xi), \quad \text{with} \quad \xi = ct - x,$$

where $S$ and $p$ are the wave profiles of saturation and pressure and $c \in \mathbb{R}$ the wave-speed. We seek travelling waves that satisfy

$$\begin{align*}
\lim_{\xi \to -\infty} S(\xi) &= S_B, \quad \lim_{\xi \to -\infty} \frac{dS}{d\xi} = S_T, \\
\lim_{\xi \to -\infty} \frac{dp}{d\xi} &= \lim_{\xi \to -\infty} \frac{dp}{d\xi} = 0,
\end{align*}$$

(7.20)

where $S_B$ corresponds to an initial saturation and $S_T$ to the injected saturation. The choice of $p'(\pm \infty) = 0$ ensures that the diffusive flux vanishes at $\xi = \pm \infty$. Substituting (7.20) into (7.18a) and (7.18b), and integrating (7.18a) one obtains

$$\begin{align*}
cS - (F(S, p) - h(S, p)p') &= A, \\
cS' &= \frac{1}{r} \mathcal{D}(S, p),
\end{align*}$$

(7.21a, 7.21b)
where $\xi \in \mathbb{R}$ and $A$ is a constant of integration.

As was shown in [251] for the Richards equation, (7.20) and (7.21) do not automatically guarantee the existence of $\lim_{\xi \to \pm \infty} p(\xi)$. But if $p(\pm \infty)$ is well-defined then (7.21b) and the existence of $S(\pm \infty)$ forces $\lim_{\xi \to \pm \infty} \mathcal{F}(S(\xi), p(\xi)) = 0$. Recalling that $\mathcal{F}(S, p) = 0$ iff $(S, p) \in \mathcal{H}$ we then have

$$\lim_{\xi \to -\infty} p(\xi) = p_B \in [p_c^{(f)}(S_B), p_c^{(d)}(S_B)] \quad \text{and} \quad \lim_{\xi \to \infty} p(\xi) = p_T \in [p_c^{(i)}(S_T), p_c^{(d)}(S_T)].$$

We show later that $p_B$, interpreted as the initial pressure, can sometimes be chosen independently, whereas, $p_T$, when existing, is always fixed by the choice of $S_B, S_T$ and $p_B$. Following the steps in Chapters 4 and 5 we obtain the Rankine-Hugoniot condition for wave-speed $c$, i.e.

$$c = \frac{F(S_T, p_T) - F(S_B, p_B)}{S_T - S_B}. \quad (7.22)$$

With this, System (7.21) can be rewritten as a dynamical system,

$$\begin{align*}
(TW) \quad & S' = \frac{1}{cT} \mathcal{F}(S, p), \quad (7.23a) \\
& p' = \mathcal{G}(S, p). \quad (7.23b)
\end{align*}$$

where

$$\mathcal{G}(S, p) := \frac{F(S, p) - \ell(S)}{h(S, p)} \quad \text{with} \quad \ell(S) = F(S_B, p_B) + c(S - S_B). \quad (7.24)$$

Note that when $F$ is non-monotone (e.g. $N_g > M$ in Remark 7.2), the wave-speed $c$ can be positive or negative depending on the values of $S_B$ and $S_T$.

We study all possible solutions of System (TW) for $\tau > 0$. They serve as viscous profiles of admissible shocks of the limiting Buckley-Leverett equation. Existence conditions for solutions of (TW) act as admissibility/entropy conditions for the corresponding shocks.

The solutions of (TW) are investigated under two different scenarios.

A: No hysteresis in relative permeabilities, i.e. $\zeta^{(i)} = \zeta^{(d)}$ for $\zeta \in \{f, h\}$. Furthermore, $N_g$ is sufficiently small so that $F$ satisfies properties stated for $f^{(j)}$ in (P5). For $F$ as in Remark 7.2 this is satisfied if $N_g \leq M$.

B: $N_g$ and $\tau$ sufficiently small; relative permeabilities are hysteretic.

A third scenario where $N_g$ is large so that $F$ is non-monotone is discussed briefly at the end of Section 7.3.
7.3 No relative permeability hysteresis and small $N_g$ (Scenario A)

In the absence of relative permeability hysteresis, the functions $f$, $h$, $F$ and $G$ depend on $S$ only. We explicitly state the properties of $F$ as a result of (P5), (P6) and Remark 7.2.

(A1) $F \in C^2([0,1])$, $F'(S) > 0$ for $0 < S < 1$, $F(0) = 0$, $F(1) = 1$. Moreover, a unique $S_o \in (0,1)$ exists such that $F''(S_o) = 0$, $F''(S) > 0$ for $0 < S < S_o$ and $F''(S) < 0$ for $S_o < S < 1$.

Figure 7.3: (left) The saturations $S_B$, $S$, $\bar{S}$, $\tilde{S}$, $\alpha$ and $\beta$ for Scenario A. (right) The functions $\beta$, $\gamma$ and $\bar{\gamma}$ (assuming (7.31)) and the definitions of $S_*$ and $S^*$ for $S_B < \bar{S}$.

7.3.1 Preliminaries

Throughout this paper we restrict ourselves to relatively small values of $S_B$. Specifically, we assume

$$0 < S_B < S_o.$$  

First let us take $S_B \leq \bar{S}$, where $\bar{S}$ is the saturation at which $F'(\bar{S}) = \frac{1-F(\bar{S})}{1-S_B}$. The convex-concave behaviour of $F$ implies $S < S_o$. For later purpose, and with reference to Figure 7.3 (left), we introduce the additional saturations $S_B < \widetilde{S} < S < 1$, where $\widetilde{S}$ is the saturation at which $F(S)$ intersects the line connecting $(S_B, F(S_B))$ and $(1,1)$, and where $\bar{S}$ is the saturation for which $F'(\bar{S}) = \frac{F(\bar{S})-F(S_B)}{S-S_B}$. Then to each $\alpha \in [\bar{S},1]$ corresponds a unique $\beta \in [\bar{S},1]$ such that $(\beta, F(\beta))$ is the third intersection point between
the graph of $F$ and the chord through $(S_B, F(S_B))$ and $(α, F(α))$, see Figure 7.3 (left). This defines the function

$$β : [\tilde{S}, 1] \rightarrow [\tilde{S}, 1], \quad β(\tilde{S}) = \tilde{S}, \quad β(1) = 1,$$

$$β(α) = α \text{ and } β(α) \text{ is strictly decreasing.} \quad (7.26)$$

In Scenario A we restrict our discussions to $α \in [\tilde{S}, \bar{S}]$ giving $β \in [\bar{S}, 1]$.

Later in this section a second function $γ = γ(α)$ is introduced as one of the roots of the equation

$$\int_{S_B}^{γ(α)} G(S; S_B, α) dS = 0 \text{ for } S_B ≤ α ≤ \bar{S}. \quad (7.27)$$

Here $G(S; S_B, α)$ is the expanded notation of $G$ from (7.24) for the $p$ independent case:

$$G(S; S_B, α) = \frac{F(S) - ℓ(S; S_B, α)}{h(S)} \text{ with } ℓ(S; S_B, α) = F(S_B) + \frac{F(α) - F(S_B)}{α - S_B} (S - S_B).$$

A typical sketch of $G(S; S_B, α)$ for different values of $α$ is shown in Figure 7.4. Note that

$$G(S; S_B, α) \text{ decreases with respect to } α \in [S_B, \tilde{S}] \text{ and }$$

$$G(S; S_B, α) = \begin{cases} < 0 & \text{for } S_B < S < α \\ > 0 & \text{for } S > α \end{cases} \quad \text{when } S_B < α < \tilde{S},$$

$$G(S; S_B, α) = \begin{cases} < 0 & \text{for } \beta(α) < S < α \\ > 0 & \text{for } S_B < S < \beta(α) \\ < 0 & \text{for } α < S < \beta(α) \end{cases} \quad \text{when } \tilde{S} < α < \bar{S}. \quad (7.29)$$

Since,

$$G(S; S_B, α) = \begin{cases} G(1) \left(\frac{1}{k_B} S\right) & \text{when } α ≠ \tilde{S}, \\ G(1) \left(\frac{1}{k_B} \bar{S}\right) & \text{when } α = \tilde{S}, \end{cases} \quad (7.30)$$

as $S \nearrow 1$, we have for most practical applications

$$G(S; S_B, α) \text{ is non-integrable near } S = 1 \text{ for each } S_B ≤ α ≤ \tilde{S}. \quad (7.31)$$

This is the case for Brooks-Corey permeabilities with $q ≥ 2$, see Remark 7.2.
Returning to equation (7.27), we note that \( \gamma = S_B \) is the trivial solution. Properties (7.29) and (7.30) imply the existence of a second (non-trivial) solution \( \gamma = \gamma(a) \) for \( a \geq S_B \). It satisfies \( \gamma(S_B) = S_B \), \( \gamma(a) \) increases, \( \gamma(a) > a \) for \( a > S_B \). Moreover, if (7.31) is satisfied then \( \gamma(\bar{S}) < 1 \). This shows the existence of \( \gamma(a) \) in a right neighbourhood of \( S = \bar{S} \). The solution in this case exists up to \( a = S_* \in (\bar{S}, \hat{S}) \) where \( \gamma(a) \) and \( \beta(a) \) intersect: \( \gamma(S_*) = \beta(S_*) =: \gamma_* \). Further, if (7.31) holds, then a third solution \( \gamma = \hat{\gamma} \) exists for \( \bar{S} < a < S_* \). It decreases in \( a \) with \( \hat{\gamma}(\bar{S}) = 1 \) and \( \hat{\gamma}(S_*) = \gamma_* \). When (7.31) is not satisfied, the existence of \( S_* \) and a third solution depends on the specific form of \( k_{rn}(S) \). The solutions of (7.27) and the function \( \beta(a) \) are sketched in Figure 7.3 (right).

For \( S_B \in (\bar{S}, S_o) \), \( \beta(a) \) and \( \gamma(a) \) can similarly be defined, although the domain where \( \beta(a) \) is defined is different. In this case the intersection of \( \beta(a) \) and the second solution \( \gamma(a) \) is guaranteed irrespective of (7.31) since \( \int_{S_B}^{\hat{S}} \varrho(S; S_B, \bar{S}) dS < 0 \) and \( \int_{S_B}^{\hat{S}} \varrho(S; S_B, 1) dS > 0 \). This is because \( \varrho(S; S_B, \bar{S}) < 0 \) for \( S_B < \bar{S} < \hat{S} \) and \( \varrho(S; S_B, 1) > 0 \) for \( S_B < \bar{S} < 1 \). Since we use the second solution \( \gamma = \gamma(a) \) only, we summarize its properties in the following proposition.

**Proposition 7.1.** Assume either (7.31) or \( S_B \in (\bar{S}, S_o) \). Let \( \gamma \) be the increasing (unique) solution of (7.27). Then it is defined in the interval \([S_B, S_*]\) where \( S_* \in (\bar{S}, \hat{S}) \) is such that \( \gamma(S_*) = \beta(S_*) =: \gamma_* \). Further, \( \gamma(S_B) = S_B \), \( \gamma(a) > a \) for \( a > S_B \) and \( \gamma(a) < \beta(a) \) for \( a < S_* \) in the common domain of definition of \( \beta \) and \( \gamma \).

**Remark 7.3.** For simplicity, we assume (7.31) for the rest of the discussion. This guarantees the existence of a \((S_*', S_*)\) pair. The methods presented in this paper can also be applied to analyse the case when \( \beta(a) \) and \( \gamma(a) \) are not intersecting. The results are briefly discussed in Section 7.3.2.
Since (TW) is autonomous, it is convenient to represent solutions as orbits in the $(S,p)$-plane, or rather, in the strip $\{(S,p) : 0 \leq S \leq 1, p \in \mathbb{R}\}$. Moreover, orbits are same for any shift in the independent variable $\xi$. Therefore we may set without loss of generality, see Chapters 4 and 5,

$$S(0) = \frac{1}{2}(S_B + S_T). \quad (7.33)$$

Equilibrium points of (TW) are

$$E^i_K \equiv (K, p^{(j)}_c(K)), \text{ where } K \in \{S_B, S_T\} \text{ and } j \in \{i, d\}. \quad (7.34a)$$

If $\tilde{S} \leq S_T < \tilde{S}$, a third pair exists for $K = \beta(S_T)$. The points $E^i_K$ and the direction of the orbits are indicated in Figure 7.5. By the special nature of the function $F$, we have in fact that all points of the segments $E^i_K E^d_K$ are equilibrium points. Boundary conditions (7.20) are satisfied if an orbit connects the segments $E^i_{S_B} E^d_{S_B}$ and $E^i_{S_T} E^d_{S_T}$.

As shown in Chapter 5, an orbit can leave $E^i_{S_B}$ only from the lowest point $E^i_{S_B}$. Then it enters region $H^{(i)}$ where it moves monotonically with respect to $S$ as a consequence of the sign in the right hand side of equation (7.23a): if $p < p^{(i)}_c(S)$ we have $S' > 0$.

Due to this monotonicity one can alternatively describe an orbit leaving $E^i_{S_B}$ as a function of the saturation as long as it belongs to $H^{(i)}$. For given $\tau > 0$ and $S_T$ satisfying (7.32), let $w(S) = w(S; \tau, S_T)$ denote this function. Then

$$w(S_B; \tau, S_T) = p^{(i)}_c(S_B) \quad (7.34a)$$

and

$$w(S; \tau, S_T) < p^{(i)}_c(S) \text{ in a right neighbourhood of } S_B. \quad (7.34b)$$
As in Chapters 4 and 5, we deduce from (TW) that \( w \) should satisfy

\[
w'(S; \tau, S_T) = \frac{ct\phi(S; S_B, S_T)}{p_c^{(i)}(S) - w(S; \tau, S_T)} \quad \text{for } S > S_B.
\]

(7.35)

Using techniques from Chapters 4 and 5, one can show that initial value problem (7.35), (7.34a) has a unique local solution \( w(S; \tau, S_T) \) that satisfies (7.34b).

Remark 7.4. Conversely one recovers the orbit \((S(\xi), p(\xi))\) by substituting \( w \) into (7.23a). Using (7.33) this gives

\[
\xi = cT \int_{S(B + S_T)}^{S(\xi)} \frac{d\rho}{p_c^{(i)}(\rho) - w(\rho; \tau, S_T)} d\rho \text{ and } p(\xi) = w(S(\xi); \tau, S_T).
\]

Rewriting (7.35) as

\[
(p_c^{(i)}(w - p_c^{(i)}))' + (p_c^{(i)} - w)p_c^{(i)}' = ct\phi(S; S_B, S_T),
\]

we find recalling (P1) that

\[
((p_c^{(i)} - w)^2)' = 2(p_c^{(i)} - w)p_c^{(i)} - 2ct\phi(S; S_B, S_T) \leq -2ct\phi(S; S_B, S_T) \quad \text{in } \{w < p_c^{(i)}\}.
\]

(7.36)

Integrating this inequality from \( S_B \) to \( S \) gives the lower bound

\[
w(S; \tau, S_T) > p_c^{(i)}(S) - \sqrt{2ct\phi(S)} \quad \text{in } \{w < p_c^{(i)}\},
\]

where \( \Phi(S) = \Phi(S; S_B, S_T) := -\int_{S_B}^{S} \phi(\rho; S_B, S_T) d\rho \).}

(7.37)

With \( S_T \) satisfying (7.32), properties (7.29)-(7.31) and Proposition 7.1 imply

\[
\begin{align*}
\Phi(S) > 0 & \text{ for } S_B < S < \gamma(S_T) \\
\Phi(S_B) = \Phi(\gamma(S_T)) = 0 & \text{ when } S_B < S_T \leq S_*, \\
\Phi(S) > 0 & \text{ for } S_B < S < 1 \\
\lim_{S \to 1} \Phi(S) = +\infty & \text{ when } S_* < S_T \leq \bar{S}.
\end{align*}
\]

(7.38a)

\[
\begin{align*}
\Phi(S) > 0 & \text{ for } S_B < S < \gamma(S_T) \\
\Phi(S_B) = \Phi(\gamma(S_T)) = 0 & \text{ when } S_B < S_T \leq S_*, \\
\Phi(S) > 0 & \text{ for } S_B < S < 1 \\
\lim_{S \to 1} \Phi(S) = +\infty & \text{ when } S_* < S_T \leq \bar{S}.
\end{align*}
\]

(7.38b)

Observe that, depending on \( S_B, S_T \) and \( \tau \), the interval where \( w(S) < p_c^{(i)}(S) \) is either \((S_B,1)\) if \( w(S) \) and \( p_c^{(i)}(S) \) do not intersect, or \((S_B,S_i)\) with \( S_i \leq 1 \) in case there is an intersection at \( S = S_i \). In the latter case, it follows immediately from (7.36) that we must have,

**Proposition 7.2.** Suppose there exists \( S_i \in (S_B,1) \) such that \( w(S) < p_c^{(i)}(S) \) for \( S_B < S < S_i \) and \( w(S_i) = p_c^{(i)}(S_i) \). Then \( \phi(S_i; S_B, S_T) \geq 0 \).
Hence, if the orbit exits through the capillary pressure curve $p_c^{(i)}$, it can only do so at points where $\mathcal{G} \geq 0$. From the discussion above, one defines

$$S_m(\tau, S_T) = \sup\{S \in (S_B, 1) : w(z; \tau, S_T) < p_c^{(i)}(z) \text{ for all } S_B < z < S\}, \quad (7.39)$$

which is the upper limit of the interval on which $w$ exists. Then we have

**Proposition 7.3.**

(a) If $S_B < S_T \leq S_*$, then $S_T \leq S_m(\tau, S_T) < \gamma(S_T)$ for all $\tau > 0$;

(b) If $S_* < S_T \leq S$ and $w(\beta(S_T); \tau, S_T) < p_c^{(i)}(\beta(S_T))$, then $S_m(\tau, S_T) = 1$ and $\lim_{S/1} w(S) = -\infty$.

**Proof.** (a) The lower bound follows from Proposition 7.2. To show the upper bound, observe that if $S_m(\tau, S_T) \geq \gamma(S_T)$, then $w(\gamma(S_T)) \leq p_c^{(i)}(\gamma(S_T))$. This directly contradicts the strict inequality in (7.37) since $\Phi(\gamma(S_T)) = 0$.

(b) Since $\mathcal{G}(S, S_B, S_T) < 0$ in $(\beta(S_T), 1)$, Proposition 7.2 and (7.37), (7.38a) imply $S_m(\tau, S_T) = 1$. Since (7.37) holds for all $S < 1$ and since $w' < 0$ in a left neighbourhood of $S = 1$, let us suppose that $\lim_{S/1} w(S; \tau, S_T) = -L$ $(L > 0)$. Then equation (7.35) and property (7.31) give $w' \notin L^1$ near $S = 1$, contradicting the boundedness of $w$. \hfill $\square$

In Figure 7.6 we sketch the behaviour of $w(S; \tau, S_T)$ in $\mathcal{H}^{(i)}$. The existence of orbits as in Figure 7.6 (left) is a direct consequence of the behaviour of the lower bound (7.37). Orbits as in Figure 7.6 (right) need more attention since the case $S_m(\tau, S_T) < \beta(S_T)$, represented by $\tau_3$, remains to be discussed. We make the behaviour as sketched in Figure 7.6 (right) precise in a number of steps.

![Figure 7.6: Sketch of orbits represented by $w(S; \tau, S_T)$. (left) $S_{T,1} \in (S_B, S_*)$, $\tau_1 < \tau_2$; (right) $S_{T,2} \in (S_*, S)$, $\tau_3 < \tau_4$.](image)

We start with the following

**Remark 7.5.** In the context of this section the wave-speed (7.22) reduces to

$$c = \frac{F(S_T) - F(S_B)}{S_T - S_B}.$$
7.3 No relative permeability hysteresis and small $N_g$ (Scenario A)

From assumption (A1) it follows that there is a one-to-one correspondence between $c$ and $S_T \in [S_B, \bar{S}]$. Writing $c = c(S_T)$, we have $c(S_B) = F'(S_B)$, $c(\bar{S}) = F'(\bar{S})$ and $\frac{dc}{dS_T} > 0$ in $[S_B, \bar{S}]$.

Next, we give a general monotonicity result.

**Proposition 7.4 (Monotonicity).** Let $S_B$ satisfy (7.25).

(a) For a fixed $S_T \in (S_B, \bar{S})$ and any pair $0 < \tau_1 < \tau_2$, 
\[ w(\cdot; \tau_2, S_T) < w(\cdot; \tau_1, S_T) \text{ in } \{w(\cdot; \tau_1, S_T) < p_c^{(i)}(\cdot)\} \]

and 
\[ S_m(\tau_1, S_T) < S_m(\tau_2, S_T) \text{ if } S_T < S_m(\tau_2, S_T) \leq \beta(S_T); \]

(b) For fixed $\tau > 0$ and any pair $S_B < S_{T,1} < S_{T,2} \leq \bar{S}$, 
\[ w(\cdot; \tau, S_{T,2}) < w(\cdot; \tau, S_{T,1}) \text{ in } \{w(\cdot; \tau, S_{T,1}) < p_c^{(i)}(\cdot)\} \]

and  
\[ S_m(\tau, S_{T,1}) < S_m(\tau, S_{T,2}) \text{ if } S_{T,2} \leq S_m(\tau, S_{T,2}) \leq \beta(S_{T,2}). \]

**Proof.** We argue as in [248, 254]. The key idea is to introduce the function 
\[ u = \frac{(p_c^{(i)} - w)}{\sqrt{d}} \text{ with } d = c(S_T) \tau. \]

Using (7.35) one obtains for $u$ the equation 
\[ u'(S; \tau, S_T) = \frac{1}{\sqrt{d}} p_c^{(i)'(S)} - \frac{\mathcal{G}(S; S_B, S_T)}{u(S; \tau, S_T)}. \]

Clearly, $u|_{S_B} = 0$ and $u > 0$ in a right neighbourhood of $S_B$. Since $\mathcal{G}|_{S_B} = 0$ as well, one finds from (7.41) and the sign of $u$ 
\[ u'(S_B; \tau, S_T) = -\frac{p_c^{(i)'(S_B)}}{2} \left[ \frac{1}{d} - 4 \frac{\mathcal{G}'(S_B; S_B, S_T)}{(p_c^{(i)}(S_B))^2} - \frac{1}{\sqrt{d}} \right] > 0 \]

since $\mathcal{G}'(S_B; S_B, S_T) < 0$. Using (7.28), Remark 7.5 and some elementary algebra 
\[ u'(S_B; \tau_1, S_T) < u'(S_B; \tau_2, S_T) \text{ in case (a),} \quad (7.42a) \]
\[ u'(S_B; \tau, S_{T,1}) < u'(S_B; \tau, S_{T,2}) \text{ in case (b).} \quad (7.42b) \]

(a) From (7.42a), $u_1(\cdot) := u(\cdot; \tau_1, S_T) < u(\cdot; \tau_2, S_T) =: u_2(\cdot)$ in a right neighbourhood of $S_B$. We claim that $u_1$ and $u_2$ do not intersect in $[u_1 > 0]$. Suppose, to the contrary,
There exists $S_i > S_B$ such that $u_1(S) < u_2(S)$ for $S_B < S < S_i$ and $u_1(S_i) = u_2(S_i)$. Thus $u_1'(S_i) \geq u_2'(S_i)$. Evaluating (7.41) at $S_i$ gives

$$u_1'(S_i) = \frac{p_c^{(i)}(S_i)}{\sqrt{d_1}} - \frac{\beta(S_i; S_B, S_{T,1})}{u_1(S_i)} < \frac{p_c^{(i)}(S_i)}{\sqrt{d_2}} - \frac{\beta(S_i; S_B, S_{T,2})}{u_2(S_i)} = u_2'(S_i),$$

a contradiction.

If $S_T < S_m(\tau_2, S_T) \leq \beta(S_T)$, the $u$-monotonicity gives $S_m(\tau_1, S_T) \leq S_m(\tau_2, S_T)$. We rule out the equality by contradiction. Suppose $S_m(\tau_1, S_T) = S_m(\tau_2, S_T) = S$. Then

$$u_1 < u_2 \text{ in } (S_B, S_m).$$

Integrating equation (7.41) from $S_T$ to $S_m$ gives

$$u_2(S_T) - u_1(S_T) = \left(\frac{p_c^{(i)}(S_T)}{\sqrt{d_2}} - \frac{\beta(S_T)}{u_1(S_T)}\right) \left(1 - \frac{1}{\sqrt{d_1}}\right) + \int_{S_T}^{S_m} \beta \left(\frac{1}{u_2} - \frac{1}{u_1}\right). \quad (7.43)$$

Since $\beta > 0$ in $(S_T, S_m)$ for $S_T < S_m \leq \beta(S_T)$, the term in the right of (7.43) is negative, yielding a contradiction.

(b) Using (7.28) this part is demonstrated along the same lines. Details are omitted.

**Remark 7.6.** To complement Proposition 7.4, we further state that

$$S_m(\tau_1, S_T) = S_T \text{ if } S_m(\tau_2, S_T) = S_T \text{ and } S_m(\tau_2, S_T) = 1 \text{ if } S_m(\tau_1, S_T) = 1.$$  

The statements follow directly from the ordering of the orbits.

So far we have shown the monotonicity of the orbits. However, the question of continuous variation is still open. This is addressed in the following results.

**Proposition 7.5** (Continuous dependence of $u$). Let $v = p_c^{(i)} - w$. In the context of Proposition 7.4 and with $\Phi$ defined in (7.37) we have

(a) $0 < v^2(\tau_2, S_T) - v^2(\tau_1, S_T) < 2c(\tau_2 - \tau_1)\Phi(S)$ for $S_B < S \leq S_m(\tau_1, S_T)$;

(b) $0 < v^2(\tau, S_{T,2}) - v^2(\tau, S_{T,1}) < 2c(S_{T,2})\Phi(S; S_B, S_{T,2}) - c(S_{T,1})\Phi(S; S_B, S_{T,1})$ for $S_B < S \leq S_m(\tau_1, S_T)$;

**Proof.** As shown earlier in this section, $v$ satisfies the equation

$$(v^2)' = 2vp_c^{(i)} - 2c\tau\beta \text{ in } (v > 0).$$

Integrating this equation and using Proposition 7.4 and $\Phi$ from (7.37) gives the desired inequalities.
7.3 No relative permeability hysteresis and small $N_g$ (Scenario A)

**Corollary 7.1** (Continuous dependence of $S_m$). Let $\tau_0 > 0$ and $S_{T_0}$ be fixed such that $S_m(\tau_0, S_{T_0}) \leq \beta(S_{T_0})$. Then for any small $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon; \tau_0, S_{T_0})$ so that $|S_m(\tau, S_T) - S_m(\tau_0, S_{T_0})| < \varepsilon$ if $\max|\tau - \tau_0, |S_T - S_{T_0}| | < \delta$ and $S_m(\tau, S_T) < \beta(S_T)$.

**Proof.** We only demonstrate continuity with respect to $\tau$. Proving the continuity with respect to $S_T$ follows the same lines. We therefore take $S_T = S_{T_0}$ and drop its dependence from the notation for simplicity. Consider first $\tau > \tau_0$ and $S_{T_0} < \beta(S_{T_0})$.

Recalling $v(S_m(\tau_0); \tau_0) = 0$, Proposition 7.5 gives

$$0 < v(S_m(\tau_0); \tau) < \sqrt{2c(\tau - \tau_0)\Phi(S_m(\tau_0))},$$

where $\Phi(S_m(\tau_0)) > 0$ by (7.38) and Proposition 7.3. For any given (small) $\varepsilon > 0$ and with reference to Figure 7.7 choosing $\delta < \frac{p^{(i)}_c(S_m(\tau_0)) - p^{(i)}(S_m(\tau_0) + \varepsilon)}{2c\Phi(S_m(\tau_0))}$ we have

$$w(S_m(\tau_0), \tau) > p^{(i)}_c(S_m(\tau) + \varepsilon)$$

for all $\tau - \tau_0 < \delta$.

Since $w' > 0$, this implies the continuity of $\tau > \tau_0$.

Next let $\tau < \tau_0$ and $S_{T_0} \leq \beta(S_{T_0})$. Now we have from Proposition 7.5

$$0 < v(S_m(\tau); \tau_0) < \sqrt{2c(\tau_0 - \tau)\Phi(S_m(\tau))}. \quad (7.44)$$

Since $v(\cdot, \tau_0) \in C(\{S_B, S_m(\tau_0)\})$, $v(S_m(\tau_0), \tau_0) = 0$ and $v(\cdot, \tau_0) > 0$ in $(S_B, S_m(\tau_0))$, the continuity of $S_m$ follows directly from (7.44). \qed

![Figure 7.7: Behaviour of $w$ close to $S_m$: (left) $\tau > \tau_0$ and (right) $\tau < \tau_0$.](image)

Now we are in a position to describe how $S_m$ behaves for different combinations of $S_T$ and $\tau$.

**Proposition 7.6.** Let $S_B$ satisfy (7.25) and fix $S_T \in (S_B, \bar{S})$. Then there exists a $\tau_m(S_T) > 0$ such that

$$S_m(\tau, S_T) = S_T \text{ for all } 0 < \tau \leq \tau_m(S_T) \text{ and } S_m(\tau, S_T) > S_T \text{ for all } \tau > \tau_m(S_T).$$
Proof. The proof is based on Proposition 5.1 and [254, Lemma 4.5]. Define the function \( \ell_r(S) = p_c^{(i)}(S) + r(S - S_T) \), \( r > 0 \) and the constant \( P = \min_{S \in (S_B, S_T)} \{-p_c^{(i)}(S)\} > 0 \). We show that for \( \tau \) small enough there exists an \( r > 0 \) for which the curves \( w(S) \) and \( \ell_r(S) \) do not intersect. Specifically

\[
\ell_r(S) < w(S; \tau, S_T) < p_c^{(i)}(S) \quad \text{for all } r > r^- \in (0, P) \text{ and } S \in (S_B, S_T).
\]

This directly shows that \( w(S_T) = p_c^{(i)}(S_T) = \ell_r(S_T) \) meaning \( S_m(\tau, S_T) = S_T \).

Assuming the contrary, let \( S_i \in (S_B, S_T) \) be the coordinate at which \( w(S; \tau, S_T) \) and \( \ell_r(S) \) intersect for the first time. Since \( w(S_B) = p_c^{(i)}(S_B) > \ell_r(S_B) \), one gets

\[
(p_c^{(i)} - w)(S_i) = r(S_T - S_i) \quad \text{and} \quad \frac{cr'(S_i)}{r(S_T - S_i)} = w'(S_i) \leq \ell_r'(S_i) = p_c^{(i)}'(S_i) + r.
\]

Recalling that \( \gamma(S_T) = 0 \) and taking \( m_0(S_T) = \sup_{S \in (S_B, S_T)} \gamma'(S; S_B, S_T) < \infty \) we get

\[
\frac{\gamma(S_i)}{S_i - S_T} \leq m_0(S_T).
\]

Combining this with (7.46) results in the inequality \( r^2 - rP + crm_0 \geq 0 \). The roots of the quadratic expression on the left hand side of this inequality motivates us to define

\[
\bar{\tau}_m = \frac{P^2}{4cm_0(S_T)} \quad \text{and} \quad r^\pm = \frac{P}{2} \left[ 1 \pm \sqrt{1 - \frac{r}{\bar{\tau}_m}} \right].
\]

It directly follows that the inequality in (7.46) is not satisfied if \( 0 < \tau < \bar{\tau}_m \) and \( r \in (r^-, r^+) \subset (0, P) \). In this case one has \( w(S) > \ell_r(S) \) for \( S \in (S_B, S_T) \) and consequently, (7.45) holds. Note that \( \tau > \bar{\tau}_m \) does not necessarily imply that \( S_m(S_T, \tau) > S_T \). For this purpose, we define

\[
\tau_m(S_T) := \sup\{\tau : S_m(S_T, \tau) = S_T\} \geq \bar{\tau}_m > 0.
\]

Using Proposition 4.9, which states that

\[
w(S; \tau, S_T) \to -\infty \text{ as } \tau \to -\infty \text{ for all } S \in (S_B, S_T),
\]

we get from Corollary 7.1, \( \tau_m(S_T) < \infty \).

Proposition 7.7. Let \( S_B \) satisfy (7.25). Then the following holds:
(a) For each $S_T \in (S_*, \bar{S})$, there exists a unique $\tau_c = \tau_c(S_T)$ such that

$$S_m(\tau_c, S_T) = \beta(S_T).$$

(b) The function $\tau_c(\cdot)$ is strictly decreasing and continuous on $[S_*, \bar{S}]$. One has $\tau_c(S_T) \to \infty$ as $S_T \searrow S_*$ and $\tau_c(\bar{S}) = \bar{\tau} = \tau_m(\bar{S}) > 0$.

**Proof.** (a) Suppose no $\tau_c(S_T)$ exists such that $S_m(\tau_c, S_T) = \beta(S_T)$, meaning $S_m(\tau, S_T) < \beta(S_T)$ for all $\tau > 0$. Combined with (7.49), this implies that for large enough $\tau$, a $S_1 \in [S_T, \beta(S_T)]$ exists for which $w(S_1) = 0$. From (7.35) it is evident that $w(S_T) \leq w(S)$, in particular $w(S_T) < w(S_1) = 0$. Moreover, (7.37) gives the lower bound $w(S) \geq w(S_T) \geq p_c^{(i)}(S_T) - \sqrt{2ct\Phi(S_T)} \geq -\sqrt{2ct\Phi(S_T)}$ for all $S \in [S_B, S_m(\tau, S_T)]$. Multiplying both sides of (7.41) by $u$, integrating from $S_B$ to $S_1$ and using the above inequality we get

$$-\frac{1}{2} p_c^{(i)}(S_B)^2 = \int_{S_B}^{S_1} (ct\Phi(S; S_B, S_T) + p_c^{(i)}(S) w(S; c, \tau)) dS \leq -ct\Phi(S_1) + (p_c^{(i)}(S_B) - p_c^{(i)}(S_1)) \sqrt{2ct\Phi(S_T)}.$$

Since $\Phi(S_1) > 0$ (as stated in (7.38b)), this leads to a contradiction for $\tau \to \infty$. Hence, $S_m(\tau, S_T) = \beta(S_T)$ for some $\tau > 0$. The uniqueness follows from Proposition 7.4.

(b) The monotonicity and continuity follows from Propositions 7.4 and 7.5 and Corollary 7.1. To show the limit for $S_T \searrow S_*$, assume that $\lim_{S_T \searrow S_*} \tau_c(S_T) = \tau_\infty < \infty$. Let then $\tau > \tau_\infty$. Proposition 7.3 implies that $S_m(\tau, S_* < \beta(S_*)$. Choose an $S_T > S_*$ such that $\beta(S_T) \geq S_m(\tau, S_*)$. Since $\tau > \tau_c(S_T)$, we get that $w(S_m(\tau, S_*) ; \tau, S_T) \leq w(\beta(S_T) ; \tau, S_T) \leq p_c^{(i)}(\beta(S_T))$, implying $w(S_m(\tau, S_*) ; \tau, S_T) - w(S_m(\tau, S_*)) \geq p_c^{(i)}(S_m(\tau, S_*)) - p_c^{(i)}(\beta(S_T))$. This gives a contradiction when $S_T \searrow S_*$ since the right hand side goes
to \( p_c^{(i)}(S_m(\tau, S_*)) - p_c^{(i)}(\beta(S_*)) > 0 \), whereas the left hand side converges to 0 from Corollary 7.1.

The existence of a \( \bar{\tau} > 0 \) is a consequence of the continuity of \( \tau_c \) with \( \bar{\tau} = \tau_m(\bar{S}) \) following from Proposition 7.6.

After the preliminary statements we are in a position to consider the solvability of (TW) for different ranges of \( S_T \).

### 7.3.2 Problem (TW) with \( S_B < S_T \leq \bar{S} \)

![Figure 7.9: The different cases of \( S_B < S_T < S_* \). The orbits are plotted for \( 0 < \tau_1 < \tau_i < \tau_2 < \tau_d < \tau_3 \).](image)

We investigate the existence of an orbit connecting \((S_B, p_c^{(i)}(S_B))\) and the segment \(E^i_{S_T} \) to \( E^d_{S_T}\). Defining

\[
\tau_j = \frac{(p_c^{(j)'(S_T)})^2}{4cG'(S_T; S_B, S_T)} > 0, \quad j \in \{i, d\},
\]

the eigenvalues of the (TW) System associated with the equilibrium points \( E^j_{S_T}, j \in \{i, d\} \) are

\[
\lambda_{j}^i = \frac{p^{(j)'(S_T)}}{2c\tau} \left[ 1 \pm \sqrt{1 - \frac{\tau}{\tau_j}} \right] \text{ implying } \begin{cases} E^j_{S_T} \text{ is stable sink for } \tau \leq \tau_j, \\ E^j_{S_T} \text{ is stable spiral sink for } \tau > \tau_j. \end{cases}
\]

This immediately gives \( \tau_i \geq \tau_m \) as no monotone orbit can connect with \( E^i_{S_T} \) for \( \tau > \tau_i \).

The general behaviour of the orbits for \( S_T \in (S_B, S_*) \) are stated in

**Theorem 7.1.** Under the assumptions of Scenario A, consider \( S_B \) satisfying (7.25), \( S_T \in (S_B, S_*) \) and \( \tau_i < \tau_d \). Let \((S, p)\) be the orbit originating from \((S_B, p_c^{(i)}(S_B))\) satisfying (TW). Then, with reference to Figure 7.9, as \( \xi \to \infty \) one gets
(a) If $0 < \tau \leq \tau_i$, then either $S \to S_T$ and $p \to p_c^{(i)}(S_T)$ monotonically with respect to $\xi$ through $\mathcal{H}^{(i)}$ (when $\tau \leq \tau_m$) or the orbit $(S, p)$ goes around $E^i_{S_T}E^d_{S_T}$ finitely many times and ends up in either $E^i_{S_T}$ or $E^d_{S_T}$ (when $\tau_m < \tau \leq \tau_i$).

(b) If $\tau_i < \tau \leq \tau_d$, $(S, p) \to E^d_{S_T}$ after finitely many turns around $E^i_{S_T}E^d_{S_T}$.

(c) If $\tau_d < \tau$, then $(S, p)$ revolves infinitely many times around $E^i_{S_T}E^d_{S_T}$ while approaching it and (7.20) is satisfied.

Figure 7.10: Typical behaviour of $S(\xi)$ (left) and $p(\xi)$ (right) for different values of $\tau$. Here profiles for three different $\tau$ values are plotted satisfying $0 < \tau_1 < \tau_i < \tau_2 < \tau_d < \tau_3$. The $\xi = 0$ coordinate is fixed by (7.33).

These statements are demonstrated by arguments from Theorem 5.1 and Lemmas 5.1 and 5.2. We omit the details here. In Theorem 7.1 we have taken $\tau_i < \tau_d$ without loss of generality. In the $\tau_i > \tau_d$ case, the roles of the equilibrium points $E^i_{S_T}$ and $E^d_{S_T}$ are reversed. The typical behaviour of the $S$ and the $p$ profiles with respect to $\xi$ is given in Figure 7.10. Both $S$ and $p$ are monotone for $\tau < \tau_i$, whereas for $\tau_i < \tau < \tau_d$ they have finite number of local extrema and $p(+\infty) = p_c^{(d)}(S_T)$. For $\tau > \tau_d$, $S$ has infinitely many decaying local extrema, whereas $p$ has no limit. In particular, each $S$ maximum corresponds to a saturation overshoot. On the other hand, the oscillations in $p$ become wider, in line with the assumption $\lim_{\xi \to \infty} p'(\xi) = 0$. In this case, the segment $E^i_{S_T}E^d_{S_T}$ becomes an $\omega$-limit set of the orbit.

Turning to the case, $S_T \in (S_*, \bar{S})$, we define the two functions which will be used extensively below

**Definition 7.1.** The functions $\hat{S}_B, \bar{S}_B : [0, \infty) \to (0, 1]$ are such that

\[
\hat{S}_B(\tau) = \begin{cases} 
(\tau_c)^{-1}(\tau) & \text{for } \tau > \bar{\tau}, \\
\bar{S} & \text{for } 0 \leq \tau \leq \bar{\tau},
\end{cases}
\text{ and } \hat{S}_B(\tau) = \beta(\bar{S}_B(\tau)).
\]
Observe that, $\tilde{S}_B(\tau)$ is a strictly decreasing function whereas $\hat{S}_B(\tau)$ is a strictly increasing function for $\tau > \bar{\tau}$ and $\tilde{S}_B(\tau) = \hat{S}_B(\tau) = \bar{S}$ for $\tau \leq \bar{\tau}$. This is sketched in Figure 7.11. Numerically computed $\tilde{S}_B(\tau)$ and $\hat{S}_B(\tau)$ functions are shown in Figure 7.19.

Remark 7.7. The case when $\beta(\alpha)$ does not intersect $\gamma(\alpha)$ is treated in a similar way. However, since orbits may intersect the line segment $\{S = 1, p \leq 0\}$ in this case, a multivalued extension of $p_c^{(i)}$ at $S = 1$ needs to be introduced, see [169, 251] for further details. With this, one shows that the function $\tau_c(S_T)$ is well-defined in $[\tilde{S}, \bar{S}]$. Then a $\tau_B > 0$ exists such that $\tilde{S}_B(\tau_B) = \tilde{S}$ and $\hat{S}_B(\tau_B) = 1$. The subsequent results remain valid if $\hat{S}_B$ and $\tilde{S}_B$ are extended by $\tilde{S}_B(\tau_B) = \tilde{S}$ for $\tau > \tau_B$, and $\hat{S}_B(\tau) = 1$ for $\tau > \tau_B$.

With this in mind, we define the following sets:

$$\mathcal{A} = \{(S_T, \tau) : S_B < S_T < \tilde{S}, \tau < \tau_c(S_T)\},$$
$$\mathcal{B} = \{(S_T, \tau) : \tau > \bar{\tau}, \tilde{S}_B(\tau) < S_T < \hat{S}_B(\tau)\},$$
$$\mathcal{C} = \{(S_T, \tau) : \tilde{S} < S < S_*, \tau < \tau_c(\beta^{-1}(S_T))\}. \quad (7.51)$$

Figure 7.11: The sets $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ and the functions $\tilde{S}_B(\tau)$, $\hat{S}_B(\tau)$.

Observe that, if $S_T < \tilde{S}$ then only regions $\mathcal{A}$ and $\mathcal{B}$ are relevant. With $S_o$ defined in (A1), for $(S_T, \tau) \in \mathcal{A}$ one has

**Proposition 7.8.** For a fixed $S_B \in (0, S_o)$ and $(S_T, \tau) \in \mathcal{A}$, the orbit $(S, p)$ entering $\mathcal{H}^{(i)}$ from $E_{S_B}^i$ behaves according to statements (a), (b) and (c) of Theorem 7.1.

We discuss the remaining situations, $(S_T, \tau) \in \mathcal{B}$ and $(S_T, \tau) \in \mathcal{C}$ in the next section.
7.3 No relative permeability hysteresis and small $N_g$ (Scenario A)

7.3.3 $(S_T, \tau) \not\in \mathcal{A}$

Since $S_T > \hat{S}_B(\tau)$, a TW cannot connect $S_B$ and $S_T$. However, a different class of waves is possible when $(S_T, \tau) \in \mathcal{B}$.

**Proposition 7.9.** For a fixed $S_B \in (0, S_o)$ and $(S_T, \tau) \in \mathcal{B}$, consider the system

$$
\begin{cases}
    S' = \frac{1}{c_d} \mathcal{F}(S, p), \\
    p' = \mathcal{G}(S; \hat{S}_B(\tau), S_T),
\end{cases}
$$

with $c_d = \frac{F(\hat{S}_B(\tau)) - F(S_T)}{\hat{S}_B(\tau) - S_T}$. (7.52)

For this system an orbit $(S_d, p_d)$ exists that connects $E^d_{\hat{S}_B(\tau)}$ for $\xi \to -\infty$ to $E^d_{S_T}$ for $\xi \to \infty$.

**Proof.** Upon inspection of the eigen-directions for the system (7.52) around the equilibrium point $E^d_{\hat{S}_B(\tau)}$ one concludes that there is indeed an orbit $(S_d, p_d)$ that connects to $E^d_{\hat{S}_B(\tau)}$ as $\xi \to -\infty$ from the set $\mathcal{H}^{(d)}$ defined in (7.8). Moreover, from the direction of the orbits in this case, as shown in Figure 7.12 (left), it is apparent that after leaving $E^d_{\hat{S}_B(\tau)}$, $S_d$ decreases monotonically till the orbit either hits the curve $p = p^{(d)}_c(S)$ for some $S \leq S_T$ or exits $\{S > S_B\}$ through the line $S = S_B$. We prove that it is not possible for the orbit to escape through $S = S_B$.

To show this, consider the orbit $(S_1, p_1)$ that satisfies the original (TW) equations and enters $\mathcal{H}^{(d)}$ from $E^d_{\hat{S}_B(\tau)}$. We show that this orbit cannot cross the line $S = S_B$.

