Numerical solution for subsurface reservoir simulation

BY

Kossi Etekpo

Thesis presented for the degree of Master of Science in the Department of Mathematics and Applied Mathematics

University of Cape Town

March 2017

Supervised by

Dr. Antoine Tambue[1] (Main supervisor)

Prof. Daya Reddy[2] (Co-supervisor)

[1]Department of Mathematics and Applied Mathematics at UCT, CERECAM and AIMS Research Centre

[2]Department of Mathematics and Applied Mathematics at UCT and CERECAM

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to attributed to the NRF.

Declaration

I know the meaning of plagiarism and declare that all of the work in this thesis, save for that which is properly acknowledged, is my own.

Signed by candidate

Acknowledgment

I hereby express my gratitudes to my two supervisors, Dr. Antoine Tambue and Prof. Daya Reddy, who devoted great time and attention to my work. Your several remarks and suggestions gave me more insights in my research.

I also express my sincere thanks to my two founders, the ***Robert Bosch Stiftung*** and the ***National Research Foundation*** for the financial support they gave during my work.

Finally, I express my deep gratitudes to my host institutions, namely the University of Cape Town (UCT), the Centre for Research in Computational and Applied Mechanics (CERECAM) and the African Institute for Mathematical Sciences Research Centre (AIMS-RC) for the opportunity of study and the facilities they offered me.

Contents

Abstract

Transport problems in porous media constitute an important field of scientific research in modern world, due to their broad applications in area such as petroleum engineering, water resources, pollutants transport and greenhouse gases sequestration to just mention few. The mathematical models that describe such problems have been developed and form one of the main classes of partial differential equations (PDEs) that scientists encounter in the real world modeling. Nevertheless, in most of the cases, the exact solutions in the classical sense of those models are not available. The study of numerical approximation of PDEs is therefore an active research area and there is an extensive literature on numerical methods for PDEs.

In this work, after providing the well-posedness results, we review some numerical techniques to approximate the model equations, more precisely we present finite volume method with two-point flux approximation and mixed finite volume method for spatial discretization of elliptic and parabolic PDEs modeling transport flow in porous media. We then present some standard explicit and implicit methods, Rosenbrock schemes and exponential time stepping schemes for temporal discretization. We finally run some numerical simulations of advection-diffusion-reaction problems in heterogeneous and anisotropic porous media.

Chapter 1

Introduction

Many physical processes can be described by mathematical models in the form of partial differential equations (PDEs). Nevertheless, to get a better knowledge of those problems and provide better solutions, the analysis of the PDEs has become an important field of research for both the engineers and the mathematicians.

The researches in petroleum engineering and the hydrology have significantly contributed to the search of the solutions to the subsurface flow problems. In recent years, the transport problems in porous media have become more attractive with some practical new challenges such as radioactive waste sequestration, geothermal energy recovery, greenhouse gases sequestration, transport of contaminant and many other more. In real situations, flow problems are often complex. On one hand, the flow can be single phase when it involves one single fluid (for example underground water flow), it can be multiphase when it involves more than one fluid (for example oil and gas), and it can be compressible or incompressible. On the other hand, the properties of the porous medium itself influences the flow. For example the medium can be anisotropic or isotropic, in the sense that the flow can be directional dependent or not. The transport of a physical phenomenon (like heat) or physical matter (like chemical substance) across a medium is not due only to the fluid flow (advection) but also to diffusion and reaction processes. Fortunately despite the complexity of the real world problems, there are some good mathematical models. The models in subsurface are described by three fundamental laws, the laws of conservation of mass, momentum and energy.

Assuming that there is not transport of energy, the Darcy's law (conservation of momentum) along with the mass conservation law yield a simple PDEs model that describe the flow of fluid in porous media. The resulting PDEs can be time dependent (for example parabolic PDEs) or time independent (for example elliptic PDEs)[1, 4]. Very often the classical solutions of those

PDEs are not known [4, 6, 10]. This lack of exact solutions leads to search for the so-called weak solutions or generalized solutions in some Sobolev spaces [10, 11]. Afterward, some numerical schemes are used to approximate those weak solutions.

This work is mainly about modeling of the transport process in porous media, the conditions of existence of the generalized solutions of the PDEs and some numerical methods that can be used to approximate the weak (or classical) solutions.

In Chapter 2, we give the definitions of some key concepts for modeling flow and transport problems. More precisely, we recall the empirical Darcy's law and the equation of the conservation of mass that allow to model single phase flow and transport equation in porous media. In Chapters 3 and 4, we recall some fundamental concepts on distributions and Sobolev spaces that are essential for the definition of weak solutions of the equations set up in chapter 2. Thereafter, we give the conditions that ensure the existence and uniqueness of the generalized solution. More precisely, we use standard variations formulation and the mixed formulation for the existence and uniqueness of the weak solution of the flow equation, while the semigroup approach is used for the existence and uniqueness of mild solution of the transport equation.

Chapters 5 and 6 are devoted to the numerical methods to approximate the weak solutions. For spatial discretization, we present the finite volume method with two-point flux approximations for both the flow equation and the transport equation. We also present the mixed finite element method for the flow equation. For the temporal discretization, we present a range of time integrators, especially some implicit methods, Rosenbrock methods and exponential methods.

In Chapter 7, we present some numerical simulations. More precisely, we run some advection-diffusion-reaction problems in heterogeneous and anisotropic porous media, along with some numerical convergence results. Finally we draw a conclusion in chapter 8.

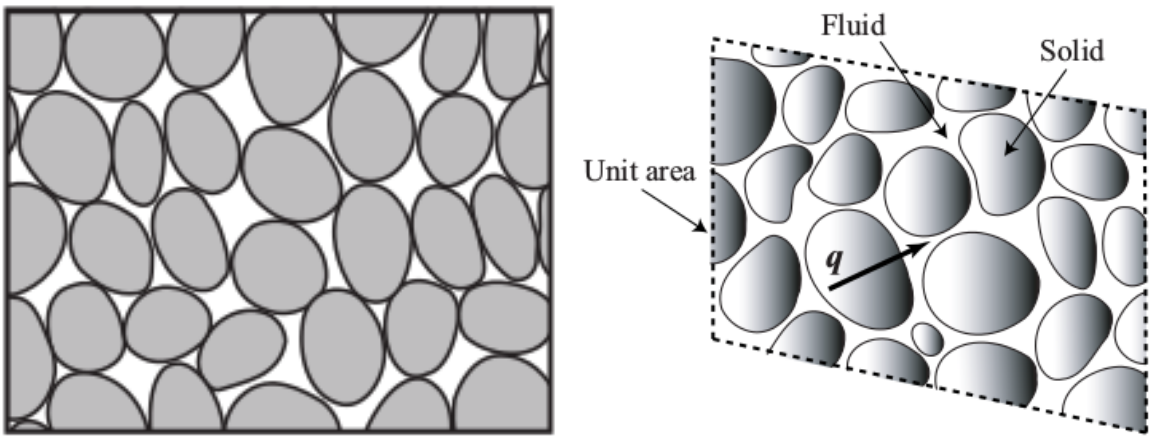
Chapter 2

Governing equations of flow and transport in porous media

In this chapter, we first give the definitions of some key concepts of porous media for a good understanding of the whole process of transport. We then formulate the governing equations for transport in porous media.

Definition (Porous media)**.**
A porous medium is medium that contains connected void spaces through which some entities can pass.

In reality every material is porous when working at the microscopic scale. In the area of fluid flow, we consider a medium as porous if the molecules of the fluid can flow through its void spaces.



(a) (b)

Figure 2.1: Cross section of a porous medium from [2]

Definition (Porosity).
The porosity of a medium is the ratio of its connected void spaces by its volume.

In this work, we denote the porosity by $\phi = \frac{\text{volume of void spaces}}{\text{total volume of the medium}}$. It is a non-dimensional quantity.

Definition (Permeability).
The permeability of a medium is its ability to allow a given material to pass through it.

It is intrinsically related to the property of porosity. In three dimensions, the permeability tensor is a 3×3 matrix:

$$\mathcal{K} = \begin{bmatrix} k_{xx} & k_{xy} & k_{xz} \\ k_{yx} & k_{yy} & k_{yz} \\ k_{zx} & k_{zy} & k_{zz} \end{bmatrix}.$$

In two dimensions it is a 2×2 matrix and in one dimension it is a scalar. In the sequel we will use the notation $\mathcal{K}$ for all dimensions.

Definition (Isotropy and anisotropy).
The isotropy is the property of being directional independent while the anisotropy is the property of being directional dependent.

In the later case, in the multi-dimensional problems, the permeability tensor is of the form $\mathcal{K} = k\mathbf{I}$ where k is a scalar and $\mathbf{I}$ is an identity matrix .

Definition (Conductivity).
The conductivity is the rate at which a given fluid flows through a medium per unit of area and time. It is related to the permeability tensor by

$$\mathbb{K} = \frac{\rho g}{\mu}\mathcal{K} \tag{2.1}$$

where ρ is the density of the fluid, μ is the viscosity of the fluid and g is the gravity acceleration. We notice that $\mathbb{K}$ is a scalar in one dimension.

2.1 Governing equations

2.1.1 Darcy's law

Henry Darcy (1803-1858) was a French engineer who made major contributions to hydraulics. He derived from experimentation an empirical law known as Darcy's law which constitutes one

of the fundamental laws of fluid flow in porous media. A modified version of Darcy's apparatus is shown in Figure 2.2

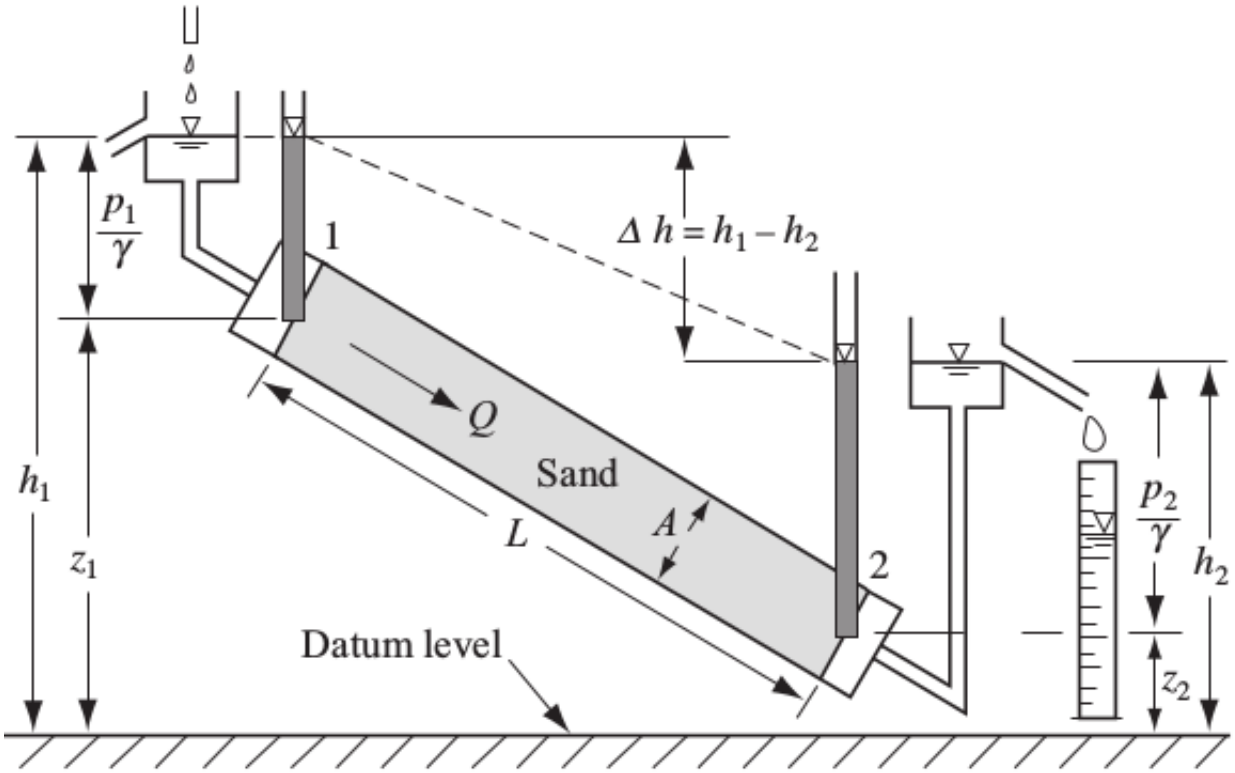


Figure 2.2: From [2]

Darcy's law in one-dimensional flow is as follows:

$$Q = \mathbb{K} \cdot A \frac{\triangle h}{L}, \tag{2.2}$$

where

- Q is the total discharge (volume of water passing per unit of time),
- A is the cross-sectional area,
- $\triangle h = h_1 - h_2$ is the difference of the piezometric head ($h = z + \frac{p}{\rho g}$, z is the altitude above datum level) and
- L is the length of sand column.

By dividing equation (2.2) by A and tending L to zero we get the so-called **equation of motion** in a porous medium for one dimensional flow:

$$\mathbf{u} = \frac{Q}{A} = -\mathbb{K}\nabla h = -\frac{\mathcal{K}}{\mu} \cdot (\nabla p + \rho g \nabla z) = -\frac{\mathcal{K}}{\mu} \cdot (\nabla p - \rho \mathbf{g}), \tag{2.3}$$

with $\mathbf{g} = -g\nabla z$. $\mathbf{u}$ is called **Darcy's velocity** or specific discharge and is related to the velocity $\mathbf{v}$ at which the fluid flows through the medium by $\mathbf{v} = \frac{\mathbf{u}}{\phi}$. In muti-dimensional problem, the equation of motion has the same structure $\mathbf{u} = -\frac{\mathcal{K}}{\mu}(\nabla p - \rho\mathbf{g})$.

2.1.2 Equation of conservation of mass

We derive the equation of conservation of mass by considering a unit volume called the **Representative Elemental Volume** (REV). The coordinates of the center of REV is (x, y, z) and those of its faces are $(x \pm \frac{\triangle x}{2}, y, z)$,$(x, y \pm \frac{\triangle y}{2}, z)$ and $(x, y, z \pm \frac{\triangle z}{2})$. We denote by $\triangle x = AB$, $\triangle y = BC$, and $\triangle z = AE$.

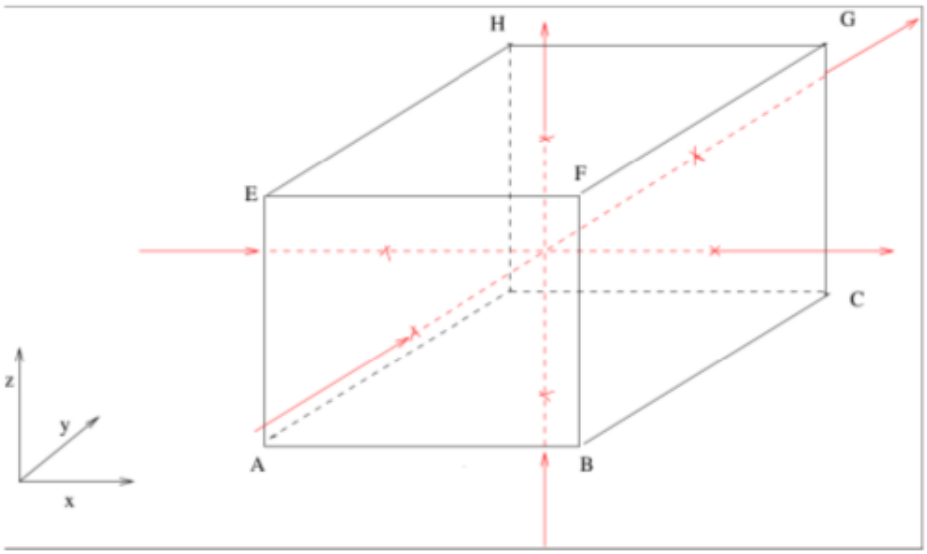


Figure 2.3: REV

Let $u_{(x-\frac{\triangle x}{2},y,z)}$, $u_{(x,y-\frac{\triangle y}{2},z)}$ and $u_{(x,y,z-\frac{\triangle z}{2})}$ be the incoming Darcy's velocity, and $J_{(x-\frac{\triangle x}{2},y,z)}$, $J_{(x,y-\frac{\triangle y}{2},z)}$ and $J_{(x,y,z-\frac{\triangle z}{2})}$ their corresponding mass flux per unit of time and area respectively. We have then

$$\begin{aligned} J_{(x-\frac{\triangle x}{2},y,z)} &= \rho u_{(x-\frac{\triangle x}{2},y,z)} \\ J_{(x,y-\frac{\triangle y}{2},z)} &= \rho u_{(x,y-\frac{\triangle y}{2},z)} \\ J_{(x,y,z-\frac{\triangle z}{2})} &= \rho u_{(x,y,z-\frac{\triangle z}{2})}. \end{aligned}$$

The total influx mass per unit of time is therefore

$$\begin{aligned} M_{in} &= J_{(x-\frac{\triangle x}{2},y,z)} \triangle y \triangle z + J_{(x,y-\frac{\triangle y}{2},z)} \triangle x \triangle z + J_{(x,y,z-\frac{\triangle z}{2})} \triangle x \triangle y \\ &= \rho u_{(x-\frac{\triangle x}{2},y,z)} \triangle y \triangle z + \rho u_{(x,y-\frac{\triangle y}{2},z)} \triangle x \triangle z + \rho u_{(x,y,z-\frac{\triangle z}{2})} \triangle x \triangle y. \end{aligned} \tag{2.4}$$

Similarly as above, by considering the outgoing Darcy's velocities and their corresponding mass fluxes, the outflux mass per unit of time is:

$$M_{out} = \rho u_{(x+\frac{\triangle x}{2},y,z)} \triangle y \triangle z + \rho u_{(x,y+\frac{\triangle y}{2},z)} \triangle x \triangle z + \rho u_{(x,y,z+\frac{\triangle z}{2})} \triangle x \triangle y. \tag{2.5}$$

We denote by Q' the distribution of source or sink density in REV; Q' is positive if the mass is produced and negative if the mass is destroyed. It generates a mass flux:

$$M_s = Q' \triangle x \triangle y \triangle z. \tag{2.6}$$

Summing up on equations (2.4,2.5,2.6), the equation of mass conservation in REV is:

$$\frac{\partial M}{\partial t} = M_{in} - M_{out} + M_s. \tag{2.7}$$

The volume of pores in the REV is $\phi \triangle x \triangle y \triangle z$ and the mass of the fluid in the pores is

$$M = \rho\phi \triangle x \triangle y \triangle z. \tag{2.8}$$

Considering equations (2.4 - 2.8) we have subsequently

$$\begin{aligned}\frac{\partial}{\partial t}(\rho\phi \triangle x \triangle y \triangle z) =& M_{in} - M_{out} + M_s \\ =& [\rho u_{(x-\frac{\triangle x}{2},y,z)} - \rho u_{(x+\frac{\triangle x}{2},y,z)}] \triangle y \triangle z \\ &+ [\rho u_{(x,y-\frac{\triangle y}{2},z)} - \rho u_{(x,y+\frac{\triangle y}{2},z)}] \triangle y \triangle z \\ &+ [\rho u_{(x,y,z-\frac{\triangle z}{2})} - \rho u_{(x,y,z+\frac{\triangle z}{2})}] \triangle x \triangle y \\ &+ Q' \triangle x \triangle y \triangle z. \end{aligned} \tag{2.9}$$

Dividing equation (2.9) by $\triangle x \triangle y \triangle z$ and taking the limit as $\triangle x, \quad \triangle y, \quad \triangle z \to 0$, we get:

$$\frac{\partial(\phi\rho)}{\partial t} = -\nabla.(\rho\mathbf{u}) + Q'. \tag{2.10}$$

We set

$$\frac{\partial(\phi\rho)}{\partial t} = \frac{\partial(\phi\rho)}{\partial p}\frac{\partial p}{\partial t} = \frac{S_s}{g}\frac{\partial p}{\partial t},$$

with $S_s = g\frac{\partial(\phi\rho)}{\partial p}$ called the **specific storage** (the volume of the fluid that can be stored by compressing the porous medium and the fluid itself). Its unit is $[m^{-1}]$.

Following closely [1] and assuming that ρ is independent of position we get the mass conservation equation in terms of pressure:

$$\frac{S_s}{\rho g}\frac{\partial p}{\partial t} = \nabla \cdot \left(\frac{\mathcal{K}}{\mu}(\nabla p - \rho\mathbf{g})\right) + \frac{Q'}{\rho}. \tag{2.11}$$

2.1.3 Transport equation

In this section we consider the transport in a porous medium of a given substance that we call solute or contaminant.

Transport of a substance involves generally the phenomena such as **diffusion** modelled by Fick's law, **convection** (advection) due to the motion of a fluid and **reaction**. If we denote by $\mathcal{C}$ the concentration, the convective flux is generally expressed by $\mathbf{j}_1 = \mathbf{u}\mathcal{C}$ and the diffusive flux by $\mathbf{j}_2 = -\boldsymbol{D}\nabla\mathcal{C}$, where $\mathbf{D}$ stands for diffusion tensor and $\mathbf{u}$ is the Darcy velocity. The total flux is therefore

$$\mathbf{j} = \mathbf{j}_1 + \mathbf{j}_2 = -\boldsymbol{D}\nabla\mathcal{C} + \mathbf{u}\mathcal{C}.$$

Transport processes can be dominated by convection or by diffusion and that phenomenon is determined by the so-called **Peclet number**. It is a dimensionless number and is the ratio of the rate of advection of a substance by the flow to the rate of diffusion of the same substance. In some problems as mentioned in [3], a basic transport model equation is formulated as: Find the concentration $\mathcal{C}$ of a matter (usually chemical reactive) such that

$$\partial_t h(\mathcal{C}) = -\nabla\cdot(\mathbf{j}_1 + \mathbf{j}_2) + Q(\mathcal{C}) = \nabla\cdot(\mathbf{D}\nabla\mathcal{C} - \mathbf{u}\mathcal{C}) + Q(\mathcal{C}). \tag{2.12}$$

Here $Q(C)$ is the source or the sink term of the contaminant and h the function of absorption of the contaminant by the medium.

In summary, transport problems in porous media lead to the following system:

$$\begin{cases} \dfrac{S_s}{\rho g}\dfrac{\partial p}{\partial t} = \nabla\cdot\left(\dfrac{\mathcal{K}}{\mu}(\nabla p - \rho\mathbf{g})\right) + \dfrac{Q'}{\rho}, \\ \qquad \mathbf{u} = -\dfrac{\mathcal{K}}{\mu}(\nabla p - \rho\mathbf{g}), \\ \partial_t h(\mathcal{C}) = \nabla\cdot(\mathbf{D}\nabla\mathcal{C} - \mathbf{u}\mathcal{C}) + Q(\mathcal{C}). \end{cases} \tag{2.13}$$

Here $p, \mathbf{u}$ and $\mathcal{C}$ are the unknowns.

Assuming that the fluid is incompressible and setting $\mathbf{K} = \dfrac{\mathcal{K}}{\mu}$, the system (2.13) is reduced to

$$-\nabla\cdot(\mathbf{K}(\nabla p - \rho\mathbf{g})) = \frac{Q'}{\rho}, \tag{2.14a}$$

$$\mathbf{u} = -\mathbf{K}(\nabla p - \rho\mathbf{g}) \tag{2.14b}$$

$$\partial_t h(\mathcal{C}) = \nabla\cdot(\mathbf{D}\nabla\mathcal{C} - \mathbf{u}\mathcal{C}) + Q(\mathcal{C}). \tag{2.14c}$$

Remark 2.1.1.

In the Darcy's experiment (water flow in a column of sand) there is no source or sink terms

therefore we have $\nabla.\mathbf{u} = -\nabla \cdot (\mathbf{K}(\nabla p - \rho \mathbf{g})) = 0$ as in fluid dynamic for incompressible flow. But in the system (2.14), we assume that the source or sink distribution which is $\frac{Q'}{\rho}$ is not null, which leads to (2.14a).

PDEs require in addition boundary conditions for stationary problems, and boundary conditions and the state at a given time usually at the initial time for evolutionary problems. Furthermore in some complex systems of PDEs, the problem can be split into subsystems of PDEs. Here we split (2.14) into two sub-problems along with their initial conditions.

Firstly we set up the stationary problem that we call the **pressure-velocity system**:

$$-\nabla \cdot (\mathbf{K}(\nabla p - \rho \mathbf{g})) = \frac{Q'}{\rho} = f, \tag{2.15a}$$

$$\mathbf{u} = -\mathbf{K}(\nabla p - \rho \mathbf{g}), \tag{2.15b}$$

$$p = f_1 \text{ on } \Gamma_D, \tag{2.15c}$$

$$\mathbf{K}\nabla p \cdot \nu = f_2 \text{ on } \Gamma_N, \tag{2.15d}$$

where Γ_D is the **Dirichlet boundary**, Γ_N is the **Neumann boundary**, Ω is the domain on which the problem is considered and f_1 and f_2 are real value functions on the boundaries of Ω. In some problems we can have mixed boundary condition in the form $\mathbf{K}\nabla p \cdot \nu + \alpha p = g_3$ also known as the **Robin boundary condition** . But we limit this work to the Dirichlet and Neumann boundary conditions and we have $\partial\Omega = \Gamma_D \cup \Gamma_N$ and $\Gamma_D \cap \Gamma_N = \varnothing$.

Secondly we set up the evolutionary problem

$$\partial_t h(\mathcal{C}) = \nabla \cdot (\mathbf{D}\nabla\mathcal{C} - \mathbf{u}\mathcal{C}) + Q(\mathcal{C}), \tag{2.16a}$$

$$\mathcal{C} = g_1 \text{ on } (0, T] \times \Gamma_D \tag{2.16b}$$

$$\mathbf{D}\nabla\mathcal{C}.\nu = g_2 \text{ on } (0, T] \times \Gamma_N, \tag{2.16c}$$

$$\mathcal{C}(0, x) = \mathcal{C}_0, \quad x \in \bar{\Omega}, \tag{2.16d}$$

where $[0, T]$ is the time interval and g_1 and g_2 are real value functions defined on the boundaries of Ω. The two sub-problems make up the full transport problem.

In general, classical solution to the problems described above are not available. The only alternative that remains to make those PDEs useful in the applications is to seek for approximate solutions of the weak forms of the PDEs. To address the weak formulations of the PDEs issue, the mathematical analysis of the functions involved in the equations is needed. The next chapter recalls some fundamentals on the relevant functions spaces.

Chapter 3

Functions Spaces

Sobolev spaces form the basis upon which the development of the PDEs' analysis has flourished. They provide a wide set of tools that allow to prove the existence and uniqueness of PDEs'solutions. In this chapter we give a review on the fundamentals of Sobolev spaces and some basics of distributions upon which Sobolev theory itself depends. More details can be found in [6, 7, 8, 9]. We introduce some notation that will be used throughout the rest of this work.

We set $\mathbb{N} = \{1, 2, ...\}$, $\mathbb{N}_0 = \{0, 1, 2, ...\}$. We denote by Ω a non-empty open subset in $\mathbb{R}^d$, $d \in \mathbb{N}$ and all functions on Ω are real value functions. For a function f defined on Ω, $\partial_{x_i} f = \dfrac{\partial f}{\partial x_i}$. For $\alpha = (\alpha_1, .., \alpha_d) \in \mathbb{N}_0^d$, $\partial_x^\alpha f = \dfrac{\partial^{|\alpha|} f}{\partial_{x_1}^{\alpha_1} ... \partial_{x_d}^{\alpha_d}}$, where $|\alpha| = \alpha_1 + ... + \alpha_d$ and $x = (x_1, .., x_d)$.

Definition (Vector space)**.**

Let $\mathbf{V}$ be non-empty set. The set $(\mathbf{V}, +, .)$ is a real vector space (vector space on $\mathbb{R}$) if the following conditions are met:

(i) $a + b \in \mathbf{V}$, $\forall a, b \in \mathbf{V}$ (closure),

(ii) $a + b = b + a$, $\forall a, b \in \mathbf{V}$ (commutativity),

(iii) $(a + b) + c = a + (b + c)$, $\forall a, b, c \in \mathbf{V}$ (associativity),

(iv) there exist a unique element $o_V \in \mathbf{V}$ called **identity** such that $o_V + a = a$, $\forall a \in \mathbf{V}$,

(v) for any $a \in \mathbf{V}$ there exists a unique element $-a \in \mathbf{V}$ called **inverse** of a such that $a + (-a) = o_V$,

(vi) $\lambda.a \in \mathbf{V}$, $\forall \lambda \in \mathbb{R}$, $a \in \mathbb{V}$,

(vii) $1.a = a$, $\forall a \in \mathbf{V}$,

(viii) $\lambda.(a+b) = \lambda.a + \lambda.b$, $\forall a, b \in \mathbf{V}$ and $\lambda \in \mathbb{R}$,

(viv) $(\lambda+\beta).a = \lambda.a + \beta.a$, $\forall \lambda, \beta \in \mathbb{R}$ and $a \in \mathbf{V}$,

(x) $(\lambda.\beta).a = \lambda.(\beta.a)$, $\forall \lambda, \beta \in \mathbb{R}$ and $a \in \mathbf{V}$.

Any element in $\mathbf{V}$ is called **vector** and any element in $\mathbb{R}$ is called **scalar**. We will denote $\lambda.a$ by λa and we will use the expression vector space or linear space instead of real vector space for simplicity. More details on vectors spaces can be found in [5, 6]

Definition (Norm and Semi-norm)**.**
A norm on a vector space $\mathbf{V}$ is any mapping $N : \mathbf{V} \to \mathbb{R}$ satisfying

(i) $N(\lambda a) = |\lambda| N(a)$, $\forall \lambda \in \mathbb{R}$ and $a \in \mathbf{V}$,

(ii) $N(a+b) \leq N(a) + N(b)$,

(iii) $N(a) = 0$ if only if $a = o_V$.

If N does not meet the last condition, then it is a semi-norm. A vector space endowed with a norm is called **normed vector space**. More comments on norms and semi-norms can be found in [4, 5, 6, 10, 12]

Definition (Linear and bilinear forms [4, 5, 6])**.**
Let $\mathbf{V}$ a be a normed vector space with norm $||.||$.

(i) A linear form on $\mathbf{V}$ is any mapping $P : \mathbf{V} \to \mathbb{R}$ satisfying
$P(\lambda a + \beta b) = \lambda P(a) + \beta P(b)$ for every $\lambda, \beta \in \mathbb{R}$ and $a, b \in \mathbf{V}$.
The mapping P is continuous if there exists a constant $c > 0$ such
$|P(a)| \leq c||a||$ for all $a \in \mathbf{V}$.

(ii) A bilinear form on $\mathbf{V}$ is any mapping $Q : \mathbf{V} \times \mathbf{V} \to \mathbb{R}$ which is linear on the first and the second arguments. Q is symmetric if $Q(a,b) = Q(b,a)$ for all $a, b \in \mathbf{V}$.
The bilinear form Q is continuous if there exists a constant $m > 0$ such that
$|Q(a,b)| \leq m||a||.||b||$ for all $a, b \in \mathbf{V}$.

Definition (Inner product)**.**
An inner product F on a vector space $\mathbf{V}$ is a symmetric bilinear form that is positive definite that is

(i) $F(a,a) \geq 0$ for any $a \in \mathbf{V}$,

(ii) $F(a,a) = 0$ if and only if $a = o_V$.

Definition (Banach and Hilbert spaces [4, 5, 6, 11]).
Let $\mathbf{V}$ be a vector space.

(i) $\mathbf{V}$ is a Banach space if it is normed and complete.

(ii) $\mathbf{V}$ is a Hilbert space if it is a Banach space and the norm on it is induced by an inner product.

Definition (Dual Space).
Let $\mathbf{V}$ be a vector space. The set of all linear forms on $\mathbf{V}$ called the **topological dual** of $\mathbf{V}$ is a Banach space denoted by $\mathbf{V}^*$ or $\mathbf{V}'$. See [10] for the proof that $\mathbf{V}'$ is a Banach space and comments on $\mathbf{V}'$.
Usually, for $a \in \mathbf{V}$ and $f \in \mathbf{V}'$, $f(a)$ is denoted by $\langle f, a\rangle_{\mathbf{V}',\mathbf{V}}$ and $\langle .,. \rangle_{\mathbf{V}',\mathbf{V}}$ is a bilinear form on $\mathbf{V}' \times \mathbf{V}$ called the **duality pairing** between $\mathbf{V}'$ and $\mathbf{V}$ (see [10, 6]). Sometimes, $\langle .,. \rangle_{\mathbf{V}',\mathbf{V}}$ is replaced by $\langle .,. \rangle$ for simplicity.
By means of **Riesz-Fréchet representation theorem** [10, Theorem 5.5] we can identify $\mathbf{V}'$ and $\mathbf{V}$.

Definition (Orthogonal of space).
Let $\mathbf{V}$ be a vector space and $\mathbf{M} \subset \mathbf{V}$ a subspace of $\mathbf{V}$. The orthogonal of $\mathbf{M}$ denoted by $\mathbf{M}^\perp$ is a vector space [10], and is defined as $\mathbf{M}^\perp = \{f \in \mathbf{V}' | \langle f, a\rangle = 0 \quad \forall a \in \mathbf{M}\}$.

Definition (Operator [18]).
Let $\mathbf{V}_1$ and $\mathbf{V}_2$ be vector spaces.

(i) Any mapping $\mathbf{O} : \mathbf{V}_1 \to \mathbf{V}_2$ is an operator.

(ii) The domain of $\mathbf{O}$ that we denote by $\mathcal{D}_m(\mathbf{O})$ is the set of elements in $\mathbf{V}_1$ on which the operator acts and we denote by $ker(\mathbf{O}) = \{a \in \mathcal{D}_m(O) \,|\, \mathbf{O}a = o_{V_2}\}$.

(iii) The operator $\mathbf{O}$ is linear if $\mathbf{O}(\lambda a + \beta b) = \lambda \mathbf{O}(a) + \beta \mathbf{O}(b)$ for all $\lambda, \beta \in \mathbb{R}$ and $a, b \in \mathbf{V}_1$.

(iv) The operator $\mathbf{O}$ is closed if $\mathbf{V}_1$ and $\mathbf{V}_2$ are normed spaces, and if for any sequence $(a_n)_{n\in\mathbb{N}}$ of elements in $\mathcal{D}_m(\mathbf{O})$ such that as $n \to \infty$, $a_n \to a$ and $\mathbf{O}a_n \to b$, it follows that $\mathbf{O}a = b$.

(v) The operator $\mathbf{O}$ is bounded (or continuous) if $\mathcal{D}_m(\mathbf{O}) = \mathbf{V}_1$, $\mathbf{V}_1$ and $\mathbf{V}_2$ are normed spaces and there exists a constant $M > 0$ such that $||\mathbf{O}(a)||_{V_2} \leq ||a||_{V_1}$ for all $a \in \mathcal{D}_m(\mathbf{O})$ where $||.||_{V_1}$ and $||.||_{V_2}$ are norms on $\mathbf{V}_1$ and $\mathbf{V}_2$ respectively.

Definition (Adjoint of operator)**.**

Let $\mathbf{V}_1$, $\mathbf{V}_2$ be Banach spaces, and $\mathbf{O} : \mathbf{V}_1 \to \mathbf{V}_2$ an operator. The adjoint of $\mathbf{O}$ denoted by $\mathbf{O}^*$ is the operator $\mathbf{O}^* : \mathbf{V}_2 \to \mathbf{V}_1$ defined by

$$\langle b, \mathbf{O}a\rangle_{\mathbf{V}_2',\mathbf{V}_2} = \langle \mathbf{O}^* b, a\rangle_{\mathbf{V}_1',\mathbf{V}_1} \quad \forall b \in \mathcal{D}_m(\mathbf{O}^*),\ a \in \mathcal{D}_m(\mathbf{O}).$$

Definition (Analytic semigroup [18, 40])**.**

Let $\triangle_\theta = \{z \in \mathbb{C}\,|\,|arg(z)| < \pi - \theta\}$ be a sector in the complex plane containing the positive real axis, with $0 < \theta < \pi/2$.

A family $\{S(z),\ z \in \triangle_\theta\}$ is an analytic semigroup in $\triangle_\theta$ if

(i) $S(z)$ defines a bounded linear operator from $\mathbf{V}$ to $\mathbf{V}$ for all $z \in \triangle_\theta$.

(ii) The mapping $z \longrightarrow S(z)$ is analytic in $\triangle_\theta$.

(iii) $S(0) = I$, and $\lim\limits_{z\to 0} S(z)a = a$ for every $a \in \mathbf{V}$, I being the identity operator.

(iv) $S(z_1 + z_2) = S(z_1)S(z_2)$ for all $z_1, z_2 \in \triangle_\theta$.

Definition (Generator of analytic semigroup [18])**.**

A linear operator $\mathbf{P}$ is an infinitesimal generator of an analytic semigroup $\mathbf{O}$ operating on $\mathbf{V}$ if the domain of $\mathbf{P}$ is

$$\mathcal{D}_m(\mathbf{P}) = \left\{ x \in \mathbf{V} |\ \lim_{\substack{t\to 0 \\ t\in\mathbb{R}}} \frac{\mathbf{O}(t)x - x}{t} \text{ exixts} \right\}$$

and

$$\mathbf{P}x = \lim_{\substack{t\to 0 \\ t\in\mathbb{R}}} \frac{\mathbf{O}(t)x - x}{t}.$$

3.1 Basics on distributions

Distributions are related to the regularity concept of functions defined on a given domain Ω. The notion of regularity in the distributional sense is different from the classical regularity (pointwise regularity) in differential calculus. Here we present some key definitions.