The divergence argument presented in [68, 169, 251] is used for this purpose. To

![Figure 7.12: (left) The direction of orbits for the system (7.23a), (7.52) and the orbits $(S_d, p_d)$ and $(S_1, p_1)$. Here the orbit $(S_d, p_d)$ connects $E^d_{\hat{S}_B(\tau)}$ and $E^d_{S_T}$. (right) The domain $\Omega$ used in the divergence argument for the hypothetical case where $(S_1, p_1)$ crosses the line $S = S_B$.](image-url)
elaborate, assume that \((S_1, p_1)\) intersects the line \(S = S_B\) at \(T\). Consider the region \(\Omega\), enclosed by the segments \(E^l_{S_B} T\), \(E^l_{S_B(\tau)} E^d_{S_B(\tau)}\), the orbit \((S_1, p_1)\) and the orbit \((S_2, p_2)\) that satisfies (TW) and connects \(E^l_{S_B}\) and \(E^l_{S_B(\tau)}\), see Figure 7.12 (right). Introducing the vector-valued function \(\overrightarrow{R}(S, p) = (\frac{1}{cT} \overrightarrow{S}(S, p), \mathcal{G}(S; S_B, S_T))\) and deduces from (7.9),

\[
\text{div} \overrightarrow{R} = \frac{1}{cT} \frac{\partial}{\partial S} \overrightarrow{S} (S, p) = \frac{1}{cT} \begin{cases} 
  p_c^{(j)}(S) & \text{in } \mathcal{H}^{(j)}, \\
  0 & \text{in } \mathcal{H}, \\
  p_c^{(d)}(S) & \text{in } \mathcal{H}^{(d)}.
\end{cases}
\]

This gives a contradiction when the divergence theorem is applied to \(\overrightarrow{R}\) in the domain \(\Omega\): the integral of \(\overrightarrow{R}\) over \(\partial \Omega\) is non-negative whereas \(\int_\Omega \text{div} \overrightarrow{R} < 0\) from (P1) and Figure 7.12 (right). Hence, the orbit \((S_1, p_1)\) intersects \(p_c^{(d)}(S)\) at some \(S \in (S_B, \hat{S}_B(\tau))\).

The wave-speed corresponding to the orbit \((S_d, p_d)\) satisfies

\[
c_d < \frac{F(\hat{S}_B(\tau)) - F(S_B)}{\hat{S}_B(\tau) - S_B} = c_i,
\]

\((7.53)\)

\(c_i\) being the speed of both \((S_1, p_1)\) and \((S_2, p_2)\) waves. Hence, by the continuity of the orbits with respect to \(c\), as shown in Proposition 7.4, it is evident that \((S_d, p_d)\) intersects \(p_c^{(d)}(S)\) for some \(S_d > S_B\). From here, the rest of the proof is identical to the proof of Theorem 7.1, and follows the arguments in Theorem 5.1 and Lemmas 5.1 and 5.2.

\[\square\]

From the results of Theorem 7.1 we further state

**Corollary 7.2.** The orbit \((S_d, p_d)\) can monotonically go to \(E^d_{S_T}\) only if \(\tau \leq \frac{(p_c^{(d)}(S_T))^2}{4c_d \mathcal{G}'(S_T; S_B(\tau), S_T)}\).

For \(\tau\) large enough, the orbit \((S_d, p_d)\) goes around \(E^l_{S_T} E^d_{S_T}\) infinitely many times while approaching it, and \(\lim_\xi \rightarrow \pm \infty p_d'(\xi) = 0\).

Observe that, if \((S_T, \tau) \in \mathcal{C}\) then travelling waves do not exist between \(S_T\) and \(\hat{S}_B(\tau)\) since both are in the concave part of \(F\) with \(S_T > \hat{S}_B(\tau)\). Thus we have exhausted all the possibilities of connecting \(S_B\) and \(S_T\) with Theorem 7.1 and Proposition 7.9.
7.3 No relative permeability hysteresis and small \( N_g \) (Scenario A)

7.3.4 Entropy solutions to hyperbolic conservation laws

![Figure 7.13: The entropy solutions for (left) \((S_T, \tau) \in \mathcal{A}\), (center) \((S_T, \tau) \in \mathcal{B}\) and (right) \((S_T, \tau) \in \mathcal{C}\). Note that the solutions in the center and the right figures include non-classical shocks.]

Under the conditions of Scenario A, we consider the Riemann problem

\[
\partial_t S + \partial_x F(S) = 0 \text{ in } \mathbb{R} \times [0, \infty) \tag{7.54a}
\]

with \( S(x, 0) = \begin{cases} S_T & \text{for } x < 0, \\ S_B & \text{for } x > 0. \end{cases} \) \tag{7.54b}

In the context of the viscous model discussed in this paper, we consider the Buckley-Leverett equation (7.54a) as the limit of System (7.17) for \( N_c \downarrow 0 \). As a consequence, we only take into account those shock solutions of (7.54a) that have a viscous profile in the form of a travelling wave satisfying (TW). Such shocks are called admissible because they arise as the \( N_c \to 0 \) limit of TWs. In this sense, the entropy condition for shocks satisfying (7.54a) are equivalent to existence conditions for travelling waves satisfying (TW). This may lead to non-classical shocks violating the well-known Oleinik entropy conditions, see e.g. [254].

Here, we assume

\[
0 < S_B < S_T < 1, \tag{7.55}
\]

which is more general compared to (7.25) where the additional constraint of \( S_B < S_o \) was imposed. This generalisation is possible since \( S_B > S_o \) simply implies that the sets \( \mathcal{A}, \mathcal{B} \) are empty. Our analysis can also be applied to derive the entropy conditions for the case \( S_B > S_T \), however, for simplicity we restrict our discussion to (7.55).

\((S_T, \tau) \in \mathcal{A}\)

As in the usual Buckley-Leverett case (i.e. without dynamic capillarity and hysteresis in the regularised models) the solution is given by

\[
S(x, t) = \begin{cases} S_T & \text{for } x < ct, \\ S_B & \text{for } x > ct, \end{cases} \quad \text{where } c = \frac{F(S_T) - F(S_B)}{S_T - S_B}. \tag{7.56}
\]
Here, the shock satisfies the classical Oleinik condition.

\[(S_T, \tau) \in \mathcal{B}\]

In this case the admissible solution is composed of two shocks: an imbibition shock from \(S_B\) to \(\hat{S}_B(\tau)\), followed by a drainage shock from \(\hat{S}_B(\tau)\) to \(S_T\).

\[
S(x, t) = \begin{cases} 
S_T & \text{for } x < c_d t, \\
\hat{S}_B(\tau) & \text{for } c_d t < x < c_i t, \\
S_B & \text{for } x > c_i t,
\end{cases}
\]

with \(c_i = \frac{F(\hat{S}_B(\tau)) - F(S_B)}{\hat{S}_B(\tau) - S_B}\), \(c_d = \frac{F(\hat{S}_B(\tau)) - F(S_T)}{\hat{S}_B(\tau) - S_T}\). (7.57)

Note that this solution violates the Oleinik condition [180]. Both shocks are under compressive.

\[(S_T, \tau) \in \mathcal{C}\]

The solution in this case violates again the Oleinik entropy condition. It consists of an imbibition shock from \(S_B\) to \(\hat{S}_B(\tau)\) followed by a rarefaction wave from \(\hat{S}_B(\tau)\) to \(S_T\),

\[
S(x, t) = \begin{cases} 
S_T & \text{for } x < F'(S_T) t, \\
r(x/t) & \text{for } F'(S_T) t < x < F'(\hat{S}_B(\tau)) t, \\
\hat{S}_B(\tau) & \text{for } F'(\hat{S}_B(\tau)) t < x < c_i t, \\
S_B & \text{for } x > c_i t,
\end{cases}
\]

with \(r(\cdot)\) satisfying

\[F'(r(\zeta)) = \zeta, \text{ for } F'(S_T) \leq \zeta \leq F'(\hat{S}_B(\tau)).\] (7.58)

Since \(F\) is concave for \(S \in [\hat{S}_B(\tau), S_T]\), \(F'\) is monotone implying that \(r(\cdot)\) is well-defined. We observe that in the last two cases the solution features a plateau-like region. This plateau appears and grows in time since the speeds of the drainage shock and of the end point of the rarefaction wave are lesser than the speed of the imbibition shock. Interestingly, the saturation of the plateau only depends on \(p_c^{(i)}\) and not on \(p_c^{(d)}\). To be more specific, although the viscous profile consisting of a travelling wave connecting \(E_{d_{S_B}}^d\) and \(E_{S_T}^d\) depends on \(p_c^{(d)}\), the shock resulting from it, in the hyperbolic limit, does not. However, the role of the drainage curve in the entropy solutions becomes evident in Scenario B, which is discussed later.

In the absence of hysteresis and for linear higher order terms, which corresponds to constant \(k\) and linear \(p_c-S\) dependence, in [254, Section 6] it is proved that the non-standard entropy conditions discussed here are entropy dissipative for the entropy \(U(s) = \frac{1}{2} s^2\). However, such an analysis is beyond the scope of this paper. The solution profiles for the Riemann problem are shown in Figure 7.13.
Extension to the non-monotone $F$ case

The analysis so far can be extended to the case where $N_g$ is large resulting in $F$ being non-monotone. If $S_F \in (0,1)$ is the saturation where $F(S)$ attains its maximum (see Remark 7.2 and Figure 7.2), then the results obtained so far cover the case when $S_T$ and $S^*$ are below $S_F$. However, if $S_T > S_F$ then the TW study has to be conducted also from a $S_T$ perspective, not only from the $S_B$ one. In this scenario, since fronts having negative speeds and thus moving towards $S_T$ become possible, one has to consider the functions $\hat{S}_T(\tau)$, $\check{S}_T(\tau)$ for a fixed $S_T$, similar to $\hat{S}_B(\tau)$, $\check{S}_B(\tau)$ from Definition 7.1 for fixed $S_B$. Due to the symmetry in the behaviour of the fronts approaching $S_B$, respectively $S_T$, some of the results obtained so far extend straightforwardly to the non-monotone case. However, a detailed analysis is much more involved and therefore left for future research because of the following two reasons:

(a) Depending on the relative positions of $S_B$, $\hat{S}_B$, $S_T$ and $\check{S}_T$, there are many subcases to consider. In this case up to three shocks are possible, traveling both forward and backward. Which of these shocks are admissible and how they are connected requires further analysis.

(b) For a non-monotone $F$, when considering the hyperbolic limit in the absence of hysteresis or dynamic effects, the entropy solutions may include rarefaction waves with endpoints moving in opposite directions, forward and backward. When capillary hysteresis is included, preliminary numerical results have provided solutions incorporating two rarefaction waves, one with endpoints travelling backward and another one with endpoints travelling forward, and a stationary shock at $x = 0$. Such solutions still need to be analysed further.

7.4 Hysteretic relative permeabilities and small $N_g$ (Scenario B)

For Scenario B, the flux function $F(S, p)$ is composed of $F^{(j)} = f^{(j)} + N_g h^{(j)}$ for $j \in \{i, d\}$ and $F = \bar{f} + N_g \bar{h}$ such that

$$F(S, p) = \begin{cases} F^{(d)}(S) & \text{if } (S, p) \in \mathcal{H}^{(d)}, \\ \hat{F}(S, p) & \text{if } (S, p) \in \mathcal{H}, \\ F^{(i)}(S) & \text{if } (S, p) \in \mathcal{H}^{(i)}. \end{cases}$$  \hspace{1cm} (7.59)

It has the following properties

(A2) $F \in C(\mathcal{W})$, $\bar{F} \in C^2(\mathcal{H})$, $\partial_p F > 0$ in $\mathcal{H}$ and $F^{(i)}$, $F^{(d)}$ satisfy properties stated for $F$ in (A1). Additionally, $F^{(d)}(S) > F^{(i)}(S)$ for $0 < S < 1$. 

In this scenario, $S_B$ can be taken in the entire interval $(0,1)$ and $p_B$ can be chosen independently as long as $(S_B, p_B) \in \mathcal{H}$, i.e.

$$0 < S_B < 1 \text{ and } p_B \in [p^{(i)}_c(S_B), p^{(d)}_c(S_B)].$$ \hspace{1cm} (7.60)
This is different from Scenario A where $S_B$ is restricted to the interval $(0, S_o)$ and $p_B$ is fixed to $p_B = p_c^{(i)}(S_B)$.

We first introduce some notation.

**Definition 7.2.** For $k \in \{B, T\}$ let $E_k = (S_k, p_k)$ and $U_k = (S_k, F(S_k, p_k))$ (see Figure 7.14 (left)). We define the saturations $\tilde{S}_j$, $j \in \{i, d\}$ as the $S$-coordinates of the tangent points to $F^{(j)}(S)$ from $U_B$ such that $\tilde{S}_i \geq S_B$ and $\tilde{S}_d \leq S_B$.

Observe that, the saturations $\tilde{S}_j$, for $j \in \{i, d\}$, are functions of $U_B$. The properties of $F^{(j)}$ further ensure that they are well defined. If $S_B$ is such that $F^{(i)''}(S_B) \leq 0$ and $p_B = p_c^{(i)}(S_B)$ then $\tilde{S}_i = S_B$. Similarly if $F^{(d)''}(S_B) \geq 0$ and $p_B = p_c^{(d)}(S_B)$ then $\tilde{S}_d = S_B$.

The existence of travelling waves is analysed for the following two cases:

Case (i): $S_B < S_T \leq \tilde{S}_i$, and Case (ii) : $\tilde{S}_d \leq S_T < S_B$.

![Figure 7.14: (left) The graphs of $F^{(i)}$ and $F^{(d)}$, together with the saturations $\tilde{S}_i$, $\tilde{S}_d$ and the points $U_B$, $U_T$. (right) The orbit directions for Case (i) for two equilibrium points $E_B$ and $E_T$. The black dotted curve represents points where $\Theta(S, p) = 0$, implying $p' = 0$.](image)

Regarding the choice of $p_T$, we have the following

**Proposition 7.10.** Let $S_B$ and $S_T$ be as in Case (i) or (ii). Then any solution of (TW) that connects $E_B$ and $E_T$ can only exist if $p_T = p_c^{(i)}(S_T)$ or $p_T = p_c^{(d)}(S_T)$.

**Proof.** Since $E_T$ is an equilibrium point, $\Theta(S_T, p_T) = 0$, which implies that $p_T \in [p_c^{(i)}(S_T), p_c^{(d)}(S_T)]$. The directions of the orbits for $p_T$ in this interval are displayed in Figure 7.14 (right). We proceed by introducing the set 

$$H_0 = \{(S, p) : S \in (0, 1), p \in \mathbb{R} \text{ such that } \Theta(S, p) = 0\}.$$
7.4 Hysteretic relative permeabilities and small $N_g$ (Scenario B)

It corresponds to the black dotted curve in Figure 7.14 (right). Let $\ell = \ell(S)$, defined in (7.24), be the line passing through $U_B$ and $U_T$. If $\ell$ intersects $F^{(i)}$ at $S = S_H$, then the vertical half-line $\{(S_H, p) : p < p_c^{(i)}(S_H)\}$ lies in $\mathcal{H}_0$ due to the definition of $F$ in (7.59). Concerning $F^{(d)}$, $\ell$ has either zero, one or two intersection points, see Figure 7.15 (left). In the latter case, as before, $\mathcal{H}_0$ contains one or two vertical half-lines as shown in the (right) plot of Figure 7.15. However, this aspect plays no major role in the analysis below.

Every point in the set $\mathcal{H}_0 \cap \mathcal{H}$ is an equilibrium point. However, all points in the set $\mathcal{H}_0 \cap \text{int}(\mathcal{H})$ (the interior of $\mathcal{H}$ being referred to as int($\mathcal{H}$) here) are unstable and as follows from Figure 7.14 (right), no orbit can reach these points as $\xi \to \infty$. This eliminates all other possibilities to reach $E_T$ as $\xi \to \infty$ except for $p_T = p_c^{(i)}(S_T)$ and $p_T = p_c^{(d)}(S_T)$. □

We now consider the two cases separately.

7.4.1 Case (i): $S_B < S_T \leq \tilde{S}_i$

The main result of this section is

**Proposition 7.11.** Assume (7.60) and let $S_T \in (S_B, \tilde{S}_i)$, $p_T = p_c^{(i)}(S_T)$ and $F^{(i)}(S_T) > F(S_B, p_B)$. Then a $\tau_T^*(S_T) > 0$ exists such that for all $\tau < \tau_T^*(S_T)$ there is an orbit satisfying (TW) and connecting $E_B$ to $E_T$.\footnote{Note that, $\mathcal{G}_i(S) = (F^{(i)}(S) - \ell(S))/h^{(i)}(S)$. The system (7.61) has exactly the same structure as (TW) described in Section 7.3. Defining $\tau_T^*(S_T)$ similar to $\tau_m$ in Proposition 7.6, the result follows directly. □}

**Proof.** Consider the orbit $(S_{(i)}, p_{(i)})$ that leaves $E_B$ vertically through the half-line $\{(S = S_B, p < p_B)\}$. The directions of the orbits in $\mathcal{H}$ imply that $(S_{(i)}, p_{(i)})$ intersects $p_c^{(i)}(S)$ and enters $\mathcal{H}^{(i)}$ (the region under the graph of $p_c^{(i)}(S)$ at some finite $\xi \in \mathbb{R}$, see Figure 7.15 (right). In $\mathcal{H}^{(i)}$ its motion is governed by the system

$$
\begin{align*}
S_{(i)\prime} &= \frac{1}{c_{(i)}} (p^{(i)}_c(S_{(i)}) - p_{(i)}), \\
p_{(i)\prime} &= \mathcal{G}_i(S_{(i)}) := \mathcal{G}(S_{(i)}, p^{(i)}_c(S_{(i)})),
\end{align*}
$$

with $c_{(i)} = \frac{F^{(i)}(S_T) - F(S_B, p_B)}{S_T - S_B} > 0$. (7.61)

Note that, $\mathcal{G}_i(S) = (F^{(i)}(S) - \ell(S))/h^{(i)}(S)$. The system (7.61) has exactly the same structure as (TW) described in Section 7.3. Defining $\tau_T^*(S_T)$ similar to $\tau_m$ in Proposition 7.6, the result follows directly. □

**Remark 7.8.** Observe that, the construction fails if $F^{(i)}(S_T) < F(S_B, p_B)$ which is intuitive since the overall process is not imbibition in this case. If one prescribes a flux $F = F_T$ at $\xi = \infty$ which is less than $F(S_B, p_B)$, then Propositions 7.10 and 7.11 forces the saturation at $\xi = \infty$ to be $S_T = (F^{(d)})^{-1}(F_T) < S_B$, reducing the problem to Case (ii). However, if one fixes the saturation $S_T$ so that $F(S_B, p_B) > F^{(i)}(S_T)$, then we get a frozen profile with a $p_T \in (p_c^{(i)}(S_T), p_c^{(d)}(S_T))$ that satisfies $F(S_T, p_T) = F(S_B, p_B)$. This is explained further in Section 7.5.2. We set $\tau_T^*(S_T) = \infty$ in this case.
Proposition 7.11 implies the following:

**Corollary 7.3.** Under the assumptions of Proposition 7.11, let $S_{(i)}(\xi) = S$ for some $S \in (S_B, S_T)$ and $\xi \in \mathbb{R}$. Define $w(S; \tau) := p_{(i)}(\xi) < p_{c(i)}(S)$. Then $\lim_{\tau \to 0} w(S; \tau) = p_{c(i)}(S)$.

![Diagram](image)

Figure 7.15: (left) $U_B, U_T, \ell(S)$ used in Proposition 7.11 in the $S$-$F$ plane. (right) The $S$-$p$ plane and the orbit $(S_{(i)}, p_{(i)})$ for Case (i) with $F^{(i)}(S_T) > F(S_B, p_B)$ and $\tau < \tau^*_i(S_T)$.

Here, $w$ is the counterpart of $w$ defined in Section 7.3 for Scenario A. The proof of Corollary 7.3 is based on the inequality (7.37) which is satisfied in this case by $w$. From Corollary 7.3 one obtains that for Case (i), if $\tau \searrow 0$, meaning that if the dynamic capillarity is vanishing, then the orbit follows either the scanning curve, here the line segment $(S = S_B, p_{c(i)}(S_B) < p < p_B)$, or the imbibition curve $p_{c(i)}$. The result is analogous to the results for capillary hysteresis given in Section 4.3.

### 7.4.2 Case (ii): $\bar{S}_d \leq S_T < S_B$ and stability of plateaus

The counterpart of Proposition 7.11 for Case (ii) is (see also Figure 7.16),

**Proposition 7.12.** Assume (7.60) and let $S_T \in [\bar{S}_d, S_B)$, $p_T = p_{c(d)}(S_T)$ and $F^{(d)}(S_T) < F(S_B, p_B)$. Then a $\tau^*_d(S_T) > 0$ exists such that for all $\tau < \tau^*_d(S_T)$ there is an orbit $(S_{(d)}, p_{(d)})$ satisfying (TW) and connecting $E_B$ to $E_T$. Moreover, for a fixed $S_{(d)} = S \in [S_T, S_B)$, one has $p_{(d)} \to p_{c(d)}(S)$ as $\tau \to 0$. 
Finally, we investigate a special case related to the development of stable saturation plateaus in imbibition experiments. For $S_B \in (0,1)$, and $S_T \in (S_B,1)$ a stable plateau is formed when an imbibition wave, from $S_B$ to $S_P \in (S_T,1)$, followed by a drainage wave, from $S_P$ to $S_T$, both have the same speed resulting in the width of the plateau to remain constant. This is different from the plateaus described in (7.57) where the speeds of the imbibition and the drainage fronts are necessarily different. The existence of stable saturation plateaus has been widely studied experimentally [72,102,229] and numerically [119,216]. Although results regarding stability of the plateau are available [119,216], the mechanism behind its development is still not well understood. Here, we give an example where our analysis predicts that such a plateau will develop. Specifically, it occurs when $\tau > \tau^*_T(S_T)$ and a direct monotone orbit from $E_B$ to $(S_T, p^{(i)}_c(S_T))$ is no longer possible. This is verified numerically in Section 7.5.2.

**Proposition 7.13.** Assume (7.60) and let $S_T \in (S_B,1)$ be such that the line $\ell$ through $U_B = (S_B,F(S_B,p_B))$ and $U_T = (S_T,F^{(d)}(S_T))$ in the F-S plane intersects $F^{(i)}$ at some $S = S_P \in (S_T,\bar{S}).$ Consider the system (TW) with the wave-speed 

$$c_P = \frac{F^{(d)}(S_T) - F(S_B,p_B)}{S_T - S_B} = \frac{F^{(i)}(S_P) - F(S_B,p_B)}{S_P - S_B} = \frac{F^{(i)}(S_P) - F^{(d)}(S_T)}{S_T - S_B}.$$ 

For this system, let $(S^P_{(i)},P^P_{(i)})$ be the orbit that passes through $H^{(i)}$ and connects to the equilibrium point $(S_B,p_B)$ as $\xi \to -\infty$, described in Proposition 7.11. Similarly, let $(S^P_{(d)},P^P_{(d)})$ be the orbit passing through $H^{(d)}$ and connecting to $(S_P,p^{(i)}_c(S_P))$ as $\xi \to -\infty$, described in Proposition 7.12. Assume that $0 < \tau < \max\{\tau^*_i(S_P),\tau^*_d(S_T)\}$ where the $\tau^*_i(S_P)$ and the $\tau^*_d(S_T)$ values correspond to the orbits $(S^P_{(i)},P^P_{(i)})$ and $(S^P_{(d)},P^P_{(d)})$ respectively. Then, $(S^P_{(i)},P^P_{(i)}) \to (S_P,p^{(i)}_c(S_P))$ as $\xi \to -\infty$ and $(S^P_{(d)},P^P_{(d)}) \to (S_T,p^{(d)}_c(S_T))$ as $\xi \to -\infty$.

The proof follows directly from Propositions 7.11 and 7.12.
7.4.3 Entropy solutions

We can now discuss the entropy solutions of the Riemann problem (7.54) under the assumptions of Scenario B. To be more specific, we give a selection criteria for the solutions of the system

\[
\begin{aligned}
\partial_t S + \partial_x F(S,p) &= 0, \\
p &\in \frac{1}{2}(p^{(d)}_c(S) + p^{(i)}_c(S)) - \frac{1}{2}(p^{(d)}_c(S) - p^{(i)}_c(S)) \cdot \text{sign}(\partial_t S), \\
\end{aligned}
\]  

in \( \mathbb{R} \times [0,\infty) \)  

(7.62)

with \( S(x,0) = \begin{cases} S_T & \text{for } x < 0, \\ S_B & \text{for } x > 0, \end{cases} \)  

and \( p(x,0) = p_B \) for \( x > 0 \).  

(7.63)

We view (7.62) as the limit of (\( \mathcal{P} \)) (System (7.17)) when the capillary effects vanish. However, hysteresis is still present in the model. Note that, \( \tau \) still plays a role in determining the entropy solution despite being absent in (7.62). This is similar to what we saw in Section 7.3. However, the focus here being hysteresis in permeability and capillary pressure, for a fixed \( S_B \in (0,1) \) we take

\[
0 < \tau < \min \left\{ \inf_{S_T \in (S_B,\bar{S}_i)} \tau^*_i(S_T), \inf_{S_T \in (\bar{S}_d, S_B)} \tau^*_d(S_T) \right\}. 
\]

(7.64)

Observe that, (7.64) does not provide a void interval for \( \tau \). To see this, note that \( \tau^*_r(S_T) \) is defined similar to \( \tau^*_m \) in Proposition 7.6 and thus, it satisfies the inequality in (7.48), i.e. it has the positive quantity \( \bar{\tau}_m \) as its lower bound. Although \( \bar{\tau}_m \) in Proposition 7.6 actually depends on \( S_T \), one sees from (7.47) that the values of \( \bar{\tau}_m \) are bounded away from 0 uniformly with respect to \( S_T \). Hence, \( \tau^*_r(S_T) \) is also bounded uniformly away from 0. Similar argument holds for \( \tau^*_d(S_T) \).

We now consider the cases \( S_T > S_B \) and \( S_T < S_B \) separately.

\( S_T > S_B \)

If \( S_T \leq \bar{S}_i \) (\( \bar{S}_i \) introduced in Definition 7.2) and \( F^{(i)}(S_T) > F(S_B,p_B) \) then the entropy solution is a shock:

\[
S(x,t) = \begin{cases} S_T & \text{for } x < c(i)_t, \\ S_B & \text{for } x > c(i)_t, \end{cases} \quad \text{with } c(i)_t = \frac{F^{(i)}(S_T) - F(S_B,p_B)}{S_T - S_B}. 
\]

(7.65)

For \( F^{(i)}(S_T) < F(S_B,p_B) \), from Remark 7.8, the solution is (7.65) but with \( c(i)_t = 0 \), i.e. it is a stationary shock. However, if \( S_T > \bar{S}_i \) then the solution becomes more complex,
combining a rarefaction wave with a shock:

\[
S(x, t) = \begin{cases} 
S_T & \text{for } x < F^{(i)}(S_T)t, \\
F^{(i)}(S_T)t & \text{for } F^{(i)}(S_T)t < x < F^{(i)}(S)_{\ell}t, \\
S_{\ell} & \text{for } F^{(i)}(S)_{\ell}t < x < c_i t, \\
S_B & \text{for } x > c_i t.
\end{cases} \tag{7.66}
\]

Here \(r^{(i)}(\cdot)\) satisfies

\[F^{(i)}(r^{(i)}(\zeta)) = \zeta, \text{ for } F^{(i)}(S_T) \leq \zeta \leq F^{(i)}(S_{\ell}).\]

If \(S_T < S_B\)

If \(S_T \geq S_d\) then the entropy solution for \(F(S_B, p_B) > F^{(d)}(S_T)\) is the shock

\[
S(x, t) = \begin{cases} 
S_T & \text{for } x < c_d t, \\
S_B & \text{for } x > c_d t, \text{ with } c_d = \frac{F(S_B, p_B) - F^{(d)}(S_T)}{S_B - S_T},
\end{cases} \tag{7.67}
\]

and for \(F(S_B, p_B) < F^{(d)}(S_T)\) it is (7.67) with \(c_d = 0\). If \(S_T < S_d\) then the solution has a similar structure to (7.66), i.e.

\[
S(x, t) = \begin{cases} 
S_T & \text{for } x < F^{(d)}(S_T)t, \\
r^{(d)}(x/t) & \text{for } F^{(d)}(S_T)t < x < F^{(d)}(S_d)t, \\
S_d & \text{for } F^{(d)}(S_d)t < x < c_d t, \\
S_B & \text{for } x > c_d t, \\
\end{cases} \tag{7.68}
\]

with the function \(r^{(d)}(\cdot)\) satisfying

\[F^{(d)}(r^{(d)}(\zeta)) = \zeta, \text{ for } F^{(d)}(S_T) \leq \zeta \leq F^{(d)}(S_d).\]

### 7.5 Numerical results

For the numerical experiments, we solve \((\tilde{\mathcal{P}})\) (System (7.18)) in a domain \((x_{in}, x_{out})\), where \(x_{in} < 0\) and \(x_{out} > 0\). As an initial condition for the saturation variable, we choose a smooth and monotone approximation of the Riemann data as in Chapters 4 and 5:

\[
S(x, 0) = \begin{cases} 
S_T & \text{for } x < -l, \\ \
\frac{(S_B + S_{\ell})}{2} + \frac{(S_{\ell} - S_B)}{4l^2} x(x^2 - 3l^2) & \text{for } -l \leq x \leq l, \\ \
S_B & \text{for } x > l.
\end{cases} \tag{7.69}
\]
Here, \( l \) is a smoothing parameter, \( S_T \) denotes the saturation induced by a certain injection rate and \( S_B \) is the initial saturation within the porous medium. In order to model the capillary pressure, a van Genuchten parametrisation (1.7) is considered, i.e.

\[
p_{c}^{(j)}(S) = \Lambda_{j}(S^{\frac{1}{m_{j}}}-1)^{1-m_{j}}, \quad j \in \{i, d\}.
\]

In the remainder of this section we use the following parameter set: \( \Lambda_{i} = 3.5, m_{i} = 0.92, \Lambda_{d} = 7 \) and \( m_{d} = 0.9 \). To solve \( \mathcal{P} \) numerically, for \( n \in \mathbb{N} \cup \{0\} \) and \( t_{0} = 0 \), we solve within the time step \([t_{n}, t_{n+1}]\) of width \( \Delta t_{n} = t_{n+1} - t_{n} \), the elliptic problem

\[
-\partial_{x}(F(S, p) + h(S, p) \partial_{x} p) = \frac{1}{t} \mathcal{F}(S, p),
\]

with respect to the pressure variable \( p \). For a given \( S \), this is a nonlinear elliptic problem and to solve it, a linear iterative scheme is employed which is referred to as the L-scheme in literature [155,192]:

\[
L[p_{n}^{i} - p_{n}^{i-1}] - \partial_{x}(F(S_{n}, p_{n}^{i-1}) + h(S_{n}, p_{n}^{i-1}) \partial_{x} p_{n}^{i}) = \frac{1}{t} \mathcal{F}(S_{n}, p_{n}^{i-1}),
\]

see also Chapter 9. Here, \( p_{n}^{i} \) denotes the pressure at the \( i \)th iteration and \( p_{n}^{0} = p(x, t_{n}) \). On closer examination, the L-scheme corresponds to a linearisation of the nonlinear problem, since for each iteration a linear equation in the unknown pressure variable \( p_{n}^{i} \) is solved. For Scenario A, the parameter \( L \) is set to \( L = \frac{1}{t} \) to ensure convergence of the L-scheme [169,192] and for Scenario B, the modified variant of the L-scheme is used, proposed in Chapter 9, to speed up the convergence, since in this scenario the stiffness matrix has to be recomputed in every iteration. A standard cell centered finite volume scheme is considered for discretizing the linearised elliptic problem in space. Having the pressure variable \( p_{n} \) and the saturation variable \( S_{n} \) for \( t = t_{n} \) at hand, we update the saturation as follows:

\[
S_{n+1} = S_{n} + \frac{\Delta t_{n}}{t} \mathcal{F}(S_{n}, p_{n}).
\]

### 7.5.1 Numerical results for Scenario A

First we illustrate the theoretical findings of Scenario A. The boundary conditions with respect to the pressure variable are of Neumann type at \( x = x_{in} \) and of Dirichlet type at \( x = x_{out} \):

\[
p^{i}(x_{in}, t) = 0 \quad \text{and} \quad p(x_{out}, t) = p_{c}^{(i)}(S_{B}) \quad \text{for all} \quad t > 0.
\]
The boundaries of the domain are given by: \( x_{in} = -10 \) and \( x_{out} = 500 \). Since we do not include hysteresis in the relative permeabilities, the flux function \( F \) depends only on \( S \) and is determined by:

\[
 f(S) = \frac{S^2}{S^2 + (1 - S)^2} \quad \text{and} \quad N_g = 1.
\]

The numerical results presented in this subsection are related to \( t = t_{end} = 300 \). For the parameters of the initial condition, we take:

\[
 S_B = 0.1, \quad S_T = 0.4 \quad \text{and} \quad l = 1.
\]

Based on these data, some of the variables and constants occurring in Section 7.3.1 and Figure 7.3 are computed, i.e:

\[
 \tilde{S} \approx 0.3138, \quad \bar{S} \approx 0.5909, \quad S_o \approx 0.4393, \quad S^* \approx 0.4111 \quad \text{and} \quad \tilde{S}^* \approx 0.8132.
\]  
(7.71)

Moreover, the curves for \( \gamma \) and \( \beta \) are determined (see Figure 7.17). Observe that, from our choice, \( S_B < S_o \) and \( S_T \in (\tilde{S}, S^*) \). Next, the characteristic \( \tau \)-values for drainage and imbibition are computed. Using (7.50) and given parameters, we obtain:

\[
 \tau_i = 0.0452 \quad \text{and} \quad \tau_d = 0.2620.
\]

Since the requirements listed in Theorem 7.1 are all fulfilled, we can compare the numerical results with the claims contained in the theorem. For this purpose, we choose \( \tau \) from the following set:

\[
 \tau \in \{0.045, 0.25, 1.0, 2.0\},
\]
and study the resulting $S$-$p$ orbits. Considering Figure 7.18, it can be observed that for $\tau < \tau_i$ monotone saturation waves are produced by the numerical model linking $E^i_{SB}$ and $E^d_{ST}$. In the other cases, a saturation overshoot can be detected, where for $\tau_i < \tau < \tau_d$ the orbit ends up at the equilibrium point $E^d_{ST}$ and for $\tau > \tau_d$ the orbits spiral around the segment $E^i_{ST}E^d_{ST}$. If we choose larger values of $\tau$, the corresponding $S_m(\tau, S_T)$ value of the orbit increases. This supports the claims of Corollary 7.1 and Proposition 7.4. Similar results including variation of saturation with $\xi$ can be found in [169]. The parameter choice considered so far, corresponds to the solution class $A$ (see (7.51)), whose entropy solution consists of a single shock without any saturation overshoots (see Figure 7.20 (top)). However, there are two further solution classes, $B$ and $C$ (see (7.51)), arising in the context of Scenario A, represented by entropy solutions (7.57) and (7.58). In case of solution class $B$, the entropy solution is given by saturation plateau that is formed by an imbibition wave followed by a drainage wave. The saturation at plateau level is denoted by $\hat{S}_B(\tau)$. For solution class $C$, the entropy solution exhibits a rarefaction wave connecting $S_T$ with $\hat{S}_B(\tau)$, which is connected to $S_B$ by a shock. To observe these cases numerically, we compute the $\hat{S}_B(\tau)$ and $\hat{S}_B(\tau)$ curves introduced in Definition 7.1, see Figure 7.19.
In the figure we fix $\tau = 1$ and vary $S_T$ so that the pairs $(S_T, \tau)$ belong to one of the sets $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$. The results are shown in Figure 7.20 with the (left) plot showing the variation of $S$ with $x$, and the (right) plot showing the profiles in the $S$-$p$ phase plane. The curves corresponding to Set $\mathcal{A}$ show a direct travelling wave connecting $S_B$ and $S_T = 0.35$. Some oscillatory behaviour around $S_T$ can be observed since $\tau$ is comparatively large, however, the existence of a single travelling wave between $S_B$ and $S_T$ implies that these states are connectable by an admissible shock in the hyperbolic limit. Next, choosing $S_T = 0.55$, $(S_T, \tau)$ lies in Set $\mathcal{B}$, and a solution consisting of an imbibition wave followed by a drainage wave is computed in accordance with the theory. Again, small oscillations are seen in the drainage wave part which is expected from Corollary 7.2 since $\tau$ is large. The resulting plateau has saturation $.7158$, whereas, the prediction from Figure 7.19 is $\hat{S}_B(\tau) = .7254$. Finally, for $S_T = 0.8$, the pair $(S_T, \tau)$ belongs to the Set $\mathcal{C}$. The numerical solution exhibits a shock-like structure followed by a plateau and they coincide with the imbibition wave of Set $\mathcal{B}$ on both plots of Figure 7.20. Moreover, a rarefaction wave between $\hat{S}_B(\tau)$ and $S_T$ is detected. Thus, we conclude that the saturation profiles in Figure 7.20 correspond to the entropy solutions depicted in Figure 7.13 and the numerical results are in agreement with the theory.

### 7.5.2 Numerical results for Scenario B

In case of Scenario B, we choose the following boundary conditions with respect to the pressure variable. As in the previous subsection, they are of Neumann type at
Fronts in two-phase porous flow problems: effects of hysteresis and dynamic capillarity

Figure 7.20: Numerical solutions corresponding to different \((S_T, \tau)\) pairs from solution classes \(A\), \(B\) and \(C\), marked in Figure 7.19. Here, \(\tau = 1\) is fixed and \(S_T\) is chosen from \((0.35, 0.55, 0.8)\). The (left) plot shows the variation of \(S\) with \(x\), whereas, the (right) plot shows \(p\) vs. \(S\). The saturation plateau for the Sets \(B\) and \(C\) is observed at \(\hat{S}_B = 0.7158\).

Moreover, the boundaries of the domain are given by: \(x_{in} = -10\) and \(x_{out} = 190\). To make matters interesting, contrary to the previous subsection, we do not start with an imbibition state for \(S_B\), but with a drainage state. Due to the fact that we consider hysteresis both in the capillary pressure and relative permeabilities, fractional flow functions are introduced both for imbibition and for drainage. We use

\[
\frac{f^{(i)}}{f^{(d)}}(S) = \frac{S^2}{S^2 + 3(1 - S^2)}, \quad \frac{f^{(d)}}{f^{(d)}}(S) = \frac{S^2}{S^2 + 2(1 - S^2)} \quad \text{with } N_g = 0,
\]

and define \(F^{(i)}\) and \(F^{(d)}\) accordingly. We verified numerically that if \(S_T > S_B\) and \(F^{(i)}(S_T) < F(S_B, p_B)\) then the solution is frozen in time in the sense that \(S(x, t) = S(x, 0)\) for all \(t > 0\). This is what was discussed in Remark 7.8. To verify Propositions 7.11 and 7.12 and entropy solutions (7.65)-(7.68), we show two results: \(S_B = S_{B,1} = 0.3\), \(S_T = S_{T,1} = 0.95\) and \(S_B = S_{B,2} = 0.95\), \(S_T = S_{T,2} = 0.3\) both for \(\tau = 0.02\). Let the corresponding solutions be \((S^{(i)}, p^{(i)})\) and \((S^{(d)}, p^{(d)})\). Since \(S_{T,1} > S_l\) for the first case (see Definition 7.2) and \(\tau\) is small, from (7.66) it is expected that the entropy solution will have a shock from \(S_{B,1}\) to \(S_l\), followed by a rarefaction wave from \(S_l\) to \(S_{T,1}\). This is exactly what is seen from the viscous profiles obtained numerically, see Figure 7.21. Similarly, for the second case, since \(S_{T,2} < S_d\) and \(\tau\) is small, we see from Figure 7.21 a viscous solution resembling a drainage shock followed by a rarefaction wave, as predicted in (7.68). Next, we investigate whether a stable plateau is formed.
Figure 7.21: The viscous solutions for $S_{B,1} = 0.3$, $S_{T,1} = 0.95$ denoted by $(S_{B}(i), p_i)$ and $S_{B,2} = 0.95$, $S_{T,2} = 0.3$ denoted by $(S_{d}(d), p_d)$ with boundary conditions (7.72) and $\tau = 0.02$ fixed. In the (left) plot, the solutions are shown in the $F-S$ plane and in the (right) plot the saturations are plotted as functions of $x$. The points $U_{B,1}$ and $U_{B,2}$ and the saturations $\bar{S}_{i}(U_{B,1})$ and $\bar{S}_{d}(U_{B,2})$, introduced in Definition 7.2, are marked. The results agree with the predictions of Propositions 7.11 and 7.12 and Section 7.4.3.

for suitable parameter values by an imbibition wave and an ensuing drainage wave, as predicted in Proposition 7.13. This happens only if $\tau > \tau^*_i(S_T)$, since in this case, a monotone connection between $(S_B, p_B)$ and $(S_T, p_c(S_T))$ does not exist. To this end, in the numerical experiment we have used the following parameters:

$$S_B = 0.3, \quad S_T = 0.5 \quad \text{and} \quad \tau = 0.5.$$  

For a stable saturation plateau, the velocities of the imbibition wave, connecting $S_B$ and $S_P$, and the drainage wave, connecting $S_P$ and $S_T$, have to be equal, i.e. if $c^P(i)$ and $c^P(d)$ are denoting the two wave-speeds, then

$$c^P(i) = \frac{F(i)(S_P) - F(d)(S_B)}{S_P - S_B} = \frac{F(i)(S_P) - F(d)(S_T)}{S_P - S_T} = c^P(d),$$

where $S_P$ stands for the saturation of the plateau. Geometrically, this equality is fulfilled, if the points

$$(S_B, F(S_B, p_B)), (S_T, F(d)(S_T)) \quad \text{and} \quad (S_P, F(i)(S_P))$$

are located on the same line. This is precisely the condition that the solutions $(S^P(i), p^P(i))$ and $(S^P(d), p^P(d))$ of Proposition 7.13 satisfy. Drawing a line through the given points for $S_B = 0.3$ and $S_T = 0.5$ (see Figure 7.22), we obtain that a stable plateau should be located at $S_P \approx 0.634$. As seen from Figure 7.22, the orbit in the
$S$-$F$ plane stabilizes exactly at $S_p \approx 0.634$ and all the three points line up. Considering Figure 7.23, we observe that the saturation plateau is in a transient state in the beginning, but it stabilizes at $S_p \approx 0.634$ for longer times, as the speeds of the imbibition and drainage waves match.

![Figure 7.22: The orbit in the $S$-$F$ plane representing a stable saturation plateau for Scenario B. The equilibrium points for this orbit are shown on the flux curves.](image)

### 7.6 Conclusion

In this work, a one-dimensional two-phase flow model has been analysed for infiltration problems. For simplicity, we have assumed that the medium is homogeneous and a constant total velocity is prescribed at the boundary. Dynamic and hysteretic effects are included in the capillary pressure with transitions between drainage and imbibition processes being modelled by a play-type hysteresis model having vertical scanning curves. Relative permeabilities are modelled as functions of saturation and capillary pressure in order to make their hysteretic nature explicit.

The focus being on travelling waves (TW), the system of partial differential equations is transformed into a dynamical system. This system is analysed for two different scenarios, A and B. In Scenario A, the hysteresis appears only in the capillary pressure, and we consider a broad range of dynamic capillarity terms, from small to large ones. In Scenario B, hysteresis is included in both the relative permeabilities and in the capillary pressure, whereas the dynamic capillary effects are kept small. For each scenario the existence of TW solutions is studied. In particular, we show that if the dynamic capillary effects exceed a certain threshold value, the TW profiles...
become non-monotonic. Such results complement the analysis in Chapters 4 and 5 done for the unsaturated flow case, respectively in [248, 254] for two-phase flow but without hysteresis. From practical point of view, the present analysis provides a criterion for the occurrence of overshoots in two-phase infiltration experiments.

Based on the TW analysis, we give admissibility conditions for shock solutions to the hyperbolic limit of the system. Motivated by the hysteretic and dynamic capillarity effects, such solutions do not satisfy the classical entropy condition. This is because the standard entropy solutions to hyperbolic two-phase flow models are obtained as limits of solutions to classical two-phase flow models, thus not including hysteresis and dynamic capillarity. In particular, for the imbibition case of Scenario A, apart from the classical solutions, there can be solutions consisting of (i) an imbibition shock followed by a rarefaction wave having non-matching speeds, or (ii) an imbibition shock followed by a drainage shock resulting in a growing saturation plateau.

Figure 7.23: Saturation profiles for different time points. Besides the initial condition (top left) and the final saturation profile (bottom right), two intermediate profiles are shown, which have the form of a plateau. Contrary to the final saturation profile their plateaus are not stable, since the speeds of the imbibition and the drainage fronts are equal only for $S_P = 0.634$. 
(overshoot) in between. This is similar to the results in [248, 254]. In Scenario B, the entropy solutions are shown to depend also on the initial pressure. In particular, if certain parametric conditions are met, the solutions may include ones featuring a stable saturation plateau between an imbibition front and a drainage front, both travelling with the same velocity. Such solutions are obtained e.g. in [216], but only after generating the overshoot through a change in the boundary condition. All cases mentioned above have been reproduced by numerical experiments, in which a good resemblance has been observed between the TW results and the long time behaviour of the solutions to the original system of partial differential equations.

A discussion on the implication of these results can be found in Chapter 11 where we also compare them with experimental observations.
Part II:
Numerical methods
Chapter 8

A linear domain decomposition method for partially saturated flow in layered porous media

8.1 Introduction

In this work we focus on the unsaturated flow of water through a porous medium (e.g. the subsurface) occupying the domain $\Omega \subset \mathbb{R}^d$ ($d \in \{1, 2, 3\}$). The medium is possibly heterogeneous and/or layered. Besides water, a second phase (air) is present, which is assumed to be at a constant (atmospheric) pressure. This situation is described by the Richards equation, the dimensional form of which is given in (1.8)-(1.9). Here it is written in the pressure formulation as

$$\phi \partial_t S(p) - \nabla \cdot \left[ \frac{K}{\mu} k_r(S(p)) \nabla (p + z) \right] = 0, \quad (8.1)$$

see e.g. [115], originally [212]. In the above, $\phi$ denotes the porosity, $S$ is the water saturation, $p$ the water pressure, $k_r$ the relative permeability, $K$ the intrinsic permeability and $z = -\rho_w g x_1$ is the gravitational term in the direction of the $x_1$-axis. Finally, $g$ is the gravitational acceleration, $\rho_w$ the water density and $\mu$ its viscosity. With $T > 0$ being a maximal time, the equation is defined for the time $t \in (0, T)$ on

the bounded Lipschitz domain $\Omega$. Observe that, instead of writing the pressure as a function of saturation as was done in (1.9), in the pressure formulation the relation is inverted and saturation is written as a function of pressure, i.e.

$$S = (p_c)^{-1},$$

with $p_c$ being the capillary pressure introduced in Section 1.1. For this chapter, we disregard the non-equilibrium effects discussed in Section 1.2 and focus on the standard model where (8.1) holds.

Below we propose a domain decomposition (DD) scheme for the numerical solution of (8.1). To this aim we assume that $\Omega$ is partitioned into two subdomains $\Omega_l$ ($l \in \{1,2\}$) separated by a Lipschitz-continuous interface $\Gamma$, see Figure 8.1. In other words, one has $\Omega = \Omega_1 \cup \Omega_2 \cup \Gamma$. The restriction to two subdomains is made for the ease of presentation, but the scheme can be extended straightforwardly to more subdomains, see Remark 8.3 and Section 8.4.5. In each $\Omega_l$ ($l \in \{1,2\}$) we use the physical pressure $p_l$ as the primary variable. Furthermore, the permeability and porosity in each of the subdomains may be different and even discontinuous, which is the case in a heterogeneous medium consisting of block-type heterogeneity (like a fractured medium).

![Illustration of a layered soil domain](image)

**Figure 8.1**: Illustration of a layered soil domain $\Omega = \Omega_1 \cup \Omega_2 \subset \mathbb{R}^d$ with fixed interface $\Gamma$. Also shown are the normal vectors along the interface.

In view of its relevance in diverse applications, Richards equation has been studied extensively, both analytically and numerically. We restrict ourselves here by mentioning [9, 10] for the existence of weak solutions and [181] for the uniqueness. Numerical schemes for the Richards equation, or in general for degenerate parabolic equations, are analysed in [13, 140, 177, 191, 202, 205, 207, 268]. Most of the papers are considering the backward Euler method for the time discretization in view of the low regularity of the solution, see [9], and to avoid restrictions on the time step size.

Different approaches with regard to spatial discretization have been considered. Galerkin finite elements were used in [27, 177, 234]. Discontinuous Galerkin finite element schemes for flows through (heterogeneous) porous media have been studied...
in [15, 16, 86]. Finite volume schemes including multipoint flux approximation ones for the Richards equation are analysed in [92, 140], and mixed finite elements are analysed in [13, 18, 202, 205, 207, 268]. Such schemes are locally mass conservative.

Applying the Kirchhoff transformation [9] brings the mathematical model to a form that simplifies mathematical and numerical analysis, see e.g. [13, 177, 205, 207]. However, the transformed unknown is not directly related to a physical quantity like the pressure, and therefore a postprocessing step is required after a numerical approximation of the solution has been obtained. Alternatively, one may develop numerical schemes for the original formulation and in terms of the physical quantities. Nevertheless, when proving the convergence rigorously, one often resorts to a Kirchhoff transformed formulation as an intermediate step. Otherwise, sufficient regularity of the solution has to be assumed, e.g. by avoiding cases where the medium is completely saturated, or completely dry. We point out that in this work we will not make use of the Kirchhoff transformation, keeping the equation in a more physically relevant form, suitable for applications.

If implicit methods are adopted for the time discretization, the (elliptic or fully discrete) problems obtained at each time step are nonlinear. For solving these, different approaches have been proposed. Examples are the Newton method [26, 39, 183], the Picard/modified Picard method [57, 158], and the Jäger-Kacur method [124, 133]. We refer to [206] for the convergence analysis of such nonlinear schemes. Assuming that the initial guess is the solution from the previous time step, the convergence of such schemes can only be guaranteed under severe restriction of the time step in terms of the mesh size. Additionally, regularisation of the problem is required, which prevents the Jacobian from becoming singular, see [206]. Such difficulties do not appear when the L-scheme is being used, which is a fixed point scheme transforming the iteration into a contraction, [192, 203, 234]. The convergence is merely linear but in a better norm ($H^1$) and requires no regularisation or severe constraint on the time step. We also refer to [155] for a combination of the Newton method and the L-scheme, and to Chapter 9 for a modified version of the L-scheme which gives much faster convergence rates while preserving its stability. Moreover, we mention [202] for the application of the L-scheme to Hölder instead of Lipschitz continuous nonlinearities.

Independently of the chosen discretization method and of the linearisation scheme, domain decomposition (DD) methods offer an efficient way to reduce the computational complexity of the problem, and to perform calculations in parallel. This is in particular interesting whenever domains with block type heterogeneities are considered, as DD schemes allow decoupling the problem into subproblems defined in different homogeneous subdomains and solving these numerically in parallel. We refer to [199] for a detailed discussion of linear DD methods and to [75] for a general introduction into the subject. Comprehensive studies of nonlinear DD schemes in the field of fluid dynamics can be found in [79, 104, 239]. For articles strictly related to porous media flow models, we refer to [232, 233] for an overview of different over-
lapping domain decomposition strategies. Linear and nonlinear additive Schwartz methods are compared, and the use of such methods as linear and nonlinear preconditioners is discussed. Regardless of the type of the DD scheme, choosing the optimal parameters is a key issue. Such aspects are analysed e.g. in [97, 98]. We also refer to [5] for a DD algorithm for porous media flow models, where a-posteriori estimates are used to optimise the parameters and the number of iterations.

Recall that the Richards equation is a nonlinear evolution equation. For solving this type of equation, methods like parareal [101] and wave-form relaxation [99, 100] have been proposed. The main ideas there are to decompose the problem into separate problems defined in time/space-time domains. DD methods for the Richards equation are discussed in [28, 29]. In these papers the domain is decomposed into multiple layers and the Richards models restricted to adjacent layers are coupled by Robin type boundary conditions. The approach uses nonoverlapping domain-decomposition and generalises the ideas of the method introduced in [154] for linear elliptic problems (see also [120, 198]), leading to decoupled, nonlinear problems in the subdomains.

Here we consider a linear DD scheme that computes the time-discrete solution in the full domain by solving linear elliptic subproblems in the subdomains. A nonoverlapping DD scheme (referred to henceforth as LDD scheme) inspired by the DD method introduced in [154] is defined. The LDD iterations are linear, based on the L-scheme. This approach differs from the one commonly used when dealing with nonlinear elliptic problems in the context of DD. In most cases, the DD iterations lead to nonlinear subproblems. For solving these, iterative methods in each subdomain are applied. In our approach, the linearisation step is part of the DD iterations, which reduces the computational time. More precisely, the L-scheme idea is combined with the nonoverlapping DD scheme such that the equations defined in each subdomain along with the Robin type coupling conditions on the interface become linear. For the resulting scheme, we prove rigorously the convergence under mild restrictions on the time step, and provide numerical examples supporting the theoretical findings and demonstrating its effectiveness.

The paper is structured as follows. In Section 8.2 we present the mathematical model and introduce the DD scheme. Section 8.3 contains the analysis of the scheme. Finally, Section 8.4 provides numerical experiments in two spatial dimensions, together with a study of the practical performance of the scheme. This includes a comprehensive comparison (including robustness and efficiency) between the proposed DD scheme and standard monolithic schemes based on Newton, modified Picard as well as the L-scheme. Moreover, we give an example of a four subdomain implementation of the LDD scheme and illustrate the benefit of distributing the workload of each subdomain to dedicated CPUs.
8.2 Problem formulation and iterative scheme

8.2.1 Problem formulation

Recall that $T > 0$ and $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain partitioned in two subdomains $\Omega_{1,2}$, separated by the Lipschitz-continuous interface $\Gamma$. The boundary of $\Omega$ is denoted by $\partial \Omega$ and the portions of $\partial \Omega$ that are also boundaries of $\Omega_j$ are denoted by $\partial \Omega_j$ (see also Figure 8.1). To ease the presentation, the two subdomains are assumed to be homogeneous and isotropic, i.e. for $l \in \{1,2\}$ we have different relative permeabilities $k_{rl} = k_{rl}$ on each subdomain $\Omega_l$, different capillary relationships $S_l = S_l$, the intrinsic permeabilities $K_l = K_l$ are scalars and the two porosities $\phi_l$ are constants. Our analysis holds in the more general context too, if (A1)-(A4), listed in Chapter 1, are assumed. Motivated by Section 1.1.3, the product $K_l k_{rl}$ in (8.1) is abbreviated by $k_l$ henceforth. We solve equation (8.1) in $\Omega$ together with initial and homogeneous Dirichlet boundary conditions. We refer to [29,219] for more general conditions, including outflow-type ones.

On the two subdomains, the problem transforms into two subproblems, coupled through two conditions at the interface $\Gamma$: the continuity of the normal fluxes and the continuity of the pressures. With the fluxes $F_l := -k_l(S_l(p_l))\nabla(p_l + z)$, (8.1) becomes:

\[
\begin{align*}
\partial_t S_l(p_l) + \nabla \cdot F_l &= 0 & \text{in } \Omega_l \times (0,T], \quad (8.2a) \\
F_1 \cdot n_1 &= -F_2 \cdot n_2 & \text{on } \Gamma \times [0,T], \quad (8.2b) \\
p_1 &= p_2 & \text{on } \Gamma \times (0,T], \quad (8.2c) \\
p_l &= 0 & \text{on } \partial \Omega_l \times (0,T]. \quad (8.2d)
\end{align*}
\]

This is closed by the initial conditions $p_l(\cdot,0) := p_{l,0}$ in $\Omega_l$, where $p_l$ is the water pressure on $\Omega_l$, $l \in \{1,2\}$, and $k_l$, $S_l$ are (given) scaled relative permeability and saturation functions respectively. In the above, $n_l$ stands for the outer unit normal vector at $\partial \Omega_l$.

Semi-discrete formulation (discretization in time)

For the time discretization we let $\Delta t := \frac{T}{N}$ be the time step for a given $N \in \mathbb{N}$. Then $p^n_l$ is the approximation of the pressure $p_l$ at time $t_n = n\Delta t$. The Euler implicit discretization of (8.2) reads for $l \in \{1,2\}$,

\[
\begin{align*}
S_l(p^n_l) - S_l(p^{n-1}_l) + \Delta t \nabla \cdot F^n_l &= 0 & \text{in } \Omega_l, \quad (8.3a) \\
F^n_1 \cdot n_1 &= -F^n_2 \cdot n_2 & \text{on } \Gamma, \quad (8.3b) \\
p^n_1 &= p^n_2 & \text{on } \Gamma, \quad (8.3c) \\
p^n_l &= 0 & \text{on } \partial \Omega_l. \quad (8.3d)
\end{align*}
\]
where $F_l^n := -k_l(S_l(p_l^n))\nabla(p_l^n + z)$ is the flux at time step $t_n$. Observe that (8.3b) and (8.3c) are the coupling conditions at the interface $\Gamma$.

### 8.2.2 The LDD iterative scheme

Take $I \in \{1,2\}$. If $(p_1^{n-1}, p_2^{n-1})$ is known, $(p_1^n, p_2^n)$ can be obtained by solving the nonlinear problem (8.3). To this end, we define an iterative scheme that uses Robin type conditions at $\Gamma$ to decouple the subproblems in $\Omega_i$, while linearising the system by adding stabilisation terms that cancel each other in the limit (see e.g. [155,192]). Specifically, assuming that for some $i \in \mathbb{N}$ the approximations $p_l^{n,i-1}$ and $g_l^{i-1}$ are known, we seek $(p_l^{n,i}, p_2^{n,i})$ solving the problems

\begin{align}
L_ip_l^{n,i} - L_ip_l^{n,i-1} + \Delta t \nabla \cdot F_l^{n,i} &= -S_l(p_l^{n,i-1}) + S_l(p_l^{n-1}) \quad \text{in } \Omega_i, \quad (8.4a) \\
F_l^{n,i} \cdot n_I &= g_l^i + \lambda p_l^{n,i} \quad \text{on } \Gamma \times [0,T], \quad (8.4b) \\
g_l^i &= -2\lambda p_l^{n,i-1} - g_l^{i-1}. \quad (8.4c)
\end{align}

Following the previously introduced notation, $F_l^{n,i} := -k_l(S_l(p_l^{n,i-1}))\nabla(p_l^{n,i} + z)$ denotes the linearised flux at iteration $i$. By $\lambda \in (0,\infty)$, we denote a (free to be chosen) parameter used to weight the influence of the pressure on the interface conditions at $\Gamma$. The parameters $L_l > 0$ must adhere to some mild constraints in order for the scheme to converge, which will be discussed later. Other than that, they are arbitrary. The iteration starts with

$$p_l^{n,0} := p_l^{n-1}, \quad \text{and} \quad g_l^{0} := F_l^{n-1} \cdot n_I - \lambda p_l^{n-1}.$$ 

Clearly, the difference $L_ip_l^{n,i} - L_ip_l^{n,i-1}$ is vanishing in case of convergence.

**Remark 8.1.** The usage of the terms $g_l^i$ and of the parameter $\lambda$ is motivated by the following. With the notation $f_l^n := F_l^n \cdot n_I$, the transmission conditions (8.3b)-(8.3c) become $f_1^n = -f_2^n$ and $p_1^n = p_2^n$. For any $\lambda \neq 0$, these are equivalent to

\begin{align}
f_1^n &= (f_2^n - \lambda p_2^n) + \lambda p_1^n, \\
f_2^n &= (f_1^n - \lambda p_1^n) + \lambda p_2^n, \quad (8.5)
\end{align}

Denoting the terms between brackets by $g_l$, one obtains

\begin{align}
f_1^n &= g_1 + \lambda p_1^n, \quad \text{and} \quad g_1 = -2\lambda p_2^n - g_2, \\
f_2^n &= g_2 + \lambda p_2^n, \quad \text{and} \quad g_2 = -2\lambda p_1^n - g_1. \quad (8.6)
\end{align}

The conditions in (8.4b)-(8.4c) are the linearised counterparts of (8.6).
Remark 8.2 (different decoupling formulations). The decoupled conditions in (8.3b)-(8.3c) can be formulated as convex combinations of the terms \( g \) and \( p \), namely

\[
F_l^{n,i} \cdot n_l = (1 - \lambda)g_l^i + \lambda p_l^{n,i},
\]

\[
(1 - \lambda)g_l^i = -2\lambda p_{3-l}^{n,i-1} - (1 - \lambda)g_{3-l}^{i-1}.
\]

The convergence analysis, presented below, remains the same for this formulation. However, the DD scheme using this convex formulation showed a slower convergence in the numerical experiments than when (8.4b)-(8.4c) was used. Such aspects are discussed in Section 8.4. In view of this, in what follows we restrict the analysis to the initial formulation.

Remark 8.3 (multi domain formulation). The scheme and the analysis presented here for two subdomains can be straightforwardly extended to more subdomains. To see this assume that \( \Omega = \cup_{l=1}^L \overline{\Omega_l} \) is decomposed into \( W \) open subdomains \( \Omega_l \), and \( \Omega_l \cap \Omega_r = \emptyset \) for all \( l, r \in \{1, \ldots, W\} \), \( l \neq r \). For all \( l \), let \( I_l = \{ r : r \in \{1, \ldots, W\} \), \( r \neq l \), such that \( \text{meas}_{d-l}(\overline{\Omega_l} \cap \overline{\Omega_r}) > 0 \). For all \( r \in I_l \), let \( \Gamma_{l,r} = \Omega_l \cap \Omega_r \) be the interface separating the subdomains \( \Omega_l \) and \( \Omega_r \) and let \( n_{l,r} \) be the outwards pointing unit normal of \( \Omega_l \) on \( \Gamma_{l,r} \). Furthermore, define the outer boundaries of subdomains as \( \partial \Omega_l = \overline{\Omega_l} \cap \partial \Omega \). Observe that \( r \in I_l \) implies \( l \in I_r \) for all \( l, r \). Moreover, \( \Gamma_{l,r} = \Gamma_{r,l} \) and \( n_{l,r} = -n_{r,l} \). Then, the LDD scheme for \( N \) subdomains amounts to solving the following subproblems:

\[
L_l p_l^{n,i} + \Delta t \nabla \cdot F_l^{n,i} = L_l p_l^{n,i-1} - (S_l(p_l^{n,i-1}) - S_l(p_l^n)), \quad \text{on } \Omega_l,
\]

\[
F_l^{n,i} \cdot n_{l,r} = g_{l,r}^i + \lambda_{l,r} p_l^{n,i}, \quad \text{on } \Gamma_{l,r} \text{ for all } r \in I_l,
\]

\[
g_{l,r}^i = -2\lambda_{l,r} p_{3-l}^{n,i-1} - g_{3-l}^{i-1},
\]

\[
p_l^{n,i} = 0 \quad \text{on } \partial \Omega_l,
\]

for all \( l \in \{1, \ldots, W\} \). Here, \( \lambda_{l,r}, \lambda_r \) are parameters that can be chosen freely as long as they satisfy \( \lambda_{l,r} = \lambda_r \) for \( l \in \{1, \ldots, W\} \) and \( r \in I_l \). For ease of presentation we have restricted ourselves to discussing the situation on two domains.

Remark 8.4 (Applicability of the LDD scheme). The idea of the LDD scheme is by far not restricted to the Richards equation and can be extended to other models. It is well suited for problems involving monotonic nonlinearities. It has already been extended to the two-phase case [227], and to the case involving non-equilibrium effects [159]. Further possibilities for extension include the reactive transport models, and coupled flow and poromechanics models, see e.g. [33].

Before formulating the main result we specify the notation that will be used below: \( L^2(\Omega) \) is the space of Lebesgue measurable, square integrable functions over \( \Omega \). \( H^1(\Omega) \) contains functions in \( L^2(\Omega) \) having also weak derivatives in \( L^2(\Omega) \). \( H_0^1(\Omega) = \)
A linear domain decomposition method for unsaturated flow in layered porous media

\[ C_0^{\infty}(\Omega)^{H^1} \] where the completion is with respect to the standard \( H^1 \) norm and \( C_0^{\infty}(\Omega) \) is the space of smooth functions with compact support in \( \Omega \). The definition for \( H^1(\Omega,l) \) (\( l \in (1,2) \)) is similar. With \( \Gamma \) being a \((d-1)\) dimensional manifold in \( \Omega \), \( H^1(\Gamma) \) contains the traces of \( H^1 \) functions on \( \Gamma \) (see e.g. [41, 199]). Given \( u \in H^1(\Omega) \), its trace on \( \Gamma \) is denoted by \( u|_{\Gamma} \). Furthermore, the following spaces will be used

\[ \mathcal{V}_l := \left\{ u \in H^1(\Omega) \mid u|_{\partial\Omega_l} = 0 \right\}, \quad (8.8) \]

\[ \mathcal{V} := \left\{ (u_1, u_2) \in \mathcal{V}_1 \times \mathcal{V}_2 \mid u_1|_{\Gamma} \equiv u_2|_{\Gamma} \right\}, \quad (8.9) \]

\[ H^{1/2}_{00}(\Gamma) = \left\{ v \in H^{1/2}(\Gamma) \mid v = w|_{\Gamma} \text{ for } w \in H^1_0(\Omega) \right\}. \quad (8.10) \]

where the norms in the spaces \( \mathcal{V}_l \) are the standard \( H^1 \)-norms, with the straightforward extension to the product space \( \mathcal{V} \). \( \mathcal{V}' \) stands for the dual of \( \mathcal{V}_l \) with the norm

\[ \| F \|_{\mathcal{V}'_l} = \sup_{\varphi_l \in \mathcal{V}_l} \| F \varphi_l \|_{\mathcal{V}_l}. \]

Moreover, since the boundaries of \( \Omega \) and \( \Omega_l \) have a non-zero \((d-1)\)-measured intersection, the functions in \( \mathcal{V}_l \) are vanishing on this common part of the boundary. \( H^{1/2}_{00}(\Gamma)' \) denotes the dual space of \( H^{1/2}_{00}(\Gamma) \). \( \langle \cdot, \cdot \rangle_X \) will denote the \( L^2(X) \) scalar product, with \( X \) being one of the sets \( \Omega, \Omega_l \) (\( l \in (1,2) \)) or \( \Gamma \). Whenever self understood, the notation of the domain of integration \( X \) will be dropped. Furthermore, \( \langle \cdot, \cdot \rangle_\Gamma \) stands also for the duality pairing between \( H^{1/2}_{00}(\Gamma)' \) and \( H^{1/2}_{00}(\Gamma) \).

In what follows, we make the following assumptions for \( l \in (1,2) \):

(A1) \( k_l : \mathbb{R} \to [0,1] \) are Lipschitz continuous functions with Lipschitz constants \( L_{k_l} > 0 \).

(A2) There exists \( m \in \mathbb{R} \) such that \( 0 < m \leq k_1(S), k_2(S) \) for all \( S \in \mathbb{R} \).

(A3) \( S_l : \mathbb{R} \to \mathbb{R} \) are monotonically increasing and Lipschitz continuous functions with Lipschitz constants \( L_{S_l} > 0 \).

For later use we define \( L_k := \max\{ L_{k_1}, L_{k_2} \} \) and \( L_S := \max\{ L_{S_1}, L_{S_2} \} \).

In a simplified formulation, the main result in this paper is

**Theorem 8.1.** Assume (A1)-(A3). Let there exist a solution pair \((p_l^n, p_2^n)\) to (8.3) that additionally fulfills \( \sup_l \| \nabla (p_l^n + z) \|_{L^\infty} \leq M < \infty \) for \( l \in (1,2) \). Take \( L_l \) such that \( L_{S_l} < 2L_l \) and assume that the time step \( \Delta t > 0 \) is chosen small enough, so that for both \( l \) one has

\[ \Delta t < \frac{2m}{L_{k_l}^2 M^2 \left( \frac{1}{L_{S_l}} - \frac{1}{2L_l} \right)}. \quad (8.11) \]

Then the sequence of solution pairs \( \{(p_1^{n,i}, p_2^{n,i})\}_{i \geq 1} \) of (8.4) converges to \((p_1^n, p_2^n)\).

**Remark 8.5.** The precise form of Theorem 8.1 will be formulated in Section 8.3, after having defined a weak solution. The theorem above is given here for its better representation.
8.3 Analysis of the scheme

This section gives the convergence proof for the proposed scheme. The starting point is the Euler implicit discretization in Section 8.2. Assuming \((p_1^{n-1}, p_2^{n-1}) \in \mathcal{V}\) to be known, a weak formulation of (8.3) is given by

**Problem 8.1 (Semi-discrete weak formulation).** Find \((p_1^n, p_2^n) \in \mathcal{V}\) such that \(F_l^n \cdot n_l \in H^{1/2}_0(\Gamma)'\) for \(l \in \{1, 2\}\) and

\[
\langle S_l(p_1^n), \varphi_1 \rangle - \Delta t \langle F_l^n, \nabla \varphi_1 \rangle + \Delta t \langle F_{3-l}^n \cdot n_l, \varphi_1 \rangle_{\Gamma} = \langle S_l(p_1^{n-1}), \varphi_1 \rangle, \quad (8.12)
\]
for all \((\varphi_1, \varphi_2) \in \mathcal{V}\).

**Remark 8.6.** If \((p_1^n, p_2^n) \in \mathcal{V}\) is a solution of Problem 8.1, we have \(p_1^n|_{\Gamma} = p_2^n|_{\Gamma}\) by definition of \(\mathcal{V}\). Testing in (8.12) by an arbitrary \(\varphi_1 \in C_0^\infty(\Omega_l)\) shows that the distribution \(\nabla \cdot F_l^n\) is regular and in \(L^2\), yielding \(F_l^n \in H(\text{div}, \Omega_l)\) and

\[
S_l(p_1^n) - S_l(p_1^{n-1}) = -\Delta t \nabla \cdot F_l^n \quad \text{a.e. in } \Omega_l \quad (8.13)
\]

by the variational lemma. By Lemma III. 1.1 in [41], \(F_l^n \cdot n_l \in H^{-1/2}(\partial \Omega_l)\) and integrating by parts in (8.12) yields

\[
0 = -\langle F_l^n \cdot n_l, \varphi_1 \rangle_{\Gamma} + \langle F_{3-l}^n \cdot n_l, \varphi_1 \rangle_{\Gamma} \quad (8.14)
\]
for all \((\varphi_1, \varphi_2) \in \mathcal{V}\). Therefore

\[
F_l^n \cdot n_l = F_{3-l}^n \cdot n_l \quad (8.15)
\]
in \(H^{1/2}_0(\Gamma)'\) since the trace is a surjective operator.

Note additionally that Problem 8.1 is equivalent to the semi-discrete Richards equation on the whole domain, namely to find \((p_1^n, p_2^n) \in \mathcal{V}\) such that

\[
\langle S_1(p_1^n), \varphi_1 \rangle - \Delta t \langle F_1^n, \nabla \varphi_1 \rangle + \langle S_2(p_2^n), \varphi_2 \rangle - \Delta t \langle F_2^n, \nabla \varphi_2 \rangle = \langle S_1(p_1^{n-1}), \varphi_1 \rangle + \langle S_2(p_2^{n-1}), \varphi_2 \rangle, \quad (8.16)
\]
for all \((\varphi_1, \varphi_2) \in \mathcal{V}\).

**Remark 8.7.** By applying a Kirchhoff transform in each subdomain \(\Omega_l\), Problem 8.1 can be reformulated as a nonlinear transmission problem. The existence and uniqueness of a solution for such problems has been studied in [125, 126] for the case when \(\Omega_l\) is surrounded by \(\Omega_2\), and the common boundary is smooth.

Now we can give the weak form of the iterative scheme. Let \(n \in \mathbb{N}\) and assume that the pair \((p_1^{n-1}, p_2^{n-1}) \in \mathcal{V}\) is given. Furthermore, let \(\lambda > 0\) and \(L_l > 0\) \((l \in \{1, 2\})\) be fixed parameters and

\[
p_1^{n,0} := p_1^{n-1}, \quad \text{as well as} \quad s_l^0 := F_l^{n-1} \cdot n_l - \lambda p_l^{n-1}|_{\Gamma}.
\]

The iterative scheme is defined through...
Problem 8.2 (LDD, weak form). Let \( i \in \mathbb{N} \) and assume that the approximations \( \{p_i^{n,k}\}_{k=0}^{i-1} \) and \( \{g_i^k\}_{k=0}^{i-1} \) are known for \( i \in \{1,2\} \). Find \( (p_1^{n,i},p_2^{n,i}) \in \mathcal{V}_1 \times \mathcal{V}_2 \) such that

\[
L_i^{n,i}\varphi_i - \Delta t \langle F_i^{n,i} \nabla \varphi_i \rangle + \Delta t \langle \lambda p_i^{n,i} + g_i^i, \varphi_i \rangle_{\Gamma} = 0,
\]

\[
\langle g_i^i, \varphi_i \rangle_{\Gamma} = \langle -2\lambda p_{3-i}^{n,i-1} - g_{3-i}^{i-1}, \varphi_i \rangle_{\Gamma},
\]

holds for all \( (\varphi_1, \varphi_2) \in \mathcal{V}_1 \times \mathcal{V}_2 \).

8.3.1 Intuitive justification of the L-scheme

We start the analysis by taking a closer look at the formal limit of the L-scheme iterations in weak form and show that this is actually a reformulation of Problem 8.1.

Lemma 8.1 (Limit of the L-scheme). Let \( n \in \mathbb{N} \) be fixed and assume that the functions \( p_i^n \in \mathcal{V} \) and \( g_i \in H_{00}^{1/2}(\Gamma)' \) \( (i \in \{1,2\}) \) exist such that

\[
\langle S_l(p_l^n), \varphi_l \rangle - \langle S_l(p_l^{n-1}), \varphi_l \rangle - \Delta t \langle F_l^n \nabla \varphi_l \rangle + \Delta t \langle \lambda p_l^n + g_l, \varphi_l \rangle_{\Gamma} = 0,
\]

\[
\langle g_l, \varphi_l \rangle_{\Gamma} = \langle -2\lambda p_{3-l}^n - g_{3-l}^n, \varphi_l \rangle_{\Gamma},
\]

hold for all \( (\varphi_1, \varphi_2) \in \mathcal{V}_1 \times \mathcal{V}_2 \). Then the interface conditions

\[
p_l^n|_{\Gamma} = p_2^n|_{\Gamma} \quad \text{in} \quad H_{00}^{1/2}(\Gamma),
\]

\[
F_1^n \cdot n_1 = F_2^n \cdot n_1 \quad \text{in} \quad H_{00}^{1/2}(\Gamma)',
\]

are satisfied and \( (p_1^n, p_2^n) \) solves Problem 8.1. Moreover,

\[
g_l = -\lambda p_l^n|_{\Gamma} + F_l^n \cdot n_l
\]

in \( H_{00}^{1/2}(\Gamma)' \). Conversely, if \( (p_1^n, p_2^n) \in \mathcal{V} \) and \( g_l := -\lambda p_l^n|_{\Gamma} + F_l^n \cdot n_l \), then \( p_l^n \) and \( g_l \) solve the system (8.19), (8.20).

Remark 8.8. Lemma 8.1 states that solving Problem 8.1 is equivalent to finding a solution to (8.19), (8.20). This reformulation will be used to show that the L-scheme converges to a solution of Problem 8.1.

Proof. Writing out (8.20) for \( l \in \{1,2\} \) and subtracting the resulting equations yields \( p_1^n|_{\Gamma} = p_2^n|_{\Gamma} \) in the sense of traces. On the other hand, adding up these equations leads to \( (g_1 + g_2) = -\lambda (p_1^n|_{\Gamma} + p_2^n|_{\Gamma}) \). Inserting this into the sum of the equations (8.19) leads to (8.16), and by the equivalence to the semi-discrete formulation, we
We introduce the iteration errors domain is of class $C^1$ under the additional assumption that the functions $S$.

**Remark**

(The Convergence of the LDD scheme)

Theorem 8.2

**Proof.** We now prove the convergence result, which was announced in Theorem 8.1. We assume that the solution $(p^{n-1}_1, p^{n-1}_2)$ of Problem 8.1 at time step $(n-1)$ is known and let $p^{n,0}_i \in \mathcal{V}_i$ be arbitrary starting pressures (however, a natural choice is $p^{n,0}_i := p^{n-1}_i$).

Lemma 8.2 enables us to construct a sequence $\{p^{n,i}_i\}_{i \in \mathbb{N}^+} \in \mathcal{V}_i$ of solutions to Problem 8.2 and prove its convergence to the solution $(p^{n}_1, p^{n}_2)$ of Problem 8.1.

**Theorem 8.2** (Convergence of the LDD scheme). Assume there exists a solution $(p^{n}_1, p^{n}_2) \in \mathcal{V}$ to Problem 8.1 such that $\sup_i \|\nabla (p^n_i + z)\|_{L^\infty} \leq M < \infty$ for $i \in \{1, 2\}$ and let $g_i$ be as in (8.23). Let Assumptions (A1)-(A3) hold, $\lambda > 0$ and $L_i \in \mathbb{R}$ be given with $L_i s_{1/2} < L_i$. For arbitrary starting pressures $p^{n,0}_i \in \mathcal{V}_i$ and $g^{n}_i \in H^{1/2}_{00}(\Gamma')$ let $\{p^{n,i}_i, p^{n,i}_2\}_{i \in \mathbb{N}^+}$ be the sequence of solutions of Problem 8.2 and let $\{s^{n}_i\}_{i \in \mathbb{N}^+}$ be defined by (8.18). Assume further that the time step $\Delta t$ satisfies (8.11). Then $p^{n,i}_i \rightarrow p^{n}_i$ in $\mathcal{V}_i$ and $g^{n}_i \rightarrow g^{n}_i$ in $H^{1/2}_{00}(\Gamma')$ as $i \rightarrow \infty$.

**Remark 8.9.** The essential boundedness of the pressure gradients can be proven under the additional assumption that the functions $S_i$ are strictly increasing and the domain is of class $C^{1, \alpha}$, see e.g. [53, Lemma 2.1].

**Proof.** We introduce the iteration errors $e^{i}_{p, i} := p^{n,i}_p - p^{n,i}_i$ as well as $e^{i}_{g, i} := g^{n}_i - g^{n}_i$, add $L_i(p^n_i, \varphi_i) - L_i(p^n_i, \varphi_i)$ to (8.19) and subtract (8.17) to arrive at

$$
L_i(e^{i}_{p, i}, \varphi_i) + \Delta t \lambda(e^{i}_{p, i}, \varphi_i) + \Delta t(e^{i}_{g, i}, \varphi_i) + \Delta t(-F^n_i - k_i(S_i(p^{n,i-1})))\nabla(p^n_i + z) + k_i(S_i(p^{n,i-1})))\nabla(p^n_i + z) + F^n_i \nabla \varphi_i)

+ L_i(e^{i-1}_{p, i}, \varphi_i) - L_i(p^n_i, \varphi_i).

(8.25)
Inserting $q_I := e^{i}_{p,l}$ in (8.25) and noting that

$$L_I \langle e^{i}_{p,l} - e^{i-1}_{p,l}, e^{i}_{p,l} \rangle = \frac{L_I}{2} \left[ \|e^{i}_{p,l}\|^2 - \|e^{i-1}_{p,l}\|^2 + \|e^{i}_{p,l} - e^{i-1}_{p,l}\|^2 \right],$$

yields

$$\frac{L_I}{2} \left[ \|e^{i}_{p,l}\|^2 - \|e^{i-1}_{p,l}\|^2 + \|e^{i}_{p,l} - e^{i-1}_{p,l}\|^2 \right] + \langle S_I(p^n_I) - S_I(p^{n,i-1}_I), e^{i}_{p,l} - e^{i-1}_{p,l} \rangle + \Delta t \lambda \langle e^{i}_{p,l}, e^{i}_{p,l} \rangle_\Gamma =: I_1$$

$$=: I_2$$

$$- \Delta t \left( \|k_1(S_I(p^n_I)) - k_1(S_I(p^{n,i-1}_I))\| \nabla (p^n_I + z), \nabla e^{i}_{p,l} \right)$$

$$=: I_3$$

$$- \Delta t \left( k_1(S_I(p^{n,i-1}_I)) \nabla e^{i}_{p,l}, \nabla e^{i}_{p,l} \right)$$

$$=: I_4$$

We estimate now the terms $I_1$–$I_4$ in (8.26) one by one. By Assumption (A3), for $I_1$ we have

$$\frac{1}{L_S^2} \|S_I(p^n_I) - S_I(p^{n,i-1}_I)\|^2 \leq \langle S_I(p^n_I) - S_I(p^{n,i-1}_I), e^{i-1}_{p,l} \rangle.$$

(8.27)

$I_2$ is estimated by

$$|I_2| = \left| \left\langle S_I(p^n_I) - S_I(p^{n,i-1}_I), e^{i-1}_{p,l} - e^{i}_{p,l} \right\rangle \right|$$

$$\leq \frac{L_I}{2} \|e^{i-1}_{p,l} - e^{i}_{p,l}\|^2 + \frac{1}{2L_S^2} \|S_I(p^n_I) - S_I(p^{n,i-1}_I)\|^2.$$

(8.28)

For an $\epsilon_I > 0$ to be chosen below, we use Young’s inequality to deal with $I_3$, which can be estimated by

$$|I_3| = \left| \Delta t \left( k_1(S_I(p^n_I)) - k_1(S_I(p^{n,i-1}_I)) \right) \nabla (p^n_I + z), \nabla e^{i}_{p,l} \right|$$

$$\leq \Delta t \|k_1(S_I(p^n_I)) - k_1(S_I(p^{n,i-1}_I))\| \|\nabla e^{i}_{p,l}\|$$

$$\leq \Delta t L_{k_1} M \|S_I(p^n_I) - S_I(p^{n,i-1}_I)\| \|\nabla e^{i}_{p,l}\|$$

$$\leq \Delta t L_{k_1} M \epsilon_I \|S_I(p^n_I) - S_I(p^{n,i-1}_I)\|^2 + \Delta t \frac{L_{k_1} M}{4\epsilon_I} \|\nabla e^{i}_{p,l}\|^2,$$

(8.29)
where we used the Lipschitz continuity of $k_l$ and the assumption $\sup_i \|\nabla (p^l_i + z)\|_{L^\infty} < M$. Finally, by Assumption (A2) one has
\[
\Delta t \langle k_l(S_i(p_i^{n,i-1})\nabla e_{p,l}^i, \nabla e_{p,l}^i) \rangle \geq \Delta t m \|\nabla e_{p,l}^i\|^2, \tag{8.30}
\]
for $t_i$. Using the estimates (8.27)–(8.30), (8.26) becomes
\[
\frac{L_i}{2} \left[ \|e_{p,l}^i\|^2 - \|e_{p,l}^{i-1}\|^2 \right] + \frac{1}{L_{S_l}} \|S_l(p_i^{n}) - S_l(p_i^{n,i-1})\|^2 + \Delta t \lambda \langle e_{p,l}^i, e_{p,l}^i \rangle_T + \Delta t \langle e_{g,l}^i, e_{p,l}^i \rangle_T
\leq \left( \frac{1}{2L_i} + \Delta t L_k, M e_l \right) \|S_l(p_i^{n}) - S_l(p_i^{n,i-1})\|^2 + \Delta t \left( \frac{L_i M}{4 e_l} - m \right) \|\nabla e_{p,l}^i\|^2. \tag{8.26'}
\]
In order to deal with the interface term $\Delta t \langle e_{g,l}^i, e_{p,l}^i \rangle_T$, recall that $\langle \cdot, \cdot \rangle_T$ denotes the dual pairing of $H^{1/2}((\Gamma)^\prime$ and $H_{00}^{1/2}(\Gamma)$ and the $H^{1/2}(\Gamma)$-norm simultaneously. Subtracting (8.18) from (8.20), i.e. $e_{g,l}^i = -2\lambda e_{p,3-l}^i - e_{g,3-l}^i$, we get
\[
\|e_{p,l}^i\|^2_T = \frac{1}{4\lambda^2} \left( \|e_{g,l}^{i+1}\|^2_T - \|e_{g,l}^i\|^2_T - 4\lambda \langle e_{g,l}^i, e_{g,l}^i \rangle_T \right). \tag{8.31}
\]
We define $e_{p,g}^i := (e_{p,1}^i, e_{p,2}^i) \in \mathcal{V}_1 \times \mathcal{V}_2$ and $\|e_{p,g}^i\|^2 := \sum_{i=1}^2 \|e_{p,l}^i\|^2$. Similarly on $\Gamma$ we define $\langle e_{p,g}^i, e_{g}^i \rangle_T := \sum_{i=1}^2 \langle e_{p,l}^i, e_{g,l}^i \rangle_T$ and correspondingly $\|e_{g,l}^i\|^2 = \sum_{i=1}^2 \|e_{b,l}^i\|^2$. Summing in (8.31) over $l \in (1,2)$ gives
\[
\|e_{p,g}^i\|^2_T = \frac{1}{4\lambda^2} \left( \|e_{g,l}^{i+1}\|^2_T - \|e_{g,l}^i\|^2_T - 4\lambda \langle e_{g,l}^i, e_{g,l}^i \rangle_T \right). \tag{8.32}
\]
Doing the same for (8.26’) and inserting (8.32) leaves us with
\[
\frac{L_i}{2} \left[ \|e_{p,g}^i\|^2 - \|e_{p,g}^{i-1}\|^2 \right] + \sum_{i=1}^2 \frac{1}{L_{S_l}} \|S_l(p_i^{n}) - S_l(p_i^{n,i-1})\|^2
\] + \Delta t \frac{2}{4\lambda} \left( \|e_{g,l}^i\|^2_T - \|e_{g,l}^i\|^2_T \right) + \Delta t \left( \frac{L_i M}{4 e_l} - m \right) \|\nabla e_{p,l}^i\|^2
\leq \sum_{i=1}^2 \left( \frac{1}{2L_i} + \Delta t L_k, M e_l \right) \|S_l(p_i^{n}) - S_l(p_i^{n,i-1})\|^2. \tag{8.33}
\]
Now, summing for the iteration index $i = 1, \ldots, r$ and noticing the telescopic property of the sums, one gets
\[
\sum_{i=1}^r \sum_{i=1}^2 \left( \frac{1}{L_{S_l}} - \frac{1}{2L_i} - \Delta t L_k, M e_l \right) \|S_l(p_i^{n}) - S_l(p_i^{n,i-1})\|^2
\] + \Delta t \sum_{i=1}^r \sum_{i=1}^2 \left( \frac{L_i M}{4 e_l} - m \right) \|\nabla e_{p,l}^i\|^2
\leq \frac{L_i}{2} \left[ \|e_{p,l}^i\|^2 - \|e_{p,l}^{i-1}\|^2 + \Delta t \|e_{g,l}^i\|^2_T - \|e_{g,l}^{i+1}\|^2_T \right]. \tag{8.34}
\]
Now we choose \( \epsilon_l = \frac{L_{k,M}}{2m} \), hence \( m - \frac{L_{k,M}}{4\epsilon_l} = \frac{m}{2} > 0 \) for both \( l \). Recalling the restriction on \( L_l, \frac{1}{L_S} < \frac{1}{L_l} > 0 \), as well as the time step restriction \( \frac{1}{L_S} < \frac{1}{2L_l} - \Delta t \frac{L_{k,M}^2}{2m} > 0 \) for \( l \in \{1,2\} \), the estimates

\[
\sum_{i=1}^r \sum_{l=1}^m \left( \frac{1}{L_S} - \Delta t \frac{L_{k,M}^2}{2m} \right) \| S_l(p_l^0) - S_l(p_l^{n,i-1}) \|^2 \leq \frac{L_l}{2} \| e_p \|^2 + \frac{\Delta t}{4\lambda} \| e^1_p \|^2, \quad (8.35)
\]

\[
\Delta t \sum_{i=1}^r \sum_{l=1}^m \| \nabla e_p \|^2 \leq \frac{L_l}{2} \| e_p \|^2 + \frac{\Delta t}{4\lambda} \| e^1_p \|^2, \quad (8.36)
\]

follow for any \( r \in \mathbb{N} \). Since the right hand sides are independent of \( r \), we conclude that the series on the left are absolutely convergent and therefore \( \| S_l(p_l^n) - S_l(p_l^{n,i-1}) \|, \| \nabla e_p \| \to 0 \) as \( i \to \infty \). Moreover, (8.36) implies \( \| e^i_{p,l} \| \to 0 \) \( (i \to \infty) \) as well, by the Poincaré inequality. To show that \( e^i_{g,l} \to 0 \) in \( H_0^{1/2}(\Gamma) \) we subtract again (8.17) from (8.19) and consider test functions \( \varphi_l \in C_0^\infty(\Omega_l) \) to get

\[
-\Delta t \left( p^n_l - p^{n,i}_l, \nabla \varphi_l \right) = -L_l \left( e^i_{p,l}, \varphi_l \right) + \sum_{i} \left( e^{i-1}_{p,l}, \varphi_l \right) - \left( S_l(p_l^n) - S_l(p_l^{n,i-1}), \varphi_l \right).
\] (8.37)

Thus, \( \nabla \left( p^n_l - p^{n,i}_l \right) \) exists in \( L^2 \) and

\[
-\Delta t \nabla \left( p^n_l - p^{n,i}_l \right) = L_l \left( e^i_{p,l} - e^{i-1}_{p,l} \right) + S_l(p_l^n) - S_l(p_l^{n,i-1})
\] (8.38)

almost everywhere. Therefore, for any \( \varphi_l \in \mathcal{V}_l \) one has

\[
\left| \left( \nabla \left( p^n_l - p^{n,i}_l \right), \varphi_l \right) \right| \leq \frac{L_l}{\Delta t} \| e^i_{p,l} - e^{i-1}_{p,l} \| \| \varphi_l \| + \frac{1}{\Delta t} \| S_l(p_l^n) - S_l(p_l^{n,i-1}) \| \| \varphi_l \|. \quad (8.39)
\]

Abbreviating the left hand side of (8.39) as \( \psi^n_{i,l}(\varphi_l) \), (8.39) means for \( i \to \infty \),

\[
\sup_{\varphi_l, \varphi_l \neq 0} \frac{\| \psi^n_{i,l}(\varphi_l) \|}{\| \varphi_l \|_{H^1(\Omega_l)}} \leq \frac{L_l}{\Delta t} \| e^i_{p,l} - e^{i-1}_{p,l} \| + \frac{1}{\Delta t} \| S_l(p_l^n) - S_l(p_l^{n,i-1}) \| \to 0 \quad (8.40)
\]

as a consequence of (8.36). In other words \( \| \psi^n_{i,l} \|_{\mathcal{V}_l} \to 0 \) as \( i \to \infty \).