Definition 3.1.1.

Let $k \in \mathbb{N}_0$. A function $f \in C^k(\Omega)$ if one of the following conditions holds;

(i) $k = 0$ and f is continuous on Ω.

(ii) $k \geq 1$ and f has partial derivatives $\frac{\partial f}{\partial x_i}$, $i = 1, .., d$ in $C^{k-1}(\Omega)$.

Moreover $f \in C^\infty(\Omega)$ if $f \in C^k(\Omega)$ for all $k \in \mathbb{N}_0$. Here in this definition, the differentiability is pointwise.

Definition 3.1.2 (Support of function and test function [12])**.**

(i) The support of a function f denoted by $\mathbf{supp}\, f$ is defined by $\mathbf{supp}\, f = \overline{\{x \in \Omega |\, f(x) \neq 0\}}$.

(ii) A test function is a function in $C^\infty(\Omega)$ having compact support. The set of all test functions on Ω is a vector space and is denoted by $\mathcal{D}(\Omega)$. We recall that compact sets on $\mathbb{R}^d$ are bounded and closed.

Definition 3.1.3 (Convergence [7, 8, 9])**.**
The sequence $(u_n)_{n\in\mathbb{N}}$ of elements in $\mathcal{D}(\Omega)$ converges to $u \in \mathcal{D}(\Omega)$ if the following conditions hold:

(i) There is a compact $K \subset \Omega$ such that $\mathbf{supp}\, u_n \subset K\ \forall n \in \mathbb{N}$.

(ii) For all $\alpha \in \mathbb{N}^d$, $\lim\limits_{n\to\infty} ||\partial_x^\alpha (u_n - u)||_\infty = 0$.

The norm $||.||_\infty$ is defined by $||u||_\infty = ||u||_{L^\infty(\Omega)} = \inf\{c \geq 0 \,|\, |u(x)| \leq c \text{ a.e}\}$ and is called the ***uniform norm***.

Definition 3.1.4 (Distribution [7, 8])**.**
A distribution $\mathcal{T}$ on $\mathcal{D}(\Omega)$ is a linear map from $\mathcal{D}(\Omega)$ to $\mathbb{R}$ that meets one of the following conditions:

(i) For every sequence $(u_n)_{n\in\mathbb{N}}$ in $\mathcal{D}(\Omega)$ converging to zero with respect to the norm $||.||_\infty$, $\mathcal{T}(u_n)$ converges to zero.

(ii) For every compact $K \subset \Omega$, there are some constant c and integer p such that for every $\theta \in \mathcal{D}(\Omega)$, with $\mathbf{supp}\, \theta \subset K$ we have $|\mathcal{T}(\theta)| \leq c \sum\limits_{|\alpha|\leq p} ||\partial_x^\alpha \theta||_\infty$.

We denote by $\mathcal{D}'(\Omega)$ the set of all distributions on $\mathcal{D}(\Omega)$ and $\mathcal{D}'(\Omega)$ is a vector space. The notation $\langle \mathcal{T}, \theta \rangle$ is commonly used rather than $\mathcal{T}(\theta)$.

Theorem 3.1.1.

Every locally integrable function f on Ω defines a distribution T_f by setting

$$\langle T_f, \theta \rangle = \int f\theta dx, \qquad \forall \theta \in \mathcal{D}(\Omega).$$

See [8] for the proof of this theorem.

Corollary 3.1.1.

Every continuous function f defines a distribution T_f on $\mathcal{D}(\Omega)$. Moreover if $\langle T_f, \theta\rangle = 0\ \forall\theta \in \mathcal{D}(\Omega)$, then $f = 0$ almost everywhere [6, 8].

Definition 3.1.5 (Derivatives of distributions)**.**

Let $\mathcal{T} \in \mathcal{D}'(\Omega)$ and $\theta \in \mathcal{D}(\Omega)$. The partial derivative of $\mathcal{T}$ respect to x_i denoted by $\partial_{x_i}\mathcal{T}$ is defined by $\langle \partial_{x_i}\mathcal{T}, \theta\rangle = -\langle \mathcal{T}, \partial_{x_i}\theta\rangle$. This definition is a direct consequence of integration by part. More generally for $\alpha \in \mathbb{N}^d$, $\partial_x^\alpha \mathcal{T}$ is defined by $\langle \partial_x^\alpha \mathcal{T}, \theta\rangle = (-1)^{|\alpha|}\langle \mathcal{T}, \partial_x^\alpha\theta\rangle$.

Definition 3.1.6 (Derivatives of functions in distributional sense)**.**

Let f be a function on Ω.The function f is differentiable in the distributional sense with respect to x_i denoted by $\partial_{x_i} f$ if f defines a distribution and then we have $\langle \partial_{x_i} f, \theta\rangle = -\langle f, \partial_{x_i}\theta\rangle$. $\partial_{x_i} f$ is called ***weak partial derivative*** or ***generalized partial derivative*** of first order with respect to x_i. We have the following corollary from this definition

Corollary 3.1.2.

(i) Let f be a real function on Ω. If f is weakly differentiable with respect to x_i to order $k \geq 1$, then it is infinitely weakly differentiable at any order with respect to x_i (see [7] for more information).

(ii) If $f \in C^k(\Omega)$ then the weak derivatives $\partial_x^\alpha f$ coincide with the classical derivation for $|\alpha| \leq k$ (see [7, 8] for the proof and more comment).

Definition 3.1.7 (Multiplication of distribution by functions)**.**

Let $\mathcal{T}$ be a distribution. The multiplication of $\mathcal{T}$ by a function $\theta \in C^\infty(\Omega)$ is the distribution $\theta\mathcal{T}$ defined by $\langle \theta\mathcal{T}, \Phi\rangle = \langle \mathcal{T}, \theta\Phi\rangle$. Indeed $\theta\Phi$ is a test function since we have $\mathbf{supp}\,\theta\Phi \subset \mathbf{supp}\,\theta$.

3.2 Fundamentals on $W^{m,p}(\Omega)$ spaces

The solutions of PDEs are commonly seeked in the Sobolev spaces. We give the general definition of those spaces before we focus on a particular space in which we are interested. But first all we introduce the $L^p(\Omega)$ spaces. Details can be found in [6, 10, 11, 12]

Definition 3.2.1.

Let $1 \leq p \leq \infty$. The set of all functions f such that $\displaystyle\int_\Omega |f|^p dx < \infty$ if $p < \infty$ is denoted by $L^p(\Omega)$, and the set of essentially bounded functions is denoted by $L^\infty(\Omega)$, that is

$$L^\infty(\Omega) = \{f : \Omega \to \mathbb{R} | \sup_{x\in\Omega} f(x) < \infty \text{ a.e}\}.$$

Indeed f is regarded as a representative of class of functions that are equal to f almost everywhere (see [12], Chapter 2).

Theorem 3.2.1.

$L^p(\Omega)$ is a vector space. Moreover endowed with the norm

$$||f||_{L^p(\Omega)} = \left(\int_\Omega |f|^p dx\right)^{1/p} \text{ if } p < \infty,$$

$$||f||_{L^\infty(\Omega)} = \inf\{c \geq 0 \,|\, |f(x)| \leq c \text{ a.e}\} \text{ if } p = \infty,$$

$L^p(\Omega)$ is a Banach space .

See [10] for the proof and more comment on this theorem.

Corollary 3.2.1.

The space $L^2(\Omega)$ endowed with the scalar product defined on $L^2(\Omega)' \times L^2(\Omega)$ by

$$\langle f, g\rangle_0 = \int_\Omega fg dx$$

for all $f, g \in L^2(\Omega)$ is a Hilbert space. Indeed we have for all $f \in L^2(\Omega)$

$$\langle f, f\rangle_0 = \int_\Omega f^2 dx = ||f||_0^2.$$

So here we identify $L^2(\Omega)'$ and $L^2(\Omega)$. See definition of dual space above at the beginning of this Chapter.

Definition 3.2.2 (Sobolev spaces)**.**

The sobolev spaces $W^{m,p}(\Omega) = W_p^m(\Omega)$ where $m \in \mathbb{N}_0, 1 \leq p \leq \infty$ are defined by

$$W^{m,p} = \{u \in L^p(\Omega) |\, \partial_x^\alpha u \in L^p(\Omega), |\alpha| \leq m\} \text{ if } p < \infty,$$

$$W^{m,\infty} = \{u \in L^\infty(\Omega) |\, \partial_x^\alpha u \in L^\infty(\Omega), |\alpha| \leq m\}.$$

The differentiations ∂_x^α are understood in the distributional sense. We denote by $H^m(\Omega) = W^{m,2}$ and these spaces are of great interest in PDEs problems. Hence it is obvious that $H^0(\Omega) = W^{0,2}(\Omega) = L^2(\Omega)$.

Theorem 3.2.2.

The Sobolev space $W^{m,p}$ endowed with the norm

$$||.||_{W^{m,p}(\Omega)} = ||.||_{m,p} = \Big(\sum_{|\alpha|\leq m} \int_\Omega |\partial_x^\alpha .|^p dx\Big)^{1/p} \text{ if } p < \infty,$$

$$||u||_{W^{m,\infty}(\Omega)} = ||u||_{m,\infty} = \max_{|\alpha|\leq m}\{||\partial_x^\alpha u||_{L^\infty(\Omega)}\}$$

is a Banach space. Moreover if Ω is a Lipschitz domain then $C^\infty(\bar{\Omega})$ is dense in the space with respect to the above norm.

A Lipschitz domain is a bounded domain with some regularity on the boundary. See [4, 6, 12, 17] for the mathematical definition of Lipschitz domain and the proof of this theorem. For $p = 2$ we simply denote the norm $||.||_{m,p}$ by $||.||_m$. For instance we have

$$||.||_{H^0(\Omega)} = ||.||_{L^2(\Omega)} = ||.||_0.$$

Corollary 3.2.2.
The space $H^m(\Omega)$ endowed with the scalar product

$$\langle u, v\rangle_m = \int_\Omega \sum_{|\alpha|\leq m} \partial_x^\alpha u \partial_x^\alpha v dx$$

is a Hilbert space. We notice that for $p \neq 2$ we can not define a scalar product on $W^{m,p}(\Omega)$ [4].

Definition 3.2.3 (Embedding [6, 40])**.**
Let X and Y be vector spaces with $||.||_X$, $||.||_Y$ their respective norms and $X \subset Y$. X is continuously embedded in Y and we write $X \hookrightarrow Y$ if there is $c > 0$ such that $||u||_Y \leq c||u||_X$ for all $u \in X$.

Example 3.2.1. We have $(H^2(\Omega), ||.||_2) \hookrightarrow (H^1(\Omega), ||.||_1)$.
Indeed by the definitions of $H^2(\Omega)$ and $H^1(\Omega)$ we have

$$H^2(\Omega) \subset H^1(\Omega).$$

Moreover by the definition of the norms $||.||_m$ we have for all $u \in H^2(\Omega)$,

$$||u||_1 = \Big(\sum_{\alpha\leq 1} ||\partial_x^\alpha u||_0\Big)^{1/2} \leq \Big(\sum_{\alpha\leq 2} ||\partial_x^\alpha u||_{0,}\Big)^{1/2} = ||u||_2.$$

In general we have $H^p(\Omega) \hookrightarrow H^q(\Omega)$ for $p \geq q$.

Theorem 3.2.3 (Sobolev Embedding)**.**
For a lipschitz domain Ω, the space $H^m(\Omega)$ is continuously embedded in $C^k(\Omega)$ where $m > k + d/2$ (see [6, 12] for the proof).

This theorem is important to know under which conditions a weak solution of a PDE chosen in H^m is regular in the classical sense.

Theorem 3.2.4 (Trace theorem)**.**
Let Ω be a bounded Lipschitz domain. There is a unique linear and continuous operator

$$\gamma_0 : \Big(H^1(\Omega), ||.||_{H^1(\Omega)}\Big) \longrightarrow \Big(L^2(\partial\Omega), ||.||_{L^2(\Gamma)}\Big).,$$

that is for any $v \in H^1(\Omega)$. $||\gamma_0 v||_{L^2(\Gamma)} \leq c||v||_1$ for some constant $c > 0$.

See [4, 6, 12] for the proof. The operator γ_0 is called the **trace operator** and $\gamma_0 v$ is called the **trace** of v (the boundary value of v) for all $v \in H^1(\Omega)$. We denote by
$H_0^1(\Omega) = \{v \in H^1(\Omega) |\, \gamma_0(v) = 0\} = ker(\gamma_0)$.

Definition 3.2.4 (***$H^{1/2}(\partial\Omega)$*** and ***$H^{3/2}(\partial\Omega)$***)**.**
The space $H^{1/2}(\partial\Omega)$ is the range of the trace operator defined in the theorem (3.2.4) that is

$$H^{1/2}(\partial\Omega) = \gamma_0(H^1(\Omega)),$$

and is endowed with the norm

$$||u||_{H^{1/2}(\partial\Omega)} = ||u||_{1/2,\partial\Omega} = \inf_{v \in \gamma_0^{-1}(\{u\})} ||v||_{H^1(\Omega)}.$$

The space $H^{1/2}(\partial\Omega)$ is a proper dense space of $L^2(\partial\Omega)$. Its dual is $H^{-1/2}(\partial\Omega)$ with the dual norm

$$||u^*||_{H^{-1/2}(\partial\Omega)} = ||u^*||_{-1/2,\partial\Omega} = \sup_{u \in H^{1/2}(\partial\Omega)} \frac{\ll u^*, u \gg}{||u||_{1/2,\partial\Omega}},$$

where $\ll ., . \gg$ is the duality pairing between $H^{1/2}(\partial\Omega)$ and $H^{-1/2}(\partial\Omega)$. Likewise as $H^{1/2}(\partial\Omega)$, with sufficient regularity condition on $\partial\Omega$ we define the space $H^{3/2}(\partial\Omega)$ to be

$$H^{3/2}(\partial\Omega) = \gamma_0(H^2(\Omega)).$$

The spaces $H^2(\Omega), H^1(\Omega), H^{1/2}(\partial\Omega)$ and $H^{3/2}(\partial\Omega)$ will allow us to define the **normal trace** (the normal derivative).

Theorem 3.2.5.
Let Ω be a bounded Lipschitz domain. The outer unit normal vector $\nu = (\nu_i)_{i,=1,\ldots,d}$ is defined almost everywhere and $\nu_i \in L^\infty(\Gamma)$ (see [4]).

Definition 3.2.5. Let $u \in H^2(\Omega)$ where Ω is assumed to be a bounded Lipschitz domain. We have $\nabla u \in (H^1(\Omega))^d$ and we define the **normal derivative** of u, denoted by $\partial_\nu u$ to be the function

$$\partial_\nu u := \sum_{i=1}^{d} \partial_{x_i} u|_\Gamma \nu_i.$$

Corollary 3.2.3.
Let Ω be bounded Lipschitz domain.

(i) For $u \in H^1(\Omega)$ and $q \in (H^1(\Omega))^d$, we have

$$\int_\Omega q \cdot \nabla u \, dx = -\int_\Omega \nabla \cdot qu \, dx + \int_{\partial\Omega} q \cdot \nu u \, d\sigma.$$

(ii) For $u \in H^2(\Omega)$ and $v \in H^1(\Omega)$ we have

$$\int_\Omega \nabla u \cdot \nabla v \, dx = -\int_\Omega \Delta u v \, dx + \int_{\partial\Omega} \partial_\nu u v \, d\sigma.$$

(iii) For $u \in H^2(\Omega)$, $v \in H^1(\Omega)$ and $k_{ij} \in W^{1,\infty}(\Omega)$ where $(k_{ij})_{1\leq i,j\leq d} = K$ we have

$$\int_\Omega K\nabla u \cdot \nabla v \, dx = -\int_\Omega \nabla.(K\nabla u) v dx + \int_{\partial\Omega} K\nabla u \cdot \nu v \, d\sigma.$$

See [4] for the proof.

Theorem 3.2.6 (Lax-Milgram [4, 12, 15])**.**
let $\mathcal{A} : V \times V \longrightarrow \mathbb{R}$ be a bilinear form not necessarily symmetric, $\mathcal{L} : V \longrightarrow \mathbb{R}$ a linear form. If :

(i) $\mathcal{A}$ is continuous, that is, there exists $M > 0$ such that $|\mathcal{A}(v,u)| \leq M||u||_V||v||_V$,

(ii) $\mathcal{A}$ is $V-$elliptic, that is, there exists $\alpha > 0$ such that $A(u,u) \geq \alpha||u||_V^2$, and

(iii) $\mathcal{L}$ is continuous, that is, there exists $m > 0$ such that $|\mathcal{L}(u)| \leq m||u||_V$,

Then there is a unique solution to the problem: Find $u \in V$ such that

$$\mathcal{A}(u,v) = \mathcal{L}(v)$$

for all $v \in V$. Furthermore we have

$$||u||_V \leq \frac{1}{\alpha}||\mathcal{L}||_{V'}$$

.

3.3 $H(\mathbf{div}, \Omega)$ space

The spaces we set up in the previous section in this chapter are suitable for **standard variational formulation** that we will develop in the next chapter. In some problems, where there are two dependent unknowns to find, the so-called **mixed formulation** appears to be advantageous in some ways and this leads to introduce some new spaces.

Definition 3.3.1.

(i) The space $H(\mathrm{div}, \Omega)$ is defined by $H(\mathrm{div}, \Omega) = \{q \in (L^2(\Omega))^d \,|\, \mathrm{div} q \in L^2(\Omega)\}$.

(ii) The space $H_{0,N}(\mathrm{div}, \Omega)$ is defined by $H(\mathrm{div}, \Omega) = \{q \in H(\mathrm{div}, \Omega) \,|\, q \cdot \nu = 0 \text{ on } \Gamma_N\}$.

Theorem 3.3.1.

The space $H(div,\Omega)$ endowed with the norm

$$||q||_{div} = \left(||q||_0^2 + ||\nabla.q||_0^2\right)^{1/2}, \forall q \in H(div,\Omega)$$

is a Hilbert space.

For the proof see [17]

Corollary 3.3.1.

For any $q \in H(\text{div},\Omega)$, $||q||_0 \leq ||q||_{\text{div}}$ that is, the norm $||.||_{\text{div}}$ is stronger than $||.||_0$ in the space $H(\text{div},\Omega)$

3.4 $L^2((0,T),\mathbf{V})$ space

The spaces $W^{m,p}(\Omega)$ and $H(\text{div},\Omega)$ we defined earlier are suitable for time-independent problems. For time-dependent problems, we need to define new function spaces.