Starting again from (8.25) (without the added zero term), this time however inserting \( \varphi_l \in \mathcal{V}_l \), integrating by parts and keeping in mind (8.38) one gets

\[
\left( e^i_{g,l} , \varphi_l \right) \Gamma = -\lambda \left( e^i_{p,l} , \varphi_l \right) \Gamma + \left( \left[ F^n_l - F^{n,i}_l \right] \cdot \mathbf{n} , \varphi_l \right) \Gamma.
\] (8.41)
8.4 Numerical experiments

We already know that \( e_{p,l}^{i} \to 0 \) as \( i \to \infty \) so by the continuity of the trace operator, the first term on the right vanishes in the limit. For the last summand in (8.41) we use the integration by parts formula to obtain

\[
\left\langle \left[ F_{l}^{n} - F_{l}^{n,i} \right] \cdot n_{l}, \varphi_{l} \right\rangle_{\Gamma} = \psi_{l}^{n,i}(\varphi_{l}) + \left\langle F_{l}^{n} - F_{l}^{n,i}, \nabla \varphi_{l} \right\rangle.
\]  

(8.42)

While the first term on the right approaches 0, the second can be estimated by

\[
\left| \left\langle k_{l}(S_{l}(p_{l}^{n}))\nabla (p_{n}^{n} + z) - k_{l}(S_{l}(p_{l}^{n,i-1}))\nabla (p_{l}^{n,i} + z), \nabla \varphi_{l} \right\rangle \right| 
\leq L_{k_{l}} M \| S(p_{l}^{n}) - S(p_{l}^{n,i-1}) \| \| \varphi_{l} \|_{H^{1}(\Omega_{l})} + \| \nabla e_{p,l}^{i} \| \| \varphi_{l} \|_{H^{1}(\Omega_{l})},
\]  

(8.43)

where we used the same reasoning as in (8.29). With this we let \( i \to \infty \) in (8.42) to obtain

\[
\sup_{\varphi_{l} \in V_{l}} \left| \left\langle \left[ F_{l}^{n} - F_{l}^{n,i} \right] \cdot n_{l}, \varphi_{l} \right\rangle_{\Gamma} \right| \leq \| \psi_{l}^{n,i} \|_{V_{l}'} + L_{k_{l}} M \| S(p_{l}^{n}) - S(p_{l}^{n,i-1}) \| + \| \nabla e_{p,l}^{i} \| \longrightarrow 0.
\]

(8.44)

Finally, using the above and letting \( i \to \infty \) in (8.41) gives

\[
\sup_{\varphi_{l} \in V_{l}} \frac{\left| \left\langle e_{g,l}^{i}, \varphi_{l} \right\rangle_{\Gamma} \right|}{\| \varphi_{l} \|_{H^{1}(\Omega_{l})}} \longrightarrow 0.
\]

This shows that \( e_{g,l}^{i} \to 0 \) in \( H^{1/2}_{00}(\Gamma)' \) for both \( l \) and thus, concludes the proof. \( \square \)

Remark 8.10. Note that Theorem 8.2 states that if a solution to the semi-discrete coupled problem exists, then it is the limit of the iteration scheme. Since in the convergence proof we use the existence of a solution to Problem 8.1, the argument cannot be used to prove existence. The difficulty lies in the fact that the nonlinearities encountered in the diffusion terms are space dependent and may be discontinuous with respect to \( x \) at the interface.

8.4 Numerical experiments

This section is devoted to numerical experiments and the implementation of the proposed domain decomposition L-scheme (LDD). As our formulation and analysis did not specialise to a particular spatial discretization, the numerical implementation of the LDD scheme can in principal be done with finite differences, finite elements as well as finite volume schemes. Since mass conservation is an essential feature of porous media flow models, we adopted a cell-centred two point flux approximation variant of the finite volume scheme to reflect this on a numerical level. The domain \( \Omega \) is assumed to be rectangular and a rectangular uniform mesh was used.
Remark 8.11 (different decoupling formulations revisited). We saw in Remark 8.2 that another decoupling formulation is possible. In fact, this can be taken a step further. Equations (8.4b), (8.4c) as well as (8.4b’), (8.4c’) can be embedded into a combined formulation. For some $0 < \eta < 1$ and $M > 0$, consider the generalised decoupling

$$F^{n,i}_l \cdot n_l = M \left[ (1 - \eta) g^{i}_l + \eta p^{n,i}_l \right],$$

(8.4b’’)

$$\eta = -2\eta p^{n,i-1}_l - (1 - \eta) g^{i-1}_l, \quad (8.4c’’)

Observe that the $\lambda$-formulation (8.4b), (8.4c), as well as the convex-combination formulation (8.4b’), (8.4c’), are special cases of this general formulation: In particular, $M = (1 - \eta)^{-1}$ and $\lambda = \eta(1 - \eta)^{-1}$ recovers the $\lambda$-formulation; $M = 1$ and $\eta = \lambda$ yields the convex-combination formulation. Although (8.4b’’) and (8.4c’’) might give even greater parametric control over the numerics, in this paper we adhere to the $\lambda$-formulation because of its simplicity. Figures 8.9 and 8.10 show the influence of $\lambda$ and $\eta$ in both formulations.

We start by considering an analytically solvable example. The LDD scheme is tested against other frequently used schemes that do not use a domain decomposition ansatz. All of them are defined on the entire domain and the continuity of normal flux and pressure over $\Gamma$ is maintained implicitly. The first scheme to be compared is a finite volume implementation of the original L-scheme on the whole domain (see [155, 192, 234]), henceforth referred to as the LV scheme. Comparison is also drawn with the modified Picard scheme, (which performs better than the Picard method, see [57]), which is given by

$$S_l(p^{n,i-1}_l)\left(p^{n,i}_l - p^{n,i-1}_l\right) + \Delta t \nabla \cdot F^{n,i}_l = \Delta t f_l - \left( S_l(p^{n,i-1}_l) - S_l(p^{n-1}_l) \right) \quad \text{on } \Omega_l,$$

$$\left[ F^{n,i}_l \cdot n_l \right] = 0 \quad \text{on } \Gamma.$$

Here, the brackets $[\cdot]$ denote the jump over the interface $\Gamma$, and a source term $f_l$ is added in (8.2) for the domain $\Omega_l$ for reasons mentioned afterwards. Finally, a comparison with the quadratically convergent Newton scheme is made. Writing $\delta p^{n,i}_l = p^{n,i}_l - p^{n,i-1}_l$, it reads as follows:

$$S_l'(p^{n,i-1}_l)\delta p^{n,i}_l - \Delta t \nabla \cdot \left[ k_l(S_l(p^{n,i-1}_l))\nabla \delta p^{n,i}_l + k_l'(S_l(p^{n,i-1}_l))S_l'(p^{n,i-1}_l)\delta p^{n,i}_l \nabla (p^{n,i}_l + z) \right] = \Delta t f_l - \left( S_l(p^{n,i-1}_l) - S_l(p^{n-1}_l) \right) - \Delta t \nabla \cdot \left[ k_l(S_l(p^{n,i-1}_l))\nabla (p^{n,i-1}_l + z) \right] \quad \text{on } \Omega_l,$$

$$\left[ k_l(S_l(p^{n,i-1}_l))\nabla \delta p^{n,i}_l \cdot n_1 \right] + \left[ k_l(S_l(p^{n,i-1}_l))'\delta p^{n,i}_l \nabla (p^{n,i-1}_l + z) \cdot n_1 \right] = - \left[ k_l(S_l(p^{n,i-1}_l))\nabla (p^{n,i-1}_l + z) \cdot n_1 \right] \quad \text{on } \Gamma.$$

We refer to [155] and Chapter 9 for a study on linearisation schemes for Richards equation.
8.4 Numerical experiments

8.4.1 Results for a case with known exact solution

To demonstrate the robustness of the proposed scheme, we solve (8.2) with both Dirichlet and Neumann type boundary conditions. In the first case we disregard gravity. Specifically, we consider

\[ \Omega_1 = (-1, 0) \times (0, 1), \quad \Omega_2 = (0, 1) \times (0, 1), \quad \text{and} \quad \Gamma = [0] \times [0, 1]. \quad (8.45) \]

The relative permeabilities are \( k_1(S_1) = S_1^2 \) on \( \Omega_1 \), \( k_2(S_2) = S_2^3 \) on \( \Omega_2 \) and the saturation functions are

\[ S_l(p) = \begin{cases} \frac{1}{(1 - p)^{\frac{1}{2l}}} & \text{for } p < 0, \\ 1 & \text{for } p \geq 0, \end{cases} \quad l \in \{1, 2\}. \quad (8.46) \]

The boundaries and right hand sides are chosen to make the exact solution

\[ p_1(x, y, t) = 1 - (1 + t^2)(1 + x^2 + y^2), \quad t > 0, \quad (x, y) \in \Omega_1, \]
\[ p_2(x, y, t) = 1 - (1 + t^2)(1 + y^2), \quad t > 0, \quad (x, y) \in \Omega_2, \]

and this corresponds to the right hand sides

\[ f_1(x, y, t) = \frac{4}{(1 + x^2 + y^2)^2} - \frac{t}{\sqrt{(1 + t^2)^3(1 + x^2 + y^2)}}, \]
\[ f_2(x, y, t) = \frac{2(1 - y^2)}{(1 + y^2)^2} - \frac{2t}{3^3(1 + t^2)^4(1 + y^2)}, \]

for \( t > 0 \), and \( (x, y) \in \Omega_1 \) respectively. The boundary and initial conditions are summed up in Table 8.1.

<table>
<thead>
<tr>
<th>BC</th>
<th>( \Omega_1 )</th>
<th>( \Omega_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t = 0 )</td>
<td>( p_1(x, y, 0) = -(x^2 + y^2) )</td>
<td>( p_2(x, y, 0) = -y^2 )</td>
</tr>
<tr>
<td>( y = 0 )</td>
<td>( \partial_y p_1 = 0 )</td>
<td>( \partial_y p_2 = 0 )</td>
</tr>
<tr>
<td>( y = 1 )</td>
<td>( k_1(S_1(p_1)) \partial_y p_1 = \frac{2}{2 + x^2} )</td>
<td>( k_2(S_2(p_2)) \partial_y p_2 = 1 )</td>
</tr>
<tr>
<td>( x = -1 )</td>
<td>( p_1(-1, y, t) = 1 - (1 + t^2)(2 + y^2) )</td>
<td>( p_2(1, y, t) = 1 - (1 + t^2)(1 + y^2) )</td>
</tr>
</tbody>
</table>
| \( x = 1 \) | \[ \text{Table 8.1: Initial and boundary conditions for the example with exact solution.} \]

All linear systems were solved using a restarted generalised minimum residual method (GMRES) and the ILU[0] preconditioner has been used for all computations, [14]¹.

¹ see also the project’s website https://people.sc.fsu.edu/~jburkardt/cpp_src/mgmres/mgmres.html
To boost up speed, a sparse triplet format has been used in the matrix computations. The programs are implemented in ANSI C. The role of the different schemes in the numerical approach is explained in Figure 8.2. For the implementation we took the same $L_f$ in both sub-domains, i.e. $L := L_1 = L_2$. The results are shown in Figure 8.3 and Figure 8.4(a). For mesh size $h = 10^{-2}$, $\Delta t = 2 \cdot 10^{-4}$ as well as parameters $L = 0.25$ and $\lambda = 4$, the maximum relative error between numerical and analytical solution was less than 0.03%, i.e. $\|p^n_i - p^n_{\text{exact}}\|_{L^\infty(\Omega)} < 0.0003$. This is illustrated in Figure 8.3, where the difference $p^n_i - p^n_{\text{exact}}$ is shown. Observe, that the pressure continuity is met and the error is largest at the boundary. The relative errors of the LDD, LFV and Newton schemes at the mid-line $y = 0.5$ are plotted in Figure 8.4(a). The LDD scheme preserves the flux continuity and pressure continuity at the interface at every time step without having to solve for the entire domain. We test this theory numerically. Figure 8.4(b) shows how different kinds of errors behave within one time step. The errors $\|p^n_i - p^n_{i-1}\|_{L^2(\Omega)}$, $\|p^n_i - p^n_{i-1}\|_{L^\infty(\Omega)}$ defined on the
Figure 8.3: Difference function $p^{n,i} - p_{\text{exact}}^{n}$ over the computational domain.

(a) Comparison between the numerical solutions provided by the LDD, LFV and the Newton schemes. Plotted are the relative errors $\| p^{n,i} - p_{\text{exact}}^{n} \|_{\mathcal{L}_2(\Gamma)}$ as functions of $x$, for $y = 0.5$ and $t = 1$.

(b) Different errors vs. inner iterations for the case with exact solution. Here $t = 0.2$, $L = 0.25$ and $\lambda = 4$.

Figure 8.4: Comparison of schemes and illustration of convergence for the analytically solvable case.

domain $\Omega$, as well as $\| p^{n,i} \|_{\mathcal{L}_2(\Gamma)}$ and $\| [\mathbf{F}_1^{n,i} \cdot n_t] \|_{\mathcal{L}_2(\Gamma)}$ defined on the interface $\Gamma$, are shown. We observe that the flux and pressure jump tend to zero which implies that flux and pressure continuity is achieved. Note that the flux at $x = 0$ from the
exact solution is 0.

Next, we compare the LDD scheme with other schemes and study their dependence on discretization parameters. We compare the Newton scheme, the modified Picard iteration (called from now on simply the Picard scheme), the already mentioned LFV scheme and the LDD scheme, investigating the dependence of time step refinement and space grid refinement separately. For any of the schemes discussed in what follows, we refer to the iterations defining the nonlinear solver as inner iterations. For the LDD scheme, this means solving Problem 8.2 and similarly for the LFV and the Picard scheme. For the Newton scheme, it means one step of the Newton iteration, see the grey block in Figure 8.2.

The first study, shown in Figure 8.5, plots \( \log_{10}(\| p^{n,i} - p^{n,i-1} \|_{L^2(\Omega)}) \) for all schemes, at the fixed time step corresponding to \( t = 0.2 \). As expected, Newton is the fastest and shows a quadratic convergence rate. However, this behaviour may change once the mesh is refined. Observe in Figure 8.5 that the convergence behaviour, though still quadratic, depreciates for the finest mesh. The convergence rate of the Picard iteration is linear, faster than both the L-schemes and is stable with respect to variation in mesh size. Both L-schemes exhibit linear convergence as well, albeit slower than Picard, and the convergence speed does not vary much with the mesh size. This is in accordance with the restriction (8.11) not involving spatial discretization parameters. A similar restriction holds for the monolithic scheme, see [155]. LFV and LDD schemes have practically the same convergence rate. Table 8.2 complements the plot in Figure 8.5 and lists average convergence rates, defined as the geometric average of \( \| e_p^{n,i+1} \| / \| e_p^{n,i} \| \), for all the schemes (Newton data is not shown for \( h = 0.1, 0.05, 0.02 \) as it reaches an error lower than \( 10^{-10} \) in 3 iterations).

<table>
<thead>
<tr>
<th>( h )</th>
<th>Newton</th>
<th>Picard</th>
<th>LFV</th>
<th>LDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>-</td>
<td>0.0504</td>
<td>0.4046</td>
<td>0.4400</td>
</tr>
<tr>
<td>0.05</td>
<td>-</td>
<td>0.0504</td>
<td>0.3906</td>
<td>0.4270</td>
</tr>
<tr>
<td>0.02</td>
<td>-</td>
<td>0.0505</td>
<td>0.3909</td>
<td>0.4221</td>
</tr>
<tr>
<td>0.01</td>
<td>0.0113</td>
<td>0.0567</td>
<td>0.3910</td>
<td>0.4221</td>
</tr>
</tbody>
</table>

Table 8.2: The average convergence rate, \( \| e_p^{n,i+1} \| / \| e_p^{n,i} \| \), for the different schemes and with respect to the mesh-size.

Secondly, we study the dependence of the convergence rates on time step size for a fixed mesh size (\( h = 0.02 \)). The error characteristics of all four schemes in Figure 8.6 are shown for \( t = 0.5 \). In Figure 8.6(a) both, Newton and Picard schemes, diverge, whereas both the L-schemes converge for \( L = 0.25 \). The LFV scheme, however, exhibits some oscillations, which are due to the dependence of the choice of \( L \) on the time step size. The convergence of both the L-schemes is guaranteed under...
(mild) restrictions, involving $L$, on the time step. In the experiment reported in Figure 8.6(a), $\Delta t = 0.5$ is quite large and so the choice of $L = 0.25$ in the LFV scheme is inappropriate for this time step, resulting in oscillations and a very slow convergence. Indeed, observe that LFV becomes convergent if either the $L$ parameter is enlarged (the LFV* results) or the time step size is decreased (Figure 8.6(b) and Figure 8.6(c)), which sustains the above interpretation. Note, that the LDD scheme converges for all $\Delta t$ and is approximately as fast as the LFV scheme in all the cases. For smaller values of $\Delta t$ the Newton and Picard schemes converge faster than both the L-schemes, as shown in Figure 8.6(b) and 8.6(c).

According to the theory, the convergence of the Newton and Picard schemes is only guaranteed if the initial guess is close enough to the exact solution. Therefore, when the iteration is started with the numerical solution of the previous time step, this suggests that the time step should be taken small enough to guarantee convergence (see [155,183,206]). On the other hand, L-schemes are free of this constraint.

To illustrate this behaviour, we have investigated the convergence of the schemes for a constant initial guess. Specifically, $p^{n,0} = -5$ has been used instead of $p^{n,0} = p^{n-1}$. In this case, the Newton and Picard schemes are divergent whereas both L-schemes still converge. This is displayed in Figure 8.6(d). A similar behaviour will be observed again while discussing a numerical example with realistic parameters.
A linear domain decomposition method for unsaturated flow in layered porous media

Figure 8.6: Convergence study for the time steps $\Delta t = 0.1, 0.01, 0.001$. Here, $L = 0.25$ for the LFV scheme and $L = 0.5$ for the LFV* scheme. For the LDD scheme one has $L = 0.25, \lambda = 2$ in case (a); $L = 0.25, \lambda = 4$ in case (b); and $L = 0.25, \lambda = 10$ in case (c).

Remark 8.12. The convergence behaviour of the LDD scheme can be optimised by choosing $\lambda$ properly. In the above comparison $\lambda$ was chosen differently for every choice of mesh size. The optimality of $\lambda$ is dependent on the mesh size and the time step size. With a good choice of $\lambda$, one can make the LDD scheme approximately as fast as the LFV scheme. This is discussed in more detail in Section 8.4.4.

8.4.2 Results for a realistic case with van Genuchten parameters

We demonstrate the applicability of the LDD scheme for a case with realistic parameters, incorporating also gravity effects. We consider a van-Genuchten-Mualem
parametrisation [259] with the curves $k$ and $S$

$$
S_I(p) = S_{w,r,l} + (1 - S_{n,r,l} - S_{w,r,l})S_{e,l}(p),
$$

$$
S_{e,l}(p) = \frac{1}{1 + (-\alpha(p)\hat{n}_l)m_l},
\quad m_l = 1 - \frac{1}{\hat{n}_l},
$$

$$
k_l(S_i(p)) = \sqrt{S_{e,l}(p)(1 - (1 - S_{e,l}(p)^{\frac{1}{m_l}})^2)}.
$$

(8.47)

The specific parameter values are listed in Table 8.3 and are characteristic for particular types of materials, *silt loam G.E. 3* ($\Omega_1$) and *sandstone* ($\Omega_2$). These materials have very different absolute permeabilities $K_1, K_2$, which makes the numerical calculations more challenging.

The dimensional governing equations and boundary conditions become ($l \in (1, 2)$)

$$
L_l p^{n,i}_l + \Delta t \nabla \cdot F^{n,i}_l = L_l p^{n,i-1}_l - \phi_l(S_l(p^{n,i}_l) - S_l(p^{n,-1}_l)), \quad \text{on } \Omega_l,
$$

(8.48a)

$$
F^{n,i}_l \cdot n_l = g_l + 2\lambda p^{n,i}_l,
$$

(8.48b)

$$
p^{n,i}_l = 0
$$

(8.48c)

In this case $F^{n,i}_l = -\frac{K_l}{\mu}k_{r,l}(S_l(p^{n,i}_l - \rho g)$. Here $g = g \hat{x}$ is the gravitational acceleration aligned with the positive $x$-direction; $\rho, \mu$ are the density and the viscosity of the fluid and $K_l, \phi_l$ are the absolute permeability as well as the porosity of the medium.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unit</th>
<th>Silt Loam G.E. 3 ($\Omega_1$)</th>
<th>Sandstone ($\Omega_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Porosity ($\phi_l$)</td>
<td>-</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>Water Density ($\rho$)</td>
<td>kg m$^{-3}$</td>
<td>$1 \times 10^3$</td>
<td>$1 \times 10^3$</td>
</tr>
<tr>
<td>Water Viscosity ($\mu$)</td>
<td>Pa s</td>
<td>$1 \times 10^{-3}$</td>
<td>$1 \times 10^{-3}$</td>
</tr>
<tr>
<td>Absolute permeability ($K_l$)</td>
<td>m$^2$</td>
<td>$5.7407 \times 10^{-14}$</td>
<td>$1.2500 \times 10^{-12}$</td>
</tr>
<tr>
<td>Retention exponent ($\hat{n}_l$)</td>
<td>-</td>
<td>2.06</td>
<td>10.4</td>
</tr>
<tr>
<td>Retention parameter ($\alpha_l$)</td>
<td>Pa$^{-1}$</td>
<td>$4.23 \times 10^{-5}$</td>
<td>$7.90 \times 10^{-5}$</td>
</tr>
<tr>
<td>Irreducible water saturation ($S_{w,r,l}$)</td>
<td>-</td>
<td>0.131</td>
<td>0.153</td>
</tr>
<tr>
<td>Irreducible air saturation ($S_{n,r,l}$)</td>
<td>-</td>
<td>0.604</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 8.3: The van Genuchten-Mualem parameters in the realistic test case.

The problem is non-dimensionalized by using the characteristic pressure $p^* := -14.8 \times 10^3$ Pa, length 1.48 m and time $t^* = 41.440$ s. This leads to the non-dimensional quantities $\bar{p} := \frac{p}{p^*}, (x, y)$ and $t := \frac{t}{t^*}$, where $t$ denotes the physical time. After non-dimensionalization, the domain used is again taken to be $\Omega_1 = (-1, 0) \times (0, 1), \Omega_2 = (0, 1) \times (0, 1)$. The initial condition used is

$$
\bar{p}(x, y, 0) = -1,
$$

(8.49)
A linear domain decomposition method for unsaturated flow in layered porous media

and boundary conditions are

\[
\tilde{p}(-1, y, t) = \begin{cases} 
-1 + ty & \text{if } y < (1 - \epsilon)t^{-1}, \\
-\epsilon & \text{if } y \geq (1 - \epsilon)t^{-1}, 
\end{cases}
\]

\[
\tilde{p}(1, y, t) = -1,
\]

together with a no-flow condition at \(y = 0, 1\). We take \(\epsilon > 0\) to avoid degeneracy.

Figure 8.7: Error plots and scheme comparison for the realistic case.

Figure 8.7(a) shows the different errors for this case and it can be seen that all the errors are decreasing for the LDD scheme. Errors at the interface and inside the domain tend to 0, the convergence is slower compared to the case with exact solution, however. This is due to the large variance of the parameters as well as the highly nonlinear nature of the associated functions. Because of this, both Newton and Picard schemes diverge. The behaviour of different schemes for the same set of parameters is shown in Figure 8.7(b). Observe that for the Newton scheme the starting error as well as the number of iterations required increases steadily with \(t\) until \(t = 0.94\), at which point the errors start diverging. The Picard scheme becomes divergent even before \(t = 0.2\). In contrast, both L-schemes remain stable in this case.

8.4.3 Time performance

This section is devoted to the comparison of time performance of the schemes. We have seen that L-schemes are more stable than Newton and Picard. But if they are
converging, then Newton and Picard schemes converge faster than the L-schemes. Below we investigate how the schemes compare to one another with respect to actual computational time. We set an error tolerance for the schemes that stops the iterations within one time step, after reaching an error lower than $10^{-6}$, i.e. $\| p^{n,i} - p^{n,i-1} \|_{L^2(\Omega)} < 10^{-6}$. This is to ensure that we get comparative CPU-clock-time for different schemes for the same degree of accuracy. We computed the exactly solvable case on a LINUX server (mammoth.win.tue.nl) for all four schemes using the same set of parameters ($h = 0.02$, $\Delta t = 0.001$, $L = 0.25$ and $\lambda = 10$). Figure 8.8 illustrates the time-performances of these schemes over the whole computational time domain. No parallelisation has been implemented in this study, except for the LDD parallel plot in Figure 8.8, where each domain in the LDD scheme has been calculated on a different processor. Table 8.4 shows how many inner iterations are required on average for different schemes to reach the error criterion at different points in time. The number of iterations per time step increases for all schemes as the boundary conditions change more rapidly with time. Table 8.4 also shows the average time per inner iteration. Additionally, the last row of Table 8.4 shows how many GMRES iterations were required by each scheme to execute one inner iteration.

Unsurprisingly, excluding the parallel version of LDD, the Newton scheme is the fastest, followed by the Picard and the LDD scheme. However, the latter competes closely with the Newton and the Picard schemes. Even more surprising is the fact that the LFV scheme takes considerably more time to reach the desired accuracy compared to the LDD scheme, despite both having almost the same convergence rate. The reason becomes apparent from Table 8.4: The LDD scheme requires much less time per inner iteration than all other schemes. The LFV scheme has the second fastest average time per iteration. For the Picard iteration, the derivative of the saturation function needs to be evaluated which in turn costs more time than an iteration in the LFV scheme. The Newton scheme is computationally most expensive per
iteration because it calculates the Jacobian at every iteration. The schemes that do

<table>
<thead>
<tr>
<th>Time-step/Scheme</th>
<th>LDD (parallel)</th>
<th>LFV</th>
<th>Picard</th>
<th>Newton</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>7.3</td>
<td>7.5</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>50</td>
<td>9.880</td>
<td>10.72</td>
<td>2.520</td>
<td>2.060</td>
</tr>
<tr>
<td>100</td>
<td>11.26</td>
<td>12.31</td>
<td>2.760</td>
<td>2.030</td>
</tr>
<tr>
<td>500</td>
<td>11.19</td>
<td>11.79</td>
<td>2.952</td>
<td>2.006</td>
</tr>
<tr>
<td>1000</td>
<td>14.18</td>
<td>14.35</td>
<td>2.946</td>
<td>2.408</td>
</tr>
<tr>
<td>Avg. time per iter.</td>
<td>0.1965</td>
<td>0.5392</td>
<td>0.6591</td>
<td>0.6722</td>
</tr>
<tr>
<td>Avg. GMRES iterations</td>
<td>119 + 123</td>
<td>396.6</td>
<td>390.9</td>
<td>397.7</td>
</tr>
</tbody>
</table>

Table 8.4: The average number of inner iterations per time step required by the different schemes to reach the stopping criterion \( \| p^{n,i} - p^{n,i-1} \|_{L^2(\Omega)} < 10^{-6} \). The last two rows give the average time and the average number of GMRES-iterations per inner iteration. Observe that for the LDD scheme, the two numbers given in the last row reflect average numbers per subdomain.

not decouple the domain require much more time and many more GMRES-iterations per inner iteration. The reason is that the domain decomposition schemes involve smaller matrices and they have smaller condition numbers. This is illustrated by the last row of Table 8.4. The LDD scheme requires on average 119 GMRES-iterations on \( \Omega_1 \) and 123 GMRES-iterations on \( \Omega_2 \) and both domains have 52 \( \times \) 50 elements. Compare this with Newton, which takes almost 400 GMRES-iterations and deals with 102 \( \times \) 50 variables on each GMRES-iteration. This explains why the LDD scheme takes so much less time per inner iteration. Table 8.5 compares the condition numbers of the LDD and the LFV scheme. It shows that the matrices for the LFV scheme are worse conditioned than the ones of the LDD scheme. The latter has two condition numbers, one for each domain. The 2-norm condition numbers were calculated with MATLAB’s build-in cond() function.

Finally, the solid line plots the computational time of the LDD scheme which has been parallelised using OpenMP. The two different domains are run on two different processors and a speedup ratio of 1.79 is obtained. This makes the scheme actually slightly faster than the Newton scheme. Observe that the average number of required inner iterations and GMRES-iterations of the parallelised LDD scheme is the same as for the LDD scheme without parallelisation.

Remark 8.13. The fact that the LDD scheme performance competes closely with Newton and Picard, even without distributing the calculation on the subdomains to different processors, along with its global convergence property and robustness are the key advantages of the LDD scheme. Parallelisation helps even further, as this study suggests, although of course, the linear algebraic solver can also be parallelised, which can lead to a further computational speedup. This is a general feature of all schemes,
8.4 Numerical experiments

Condition number

<table>
<thead>
<tr>
<th>$h$</th>
<th>0.1</th>
<th>0.05</th>
<th>0.02</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDD ($\Omega_1$)</td>
<td>7.6191</td>
<td>11.8947</td>
<td>73.362</td>
</tr>
<tr>
<td>LDD ($\Omega_2$)</td>
<td>7.0219</td>
<td>12.3557</td>
<td>74.519</td>
</tr>
<tr>
<td>LFV ($\Omega$)</td>
<td>94.8158</td>
<td>171.47</td>
<td>397.34</td>
</tr>
</tbody>
</table>

Table 8.5: The condition number vs mesh size for the LDD and LFV schemes. Here, $\Delta t = 0.001$, $t = 0.2$, $L = 0.25$, $\lambda = 10$. The condition numbers are calculated for the 200th time step for the matrices of the first inner iteration.

whereas an additional level of parallelisation is possible for the LDD scheme because of the inherent decoupling of the subproblems.

8.4.4 Parameter dependence and key features

Having outlined the robustness and speed of the proposed LDD scheme we turn to investigate some of its properties. Two important parameters have been introduced in the LDD scheme, i.e. $L_l$ and $\lambda$, and apart from a lower bound on $L_l$ nothing has been specified about these parameters. This means that they can freely be adjusted to give optimal convergence rate. In fact, in this section we will see that the convergence rate depends strongly on these parameters.

The influence of $\lambda$

![Graph](image)

(a) The decay of the $p$ error in terms of $\lambda$.

![Graph](image)

(b) The decay of the $g$-error in terms of $\lambda$.

Figure 8.9: The influence of $\lambda$ on the convergence rate. The parameters for the LDD scheme are $\Delta t = 0.01$, $h = 0.02$, $L_l = 0.25$ at $t = 0.2$.

Figure 8.9 shows the influence of the parameter $\lambda$ on error characteristics. All the results shown are for the case with exact solution. Figure 8.9(a) focuses on the
errors $\|p^n_i - p^{n-1}_i\|_{L^2(\Omega)}$ on the domain $\Omega$, while Figure 8.9(b) depicts the $L^2$-errors $\|g^n_i - g^{i-1}\|_{L^2(\Gamma)}$ on the interface for the same time step. Clearly, $\lambda$ has tremendous impact on the convergence rate. The convergence rate rapidly increases with $\lambda$ at first but after a certain point the convergence rate starts decreasing. This trend is noticeable in both plots of Figure 8.9. This indicates that there is an optimal lambda $\lambda_{opt}$ for which the whole scheme has a fastest convergence rate. The optimality of $\lambda$ is actually a well studied behaviour in the domain decomposition literature. In [157,198] it has been shown that $\lambda_{opt}$ depends at least on mesh size and sub-domain size. Later we will show that it also depends on $L_l$ and $\Delta t$ in our case. This control over the convergence rate is the reason why the $\lambda$-formulation was chosen over the convex-combination formulation given in Remark 8.2. To illustrate this, Figure 8.10 shows the same plots as Figure 8.9 but for the convex-combination formulation. In order to differentiate between plots more easily, we use the combined formulation (8.4b’), (8.4c’) and set $M = 1$. For $\eta = 0.01$ the convex-combination formulation even fails to converge. In all other cases the convergence is considerably slower.

The influence of $L_l$

We briefly give an overview of the influence of $L_l$ on the convergence rate. Figure 8.11(a) depicts this for $L = L_1 = L_2$. For $L$-schemes it is common to diverge if $L$ is too small, which seems to be the case for $L = 0.1$. On the other hand, the convergence rate decreases significantly for very large $L$, a behaviour that is common to all $L$-schemes, see [205]. It is best to choose $L$ as small as possible, yet great enough
to ensure convergence of the scheme. Note that \( L_I = 0 \) represents the original (non-modified) Picard iteration case and Figure 8.11(a) suggests that the original Picard scheme fails for these problems.

The dependence of \( \lambda_{\text{opt}} \) on \( L_I, \Delta t \) and \( h \)

![Graph showing the influence of \( L \) on the convergence rate, as obtained from the inner iterations for the 50th time step.](image1)

![Graph showing the convergence rate vs. \( \lambda \) for \( L = 0.25 \) and \( L = 1 \). For \( L = 0.25 \), \( \lambda_{\text{opt}} \approx 4 \).](image2)

Figure 8.11: Influence of \( L \) on the convergence rate (a) and optimality study for \( \lambda_{\text{opt}} \) (b).

<table>
<thead>
<tr>
<th>( L )</th>
<th>( \lambda = 0.1 )</th>
<th>( \lambda = 1 )</th>
<th>( \lambda = 10 )</th>
<th>( \lambda = 100 )</th>
<th>( \lambda_{\text{opt}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>diverged</td>
<td>diverged</td>
<td>diverged</td>
<td>diverged</td>
<td>-</td>
</tr>
<tr>
<td>0.25</td>
<td>0.9020</td>
<td>0.6223</td>
<td>0.5480</td>
<td>0.7721</td>
<td>(1,10)</td>
</tr>
<tr>
<td>1</td>
<td>diverged</td>
<td>0.7675</td>
<td>0.7750</td>
<td>0.8138</td>
<td>(1,10)</td>
</tr>
<tr>
<td>5</td>
<td>diverged</td>
<td>0.8993</td>
<td>0.8718</td>
<td>0.8708</td>
<td>(10,100)</td>
</tr>
</tbody>
</table>

Table 8.6: The dependence of the convergence rates on \( \lambda \) and \( L \): the geometric average of the contraction rates over the first 20 iterations and for different \( (L, \lambda) \) pairs is given in the first columns, whereas the last gives the interval for \( \lambda_{\text{opt}} \). Here, \( h = 0.02 \), \( \Delta t = 0.01 \), \( t = 0.2 \).

In this section we investigate numerically how \( \lambda_{\text{opt}} \) depends on the choice of \( L \), \( \Delta t \) and \( h \). For a fixed grid in time and space, Table 8.6 lists convergence rates for different \( \lambda \) and \( L \). With this table we can guess the interval in which \( \lambda_{\text{opt}} \) lies. Within this estimated interval, Figure 8.11(b) shows how the convergence rate varies with \( \lambda \) for fixed \( L \). For \( L = 0.25 \), \( h = 0.02 \) and \( \Delta t = 0.01 \) the fastest convergence is achieved for \( \lambda = 4 \) (this is why \( \lambda = 4 \) was chosen for the above comparisons, wherever the specified \( L, \Delta t, h \) set was used). The \( \lambda \) dependence for higher values of \( L \) is less pronounced.
For a fixed $L$, Table 8.7(a) and Table 8.7(b) show the dependence of $\lambda_{opt}$ on the time step and the mesh size respectively. The second row of each table shows estimates of $\lambda_{opt}$ based on a study like the one shown in Figure 8.11(b), which minimizes the average convergence rates. The results indicate a quite strong correlation of $\lambda_{opt}$ with the time step size, contrasted by a rather minor correlation with the mesh size.

<table>
<thead>
<tr>
<th>$\Delta t$</th>
<th>0.1</th>
<th>0.01</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_{opt}$</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Avg. CR</td>
<td>0.4444</td>
<td>0.4221</td>
<td>0.5408</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$h$</th>
<th>0.1</th>
<th>0.05</th>
<th>0.02</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_{opt}$</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Avg. CR</td>
<td>0.4398</td>
<td>0.4270</td>
<td>0.4221</td>
<td>0.4221</td>
</tr>
</tbody>
</table>

(a) Estimation of $\lambda_{opt}$ for fixed $h = 0.02$, $L_l = 0.25$ and varying $\Delta t$. In this case $\lambda_{opt}$ needs to be adjusted to keep Avg. CR low.

(b) Estimation of $\lambda_{opt}$ for fixed $\Delta t = 0.01$, $L_l = 0.25$ and varying $h$. In this case no adjustment of $\lambda_{opt}$ is needed to keep Avg. CR low.

Table 8.7: Influence of $\lambda_{opt}$ on the average convergence rates (Avg. CR) for varying $h$ and $\Delta t$. $\lambda_{opt}$ in each case is estimated, based on an optimality study similar to the one shown in 8.11(b).

### 8.4.5 Multi domain and parallelisation study

Finally, we give an example where the LDD scheme is applied to four subdomains. For this purpose, the numerical domain as well as the solution defined in the test case with exact solution, is extended by reflection of $\Omega$ along the $y$ axis. With this extension, the four domains are

$$
\Omega_1 = (-1,0) \times (0,1), \quad \Omega_2 = (0,1) \times (0,1), \\
\Omega_3 = (0,1) \times (-1,0), \quad \Omega_4 = (-1,0) \times (-1,0),
$$

and interfaces are defined as in Remark 8.3. Now, $\Omega = \bigcup_{l=1}^{4} \Omega_l$. To prevent all the interfaces from having zero flux, gravity is included which modifies the definition of one of the source terms such that it reads

$$
f_1(x,y,t) = \frac{2(1-y^2)}{(1+y^2)^2} - \frac{2t}{3\sqrt{(1+t^2)^4(1+y^2)}} - \frac{2x}{(1+t^2)(1+x^2+y^2)^2} \quad \text{in} \quad \overline{\Omega_1} \cup \overline{\Omega_4}.
$$

(8.50)

All other functions and exact solutions defined in Section 4.1 remain essentially the same, the only difference being that all definitions on $\Omega_1$ are extended to $\Omega_1 \cup \Omega_4$ and all definitions on $\Omega_2$ are extended to $\Omega_2 \cup \Omega_3$. In line with the optimal choice of $L$ and $\lambda$ found for the two domain studies, we have kept $L_l := L = 0.25$ and $\lambda_{l,r} := \lambda_{opt} = 4$ for the multi domain experiment and $l \in \{1,2,3,4\}, \ r \in I_l$.

For this experiment, triangular unstructured meshes $\mathcal{T}_h$ have been used on all domains, namely, we used FVCA8 benchmark meshes in order to better resolve the
Table 8.8: Speedup for the parallelisation of the multi-domain example on up to 4CPUs. The computation was done on the four subdomains with \( L_i := 0.25 \), \( \lambda_j \), \( r = 4 \), \( h = 0.05 \), \( \Delta t = 0.001 \) until an accuracy of \( 10^{-6} \) was reached. Times are compared at \( t = 1 \).

<table>
<thead>
<tr>
<th></th>
<th>1 CPU</th>
<th>2 CPUs</th>
<th>4 CPUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>753s</td>
<td>411s</td>
<td>207s</td>
</tr>
<tr>
<td>Speedup</td>
<td>-</td>
<td>1.83</td>
<td>3.64</td>
</tr>
</tbody>
</table>

To illustrate the possible speedup due to parallelisation of the LDD scheme on multiple domains, we computed the same problem as described above on one, two and four processors. The results are shown in Table 8.8. For two and four CPUs, the computations are 1.83 and 3.64 times faster than on a single CPU, which shows that the computational speeds scale almost perfectly with the number of CPUs. It should however be mentioned that this is in the most optimum case of perfect load balancing since all four domains are of the same size.


8.5 Conclusion

We considered a nonlinear parabolic problem appearing as mathematical model for variably saturated flow in porous media. For the numerical solution of the nonlinear, time discrete problems we proposed a combined scheme (LDD) that is based on a fixed point iteration (the L-scheme), and on a domain decomposition scheme involving Robin type coupling conditions at the interface separating different subdomains. The result is a scheme featuring the advantages of both approaches: convergence, analytically proven under mild restrictions on the time step size, and decoupling of the time discrete problems into subproblems, that lends itself naturally to parallelisation, speeding up the computations. The stability, robustness and efficiency of the method are tested numerically for various cases and also compared to the Newton and Picard schemes. The tests include situations where the latter diverge whereas the proposed scheme is converging. In summary, the key advantages of this method are:

- The LDD scheme converges under mild restrictions on the time step size, i.e. the restriction does not depend on the space discretization parameters. It can provide accurate results even in situations where the Picard or Newton iterations fail.

- In conjunction with a suitable space discretization, it provides a decoupled, mass conservative approach. This is very useful, in particular, when dealing with models defined in media with block-type heterogeneities, where the material properties in different blocks may vary significantly.

- Though linearly convergent, the computational time required by the LDD scheme for achieving a certain accuracy of the approximation is comparable to the time needed by the Newton and Picard schemes. LDD scheme is much faster than a standard L-scheme applied to the model in the entire domain. This efficiency is due to the fact that LDD needs less time per inner iteration than a scheme defined in the entire domain. Moreover, LDD offers inherently a way of parallelisation due to the decoupling of the subproblems, see Remark 8.13, which makes it possible to increase its speed even further.

- The convergence rate of LDD schemes depends on the choice of \( L \) and \( \lambda \). With the optimal choice of parameters, the convergence order can be reduced significantly.
A linear iterative scheme for nonlinear diffusion problems

9.1 Introduction

In this chapter, a linearisation technique is considered for the generalised nonlinear advection diffusion equations of the type

\[ \partial_t b(u) + \nabla \cdot \mathbf{F}(\mathbf{x}, u) = \nabla \cdot [\mathcal{D}(\mathbf{x}, u) \nabla u] + f(\mathbf{x}, t, u); \] (9.1)

completed with suitable boundary and initial conditions. Equation (9.1) appears as mathematical model for many real world applications, like flow through porous media or reactive transport. For the discretization in time, the backward Euler method is often used due to its stability. This changes (9.1) into a sequence of nonlinear elliptic equations. For solving these, linear iterative schemes are required.

A commonly used linearisation technique is the Newton scheme (NS). Being quadratically convergent, it is widely used for solving nonlinear equations [26,151]. However, this quadratic convergence is featured under certain restrictions. In particular, degenerate problems do not fulfill these restrictions [206]. Another drawback of the NS is that it is only locally convergent, meaning that the initial guess for the iterations has to be close enough to the actual solution for the scheme to be convergent [26,206]. In many cases, this requires extremely small time step sizes limiting the applicability of the NS. For this reason, pre-conditioners, line-search

---

methods and different parametrizations are often used to enhance the robustness of the NS [39, 130].

An alternative to the NS is a modified Picard scheme (PS), proposed in [57]. In [151, 155, 158] it is shown that this scheme is quite fast despite having linear convergence. Another linearly converging scheme is the Jäger-Kačur scheme (see [123, 124, 133]). A sufficient condition for convergence was derived in [206] for all the schemes mentioned above. When applied to the Euler implicit discretization of (9.1) one needs to take

$$\Delta t < C m b^r h^d,$$

(9.2)
to guarantee the convergence of these schemes. Here $\Delta t$ is the time step size, $h$ is the mesh size, $d$ is the spatial dimension of the problem, $m_b \geq 0$ is the lower bound of $b'$ and $C, r > 0$ are constants that depend on the nonlinear functions. For $d \geq 2$ this imposes a severe restriction on the time step size, which can increase the computation time to a great extent. Moreover, in the degenerate case $m_b = 0$, either a regularised version of the function $b$ has to be used [123, 124, 160] or the initial/boundary data has to be shifted [195] to ensure convergence.

For porous media applications, where all the associated functions are nonlinear and the problem might become degenerate, stability is an important issue. Also, extremely large timescales are involved for such processes and so condition (9.2) cannot always be satisfied. To address this, a fixed point iteration scheme, termed the LS or simply LS in the context of this discussion, was proposed in [192, 194, 203]. The LS is linearly convergent but it has the interesting property of unconditional convergence, meaning that it converges to the time-discrete solution irrespective of the choice of initial guess. This is due to the fact that in the LS, the stabilization terms are estimated globally as opposed to the local estimations used in the NS, the PS or the Jäger-Kačur scheme. However, as shown in [155, 226], this increases the convergence rate of the LS, making it slower when compared to the NS or the PS. This has motivated authors to either use the LS to provide initial guesses for the NS [155] or to apply it in a domain decomposition type approach that boosts the speed of the LS [226].