Definition 3.4.1. [4]

Let $\mathbf{V}$ be a Banach space. The space $L^2((0,T),\mathbf{V})$ is the set of real valued function v on $(0,T)\times\Omega$ such that

(i) $v(t,.) \in \mathbf{V}$, $\forall t \in (0,T)$

(ii) The mapping $f_v : (0,T) \longrightarrow \mathbb{R}$ defined by
$f_v(t) = ||v(t,.)||_{\mathbf{V}}$ is such that, $f_v \in L^2((0,T))$.

We define on $L^2((0,T),\mathbf{V})$ the norm

$$||v||_{L^2((0,T),\mathbf{V})} = ||f_v||_{L^2((0,T))},$$

Definition 3.4.2. [4]

A function $u \in L^2((0,T),V)$ is said to have a **weak derivative** w if the following holds:

$$\int_0^T u(t)\frac{d}{dt}\psi(t)dt = -\int_0^T w(t)\psi(t)dt \quad \psi \in \mathcal{D}((0,T)).$$

We will also need some key inequalities for the next Chapter.

Definition 3.4.3 (Gårding's inequality)**.**

Let $\mathbf{H}$ and $\mathbf{V}$ be two Hilbert spaces such that $\mathbf{V} \hookrightarrow \mathbf{H}$. Identifying $\mathbf{H}$ and $\mathbf{H}'$ so that $\mathbf{V} \subset \mathbf{H} = \mathbf{H}' \subset \mathbf{V}'$ and the pairing duality $\langle .,. \rangle_{\mathbf{V}',\mathbf{V}}$ between $\mathbf{V}'$ and $\mathbf{V}$ coincides on $\mathbf{H}\times\mathbf{V}$ with the

scalar product (or inner product) and defining an operator $\mathbf{O_w} : \mathbf{V} \to \mathbf{V}'$ with its associated bilinear form $\mathcal{A} = \mathcal{A}(\mathbf{w}; ., .) : \mathbf{V} \times \mathbf{V} \to \mathbb{R}$ so that

$$\mathcal{A}(\mathbf{w}; \mathbf{v}, \mathbf{u}) = \langle \mathbf{O_w v}, \mathbf{u} \rangle_{\mathbf{V}', \mathbf{V}}, \quad \forall \mathbf{v}, \mathbf{u} \in \mathbf{V},$$

the Gårding's inequality for $\mathbf{v} \in \mathbf{V}$ is defined by

$$\mathcal{A}(\mathbf{w}; \mathbf{v}, \mathbf{v}) \geqslant \alpha ||\mathbf{v}||_{\mathbf{V}} - \beta ||\mathbf{v}||_{\mathbf{H}}, \tag{3.1}$$

for some constants α and β.

Some comments on this definition can be found in [26, 40].

Definition 3.4.4 (Young's inequality [13])**.**
Let a, b, p and q be positive real numbers such that $\frac{1}{p} + \frac{1}{q} = 1$, we have the inequality

$$ab \leqslant \frac{a^p}{p} + \frac{b^q}{q}. \tag{3.2}$$

Chapter 4

Variational formulations and well posedness of equations occurring in transport problems

Several phenomena comprise the transport processes in porous media. They include the diffusion described by Ficks' law, advection, described by Darcy's law, and reaction due the destruction or the production of mass of the substance under study. All of these physical phenomena are modelled by PDEs as shown in Chapter 2. Before looking for the solutions of these PDEs, we first have to investigate whether they have solutions . This investigation leads to the concepts of well posedness and the variational formulations of a PDE.

A problem is said to be well-posed if :

- it has a solution,
- the solution is unique, and
- the solution depends continuously on the given data (initial condition).

The well posedness does includes information about whether the solution is **analytic** (infinitely differentiable), **classical** (continuously differentiable at least at the order of the PDE) or a generalized solution (a solution with less regularity conditions). In this work the well posedness is intended to find weak solutions which need the concept of a variational formulations.

4.1 Standard variational formulation of the elliptic boundary problem

Let us consider the elliptic problem (2.15a)-(2.15d). In the generalized solution view, the weak solution $\tilde{p}$, if it does exist, has to fulfill the conditions: $\tilde{p} \in H^2(\Omega) \cap \mathbf{V}$ where $\mathbf{V}$ is a space that will be defined later, $f \in L^2(\Omega)$, $K_{i,j} \in W^{1,\infty}(\Omega)$ and there are $F_1 \in H^1(\Omega)$, $F_2 \in H^2(\Omega)$ such that $F_1|_{\partial\Omega_D} = f_1$ and $F_2|_{\partial\Omega_N} = f_2$. Unfortunately $H^{3/2}(\partial\Omega) \subset H^{1/2}(\partial\Omega)$ and $H^{1/2}(\partial\Omega)$ is a proper dense space of $L^2(\partial\Omega)$. So we assume the existence of F_1 and F_2 (see [4]).

Now we want to put problem (2.15a)-(2.15d) in its standard variational form. To do so, we identify $\nabla\tilde{p} - \rho\mathbf{g}$ with $\nabla\tilde{u}$ and we use (iii) of corollary (3.2.3), that is we multiply (2.15a) by a test function $q \in H^1(\Omega)$ and do integration by parts to obtain

$$\int_\Omega \mathbf{K}(\nabla\tilde{p} - \rho\mathbf{g}) \cdot \nabla q \, dx = \int_\Omega fq \, dx + \int_\Gamma \mathbf{K}(\nabla\tilde{p} - \rho\mathbf{g}) \cdot \nu q \, d\sigma, \tag{4.1}$$

where ν is the outward normal unit vector to $\partial\Omega$. From (4.1) the requirements for $\tilde{p}$ and $\mathbf{K}$ are now $\tilde{p} \in H^1(\Omega) \cap \mathbf{V}$ and $K_{i,j} \in L^\infty(\Omega)$ for $1 \leq i, j \leq d$; and $F_1 \in H^1(\Omega)$. Now we transform (4.1) into

$$\begin{aligned}\int_\Omega \mathbf{K}\nabla\tilde{p} \cdot \nabla q \, dx = &\int_\Omega fq \, dx + \int_\Omega \mathbf{K}\rho\mathbf{g} \cdot \nabla q \, dx + \int_{\Gamma_N} f_2 q \, d\sigma \\ &+ \int_{\Gamma_D} \mathbf{K}\nabla\tilde{p} \cdot \nu q \, d\sigma - \int_\Gamma \mathbf{K}\rho\mathbf{g} \cdot \nu q \, d\sigma.\end{aligned} \tag{4.2}$$

To write the variational formulation and set up the space $\mathbf{V}$, we consider the case where $f_1 = 0$ and the case where $f_1 \neq 0$ because the test function is chosen according to the Dirichlet boundary.

(1) **Case $\mathbf{f_1 = 0}$.** We set $\mathbf{V} = H^1_{0,D} = \{q \in H^1(\Omega) \,|\, \gamma_0 q = 0 \text{ on } \Gamma_D\}$. Taking $q \in H^1_{0,D}$ the formulation (4.2) becomes

$$\int_\Omega \mathbf{K}\nabla\tilde{p} \cdot \nabla q \, dx = \int_\Omega fq \, dx + \int_\Omega \mathbf{K}\rho\mathbf{g} \cdot \nabla q \, dx + \int_{\Gamma_N} f_2 q \, d\sigma - \int_{\Gamma_N} \mathbf{K}\rho\mathbf{g} \cdot \nu q \, d\sigma. \tag{4.3}$$

Hence the variational formulation in this case is: Find $\tilde{p} \in H^1_{0,D}$ such that for every $q \in H^1_{0,D}$ the equation (4.3) holds.

(2) **Case $\mathbf{f_1 \neq 0}$.** We assume the existence of F_1 as said earlier with $\gamma_0 w = F_1$ and we set the change of variable $\tilde{p_1} = \tilde{p} + w$ where $\tilde{p} \in \mathbf{V} = H^1_{0,D} = \{q \in H^1(\Omega) \,|\, \gamma_0 q = 0 \text{ on } \Gamma_D\}$. The variational formulation in this case is: Find $\tilde{p} \in H^1_{0,D}$ such that for every $q \in H^1_{0,D}$ the following equation holds

$$\int_\Omega \mathbf{K}\nabla(\tilde{p_1})\cdot\nabla q\,dx = \int_\Omega fq\,dx + \int_\Omega \mathbf{K}\rho\mathbf{g}\cdot\nabla q\,dx + \int_{\Gamma_N} f_2 q\,d\sigma - \int_{\Gamma_N}\mathbf{K}\rho\mathbf{g}\cdot\nu q\,d\sigma,$$

which is the same as

$$\begin{aligned}\int_\Omega \mathbf{K}\nabla\tilde{p}\cdot\nabla q\,dx = &-\int_\Omega \mathbf{K}\nabla w\cdot\nabla q\,dx + \int_\Omega fq\,dx \\ &+\int_\Omega \mathbf{K}\rho\mathbf{g}\cdot\nabla q\,dx + \int_{\Gamma_N} f_2 q\,d\sigma - \int_{\Gamma_N}\mathbf{K}\rho\mathbf{g}\cdot\nu q\,d\sigma. \qquad (4.4)\end{aligned}$$

Summarizing the above two cases, the primal variation formulation can be formulated as find $\tilde{p}\in H^1_{0,D}$ such that for every $p\in H^1_{0,D}$ the following equation holds

$$\mathcal{A}(\tilde{p},q) = \mathcal{L}(q), \qquad (4.5)$$

where

(i) $H^1_{0,D} = \{u\in H^1(\Omega)\,|\,\gamma_0 u = 0 \text{ on } \Gamma_D\}$,

(ii) $\mathcal{A}$ is the bilinear form defined on $H^1_{0,D}\times H^1_{0,D}$ by

$$\mathcal{A}(p,q) = \int_\Omega \mathbf{K}\nabla p\cdot\nabla q\,dx. \qquad (4.6)$$

(iii) $\mathcal{L}$ is a linear form defined on $H^1_{0,D}$ by

$$\mathcal{L}(q) = \begin{cases} \displaystyle\int_\Omega fq\,dx - \int_\Omega \mathbf{K}\rho\mathbf{g}\cdot\nabla q\,dx + \int_{\Gamma_N} f_2 q\,d\sigma + \int_{\Gamma_N}\mathbf{K}\rho\mathbf{g}\cdot\nu q\,d\sigma, & \text{if } f_1 = 0, \\ \displaystyle -\mathcal{A}(w,q) + \int_\Omega fq\,dx - \int_\Omega \mathbf{K}\rho\mathbf{g}\cdot\nabla q\,dx + \int_{\Gamma_N} f_2 q\,d\sigma + \int_{\Gamma_N}\mathbf{K}\rho\mathbf{g}\cdot\nu q d\sigma & \text{if } f_1\neq 0, \end{cases} \qquad (4.7)$$

with $w\in H^1(\Omega)$ and $\gamma_0 w = f_1$.

4.2 Mixed formulation of the elliptic boundary problem

In the primal variational formulation there is only one variable which is the pressure $\tilde{p}$, and then one solves for the velocity field once we obtain the pressure. In the mixed formulation, both the pressure p and the Darcy's velocity field $\mathbf{u}$ are obtained simultaneously. So we reconsider the problem (2.15a-2.15d) and we transform it into

$$\mathbf{u} = \mathbf{K}(\nabla p - \rho \mathbf{g}), \tag{4.8a}$$

$$-\nabla \cdot \mathbf{u} = f \tag{4.8b}$$

$$p = f_1 \text{ on } \Gamma_D, \tag{4.8c}$$

$$\mathbf{u} \cdot \nu = f_3 \text{ on } \Gamma_N. \tag{4.8d}$$

Similarly as in standard variation formulation, since the Neumann boundary condition becomes essential boundary we begin by case $f_3 = 0$ and end up with $f_3 \neq 0$.

(1) **Case $\mathbf{f_3} = \mathbf{0}$**. We consider the variational formulation: Find $\tilde{p} \in H^1_{f_1,D} = \{u \in H^1(\Omega) | u_{|\Gamma_D} = f_1\}$ and $\tilde{\mathbf{u}} \in (L^2(\Omega))^d$ such that for all $\mathbf{v} \in (L^2(\Omega))^d$ and $q \in H^1_{0,D}$ we have:

$$\int_\Omega \mathbf{K}^{-1}\tilde{\mathbf{u}} \cdot \mathbf{v}\, dx - \int_\Omega (\nabla \tilde{p} - \rho \mathbf{g}) \cdot \mathbf{v}\, dx = 0, \tag{4.9a}$$

$$\int_\Omega \tilde{\mathbf{u}} \cdot \nabla q\, dx - \int_\Omega f q\, dx = 0 \tag{4.9b}$$

By integrations by parts the variational formulation (4.9a-4.9b) we get the variational formulation known as ***Mixed Formulation***; find $\tilde{p} \in L^2(\Omega)$ and $\tilde{\mathbf{u}} \in H_{0,N}(\mathrm{div}, \Omega)$ such that for all $q \in L^2(\Omega)$ and $\mathbf{v} \in H_{0,N}(\mathrm{div}, \Omega)$ we have:

$$\begin{cases} \displaystyle\int_\Omega \mathbf{K}^{-1}\tilde{\mathbf{u}} \cdot \mathbf{v}\, dx + \int_\Omega \tilde{p} \nabla \cdot \mathbf{v}\, dx &= -\displaystyle\int_\Omega \rho \mathbf{g} \cdot \mathbf{v}\, dx + \int_{\Gamma_N} f_1 \mathbf{v} \cdot \nu d\sigma, \\ \displaystyle\int_\Omega \nabla \cdot \tilde{\mathbf{u}} q\, dx &= -\displaystyle\int_\Omega f q\, dx. \end{cases} \tag{4.10}$$

(2) **Case $\mathbf{f_3} \neq \mathbf{0}$**. Similarly as in primal formulation, we assume that there is $\tilde{\mathbf{u}}_0 \in H_{0,N}(\mathrm{div}, \Omega)$ such that $\tilde{\mathbf{u}}_0 \cdot \nu = f_3$ on Γ_N and we make change of variable $\tilde{\mathbf{u}}_1 = \tilde{\mathbf{u}}_0 + \tilde{\mathbf{u}}$ with $\tilde{\mathbf{u}} \in H_{0,N}(\mathrm{div}, \Omega)$. Thus the ***Mixed Formulation*** in this case is: Find $\tilde{p} \in L^2(\Omega)$ and $\tilde{\mathbf{u}} \in H_{0,N}(\mathrm{div}, \Omega)$ such that for all $q \in L^2(\Omega)$ and $\mathbf{v} \in H_{0,N}(\mathrm{div}, \Omega)$ we have:

$$\begin{cases} \displaystyle\int_\Omega \mathbf{K}^{-1}\tilde{\mathbf{u}} \cdot \mathbf{v}\, dx + \int_\Omega \tilde{p} \nabla \cdot \mathbf{v}\, dx = -\int_\Omega \rho \mathbf{g} \cdot \mathbf{v}\, dx + \int_{\Gamma_N} f_1 \mathbf{v} \cdot \nu d\sigma - \int_\Omega \mathbf{K}^{-1}\tilde{\mathbf{u}}_0 \cdot \mathbf{v}\, dx, \\ \displaystyle\int_\Omega \nabla \cdot \tilde{\mathbf{u}} q\, dx = -\int_\Omega f q\, dx - \int_\Omega \nabla \cdot \tilde{\mathbf{u}}_0 q\, dx \end{cases} \tag{4.11}$$

The formulations (4.10-4.11) can be put in the form: Find $\tilde{p} \in L^2(\Omega)$ and $\tilde{\mathbf{u}} \in H_{0,N}(\mathrm{div}, \Omega)$ such that for all $q \in L^2(\Omega)$ and $\mathbf{v} \in H_{0,N}(\mathrm{div}, \Omega)$ we have:

$$\begin{cases} A(\tilde{\mathbf{u}}, \mathbf{v}) + B(\mathbf{v}, \tilde{p}) = D_1(\mathbf{v}), \\ B(\tilde{\mathbf{u}}, q) = D_2(q), \end{cases} \tag{4.12}$$

where

(i) A is a bounded bilinear form on $H_{0,N}(\mathrm{div}, \Omega) \times H_{0,N}(\mathrm{div}, \Omega)$ defined by

$$A(\mathbf{u}, \mathbf{v}) = \int_\Omega \mathbf{K}^{-1}\mathbf{u} \cdot \mathbf{v}\, dx, \tag{4.13}$$

(ii) B is a bounded bilinear form on $H_{0,N}(\mathrm{div}, \Omega) \times L^2(\Omega)$ defined by

$$B(\mathbf{v}, q) = \int_\Omega \nabla.\mathbf{v} q\, dx, \tag{4.14}$$

(iii) D_1 is a bounded linear form on $H_{0,N}(\mathrm{div}, \Omega)$ defined by

$$D_1(\mathbf{v}) = \begin{cases} -\int_\Omega \rho \mathbf{g} \cdot \mathbf{v}\, dx + \int_{\Gamma_N} f_1 \mathbf{v} \cdot \nu\, d\sigma & \text{if } f_3 = 0, \\ -\int_\Omega \rho \mathbf{g} \cdot \mathbf{v}\, dx + \int_{\Gamma_N} f_1 \mathbf{v} \cdot \nu\, d\sigma - \int_\Omega \mathbf{K}^{-1}\tilde{\mathbf{u}}_0 \cdot \mathbf{v}\, dx, & \text{if } f_3 \neq 0, \end{cases} \tag{4.15}$$

(iv) D_2 is a bounded linear form on $L^2(\Omega)$ defined by

$$D_2(q) = \begin{cases} -\int_\Omega f q\, dx & \text{if } f_3 = 0, \\ -\int_\Omega f q\, dx - \int_\Omega \nabla \cdot \tilde{\mathbf{u}}_0 q\, dx & \text{if } f_3 \neq 0. \end{cases} \tag{4.16}$$

In the sequel we set $H_{0,N}(\mathrm{div}, \Omega) = \mathbf{Q}$. It is clear that we have the estimate

$$||q||_{\mathbf{Q}} \geq ||q||_0. \tag{4.17}$$

Now let us prove the assumptions made on A, B, D_1 and D_2 previously.

Proof.

(i) **Bilinearity and boundness of** A

$\forall \alpha_1, \alpha_2 \in \mathbb{R}$ and $\mathbf{v}_1, \mathbf{v}_2, \mathbf{u} \in \mathbf{Q}$ we have:

$$A(\alpha_1\mathbf{v}_1+\alpha_2\mathbf{v}_2,\mathbf{u}) = \int_\Omega \mathbf{K}^{-1}(\alpha_1\mathbf{v}_1+\alpha_2\mathbf{v}_2)\cdot\mathbf{u}\,dx = \alpha_1\int_\Omega \mathbf{K}^{-1}\mathbf{v}_1\cdot\mathbf{u}\,dx + \alpha_2\int_\Omega \mathbf{K}^{-1}\mathbf{v}_2\cdot\mathbf{u}\,dx$$
$$= \alpha_1 A(\mathbf{v}_1,\mathbf{u}) + \alpha_2 A(\mathbf{v}_2,\mathbf{u}).$$

A is linear in the first variable and since it is symmetric (because $\mathbf{K}$ and $\mathbf{K}^{-1}$ are symmetric) A is bilinear. To show the boundedness, consider

$$\begin{aligned}
|A(\mathbf{v}_1,\mathbf{u})| &= \Big|\int_\Omega \mathbf{K}^{-1}\mathbf{v}_1\cdot\mathbf{u}\,dx\Big| \\
&\le ||\mathbf{K}^{-1}||_\infty \int_\Omega |\mathbf{v}_1|\cdot|\mathbf{u}|\,dx \\
&\le ||\mathbf{K}^{-1}||_\infty ||\mathbf{v}_1||_0||\mathbf{u}||_0 \quad \text{(from Cauchy-Schwarz inequality)} \\
&\le ||\mathbf{K}^{-1}||_\infty ||\mathbf{v}_1||_{\mathbf{Q}}||\mathbf{u}||_{\mathbf{Q}} \quad \text{(from inequality) (4.17).}
\end{aligned}$$

Hence A is a bounded bilinear form.

(ii) **Bilinearity and a boundness of B**

Let $\alpha_1,\alpha_2\in\mathbb{R}$, $\mathbf{v}_1,\mathbf{v}_2\in\mathbf{Q}$ and $p_1,p_2\in L^2(\Omega)$. On the one hand we have:

$$B(\alpha_1\mathbf{v}_1+\alpha_2\mathbf{v}_2,p_1) = \int_\Omega \nabla(\alpha_1\mathbf{v}_1+\alpha_2\mathbf{v}_2)\cdot p_1\,dx = \alpha_1\int_\Omega \nabla\mathbf{v}_1\cdot p_1\,dx + \alpha_2\int_\Omega \nabla\mathbf{v}_2\cdot p_1\,dx,$$
$$= \alpha_1 B(\mathbf{v}_1,p_1) + \alpha_2 B(\mathbf{v}_2,p_1),$$

and

$$B(\mathbf{v}_1,\alpha_1p_1+\alpha_2p_2) = \alpha_1\int_\Omega \nabla\mathbf{v}_1\cdot p_1\,dx + \alpha_2\int_\Omega \nabla\mathbf{v}_1\cdot p_2\,dx,$$
$$= \alpha_1 B(\mathbf{v}_1,p_1) + \alpha_2 B(\mathbf{v}_1,p_2),$$

that proves that B is bilinear. On the other hand, we have

$$\begin{aligned}
|B(\mathbf{v}_1,p_1)| &= \Big|\int_\Omega \nabla\cdot\mathbf{v}_1p_1\,dx\Big| \\
&\le \int_\Omega |\nabla\cdot\mathbf{v}_1||p_1|dx, \\
&\le ||\nabla\cdot\mathbf{v}_1||_0||p_1||_0, \\
&\le ||\mathbf{v}_1||_{\mathbf{Q}}||p_1||_0,
\end{aligned}$$

so that B is bounded.

(iii) **Linearity and boundness of D_1 .**

We recall that the normal trace is continuous that is for $\mathbf{v} \in H_{0,N}(\mathrm{div}, \Omega)$, there is $c > 0$ such that $||\mathbf{v}||_{L^2(\partial\Omega)} \leq c||\mathbf{v}||_{\mathrm{div}}$. Let $\alpha_1, \alpha_2 \in \mathbb{R}$, $\mathbf{v}, \mathbf{u} \in \mathbf{Q}$. We have

$$\begin{aligned} D_1(\alpha_1\mathbf{v} + \alpha_2\mathbf{u}) &= \int_\Omega -\rho\mathbf{g} \cdot (\alpha_1\mathbf{v} + \alpha_2\mathbf{u})\, dx + \int_{\Gamma_N} f_1(\alpha_1\mathbf{v} + \alpha_2\mathbf{u}) \cdot \nu\, d\sigma - \int_\Omega \mathbf{K}^{-1}\tilde{\mathbf{u}}_0 \cdot (\alpha_1\mathbf{v} + \alpha_2\mathbf{u})\, dx, \\ &= \alpha_1 \int_\Omega -\rho\mathbf{g} \cdot \mathbf{v}\, dx + \alpha_1 \int_{\Gamma_N} f_1\mathbf{v} \cdot \nu\, d\sigma - \alpha_1 \int_\Omega \mathbf{K}^{-1}\tilde{\mathbf{u}}_0 \cdot \mathbf{u}\, dx \\ &\quad + \alpha_2 \int_\Omega -\rho\mathbf{g} \cdot \mathbf{u}\, dx + \alpha_2 \int_{\Gamma_N} f_1\mathbf{u} \cdot \nu\, d\sigma - \alpha_2 \int_\Omega \mathbf{K}^{-1}\tilde{\mathbf{u}}_0 \cdot \mathbf{v}\, dx, \\ &= \alpha_1 D_1(\mathbf{u}) + \alpha_1 D_1(\mathbf{v}), \end{aligned}$$

that is D_1 is linear. Furthermore,

$$\begin{aligned} |D_1(\mathbf{v})| &= \Big| \int_\Omega -\rho\mathbf{g}.\mathbf{v}\, dx + \int_{\Gamma_N} f_1\mathbf{v}.\nu\, d\sigma - \int_\Omega \mathbf{K}^{-1}\tilde{\mathbf{u}}_0.\mathbf{v}\, dx \Big| \\ &\leq |\rho\mathbf{g}| \int_\Omega |\mathbf{v}|\, dx + \int_\Gamma |f_1||\mathbf{v}.\nu|\, d\sigma + ||\mathbf{K}^{-1}||_\infty \int_\Omega |\tilde{\mathbf{u}}_0||\mathbf{v}|\, dx \\ &\leq |\rho\mathbf{g}|||\mathbf{v}||_0 + ||f_1||_{L^2(\partial\Omega)}||\mathbf{v}||_{L^2(\partial\Omega)} + ||\mathbf{K}^{-1}||_\infty||\tilde{\mathbf{u}}_0||_0||\mathbf{v}||_0 \\ &\leq |\rho\mathbf{g}|||\mathbf{v}||_{\mathbf{Q}} + c||f_1||_{L^2(\partial\Omega)}||\mathbf{v}||_{\mathbf{Q}} + ||\mathbf{K}^{-1}||_\infty||\tilde{\mathbf{u}}_0||_0||\mathbf{v}||_{\mathbf{Q}}, \\ &= M||\mathbf{v}||_{\mathbf{Q}} \end{aligned}$$

where $M = |\rho\mathbf{g}| + c||f_1||_{L^2(\partial\Omega)} + ||\mathbf{K}^{-1}||_\infty||\tilde{\mathbf{u}}_0||_0$. Hence D_1 is a linear bounded form

(iv) **Linearity and boundness of D_2 .**

Let $p_1, p_2 \in L^2(\Omega)$, $\alpha_1, \alpha_2 \in \mathbb{R}$. We have

$$\begin{aligned} D_2(\alpha_1 p_1 + \alpha_2 p_2) &= -\int_\Omega f(\alpha_1 p_1 + \alpha_2 p_2) dx - \int_\Omega \nabla.\tilde{\mathbf{u}}_0(\alpha_1 p_1 + \alpha_2 p_2)\, dx \\ &= \alpha_1 \int_\Omega f p_1 dx - \alpha_1 \int_\Omega \nabla.\tilde{\mathbf{u}}_0 p_1 dx - \alpha_2 \int_\Omega f p_2\, dx - \alpha_2 \int_\Omega \nabla.\tilde{\mathbf{u}}_0 p_2\, dx \\ &= \alpha_1 D_2(p_1) + \alpha_2 D_2(p_2), \end{aligned}$$

that is D_2 is linear. Furthermore,

$$\begin{aligned} |D_2(p_1)| &= \Big| -\int_\Omega f p_1 dx - \int_\Omega \nabla.\tilde{\mathbf{u}}_0 p_1\, dx \Big| \\ &\leq \int_\Omega |f||p_1|\, dx + \int_\Omega |\nabla.\tilde{\mathbf{u}}_0||p_1|\, dx \\ &\leq ||f||_0||p_1||_0 + ||\nabla.\tilde{\mathbf{u}}_0||_0||p_1||_0 \\ &\leq \Big(||f||_0 + ||\nabla.\tilde{\mathbf{u}}_0||_0\Big)||p_1||_0, \end{aligned}$$

that is D_2 is a bounded linear form. We only consider the case $f_3 \neq 0$, since the case $f_3 = 0$ is a direct consequence of the first case.

□

Let $\mathcal{O}$ be the bounded linear operator from $\mathbf{Q}$ to $L^2(\Omega)$ defined by the bilinear B. Such an operator always exists (see [15]). We denote by $\mathfrak{O} = ker(\mathcal{O}) = \{\mathbf{v} \in \mathbf{Q} \,|\, \nabla.\mathbf{v} = 0\}$.

Remark 4.2.1.

In our case the operator $\mathcal{O} =$ div. Indeed we have for $\mathbf{v} \in \mathbf{Q}$ and any $q \in L^2(\Omega)$

$$\begin{aligned} B(\mathbf{v}, q) &= \langle \mathcal{O}\mathbf{v}, q \rangle_0 \\ &= \int_\Omega \mathcal{O}\mathbf{v} q \, dx \\ &= \int_\Omega \nabla.\mathbf{v} q \, dx. \end{aligned}$$

Hence $\mathcal{O}\mathbf{v} = \nabla.\mathbf{v} \in L^2(\Omega)' = L^2(\Omega)$ that is $\mathcal{O} =$ div and $\mathfrak{O} = ker(\mathcal{O})$. As consequence, it follows that $||\mathbf{v}||_{\text{ div}} = ||\mathbf{v}||_{\mathbf{Q}} = ||\mathbf{v}||_0$ for $\mathbf{v} \in \mathfrak{O}$.

The existence and the uniqueness of the solution of the problem (4.12) is given by the following theorem.

Theorem 4.2.1. *[15]*

Let $\mathcal{H}$ and $\mathcal{Q}$ be real Hilbert spaces, $\mathfrak{a}$ a bounded bilinear form on $\mathcal{Q} \times \mathcal{Q}$, $\mathfrak{b}$ a bounded bilinear form on $\mathcal{Q} \times \mathcal{H}$, $\mathfrak{l}_1$ a bounded linear form on $\mathcal{Q}$, $\mathfrak{l}_2$ a bounded linear form on $\mathcal{H}$ and $\mathcal{N} = ker(\mathfrak{o})$ where $\mathfrak{o}$ is the operator defined by $\mathfrak{b}$. If

(i) the bilinear form $\mathfrak{a}$ is $\mathcal{N}$-elliptic,

(ii) the bilinear form $\mathfrak{b}$ satisfies the inf-sup condition; that is there is $\beta > 0$ such that

$$\sup_{\substack{\mathbf{v} \in \mathcal{Q} \\ p \neq 0}} \frac{\mathfrak{b}(\mathbf{v}, p)}{||\mathbf{v}||_{\mathcal{Q}}} \geq \beta ||p||_{L^2(\Omega)} \text{ or equivalently } \inf_{\substack{p \in L^2(\Omega) \\ p \neq 0}} \sup_{\substack{\mathbf{v} \in \mathcal{Q} \\ \mathbf{v} \neq 0}} \frac{\mathfrak{b}(\mathbf{v}, p)}{||p||_{L^2(\Omega)} ||\mathbf{v}||_{\mathcal{Q}}} \geq \beta,$$

then there is an unique solution $(x, y) \in (\mathcal{Q} \times L^2(\Omega))$ for the problem:

$$\begin{cases} \mathfrak{a}(\mathbf{x}, \mathbf{w}) + \mathfrak{b}(\mathbf{w}, y) &= \mathfrak{l}_1(\mathbf{w}) \quad \forall \mathbf{w} \in \mathcal{Q}, \\ \mathfrak{b}(\mathbf{x}, z) &= \mathfrak{l}_2(z) \quad \forall z \in L^2(\Omega). \end{cases}$$

Instead of using inf-sup condition (ii) in the theorem, we rather use its equivalences in the following Lemma.

Lemma 4.2.2. *[15]*

Let $\mathfrak{b}$ be a bilinear form on $\mathcal{Q} \times \mathcal{H}$ and W *its induced operator on $\mathcal{Q}$ to $L^2(\Omega)$. The following statement are equivalent:*

(i) There exists $\beta > 0$ *such that*

$$\sup_{\substack{\mathbf{v}\in\mathcal{Q}\\ \mathbf{v}\neq 0}} \frac{\mathfrak{b}(\mathbf{v},p)}{||\mathbf{v}||_{\mathcal{Q}}} \geq \beta||p||_{L^2(\Omega)} \quad \forall p \in L^2(\Omega).$$

(ii) The operator W^**, the adjoint of* W *is an isomorphism from* $L^2(\Omega)$ *to* $ker(W)^\perp$*, and*

$$||\mathrm{W}^*(p)||_{\mathcal{Q}} \geq \beta||p||_{L^2(\Omega)} \quad \forall p \in L^2(\Omega) \quad .$$

(iii) The operator W *is an isomorphism from* $ker(\mathrm{W}^*)^\perp$ *into* $L^2(\Omega)$ *and*

$$||\mathrm{W}(\mathbf{v})||_{L^2(\Omega)} \geq \beta||\mathbf{v}||_{\mathcal{Q}} \quad \forall \mathbf{v} \in ker(\mathrm{W})^\perp.$$

(iv) The operator $\mathrm{W} : L^2(\Omega) \longrightarrow \mathrm{Q}$ *is surjective.*

The proof of theorem 4.2.1 and lemma 4.2.2 can be found in [15]. Now we let's prove that the mixed formulation 4.12 is well-posed, that is it satisfies the conditions of the theorem 4.2.1.

Proof.
(i) We assume that $\mathbf{K}^{-1}$ is $(L^2(\Omega))^d$-elliptic that is there exist $\alpha_0 > 0$ such that
$\mathbf{K}^{-1}\mathbf{v}.\mathbf{v} \geq \alpha_0|\mathbf{v}|^2 \quad \forall \mathbf{v} \in (L^2(\Omega))^d$. So we have

$$|A(\mathbf{v},\mathbf{v})| = |\int_\Omega \mathbf{K}^{-1}\mathbf{v}.\mathbf{v}\, dx| \geq \alpha_0||\mathbf{v}||_0^2 = \alpha_0||\mathbf{v}||_{\mathbf{Q}}^2 \quad \forall \mathbf{v} \in \mathfrak{O}.$$

Hence the bilinear A defined in (4.12) is $\mathfrak{O}$-elliptic.

(ii) Instead of using the inf-sup condition, we use condition (iv) in lemma (4.2.2).
So let $q \in L^2(\Omega)$, we consider the problem

$$\begin{cases} -\nabla.\nabla p = q \text{ in } \Omega \\ \qquad p = 0 \text{ on } \Gamma_D \\ \nabla p.\nu = 0 \text{ on } \Gamma_N. \end{cases} \tag{4.18}$$

The primal variation formulation of problem (4.18) is: Find $\tilde{u} \in H^1_{0,D} = \{p \in H^1(\Omega) | p = 0 \text{ on } \Gamma_D\}$ such that

$$\mathcal{A}(\tilde{p},w) = \mathcal{L}(w) \quad \forall w \in H^1_{0,D}, \tag{4.19}$$

where $\mathcal{A}(\tilde{p},w) = \int_\Omega \tilde{p}w\, dx$ and $\mathcal{L}(w) = \int_\Omega qw\, dx$ (see 4.6-4.7). Similarly to the proof of the assumptions made on A, and D_1 previously, we prove that $\mathcal{A}$ is a bounded bilinear form on $H^1_{0,D} \times H^1_{0,D}$ and $\mathcal{L}$ is a bounded linear form on $H^1_{0,D}$. From the Lax-Milgram Theorem we deduce that problem (4.19) has a unique solution $\tilde{p}$. Then defining $\tilde{\mathbf{v}} = -\nabla\tilde{p}$, we have $\nabla.\tilde{\mathbf{v}} = q$; that is $\tilde{\mathbf{v}} \in \mathbf{Q}$. We conclude that the operator $\mathcal{O} =$ div is surjective. □

4.3 Well posedness of parabolic initial value problem

In this section we are interested in looking for a more general solution since in limited cases, separation of variable technique leads to classical solution (see [4] chapter 7). Considering equation (2.16a), we have

$$\partial_t h(\mathcal{C}) = \partial_{\mathcal{C}} h.\partial_t \mathcal{C}.$$

where $\partial_{\mathcal{C}} h$ can be degenerate matrix. In the case it is non-degenerate, the equation (2.16a) can be treated as in the case $\partial_{\mathcal{C}} h$ is an identity matrix. So we will consider in this work the simpler case, that is:

$$\partial_t \mathcal{C} - \nabla.(\mathbf{D}\nabla\mathcal{C} - \mathbf{u}\mathcal{C}) = Q(\mathcal{C}), \tag{4.20a}$$

$$\mathcal{C} = g_1 \text{ on } (0,T] \times \partial\Omega_D \tag{4.20b}$$

$$\mathbf{D}\nabla\mathcal{C}.\nu = g_2 \text{ on } (0,T] \times \partial\Omega_N, \tag{4.20c}$$

$$\mathcal{C}(0,x) = \mathcal{C}_0(x), \quad x \in \bar{\Omega}. \tag{4.20d}$$

In this sketch of the weak solution to the problem (4.20a-4.20d), we consider the space $L^2((0,T),\mathbf{V})$ where $\mathbf{V}$ is a space of functions on Ω that will be specified later. To deal with the weak solutions, functions in $L^2((0,T),\mathbf{V})$ are treated as the following:

(i) For a fixed time t, $\mathcal{C}(t,.)$ is treated as parameter-dependent function in $\mathbf{V}$.

(ii) For varying the time t, the function $t \longmapsto \mathcal{C}(t,.)$ is a function valued mapping with range in $\mathbf{V}$.