All of the schemes mentioned above are mostly designed to solve nonlinear elliptic problems and do not use the fact that these are the outcomes of the time discretization of a nonlinear parabolic problem. In this context, the solution from the previous time step can be used as initial guess in the iterative process. For solutions that have a good regularity in time, as being the case for parabolic problems, the changes in the solution from one time step to the next one are limited. However the standard schemes do not use this fact and thus their implementation for parabolic and elliptic problems is more or less the same.

The main idea in this work is to exploit the fact that the nonlinear elliptic problems are the result of the time discretization of (9.1). The proposed scheme is a combination of the PS and the LS and it uses local estimations to improve the con-
vergence behaviour of the LS without affecting its stability. After introducing the problem in Section 9.2 in Section 9.3 we propose the iterative scheme for simple quasilinear equations, and analyse its convergence. It can be observed that, apart from being unconditionally convergent, the scheme has a convergence rate proportional to the time step size for sufficiently small time steps. This is unlike the LS, where a lower time step increases the convergence rate. Section 9.4 generalises these ideas for (9.1). The scheme converges also in the degenerate case, however for small enough time step sizes. For the non-degenerate case, the convergence is linear and the rate is proportional to the square root of the time step size. Finally in Section 9.5 some numerical experiments are presented. These show that the scheme is more stable than the NS or the PS and converges at least as fast as the PS and sometimes comparable to the NS.

9.2 The general problem and linearisation schemes

Let \( \Omega \) be a bounded domain in \( \mathbb{R}^d \) which has a Lipschitz boundary \( \partial \Omega \) and define \( Q = \Omega \times [0, T) \) for some \( T > 0 \). For the rest of our discussion (\( \cdot, \cdot \)) and \( \| \cdot \| \) will represent \( L^2(\Omega) \) inner product and norm. Any other norm will be presented as \( \| \cdot \|_V \) with \( V \) being the corresponding space. The Sobolev space \( W^{k,p}(\Omega) \) is the set of functions \( u \) defined on \( \Omega \) such that \( D^k u \in L^p(\Omega) \) for the multi-index \( k \), equipped with the norm \( \| u \|_{W^{k,p}(\Omega)} = (\sum_{q \leq k} \int_{\Omega} |D^q u|^p)^{1/p} \) for \( 1 \leq p < \infty \). Further, \( H^k(\Omega) = W^{k,2}(\Omega) \) and \( H^k_0(\Omega) \) represents the set of elements of \( H^k(\Omega) \) which have \( 0 \) trace at the boundary \( \partial \Omega \).

The Hölder space \( C^{\ell,\delta}(\Omega) \) refers to the space of \( \delta \)-Hölder continuous functions up to the \( \ell \)th space derivative for the metric \( \text{dist}(\vec{x}, \vec{y}) = |\vec{x} - \vec{y}|, \vec{x}, \vec{y} \in \mathbb{R}^d \). The associated norms of these spaces are defined in detail in [146]. With these basic definitions stated, we introduce the problem.

9.2.1 Time-discrete formulation and properties of functions

We consider (9.1) in the space-time domain \( Q \). For simplicity homogeneous Dirichlet boundary condition is assumed on \( \partial \Omega \times [0, T) \). Further, \( u_0(\cdot) \in H^1_0(\Omega) \) be the initial condition. We refer to [9] and [181] for results on the existence and uniqueness of weak solutions to (9.1) for this case.

For the time discretization of (9.1) we let \( \Delta t = T/N, N \in \mathbb{N} \), be the time step size and \( n \in \{1, \ldots, N\} \) represent the time step. Define \( t_n = n \Delta t \). The Euler implicit discretization of (9.1) leads to the sequence of elliptic problems \( (n \in \{1, \ldots, N\}) \)

\[
\begin{align*}
& (P_1) \quad \text(b(u_n) - b(u_{n-1}) + \Delta t \nabla \cdot \vec{F}(\vec{x}, u_n) = \Delta t \nabla \cdot [\mathcal{D}(\vec{x}, u_n) \nabla u_n] + \Delta t f(\vec{x}, t_n, u_n) \text{ in } \Omega; \quad (9.3a) \\
& u_n = 0 \text{ on } \partial \Omega \quad (9.3b)
\end{align*}
\]

Below we consider weak solutions to the time discrete problem \( (P_1) \), defined as:
Definition 9.1. Let \( n \in \{1, \ldots, N\} \) and \( u_{n-1} \in L^2(\Omega) \) be given. A weak solution to Problem (P1) is a function \( u_n \in H^1_0(\Omega) \) satisfying for any test function \( \phi \in H^1_0(\Omega) \)

\[
(b(u_n) - b(u_{n-1}), \phi) + \Delta t (\mathcal{D}(\tilde{x}, u_n) \nabla u_n, \nabla \phi) = \Delta t (\mathcal{F}(\tilde{x}, u_n), \nabla \phi) + \Delta t (f(\tilde{x}, t_n, u_n), \phi). \tag{9.4}
\]

The existence of \( u_n \in H^1_0(\Omega) \) for a given \( u_{n-1} \in L^2(\Omega) \) is shown in [194]. Note that the sequence \( \{u_n\}_{n=1}^N \) can be used to approximate the solution to the original parabolic problem (9.1) (the Rothe method, [132]). Below we assume the following:

(P1) \( b : \mathbb{R} \to \mathbb{R} \) is a \( C^2(\mathbb{R}) \) function such that \( b' \geq m_b \) and \( |b'|, |b''| \leq M_b \) for some \( m_b, M_b \geq 0 \).

(P2) \( f : Q \times \mathbb{R} \to \mathbb{R} \) is twice differentiable with respect to \( u \) and satisfies \( T \partial_u f(\tilde{x}, t, u) \leq b'(u) \) and \( |f|, |\partial_u f|, |\partial_{uu} f| \leq M_f \) (\( M_f \geq 0 \)) a.e. in \( (\tilde{x}, t) \in Q \) and \( u \in \mathbb{R} \).

(P3) \( \mathcal{D} : \Omega \times \mathbb{R} \to \mathbb{R}^+ \) is a \( C^2(\Omega \times \mathbb{R}) \) function. Moreover, there exists \( \mathcal{D}_m, \mathcal{D}_M > 0 \) such that \( \mathcal{D}_m \leq \mathcal{D} \) and \( \mathcal{D}, |\partial_x^d \mathcal{D}|, |\partial_{x_j} \mathcal{D}|, |\partial_u \partial_{x_j} \mathcal{D}| \leq \mathcal{D}_M \) for \( j \in \{1, \ldots, d\} \), \( q \in \{1, 2\} \).

(P4) \( \mathcal{F} : \Omega \times \mathbb{R} \to \mathbb{R}^d \), with \( x_j \)-component denoted by \( F_j \), admits partial derivatives that satisfy \( |\partial_u^d F_j|, |\partial_{x_j} F_j|, |\partial_u \partial_{x_j} F_j| \leq M_F \) for \( j \in \{1, \ldots, d\} \), \( q \in \{1, 2\} \) and \( M_F \geq 0 \).

(P5) \( u_0 \in H^1_0(\Omega) \). In some cases we also use \( u_0 \in W^{2,\infty}(\Omega) \).

In (P2) instead of using \( \partial_u f \leq 0 \) we have used the less restrictive condition \( b' \geq T \partial_u f \). At this stage we define the following function which will be used later,

\[
z : \Omega \times \mathbb{R} \to \mathbb{R}, \quad z(\tilde{x}, t) = b(u) - \Delta t f(\tilde{x}, t, u). \tag{9.5}
\]

Observe that actually \( z \) depends on the time step size \( \Delta t \) and \( n \) as well. However the focus here is on constructing an iterative scheme for a fixed time step size \( \Delta t \) at a fixed time \( t_n \). Hence, we disregard this dependence. From (P1) and (P2), \( z \in C^2 \) and satisfies \( \partial_u z = b' - \Delta t \partial_u f \geq b' \left(1 - \frac{\Delta t}{2}\right) \geq m_b \left(1 - \frac{\Delta t}{2}\right) \geq 0 \). So by defining \( m = m_b \left(1 - \frac{\Delta t}{2}\right) \) we get the inequality

\[
\partial_u z \geq m \geq 0 \text{ a.e. for } \tilde{x} \in \Omega \text{ and } u \in \mathbb{R}. \tag{9.6}
\]

Remark 9.1 (On the properties of the functions). The situation \( m_b = 0 \) in (P1) gives rise to degeneracy, which will be discussed in detail later. The boundedness of \( b'' \) is assumed in our analysis. This is a more severe condition compared to the Lipschitz continuity assumed in [155, 192, 194, 203, 234] and the Hölder continuity assumed in [202]. However, it is true for many porous media flow models, where the form of \( b \) is as in (9.51). \( \mathcal{D} \) can also be a matrix, the only constraint being that the
condition (P3) is satisfied for all the components \( \tilde{D}_{ij} \) and that \( \tilde{D} \) is positive definite. This is the same requirement as in (A4) of Chapter 1. The bounds assumed in (P4) guarantee that \( \tilde{F} \), which corresponds to ‘advective flux’ in physics, is bounded and varies smoothly within \( \Omega \).

**Remark 9.2 (Boundary conditions).** For the ease of presentation a zero Dirichlet boundary condition is assumed at the boundary, but the results can be proved for more general boundary conditions, including non-homogeneous Dirichlet, Neumann, Robin or mixed type ones.

### 9.2.2 Standard linearisation techniques

For resolving the nonlinear terms in problem \((\mathcal{P}1)\), Newton scheme (NS) uses iterations where the values of the nonlinear terms in the current iteration are approximated by Taylor series expansion. For \( i \in \mathbb{N} \) let \( u_n^i \) stand for the \( i^{th} \) iterate at the \( n^{th} \) time step and let \( \delta u_n^i = u_n^i - u_n^{i-1} \) for \( i > 0 \). If \( \phi_n^i \) is one of the functions \( b, f, z, \mathcal{D}, F_j \) for \( u = u_n^i \) and \( t = t_n \), then one takes \( \phi_n^{i+1} = \phi_n^{i-1} + \partial \phi_n\delta u^{i-1}_{n} \), \( \partial \phi_n\delta u^{i-1}_{n} \) being the partial derivative of \( \phi \) with respect to \( u \) at \( u = u_n^i \). With this substitution and given \( u_n^{i-1} \), NS resumes to find \( \delta u_n^i \) that solves

\[
(b'(u_n^{i-1}) - \Delta t \partial_u f^{i-1} \delta u_n^i - \Delta t \nabla \cdot [\mathcal{D}_n^{i-1} \nabla \delta u_n^i + (\partial_u \mathcal{D}_n^{i-1} \nabla u_n^{i-1} - \partial_u F_n^{i-1}) \delta u_n^i]) = -(b(u_n^{i-1}) - b(u_n^{i-1})) + \Delta t \left[ \nabla \cdot (\mathcal{D}_n^{i-1} \nabla u_n^{i-1}) - \nabla \cdot F_n^{i-1} + f_n^{i-1} \right] \quad \text{in } \Omega, \quad (9.7a)
\]

\[
\delta u_n^i = 0 \text{ on } \partial \Omega. \quad (9.7b)
\]

For Problem \((\mathcal{P}_N^{n,i})\), one of the natural initial guesses is \( u_n^0 = u_{n-1} \).

The modified Picard scheme (PS) can be interpreted as a simplified version of the NS [57]. Here the Taylor expansion is used only for the function \( z = b - \Delta t f \). For \( i \in \mathbb{N} \) and with given \( u_n^{i-1} \) one seeks \( u_n^{i} \) solving

\[
\begin{cases}
\partial u_n^{i-1} u_n^{i-1} - \Delta t \nabla \cdot \left[ \mathcal{D}_n^{i-1} \nabla u_n^i \right] = \partial u_n^{i-1} u_n^{i-1} - (b(u_n^{i-1}) - b(u_n^{i-1})) + \Delta t [-v \cdot \tilde{F}_{n-1} + f_n^{i-1}] \quad \text{in } \Omega, \quad (9.8a)
\end{cases}
\]

\[
\begin{array}{c}
u_n^{i} = 0 \quad \text{on } \partial \Omega. \quad (9.8b)
\end{array}
\]

The computational cost per iteration of the PS is lower than that for the NS, because the number of gradients to be calculated is less. However, the scheme is linearly convergent and also only conditionally stable. A convergence proof can be found in [206].

The PS takes advantage of the fact that the function \( z \) is increasing with respect to \( u \). This insight is taken a step further in the L-scheme (LS). For a \( \lambda \geq \max_{u \in \mathbb{R}, (\tilde{x}, t) \in Q} |\partial_u z| = \ldots \)
\[
\max_{u \in \mathbb{R}, (\bar{x}, t) \in \Omega} [(b'(u) - \Delta t \partial_u f(\bar{x}, t, u)]] \text{ and a given } u_{n-1}^i \text{ the scheme reads (see [192, 194, 203])}
\]
\[
(\mathcal{R}^{n,i}_L) \begin{cases}
L u_n^i - \Delta t \nabla \cdot \left[ \mathcal{D}^{i-1}_n \nabla u_n^i \right] = Lu_n^{i-1} - b(u_n^{i-1}) - b(u_{n-1}) + \Delta t [\nabla b^{i-1} + f^{i-1}] \quad \text{in } \Omega, \\
u_n^i = 0 \text{ on } \partial \Omega.
\end{cases}
\]
\]
\begin{equation}
(9.9a)
\end{equation}
\begin{equation}
(9.9b)
\end{equation}

The difference between the LS and the PS is that, instead of using the factor \(\partial_u z_n^{i-1}\), the former uses an upper bound of it, \(L\). This can affect the convergence order [155, 226], but transforms the scheme to a \(H^1\)-contraction which converges irrespective of the spatial discretization or the initial guess. Convergence can also be achieved for smaller values of \(L\), e.g. for \(L = \frac{1}{2} \max_{u \in \mathbb{R}, (\bar{x}, t) \in \Omega} (\partial u z)\) convergence is guaranteed (see [155, 226]) although the contraction property cannot be ensured in this case. In the next section we show how to improve the speed of the LS while preserving this stability.

Observe that each of the problems, \((\mathcal{R}^{n,i}_N), (\mathcal{R}^{n,i}_P)\) and \((\mathcal{R}^{n,i}_L)\), has unique solution in \(H^1_0(\Omega)\) as they are linear and coercive. Also, for the schemes mentioned above, if \(u_n^i \to u_n\) in strong sense in \(H^1_0(\Omega)\) then \(u_n\) is a weak solution of \((\mathcal{R}_I)\) in the sense of Definition 9.1.

For the convergence analysis we let \(e_n^i\) denote the error of the iterate at the \(n^{th}\) time step,
\[
e_n^i = u_n^i - u_n.
\]
\begin{equation}
(9.10)
\end{equation}

The convergence rate (linear) of the schemes is defined as
\[
\alpha = \lim_{i \to \infty} \sup \frac{\|e_n^i\|_{H^1(\Omega)}}{\|e_n^{i-1}\|_{H^1(\Omega)}}.
\]
\begin{equation}
(9.11)
\end{equation}

Notice that \(\alpha < 1\) implies convergence. Also worth pointing out is that the \(H^1\)-norm has been used to define the convergence rate in (9.11) but the convergence rate will roughly be the same if defined with respect to the \(L^2\) norm [155, 192]. Hence, in Section 9.5 the \(L^2\)-norm has been used to estimate \(\alpha\).

**Remark 9.3.** The convergence rate of the LS has the form \(\alpha = (L - m)/(L + C \Delta t) < 1\), for some \(C > 0\) and \(L > m\) (see [155, 192, 194, 234]). However, for large \(L\), or small \(\Delta t\) or \(m\), \(\alpha\) approaches 1. This leads to extremely slow convergence of the L-scheme. This issue has been reported for a variety of problems in literature [33, 155, 226].

The proofs given below make use of the following notations: \(|\cdot|_+ = \max(\cdot, 0)\) and \(|\cdot|_- = \min(\cdot, 0)\) are the positive and negative cut functions respectively. For any \(a, b \in \mathbb{R}\) the interval \(\mathcal{I}(a, b)\) is defined as
\[
\mathcal{I}(a, b) = \{x : \min(a, b) < x < \max(a, b)\}.
\]
9.3 The modified L-scheme

We will further use Young’s inequality: for \( a, b \in \mathbb{R} \) and \( \rho > 0 \) one has

\[
ab \leq \frac{1}{2\rho} a^2 + \frac{\rho}{2} b^2.
\]

(9.12)

Finally, \( C_\Omega \) is the constant appearing in the Poincaré inequality.

9.3 The modified L-scheme

In what follows we discuss a modified form of the L-scheme that preserves its stability property while having an improved convergence rate. The idea is to replace the constant \( L \) with a function \( L^m \) : \( \Omega \rightarrow \mathbb{R}^+ \) defined at each iteration. This leads to the modified L-scheme (MS). Let \( n \in \{1, \ldots, N\} \) and \( i \in \mathbb{N} \) be fixed and assume that \( u_{n-1}, u_{n}^{i-1} \in H^1_0(\Omega) \) are given. Find \( u_{n}^i \in H^1_0(\Omega) \) satisfying

\[
(L^m_n u^i_n, \phi) + b(u_{n}^{i-1}) - b(u_{n-1}, \phi) + \Delta t \frac{\partial}{\partial t} f(t_n, u_{n}^{i-1}, \phi) = (L^m_n u^i_n, \phi),
\]

(9.13)

for all \( \phi \in H^1_0(\Omega) \), where \( L^m_n : \Omega \rightarrow \mathbb{R} \) is defined as

\[
L^m_n(\hat{x}) = \max(|b'(u_{n}^{i-1}(\hat{x})) - b(u_{n-1}, \phi) + \Delta t f(t_n, u_{n}^{i-1}, \phi) + \Delta t f(t_n, u_{n}^{i-1}, \phi)|, 2\mathcal{H} \Delta t).
\]

(9.14)

Here \( u^0_n = u_{n-1} \), and \( \mathcal{H} \) is a positive constant that will be specified later. Observe that \( \mathcal{H} = 0 \) corresponds to the PS, and disregarding the \( b' - \Delta t \partial u f \) term in the definition of \( L^m_n \) leads to the LS. In this sense the scheme is a combination of the PS and the LS. Below we show that it inherits the qualities of both schemes. The convergence results and estimates are first obtained for a simple version of (P1) where \( \partial_u \mathcal{D} \equiv 0 \) and \( \partial_u F_j \equiv 0 \) and then for the general problem.

9.3.1 Quasilinear equations with linear diffusivity and flux

To analyse the scheme we first look at the simpler quasilinear parabolic equation,

\[
\partial_t b(u) + \nabla \cdot \vec{F}(\hat{x}) = \nabla \cdot (\partial_u \mathcal{D}(\hat{x}) \nabla u) + f
\]

(9.15)

which corresponds to \( \partial_u \mathcal{D} = 0 \) and \( \partial_u F_j = 0 \) in (9.1). Problems where \( \partial_u \mathcal{D} \neq 0 \) can also be reduced to this case if \( \mathcal{D} \) is separable in the variables, i.e. \( \mathcal{D}(\hat{x}, u) = \mathcal{D}_1(\hat{x})\mathcal{D}_2(u) \).

By (P3), \( \mathcal{D}_2 > 0 \) and by using the Kirchhoff transform \( U = \int u \mathcal{D}_2(d\bar{\varrho}) \) (see [9, 192, 263]) one obtains \( \nabla \cdot [\mathcal{D}_1(\hat{x}, u) \nabla U] = \nabla \cdot (\partial_u \mathcal{D}_1(\hat{x}) \nabla U) \). Using \( U \) as the primary unknown one gets an equation similar to (9.15). Well-known examples, such as the porous medium equation or the Richards equation, can be reduced to this form.
Recalling the homogeneous boundary conditions, a weak solution to the time discrete version of (9.15) satisfies
\[
(b(u_n) - b(u_{n-1}), \phi) + \Delta t(\nabla \nabla u_n, \nabla \phi) = \Delta t(f_n, \phi) + \Delta t(F, \nabla \phi),
\] (9.16)
for all \( \phi \in H^1_0(\Omega) \). For the subsequent analysis we assume the following:

(A1) There exists a \( \Lambda > 0 \) such that \( \|u_n - u_{n-1}\|_{L^\infty(\Omega)} \leq \Lambda \Delta t \) for all \( n \in \mathbb{N} \).

Note that since \( u_n \) is, in fact, the time discrete approximation of the solution to the parabolic problem (9.15), Assumption (A1) is similar to saying that \( \partial_t u \in L^\infty(Q) \). Sufficient conditions for boundedness of \( \partial_t u \) in the \( L^\infty(\Omega) \)-norm can be found in [146, Chapter 5]. Below we present a result that shows the validity of Assumption (A1), in this sense.

**Proposition 9.1.** For a fixed \( n \in \{1, \ldots, N\} \) let \( u_n \) solve (9.16) and assume that \( m > 0 \) and \( \nabla \cdot (\partial_t \nabla u_{n-1}) \in L^\infty(\Omega) \). Then \( \nabla \cdot (\partial_t \nabla u_n) \in L^\infty(\Omega) \) and a \( \Lambda \geq 0 \), independent of \( \Delta t \), exists such that \( \|u_n - u_{n-1}\|_{L^\infty(\Omega)} \leq \Lambda \Delta t \).

The proof is given in Appendix B.1. Observe that Proposition 9.1 is valid if \( u_0 \in W^{2,\infty}(\Omega) \), which extends then to all time steps. Based on Assumption (A1) we have

**Lemma 9.1.** Assume (A1) and that \( L^j_n \) satisfies a.e. in \( \Omega \)
\[
L^j_n \geq \sup \{ |b'(|z|) - \Delta t \partial_u f(\tilde{x}, t_n, \zeta)| : \zeta \in \mathcal{S}(u_n^{-1}, u_n) \}.
\]
Then \( \|u^j_n - u_n\|_{L^\infty(\Omega)} \leq \Lambda \Delta t \) for all \( i \in \mathbb{N} \), \( u_n^j \) being the solution of (9.13).

Before giving the proof we observe that the result is quite general with respect to the choice of the function \( L^j_n \). However, it will be used for \( L^j_n \) either constant or as in (9.14), which corresponds to the LS and the MS.

**Proof.** The proof is by induction. Assumption (A1) guarantees that the assertion holds for \( i = 0 \). Let the assertion hold for \( u_n^i - u_{n-1} \|_{L^\infty(\Omega)} = \|e_n^{i-1}\|_{L^\infty(\Omega)} \leq \Lambda \Delta t \). We prove that this implies \( \|u^j_n - u_n\|_{L^\infty(\Omega)} = \|e_n^j\|_{L^\infty(\Omega)} \leq \Lambda \Delta t \).

As \( \partial_u \mathcal{S}, \partial_u F_j = 0 \), subtracting (9.16) from (9.13), a \( \zeta \in \mathcal{S}(u_n^{-1}, u_n) \) exists such that
\[
(L^j_n e_n^{i-1}, \phi) + \Delta t(\nabla \nabla e_n^{i-1}, \nabla \phi) = (L^j_n e_n^{i-1}, \phi) - ((b(u_n^{-1}) - b(u_n)), \phi) + \Delta t(f_n^{i-1} - f_n, \phi)
\]
\[
= ((L^j_n - \partial_u \mathcal{S}(\zeta)) e_n^{i-1}, \phi).
\]
(9.17)
Here we used the definition of \( z \) from (9.5). With \( \phi = |e_n^i - \Lambda \Delta t|_+ \) one gets
\[
(L^j_n |e_n^i - \Lambda \Delta t|_+, |e_n^i - \Lambda \Delta t|_+) + \Lambda \Delta t(L^j_n, |e_n^i - \Lambda \Delta t|_+) + \Delta t \mathcal{S}_m \|\nabla |e_n^i - \Lambda \Delta t|_+ \|^2
\]
\[
\leq (L^j_n e_n^i, |e_n^i - \Lambda \Delta t|_+) + \Delta t(\nabla \nabla e_n^i, \nabla |e_n^i - \Lambda \Delta t|_+) = ((L^j_n - \partial_u \mathcal{S}(\zeta)) e_n^{i-1}, |e_n^i - \Lambda \Delta t|_+)
\]
\[
\leq \int_\Omega |L^j_n - \partial_u \mathcal{S}(\zeta)| e_n^{i-1} \|e_n^i - \Lambda \Delta t|_+ \leq \int_\Omega (L^j_n - m) \Delta \Delta t |e_n^i - \Lambda \Delta t|_+ \leq \Lambda \Delta t (L^j_n, |e_n^i - \Lambda \Delta t|_+).
\]
In the last estimates we used the inequalities $|e_n^{i-1}| \leq \Lambda \Delta t$ a.e. and $0 \leq (L_n^i - \partial_u z(\zeta)) \leq L_n^i - m$ for $\zeta \in \mathcal{F}(u_n^{i-1}, u_n)$. Cancelling the common terms in both sides gives

$$(L_n^i, [e_n^i - \Lambda \Delta t]_+^2) + \Delta t M \|e_n^i - \Lambda \Delta t\| \leq 0,$$

which implies that $e_n^i < \Lambda \Delta t$ a.e. in $\Omega$. Similarly taking $\phi = [e_n^i + \Lambda \Delta t]_-$ one gets that $e_n^i > -\Lambda \Delta t$ a.e. which concludes the proof. \qed

For the LS, the condition stated in Lemma 9.1 is satisfied as $L_n^i = L \geq \text{sup}(|\partial_u z(x, \zeta)| : \zeta \in \mathbb{R})$. However this leads to an overestimation of $L$ at most points. Below we show that this estimation is improved significantly for the MS, resulting in much better convergence rates.

**Theorem 9.1.** Under the Assumptions (A1) and (P1)-(P5) and for $\mathcal{M}_0 = \Lambda \max_{u \in \mathbb{R}} |b''| + \Delta t |\partial_{uu} f| \geq 0$, the MS for equation (9.16) converges linearly in $H^1_0(\Omega)$ for all $\mathcal{M} \geq \mathcal{M}_0$ and $\Delta t > 0$. More precisely, for $\mathcal{M} \geq \mathcal{M}_0$ and $\Delta t > 0$ the limit $u_n = \lim_{i \to \infty} u_n^i$ exists and is a solution to (9.16), whereas for the convergence rate $\alpha$ in (9.11) one has $\alpha < 1$. Moreover if $m > 0$ (the non-degenerate case) then $\alpha = \Theta(\Delta t)$ for $\Delta t$ small enough.

**Proof.** By (P1) and (P2), $\mathcal{M}_0$ is well defined. Observe that, for any $\zeta \in \mathcal{F}(u_n^{i-1}, u_n)$ there exists a $\zeta_1 \in \mathcal{F}(u_n^{i-1}, \zeta)$ such that

$$|b'(u_n^{i-1}) - \Delta t \partial_u f(x, t_n, u_n^{i-1}) - b'\zeta_1 - \Delta t \partial_u f(x, t_n, \zeta)| = |\partial_u z(u_n^{i-1}) - \partial_u z(\zeta)|$$

$$= |\partial_{uu} z(\zeta)| |(u_n^{i-1} - \zeta)| \leq \max_{u \in \mathbb{R}} |b''| + \Delta t |\partial_{uu} f| \Lambda \Delta t = \mathcal{M}_0 \Lambda \Delta t.$$

This implies that if $\mathcal{M} \geq \mathcal{M}_0$ and $L_n^i = \partial_u z(u_n^{i-1}) + \mathcal{M}_0 \Lambda \Delta t$, then $L_n^i - \partial_u z(\zeta) \geq 0$. Moreover, if $L_n^i = 2\mathcal{M} \Delta t$ then $\partial_u z(u_n^{i-1}) \leq 2\mathcal{M} \Delta t$, which means that $\partial_u z(\zeta) \leq \partial_u z(u_n^{i-1}) + \mathcal{M}_0 \Delta t \leq 2\mathcal{M} \Delta t$, giving $L_n^i - \partial_u z(\zeta) \geq 0$. Hence, for $\mathcal{M} \geq \mathcal{M}_0$ one has

$$L_n^i - b'(\zeta) + \Delta t \partial_u f(x, t_n, \zeta) \geq 0,$$

which by Lemma 9.1 implies that $\|e_n^i\|_{L^\infty(\Omega)} \leq \Lambda \Delta t$ for all $i \in \mathbb{N}$.

Using similar arguments, if $L_n^i = \partial_u z(u_n^{i-1}) + \mathcal{M} \Delta t$ and $\mathcal{M} \geq \mathcal{M}_0$ then one gets

$$L_n^i - b'(\zeta) + \Delta t \partial_u f(x, t_n, \zeta) = \partial_u z(\zeta)(u_n^{i-1} - \zeta) + \Lambda \Delta t |\partial_{uu} z(\zeta)| + 2\mathcal{M} \Delta t \leq 2\mathcal{M} \Delta t.$$

If $L_n^i = 2\mathcal{M} \Delta t$ then $L_n^i - \partial_u z(\zeta) \leq L_n^i - m \leq 2\mathcal{M} \Delta t$. Combining both gives

$$0 \leq L_n^i - b'(\zeta) + \Delta t \partial_u f(x, t_n, \zeta) \leq 2\mathcal{M} \Delta t,$$

or, more strongly,

$$0 \leq (\mathcal{M} - \mathcal{M}_0) \Delta t \leq L_n^i - b'(\zeta) + \Delta t \partial_u f(x, t_n, \zeta) \leq 2\mathcal{M} \Delta t.$$
These two inequalities will be used repeatedly in the proofs that follow.

First we prove the convergence of the scheme in $H^1_0(\Omega)$. Taking $\phi = e^i_n$ in (9.17) yields

$$2\mathcal{M}\Delta t\|e^i_n\|^2 + \Delta t\mathcal{D}m\|\nabla e^i_n\|^2 \leq (L^i_ne^i_n, e^i_n) + \Delta t(\mathcal{D}\nabla e^i_n, \nabla e^i_n)$$

$$= ((L^i_n - \partial_u z(\zeta))e^{i-1}_n, e^i_n) \leq 2\mathcal{M}\Delta t(|e^{i-1}_n|, |e^i_n|) \leq 2\mathcal{M}\Delta t\|e^{i-1}_n\|^2 + 2\mathcal{M}\Delta t\|e^i_n\|^2. \quad (9.20)$$

Cancelling common terms on both sides and applying the Poincaré inequality one gets

$$\left(\mathcal{M} + \frac{\mathcal{D}m}{2} C_\Omega\right)\|e^i_n\|^2 + \frac{\mathcal{D}m}{2} \|\nabla e^i_n\|^2 \leq 2\mathcal{M}\|e^{i-1}_n\|^2,$$

which gives

$$\|e^i_n\|^2 + \frac{\mathcal{D}m}{2(\mathcal{M} + C_{11}\mathcal{D}m)} \|\nabla e^i_n\|^2 \leq \frac{2\mathcal{M}}{(\mathcal{M} + C_{11}\mathcal{D}m)} \left(\|e^{i-1}_n\|^2 + \frac{\mathcal{D}m}{(\mathcal{M} + C_{11}\mathcal{D}m)} \|\nabla e^{i-1}_n\|^2\right).$$

As $\left(\|u\|^2 + \frac{\mathcal{D}m}{2(\mathcal{M} + C_{11}\mathcal{D}m)} \|\nabla u\|^2\right)^{1/2}$ is equivalent to the $H^1(\Omega)$-norm it follows that $(u^i_n)_{i \in \mathbb{N}}$ converges in $H^1(\Omega)$. The convergence rate in this equivalent norm is

$$\alpha = \sqrt{2\mathcal{M}/(\mathcal{M} + C_{11}\mathcal{D}m)} < 1. \quad (9.21)$$

Observe that the convergence does not depend on the spatial discretization and also holds in the degenerate case when $\partial_u z$ may vanish.

In the non-degenerate case, when $\partial_u z \equiv m > 0$, substituting $\phi = e^i_n$ in (9.17) one gets

$$(m + \mathcal{M}\Delta t)\|e^i_n\|^2 + \Delta t\mathcal{D}m\|\nabla e^i_n\|^2 \leq (L^i_ne^i_n, e^i_n) + \mathcal{D}m\Delta t\|\nabla e^i_n\|^2$$

$$= ((L - \partial_u z(\zeta))e^{i-1}_n, e^i_n) \leq 2\mathcal{M}\Delta t(|e^{i-1}_n|, |e^i_n|) \leq \frac{2\mathcal{M}^2\Delta t^2}{m} \|e^{i-1}_n\|^2 + \frac{m}{2} \|e^i_n\|^2.$$

Cancelling the common term in both sides we obtain

$$\|e^i_n\|^2 + \frac{2\Delta t \mathcal{D}m}{(m + 2\mathcal{M}\Delta t)^2} \|\nabla e^i_n\|^2 \leq \frac{4\mathcal{M}^2\Delta t^2}{m(m + 2\mathcal{M}\Delta t)} \left(\|e^{i-1}_n\|^2 + \frac{2\Delta t \mathcal{D}m}{(m + 2\mathcal{M}\Delta t)^2} \|\nabla e^{i-1}_n\|^2\right).$$

With $\Delta t$ small enough, one obtains as before, the convergence of $u^i_n$ in $H^1(\Omega)$. The convergence rate is

$$\alpha = \frac{2\mathcal{M}\Delta t}{\sqrt{m(m + 2\mathcal{M}\Delta t)}} < \frac{2\mathcal{M}\Delta t}{m}. \quad (9.23)$$
which is less than 1 for $\Delta t < \frac{m}{2M}$. One can also use the inequality $2M\Delta t(|e_{n}^{i-1}|, |e_{n}^{i}|) < 2M\Delta t(\|e_{n}^{i-1}\|^2 + \|e_{n}^{i}\|^2)$ to prove contraction with respect to a different $H^1(\Omega)$-norm with $\alpha = \sqrt{\frac{2M}{m}}$. Hence, the actual convergence rate is

$$\alpha = \min \left( \frac{2M\Delta t}{\sqrt{m(m+2M\Delta t)}}, \frac{\sqrt{M\Delta t}}{m}, \frac{2M}{2M + C_{\Omega}Dm} \right).$$  (9.24)

We conclude this section with a result that shows that for the non-degenerate case even the $L^\infty(\Omega)$ errors, i.e. $\|e_{n}^{i}\|_{L^\infty(\Omega)}$, decrease linearly for $\Delta t$ sufficiently small.

**Proposition 9.2.** For a fixed $n \in \{1,..,N\}$, let $\{u_{i}^{n}\}_{i \in \mathbb{N}}$ be the sequence of functions resulting from the modified L-scheme for (9.16). Assume (A1) and (P1)-(P5). If $m > 0$, then for small enough $\Delta t$ and $\mathfrak{M} \geq \mathfrak{M}_0$,

$$\|u_{i}^{n} - u_n\|_{L^\infty(\Omega)} < \beta \|u_{i}^{n-1} - u_n\|_{L^\infty(\Omega)},$$

for $i \in \mathbb{N}$ and a constant $\beta \in (0,1)$. Moreover, $\beta = O(\Delta t)$.

The proof is given in Appendix B.1.

**Remark 9.4 (Robustness of the MS).** Theorem 9.1 shows that the MS converges irrespective of the spatial discretization and the time step. It converges in the degenerate case $m = 0$ too. Unlike the LS, the convergence rate $\alpha$ is independent of $M_h$, and scales with $\Delta t$ for small $\Delta t$. This robustness is extremely helpful for computations, as it is difficult to satisfy condition (9.2), which guarantees the convergence of schemes like NS, PS and Jäger-Kačur, for higher dimensional computations, i.e. $d > 1$. Moreover, the $\alpha = O(\Delta t)$ property makes the scheme faster as the time step size is made smaller.

### 9.4 General nonlinear diffusion equation

For the general problem, when $\bar{F}$ and $\bar{F}$ are functions of $u$, convergence can be shown for both the LS and the MS when $\Delta t$ is small enough. For both schemes, the convergence rate does not depend on the spatial discretization. Moreover, for the MS the convergence rate scales with $\sqrt{\Delta t}$ for small $\Delta t$ values.

For proving the main theorem of this section, we assume

(A2) $\|\nabla u_n\|_{L^\infty(\Omega)} \leq \Lambda_1$ for all $n \in \{1,..,N\}$ and some $\Lambda_1 > 0$.

Due to (P3), this is equivalent to assuming that the flux is bounded.

Similar to Section 9.3, the MS works if for all $i \in \mathbb{N}$ one has

$$\|u_{i}^{n} - u_n\|_{L^\infty(\Omega)} \leq \Lambda \Delta t.$$  (9.25)
Though non-trivial, this condition is fulfilled under certain assumptions, as follows from

**Lemma 9.2.** Let \( n \in \{1, \ldots, N\} \) be fixed and \( \Omega \) be a \( C^2 \) domain. Assume (A1), \( m > 0 \) and let \( \Lambda_2 > 0 \) be such that \( \|u_n\|_{W^{2,2q}(\Omega)} \leq \Lambda_2 \) for some \( q > \frac{d}{2} \), \( q \in \mathbb{N} \). Further, assume that a \( \Lambda_3 > 0 \) exists such that

\[
\|u_n - u_{n-1}\|_{W^{1,2q}(\Omega)} \leq \Lambda_3 \Delta t. \tag{9.26}
\]

Let \( \{u^i_n\}_{i \in \mathbb{N}} \) be the sequence generated by the MS. Then there exists a \( \Delta t > 0 \) such that for all \( \Delta t < \Delta \) and \( i \in \mathbb{N} \), \( \|u^i_n - u_{n+1}\|_{L^{\infty}(\Omega)} \leq \Lambda_3 \Delta t \) and \( u^i_n \in W^{2,2q}(\Omega) \) for small \( \Delta t \). More specifically, we show that there exists a \( C > 0 \) such that

\[
\|e^i_n\| \leq C\Delta t^2, \quad \|e^i_n\|_{L^\infty(\Omega)} \leq C\Delta t^{1+\frac{1}{2d}} \quad \text{and} \quad \|e^i_n\|_{W^{1,p}(\Omega)} \leq C\Delta t^{1+\frac{1}{p}}, \quad \|e^i_n\|_{W^{2,2q}(\Omega)} \leq C\Delta t, \tag{9.27}
\]

for all \( p \geq 2 \) and \( \Delta t > 0 \). This proves the lemma for \( \Delta t \) small enough.

First, observe that inequality (9.18) holds for \( L^i_n \) defined in (9.14) Moreover, the assumption \( u_n \in W^{2,2q}(\Omega) \) implies (A2), i.e. there exists a \( \Lambda_1 > 0 \) such that \( \|\nabla u_n\|_{L^{\infty}(\Omega)} \leq \Lambda_1 \) This is a direct consequence of Morrey’s inequality [89, Chapter 5] and \( 2q > d \). By the same argument \( u^i_n \in W^{1,\infty}(\Omega) \in H^{1}(\Omega) \). Hence, by Theorem 5 of [89, Chapter 6] it follows that \( u^i_n \) is a classical solution to

\[
L_n^i u^i_n - \Delta t \nabla \cdot (\nabla u^i_n - b^i_n) = L_n^i u^i_n - (b^i_n - b_{n-1}) + \Delta t \nabla \cdot F^i_n + \Delta t f^i_n. \tag{9.28}
\]

In particular, \( u^i_n \) is essentially bounded and lies in \( H^1_0 \) as well.

Subtracting (9.3a) from (9.28) and rearranging the terms leads to

\[
L_n^i e^i_n - \Delta t \nabla \cdot (\nabla u^i_n - b^i_n) = L_n^i e^i_n - \Delta t \nabla \cdot (\nabla u^i_n - \nabla u_n) - \Delta t \nabla \cdot (F^i_n - F_n), \tag{9.29}
\]
9.4 General nonlinear diffusion equation

where \( \zeta \in \mathcal{S}(u_n^{i-1}, u_n) \). Denoting the terms on the right by \( I_1, I_2, I_3 \) we have

\[
|I_1| = |(I_n^i - \partial u \zeta(\zeta))e_n^{i-1}| \leq 2 \Lambda M \Delta t^2;
\]

\[
|I_2| = \Delta t|\mathcal{D}_{n}^{i-1} - \mathcal{D}_{n}| \Delta u_n + (\partial u \mathcal{D}^{i-1}_n \nabla u_n^i - \partial u \mathcal{D}_n \nabla u_n) \cdot \nabla u_n + \sum_{j=1}^{d} (\partial_{x_j} \mathcal{D}^{i-1}_n - \partial_{x_j} \mathcal{D}_n) \partial_{x_j} u_n |
\]

\[
\leq \Lambda \mathcal{D}_M \Delta t^2 |\Delta u_n| + \Delta t|\nabla u_n||\partial u \mathcal{D}^{i-1}_n \nabla e_n^{i-1} + \partial u \mathcal{D}^{i-1}_n - \partial u \mathcal{D}_n \nabla u_n| + \Delta t|\nabla u_n||\sum_{j=1}^{d} (\partial_{x_j} \mathcal{D}^{i-1}_n - \partial_{x_j} \mathcal{D}_n)|
\]

\[
\leq \Lambda \mathcal{D}_M \Delta t^2 |\Delta u_n| + \Delta t \mathcal{D}_M \Lambda |\nabla e_n^{i-1}| + \mathcal{D}_M \Lambda \Delta t^2 + d \Lambda_1 \mathcal{D}_M \Delta t^2;
\]

\[
|I_3| = \Delta t|\mathbf{n} \cdot (\mathcal{D}_n^{i-1} - \mathcal{D}_n)| = \Delta t|\partial u \mathcal{D}^{i-1}_n \nabla u_n^i - \partial u \mathcal{D}_n \nabla u_n + \sum_{j=1}^{d} (\partial_{x_j} \mathcal{D}^{i-1}_n - \partial_{x_j} \mathcal{D}_n)|
\]

\[
\leq \Delta t|\partial u \mathcal{D}^{i-1}_n \nabla u_n^i - \partial u \mathcal{D}_n \nabla u_n| + \Delta t|\partial u \mathcal{D}^{i-1}_n - \partial u \mathcal{D}_n| + \Delta t|\sum_{j=1}^{d} (\partial_{x_j} \mathcal{D}^{i-1}_n - \partial_{x_j} \mathcal{D}_n)|
\]

\[
\leq \Delta t |\partial u \mathcal{D}^{i-1}_n \nabla u_n^i| + \Delta t \mathcal{D}_M |\nabla e_n^{i-1}| + \mathcal{D}_M \Lambda x^2 \Delta t^2 + d \Lambda_1 \mathcal{D}_M \Delta t^2.
\]

Define \( S_i := I_1 + I_2 + I_3 \). As \( |e_n^{i-1}| \|W^{1,2q}(\Omega)\| \leq \Lambda_3 \Delta t \) and \( \|\Delta u_n\|_{L^{2q}(\Omega)} \leq \Lambda_2 \) it follows that a \( C_1 > 0 \) exists such that

\[
\|S_i\|_{L^{2q}(\Omega)} \leq C_1 \Delta t^2. \tag{9.30}
\]

Now we test (9.29) with the test function \( \phi = (e_n^{i-1})^{2q-1} \in H_0^1(\Omega) \). This gives

\[
m \|e_n^{i-1}\|_{L^{2q}(\Omega)}^{2q} + \Delta t \mathcal{D}_M (2q-1) \int_{\Omega} |e_n^{i-1}|^{2q-1} \nabla e_n^{i-1} |^2 \leq (|S_i|, |e_n^{i-1}|^{2q-1}) \]

\[
\leq \frac{1}{2qm^{2q-2}} \|S_i\|_{L^{2q}(\Omega)}^{2q} + \frac{(2q-1)m}{2q} \|e_n^{i-1}\|_{L^{2q}(\Omega)}^{2q}.
\]

The last inequality follows from repeated application of Young's inequality. This implies

\[
\|e_n^{i-1}\|_{L^{2q}(\Omega)} \leq \frac{1}{m} \|S_i\|_{L^{2q}(\Omega)} \leq \frac{C_1}{m} \Delta t^2.
\]

Rewriting (9.29) as

\[
\nabla \cdot (\mathcal{D}^{i-1}_n \nabla e_n^{i}) = I_n^i \left( \frac{e_n^{i}}{\Delta t} \right) - \frac{1}{\Delta t} S_i, \tag{9.31}
\]

we see that the \( L^{2q}(\Omega) \)-norm of the right hand side is bounded by some constant times \( \Delta t \). As \( |\nabla \mathcal{D}^{i-1}_n| \leq |\partial u \mathcal{D}^{i-1}_n \nabla u_n^i| + \sum_{j=1}^{d} |\partial_{x_j} \mathcal{D}^{i-1}_n| \leq \mathcal{D}_M (|\nabla e_n^{i-1}| + \Lambda_1 + d), |\nabla \mathcal{D}^{i-1}_n| \in L^{2q}(\Omega) \). Hence, we apply Theorem 15.1 of [147, Chapter 2] to get that \( e_n^{i} \in W^{2,2q}(\Omega) \cap C^{1,\gamma}(\Omega) \) for \( \gamma = 1 - \frac{d}{2q} \). Moreover, there exists a \( C > 0 \) such that

\[
\|e_n^{i}\|_{W^{2,2q}(\Omega)} \leq C \Delta t \text{ and } \|\nabla e_n^{i}\|_{L^{\infty}(\Omega)} \leq \|e_n^{i}\|_{C^{1,\gamma}(\Omega)} \leq C \Delta t. \tag{9.32}
\]

This proves the last statement of (9.27).
Next, we test (9.29) with \( \phi = e^j_n \) to get
\[
m\|e^j_n\|^2 + \Delta t m \|\nabla e^j_n\|^2 \leq (S_j, e^j_n) \leq \frac{1}{2m} \|S_j\|^2 + \frac{m}{2} \|e^j_n\|^2.
\]
(9.33)

As \( S_j \in L^{2q}(\Omega) \subseteq L^2(\Omega) \) we get using (9.30) that for some constant \( C_2 > 0 \)
\[
\|\nabla e^j_n\|^2 \leq C_2 \Delta t^3.
\]
(9.34)

This implies that for \( p \geq 2, \)
\[
\|\nabla e^j_n\|_{L^p(\Omega)} \leq \left( \int_\Omega |\nabla e^j_n|^p \right)^{1/p} \leq \left( \|\nabla e^j_n\|_{L^\infty(\Omega)}^{p-2} \int_\Omega |\nabla e^j_n|^2 \right)^{1/p} \leq (C^{p-2} C_2)^{1/p} \Delta t^{1 + \frac{1}{p}}.
\]
(9.35)

Finally, as \( C^0(\Omega) \subseteq W^{1,2q}(\Omega) \), see Morrey’s inequality (Theorem 5 of [89, Chapter 5]), we get that
\[
\|e^j_n\|_{L^\infty(\Omega)} \leq \|e^j_n\|_{C^0(\Omega)} \leq C \Delta t^{1 + \frac{1}{2q}}.
\]
(9.36)

From this we get \( \|e^j_n\|_{L^{2q}(\Omega)} \leq \Lambda \Delta t \) if \( \Delta t \leq (\Lambda/C)^{2q} \). Similarly one obtains \( \|\nabla e^j_n\|_{L^2(\Omega)} \leq \Lambda_3 \Delta t \) for small \( \Delta t \).