In addition to the space $\mathbf{V}$, we require that $\mathcal{C}_0 \in L^2(\Omega) = L^2(\Omega)$. Now to put the problem (4.20a-4.20d) in the weak formulation, we start by the Dirichlet homogeneous boundary condition on Γ_Ω. Hence the obvious choice of $\mathbf{V}$ is $\mathbf{V} = H_0^1(\Omega) = \{p \in H^1(\Omega) | \, \gamma_0 p = 0\}$. So we multiply equation (4.20a) by a function $q \in \mathbf{V}$, fix the time variable and do integration by parts and we get:

$$\int_\Omega \mathcal{C}_t q \, dx + \int_\Omega \mathbf{D}\nabla\mathcal{C}.\nabla q \, dx - \int_\Omega \nabla.(\mathbf{u}\mathcal{C}) q \, dx = \int_\Omega Q(\mathcal{C}) q \, dx. \tag{4.21}$$

We consider on $H_0^1(\Omega) \times H^1(\Omega)_0$ the bilinear form defined by

$$\breve{A}(p,q) = \int_\Omega \mathbf{D}\nabla p.\nabla q \, dx - \int_\Omega \nabla.(\mathbf{u}p) q \, dx, \tag{4.22}$$

associated with the operator

$$\breve{O}\mathcal{C} = -\nabla.(\mathbf{D}\nabla\mathcal{C} - \mathbf{u}\mathcal{C}). \tag{4.23}$$

In the homogeneous Dirichlet boundary conditon that leads to equation (4.21), we have

$$\breve{A}(\mathcal{C}, p) = \langle \breve{O}\mathcal{C}, p\rangle_0 \tag{4.24}$$

for all $p \in H_0^1(\Omega)$ and $\mathcal{C} \in \mathcal{D}_m(\breve{O}) = H_0^1(\Omega) \cap H^2(\Omega)$. Hence the **weak formulation** of the problem (4.20a-4.20d) in the homogeneous Dirichlet boundary condition case is: Find $\mathcal{C} \in \mathcal{D}_m(\breve{O})$ such that for all $p \in \mathcal{D}_m(\breve{O})$

$$\begin{cases} \langle \mathcal{C}_t, p\rangle_0 + \langle \breve{O}\mathcal{C}, p\rangle_0 = \langle Q(\mathcal{C}), p\rangle_0, \\ \mathcal{C}(0) = \mathcal{C}_0. \end{cases} \tag{4.25}$$

The existence and uniqueness of the solution of system (4.25) within the semigroup framework is based on operator's theory. Here we give briefly some key concepts that are necessary for this work. A complete understanding of operators and semigroups can be found in [6, 18, 19]. Let V be a Banach space. If for a linear operator $\mathbf{O}$, $-\mathbf{O}$ is an infinitesimal generator of a bounded analytic semigroup, and $-\mathbf{O}$ is invertible, the operator $\mathbf{O}^\alpha$ is defined and is a closed linear invertible operator for $0 \leq \alpha \leq 1$. The domain $\mathcal{D}_m(\mathbf{O}^\alpha) = V_\alpha$ is dense in V, and endowed with the norm $||x||_\alpha = ||\mathbf{O}^\alpha x||_V$ is a Banach space (see [18, Chapter 6]).

Theorem 4.3.1.

Let $-\mathbf{O}$ be an infinitesimal generator of a bounded analytic semigroup $S(t)$ and $-\mathbf{O}$ invertible. Let U be an open subset of $\mathbb{R}_+ \times V_\alpha$ and assume that $f : U \longrightarrow V$ is Hölder Lipschitz in $t \in \mathbb{R}_+$ and Lipschitz with respect to $x \in V_\alpha$, that is for $(t_1, x_1) \in U$ there is a neighborhood $U' \subset U$ of (t_1, x_1) such that for every $(t, x), (t', x') \in U$ we have
$||f(t, x) - f(t', x')|| \leq L(|t - t'|^\beta + ||x - x'||_\alpha)$, for some constants $\beta > 0$ and $L > 0$.
Then the Cauchy problem

$$\begin{cases} Y_t + \mathbf{O}Y = f(t, Y), \\ Y(t_0) = Y_0, \end{cases} \tag{4.26}$$

has locally an unique solution $Y \in C([t_0, T], V) \cap C^1((t_0, T), V)$ where $T = T(t_0, Y_0)$.

For the proof of the theorem see ([18],Chapter 6). The unique solution can be extended to the interval of time $[0, +\infty)$ for some additional condition on the function f (see [18],Chapter 6, Theorem 3.3).

Remark 4.3.1.

From the semigroup theory, the solution of the Cauchy problem is in the form

$$Y(t) = S(t)Y(t_0) + \int_{t_0}^{t} S(t - s)f(s, Y(s))ds, \quad t \in [t_0, T]. \tag{4.27}$$

This solution is called a **mild solution** of the Cauchy problem.

The remaining work is to prove that problem (4.25) lies in the framework of theorem (4.3.1). Actually we will use a transformed problem that has the same solution. For that purpose we need the following assumptions and theorem.

Assumption 4.3.2.

The nonlinear term $Q(\mathcal{C})$ is continuous and satisfies the Lipschitz conditions of theorem (4.3.1).

Assumption 4.3.3.

The diffusion tensor $\mathbf{D}$ is symmetric and positive definite, and all its coefficients are bounded; that is, $D_{i,j} \in L^\infty(\Omega)$, and there exists $k_1 > 0$ such that

$$\sum_{i,j=1}^{d} D_{i,j}\xi_i\xi_j \geq k_1|\xi|, \quad \xi \in \mathbb{R}^d.$$

Theorem 4.3.4 (Gårding's inequality).

Under the assumption (4.3.3), there exists a constant $k_0 > 0$ such that

$$\breve{A}(p,p) + k_0||p||_0^2 \geq \frac{k_1}{2}||p||_1^2. \tag{4.28}$$

Proof.

On the one hand we have

$$\begin{aligned}
\breve{A}(p,p) &= \int_\Omega \mathbf{D}\nabla p.\nabla p\, dx - \int_\Omega \nabla.(\mathbf{u}p)p\, dx \\
&\geq k_1 \int_\Omega |\nabla p|^2\, dx - \int_\Omega \nabla.(\mathbf{u}p)p\, dx \\
&\geq k_1|p|_0^2 - \left|\int_\Omega \nabla.(\mathbf{u}p)p\, dx\right|.
\end{aligned} \tag{4.29}$$

On the other hand we have

$$\begin{aligned}
\left|\int_\Omega \nabla.(\mathbf{u}p)p\, dx\right| &= \left|\int_\Omega \nabla.\mathbf{u}p^2\, dx + \int_\Omega \mathbf{u}.\nabla pp\, dx\right| \\
&\leq \left|\int_\Omega \nabla.\mathbf{u}p^2\, dx\right| + \left|\int_\Omega \mathbf{u}.\nabla pp\, dx\right| \\
&\leq \int_\Omega |\nabla.\mathbf{u}|p^2\, dx + \left|\int_\Omega \mathbf{u}.\nabla pp\, dx\right|.
\end{aligned} \tag{4.30}$$

Indeed $\nabla.\mathbf{u} = f \in L^2(\Omega)$, $\mathbf{u}, \nabla p \in (L^2(\Omega))^d$ and $|\mathbf{u}.\nabla p| \leq |\mathbf{u}||\nabla p|$; we denote by $M_1 = ||f||_0$ and $M_2 = \Big(\int_\Omega \mathbf{u}^2\, dx\Big)^{\frac{1}{2}}$, and by means of the Cauchy-Schwarz inequality, we get from (4.30)

$$\begin{aligned}
\left|\int_\Omega \nabla.(\mathbf{u}p)p\, dx\right| &\leq M_1||p||_0^2 + \int_\Omega |\mathbf{u}||\nabla p||p|\, dx, \\
&\leq M_1||p||_0^2 + M_2|p|_1||p||_0.
\end{aligned} \tag{4.31}$$

Now we want to use **Young's inequality** (see Chapter 3) for the term $M_2|p|_1||p||_0$. So we have

$$\begin{aligned} M_2|p|_1||p||_0 &= \Big(\sqrt{k_1}|p|_1\Big)\Big(\frac{M_2}{\sqrt{k_1}}||p||_0\Big) \\ &\leq \frac{1}{2}\Big(\sqrt{k_1}|p|_1\Big)^2 + \frac{1}{2}\Big(\frac{M_2}{\sqrt{k_1}}||p||_0\Big)^2. \end{aligned} \tag{4.32}$$

Introducing (4.32) in (4.31) we get

$$\left|\int_\Omega \nabla.(\mathbf{u}p)p\,dx\right| \leq M_1||p||_0^2 + \frac{1}{2}\Big(\sqrt{k_1}|p|_1\Big)^2 + \frac{1}{2}\Big(\frac{M_2}{\sqrt{k_1}}||p||_0\Big)^2. \tag{4.33}$$

Using (4.33) in (4.29) we get

$$\breve{A}(p,p) \geq \frac{k_1}{2}|p|_1^2 - (M_1 + \frac{M_2}{2k_1})||p||_0^2. \tag{4.34}$$

Taking $k_0 = M_1 + \dfrac{M_2}{2k_1} + \dfrac{k_1}{2}$, we finally get the Gårding's inequality

$$\breve{A}(p,p) + k_0||p||_0^2 \geq \frac{k_1}{2}|p|_1^2 + k_0||p||_0^2 - (M_1 + \frac{M_2}{2k_1})||p||_0^2 = \frac{k_1}{2}||p||_1^2. \tag{4.35}$$

□

We now add k_0X to both sides of the equation (4.20a) to get a transformed problem that we will refer to as (*). It is clear that the existence and the uniqueness of the solution of (*) forces that of (4.20a-4.20d). In the problem (*), the right hand term $Q(X) + k_0X$ satisfies the same Lipschitz condition as $Q(X)$ and the bilinear form $\breve{A}_{k_0}$ associated with its operator $\breve{O}_{k_0} = \breve{O} + k_0I$ is $H^1(\Omega)$- elliptic. In fact we have

$$\breve{A}_{k_0}(p,p) = \langle \breve{O}_{k_0}p, p\rangle_0 = \langle \breve{O} + k_0p, p\rangle_0 = \breve{A}(p,p) + k_0||p||_0^2 \geq \frac{k_1}{2}||p||_1^2. \tag{4.36}$$

From result (4.36), it follows that $-\breve{O}_{k_0}$ is a sectorial operator on $L^2(\Omega)$ and then is the generator of analytic semigroup (see [1]). As a consequence, the problem (*) has locally an unique solution, and this implies that (4.20a-4.20d) also has locally an unique solution.

Chapter 5

Spatial discretization

After the investigation of some weak solutions of a PDEs in some Sobolev spaces, the next step is to use some numerical schemes for their approximations. There are various numerical schemes, and their choice depends in part on the type of PDE in consideration. For example, an elliptic problem requires only spatial discretization while a parabolic problem requires both spatial and temporal discretizations. Among the spatial schemes, the finite volume method (FVM), finite element method (FEM) and finite difference method (FDM) are heavily used. In this chapter we develop briefly two of those schemes.

In Section 5.1, we present finite volume method and a mixed-finite element method for an elliptic problem, and in Section 5.2, we present the finite volume method for a parabolic problem.

5.1 Elliptic boundary value problem

Elliptic problems need only spacial discretization. This entails dividing the physical domain into cells that are usually called **elements** or **control volumes**, then the solutions of the problems are approximated on each cell. With cells are associated some points usually called grid points, nodes or centres. The cells together with the points are known as **meshes**. The way of approximating a problem defines the type of mesh to generate and thus defines also whether the method is FVM, FEM or FDM. Here we present the FVM with two-point flux and a mixed finite element method.

5.1.1 Finite volume method

The finite volume method is similar to finite element method but have an additional property of local conservation of flux. In problems where discontinuities occur,the FVM can be used with a suitable mesh while FDM is not appropriate (see [23]). These facts make FVM one of the most used methods for the problems with conservation laws like fluid mechanics and heat transfer. The following is the description of the FVM with two-point flux in the so called K-orthogonal mesh.

Admissible mesh

Definition 5.1.1.

Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$. An admissible finite volume mesh of Ω, is given by a family of open polygonal convex subsets of Ω, denoted by $\mathcal{M}$ (called control volumes or cells), a family of subsets of $\bar{\Omega}$ contained on the hyperplane of $\mathbb{R}^d$ denoted by $\mathcal{E}$ (elements in $\mathcal{E}$ are edges for $d = 2$ and faces for $d = 3$ of the control volumes) with strictly positive $(d-1)$-dimensional Lebesgue measure, and a family of points, denoted by $\mathcal{P}$, satisfying the following properties:

(i) The closure of the union of the control volumes is $\bar{\Omega}$.

(ii) For any $I \in \mathcal{M}$, there is a subset $\mathcal{E}_I$ of $\mathcal{E}$ such that $\partial I = \bar{I} - I = \cup_{\sigma \in \mathcal{E}_I} \bar{\sigma}$ and $\mathcal{E} = \cup_{I \in \mathcal{M}} \mathcal{E}_I$.

(iii) For any $I, J \in \mathcal{M}$ with $I \neq J$, either $\bar{I} \cap \bar{J} = \bar{\sigma}$ for some $\sigma \in \mathcal{E}$ or the $(d-1)$-dimensional Lebesgue measure of $\bar{I} \cap \bar{J}$ is zero.

(iv) The set $\mathcal{M}$ is such that:
(1) The restrictions of f_1 and f_2 to each $\sigma \in \mathcal{E}_{\text{ext}}$ are continuous, where $\mathcal{E}_{\text{ext}} = \{\sigma \in \mathcal{E} | \, \sigma \subset \partial\Omega\}$, and for $\sigma \subset \partial\Omega$ either $\sigma \subset \Gamma_D$ or $\sigma \subset \Gamma_N$.
(2) The family of points $\mathcal{P} = (x_I)_{I \in \mathcal{M}}$ is defined as $x_I \in \bar{I}$ (for all $I \in \mathcal{M}$), the strait line from x_I to σ ($\sigma \in \mathcal{E}_I$) denoted by $D_{I,\sigma}$, is orthogonal to σ and $x_I \neq x_J$ for $I \neq J$. We set $x_\sigma = \sigma \cap D_{I,\sigma}$.

(v) For any $\sigma \in \mathcal{E}_{\text{ext}}$, let I be the control volume such that $\sigma \in \mathcal{E}_I$, then there exists $x_\sigma \in \sigma \cap D_{I,\sigma}$.
We associate with this definition the following notations. For $\sigma = \bar{I} \cap \bar{J}$, $d_{I,\sigma}$ is the Euclidean distance from x_I to σ, $Size(\mathcal{M}) = \sup\{\text{diam}(I), I \in \mathcal{M}\}$, $m(I)$ is the $d-$dimension measure of $I \in \mathcal{M}$, $m(\sigma)$ is the $(d-1)$-dimensional measure of $\sigma \in \mathcal{E}$, and $\mathcal{E}_{\text{int}} = \mathcal{E} - \mathcal{E}_{\text{ext}}$.

We highlight in the figure 5.1 two control volumes I and K among four control volumes I, J, K and L.

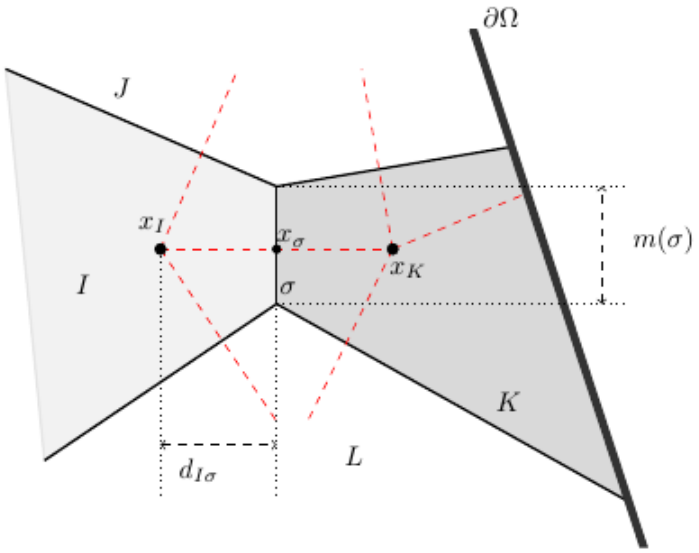


Figure 5.1: Admissible mesh

Discretization

We discretize over the mesh set up in the definition 5.1.1. The discretization consists of approximating the unknown (the pressure p) over each control volume a constant. For x_I, and x_σ we denote by p_I and p_σ the approximations of $p(x_I)$ and $p(x_\sigma)$ respectively, p_I is taken as the approximate solution over the entire control volume I and p_σ is just an auxiliary unknown which does not appear in the final discretized equation. To get the discretized equation, we integrate over each control volume the equation (2.15a) and we use the divergence theorem. For a control volume I, we have

$$-\int_I \nabla\cdot(\mathbf{K}(\nabla p+\rho\mathbf{g}))\,dx=-\sum_{\sigma\subset\partial I}\int_\sigma \mathbf{K}(\nabla p+\rho\mathbf{g})\cdot\nu_{I,\sigma}\,ds=\int_I f\,dx, \tag{5.1}$$

where $\nu_{I,\sigma}$ is the unit normal vector to σ outward of I. Because of the symmetry of $\mathbf{K}$, we note that $\mathbf{K}\nabla p\cdot\nu_{I,\sigma}=\mathbf{K}\nu_{I,\sigma}\cdot\nabla p$ and we approximate $\mathbf{K}\nu_{I,\sigma}$ by $\mathbf{K}_I\nu_{I,\sigma}$ at each point of σ and denote by $\mathbf{K}_{I,\sigma}=|\mathbf{K}_I\nu_{I,\sigma}|$. We approximate $\int_I f dx$ by $m(I)f(x_I)$. Now the approximation of $\int_\sigma \mathbf{K}\nu_{I,\sigma}\cdot\nabla p\,ds$ is made as follows:

(i) For $\sigma\subset\Gamma_N$, $\int_\sigma \mathbf{K}\nu_{I,\sigma}\cdot\nabla p\,ds=m(\sigma)f_2(x_\sigma)$;

(ii) For $\sigma\subset\Gamma_D$, $\int_\sigma \mathbf{K}\nu_{I,\sigma}\cdot\nabla p\,ds=m(\sigma)\mathbf{K}_{I,\sigma}\dfrac{f_1(x_\sigma)-p_I}{d_{I,\sigma}}$;

(iii) For $\sigma = \bar{I} \cap \bar{J}$, $\displaystyle\int_\sigma \mathbf{K}\nu_{I,\sigma} \cdot \nabla p \, ds = m(\sigma)\mathbf{K}_{I,\sigma}\frac{p_\sigma - p_I}{d_{I,\sigma}}$.