Before giving the main theorem of this section we give a context that can lead to improved convergence behaviour of the MS. Specifically, we assume
\[
|\partial_u \mathcal{D}(\bar{x}, u) + \sum_{j=1}^d |\partial_u F_j(\bar{x}, u)| |Y \partial_u z(\bar{x}, u)| \text{ a.e. in } \bar{x} \in \Omega \text{ and } u \in \mathbb{R}.
\]
(9.37)

Remark 9.5. The bound (9.37) is true, e.g. for the Richards equation (Ri). In this case the diffusivity and the flux terms are of the form \( \mathcal{D} = k(b(u)) \) and \( F = k(b(u)) \hat{e}_g \) (\( \hat{e}_g \) is a constant unit vector) with \( b : \mathbb{R} \to [0,1] \) and \( k \in C^1([0,1]) \) giving \( |\partial_u \mathcal{D}|, |\partial_u F_j| \leq \sup_{s \in [0,1]} |k'(s)| b'(u) \). Also the quasilinear system discussed in Section 9.3 is just a special case when \( Y = 0 \).

Theorem 9.2. For a fixed \( n \in \{1, \ldots, N\} \) let \( \{u^j_n\} \), \( j \in \mathbb{N} \) be the sequence provided by the MS. Assume (9.25) holds for \( i \in \mathbb{N} \). If Assumptions (A1)-(A2) and (P1)-(P5) are satisfied then the following hold:

(a) If inequality (9.37) is satisfied then there exist a \( M_1 > 0 \) and \( \Delta t^* > 0 \) independent of \( m \) such that if \( \mathcal{M} \geq M_1 \) and \( \Delta t \leq \Delta t^* \) then \( u^j_n \rightarrow u_n \) in the strong sense in \( H^1(\Omega) \).

(b) If \( m > 0 \) then there exists a \( \Delta \bar{t} > 0 \) such that for \( \Delta t < \Delta \bar{t} \) and \( \mathcal{M} \geq M_0 \), \( u^j_n \) converges linearly to \( u_n \) in \( H^1(\Omega) \). Moreover \( a = \mathcal{O}(\sqrt{\Delta t}) \) for this case.
Proof (a). We follow the line of arguments presented in [155] for proving this part. Subtracting (9.4) from (9.13) and taking \( \phi = e_n^i \) gives

\[
(L_n^i(e_n^i - e_n^{i-1}), e_n^i) + \Delta t (D_n^{i-1} \nabla u_n^i - D_n \nabla u_n, \nabla e_n^i) = -(z(u_n^{i-1}) - z(u_n^i), e_n^i) + \Delta t (\tilde{F}_n^{i-1} - \tilde{F}_n, \nabla e_n^i).
\]

(9.38)

This can be rearranged into the form \( T_1 + \Delta t T_2 + T_3 = T_4 + \Delta t T_5 + \Delta t T_6 \) where the terms \( T_j \) are estimated as:

\[
T_1 := (L_n^i(e_n^i - e_n^{i-1}), e_n^i) = \frac{1}{2} \int_\Omega L_n^i |e_n^i|^2 - \frac{1}{2} \int_\Omega L_n^i |e_n^{i-1}|^2 + \frac{1}{2} \int_\Omega L_n^i |e_n^i - e_n^{i-1}|^2;
\]

\[
T_2 := \int_\Omega D_n^{i-1} |\nabla e_n^i|^2 \geq \frac{D_m}{2} \| \nabla e_n^i \|^2;
\]

\[
T_3 := (z(u_n^{i-1}) - z(u_n^i), e_n^{i-1}) = \int_\Omega \frac{1}{\partial u z(\zeta)} |z(u_n^{i-1}) - z(u_n)|^2;
\]

\[
T_4 := (z(u_n^{i-1}) - z(u_n^i), e_n^{i-1} - e_n^i) \leq \frac{1}{2} \int_\Omega \frac{1}{\partial u z(\zeta)} |z(u_n^{i-1}) - z(u_n)|^2 + \frac{1}{2} \int_\Omega |e_n^{i-1} - e_n^i|^2;
\]

\[
T_5 := -((D_n^{i-1} - D_n) \nabla u_n, \nabla e_n^i) \leq \frac{1}{D_m} \| (D_n^{i-1} - D_n) \nabla u_n \|^2 + \frac{D_m}{4} \| \nabla e_n^i \|^2
\]

\[
\leq \frac{\Lambda^2}{D_m} \| \frac{\partial u \zeta(1)}{\partial u z(\zeta)} (z(u_n^{i-1}) - z(u_n)) \|^2 + \frac{D_m}{4} \| \nabla e_n^i \|^2 \leq \frac{(Y \Lambda)^2}{D_m} \| z(u_n^{i-1}) - z(u_n) \|^2 + \frac{D_m}{4} \| \nabla e_n^i \|^2;
\]

\[
T_6 := (\tilde{F}_n^{i-1} - \tilde{F}_n, \nabla e_n^i) \leq \frac{1}{D_m} \| F_n^{i-1} - F_n \|^2 + \frac{D_m}{4} \| \nabla e_n^i \|^2
\]

\[
\leq \frac{1}{D_m} \sum_{j=1}^{d} \left( \frac{\partial u \zeta_j(2)}{\partial u z(\zeta)} (z(u_n^{i-1}) - z(u_n)) \right)^2 + \frac{D_m}{4} \| \nabla e_n^i \|^2 \leq \frac{Y^2}{D_m} \| z(u_n^{i-1}) - z(u_n) \|^2 + \frac{D_m}{4} \| \nabla e_n^i \|^2.
\]

For the functions \( \zeta_1, \zeta_2 : \Omega \to \mathbb{R} \) one has \( \zeta_{1,2} \in \mathcal{F}(u_n^{i-1}, u_n) \). For \( T_3 \), if \( z(u_n^{i-1}) = z(u_n) \) then the equality holds for any \( \zeta_0 \in \mathbb{R} \) for which \( \partial u z(\zeta_0) \neq 0 \). If \( z(u_n^{i-1}) \neq z(u_n) \) then by the Mean Value Theorem there exists a \( \zeta_0 \in \mathcal{F}(u_n^{i-1}, u_n) \) such that \( \partial u z(\zeta_0) \neq 0 \). Hence, there exists a \( \zeta_0 : \Omega \to \mathbb{R} \) such that equality for \( T_3 \) is satisfied and \( \partial u z(\zeta_0) \neq 0 \) a.e. In the following, \( T_3 \) will only induce an upper bound for \( \Delta t \) which will depend on the lower bound of \( \frac{1}{\partial u z(\zeta_0)} \). Therefore we claim that the choice of \( \zeta_0 \), discussed above, has no influence on the result.

Putting everything together we get the following inequality

\[
\int_\Omega L_n^i |e_n^i|^2 + \Delta t D_m \| \nabla e_n^i \|^2 + \int_\Omega \left( \frac{2}{\partial u z(\zeta_0)} - \frac{1}{L_n^i} - \frac{2 \Delta t Y^2}{D_m} (1 + \Lambda^2) \right) |z(u_n^{i-1}) - z(u_n)|^2
\]

\[
\leq \int_\Omega L_n^i |e_n^{i-1}|^2 \leq \int_\Omega L_n^{i-1} |e_n^{i-1}|^2 + \int_\Omega |L_n^i - L_n^{i-1}| |e_n^{i-1}|^2.
\]

(9.39)
This inequality is useful if all the terms on the left are positive. This is achieved if
\[ \frac{2}{\partial u_x(z)} - \frac{1}{L_n^0} - \frac{2\Delta t Y^2}{\partial_m} (1 + \lambda_1^2) \geq 0, \]
which will add some restrictions on \( \Delta t \). To see this we define
\[ M_1 = \max\{4M_0, 2M_0 + G\}, \] (9.40)
where \( M_0 = \Lambda \max_{u \in \Omega} |b''(u) + \Delta t \partial uu f(u)| \) is defined just as in the proof of Theorem 9.1 and the value of \( G \) will be clarified later. Observe that \( \left( \frac{2}{\partial u_x} - \frac{1}{L_n^0} \right) = \frac{1}{\partial u_x} + \frac{L_n^0 - \partial u_x}{L_n^0 \partial u_x} \). From (9.19) we get \( L_n^0 - \partial u_x(z) > (M - M_0) \Delta t \). Moreover, from (P1)-(P2), \( \partial u_x \leq M_b + TM_f \) and \( L_n^0 = M_b + TM_f + M \Delta t \). This gives
\[ \frac{L_n^0 - \partial u_x}{L_n^0 \partial u_x} (M - M_0) \Delta t \]
(9.41)
To simplify the analysis we note that \( \frac{L_n^0 - \partial u_x}{L_n^0 \partial u_x} - \frac{2\Delta t Y^2}{\partial_m} (1 + \lambda_1^2) \geq 0 \) is a sufficient condition for the positivity of the last term on the left hand side of (9.39). Inequality (9.41) implies that this is satisfied when
\[ \left( 1 - \frac{2Y^2}{\partial_m} (1 + \lambda_1^2) (M_b + TM_f) \Delta t \right) M \geq M_0 + \frac{2Y^2}{\partial_m} (1 + \lambda_1^2) (M_b + TM_f)^2. \]
Hence by defining
\[ \Delta t^* = \frac{\partial_m}{4Y^2(1 + \lambda_1^2)(M_b + TM_f)} \]
and \( G = \frac{4Y^2}{\partial_m} (1 + \lambda_1^2)(M_b + TM_f)^2, \)
we get that for all \( \Delta t \leq \Delta t^* \) and \( M \geq M_1, \)
\[ \frac{L_n^0 - \partial u_x}{L_n^0 \partial u_x} - \frac{2\Delta t Y^2}{\partial_m} (1 + \lambda_1^2) \geq 0. \]
Now consider \( \Delta t \leq \Delta t^* \) and \( M \geq M_1 \). Inequality (9.39) is restated as:
\[ \int_\Omega L_n^1 |e_n|^2 + \Delta t \partial_m \| \nabla e_n \|^2 + \int_\Omega \partial_u z(z_0) |e_n|^{l-2} + \int_\Omega L_n^1 |e_n|^{l-1} |e_n-1|^2 + \int_\Omega |L_n^1 - L_{n-1}^1 |e_n-1|^2. \]
(9.42)
Observe that \( \Omega \) can be split into two disjoint sets defined as \( \Omega_1, \Omega_2 \) such that \( \Omega_1 = \{ \bar{x} \in \Omega : |\partial u z(z)| < \frac{1}{2} M \Delta t \} \) and \( \Omega_2 = \{ \bar{x} \in \Omega : |\partial u z(z)| \geq \frac{1}{2} M \Delta t \} \). If \( \bar{x} \in \Omega_1 \), then from \( M \geq M_1 \geq 4M_0 \) we get
\[ \partial_u z(u_n^{l-2}) \leq \partial_u z(z_0) + \max_{\mathcal{I}(u_n^{l-1}, u_n^{l-2})} |\partial u z| |z_0 - u_n| + \max_{\mathcal{I}(u_n^{l-1}, u_n^{l-2})} |\partial u z| |u_n - u_n^{l-2}| \]
\[ \leq \partial_u z(z_0) + M_0 \Delta t + M \Delta t \leq \frac{1}{2} M \Delta t + \frac{1}{2} M \Delta t \]
which implies that \( L_{n}^{i-1} = \max(\partial_u z(u_n^{i-2}) + \mathcal{M} \Delta t, 2\mathcal{M} \Delta t) = 2\mathcal{M} \Delta t \). By the same logic, as \(|\partial_u z(\zeta_0)| < \frac{1}{2} \mathcal{M} \Delta t \), one has \( L_{n}^{0} = 2\mathcal{M} \Delta t \). This means that

\[
\int_{\Omega} |L_{n}^{i} - L_{n}^{i-1}| \|e_{n}^{i-1}\|^2 = 0.
\]

If \( x \in \Omega_2 \) then \( \partial_u z(\zeta_0) \geq \frac{1}{2} \mathcal{M} \Delta t \geq 2\mathcal{M} \Delta t \). There can be four cases. If \( \partial_u z(u_{n}^{i-1}) \leq \mathcal{M} \Delta t \) and \( \partial_u z(u_{n}^{i-2}) > \mathcal{M} \Delta t \) then \( |L_{n}^{i} - L_{n}^{i-1}| = \partial_u z(u_{n}^{i-2}) - \mathcal{M} \Delta t \leq \partial_u z(u_{n}^{i-2}) - \partial_u z(u_{n}^{i-1}) \).

Similarly if \( \partial_u z(u_{n}^{i-1}) > \mathcal{M} \Delta t \) and \( \partial_u z(u_{n}^{i-2}) \leq \mathcal{M} \Delta t \) then \( |L_{n}^{i} - L_{n}^{i-1}| \leq \partial_u z(u_{n}^{i-1}) - \partial_u z(u_{n}^{i-2}) \).

Combining everything we can state that

\[
|L_{n}^{i} - L_{n}^{i-1}| \leq |\partial_u z(u_{n}^{i-1}) - \partial_u z(u_{n}^{i-2})| \leq \max_{\mathcal{F}(u_{n}^{i-1}, u_{n}^{i-2})} |\partial_u u| |u_{n}^{i-1} - u_{n}^{i-2}| \leq 2\mathcal{M} \Delta t.
\]

From the above, one obtains

\[
\int_{\Omega_2} |L_{n}^{i} - L_{n}^{i-1}| \|e_{n}^{i-1}\|^2 = \int_{\Omega_2} |L_{n}^{i} - L_{n}^{i-1}| \|e_{n}^{i-1}\|^2 \leq 2\mathcal{M} \Delta t \int_{\Omega_2} \|e_{n}^{i-1}\|^2.
\]

Using this result in (9.42) one gets

\[
\int_{\Omega} L_{n}^{i} \|e_{n}^{i}\|^2 + \Delta t \mathcal{D} m \|\nabla e_{n}^{i}\|^2 \leq \int_{\Omega} L_{n}^{i-1} \|e_{n}^{i-1}\|^2.
\] (9.43)

Taking the sum of (9.43) from \( i = 1 \) to \( i = p \) gives

\[
\int_{\Omega} L_{n}^{p} \|e_{n}^{0}\|^2 + \Delta t \mathcal{D} m \sum_{i=1}^{p} \|\nabla e_{n}^{i}\|^2 \leq \int_{\Omega} L_{n}^{0} \|e_{n}^{0}\|^2 \leq \|e_{n}^{0}\|_{L^{1}(\Omega)}^2 \Delta t^2.
\] (9.44)

with \( L_{n}^{0} = \max(\partial_u z(x, u_{n-1}) + \mathcal{M} \Delta t, 2\mathcal{M} \Delta t) \). Similar estimates are given in [123, 160, 195]. In other words, the series \( \sum_{i=1}^{\infty} \|\nabla e_{n}^{i}\|^2 \) is convergent implying that \( \|\nabla e_{n}^{i}\| \to 0 \) as \( i \to \infty \). This concludes the proof of the first part of Theorem 9.2.

(b) To prove (b) we rearrange (9.38) as \( T_{j} + \Delta t T_{j} = T_{3} + \Delta t T_{5} + \Delta t T_{6} \), where \( T_{j}, j \in \{2, 5, 6\} \) have been defined before, and \( T_{1}, T_{3} \) are

\[
T_{1} := \int_{\Omega} L_{n}^{i} \|e_{n}^{i}\|^2 \geq (m + \mathcal{M} \Delta t) \|e_{n}^{i}\|^2;
\]

\[
T_{3} := \|(L_{n}^{i} - \partial_u z(\zeta))e_{n}^{i-1}, e_{n}^{i}\) \leq \mathcal{M} \Delta t \|e_{n}^{i}\|^2 + \mathcal{M} \Delta t \|e_{n}^{i-1}\|^2;
\]
the last inequality following from (9.18) and Young’s inequality. Moreover, a different estimate for $T_5$, $T_6$ can be obtained as

$$T_5 \leq \frac{1}{\mathcal{D}_m} \| (\mathcal{D}^{-1}_n - \mathcal{D}_n) \nabla u_n \|^2 + \frac{\mathcal{D}_m}{4} \| \nabla e_n \|^2 \leq \frac{\Lambda^2}{\mathcal{D}_m} \| e^{-1}_n \|^2 + \frac{\mathcal{D}_m}{4} \| \nabla e_n \|^2;$$

$$T_6 \leq \frac{1}{\mathcal{D}_m} \| \tilde{F}_n^{-1} - \tilde{F}_n \|^2 + \frac{\mathcal{D}_m}{4} \| \nabla e_n \|^2 \leq \frac{dM_F^2}{\mathcal{D}_m} \| e^{-1}_n \|^2 + \frac{\mathcal{D}_m}{4} \| \nabla e_n \|^2.$$

Adding in the estimates for all remaining terms, one gets from (9.38),

$$m \| e_n \|^2 + \frac{\Delta t}{2} \mathcal{D}_m \| \nabla e_n \|^2 \leq \Delta t \left[ \frac{\mathbb{M} + (dM_F^2 + \mathcal{D}_M^2 \Lambda^2_1)}{\mathcal{D}_m} \right] \| e^{-1}_n \|^2,$$

which rewrites as

$$\| e_n \|^2 + \frac{\Delta t \mathcal{D}_m}{2m} \| \nabla e_n \|^2 \leq \frac{\Delta t}{m} \left[ \frac{\mathbb{M} + (dM_F^2 + \mathcal{D}_M^2 \Lambda^2_1)}{\mathcal{D}_m} \right] \left( \| e^{-1}_n \|^2 + \frac{\Delta t \mathcal{D}_m}{2m} \| \nabla e^{-1}_n \|^2 \right).$$

Taking

$$\Delta t = \frac{m \mathcal{D}_m}{\mathbb{M} \mathcal{D}_m + (dM_F^2 + \mathcal{D}_M^2 \Lambda^2_1)} \text{ and } \alpha = \sqrt{\frac{\Delta t}{m} \left[ \frac{\mathbb{M} + (dM_F^2 + \mathcal{D}_M^2 \Lambda^2_1)}{\mathcal{D}_m} \right]},$$

(9.45)

one observes that the iteration converges in the equivalent $H^1(\Omega)$-norm $\| u \|_{H^1(\Omega)} = \sqrt{\| u \|^2 + \frac{\Delta t \mathcal{D}_m}{2m} \| \nabla u \|^2}$ and has the convergence rate $\alpha = \theta(\sqrt{\Delta t}) < 1$ if $\Delta t < \Delta \tilde{t}$. □

### 9.5 Numerical results

For the numerical discussions we consider the Richards equation, which is widely used in groundwater modelling. In terms of the capillary pressure $p$, the non-dimensional Richards equation for homogeneous isotropic media reads:

$$\partial_t S(p) = \nabla \cdot \{ k(S(p)) (\nabla p - \hat{g}) \} + R_w,$$

(9.46)

where $\hat{g}$ is the unit vector along the direction of gravity and $R_w$ is the source term, see (Ri) in Chapter 1. Richards equation involves nonlinearities in all the terms. The saturation function $S$ is increasing and the permeability function $k$ takes non-negative values. Without entering into the details, as the specific forms will be given later, we mention that in general one has $k(0) = 0$ and $S'(p) \to 0$ as $p \to -\infty$. Further $S'(p) = 0$ for all $p \geq 0$. However, if the flow does not become completely unsaturated, meaning that $S(p) \to 0$ (this being the case when the initial and boundary conditions
are taken accordingly), Assumption (P3) will be satisfied. Also (9.37) is satisfied as discussed in Remark 9.5.

The theoretical results presented in the previous sections do not depend upon the spatial discretization. Hence, for the numerical results, different methods like finite difference, finite element or finite volume, can be used to discretize (9.46) in space. Here we have used a two point flux approximation finite volume scheme \[91\] defined on two dimensional triangular unstructured grids. We take

\[
\Omega = (0, 1) \times (0, 1) \quad \text{and} \quad T = 1, \tag{9.47}
\]

and use FVCA8 benchmark meshes of different sizes. For the triangulation \( \mathcal{T} \) the mesh size is \( h = \sup\{\text{diam}(T) : T \in \mathcal{T}\} \).

With \( S \) and \( k \) defined as

\[
S(p) = \begin{cases} 
\frac{1}{(1-p)^3} & \text{if } p < 0 \\
1 & \text{if } p \geq 0 
\end{cases}, \quad k(S) = S^3, \tag{9.48}
\]

the first numerical example is constructed in such a way that

\[
\tilde{p}(x, y, t) = 1 - (1 + t^2)(1 + x^2 + y^2), \tag{9.49}
\]

is the exact solution (see also [226]). The source term is

\[
R_w(x, y, t) = \frac{2(1 - y^2)}{(1 + y^2)^2} - \frac{2t}{3\sqrt{1 + t^2}} - \frac{2x}{(1 + t^2)(1 + x^2 + y^2)^2}. \tag{9.50}
\]

The boundary and initial conditions are as in Table 9.1. \( \hat{g} \) points along the positive \( x \)-axis.

<table>
<thead>
<tr>
<th>IC</th>
<th>( t = 0 )</th>
<th>( p(x, y, 0) = \tilde{p}(x, y, 0) ) on ( \Omega )</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC</td>
<td>( x = 0 : )</td>
<td>( p(0, y, t) = \tilde{p}(0, y, t), \quad x = 1 : \quad p(1, y, t) = \tilde{p}(1, y, t), )</td>
</tr>
<tr>
<td></td>
<td>( y = 0 : )</td>
<td>( \partial_y p = 0, \quad y = 1 : \quad k(S(p))\partial_y p = k(S(\tilde{p}(x, 1, t))\partial_y \tilde{p}(x, 1, t). )</td>
</tr>
</tbody>
</table>

Figure 9.1 (left) shows the comparison of the analytical solution \( \tilde{p} \) with the numerical solution. The numerical solution is obtained from the MS with \( M = 1 \). The maximum relative error \( \| \frac{p - \tilde{p}}{p} \|_{l^\infty(\Omega)} \) is 1.38\% and the \( L^2(\Omega) \) error is 0.0116 which is of the order of the discretization errors, implying that the computational results are accurate. To ensure correctness, such kind of profile match has been conducted for all the results shown afterwards.
A linear iterative scheme for nonlinear diffusion problems

Figure 9.1: (left) Analytical solution vs. numerical solution for $S$, $k$ and $R_w$ given in (9.48), (9.50) and for initial and boundary condition stated in Table 9.1. The computational parameters are $\Omega = 1$, $\Delta t = 0.001$ and $h = 0.02$. (right) The error decay of the same computation at $t = 0.5$ for different schemes. Both $L^2$ error and $L^\infty$ error of the MS have been plotted and $L = 0.2$ has been used for the LS.

Figure 9.1 (right) shows how the errors decrease for the same computation for different schemes. Both $L^2$ and $L^\infty$ errors of the MS decrease monotonically. The decrease of the $L^\infty$ error shows that Assumption (A1) is valid and the linear profile of the $L^2$ and $L^\infty$ errors shows that the convergence is indeed linear as pointed out in Theorem 9.2 and Proposition 9.2. Observe that the error plot of the PS almost coincides with the error plot of the MS. This is because $\Delta t = 0.001$ is quite small, and for reasons explained afterwards this is true in general for small values of $\tau$. The NS converges faster than the PS and the MS. The error of the LS decreases linearly but the speed of convergence is considerably less than the other schemes shown. The values of $\Omega$ and $L$ used are the optimal values that give fastest convergence in their respective cases. This is explained in detail later.

To illustrate the strength of the MS compared to the other schemes mentioned above, we increase the complexity of the problem by using relations used for real-life simulations of such processes. For this purpose we replace the expressions in (9.48) with the standard van Genuchten relationship (1.7): for $\ell > 1$ and $q = 1 - \frac{1}{\ell}$,

$$S(p) = \begin{cases} \frac{1}{(1+(-p)^\ell)^{1/q}}, & \text{if } p < 0 \\ 1, & \text{if } p \geq 0 \end{cases}, \quad k(S) = \sqrt{S}(1 - S^{1/q})^2. \quad (9.51)$$

In the computations, $\ell = 3$ has been used throughout. Other conditions and definitions remain as in Table 9.1. The problem is degenerate as inside the domain there are regions where $p > 0$ or $S'(p) = 0$. For the numerical results $\Omega = 10$ has been used for the MS and $L = 0.4$ has been used for the LS, unless specified otherwise. These values give optimal convergence for both the schemes. The choice is motivated later.
First a mesh study is conducted where the time step is fixed and the mesh size \( h \) is varied from \( h = 0.1 \) to \( h = 0.02 \). The results are shown in Figure 9.2 for two time step sizes \( \Delta t = 0.01 \) and \( \Delta t = 0.001 \) for a fixed time \( t = 0.5 \).

![Figure 9.2](image.png)

Figure 9.2: Mesh study of the NS, the PS, the LS and the MS at \( t = 0.5 \). (left) Results for \( \Delta t = 0.01 \); (right) results for \( \Delta t = 0.001 \). The mesh sizes used are \( h = 0.1, 0.05, 0.02 \). The \( L = 1 \) curve on the left figure corresponds to the error of the LS for \( L = 1 \) and \( h = 0.05 \). In all other computations \( \mathcal{M} = 10 \) and \( L = 0.4 \) and the PS corresponds to \( \mathcal{M} = 0 \).

It follows from Figure 9.2 that the MS, the LS and the PS show linear convergence. The MS is faster than the PS for all mesh sizes for \( \Delta t = 0.01 \) (Figure 9.2 (left)). The convergence rate of the LS is higher than that for other schemes and in consequence, LS converges slower. For \( h = 0.05 \) and \( h = 0.02 \) no further decrease in error is visible after a few iterations in case of the NS. The convergence rates for all the schemes vary with the mesh size. The dependence of the convergence rates of the MS and the LS on mesh size stems from the fact that \( \mathcal{M} \) or \( L \) chosen in these cases is not the \( \mathcal{M}_0 \) or Lipschitz constant value (which in this case is \( L \approx 0.65 \)) but the values that give the optimal convergence properties. This is also seen in Figure 9.6. In fact, taking \( L = 1 \) for the LS results in all the error plots for different mesh sizes being on top of each other. This error characteristic is presented by the curve labelled \( L = 1 \). Also, it appears that for a fixed time step size, decreasing \( h \) makes the convergence slower. This will be explained in detail later.

Figure 9.2 (right) shows the results for \( \Delta t = 0.001 \). The convergence is much faster than the \( \Delta t = 0.01 \) case for the MS as it was shown in Theorem 9.2 that \( \alpha = O(\sqrt{\Delta t}) \). The difference between the convergence rates of the PS and of the MS is very small. This is because \( \Delta t << 1 \) and \( \mathcal{M} = 10 \) so that \( \mathcal{M} \Delta t << S'(p) \sim O(1) \) in most of the domain \( \Omega \). So difference between the MS and the PS is small. Figure 9.2 (right) also shows that convergence rate of the LS is stable with respect to the mesh size and is not impacted greatly by the change in time step size. This is probably due to Condition
16 of [155] being satisfied for this value of $\Delta t = 0.001$. However, the convergence is slower when compared to the other schemes. As for the NS, for $h = 0.1$ it is marginally faster than the PS and the MS. But for finer meshes, NS becomes slower than the PS or the MS.

The quadratic convergence of the NS is not observed from Figure 9.2. This might be due to the small time step sizes required for the NS to show quadratic convergence in degenerate cases (see [206] and Table 9.2). It could also be attributed to the errors of the linear solver as it is known from literature [130, 155, 226] that the stiffness matrices for the NS are relatively ill-conditioned. Indeed, the GMRES solver gives higher residual errors for the NS compared to other methods. The errors introduced in computing the Jacobian could possibly be another reason that causes this deviation.

One of the advantages of the MS is its stability and for this reason we must look at larger time step sizes. Figure 9.3 shows the results for $\Delta t = 0.1$. The PS converges, albeit much slower than the MS, for $h = 0.1$. For $h = 0.05$ the convergence is very slow for the PS and the errors are not decreasing monotonically. In fact for $h = 0.05$ the iterations fail to converge starting from $t = 0.7$. For $h = 0.02$ the PS fails to converge even at $t = 0.5$. Similar behaviour is observed for the NS. For $h = 0.05$ the errors start to increase after a few iterations and the solution diverges at $t = 0.8$. The reason for this behaviour of the PS and the NS is the bound given in (9.2). As the mesh size decreases the time step size has to be reduced in order to guarantee the convergence for the PS and the NS [206]. However this constraint is not there for the MS. Although there is a threshold on the time step size for the MS’s convergence, this threshold does not depend on the mesh size and so in all the cases shown in
Figure 9.3, the errors for the MS are decreasing and the convergence is faster when compared to the NS or the PS. This shows that for numerical computations the MS is more stable than the NS or the PS in general.

As the convergence rate of the LS is quite stable with respect to mesh sizes for $L = 1$, the error behaviour of the LS has been plotted only for $h = 0.05$ and $L = 1$ in Figure 9.3. LS also has an upper bound on the time steps for linear convergence that does not depend upon the spatial discretization [155]. An interesting observation is that for larger time step sizes the convergence of the LS can actually be faster than the convergence of the MS. This is due to the fact that in the computations for $\Delta t = 0.1$, $\frac{\Delta t}{L} \sim L$ and so $\frac{\Delta t}{L} + \delta'(p)$ can be greater than $L$. So the apparent $L$ value for the MS can be larger than the value of $L$ required to be imposed on the LS. This might make the convergence of the MS slower than of the LS.

Figure 9.4 shows how the convergence rate $\alpha$, calculated here as the geometric average of $\frac{\|p_{n+1}^i - p_n^i\|}{\|p_n^i - p_{n-1}^i\|}$ over the first 10 iterations, changes with $\Delta t$. From Theorem 9.2 one expects that for small enough $\Delta t$ the convergence rate should scale with $\sqrt{\Delta t}$ which implies that the slope of the solid line in Figure 9.4 should nearly be $\frac{1}{2}$ for small $\Delta t$ values. This is indeed the case. For $\Delta t \leq 0.01$ the line is almost parallel to the reference dashed line representing a slope of $\frac{1}{2}$. It is to be noted though that Theorem 9.2 was proved for the non-degenerate case whereas this test case is degenerate.

![Figure 9.4: Convergence rate ($\alpha$) vs. time step size ($\Delta t$) for $h = 0.05$ at $t = 0.5$. The convergence rates are calculated as the geometric mean of $\|p_{n+1}^i - p_n^i\|/\|p_n^i - p_{n-1}^i\|$ over first 10 iterations. The dotted black line shows the slope of 0.5.](image)

Next we study the effect of the choice of $\Omega$ on the rate of convergence. One expects from Theorem 9.2 and (9.45) that for small values of $\Omega$ the condition $L_n^i -
A linear iterative scheme for nonlinear diffusion problems

∂u/∂t > 0 will not be satisfied whereas large values of \( M \) would result in a slower convergence. This behaviour is observed precisely if one varies \( L \) in the case of the LS [155, 226]. From Figure 9.5 we see that it describes the MS as well. For both time step sizes \( \Delta t = 0.01 \) and \( \Delta t = 0.1 \) we get that the optimal \( M \) is close to 10. If one chooses \( M \) value away from the optimal \( M \) then \( \alpha \) increases. As \( \Delta t \to 0 \) the rates tend towards the rate of the PS. In the \( \Delta t = 0.01 \) case the effect of an optimal \( M \) is less pronounced than in the \( \Delta t = 0.1 \) case and from Figure 9.2 (right) one can speculate that it is even less important in the \( \tau = 0.001 \) case.

This a-priori knowledge can be useful when developing the linearisation scheme as one can decide what \( M \) to choose by running the computation for only the first time step with a coarser grid. In our case, we chose \( M = 10 \) for the results presented earlier for this reason.

Finally, we present a comparison study with Example 1 of [155]. The functional relationships are taken exactly as stated in [155] but different meshes and linear solvers are used. The initial condition used in this example is

\[
p(x, y, 0) = \begin{cases} 
  p_{\text{vad}} & \text{in } \Omega \cap \{ y < \frac{1}{4} \}, \\
  -y + \frac{1}{4} & \text{in } \Omega \cap \{ y > \frac{1}{4} \}.
\end{cases}
\]  \hspace{1cm} (9.52)

Here \( p_{\text{vad}} < 0 \) is a constant. The gravity points towards negative \( y \) direction and the results are for the first time step, i.e. \( n = 1 \). As the initial condition is discontinuous, the NS fails to converge in all cases in our computations. However, the NS converged in most cases if the discontinuity was regularised, e.g. by considering a linear interpolation over a small interval of length 0.1. After this step the results obtained
match closely the values given in [155]. In particular, we partly reproduce [155, Figure 1] with $p_{vad} = -2$ and [155, Table 1] with $p_{vad} = -3$ in Figure 9.6 and Table 9.2 respectively.

![Figure 9.6: Iterations required by different schemes to have $\|p_h^n - p_h^{n-1}\| < 10^{-5}$, for a fixed $\Delta t = 1$ and $p_{vad} = -2$. The parameters are taken from [155, Figure 1].](image)

<table>
<thead>
<tr>
<th>$\Delta t$</th>
<th>MS $\mathcal{M} = 0.01$</th>
<th>PS $\mathcal{M} = 0$</th>
<th>NS</th>
<th>LS $L = 0.25$</th>
<th>LS $L = 0.15$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>19</td>
<td>-</td>
<td>54</td>
<td>35</td>
</tr>
<tr>
<td>0.1</td>
<td>22</td>
<td>22</td>
<td>-</td>
<td>50</td>
<td>33</td>
</tr>
<tr>
<td>0.01</td>
<td>12</td>
<td>12</td>
<td>-</td>
<td>39</td>
<td>26</td>
</tr>
<tr>
<td>0.001</td>
<td>7</td>
<td>7</td>
<td>-</td>
<td>154</td>
<td>99</td>
</tr>
</tbody>
</table>

Table 9.2: Iterations required by different schemes to have $\|p_h^n - p_h^{n-1}\| < 10^{-5}$, for a fixed $h = 1/10$ and $p_{vad} = -2$. The parameters are taken from [155, Table 1].

The result shows that the NS is the fastest when it converges but it requires smoother initial conditions and smaller time step sizes to converge. The PS and the MS have comparable stability properties. The MS, as before, is at least as fast as the PS in all cases. The LS is relatively slower and taking a smaller value of $L$ ($L = 0.15$) compared to the Lipschitz constant ($L = 0.25$) speeds up the convergence, but makes the convergence rate more susceptible to changes in mesh size. This was seen in Figure 9.2 (left) too. The convergence rate decreases with the time step size for all the schemes except for the LS. A drastic increase in number of iterations required is seen for the LS at $\Delta t = 0.001$, see Table 9.2. This is explained by the fact that the convergence order is $L/(L + C\Delta t)$, which approaches 1 when $\Delta t$ goes to 0. $\mathcal{M} = 0.01$ is used for the MS in Figure 9.6 and Table 9.2 as it gives the optimal convergence rate.

### 9.6 Conclusion

In this paper we propose a linearisation scheme for a general class of nonlinear parabolic partial differential equations. We have given a rigorous convergence proof of the scheme and compared its behaviour with that of other linearisation schemes: the Newton scheme (NS), the modified Picard scheme (PS) or the L-scheme (LS), are generally used to solve the sequence of elliptic equations obtained from time-discretization of such problems. The NS and the PS have the drawback that they converge only if the initial guess for the iterations is close enough to the solution of
the elliptic problem. For the concerned sequence of elliptic equations, this leads to a severe restriction on time step size. On the other hand, the LS converges irrespective of the initial guess but is much slower than the schemes mentioned above. To resolve these issues, a combination of the LS and the PS, termed modified L-scheme (MS) in this context, is proposed.

The MS uses local estimates and the solution of the previous time step as the initial guess to improve the convergence behaviour of the LS. This is shown first for quasilinear equations that have linear diffusivity and advection terms. It is proved that the scheme converges linearly irrespective of the spatial discretization and for any time step, even in degenerate cases. Moreover, for small time step sizes, the linear convergence rate is proportional to the time step size, implying that the scheme converges faster as the time step size is made smaller.

Next, this result is generalised to the case when the diffusivity and the advection terms are nonlinear. It is proved that if the time step size is smaller than an upper-bound which is independent of the spatial discretization, the scheme converges even for degenerate cases. Linear convergence is achieved in the non-degenerate case with a convergence rate that is proportional to the square root of the time step size for sufficiently small time steps.

Finally, numerical results are presented for the Richards equation for all the schemes mentioned above. It is seen that the MS is faster than the PS and the LS. Moreover, the MS is more stable than the NS or the PS in the sense that the MS converges for larger time steps. Numerically it is observed that the MS indeed gives convergence rates proportional to the square root of the time step size. The final numerical results imply that the convergence rate of the MS can be controlled by tuning the parameter $\mathcal{M}$. 


Chapter 10

Mixed finite element discretization for the two-phase flow model with dynamic capillarity: error estimates

10.1 Introduction

In this chapter we consider the following model which is used to describe two-phase flow through porous media incorporating dynamic capillarity effects (112):

\[ \partial_t s - \nabla \cdot (k_\alpha(s) \nabla \bar{p}) = 0, \]  
\[ -\partial_t s - \nabla \cdot (k_W(s) \nabla p) = 0, \]  
\[ \bar{p} - p = p_c(s) + \tau \partial_t s, \]

complemented with suitable initial and boundary conditions. The equations hold in \( Q := (0, T] \times \Omega \). Here \( \Omega \) is a bounded domain in \( \mathbb{R}^d \) (\( d \geq 1 \)), having Lipschitz continuous boundary, and \( T > 0 \) is a given maximal time. The unknowns \( s, \bar{p} \) and \( p \) denote the non-wetting saturation, the non-wetting phase and the wetting phase pressures respectively. For simplicity, here we have assumed homogeneity and isotropy of the medium while neglecting gravity effects and source/sink terms. This does not significantly compromise the generality of the results, see Remark 10.3. Under these

---

assumptions, the two-phase equations (TP) described in (1.13) reduce to the equations (10.1a)-(10.1b). In (10.1), instead of the subscript \( n \) denoting the non-wetting phase like in previous chapters, an alternative notation is used, since the subscript \( n \) denotes time step here. Moreover, the non-wetting phase saturation \( s \) is used as the primary variable, instead of the water saturation \( 1 - s \), to use the positivity of certain functions. For this reason, capillary pressure \( p_c \), see (1.4), is reformulated as a function of \( s \), instead of the water saturation \( 1 - s \). For a detailed description of the dynamic capillary effect, represented by the capillary pressure relation (10.1c), we refer to Section 1.2.2.

In order to close the System (10.1), we prescribe the initial and boundary conditions

\[
\begin{align*}
s(0, \cdot) &= s^0, \quad \text{in } \Omega, \\
\bar{p} &= p = 0, \quad \text{at } \partial \Omega \text{ for } t > 0,
\end{align*}
\]

where \( s^0 \) is a given function, which will be specified later. The volumetric fluxes \( \bar{q} \) and \( q \) are deduced from the Darcy law (see [115]) as

\[
\begin{align*}
\bar{q} &= -k_o(s)\nabla \bar{p}, \\
q &= -k_w(s)\nabla p.
\end{align*}
\]

If the dynamic capillarity contribution is neglected, i.e. \( \tau = 0 \) in (10.1c), the algebraic relationship between phase pressures and saturation is written as

\[
\bar{p} - p = p_c(s),
\]

which is used in the standard model discussed in Section 1.1. Numerical methods for the standard model have been the subject of extensive research in the last decades. We refer to [12, 41, 46, 60, 77, 82, 86, 179, 202, 205, 208, 268], where finite element method, mixed finite element method, discontinuous Galerkin method are analysed. The linear iterative schemes e.g. Newton method, Picard and L-scheme are investigated for the Richards equation in [155] and Chapter 9. The authors present a mass conservative numerical scheme for two-phase flow in porous media including Hölder continuous nonlinearities in [202]. In [50, 93, 162], the authors analyse finite volume methods. The major challenge in developing efficient numerical schemes is related to the degenerate nature of the problem. Due to this, the solution typically lacks regularity, which makes lower order finite elements or finite volumes a natural choice for the spatial discretization. In all cases, the convergence of the numerical schemes is proved rigorously either by compactness arguments, or by obtaining a-priori error estimates. A-posteriori error estimates are obtained e.g. in [50].

Under the assumption that the total flow is known, the two-phase model can be reduced into a scalar model, see (1.18). The existence and uniqueness of a weak solution for the model including dynamic capillarity \((\tau > 0)\), is obtained in
Travelling wave analysis has been given in [248] and Chapters 4 to 6. Numerical schemes for two-phase flow through heterogeneous media are discussed in [116] for cases without an entry pressure. For situations including an entry pressure, variational inequality approaches have been considered in [117]. Further, we refer to [246] for coupling conditions between heterogeneous blocks under the dynamic effect. In [116, 187], the authors consider numerical algorithms for unsaturated flow through highly heterogeneous media with dynamic effect. For the full two-phase flow model, the existence and uniqueness of the weak solutions are proved in [53, 141], with the assumption that the equations are non-degenerate (i.e. all nonlinearities are bounded away from 0 or +∞). For this case, a finite volume-finite element method is proposed in [85], a multipoint flux approximation finite volume method is presented in [52] and a discontinuous Galerkin scheme is proposed in [135, 136]. In addition, numerical investigations for heterogeneous media have been given in [96]. For the degenerate case, we refer to [55], which proves the existence of weak solutions for the model, using an equivalent formulation.

In this chapter, we present a fully discrete numerical scheme for the model describing two-phase immiscible flow in porous media with dynamic effects in the capillary pressure. For the time discretization, the Euler implicit method is applied. The spatial discretization is based on the mixed finite element method (MFEM). Specifically, the lowest order Raviart-Thomas elements are applied. We show the existence and uniqueness of the numerical solutions. The error estimates for the saturation, fluxes and phase pressures are derived for the temporal and spatial discretizations to show the convergence of the scheme.

The rest of the paper is organized as follows. In Section Section 10.2, we present the notations and assumptions on the data and the definition of the weak solution. We introduce the mixed formulation and give error estimates for the saturation, phase pressures and fluxes for the mixed finite element scheme in Section 10.3. In the last section, we present some numerical results that confirm the theoretical findings.

## 10.2 Notations and assumptions

In what follows, we use the standard notations of functional analysis and the theory of partial differential equations. Throughout this paper, we assume the system is defined in a bounded connected domain $\Omega \subset \mathbb{R}^d(d \geq 1)$. For simplicity, one can assume that $\Omega$ is polygonal. Furthermore, by $(\cdot, \cdot)$, we denote the inner product on $L^2(\Omega)$, and let $\| \cdot \|$, $\| \cdot \|_1$ stand for the norms in $L^2(\Omega)$ and $W^{1,2}(\Omega)$, respectively. The functions in $H(\text{div}; \Omega)$ are vector valued, having a $L^2$ divergence. By $C$ we mean a generic positive constant, not depending on the unknowns or the discretization parameters.