The local conservation of the flux across σ implies that

$$\int_\sigma \mathbf{K}\nu_{I,\sigma} \cdot \nabla p \, ds = -\int_\sigma \mathbf{K}\nu_{J,\sigma} \cdot \nabla p \, ds;$$

that is,

$$m(\sigma)\mathbf{K}_{I,\sigma}\frac{p_\sigma - p_I}{d_{I,\sigma}} = -m(\sigma)\mathbf{K}_{J,\sigma}\frac{p_\sigma - p_J}{d_{J,\sigma}}. \tag{5.2}$$

Solving for p_σ in (5.2) and using it in the first equation of (iii), we obtain

$$\int_\sigma \mathbf{K}\nu_{I,\sigma} \cdot \nabla p \, ds = m(\sigma)\frac{\mathbf{K}_{I,\sigma}\mathbf{K}_{J,\sigma}}{\mathbf{K}_{I,\sigma}d_{J,\sigma} + \mathbf{K}_{J,\sigma}d_{I,\sigma}}(p_J - p_I).$$

Summing up on (i),(ii) and (iii), after approximations, the equation (5.1) is equivalent to

$$-\sum_{\sigma=\bar{I}\cap\bar{J}} \tau_{I,J}(p_J - p_I) - \sum_{\substack{\sigma\subset\Gamma_D \\ \sigma\subset\partial I}} \tau_{I,\sigma}(f_1(x_\sigma) - p_I) - \sum_{\substack{\sigma\subset\Gamma_N \\ \sigma\subset\partial I}} m(\sigma)f_2(x_\sigma) = \sum_{\sigma\subset\partial I} m(\sigma)\mathbf{K}\rho\mathbf{g}\cdot\nu_{I,\sigma} + m(I)f(x_I), \tag{5.3}$$

where $\tau_{I,J} = m(\sigma)\dfrac{\mathbf{K}_{I,\sigma}\mathbf{K}_{J,\sigma}}{\mathbf{K}_{I,\sigma}d_{J,\sigma} + \mathbf{K}_{J,\sigma}d_{I,\sigma}}$, $\tau_{I,\sigma} = m(\sigma)\frac{\mathbf{K}_{I,\sigma}}{d_{I,\sigma}}$ and the integrations on Γ_N and Γ_D are null if $\partial I \cap \partial\Omega = \emptyset$. Writing the equation (5.3) for all the control volumes, we get a system of equations that can be put in the following form:

$$\mathbf{M}_h P_h = F_h, \tag{5.4}$$

where $\mathbf{M}_h$ is a matrix whose coefficients are linear combinations of $\tau_{I,J}$ and $\tau_{I,\sigma}$, P_h is the vector of unknowns (approximated solutions over the control volumes), and F_h the contributions from the boundaries terms, the gravity terms and the function f. The subscript h denotes the size of the mesh.

5.1.2 Mixed finite element method (MFEM)

The finite volume methods are not the only methods that have the property of local conservation. The mixed finite element methods have also that property. However, instead of computing the approximation of one variable as do FVM, FEM, and FDM, it computes simultaneously the approximations of a variable and its gradient. In cases where the gradient of a variable is also relevant , MFEM gives a better approximation of the gradient [16, 21].

Mesh

The meshes for MFEM are the same as those for FEM. Here we take a nondegenerate conforming mesh from [4, 20, 22]. Let Ω be a bounded Lipschitz domain in $\mathbb{R}^d$ with polygonal or polyhedral boundary Γ. We denote by $\mathcal{P}_h$ any mesh on Ω of size h, consisting of d-simplices (polygons in two dimensions, and polyhedra in three dimensions) satisfying the following:

(i) For any d-simplex $I \in \mathcal{P}_h$, $\mathring{I} \neq \emptyset$, I is closed, and $\bar{\Omega} = \cup_{I \in \mathcal{P}_h} I$.

(ii) For any two d-simplices $I, J \in \mathcal{P}_h$, $\mathring{I} \cap \mathring{J} = \emptyset$.

(iii) If $\sigma = I \cap J$, $I, J \in \mathcal{P}_h$; then σ is a full common face, edge or vertex for both I and J,

(iv) If $\sigma \subset I \cap \Gamma$, then $\sigma \subset \Gamma_D$ or $\sigma \subset \Gamma_N$.

(v) There exists h_0 such that $\dfrac{\operatorname{diam}(I)}{\rho_I} \leq h_0$ where ρ_I is the radius of the largest closed ball contained in I.

Discretization

Let $H_h \hookrightarrow L^2(\Omega)$ and $Q_h \hookrightarrow H_{0,N}(\operatorname{div}, \Omega)$ be finite-dimensional subspaces of $L^2(\Omega)$ and $H_{0,N}(\operatorname{div}, \Omega)$ respectively. Let $N = \dim(Q_h)$, $M = \dim(H_h)$, $\{\mathbf{v}_{ih} \,|\, 1 \leq i \leq N\}$ the basis of Q_h and $\{q_{ih} \,|\, 1 \leq i \leq M\}$ the basis of H_h. We thus look for a couple $(\mathbf{u}_h, p_h) \in Q_h \times H_h$ such that

$$\begin{cases} A(\mathbf{u}_h, \mathbf{v}_h) + B(\mathbf{v}_h, p_h) = D_1(\mathbf{v_h}), \, \forall \, \mathbf{v}_h \in Q_h, \\ B(\mathbf{u}_h, q_h) = D_2(q_h), \, \forall \, q_h \in H_h \end{cases} \tag{5.5}$$

where the bilinear forms A and B, and the linear form D_1 and D_2 are defined in Chapter 3. We now define

$$a_{i,j} = A(\mathbf{v}_{ih}, \mathbf{v}_{jh}), \tag{5.6}$$

$$b_{i,j} = B(\mathbf{v}_{jh}, q_{i,j}), \tag{5.7}$$

$$g_i = D_1(\mathbf{v}_{ih}), \tag{5.8}$$

$$f_i = D_2(q_{ih}), \tag{5.9}$$

and we denote by $A_h = (a_{ij})$, $B_h = (b_{ij})$, $f = (f_i)$ and $g = (g_i)$. Writing $\mathbf{u}_h$ and p_h in their basis

$$\mathbf{u}_h = \sum_{i=1}^{N} u_i \mathbf{v}_{ih},$$

$$p_h = \sum_{i=1}^{M} p_i q_{ih},$$

and denoting (by abuse of notation) by $\mathbf{u}_h = (u_i)$ and $\mathbf{p}_h = (p_i)$, the system (5.5) can be put in the matrix form

$$\begin{cases} A_h\mathbf{u}_h + B_h^T\mathbf{p}_h = g, \\ B_h\mathbf{u}_h = f, \end{cases} \tag{5.10}$$

or

$$\begin{bmatrix} A_h & B_h^T \\ B_h & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_h \\ \mathbf{p}_h \end{bmatrix} = \begin{bmatrix} \mathbf{g} \\ \mathbf{f} \end{bmatrix} \tag{5.11}$$

The basis $\{q_{ih} \,|\, 1 \leq i \leq M\}$ are built using the FEM basis and the basis $\{\mathbf{v}_{ih} \,|\, 1 \leq i \leq N\}$ is the Raviart-Thomas basis.

In fact, for $k \in \mathbb{N}_0$ let

$$\mathbb{P}_k(\Omega) := \{L : \Omega \to \mathbb{R} |\, L \text{ is a polynomial and } \deg(L) \leq k\}.$$

The Raviart-Thomas space on Ω of order k is defined by

$$RT_k(\Omega) = [\mathbb{P}_k(\Omega)]^d + \mathbb{P}_k(\Omega)x,$$

that is $L \in RT_k(\Omega)$ if only if there exist $L_1, ..L_d \in \mathbb{P}_k(\Omega)$ such that

$$L(x) = \begin{bmatrix} L_1(x) \\ \vdots \\ L_d(x) \end{bmatrix} + L_0(x) \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}.$$

We can see that $RT_k(\Omega) \subset H(\mathrm{div}, \Omega)$. Now a projection operator $\prod$ is used on $RT_k(\Omega)$ to built a finite-dimensional subspace Q_h of $H(\mathrm{div}, \Omega)$ (as it usually performed in FEM), such that

$$\prod(L)_{|I} \in RT_k(I)$$

for any $d-$simplex I and any $L \in RT_k(\Omega)$. The basis $\{\mathbf{v}_{ih} \,|\, 1 \leq i \leq N\}$ is built by means of the induced inner product on Q_h. More details on the discretization of MFEM and the construction of the basis can be found in [15, 14, 16].

5.2 Parabolic initial value problem

Spatial discretization of parabolic problems is usually done using FEM, FVM or FDM. In this section we present FVM scheme for diffusion-advection-reaction problems. We use the same mesh presented in definition (5.1.1) but the only difference is that we replace the permeability

tensor $\mathbf{K}$ by the diffusion tensor $\mathbf{D}$. We consider the problem (4.20a)-(4.20d) and we integrate (4.20a) over an arbitrary control volume I to obtain

$$\int_I \partial_t \mathcal{C}\, dx - \int_I \nabla \cdot (\mathbf{D}\nabla \mathcal{C})\, dx + \int_I \mathbf{u}\mathcal{C}\, dx = \int_I Q(\mathcal{C})\, dx. \tag{5.12}$$

We perform the approximation of each integral on each control volume.

(i) For the integral $\int_I \nabla \cdot (\mathbf{D}\nabla \mathcal{C})\, dx$, we proceed as we did in the elliptic case. Hence the vector $\Big[\int_I \nabla\cdot(\mathbf{D}\nabla\mathcal{C})\, dx\Big]_{I\in\mathcal{M}}$ of the integrals on each control volume can be approximated by the terms

$$\Big[\int_I \nabla \cdot (\mathbf{D}\nabla \mathcal{C})\, dx\Big]_{I\in\mathcal{M}} \simeq \mathbf{M}_h \mathcal{C}_h + \mathbf{b}_h(t), \tag{5.13}$$

where $\mathbf{M}_h$ is a matrix, $\mathcal{C}_h$ is the vector of the approximations of the concentration at the centre of each control volume, and $\mathbf{b}_h$ is the vector of boundary terms.

(ii) For the integral $\int_I Q(\mathcal{C})\, dx$,performing the approximation as in the elliptic case, the vector $\Big[\int_I Q(\mathcal{C})\, dx\Big]_{I\in\mathcal{M}}$ can be approximated by

$$\Big[\int_I Q(\mathcal{C})\, dx\Big]_{I\in\mathcal{M}} \simeq \mathbf{q}_h(t, \mathbf{c}_h). \tag{5.14}$$

(iii) For the integral $\int_I \mathcal{C}\, dx$ we perform the following approximation:

$$\begin{aligned}\int_I \partial_t \mathcal{C}\, dx \simeq & m(I) \lim_{\triangle t\to 0} \frac{\mathcal{C}(x_I, t+\triangle t) - \mathcal{C}(x_I, t)}{\triangle t} \\ = \quad & m(I)\frac{d\,\mathcal{C}(x_I,t)}{dt}.\end{aligned} \tag{5.15}$$

Hence the vector $\Big[\int_I \partial_t \mathcal{C}\, dx\Big]_{I\in\mathcal{M}}$ can be approximated by

$$\Big[\int_I \partial_t \mathcal{C}\, dx\Big]_{I\in\mathcal{M}} \simeq \mathbf{V}_h \frac{d\,\mathbf{c}_h}{dt}, \tag{5.16}$$

where $\mathbf{V}_h$ is a diagonal matrix made of the volumes of the cells.

(iv) For the integral $\int_I \mathbf{u}\mathcal{C}\, dx$ we use the divergence theorem that yields

$$\int_I \mathbf{u}\mathcal{C}\, dx = \sum_{\sigma\subset\partial I} \int_\sigma \mathbf{u}\cdot\nu_{I,\sigma}\mathcal{C}\, ds. \tag{5.17}$$

We apply an upstream approximation to the term $\mathcal{C}$. Considering two control volumes I and J sharing the interface σ, we approximate $\mathcal{C}$ on σ by $\mathcal{C}_I$ if the fluid is flowing from I

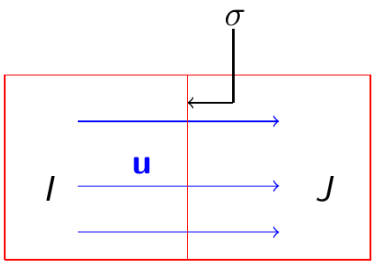


Figure 5.2: Upstream

to J, otherwise we approximate it by $\mathcal{C}_J$.

For instance, considering the Figure 5.2 and denoting by $\mathcal{C}_{\sigma+}$ the approximation of the concentration of the contaminant on σ by upstream technique, we have $\mathcal{C}_{\sigma+} \simeq \mathcal{C}_I$. From an upstream technique, we can approximate the flux across σ by,

$$\mathbf{u} \cdot \nu_{I,\sigma} \mathcal{C}_{\sigma,+} = \max(\mathbf{u} \cdot \nu_{I,\sigma}, 0)\mathcal{C}_I + \min(\mathbf{u} \cdot \nu_{I,\sigma}, 0)C_J, \text{ for } \sigma = \bar{I} \cap \bar{J}, \tag{5.18}$$

$$\mathbf{u} \cdot \nu_{I,\sigma} \mathcal{C}_{\sigma,+} = \max(\mathbf{u} \cdot \nu_{I,\sigma}, 0)\mathcal{C}_I + \min(\mathbf{u} \cdot \nu_{I,\sigma}, 0)\mathcal{C}_\sigma \text{ for } \sigma \subset \bar{I} \cap \partial\Omega\ . \tag{5.19}$$

Thus we have the approximation

$$\begin{aligned} \int_I \mathbf{u}\mathcal{C}\, dx &\simeq \sum_{\sigma \subset \partial I} \int_\sigma \mathbf{u} \cdot \nu_{I,\sigma} \mathcal{C}_{\sigma,+}\, ds, \\ &= \sum_{\sigma = \bar{I} \cap \bar{J}} m(\sigma)[\max(\mathbf{u} \cdot \nu_{I,\sigma}, 0)\mathcal{C}_I + \min(\mathbf{u} \cdot \nu_{I,\sigma}, 0)\mathcal{C}_J] \\ &+ \sum_{\sigma \subset \bar{I} \cap \partial\Omega} m(\sigma)[\max(\mathbf{u} \cdot \nu_{I,\sigma}, 0)\mathcal{C}_I + \min(\mathbf{u} \cdot \nu_{I,\sigma}, 0)\mathcal{C}_\sigma]. \end{aligned} \tag{5.20}$$

The vector $\left[\int_I \mathbf{u}\mathcal{C}\, dx\right]_{I \in \mathcal{M}}$ of the advection term on each control volume can be approximated as

$$\left[\int_I \mathbf{u}\mathcal{C}\, dx\right]_{I \in \mathcal{M}} \simeq \mathbf{N}_h \mathbf{c}_h + \mathbf{d}_h(t), \tag{5.21}$$

where $\mathbf{N}_h$ is a matrix and $\mathbf{d}_h$ is a vector from the boundaries terms.

Summing up on the approximations done in (5.13), (5.14), (5.16), and (5.21), the spatial descretization of problem (4.20a-4.20d) yields a system of ordinary differential equations that can be put in the form

$$\begin{cases} \dfrac{d\,\mathbf{c}_h}{dt} = \mathbf{U}_h \mathbf{c}_h + \mathbf{f}_h(t, \mathbf{c}_h) = \mathbf{r}_h(t, \mathbf{c}_h), \\ \mathbf{c}_h(t_0) = \mathbf{c}_0, \end{cases} \tag{5.22}$$

where $\mathbf{U}_h = \mathbf{V}_h^{-1}(\mathbf{M}_h - \mathbf{N}_h)$ and $\mathbf{f}_h = \mathbf{V}_h^{-1}(\mathbf{q}_h + \mathbf{b}_h - \mathbf{d}_h)$. In some problems, the right-hand side can be time-independent (for example in problems where the boundary condition is

time dependent). In this case the problem is called non-autonomous. If the right-hand side is independent of time, the problem is said to be autonomous. In that case the system (5.22) can be put in the simpler form:

$$\begin{cases} \dfrac{d\,\mathbf{c}_h}{dt} = \mathbf{U}_h\mathbf{c}_h + \mathbf{f}_h(\mathbf{c}_h) = \mathbf{r}_h(\mathbf{c}_h), \\ \mathbf{c}_h(t_0) = \mathbf{c}_0. \end{cases} \tag{5.23}$$

After spatial discretization of a PDE we get a system of linear or nonlinear equations. A natural problem we may encounter now is the existence and the uniqueness of the solutions of those systems. Another problem that occurs is the convergence of the numerical schemes. In this work we do not address those issues, we simply assume the existence of the numerical solutions and the convergence of the numerical schemes. Those topics are mainly discussed in [4, 14, 23]. For example in [23], it is proven that finite volume method with two-point flux approximation have one order of convergence.

Chapter 6

Temporal discretization

The spatial discretization of time-dependent problems leads in some cases to a system of ODEs in high dimension as shown in the previous chapter. The time integrators are used to the resulting ODEs to obtain the fully discretized equations. However, some time integrators perform better than others, depending on the type of the problem, in term of accuracy, efficiency and stability.

In this chapter, we first give some definitions and then present some time integrators. More precisely we will present some explicit methods, a semi-implicit method, the $\theta-$method, implicit Runge-Kutta, Rosenbrock methods and some exponential time stepping methods.

Definition (Mesh)**.**
Given an interval of time $[0, T]$, a mesh over the interval is a finite sequence $0 = t_0 < t_1 < ... < t_n = T$. We set $\tau_i = t_i - t_{i-1}$, $i = 1, .., n$.

Definition (Stiff ODEs)**.**
The definition of stiff ODEs is difficult to formulate clearly. In general, an ODE is stiff if it contains some terms that vary rapidly, that is the magnitudes of some eigenvalues of the Jacobian are very large. For a stiff problem, explicit methods do not perform well unless the step size is sufficiently small. For example the problem

$$\begin{cases} y' = ky, \\ y_0 = 1, \end{cases} \tag{6.1}$$

with exact solution $y(t) = e^{kt}$, is stiff as reported in [36]. More details on stiff problems can be found in [25].

Definition (Stability)**.**
The stability of a time integrator gives information on how well it performs on the problem

(6.1). After applying a time integrator, to this problem, we can obtain a scheme of the form

$$\begin{cases} y^{n+1} = (\phi(k\tau))^{n+1} y_0, \text{ if } \tau \text{ is constant,} \\ y^{n+1} = \phi(k\tau) y^n \text{ otherwise,} \end{cases} \tag{6.2}$$

where ϕ is called the stability function of the integrator and y^n is the approximate solution at time t_n. We note that the exact solution $y(t) = e^{kt} \longrightarrow 0$ as $t \longrightarrow +\infty$ for $k \in \mathbb{C}$ with $|Re(k)| < 1$.

The time integrator is said to be A-stable if it has the same behavior as the exact solution for constant step size; that is for a fixed τ, $y^n = (\phi(k\tau))^n y_0 \longrightarrow 0$ as $n \longrightarrow \infty$. This only happens if $|Re(\phi(k))| < 1$. The set $\{k \in \mathbb{C} || Re(\phi(k))| < 1\}$ is called the stability region.
The method is said to be L-stable if it is A-stable and $|\phi(k)| \longrightarrow 0$ as $k \longrightarrow \infty$.

Definition (Accuracy)**.**
The accuracy of a method is the rate of convergence of the numerical solution to the exact solution. A method is said to be $n-$order accurate if the error $E(\tau)$ (Error from discretization in time) is of the form $E(\tau) = C\tau^n$, where $C > 0$ and τ is the maximum step-size of the discretization.