For the spatial discretization, let $\mathcal{T}_h$ be a regular decomposition of $\Omega \subset \mathbb{R}^d$ into closed $d$–simplices; $h$ denoting the mesh diameter of the decomposition. Here we
assume $\tilde{\Omega} = \cup_{K \in \mathcal{T}_h} \tilde{K}$. Correspondingly, the discrete subspaces $W_h \subset L^2(\Omega)$ and $V_h \subset H(\text{div}; \Omega)$ are defined as:

$$W_h := \{ p \in L^2(\Omega) \mid p \text{ is constant on each element } T \in \mathcal{T}_h \},$$

$$V_h := \{ q \in H(\text{div}; \Omega) \mid q|_K(x) = a + b x, a \in \mathbb{R}^d, b \in \mathbb{R} \text{ for all } K \in \mathcal{T}_h \}.$$ 

In the following, we use the usual $L^2$ projector:

$$P_h : L^2(\Omega) \to W_h \text{ such that } (P_h w - w, w_h) = 0,$$  

(10.5)

for all $w_h \in W_h$. Furthermore, on $(W^{1,2}(\Omega))^d$ a projector $\Pi_h$ can be defined such that

$$\Pi_h : (W^{1,2}(\Omega))^d \to V_h, (\text{div}(\Pi_h v - v), w_h) = 0,$$  

(10.6)

for all $w_h \in W_h$. Following [200], page 237, this operator can be extended to $H(\text{div}; \Omega)$. For the above operators, there holds

$$\| w - P_h w \| \leq C h \| w \|_1,$$  

(10.7)

$$\| v - \Pi_h v \| \leq C h \| v \|_1,$$  

(10.8)

for any $w \in W^{1,2}(\Omega)$ and $v \in (W^{1,2}(\Omega))^d$.

For the time discretization, let $N \in \mathbb{N}$ be strictly positive. Then we define the time step $\Delta t = T/N$, as well as the time stamps $t_n = n \Delta t, (n \in \{1,2,\ldots,N\})$.

Throughout this paper, we use the following assumptions:

(A1) $\Omega$ is an open, bounded and connected, convex polygonal domain in $\mathbb{R}^d$ ($d \geq 1$) with Lipschitz continuous boundary $\partial \Omega$. $\bar{\Omega}$ denotes the closure of $\Omega$.

(A2) The functions $k_o(\cdot), k_w(\cdot): \mathbb{R} \to \mathbb{R}$ are $C^1$ and Lipschitz continuous. There exist $0 < \delta_0, \delta_w, M_o, M_w < \infty$ such that $\delta_o \leq k_o(v) \leq M_o$ and $\delta_w \leq k_w(v) \leq M_w$ for all $v \in \mathbb{R}$. We assume $k_o(\cdot)$ to be a non-decreasing function with $k_o(v) = \delta_o$ for $v \leq 0$ and $k_o(v) = M_o$ for $v \geq 1$. On the other hand, $k_w(\cdot)$ is non-increasing with $k_w(v) = M_w$ for $v \leq 0$ and $k_w(v) = \delta_w$ for $v \geq 1$.

(A3) $p_c(\cdot): \mathbb{R} \to \mathbb{R}$ is an increasing function and $p_c \in C^1$. There exist $m_p, M_p > 0$ such that $m_p \leq p_c'(\cdot) \leq M_p < \infty$.

(A4) $\tau > 0$ is a positive constant.

(A5) The initial condition $s^0$ is in $W^{1,2}_0(\Omega)$ and satisfies $0 \leq s^0 \leq 1$ a.e.

(A6) The wetting and non-wetting fluxes $\tilde{q}$ and $q$ are bounded in $L^\infty((0,T) \times \Omega)$, $\| \tilde{q} \|_{L^\infty((0,T) \times \Omega)} + \| q \|_{L^\infty((0,T) \times \Omega)} \leq C$.

Remark 10.1. Here the assumption (A6) is reasonable if $\Omega$ is a $C^1$ domain with $0 < \gamma \leq 1$ and if $s_0 \in C^{0,\gamma}(\Omega)$. The justification can be found in [54, 55].
In the following, we define a weak solution of System (10.1).

**Problem P** Find \((s, p, \bar{p}) \in W^{1,2}(0, T; L^2(\Omega)) \times L^2(0, T; W^{1,2}_0(\Omega)) \times L^2(0, T; W^{1,2}_0(\Omega))\), with \(s|_{t=0} = s^0\) such that for a.e. \(t \in [0, T]\), one has

\[
(\partial_t s, \phi) + (k_0(s)\nabla \bar{p}, \nabla \phi) = 0, \quad (10.9a)
\]

\[
- (\partial_t s, \psi) + (k_{tw}(s)\nabla p, \nabla \psi) = 0, \quad (10.9b)
\]

\[
(\bar{p} - p, \lambda) = (p_c(s), \lambda) + \tau (\partial_t s, \lambda), \quad (10.9c)
\]

for all \(\phi \in W^{1,2}_0(\Omega), \psi \in W^{1,2}_0(\Omega)\) and \(\lambda \in L^2(\Omega)\).

**Remark 10.2.** In [149], the authors obtain higher regularity for the solution: \(\partial_t s \in L^2(0, T; L^2(\Omega))\), and \(\bar{p}, p \in W^{1,2}(0, T; W^{1,2}(\Omega))\). In [53], it is also showed that \(s \in W^{1,2}(0, T; W^{1,2}(\Omega))\).

**Remark 10.3 (Generality of the analysis).** For the sake of simplicity, we have assumed homogeneous Dirichlet conditions at the boundary and neglected gravity and source terms in (10.1a) and (10.1b) while analysing the problem in Section 10.3. However, the results obtained, hold for general Dirichlet, Neumann, Robin and mixed type boundary conditions and they are valid even if gravity terms and linear source terms are included. To illustrate this, the numerical results are given for more general boundary conditions and gravity and source terms are added (see (10.39)).

### 10.3 Mixed formulation and error estimates

#### 10.3.1 Existence and uniqueness

We now present a continuous mixed variational formulation for the model (10.1) - (10.3).

**Problem 1 (Continuous Variational Formulation).** Find \((s_e, q_e, \bar{p}_e, q_e, p_e) \in W^{1,2}(0, T; L^2(\Omega)) \times L^2(0, T; H(\text{div}; \Omega)) \times L^2(0, T; L^2(\Omega)) \times L^2(0, T; H(\text{div}; \Omega)) \times L^2(0, T; L^2(\Omega))\), with \(s_e|_{t=0} = s^0\) such that

\[
(\partial_t s_e, w) + (\nabla \cdot \tilde{q}_e, w) = 0, \quad (10.10a)
\]

\[
(k_0^{-1}(s_e)\tilde{q}_e, \nabla e) - (\bar{p}_e, \nabla e) = 0, \quad (10.10b)
\]

\[
-(\partial_t s_e, w) + (\nabla \cdot q_e, w) = 0, \quad (10.10c)
\]

\[
(k_{tw}^{-1}(s_e)\tilde{q}_e, \nabla e) - (p_e, \nabla e) = 0, \quad (10.10d)
\]

\[
(\bar{p}_e - p_e, w) = (p_c(s_e), w) + \tau (\partial_t s_e, w), \quad (10.10e)
\]

for all \(w \in L^2(\Omega)\) and \(\nabla e \in H(\text{div}; \Omega)\).

In the following, we discuss the existence and uniqueness of the continuous variational formulations for the system (10.10). Since the existence and uniqueness of the model (10.9) have been proved in [53–55, 163], we follow the approach in [95, 205, 207] and show the equivalence of Problem P and Problem 1.
Proposition 10.1. Let $s \in W^{1,2}(0, T; L^2(\Omega))$ and $\bar{p}, p \in L^2(0, T; W^{1,2}_0(\Omega))$ solve Problem P. Define $(s_e, \bar{p}_e, p_e) := (s, \bar{p}, p)$, $\bar{q}_e = -k_0(s_e)\nabla \bar{p}_e$ and $q_e = -k_w(s_e)\nabla p_e$. Then $(s_e, q_e, \bar{p}_e, q_e, p_e)$ solves Problem 1. Conversely, if $(s_e, q_e, \bar{p}_e, q_e, p_e)$ solves Problem 1, then $(s, \bar{p}, p) = (s_e, \bar{p}_e, p_e)$ is the solution of Problem P.

Proof. "⇒" By the definition of $(s_e, \bar{p}_e, p_e)$, it is easy to see that $(s_e, \bar{p}_e, p_e)$ has the regularity required for solving Problem P: $s_e \in W^{1,2}(0, T; L^2(\Omega))$, and $\bar{p}_e, p_e \in L^2(0, T; W^{1,2}_0(\Omega))$. So, we have $\partial_t s_e \in L^2(0, T; L^2(\Omega))$. Then using (10.9a) and by the definition of $\bar{q}_e$, one obtains

$$\langle \partial_t s_e, \phi \rangle - (\bar{q}_e, \nabla \phi) = 0,$$

for any $\phi \in W^{1,2}_0(\Omega)$, which implies that $\nabla \cdot \bar{q}_e = -\partial_t s_e$ in a distributional sense. The regularity of $\partial_t s_e$ immediately implies $\bar{q}_e \in H(\text{div}; \Omega)$ which is required for Problem 1. Similarly, we also have $q_e \in H(\text{div}; \Omega)$.

"⇐" Clearly, if $(s_e, q_e, \bar{p}_e, q_e, p_e)$ solves Problem 1, then we have $s_e \in W^{1,2}(0, T; L^2(\Omega))$ and $\bar{p}_e \in L^2(0, T; L^2(\Omega))$, which imply $s \in W^{1,2}(0, T; L^2(\Omega))$ and $\bar{p} \in L^2(0, T; L^2(\Omega))$. Taking $v \in (C^\infty_0(\Omega))^d \subset H(\text{div}; \Omega)$ in (10.10b), one has

$$(\nabla \bar{p}_e, v) = -(k_0^{-1}(s_e)\bar{q}_e, v), \quad (10.11)$$

which gives $\nabla \bar{p}_e = -k_0^{-1}(s_e)\bar{q}_e$ in a distributional sense. Further, since $s_e \in W^{1,2}(0, T; L^2(\Omega))$ and $\bar{q}_e \in L^2(0, T; H(\text{div}); \Omega)$, one obtains $\bar{p}_e \in L^2(0, T; W^{1,2}(\Omega))$. Next, we show that $\bar{p}_e$ has vanishing trace at the boundary of $\Omega$. Taking $v \in H(\text{div}; \Omega)$ and applying (10.10b) and (10.11), one has

$$(\nabla \bar{p}_e, v) = -(k_0^{-1}(s_e)\bar{q}_e, v) = -\bar{q}_e(\nabla \cdot v).$$

By using Green’s theorem for any $v \in H(\text{div}; \Omega)$, we have

$$\int_{\partial \Omega} \bar{p}_e \cdot \mathbf{n} d\gamma = \int_{\Omega} \nabla \cdot \bar{p}_e + \int_{\Omega} \bar{p}_e \nabla \cdot v = 0.$$ 

It follows that $\bar{p}_e \in L^2(0, T; W^{1,2}_0(\Omega))$, which implies $\bar{p} \in L^2(0, T; W^{1,2}_0(\Omega))$. By the same arguments, we have $p \in L^2(0, T; W^{1,2}_0(\Omega))$.

Similarly, taking $w \in W^{1,2}_0(\Omega) \subset L^2(\Omega)$ in (10.10a) and using (10.10b), one gets that $\bar{p}$ satisfies (10.9a). Finally, from (10.10e) it is directly obtained that $\bar{p}$ and $p$ satisfy (10.9c).

We now proceed to the time discretization for Problem 1, which is achieved by the backward Euler scheme. For a given $n \in \{1, 2, \ldots, N\}$, we define the time discrete mixed variational problem at time $t_n$. We will show the existence and uniqueness of the solution to the time discrete mixed variational problem. To do so, first we give the discrete forms of Problem P and Problem 1.
Problem $P^N$ Find $(s^n, \bar{p}^n, p^n) \in L^2(\Omega) \times W^{1,2}_0(\Omega) \times W^{1,2}_0(\Omega)$, with $s^{n-1} \in L^2(\Omega)$ such that

\begin{align}
\left( \frac{s^n - s^{n-1}}{\Delta t}, \phi \right) + (k_o(s^n)\nabla \bar{p}^n, \nabla \phi) &= 0, \\
- \left( \frac{s^n - s^{n-1}}{\Delta t}, \psi \right) + (k_w(s^n)\nabla p^n, \nabla \psi) &= 0,
\end{align}

(10.12a, 10.12b)

for all $\phi, \psi \in W^{1,2}_0(\Omega)$ and $\lambda \in L^2(\Omega)$.

Problem 2 (Semi-discrete Variational Formulation). Find $(s^n_e, q^n_e, \bar{p}^n_e, q^n_e, p^n_e) \in L^2(\Omega) \times H(\text{div};\Omega) \times H(\text{div};\Omega) \times L^2(\Omega)$, with $s^{n-1}_e \in L^2(\Omega)$ such that

\begin{align}
\left( \frac{s^n_e - s^{n-1}_e}{\Delta t}, w \right) + (\nabla \cdot q^n_e, w) &= 0, \\
(k_o^{-1}(s^n_e)q^n_e, \nabla \cdot w) - (\bar{p}^n_e, \nabla \cdot w) &= 0, \\
- \left( \frac{s^n_e - s^{n-1}_e}{\Delta t}, w \right) + (\nabla \cdot q^n_e, w) &= 0, \\
(k_w^{-1}(s^n_e)q^n_e, \nabla \cdot w) - (p^n_e, \nabla \cdot w) &= 0, \\
(\bar{p}^n_e - p^n_e, w) &= (p_c(s^n), w) + \tau \left( \frac{s^n - s^{n-1}}{\Delta t}, w \right),
\end{align}

(10.13a, 10.13b, 10.13c, 10.13d, 10.13e)

for all $w \in L^2(\Omega)$ and $v \in H(\text{div};\Omega)$.

Similar to Proposition 10.1 we have the following proposition:

**Proposition 10.2.** For a fixed $n \in \mathbb{N}, n \geq 1$ let $s^{n-1} \in L^2(\Omega)$ be given. If $(s^n, \bar{p}^n, p^n) \in L^2(\Omega) \times W^{1,2}_0(\Omega) \times W^{1,2}_0(\Omega)$ is a solution of Problem $P^N$, then a solution to Problem 2 is given by $s^n_e := s^n, \bar{p}^n_e := \bar{p}^n, q^n_e = -k_o(s^n)\nabla \bar{p}^n_e, p^n_e := p^n, q^n_e = -k_w(s^n)\nabla \bar{p}^n_e$. Conversely, if $(s^n_e, q^n_e, \bar{p}^n_e, q^n_e, p^n_e)$ solves Problem 2, then $(s^n, \bar{p}^n, p^n) := (s^n_e, q^n_e, \bar{p}^n_e, q^n_e, p^n_e)$ is a solution of Problem $P^N$.

**Proof.** The proof of Proposition 10.2 is similar to the proof of Proposition 10.1 (see also [205, 208]). We therefore omit it here.

From [53], and Proposition 10.2 the existence of a solution to Problem 2 is obtained. However, the question of uniqueness is left open. This is presented below.

**Lemma 10.1.** Let $n \in \mathbb{N}, n \geq 1$ be fixed. Assume that (A2)-(A4) and (A6) holds. For time step $\Delta t$ small enough, Problem 2 has at most one solution.
Proof. Assuming there are two solutions to Problem 2 denoted by \((s^n_{e1}, \tilde{q}^n_{e1}, \tilde{p}^n_{e1}, q^n_{e2}, p^n_{e2})\) and \((s^n_{e2}, \tilde{q}^n_{e2}, \tilde{p}^n_{e2}, q^n_{e1}, p^n_{e1})\) for a given \(s^{n-1}_{e} \in L^2(\Omega)\), one has

\[
\begin{align*}
(s^n_{e1} - s^n_{e2}, w) + \Delta t (\tilde{q}^n_{e1} - \tilde{q}^n_{e2}, w) &= 0, \\
(k^{-1}_o(s^n_{e1}) \tilde{q}^n_{e1} - k^{-1}_o(s^n_{e2}) \tilde{q}^n_{e2}, v) - (\tilde{p}^n_{e1} - \tilde{p}^n_{e2}, \nabla \cdot v) &= 0, \\
-(s^n_{e1} - s^n_{e2}, w) + \Delta t (\tilde{q}^n_{e1} - q^n_{e2}, w) &= 0, \\
(k^{-1}_w(s^n_{e1}) q^n_{e1} - k^{-1}_w(s^n_{e2}) q^n_{e2}, v) - (p^n_{e1} - p^n_{e2}, \nabla \cdot v) &= 0, \\
(\tilde{p}^n_{e1} - \tilde{p}^n_{e2} - p^n_{e1} + p^n_{e2}, w) &= (p_c(s^n_{e1}) - p_c(s^n_{e2}), w) + \frac{\Delta t}{\tau}(s^n_{e1} - s^n_{e2}, w) = 0,
\end{align*}
\]

for any \(w \in L^2(\Omega)\) and \(v \in H(\text{div;}\Omega)\).

First, we set \(w = \tilde{p}^n_{e1} - \tilde{p}^n_{e2}\) in (10.14a) and \(v = \Delta t(\tilde{q}^n_{e1} - \tilde{q}^n_{e2})\) in (10.14b). Adding the results gives

\[
(s^n_{e1} - s^n_{e2}, \tilde{p}^n_{e1} - \tilde{p}^n_{e2}) + \Delta t (k^{-1}_o(s^n_{e1}) q^n_{e1} - k^{-1}_o(s^n_{e2}) \tilde{q}^n_{e2}, q^n_{e1} - \tilde{q}^n_{e2}) = 0. \tag{10.15}
\]

Similarly, taking \(w = p^n_{e1} - p^n_{e2}\) in (10.14c) and \(v = \Delta t(q^n_{e1} - q^n_{e2})\) in (10.14d) and adding the results yields

\[
-(s^n_{e1} - s^n_{e2}, p^n_{e1} - p^n_{e2}) + \Delta t (k^{-1}_w(s^n_{e1}) \tilde{q}^n_{e1} - k^{-1}_w(s^n_{e2}) q^n_{e2}, q^n_{e1} - \tilde{q}^n_{e2}) = 0. \tag{10.16}
\]

Adding (10.15) and (10.16), we have

\[
\begin{align*}
(\tilde{p}^n_{e1} - \tilde{p}^n_{e2} - p^n_{e1} + p^n_{e2}, s^n_{e1} - s^n_{e2}) + \Delta t (k^{-1}_o(s^n_{e1}) q^n_{e1} - k^{-1}_o(s^n_{e2}) \tilde{q}^n_{e2}, q^n_{e1} - \tilde{q}^n_{e2}) + \Delta t (k^{-1}_w(s^n_{e1}) \tilde{q}^n_{e1} - k^{-1}_w(s^n_{e2}) q^n_{e2}, q^n_{e1} - \tilde{q}^n_{e2}) &= 0. \tag{10.17}
\end{align*}
\]

Taking \(w = s^n_{e1} - s^n_{e2}\) in (10.14e), one obtains

\[
(\tilde{p}^n_{e1} - \tilde{p}^n_{e2} - p^n_{e1} + p^n_{e2}, s^n_{e1} - s^n_{e2}) = (p_c(s^n_{e1}) - p_c(s^n_{e2}), s^n_{e1} - s^n_{e2}) + \frac{\Delta t}{\tau}(s^n_{e1} - s^n_{e2})^2 \leq \Delta t (k^{-1}_o(s^n_{e1}) q^n_{e1} - k^{-1}_o(s^n_{e2}) \tilde{q}^n_{e2}, q^n_{e1} - \tilde{q}^n_{e2}) + \Delta t (k^{-1}_w(s^n_{e1}) \tilde{q}^n_{e1} - k^{-1}_w(s^n_{e2}) q^n_{e2}, q^n_{e1} - \tilde{q}^n_{e2}) = 0. \tag{10.18}
\]

Substituting (10.18) into (10.17) gives

\[
\begin{align*}
\frac{\tau}{\Delta t} \|s^n_{e1} - s^n_{e2}\|^2 + (p_c(s^n_{e1}) - p_c(s^n_{e2}), s^n_{e1} - s^n_{e2}) + \Delta t (k^{-1}_o(s^n_{e1}) q^n_{e1} - k^{-1}_o(s^n_{e2}) \tilde{q}^n_{e2}, q^n_{e1} - \tilde{q}^n_{e2}) + \Delta t (k^{-1}_w(s^n_{e1}) \tilde{q}^n_{e1} - k^{-1}_w(s^n_{e2}) q^n_{e2}, q^n_{e1} - \tilde{q}^n_{e2}) = 0. \tag{10.19}
\end{align*}
\]

Using (A2) and (A3), (10.19) gives

\[
\begin{align*}
\frac{\tau}{\Delta t} \|s^n_{e1} - s^n_{e2}\|^2 + m_p \|s^n_{e1} - s^n_{e2}\|^2 + \frac{\Delta t}{M_o} \|\tilde{q}^n_{e1} - q^n_{e2}\|^2 + \frac{\Delta t}{M_w} \|q^n_{e1} - \tilde{q}^n_{e2}\|^2 \\
\leq -\Delta t (k^{-1}_o(s^n_{e1}) - k^{-1}_o(s^n_{e2}), q^n_{e1} - \tilde{q}^n_{e2}) - \Delta t (k^{-1}_w(s^n_{e1}) - k^{-1}_w(s^n_{e2}), q^n_{e1} - \tilde{q}^n_{e2}).
\end{align*}
\]
Using (A6), the first term of the right hand side can be estimated as
\[ -\Delta t \left( k_o^{-1} (s^n_{e1}) - k_o^{-1} (s^n_{e2}) \right) \]
\[ \leq \frac{\Delta t}{2} M_0 \| q^n_{e1} \|_{L^\infty(\Omega)}^2 \left( k_o^{-1} (s^n_{e1}) - k_o^{-1} (s^n_{e2}) \right)^2 + \frac{\Delta t}{2 M_o} \| q^n_{e1} - q^n_{e2} \|_2^2 \]
\[ \leq \frac{\Delta t M_o k_o^2}{2 \delta_o^2} \| q^n_{e1} \|_{L^\infty(\Omega)}^2 \| s^n_{e1} - s^n_{e2} \|_2^2 + \frac{\Delta t}{2 M_w} \| q^n_{e1} - q^n_{e2} \|_2^2 \leq 0. \]

Here \( L_{k_o} \) is the Lipschitz constant of \( k_o \) and Young’s inequality (see (10.21)) has been used. A similar estimate can be obtained for the term with \( k_w \). Hence, a constant \( C > 0 \) exists such that
\[ \left( \frac{\xi}{\Delta t} + m_p - C \Delta t \right) \| s^n_{e1} - s^n_{e2} \|_2^2 + \frac{\Delta t}{2 M_o} \| q^n_{e1} - q^n_{e2} \|_2^2 + \frac{\Delta t}{2 M_w} \| q^n_{e1} - q^n_{e2} \|_2^2 \leq 0, \]
which implies that \( s^n_{e1} = s^n_{e2}, q^n_{e1} = q^n_{e2} \) and \( q^n_{e1} = q^n_{e2} \) a.e. for \( \Delta t \) sufficiently small. Substitution of this into (10.14b) and (10.14d) concludes the proof.

For the ease of presentation, in what follows, we omit the subscript in (10.10). The fully discrete mixed scheme for the problem is given by the following discrete variational formulation:

**Problem 3 (Discrete Variational Formulation).** Let \( n \in \{1, \ldots, N\} \) and \( s^{n-1}_h \in W_h \) be given. Find \( (s^n_h, q^n_h, p^n_h, q^n_h, p^n_h) \in W_h \times V_h \times W_h \times V_h \times W_h \) such that

\[ \left( \frac{s^n_h - s^{n-1}_h}{\Delta t}, w_h \right) + (\nabla \cdot q^n_h, w_h) = 0, \]  
(10.20a)

\[ (k_o^{-1} (s^n_h) q^n_h, v_h) - (p^n_h, \nabla \cdot v_h) = 0, \]  
(10.20b)

\[ \left( \frac{s^n_h - s^{n-1}_h}{\Delta t}, w_h \right) + (\nabla \cdot q^n_h, w_h) = 0, \]  
(10.20c)

\[ (k_w^{-1} (s^n_h) q^n_h, v_h) - (p^n_h, \nabla \cdot v_h) = 0, \]  
(10.20d)

\[ (\bar{p}^n_h - p^n_h, w_h) = (p_c(s^n_h), w_h) + \tau \left( \frac{s^n_h - s^{n-1}_h}{\Delta t}, w_h \right), \]  
(10.20e)

for all \( w_h \in W_h \) and \( v_h \in V_h \). We take at time \( t = 0: s^0_h = P_h s^0 \).

The existence of the solution to Problem 3 is a consequence of Proposition 10.2, while the uniqueness of the solution can be proved following the steps in Lemma 10.1.
10.3.2 Error estimates

In the following analysis, we will use the Young’s inequality

\[ ab \leq \frac{1}{2\delta} a^2 + \frac{\delta}{2} b^2, \quad \text{for any } a, b \in \mathbb{R} \quad \text{and } \delta > 0. \]  

(10.21)

We use the following identity, valid for any two families of real vectors \( a^k, b^k \in \mathbb{R}^m \) (m \( \geq 1 \))

\[ \sum_{k=1}^{N} (a^k - a^{k-1}, a^k) = \frac{1}{2} (|a^N|^2 - |a^0|^2 + \sum_{k=1}^{N} |a^k - a^{k-1}|^2). \]  

(10.22)

Furthermore, we also will use the discrete version of Gronwall’s inequality (see [179])

**Lemma 10.2** (Discrete Gronwall inequality). If \( \{y_n\}, \{f_n\} \) and \( \{g_n\} \) are non-negative sequences and

\[ y_n \leq f_n + \sum_{0 \leq k < n} g_k y_k, \quad \text{for all } n \geq 0, \]

then

\[ y_n \leq f_n + \sum_{0 \leq k < n} f_k g_k \exp \left( \sum_{k < j < n} g_j \right), \quad \text{for all } n \geq 0. \]

Finally, we state the following result from [241].

**Lemma 10.3.** If the domain \( \Omega \) is convex, for any \( f_h \in W_h \), a \( v_h \in V_h \) exists such that

\[ \nabla \cdot v_h = f_h, \quad \text{and} \quad \|v_h\| \leq C_\Omega \|f_h\|, \]  

(10.23)

where the constant \( C_\Omega \) does not depend on \( f_h \).

Now we give the convergence result, based on error estimates. To this aim we use the notations

\[ e^n_s = s_n - s^n_h, \quad e^n_q = q_n - q^n_h, \quad e^n_p = p_n - p^n_h. \]

**Theorem 10.1.** Let \( (s_n, q^n_h, p^n_h, q^n, p^n) \) solve Problem 1 and \( (s^n_h, q^n_h, p^n_h, q^n_h, p^n_h) \), \( n \in \{1, \ldots, N\} \), solve Problem 3. Assuming that assumptions (A1)-(A6) hold, with \( \Delta t \) small enough we have for any \( n \in \{1, \ldots, N\} \),

\[ \|e^n\|^2 := \|e^n_s\|^2 + \|e^n_q\|^2 + \|e^n_p\|^2 + \|e^n\_q\|^2 + \|e^n\_p\|^2 \leq C(\Delta t^2 + h^2), \]

with the constant \( C \) not depending on \( \Delta t \) or \( h \).
Proof. By subtracting (10.10a) - (10.10e) from (10.20a) - (10.20e) respectively, we obtain

\begin{align}
\left( \partial_t s - \frac{s^n_h - s^{n-1}_h}{\Delta t}, w_h \right) &+ \left( \nabla \cdot (\tilde{q}(t_n) - q^n_h), w_h \right) = 0, \\
\left( k_o^{-1}(s(t_n))\tilde{q}(t_n) - k_o^{-1}(s^n_h)\tilde{q}_h^n, v_h \right) - \left( \tilde{p}(t_n) - \tilde{p}^n_h, \nabla \cdot v_h \right) = 0, \\
- \left( \partial_t s - \frac{s^n_h - s^{n-1}_h}{\Delta t}, w_h \right) + \left( \nabla \cdot (q(t_n) - q^n_h), w_h \right) = 0, \\
\left( k_w^{-1}(s(t_n))q(t_n) - k_w^{-1}(s^n_h)q^n_h, v_h \right) - \left( p(t_n) - p^n_h, \nabla \cdot v_h \right) = 0,
\end{align}

for all $w_h \in W_h$ and $v_h \in V_h$. Using the properties of the projectors $\Pi_h$ and $P_h$, the above equations become

\begin{align}
\left( \partial_t s - \frac{s^n_h - s^{n-1}_h}{\Delta t}, w_h \right) + \left( \nabla \cdot (\Pi_h \tilde{q}(t_n) - \tilde{q}_h^n), w_h \right) = 0, \\
\left( k_o^{-1}(s(t_n))\tilde{q}(t_n) - k_o^{-1}(s^n_h)\tilde{q}_h^n, v_h \right) - \left( P_h \tilde{p}(t_n) - \tilde{p}^n_h, \nabla \cdot v_h \right) = 0, \\
- \left( \partial_t s - \frac{s^n_h - s^{n-1}_h}{\Delta t}, w_h \right) + \left( \nabla \cdot (\Pi_h q(t_n) - q^n_h), w_h \right) = 0, \\
\left( k_w^{-1}(s(t_n))q(t_n) - k_w^{-1}(s^n_h)q^n_h, v_h \right) - \left( P_h p(t_n) - p^n_h, \nabla \cdot v_h \right) = 0,
\end{align}

\begin{align}
\left( P_h \tilde{p}(t_n) - \tilde{p}^n_h - P_h p(t_n) + p^n_h, w_h \right) = \left( p_c(s(t_n)) - p_c(s^n_h), w_h \right) + \tau \left( \partial_t s - \frac{s^n_h - s^{n-1}_h}{\Delta t}, w_h \right),
\end{align}

Take $w_h = P_h \tilde{p}(t_n) - \tilde{p}^n_h$ in (10.25a) and $v_h = \Pi_h \tilde{q}(t_n) - \tilde{q}_h^n$ in (10.25b). Adding the results, we obtain

\begin{align}
\left( \partial_t s - \frac{s^n_h - s^{n-1}_h}{\Delta t}, P_h \tilde{p}(t_n) - \tilde{p}^n_h \right) + \left( k_o^{-1}(s(t_n))\tilde{q}(t_n) - k_o^{-1}(s^n_h)\tilde{q}_h^n, \Pi_h \tilde{q}(t_n) - \tilde{q}_h^n \right) = 0.
\end{align}

Similarly, setting $w_h = P_h p(t_n) - p^n_h$ in (10.25c), and $v_h = \Pi_h q(t_n) - q^n_h$ in (10.25d) we have

\begin{align}
- \left( \partial_t s - \frac{s^n_h - s^{n-1}_h}{\Delta t}, P_h p(t_n) - p^n_h \right) + \left( k_w^{-1}(s(t_n))q(t_n) - k_w^{-1}(s^n_h)q^n_h, \Pi_h q(t_n) - q^n_h \right) = 0.
\end{align}
Adding (10.26) and (10.27) yields
\[
\left( \partial_t s - \frac{s^n_h - s^{n-1}_h}{\Delta t}, P_h \bar{p}(t_n) - \bar{p}_h^n - P_h p(t_n) + p_h^n \right) + \left( k_0^{-1}(s(t_n)) \mathbf{q}(t_n) - k_0^{-1}(s_h^n) \mathbf{q}_h^n, \Pi_h \mathbf{q}(t_n) - \mathbf{q}_h^n \right) + \left( k_w^{-1}(s(t_n)) \mathbf{q}(t_n) - k_w^{-1}(s_h^n) \mathbf{q}_h^n, \Pi_h \mathbf{q}(t_n) - \mathbf{q}_h^n \right) = 0. \tag{10.28}\]

Taking \( w_h = P_h \bar{p}(t_n) - \bar{p}_h^n - P_h p(t_n) + p_h^n \) in (10.25e), one has
\[
\left( \partial_t s - \frac{s^n_h - s^{n-1}_h}{\Delta t}, P_h \bar{p}(t_n) - \bar{p}_h^n - P_h p(t_n) + p_h^n \right) = \frac{1}{\tau} \left( p_c(s(t_n)) - p_c(s_h^n), P_h \bar{p}(t_n) - \bar{p}_h^n - P_h p(t_n) + p_h^n \right) \tag{10.29}\]

This gives
\[
\frac{1}{\tau} \left( p_c(s(t_n)) - p_c(s_h^n), P_h \bar{p}(t_n) - \bar{p}_h^n - P_h p(t_n) + p_h^n \right) + \left( k_0^{-1}(s(t_n)) \mathbf{q}(t_n) - k_0^{-1}(s_h^n) \mathbf{q}_h^n, \Pi_h \mathbf{q}(t_n) - \mathbf{q}_h^n \right) + \left( k_w^{-1}(s(t_n)) \mathbf{q}(t_n) - k_w^{-1}(s_h^n) \mathbf{q}_h^n, \Pi_h \mathbf{q}(t_n) - \mathbf{q}_h^n \right) = 0. \tag{10.30}\]

Then we have by (A3)
\[
\frac{1}{2\tau} \left( P_h \bar{p}(t_n) - \bar{p}_h^n - P_h p(t_n) + p_h^n \right) + \left( k_0^{-1}(s(t_n)) \mathbf{q}(t_n) - k_0^{-1}(s_h^n) \mathbf{q}_h^n, \Pi_h \mathbf{q}(t_n) - \mathbf{q}_h^n \right) + \left( k_w^{-1}(s(t_n)) \mathbf{q}(t_n) - k_w^{-1}(s_h^n) \mathbf{q}_h^n, \Pi_h \mathbf{q}(t_n) - \mathbf{q}_h^n \right) \leq C \left\| s(t_n) - s_h^n \right\|^2. \tag{10.31}\]

From (10.31), after some manipulations, we have
\[
\frac{1}{2\tau} \left( P_h \bar{p}(t_n) - \bar{p}_h^n - P_h p(t_n) + p_h^n \right) + \left( k_0^{-1}(s(t_n)) \mathbf{q}(t_n), \mathbf{e}_q^n \right) + \left( k_0^{-1}(s_h^n) \mathbf{q}_h^n, \Pi_h \mathbf{q}(t_n) - \mathbf{q}_h^n \right) + \left( k_w^{-1}(s(t_n)) \mathbf{q}(t_n), \mathbf{e}_q^n \right) + \left( k_w^{-1}(s_h^n) \mathbf{q}_h^n, \Pi_h \mathbf{q}(t_n) - \mathbf{q}_h^n \right) \leq C \left\| \mathbf{e}_q^n \right\|^2. 
\]
10.3 Mixed formulation and error estimates

We name the terms $T_1$, $T_2$, $T_3$, $T_4$, $T_5$, $T_6$, $T_7$, $T_8$, $T_9$, $T_{10}$ from left to right. Obviously, $T_1 \geq 0$. Then by using (10.21), Cauchy-Schwarz inequality, (A2) and (A6), we obtain similar to the proof of Lemma 10.1,

$$|T_2| = \left| \left( k_0^{-1}(s(t_n)) - k_0^{-1}(s_h^n) \bar{q}(t_n), e^p_{q} \right) \right| \leq C \| e^p_{q} \|^2 + \frac{\delta_2}{2} \| e^p_{q} \|^2,$$

for a $\delta_2 > 0$. Similarly, for $T_3$ there holds

$$|T_3| = \left| \left( k_0^{-1}(s(t_n)) - k_0^{-1}(s_h^n) \bar{q}(t_n), \Pi_h \bar{q}(t_n) - \bar{q}(t_n) \right) \right| \leq C(\| e^p_{q} \|^2 + \| \bar{q}(t_n) - \Pi_h \bar{q}(t_n) \|^2).$$

Then using the boundedness of the non-wetting permeability $k_0$, we get for $T_4$

$$T_4 = (k_0^{-1}(s_h^n) e^p_{q}, e^p_{q}) \geq \frac{1}{M_0} \| e^p_{q} \|^2.$$

For the term $T_5$ we use the boundedness of the permeability $k_0$, (10.21) and Cauchy-Schwarz inequality

$$|T_5| = \left| \left( k_0^{-1}(s_h^n) e^p_{q}, \Pi_h \bar{q}(t_n) - \bar{q}(t_n) \right) \right| \leq \frac{\delta_5}{2} \| e^p_{q} \|^2 + C \| \bar{q}(t_n) - \Pi_h \bar{q}(t_n) \|^2,$$

for a $\delta_5 > 0$. Similarly, we estimate the terms $T_6$, $T_7$, $T_8$ and $T_9$

$$|T_6| = \left| \left( k_0^{-1}(s(t_n)) - k_0^{-1}(s_h^n) \bar{q}(t_n), e^p_{q} \right) \right| \leq C \| e^p_{q} \|^2 + \frac{\delta_6}{2} \sum_{n=1}^{N^*} \| e^p_{q} \|^2,$n

$$|T_7| = \left| \left( k_0^{-1}(s(t_n)) - k_0^{-1}(s_h^n) \bar{q}(t_n), \Pi_h \bar{q}(t_n) - \bar{q}(t_n) \right) \right| \leq C(\| e^p_{q} \|^2 + \| \bar{q}(t_n) - \Pi_h \bar{q}(t_n) \|^2),$$

$$T_8 = (k_0^{-1}(s_h^n) e^p_{q}, e^p_{q}) \geq \frac{1}{M_w} \| e^p_{q} \|^2,$n

$$|T_9| = \left| \left( k_0^{-1}(s_h^n) e^p_{q}, \Pi_h \bar{q}(t_n) - \bar{q}(t_n) \right) \right| \leq \frac{\delta_9}{2} \| e^p_{q} \|^2 + C \| \bar{q}(t_n) - \Pi_h \bar{q}(t_n) \|^2.$n

We gather $T_1$ to $T_{10}$ and choose $\delta_2$, $\delta_5$, $\delta_6$ and $\delta_9$ properly to have

$$\| e^p_{q} \|^2 + \| e^p_{q} \|^2 \leq C(\| e^p_{q} \|^2 + \| \bar{q}(t_n) - \Pi_h \bar{q}(t_n) \|^2 + \| \bar{q}(t_n) - \Pi_h \bar{q}(t_n) \|^2).$$

Furthermore, taking $\nabla \cdot \mathbf{v}_h = P_h \bar{p}(t_n) - \bar{p}_h^n$ in (10.24b) and applying Lemma 10.3 we obtain

$$\| \bar{p}(t_n) - \bar{p}_h^n \|^2 = \left( k_0^{-1}(s(t_n)) \bar{q}(t_n) - k_0^{-1}(s_h^n) \bar{q}(t_n), \mathbf{v}_h \right) - \left( \bar{p}(t_n) - \bar{p}_h^n, P_h \bar{p}(t_n) - \bar{p}(t_n) \right) \leq C(\| e^p_{q} \|^2 + \| e^p_{q} \|^2) \| P_h \bar{p}(t_n) - \bar{p}_h^n \| + \| \bar{p}(t_n) - \bar{p}_h^n \| \| P_h \bar{p}(t_n) - \bar{p}(t_n) \| \leq C(\| e^p_{q} \|^2 + \| e^p_{q} \|^2) + \frac{3}{2} \| P_h \bar{p}(t_n) - \bar{p}(t_n) \|^2 + \frac{3}{4} \| \bar{p}(t_n) - \bar{p}_h^n \|^2.$
which gives by applying (10.7)
\[ \| e_p^n \|^2 = \| \bar{\rho}(t_n) - \bar{\rho}_h^n \|^2 \leq C(\| e_q^n \|^2 + \| e_q^n \|^2 + h^2). \]  
(10.35)
Similarly, by taking \( V \cdot \mathbf{v}_h = P_h p(t_n) - p_h^n \) in (10.24d), one also has
\[ \| e_p^n \|^2 = p(t_n) - p_h^n \|^2 \leq C(\| e_q^n \|^2 + \| e_q^n \|^2 + h^2). \]  
(10.36)
Setting \( w_h = P_h s(t_n) - s_h^n \) in (10.25e) and summing up the results from \( n = 1 \) to \( N^* \), \( N^* \in \{1, \ldots, N\} \), we have after some manipulations,
\[
\sum_{n=1}^{N^*} \left( \partial_t s - \frac{s(t_n) - s(t_{n-1})}{\Delta t}, e_s^n \right)
+ \sum_{n=1}^{N^*} \left( \partial_t s - \frac{s_h^n - s_h^{n-1}}{\Delta t}, P_h s(t_n) - s(t_n) \right)
+ \sum_{n=1}^{N^*} \left( \frac{s(t_n) - s(t_{n-1}) - s_h^n + s_h^{n-1}}{\Delta t}, e_s^n \right)
= \frac{1}{\tau} \sum_{n=1}^{N^*} \left( P_h \bar{\rho}(t_n) - \bar{\rho}_h^n \right)
- \frac{1}{\tau} \sum_{n=1}^{N^*} \left( p_c(s(t_n)) - p_c(s_h^n), P_h s(t_n) - s_h^n \right).
\]
We proceed as before, denoting the terms by \( S_1, S_2, S_3, S_4, S_5 \). Then we have
\[ S_1 + S_2 + S_3 = S_4 + S_5. \]
We use the inequalities of (10.21) and Cauchy-Schwarz to estimate \( S_1 \):
\[
|S_1| = \left| \sum_{n=1}^{N^*} \left( \partial_t s - \frac{s(t_n) - s(t_{n-1})}{\Delta t}, e_s^n \right) \right|
\leq \frac{1}{\Delta t} \sum_{n=1}^{N^*} \| \Delta t \partial_t s - (s(t_n) - s(t_{n-1})) \| \cdot \| e_s^n \|
\leq \frac{1}{2 \Delta t^2} \sum_{n=1}^{N^*} \int_{t_{n-1}}^{t_n} (\partial_t s(t_n) - \partial_t s(z)) dz \right)^2 + \frac{1}{2} \sum_{n=1}^{N^*} \| e_s^n \|^2.
\]
Further, since \( \partial_{tt} s \in L^2(0, T; L^2(\Omega)) \) (see [149] Proposition 4.2), we estimate \( S_{11} \) by using the Bochner inequality
\[
S_{11} = \frac{1}{2 \Delta t^2} \sum_{n=1}^{N^*} \int_{t_{n-1}}^{t_n} \int_{t_{n-1}}^{t_n} (\partial_{tt} s(\eta)) d\eta dz \right)^2
\leq \frac{1}{2 \Delta t^2} \sum_{n=1}^{N^*} \int_{t_{n-1}}^{t_n} (\partial_{tt} s(z)) dz \right)^2
\leq \frac{\Delta t}{2} \sum_{n=1}^{N^*} \int_{t_{n-1}}^{t_n} \| \partial_{tt} s(\eta) \|^2 d\eta
= \frac{\Delta t}{2} \int_0^T \| \partial_{tt} s(\eta) \|^2 d\eta
\leq C\Delta t.
\]
For $S_2$, by using the properties of the projectors (10.5) and applying (10.7) one gets

$$|S_2| = \left| \sum_{n=1}^{N^*} \left( \frac{s^n_h - s^{n-1}_h}{\Delta t} - P_h s(t_n) - s(t_n) \right) \right| = \left| \sum_{n=1}^{N^*} \left( \partial_t s, P_h s(t_n) - s(t_n) \right) \right|$$

$$= \left| \sum_{n=1}^{N^*} \left( \partial_t s - P_h \partial_t s, P_h s(t_n) - s(t_n) \right) \right| \leq N^* h^2 \leq \frac{T h^2}{\Delta t}.$$

Now we estimate the term $S_3$. By using (10.22), we have

$$S_3 = \frac{1}{2} \left( \sum_{n=1}^{N^*} \left( \frac{s(t_n) - s(t_{n-1})}{\Delta t} - \frac{s^n_h + s^{n-1}_h}{2} e^n_s \right) \right)^2$$

$$\leq \frac{1}{2} \left( \sum_{n=1}^{N^*} \left( \frac{1}{\Delta t} \left( s^0 - s(t_n) \right) \right)^2 + \frac{1}{2} \left( \sum_{n=1}^{N^*} \| e^n_s - e^{n-1}_s \|^2 \right) \right).$$

For $S_4$ and $S_5$ we have

$$|S_4| = \frac{1}{2} \left( \sum_{n=1}^{N^*} \left( \left( \frac{1}{\Delta t} \left( s^0 - s(t_n) \right) \right)^2 + \frac{1}{2} \left( \sum_{n=1}^{N^*} \| e^n_s - e^{n-1}_s \|^2 \right) \right) \right).$$

Gathering the estimates for $S_1$ to $S_5$, one obtains

$$\frac{1}{2} \left( \sum_{n=1}^{N^*} \left( \frac{1}{\Delta t} \left( s^0 - s(t_n) \right) \right)^2 + \frac{1}{2} \left( \sum_{n=1}^{N^*} \| e^n_s - e^{n-1}_s \|^2 \right) \right) \leq C \Delta t + \frac{T h^2}{\Delta t} + \frac{1}{2} \left( \sum_{n=1}^{N^*} \| e^n_s \|^2 + \frac{1}{2} \left( \sum_{n=1}^{N^*} \| e^n_s - e^{n-1}_s \|^2 \right) \right).$$
By applying the properties of the operator $P_h$, one has
\[
\|e_s^{N^*}\|^2 + \frac{\Delta t}{2} \sum_{n=1}^{N^*} \|e_s^n - e_s^{n-1}\|^2
\]
\[
\leq C(\Delta t^2 + h^2) + \frac{\Delta t}{2} \sum_{n=1}^{N^*} \|e_s^n\|^2 + \frac{1}{2} \|s^0 - P_h s^0\|^2 + C \sum_{n=1}^{N^*} \left( \Delta t \|e_p^n\|^2 + \Delta t \|e_p^n\|^2 + \Delta t \|e_s^n\|^2 \right)
\]
\[
+ C \sum_{n=1}^{N^*} \left( \Delta t \|P_h \bar{p}(t_n) - \bar{p}(t_n)\|^2 + \Delta t \|P_h p(t_n) - p(t_n)\|^2 + \Delta t \|P_h s(t_n) - s(t_n)\|^2 \right)
\]
\[
\leq C(h^2 + \Delta t^2) + C \sum_{n=1}^{N^*} \Delta t \|e_s^n\|^2 + C \sum_{n=1}^{N^*} \Delta t \|e_p^n\|^2 + C \sum_{n=1}^{N^*} \Delta t \|e_s^n\|^2.
\]

Applying the estimates from (10.33), (10.35) and (10.36), and the property of the operator $\Pi_h$, we have
\[
\|e_s^{N^*}\|^2 \leq C(h^2 + \Delta t^2).
\]

By using the Gronwall’s inequality stated in Lemma 10.2, we obtain
\[
\|e_s^{N^*}\|^2 \leq C(h^2 + \Delta t^2).
\] (10.37)

The proof is completed by substituting (10.37) into (10.33), (10.35) and (10.36).

10.4 Numerical results

For the numerical results, triangular meshes from the FVCA8 benchmark mesh repository has been used. The domain selected is the unit square
\[
\Omega = (0, 1) \times (0, 1).
\] (10.38)

The results presented in Section 10.3 are for two-phase equation without gravity. However, it was mentioned in Remark 10.3 that the analysis also holds in the presence of gravity. Thus, for the numerical results we have included gravity terms in (10.4):

\[
\bar{q} = -k_o(s)(\nabla \bar{p} - \rho_o \hat{g}),
\]
\[
q = -k_w(s)(\nabla p - \rho_w \hat{g}),
\]

with $\hat{g}$ being the unit vector along the direction of gravity and $\rho_o, \rho_w$ being the non-dimensionalized densities of the non-wetting (oil) and the wetting (water) phases.
10.4 Numerical results

Gravity has been taken pointing along the \( x \)-direction. For the densities, \( \rho_o = 0.8 \) and \( \rho_w = 1 \) are used throughout. Moreover, source terms \( f_o \) and \( f_w \) have been added in (10.1a) and (10.1b) respectively.

The functions \( k_o, k_w \) and \( p_c \) used in the computations are

\[
k_o(s) = s, \quad k_w(s) = (1 - s), \quad p_c(s) = s.
\]

Simple linear forms for \( k_o, k_w \) and \( p_c \) are used because the exact solution can be computed in this case. This makes the calculated error terms exact. Otherwise, the numerical solution had to be compared with the time discrete solution computed for a very fine grid, which introduces additional errors.

The source terms \( f_o, f_w \) have the form

\[
f_o = \frac{e^{-t}}{4} [1 + x^2 + y^2 - 2\rho_o x + \tau((1 + 2x^2 + 2y^2)e^{-t} - 4)],
\]

(10.41a)

\[
f_w = -\frac{e^{-t}}{4} [1 + x^2 + y^2 - 2\rho_w x + (1 + 2x^2 + 2y^2)e^{-t}].
\]

(10.41b)

The initial and boundary conditions are shown in Table 10.1.

<table>
<thead>
<tr>
<th>IC</th>
<th>( \Omega )</th>
<th>( t = 0 )</th>
<th>( s(x, y, 0) = 1 - \frac{1}{4}(1 + x^2 + y^2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC</td>
<td>( \Omega )</td>
<td>( x = 0 )</td>
<td>( \bar{\rho}(0, y, t) = 1 + \frac{1}{4}(1 + y^2)e^{-t} ) ( \rho(0, y, t) = \frac{1}{4}(1 + y^2)e^{-t} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( x = 1 )</td>
<td>( \bar{\rho}(1, y, t) = 1 + \frac{1}{4}(2 + y^2)e^{-t} ) ( \rho(1, y, t) = \frac{1}{4}(2 + y^2)e^{-t} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( y = 0 )</td>
<td>( \bar{q}_y = 0 ) ( q_y = 0 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( y = 1 )</td>
<td>( \bar{q}_y = -\frac{1}{2}(1 - \frac{1}{4}(2 + x^2)e^{-t})e^{-t} ) ( q_y = -\frac{1}{8}(2 + x^2)e^{-2t} )</td>
</tr>
</tbody>
</table>

Observe that, both Dirichlet and Neumann boundary conditions have been used instead of the zero Dirichlet condition assumed for the analysis in Section 10.3. This is to show that the result of Theorem 10.1 stays valid for more general boundary conditions.

One can directly verify that under these conditions the exact solution of the system is given by

\[
s = 1 - \frac{1}{4}(1 + x^2 + y^2)e^{-t}, \quad \bar{\rho} = 1 + \frac{\tau}{4}(1 + x^2 + y^2)e^{-t}, \quad p = \frac{1}{4}(1 + x^2 + y^2)e^{-t},
\]

(10.42)

with \( \bar{q} \) and \( q \) calculated from (10.39).
Four different mesh sizes have been used for the computations, i.e. \( h = 0.1, 0.05, 0.02 \) and 0.01, and \( \tau = 1 \) has been taken throughout. Figure 10.1 (left) shows the saturation for the exact and the numerical solution at \( t = 1 \). The two profiles look almost superimposed on each other which implies that the computational results are accurate. As the equations are nonlinear, a linear iterative method has to be used. In our computations we have used the L-scheme as presented in [203]. Figure 10.1 (right) shows how the errors decrease with iterations of the linearisation step. An error cut-off of \( 10^{-8} \) has been used in order to make sure that the errors due to the nonlinear solver do not influence the results of the order estimate.

To recover the order of convergence of the scheme, we vary \( h \) and \( \Delta t \) so that

\[
h = \Delta t. \tag{10.43}
\]

By taking \( h \) in this form one ensures that \( \log_{10} \sqrt{(h^2 + \Delta t^2)} = \log_{10} h + \text{constant} \). So if the slope of \( \log_{10} \| e^n \| \) against \( \log_{10} h \) is 1 then it would support the analytical findings in Section 10.3. This is precisely the case as shown in Figure 10.2. It shows the error plots at two times, \( t = 0.2, 0.5 \). The lines are nearly parallel to the black dashed line, representing slope of 1.

An interesting observation is that the individual components of \( \| e^n \| \) also scale with \( h \) when \( h = \Delta t \). This is shown in Figure 10.3. For \( h = \Delta t \) the log plots of \( e^n_s, e^n_p, e^n_{\tilde{p}}, e^n_{\tilde{q}}, e^n_{\tilde{q}} \) are all parallel to the black dashed line representing slope of 1. This supports parts of the proof of Theorem 10.1, i.e. (10.33), (10.37).
10.4 Numerical results

Figure 10.2: Error \( \| \mathbf{e}_n \| \) vs. \( \log_{10} h \), where \( h \) and \( \Delta t \) follow the relation (10.43). The errors are plotted at two times, \( t = 0.2, 0.5 \). The black dashed line represents the slope of 1.

Figure 10.3: Error \( \log_{10}(\| u_n - u(t_n) \|) \) vs. \( \log_{10} h \). Here \( u = p, \bar{p}, s, \bar{q}, \bar{q} \). All the plots are for \( t = 0.5 \) and \( h, \Delta t \) are such that \( h = \Delta t \). The black dashed line represents the slope of 1.

Next, \( h \) and \( \Delta t \) are varied in a general way so that \( h = \Delta t \) is not always satisfied. We plot the errors \( \| \mathbf{e}_n \|^2 \) against \( \Delta t^2 \) in Figure 10.4 (left) and against \( h^2 \) in Figure 10.4 (right). The nearly straight profiles of the errors, for a fixed \( h \) and for a fixed \( \Delta t \), implies that the errors behave as

\[
\| \mathbf{e}_n \|^2 \approx A h^2 + B \Delta t^2,
\]

where \( A, B > 0 \) are constants. This is similar to what one expects from the proof of Theorem 10.1, for small enough \( \Delta t \). Relation (10.44) still gives linear profiles in
Figure 10.2. To estimate the values of $A$ and $B$, the average slopes of the line $h = 0.1$ in Figure 10.4 (left) and the line $\Delta t = 0.5$ in Figure 10.4 (right) are calculated. This gives $A = 0.02$ and $B = 0.22$.

Figure 10.4: (left) Errors $\|e^n\|^2$ vs. $\Delta t^2$ for fixed $h$. (right) Errors $\|e^n\|^2$ vs. $h^2$ for fixed $\Delta t$. The mesh sizes are varied between $h = 0.1$, $0.05$, $0.02$ and $\Delta t = 0.5$ is fixed.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\Delta t$</th>
<th>$\Delta t = 0.1$</th>
<th>$\Delta t = 0.05$</th>
<th>$\Delta t = 0.025$</th>
<th>$\Delta t = 0.01$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0.1$</td>
<td>$-1.3179$</td>
<td>$-1.5548$</td>
<td>$-1.7191$</td>
<td>$-1.8110$</td>
<td></td>
</tr>
<tr>
<td>$0.05$</td>
<td>$-1.3099$</td>
<td>$-1.5625$</td>
<td>$-1.7359$</td>
<td>$-1.8268$</td>
<td></td>
</tr>
<tr>
<td>$0.02$</td>
<td>$-1.3277$</td>
<td>$-1.5968$</td>
<td>$-1.8235$</td>
<td>$-1.9943$</td>
<td></td>
</tr>
<tr>
<td>$0.01$</td>
<td>$-1.3239$</td>
<td>$-1.6109$</td>
<td>$-1.8635$</td>
<td>$-2.0713$</td>
<td></td>
</tr>
</tbody>
</table>

Table 10.2: $\log_{10}(\|e^n\|)$ vs. $\log_{10}(\sqrt{Ah^2 + B\Delta t^2})$ for different mesh sizes and timesteps. For the error estimation, $A = 0.02$ and $B = 0.22$ are used in (10.44) and they are calculated from Figure 10.4. In all the cases, $t = 0.5$.

We use the values of $A$ and $B$ to predict the error for any $\Delta t$ and $h$. This is shown in Table 10.2. Note that the value of $A$ and $B$ is derived from the line $h = 0.1$ and
$\Delta t = .05$ in Figure 10.4 but they are used to estimate the rest of the errors. The match is reasonably close.
Chapter 11

Summary, outlook and recommendations

11.1 Summary

To conclude, in this work, we consider unsaturated and two-phase flow through porous media and focus on modelling and numerical aspects of the problem. Accordingly, the work is divided into two parts. In the first part, we discuss different non-equilibrium models that include hysteresis and dynamic capillarity effects. These effects play a major role in processes such as generation of overshoots and plateaus, formation of fingers, redistribution, and trapping. However, they add an extra layer of complexity to the problem and thus, are often excluded from the standard model described in Section 1.1. In the second part, we look at the numerical challenges associated with the porous media flow problem. Specifically, we discuss the issues arising from heterogeneity and layered structure of the porous domains, from non-linearities and degeneracies associated with the model, and from non-equilibrium effects that were mentioned earlier. The main results for the respective parts are as follows.

11.1.1 Part I: non-equilibrium effects

In Chapter 2 we review different models that are commonly used to incorporate hysteresis in capillary pressure. In particular, the play-type hysteresis model is discussed, which is a comparatively simple and closed-form extension of the standard model as it assumes the scanning curves to be vertical (constant saturation lines). However, it is limited in a sense, because the scanning curves obtained from experiments are non-vertical. This is shown to cause the play-type hysteresis model to predict certain unphysical solutions, like frozen profiles for the horizontal redistribution problem. To circumvent this issue, we extend the play-type model by incorporating non-vertical
scanning curves in its closed-form expression. This model can be fitted to approximate experimentally obtained scanning curves closely, and it is thermodynamically consistent. Finally, all the cases of horizontal redistribution are studied and categorized using this model. Unconventional flow cases, where water flows from low to high saturation regions, theorised previously in literature, are shown to occur based on the analysis and also based on numerical experiments. In Chapter 3, it is shown that this model is well-posed i.e., there exists a weak solution of the extended play-type model complemented by the Richards equation (Ri) (equation (1.20)), even in the degenerate case, provided that the initial conditions are bounded away from the points of degeneracy. A maximum principle is also proved for the model.

In Chapter 4, we look at the infiltration experiment set up, an example of which is shown in Figure 7.1, for unsaturated flow problems. The imbibition fronts are investigated using travelling wave (TW) solutions with the focus being on the influence of capillary hysteresis and dynamic capillarity effects on the behaviour of the fronts. In the pure hysteresis case, it is shown that TWs exist and they lie on the imbibition capillary pressure curve. In the dynamic capillarity case, TWs are also shown to exist. However, they can be non-monotone, explaining the occurrence of overshoots. Thresholds are provided that determine the behaviour of the TWs. For example, if the dynamic capillary coefficient \( \tau \) exceeds a certain threshold, the profiles are non-monotone, and when it exceeds another threshold, full saturation will be reached. Also, the consequences of having a nonlinear dynamic coefficient are investigated, which show that the saturation stays away from its maximum value, \( S = 1 \), if the nonlinearity is non-integrable near 1. Chapter 5 generalises these results to the combined capillary hysteresis and dynamic capillarity case. The TWs are proven to exist for the play-type model and plateau-like oscillations are shown to be possible in this case. If extended play-type hysteresis model is used, then the existence of the TWs is proved under a restriction on \( \tau \). However, for the extended play-type model, there can only be finitely many oscillations, which is observed from experiments.

Chapter 6 discusses the role of hysteresis and dynamic capillarity in the propagation of viscous fingers. Using TWs, it is shown that for a fixed flux at the inlet boundary, a monotone TW solution is possible representing the one-dimensional flow inside a finger. The corresponding wave-speed of the TW gives the speed of propagation of the finger.

Finally, the results in Chapters 4 and 5 are extended to the two-phase case in Chapter 7, where the flux function is convex-concave. First, all possible pairs of saturation that are connectable by TWs are delineated for the case when non-equilibrium effects are present in the capillary pressure. In the hyperbolic limit, the existence conditions for the TWs provide admissibility conditions for shocks, and thus, the entropy solutions corresponding to this viscous model are recovered. They differ significantly from the entropy solution of the standard model which disregards non-equilibrium effects. For example, an imbibition shock followed by a drainage shock with a widening plateau (saturation overshoot) in between, becomes an admissible solution to the
Riemann problem. The results are further extended to the case when relative permeabilities are hysteretic. A new model for relative permeability hysteresis is proposed for this purpose, which models the relative permeabilities as a function of saturation and pressure, instead of only saturation. Using this model, existence of TWs is proved for small values of $\tau$, and corresponding entropy solutions are derived. Interestingly, the entropy solutions depend on the initial pressure. Furthermore, if certain parametric conditions are met, then the solutions can develop into a stable plateau where the imbibition and the drainage waves delimiting the plateau move at the same speed. The long time behaviour of numerical solutions is in accordance with the analytical predictions.

Overall, through Chapters 2-7, we see that the non-equilibrium effects are important in describing flow through the subsurface. In particular, this is shown for infiltration problems, development of overshoots, horizontal redistribution and propagation of viscous fingers. We see that even in the hyperbolic limit when the capillary terms are vanishing, the non-equilibrium effects play a role in determining the entropy solutions. Moreover, comparatively small effects such as hysteresis in relative permeability can have a considerable impact on the solutions. Now, we briefly discuss how our results compare with experiments.

### 11.1.2 Comparison with experiments

From a practical point of view, we note that this work can also be used to explain experimental results reported under various settings, e.g. [72, 102, 103, 134, 272]. In Chapters 4, 5 and 7, the occurrence of saturation overshoots is predicted theoretically for high enough dynamic capillary effect, namely of the $\tau$ value in (7.16). In dimensionless setting, this can be assimilated to a sufficiently large injection rate. This is in line with the experimental results in [72], where the development of plateau-like profiles was observed for high enough injection rates, as shown in Figure 5 of [72] and Figure 5.3 of [272]. Moreover, the occurrence of a finite number of overshoots, as observed in [72], is explained using the extended play-type model in Section 5.3. The analysis in Sections 4.4.2 and 5.2.3 shows that the saturation does not reach its maximum value for the unsaturated case if the nonlinear dynamic capillarity coefficient is non-integrable near $S = 1$. It is seen from experiments in [47, 275] that $\tau$ indeed blows up near $S = 1$, which might thus be a natural mechanism preventing the domain from becoming fully saturated. A numerical study, in this regard, can be found in [276].

In the water and oil case, plateaus are seen to develop and grow in Figures 5-6, 8-9, 18 of [102]. This behaviour is predicted by the analysis in Section 7.3. Further, Figure 10 of [102] might be presenting the case when the saturation has developed a plateau between two fronts travelling with the same velocity, a situation that is explained by the authors by means of hysteretic effects in the flux functions. Such solutions are investigated numerically in [119, 216], where it is shown that the plateaus
can persist in time but without explaining how they are generated. The results in Section 7.4 partly support the conclusions there, but also explain the mechanism behind the development of such plateaus. We mention [131] in this regard, where the authors conclude that a similar mechanism must be responsible for observed stable saturation plateaus inside viscous fingers.

Finally, experiments in [103, 211, 225] emphasize the importance of hysteresis in the development of fingers. The results obtained in Chapter 6 give the speed of propagation of the fingers based on this assumption. This is supported by numerical experiments in [209].

11.1.3 Part II: numerical methods

In Chapter 8, we introduce a linear domain decomposition (LDD) scheme for unsaturated flow in heterogeneous media. The method combines the linearisation iterations, required to solve the nonlinear problem, with domain decomposition iterations, that divide the problem into subproblems defined on subdomains. Moreover, the decoupled domains can be solved in parallel on different threads. The scheme is based on the L-scheme and it is shown to converge for a mild restriction on the time step that does not depend on the spatial discretization. Numerically, it is shown to conserve flux and pressure continuity at the interface of different subdomains. Although the Newton and Picard schemes require fewer iterations to converge, the LDD scheme takes less computational time per iteration. When parallelized, it is shown to outperform the monolithic schemes. Finally, a parameter study is given which shows that the convergence rate has a strong dependence on the parameters.

Next, for nonlinear diffusion problems, a modified version of the L-scheme is proposed in Chapter 9. This scheme utilizes the fact that at each time step the time-discrete solution of the previous time step is known and uses conservative local estimates for stabilization. It is proved that this turns the iterations into a contraction for a mild restriction on the time step size. The convergence rate scales up linearly with the time step size if the diffusion coefficient is linear, and proportional to the square root of the time step size otherwise. Numerically it is shown that the scheme is at least as fast as the modified Picard scheme, much faster than the conventional L-scheme and more stable than either Newton or the Picard scheme. The chapter ends with a parametric study showing that the convergence speed can be optimized by varying the main parameter.

Finally in Chapter 10, we consider the fully discrete two-phase flow model incorporating dynamic capillarity effect. For the time-discretization, the backward Euler method is used, whereas, lowest order Raviart-Thomas mixed finite elements are used for spatial discretization. Error estimates are given that show the convergence of this scheme. Numerical experiments support our analysis.
11.2 Future outlook

For the non-equilibrium effects, following this work, future research can be done in the following directions:

- Extension of the results for the infiltration problem to incorporate the extended play-type hysteresis model. As shown in Chapters 2 and 5, this potentially improves the predictions of the play-type hysteresis model.

- Considering non-monotone flux functions, arising due to gravity, in the two-phase flow case considered in Chapter 7. This is partially discussed already in Chapter 7. However, there are many cases to consider. Moreover, preliminary numerical results suggest that there is a disparity between the classical rarefaction wave whose endpoints travel in different directions, and the rarefaction wave obtained in a similar situation but with hysteresis included.

- The analysis for relative permeability hysteresis with large $\tau$ remains to be done. Numerical results suggest that a limit cycle structure in the saturation-pressure phase plane is possible in this case. However, they require further analysis.

- Stability of the TW profiles is to be investigated. The numerical results so far hint at the TW profiles being stable.

- Finally, extending the results for the non-equilibrium effects to the three dimensional full-scale problem is the long term goal of this research. For this purpose, a numerical solver has to be written that incorporates the results in this work.

For the numerical methods:

Chapter 8 The LDD scheme has already been extended to the two-phase case in [227] and incorporates the non-equilibrium effects in [159]. We mention [4] where a similar technique is being applied to study flow through fractured media.

Chapter 9 The work on extending the modified L-scheme to cover the two-phase case, non-equilibrium effects and the domain decomposition method, is ongoing. Other applications could include poromechanics [33] and surfactant transport [122].

11.3 Recommendations

To complete this work, we recommend a model that incorporates extended play-type hysteresis and dynamic capillarity in the capillary pressure and hysteresis in the relative permeabilities. The hope is to have a simple and computationally inexpensive
model that captures the behaviour of unsaturated/two-phase flow through porous media accurately. To this end, we propose:

**A general model for non-equilibrium effects.** To close equations (TP) and (Ri) (equations (1.13) and (1.20)), introduced in Chapter 1 for the two-phase and unsaturated flow cases respectively, we propose

\[
\partial_t(S_w + b(p_c)) = \mathcal{F}(S_w, p_c) = \frac{1}{\tau f(S_w)} \begin{cases} 
 p_c^{(i)}(S_w) - p_c & \text{when } p_c \leq p_c^{(i)}(S_w), \\
 0 & \text{when } p_c \in [p_c^{(i)}(S_w), p_c^{(d)}(S_w)], \\
 p_c^{(d)}(S_w) - p_c & \text{when } p_c \geq p_c^{(d)}(S_w).
\end{cases}
\]

(11.1)

with capillary pressure

\[ p_c = p_n - p_w, \]

and \( k_\alpha \in C([0,1] \times \mathbb{R}), \alpha \in \{n, w\} \) such that

\[
\begin{aligned}
 k_\alpha(S_w, p) &= k_\alpha^{(i)}(S_w) & \text{for } p < p_c^{(i)}(S_w), \\
 k_\alpha(S_w, p) &= k_\alpha^{(d)}(S_w) & \text{for } p > p_c^{(d)}(S_w),
\end{aligned}
\]

(11.2a)

and

\[
\begin{aligned}
 \partial_p k_w(S_w, p) > 0 & \quad \text{if } p \in [p_c^{(i)}(S_w), p_c^{(d)}(S_w)], \\
 \partial_p k_n(S_w, p) < 0 & \quad \text{if } p \in [p_c^{(i)}(S_w), p_c^{(d)}(S_w)].
\end{aligned}
\]

(11.2b)

Here, \( p_c^{(i)}, p_c^{(d)} : [0,1] \to \mathbb{R} \) are imbibition and drainage capillary pressure functions satisfying

(A1) \( p_c^{(i)}(1) = 0, \quad k_c^{(i)}(S) < 0 \) for \( j \in \{i, d\} \). Moreover, \( p_c^{(i)}(S) < p_c^{(d)}(S) \) for \( S \in (0,1) \).

The non-wetting and wetting relative permeabilities \( k_w^{(i)}, k_n^{(i)} : [0,1] \to \mathbb{R} \) for \( j \in \{i, d\} \) satisfy

(A2) \( k_w^{(i)}(1) = 0, \quad k_w^{(i)}(S) > 0 \) for \( 0 < S \leq 1 \), \( k_w^{(i)}(0) = 0 \) and \( k_w^{(i)} \) is strictly convex. Moreover, for \( 0 < S < 1 \), \( k_w^{(i)}(S) > k_w^{(d)}(S) \).

(A3) \( k_n^{(i)}(1) = 0, \quad k_n^{(i)}(S) < 0 \) for \( 0 < S < 1 \), \( k_n^{(i)}(1) = 0 \) and \( k_n \) is strictly convex. Moreover, for \( 0 < S < 1 \), \( k_n^{(d)}(S) < k_n^{(i)}(S) \).

The functions \( b, f \) and the dynamic capillarity coefficient \( \tau \) satisfy

(A4) \( b \in C^1(\mathbb{R}) \) with \( b(0) = 0; \lim_{S \to 0} b'(p_c^{(j)}(S))p_c^{(j)}(S) = \lim_{S \to 1} b'(p_c^{(j)}(S))p_c^{(j)}(S) < -1 \) and

\[ 0 < b'(p_c^{(j)}(S)) < -\frac{1}{p_c^{(j)}(S)} \text{ for all } 0 < S < 1 \text{ and } j \in \{i, d\}. \]
The model for capillary pressure (11.1) is the version of the extended play-type hysteresis model proposed in Chapter 3, whereas, the model for relative permeabilities (11.2) was proposed in Chapter 7. Assumptions (A1)-(A3) are taken from Chapter 7. The Assumption (A4) is required for consistency, as shown in Chapter 3, and (A5) is as analysed in Chapters 4 and 5. The relative permeabilities, capillary pressures, $b$, $f$ and $\tau$ can be fitted from experiments following procedures in [38,47,170,259] and Chapter 2.

The thermodynamical consistency of (11.1) follows from [24], particularly from equation (35). The existence of weak solutions for such a model can be shown following the arguments in Chapter 3. The results of Chapters 2 and 5 suggest that this model can give more accurate predictions compared to the combined play-type hysteresis and dynamic capillarity model given by (1.30). Moreover, the semi-implicit time discretization discussed in Chapters 2, 4 and 7 extends straightforwardly to this case along with the modified L-scheme proposed in Chapter 9. This shows that the computational cost associated with this model is not considerably greater compared to the standard model. We leave a detailed analysis for a future study.
A.1 Extended play-type hysteresis: non-vertical scanning curves

We rewrite the pressure expression (5.4) as

\[
\frac{\tau f(S)}{p_c^-(S)} \partial_t (S + \varepsilon u) + \text{sign}(\partial_t (S + \varepsilon u)) \times \frac{p_c^+(S) - u}{p_c^-(S)} + \frac{\varepsilon \tau f(S)}{p_c^-(S)} \partial_t u.
\]  

(a.1)

Introducing the function, for each \(0 < S < 1\),

\[
\Phi_s(r) = \begin{cases} 
   p_c^-(S) \tau f(S) (r - 1) & \text{for } r > 1, \\
   0 & \text{for } -1 \leq r \leq 1, \\
   p_c^-(S) \tau f(S) (r + 1) & \text{for } r < -1,
\end{cases}
\]

we deduce from (a.1)

\[
\partial_t (S + \varepsilon u) = \Phi_s \left( \frac{p_c^+(S) - u}{p_c^-(S)} + \frac{\varepsilon \tau f(S)}{p_c^-(S)} \partial_t u \right).
\]  

(a.2)

Evaluating the different cases in (a.2) one gets

\[
\partial_t S = \begin{cases} 
   \frac{p_c^{(d)}(S) - u}{\tau f(S)} & \text{when } u > p_c^{(d)}(S) + \varepsilon \tau f(S) \partial_t u, \\
   -\varepsilon \partial_t u & \text{when } p_c^{(i)}(S) + \varepsilon \tau f(S) \partial_t u \leq u \leq p_c^{(d)}(S) + \varepsilon \tau f(S) \partial_t u, \\
   \frac{p_c^{(i)}(S) - u}{\tau f(S)} & \text{when } u < p_c^{(i)}(S) + \varepsilon \tau f(S) \partial_t u.
\end{cases}
\]
Appendix B

PART II

B.1 Proof of Propositions 9.1 and 9.2

Proof of Proposition 9.1. We consider the L-scheme for (9.16). Let \( u^i_n \) be the solution of the \( i \)-th iteration of the L-scheme (9.9) initiated by \( u^0_n = u_{n-1} \). We use the properties of the sequence \( \{u^i_n\}_{i \in \mathbb{N}} \) to prove Proposition 9.1. From (9.9) and (9.13), \( u^i_n \) satisfies

\[
(Lu^i_n, \varphi) + \Delta t(\nabla u^i_n - 1, \nabla \varphi) = (Lu^i_{n-1}, \varphi) - (b(u^i_{n-1}) - b(u_{n-1}), \varphi) + \Delta t(\vec{F}, \nabla \varphi) + \Delta t(f^i_{n-1}, \varphi).
\]  

(b.1)

Let \( \rho^i \) denote the difference between consecutive iterates, i.e.

\[
\rho^i = u^{i+1}_n - u^i_n.
\]

Observe that for \( i = 0 \), \( \rho^0 \) solves the problem

\[
L(\rho^0, \varphi) + \Delta t(\nabla \rho^0, \nabla \varphi) = \Delta t(f(\vec{x}, t_n, u_{n-1}), \varphi) + (\vec{F}, \nabla \varphi) - (\nabla u_{n-1}, \nabla \varphi) + \nabla \cdot (\varphi u_{n-1}), \nabla \varphi). \]

As \( \nabla \cdot (\varphi u_{n-1}) \in L^\infty(\Omega) \) is assumed in Proposition 9.1, \( \rho^0 \) satisfies the equation

\[
L\rho^0 - \Delta t \nabla \cdot (\varphi \nabla \rho^0) = \Delta t([- \nabla \cdot \vec{F} + f(\vec{x}, t_n, u_{n-1}) + \nabla \cdot (\varphi \nabla u_{n-1})], \nabla \varphi).
\]

In the classical sense (see [89][Section 6.3]). As the term \( f(\vec{x}, t_n, u_{n-1}) - \nabla \cdot \vec{F} + \nabla \cdot (\varphi \nabla u_{n-1}) \) is bounded, the maximum principle applies [89][Section 6.4] and therefore a \( C_0 > 0 \), independent of \( \Delta t \), exists such that

\[
\|\rho^0\|_{L^\infty(\Omega)} = C_0 \Delta t.
\]

We claim that there exists a \( \beta \in (0, 1) \) such that

\[
\|\rho^k\|_{L^\infty} \leq \beta \|\rho^{k-1}\|_{L^\infty}.
\]  

(b.2)
The value of $\beta$ will be specified later. The proof of (b.2) is similar to the proof of Lemma 9.1.

Subtracting (b.1) for $i$ from the same equation written for $(i + 1)^{th}$, we get an expression similar to (9.17), i.e.

$$L(p^{i}, \phi) + \Delta t(\mathcal{D} \nabla p^{i}, \nabla \phi) = ((L - \partial u z(\zeta))p^{i-1}, \phi),$$

where $\zeta : \Omega \to \mathbb{R}$ is a function satisfying $\zeta \in \mathcal{F}(u_{n}^{i}, u_{n}^{i-1})$. Let $\mathcal{E}_{i-1} = \|p^{i-1}\|_{L^\infty(\Omega)} / \Delta t$. Taking $\phi = |p^{i} - \beta \mathcal{E}_{i-1} \Delta t|_{+}$ as test function in the above and using the bounds for $\mathcal{D}$ we get

$$L(p^{i}, |p^{i} - \beta \mathcal{E}_{i-1} \Delta t|_{+}) + \Delta t \mathcal{D} m \|\nabla |p^{i} - \beta \mathcal{E}_{i-1} \Delta t|_{+}\|^2 \leq \int_{\Omega} ((L - \partial u z(\zeta))p^{i-1}|p^{i} - \beta \mathcal{E}_{i-1} \Delta t|_{+},$$

which implies that

$$L\|p^{i} - \beta \mathcal{E}_{i-1} \Delta t\|^2 + \Delta t \mathcal{D} m \|\nabla |p^{i} - \beta \mathcal{E}_{i-1} \Delta t|_{+}\|^2 \leq \int_{\Omega} ((L - \partial u z(\zeta))p^{i-1}-L\beta \mathcal{E}_{i-1} \Delta t)|p^{i} - \beta \mathcal{E}_{i-1} \Delta t|_{+}. \tag{9.16}$$

With this, we analyse the sign of the expression $(L - \partial u z(\zeta))p^{i-1} - L\beta \mathcal{E}_{i-1} \Delta t$. If $p^{i-1} \leq 0$ then this is clearly negative. If $p^{i-1} > 0$, since $\|p^{i-1}\|_{L^\infty(\Omega)} = \mathcal{E}_{i-1} \Delta t$ and $0 \leq L - \partial u z(\zeta) \leq L - m$, see (9.6), by taking

$$\beta = (L - m) / L < 1,$$

one obtains $(L - \partial u z(\zeta))p^{i-1} - L\beta \mathcal{E}_{i-1} \Delta t \leq |(L - m) - L\beta| \mathcal{E}_{i-1} \Delta t = 0$. This shows that $p^{i} \leq \beta \mathcal{E}_{i-1} \Delta t$ a.e. The inequality $p^{i} \geq -\beta \mathcal{E}_{i-1} \Delta t$ follows in a similar fashion.

Observe that $\sum_{i=0}^{\infty} \|p^{i}\|_{L^\infty(\Omega)} \leq \mathcal{E}_{0} \Delta t \sum_{i=0}^{\infty} \beta^{i} = \mathcal{E}_{0} \Delta t / (1 - \beta)$ and as $p^{i} = u_{n}^{i+1} - u_{n}^{i}$, it implies that $u_{n}^{i} \to u_{n}$ in the $L^\infty(\Omega)$-norm as $i \to \infty$. Note that, $u_{n}^{i}$ also converges to $u_{n}$ in the $H^1(\Omega)$ norm as shown in [192] and due to the uniqueness of the weak solutions of (9.16) (see [181]), $u_{n}$ is the unique solution of (9.16). Finally, defining $\Lambda = \mathcal{E}_{0} / (1 - \beta)$ we get

$$\|u_{n} - u_{n-1}\|_{L^\infty(\Omega)} \leq \sum_{i=0}^{\infty} \|p^{i}\|_{L^\infty(\Omega)} \leq \Lambda \Delta t. \tag{9.16}$$

Rewriting (9.16) as

$$(\nabla \cdot (\mathcal{D} \nabla u_{n}), \phi) = \left( \frac{1}{\Delta t} (b(u_{n}) - b(u_{n-1})) + \nabla \cdot F - f_{n}, \phi \right),$$

we see right away that $\nabla \cdot (\mathcal{D} \nabla u_{n}) = \frac{1}{\Delta t} (b(u_{n}) - b(u_{n-1})) - f_{n} + \nabla \cdot F$ a.e. and as the terms on the right hand side are bounded in the $L^\infty(\Omega)$-norm, so is $\nabla \cdot (\mathcal{D} \nabla u_{n})$. □
Proof of Proposition 9.2. The proof is almost identical to the proof of Proposition 9.1. We subtract (9.16) from (9.17) and follow the steps of the proof of Proposition 9.1 to obtain that \[ \|u_n^i - u_n\|_{L^\infty(\Omega)} \leq \beta \|u_n^{i-1} - u_n\|_{L^\infty(\Omega)} \]
if
\[ \beta \geq \max \left\{ \frac{L_n^i - \partial_u z(\zeta)}{L_n^i}, \quad \zeta \in \mathcal{S}(u_n, u_n^{i-1}) \right\} . \]

Observe that \( L_n^i \geq m + \mathcal{M} \Delta t \). Moreover, from (9.18) we obtain, \( L_n^i - \partial_u z(\zeta) \leq 2\mathcal{M} \Delta t \) for \( \mathcal{M} \geq \mathcal{M}_0 \). Hence, \[ \frac{L_n^i - \partial_u z(\zeta)}{L_n^i} \leq \frac{2\mathcal{M} \Delta t}{m + 2\mathcal{M} \Delta t} \leq \frac{2\mathcal{M} \Delta t}{m} . \quad (b.3) \]

By defining, \[ \beta = \frac{2\mathcal{M} \Delta t}{m} , \quad (b.4) \]
we observe that for \( \Delta t < \frac{m}{2\mathcal{M}} \), \( \beta < 1 \) and \( \beta = \mathcal{O}(\Delta t) \).

\[ \square \]
Koondanibha Mitra was born on 13th December 1992 in Shibpur, West Bengal, India. After finishing his schooling from Belur High School in 2010, he pursued a dual Bachelor-Masters degree in Mechanical Engineering at IIT Kharagpur, India with additions in Physics. During this time he interned in companies like Whirlpool, CD Adapco and Hindalco, and started an entrepreneurial venture on portable diagnostic devices based on his Masters thesis project, winning several accolades. He graduated in August 2015 with an honourable mention from IIT Kharagpur.

Immediately after this, he joined Eindhoven University of Technology (TU/e) in the Netherlands as a doctoral candidate, funded by Royal Dutch Shell and the Netherlands Organisation for Scientific Research (NWO) through a CSER grant. Subsequently, his PhD project was converted into a joint PhD project between TU/e and Hasselt University the results of which are presented in this dissertation.
List of Publications

Publications based on Koondanibha Mitra's PhD are in chronological order:

Journals


Submitted


In preparation


Acknowledgments

“A dream you dream alone is only a dream. A dream you dream together is reality”

∼ Y. Ono

This is the story of how the incredible people I interacted with over the years shaped me in making my childhood dream a reality. If asked at an early age what I would like to be when I grow up, my answer was always: a scientist. The first great educator of my life was my grandfather who kindled in me an insatiable thirst for knowledge. The teachers and peers in my school inculcated within me a version of the Indian dream. During college, I almost lost interest in research for a couple of years, due to some unfortunate events. In the end however, my peers in the University and experiences in the industrial sector reignited my inner calling and after the first interview with my future supervisor, I jumped into the opportunity of doing a PhD under him.

My PhD can very easily be summarized as me following commandments passed down by the two Gods that I know of, my supervisor, Prof. Sorin Pop, being the first. It is hard to explain in words his magnanimity; always with a smiling face even after a hard day’s work, with his affectionate hand ever-extended. However, it is even harder to fathom the full extent of his scientific prowess, which spans from the minute details of computations to the deepest complexities of mathematical logic. From that first fated interview, to all the time we have spent on the desk doing mathematics, I have never been more grateful than to be your student. You are a friend, philosopher and guide and my sincerest apologies for all the times I have annoyed you. Best of wishes to you and your family.

If Sorin is my scientific father figure, then my scientific grandfather figure would surely be Prof. Hans van Duijn, both in the figurative and the literal sense. I was nervous in my first interactions with Hans, him being the former rector of the University.
However, by the end of my PhD, I communicated with him on a completely different level and I was blessed with his affection and trust. A perfectionist in every sense of the word and a scientific visionary, I have truly been inspired by the Hans van Duijn way of doing Applied Mathematics. I wish for you, your family and Feyenoord the best of times ahead.

This work would be incomplete without the contributions and influences of my other mentors and collaborators. First I would like to thank Prof. Barry Koren for his support. The dual degree would not have been possible without his help. Prof. Florin Radu aided me to his best and a lot of this work draws inspiration from his own work. Prof. Ben Schweizer is also a great patron and an incredible mathematician I had the pleasure to work with. He generously offered me a position in Dortmund and I am truly honoured for that. Xiulei was the first PhD student I interacted with scientifically and I am collaborating with him to this day despite being on different hemispheres. I wish the best for your family and your career. Many thanks to David (Seus) and Tobias for the effort they had put in to enrich this work. Best wishes for your careers too. Prof. Rainer Helmig has been a great inspiration and I am always mesmerized by his genial nature on the backdrop of monumental responsibilities. Special thanks to Prof. Christian Rohde and Prof. Elena El Behi-Gornostaeva for their support. This work also benefited greatly from our collaboration with Utrecht University, specifically with Luwen and Prof. Majid Hassanizadeh. I would like to thank my committee members Prof. Kees Vuik, Prof. Harald van Brummelen, Prof. Peter De Maesschalck and Prof. Jochen Schütz for agreeing to be in my committee and giving their valuable time to scrutinize this work. Furthermore, the project would not have existed without the financial and administrative support from the extraordinary people in Shell, NWO and Hasselt University.

Work however is only a part of what makes the workplace a CASA. I would like to thank Aurthur, Réne, Rien, Nhung and Xingang who started their PhDs with me and they made the journey a delightful experience. My early office mates, Upanshu and Saeed, helped me to get acquainted with the ups and downs of PhD life. Mikola and Jasper joined later as my office mates and had to withstand my constant blabbering. I enjoyed a lot but hope it was not too annoying for you two. Anastasia brought a lovely breeze of girl power to the office and I had the pleasure of taking a climbing and a Salsa course with her, while tagging along in a few excursions. Thank you for cheering me up at a stressful time of my life. In my first year of PhD, Deepesh helped me to socialize and I have since then tried to continue his tradition of barbecues. Giovanni (a.k.a Professor) along with Monica were an integral part of this highly nonlinear social life. Harshit, Anuj, Nitin, Bart, Melania, Behnaz, Lotte, Jim, Wim and the rest of CASA, all made my PhD life much more enjoyable. Prof. Mark Peletier was always enthusiastic and supportive, as was Prof. Oliver Tse. Also a special thanks to Prof. Georg Prokert, Prof. Jan ten Thije Boonkkamp and Prof. Michiel Hochstenbach for useful discussions. Finally, CASA would be incomplete without Enna, Marése, Jolijn and Diane, who put the extraordinary effort to make us feel at home.
My second casa has always been the CMAT group in Hasselt. We always travelled together, our projects being more intertwined, and every time I was in Belgium they made me feel like on a holiday. Soheli had been like a mother figure to me, sharing with me a bong connection. I wish nothing but great prospects to you, Mainul Da and Mahdi. Manuela always packed a punch and I dare say that she is the coolest woman I have ever met in my life. I shared my love for Metal and Mathematics with Stephan and we had great time travelling together. I would like to thank also Carina, Alex, Markus, Florian, Stefan, Hoang-An and Jeremy for their company and inputs. My experiences in going to conferences and workshops would not have been the same without the people from Bergen, specifically David (Landa), Davide, Max, Jakub, Wietse, Elyes, Ingeborg, Manuel, Jan, Kundan and the rest of the group. If there is such a thing as having too much fun, then it was with you guys.

The life part of the work-life balance was completely dominated by the awesome friend circle I made in the four years of my PhD. I met Anwesha before even coming to Netherlands and she has since then been a close friend and confidant. I am glad to be the third Musketeer, with you and Pranav, and we shared a deep connection together. Jeroen (a.k.a Mr. Bogers) for me has been the face of Eindhoven. In all the sports we played, trips we went to and events we took part in, I am truly grateful to have you as a friend. I had the courtesy to connect with Mr. Auke who is the nicest person I have ever met and will ever probably meet in this walk of life. A special mention has to go to my Band of Bros: Tarun, Stephen and Sai. Through the countless Friday nights we spent together, debating, playing music, headbanging, and just in general chilling out, I am really lucky to have like-minded individuals like you as friends, and thanks for your unrelenting support which made life ever so easier. I had great fun with Valliappan and Vishnu in all the trips we went together. Nikhil and Abhineet, made the first half of my life in Eindhoven extremely colourful. Shinde joined at the last stage of my PhD, but we shared the KGP tempo. This brings me to my friends from college, i.e., Vatsa, Kaushal, Pandit, Khoche, Prabhu and others, along with my friends in the Netherlands, Ambuj, Narsaria, Arjun. Also a special thanks to my guitar teacher Joska, and my band (Earthlings and Hola Viola) members for making my musical aspirations come true.

Finally, this soliloquy cannot end without the mention of the people who gave me, a boy from a small town in India, the strength to pursue his ambitions. I would first like to thank Payel Mon Mukherjee (buri) who has been my closest friend through thick and thin for over a decade now. She knows me better than I know myself. I miss greatly not being close to her and I only have the following lines to say

मम दुःखवर्णन, मम सप्तक स्तुपन, 
तुमि भरिबे सौरक्ष, निश्चिपली सम।

As I mentioned earlier, my grandparents, Dada and Didi, were a huge influence in my
life, and they taught me by their own hands almost everything I knew up to a certain age. I always feel within me their well wishes. My Uncle used to carry me to school on his bike in my kinder-garden years. Twenty years have passed but his affection towards me has not dropped a bit with my Aunt loving me just as much. They have also blessed me with the best brother and sister in the world, Riju and Titir. Those two along with my brother in Shibpur, Tatai, will always guide me back home even from the opposite corner of the world. The love I got from my mother’s side of the family, from Mama, Mami, BanuDidi, cannot be expressed by words. It pains me to think that my maternal grandfather is not alive to see this day. I would also like to thank the families of Kutti, Boropisi and Chotopisi for indulging me as a child. Our family friends, Ashish Kaka, Sachi Kaka, Jhintu Kaka, also had a huge influence on me growing up. In the end however, I will have to come back to my roots, to my Mother. I am her whole world and the sacrifices she made for me will only be interred with my bones. I will not dishonour you by saying thanks, nothing would have been possible without you. Forgive me for sometimes being in an irascible disposition and know that I love you. My Father has been the biggest inspiration in my life. A jack of all trades, he is a staunch incorruptible idealist who dedicated his life for the betterment of the people. My ultimate objective is to reach his level and fulfill his dreams. That is what pushes me forward to this day.

Koondi

Eindhoven, September 2019
Bibliography