6.1 Explicit Euler method and Runge-Kutta methods

6.1.1 Simplest Explicit Euler method

The oldest and simplest numerical integrator that has been successfully used for many non-stiff problems is the forward Euler method. When applied to the problem (5.22), we get the fully discretized equation of problem (4.20a)-(4.20d), which is

$$\mathbf{c}_h^{n+1} = \mathbf{c}_h^n + \tau_n \mathbf{r}_h(t_n, \mathbf{c}_h^n). \tag{6.3}$$

The scheme (6.3) is only first order accurate for the problem (5.22). To obtain a higher order of accuracy, explicit Runge-Kutta methods can be used.

6.1.2 Explicit Runge-Kutta methods (ERK)

The Runge Kutta methods have been constructed by Runge (1895) and Heun (1900) and finally formulated by Kutta (1901) [36]. They are one-step methods with internal stages. An s-stage

(ERK)for problem (5.22) is defined as follows:

$$\begin{aligned}
\mathbf{c}_h^{n+1} &= \mathbf{c}_h^n + \tau_n \sum_{i=1}^{s} b_i \mathbf{k}_{ni}, \qquad (6.4)\\
\mathbf{k}_{n1} &= \mathbf{r}_h(t_n, \mathbf{c}_n),\\
\mathbf{k}_{n2} &= \mathbf{r}_h(t_n + c_2\tau_n, \mathbf{c}_h^n + \tau_n a_{21}\mathbf{k}_{n1}),\\
\mathbf{k}_{n3} &= \mathbf{r}_h(t_n + c_3\tau_n, \mathbf{c}_h^n + \tau_n(a_{31}\mathbf{k}_{n1} + a_{32}\mathbf{k}_{n2}))\\
&\vdots\\
\mathbf{k}_{ns} &= \mathbf{r}_h(t_n + c_s\tau_n, \mathbf{c}_h^n + \tau_n(a_{s1}\mathbf{k}_{n1} + a_{s2}\mathbf{k}_{n2} + ... + a_{(s-1)1}\mathbf{k}_{n(s-1)})).
\end{aligned}$$

The vectors $\mathbf{k}_{ni}$, $i = 1, .., s$ are called internal stages. The coefficients b_i, c_i, a_{ij}, $i, j = 1, ..., s$ are the coefficients that determine the methods and can be summarized in the so-called **Butcher tableau**.

$$\begin{array}{c|ccccc}
c_1 & 0 & & & & \\
c_2 & a_{21} & 0 & & & \\
c_3 & a_{31} & a_{32} & 0 & & \\
\vdots & & & & \ddots & \\
c_s & a_{s1} & a_{s2} & \cdots & a_{s(s-1)} & 0 \\
\hline
 & b_1 & b_2 & \cdots & & b_s
\end{array}
\iff
\begin{array}{c|c}
\mathbf{c} & \mathbf{A} \\
\hline
 & \mathbf{b}^T
\end{array}$$

Table 6.1: Butcher tableau (ERKM)

The matrix $\mathbf{A}$ is a strictly lower triangle matrix, $c_i = \sum_{j=1}^{i-1} a_{ij}$, and the coefficients a_{ij} and b_i are given by the order conditions of . We do not give the order conditions here; these can be found in [24].

Example 6.1.1.

An example of an ERKM of 4-stages is given by

$$\begin{aligned}
\mathbf{c}_h^{n+1} &= \mathbf{c}_h^n + \tau_n(\frac{1}{6}\mathbf{k}_{n1} + \frac{1}{3}\mathbf{k}_{n2} + \frac{1}{3}\mathbf{k}_{n3} + \frac{1}{6}\mathbf{k}_{n4}),\\
\mathbf{k}_{n1} &= \mathbf{r}_h(t_n, \mathbf{c}_n),\\
\mathbf{k}_{n2} &= \mathbf{r}_h(t_n + \frac{1}{2}\tau_n, \mathbf{c}_h^n + \frac{1}{2}\tau_n\mathbf{k}_{n1}),\\
\mathbf{k}_{n3} &= \mathbf{r}_h(t_n + \frac{1}{2}\tau_n, \mathbf{c}_h^n + \frac{1}{2}\tau_n\mathbf{k}_{n2}),\\
\mathbf{k}_{n4} &= \mathbf{r}_h(t_n + \tau_n, \mathbf{c}_h^n + \tau_n\mathbf{k}_{n3}).
\end{aligned}$$

This method is usually denoted by (RK4) and is of order 4. More details on this method and its order equation can be found in [25, 36].

The major drawback of explicit methods is that they are unstable for stiff problems unless the time step size is sufficiently small. Some methods that can be used for stiff problems are presented in the next section

6.2 Semi-implicit method and $\theta-$methods

6.2.1 Semi-implicit method

The semi-implicit scheme for the problem (5.22) is given by

$$\mathbf{c}_h^{n+1} = \mathbf{c}_h^n + \tau_n(\mathbf{U}_h\mathbf{c}_h^{n+1} + \mathbf{f}_h(t_n, \mathbf{c}_h^n)). \tag{6.5}$$

Rearranging (6.5) we get

$$\mathbf{c}_h^{n+1} = (\mathbf{I} - \tau_n\mathbf{U}_h)(\mathbf{c}_h^n + \tau\mathbf{f}_h(t_n, \mathbf{c}_h^n)), \tag{6.6}$$

where $\mathbf{I}$ is the identity matrix. This method is more stable comparing to Euler the backward method that we have presented in the next subsection. There is only the need to solve a linear system every time step. However the method is of order one accuracy in time [1].

6.2.2 Euler backward method and $\theta-$methods

Let $\theta \in (0, 1]$. The $\theta-$method for the problem (5.22) is given by

$$\mathbf{c}_h^{n+1} = \mathbf{c}_h^n + \tau_n\Big(\theta\mathbf{r}_h(t_{n+1}, \mathbf{c}_h^{n+1}) + (1-\theta)\mathbf{r}_h(t_n, \mathbf{c}_h^n)\Big). \tag{6.7}$$

This method is implicit since we have $\mathbf{c}_h^{n+1}$ in both sides of the scheme. Every time step, we need to solve a nonlinear system of equations, and Newton's method is mostly used for that.

We note that for $\theta = 1$ we have the standard Euler backward method, which is

$$\mathbf{c}_h^{n+1} = \mathbf{c}_h^n + \tau_n\mathbf{r}_h(t_{n+1}, \mathbf{c}_h^{n+1}). \tag{6.8}$$

We can also note that for $\theta = 1/2$, we have the so-called trapezoidal rule given by

$$\mathbf{c}_h^{n+1} = \mathbf{c}_h^n + \tau_n\Big(\frac{1}{2}\mathbf{r}_h(t_{n+1}, \mathbf{c}_h^{n+1}) + \frac{1}{2}\mathbf{r}_h(t_n, \mathbf{c}_h^n)\Big). \tag{6.9}$$

The major drawback of the $\theta-$methods is that they are of order one accuracy in time, except for $\theta = 1/2$ which is of order two. Furthermore, Newton's iterations may not be convergent in some circumstances where for example the first guess for Newton's method is badly chosen. To obtain higher order of accuracy and avoid Newton iterations, some other methods are proposed in the next section.

6.3 Implicit Runge-Kutta and Rosenbrock methods

The main focus of this section is the Rosenbrock methods which are derived from implicit Runge-Kutta methods. We present first the latter.

6.3.1 Implicit Runge-Kutta methods (IRK)

An s-stage implicit Runge-Kutta method applied to the problem (5.22) is given by

$$\begin{aligned} \mathbf{c}_h^{n+1} &= \mathbf{c}_h^n + \tau_n \sum_{i=1}^{s} b_i \mathbf{k}_{ni}, \\ \mathbf{k}_{ni} &= \mathbf{r}_h(t_n + c_i\tau_n, \mathbf{c}_h^n + a_{i1}\mathbf{k}_{n1} + a_{i2}\mathbf{k}_{n2} + ... + a_{is}\mathbf{k}_{ns}),\ i = 1, .., s. \end{aligned} \tag{6.10}$$

The matrix $\mathbf{A} = (a_{ij})_{s\times s}$ is no longer lower triangular but rather a full matrix. More on this method can be found in [25].

6.3.2 Rosenbrock type methods

Among the methods which are used for stiff problems, Rosenbrock methods are the easiest to program. They replace the nonlinear system of equations by a linear system. In the literature, Rosenbrock methods are referred to as linearly implicit Rung-Kutta methods, although they are also called semi-implicit, generalized, modified, adaptive or additive Runge-Kutta methods. Here we present the steps that lead to this class of methods. We start with the diagonally implicit method scheme applied to the problem (5.22) that yields

$$\begin{cases} \mathbf{c}_h^{n+1} = \mathbf{c}_h^n + \tau_n \sum_{i=1}^{s} b_i \mathbf{k}_{ni}, \\ \mathbf{k}_{n1} = \mathbf{r}_h(t_n + c_1\tau_n, \mathbf{c}_h^n + \tau_n a_{11}\mathbf{k}_{n1} \\ \mathbf{k}_{n2} = \mathbf{r}_h(t_n + c_2\tau_n, \mathbf{c}_h^n + \tau_n(a_{21}\mathbf{k}_{n1} + a_{22}\mathbf{k}_{n2})), \\ \quad \vdots \\ \mathbf{k}_{ns} = \mathbf{r}_h(t_n + c_s\tau_n, \mathbf{c}_h^n + \tau_n(a_{s1}\mathbf{k}_{n1} + a_{s2}\mathbf{k}_{n2} + ... + a_{ss}\mathbf{k}_{ns})), \end{cases} \tag{6.11}$$

with its associated Butcher tableau

$$\begin{array}{c|cccc} c_1 & a_{11} & 0 & \cdots & 0 \\ c_2 & a_{21} & a_{22} & 0 & 0 \\ c_3 & a_{31} & a_{32} & a_{33} & \\ \vdots & & & \ddots & \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} \iff \begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^T \end{array}$$

Table 6.2: Butcher tableau (Diagonally Implicit)

Now we split the matrix $\mathbf{A}$ into two lower triangular matrices $\mathbf{A} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}$ where $\boldsymbol{\alpha} = (\alpha_{ij})$ is such that $\alpha_{ij} = 0$ for $j \geq i$, and $\boldsymbol{\Gamma} = (\gamma_{ij})$ is such that $\gamma_{ij} = 0$ for $j > i$. We set $\alpha_i = \sum_{j=1}^{i-1} \alpha_{ij}$ and $\gamma_i = \sum_{j=1}^{i} \gamma_{ij}$. From table (6.2), we have $c_i = \alpha_i + \gamma_i$ and $a_{ij} = \alpha_{ij} + \gamma_{ij}$. For the i^{th}-stage value $1 \leq i \leq s$, we have, with the matrices α and Γ,

$$\mathbf{k}_{ni} = \mathbf{r}_h(t_n + c_i\tau_n, \mathbf{c}_h^n + \tau_n \sum_{j=1}^{i-1} a_{ij}\mathbf{k}_{nj} + \tau_n a_{ii}\mathbf{k}_{ni}) \tag{6.12}$$

$$= \mathbf{r}_h(t_n + \alpha_i\tau_n + \gamma_i\tau_n, \mathbf{c}_h^n + \tau_n \sum_{j=1}^{i-1} \alpha_{ij}\mathbf{k}_{nj} + \tau_n \sum_{j=1}^{i} \gamma_{ij}\mathbf{k}_{nj}). \tag{6.13}$$

Instead of using Newton iterations to solve for $\mathbf{k}_{ni}$ in the nonlinear equation (6.12), the clue for Rosenbrock methods, is to linearize $\mathbf{k}_{ni}$ in the equation (6.13) with respect to γ_i and Γ. However, in the linearization the exact Jacobians are replaced by the Jacobian at $(t_n, \mathbf{c}_h^n)$. This approximation avoids the computation of the Jacobian at each stage; only one Jacobian is computed at each time step and used for all stages of the time step. Hence the linearization with the approximated Jacobian yields

$$\begin{aligned} \mathbf{k}_{ni} &= \mathbf{r}_h(t_n + \alpha_i\tau_n + \gamma_i\tau_n, \mathbf{c}_h^n + \tau_n \sum_{j=1}^{i-1} \alpha_{ij}\mathbf{k}_{nj} + \tau_n \sum_{j=1}^{i} \gamma_{ij}\mathbf{k}_{nj}), \\ &\simeq \mathbf{r}_h(t_n + \alpha_i\tau_n, \mathbf{c}_h^n + \tau_n \sum_{j=1}^{i-1} \alpha_{ij}\mathbf{k}_{nj}) + \tau_n \mathbf{J}_n \sum_{j=1}^{i} \gamma_{ij}\mathbf{k}_{nj} + \gamma_i\tau_n \mathbf{r}_{nt}(t_n, \mathbf{c}_h^n), \end{aligned} \tag{6.14}$$

where $\mathbf{J}_n$ is the Jacobian of $\mathbf{r}_h$ at $(t_n, \mathbf{c}_h^n)$ and $\mathbf{r}_{ht}$ is the partial derivative of $\mathbf{r}_h$ with respect to

time. Rearranging equation (6.14), the s-stage Rosenbrock method is given by

$$\begin{cases} \mathbf{c}_h^{n+1} = \mathbf{c}_h^n + \tau_n \sum_{i=1}^{s} b_i \mathbf{k}_{ni} \\ (\mathbf{I} - \tau_n \gamma_{ii} \mathbf{J}_n)\mathbf{k}_{ni} = \mathbf{r}_h(t_n + \alpha_i \tau_n, \mathbf{c}_h^n + \tau_n \sum_{j=1}^{i-1} \alpha_{ij} \mathbf{k}_{nj}) + \tau_n \mathbf{J}_n \sum_{j=1}^{i-1} \gamma_{ij} \mathbf{k}_{nj} + \gamma_i \tau_n \mathbf{r}_{nt}(t_n, \mathbf{c}_h^n). \end{cases} \tag{6.15}$$

Usually in the implementation of Rosenbrock methods on computers, we should avoid the matrix-vector multiplication $\mathbf{J}_n \sum_{i=1}^{s} \gamma_{ij} \mathbf{k}_{ni}$. Hence we introduce the variable $\mathbf{k}'_{ni} = \tau_n \sum_{i=1}^{s} \gamma_{ij} \mathbf{k}_{ni}$ (see [25, 26, 36]). From this variable we have

$$\mathbf{k}_{ni} = \frac{1}{\gamma_{ii}} \mathbf{k}'_{ni} - \frac{1}{\gamma_{ij}} \sum_{j=1}^{i-1} \mathbf{k}_{nj}. \tag{6.16}$$

Introducing (6.16) in (6.15) leads to the mainly used Rosenbock scheme

$$\begin{cases} \mathbf{c}_h^{n+1} = \mathbf{c}_h^n + \tau_n \sum_{i=1}^{s} m_i \mathbf{k}'_{ni}, \\ (\frac{\mathbf{I}}{\tau_n \gamma_{ii}} - \mathbf{J}_n)\mathbf{k}'_{ni} = \mathbf{r}_h(t_n + \alpha_i \tau_n, \mathbf{c}_h^n + \frac{1}{\tau_n} \sum_{j=1}^{i-1} q_{ij} \mathbf{k}'_{nj}) - \frac{1}{\tau_n} \sum_{j=1}^{i-1} \beta_{ij} \mathbf{k}'_{nj} + \gamma_i \tau_n \mathbf{r}_{nt}(t_n, \mathbf{c}_h^n), \end{cases} \tag{6.17}$$

where $(m_1, .., m_s) = (b_1, .., b_s)\Gamma^{-1}$, $(q_{ij}) = \mathbf{Q} = \alpha \mathbf{\Gamma}$ and $(\beta)_{ij} = \beta = \mathbf{\Gamma}^{-1}$.

To get satisfactory accuracy in time integrations, constant time steps often turn out to be inefficient. Time step control is then essential to improve the efficiency. The embedded formula is mainly used in Rosenbrock schemes. It considers two Rosenbrock schemes of order p and $\hat{p}$ with $\hat{p} = p - 1$. To get the lower-order scheme, the coefficients b_i are modified. Then the time step $\tau_n = t_{n+1} - t_n$ is chosen with respect to the norm of the difference $\mathbf{c}_h^n - \hat{\mathbf{c}}_h^n$. Here we give some examples of the Rosenbrock methods with an embedded methods.

Example 6.3.1.

We give a 2-stage method known as Ros2 and a 3-stage method referred to as Ros3p in [26, 40], and designed for nonlinear parabolic problems. They have order 2 and 3 respectively with embedded methods. The schemes are given by

$$\begin{cases} \mathbf{c}_h^{n+1} = \mathbf{c}_h^n + \tau_n \sum_{i=1}^{s} m_i \mathbf{k}'_{ni}, \\ \hat{\mathbf{c}}_h^{n+1} = \hat{\mathbf{c}}_h^n + \tau_n \sum_{i=1}^{s} \hat{m}_i \mathbf{k}'_{ni}, \\ (\frac{\mathbf{I}}{\tau_n \gamma} - \mathbf{J}_n)\mathbf{k}'_{ni} = \mathbf{r}_h(t_n + \alpha_i \tau_n, \mathbf{c}_h^n + \frac{1}{\tau_n} \sum_{j=1}^{i-1} q_{ij} \mathbf{k}'_{nj}) - \frac{1}{\tau_n} \sum_{j=1}^{i-1} \beta_{ij} \mathbf{k}'_{nj} + \gamma_i \tau_n \mathbf{r}_{nt}(t_n, \mathbf{c}_h^n). \end{cases} \tag{6.18}$$

The coefficients for Ros3p are in Table 6.3 and those for Ros2 are in Table 6.4. The order conditions that lead to these coefficients can be found in [25, 26, 40].

$\gamma = 7.886751345948129e-01$	
$q_{21} = 1.267949192431123e+00$	$\beta_{21} = 1.607695154586736e+00$
$q_{31} = 1.267949192431123e+00$	$\beta_{31} = 3.464101615137755e+00$
$q_{32} = 0.000000000000000e+00$	$\beta_{32} = 1.732050807568877e+00$
$\alpha_1 = 0.000000000000000e+00$	$\gamma_1 = 7.886751345948129e-01$
$\alpha_2 = 1.000000000000000e+00$	$\gamma_2 = -2.113248654051871e-01$
$\alpha_3 = 1.000000000000000e+00$	$\gamma_3 = -1.077350269189626e+00$
$m_1 = 2.000000000000000e+00$	$\hat{m}_1 = 2.113248654051871e+00$
$m_2 = 5.773502691896258e-01$	$\hat{m}_2 = 2.000000000000000e+00$
$m_3 = 4.226497308103742e-01$	$\hat{m}_3 = 4.226497308103742e-01$

Table 6.3: Ros3p coefficients

$\gamma = 1.707106781186547e+00$	
$q_{11} = 0.00e+00$	$\beta_{11} = 5.857864376269050e-01$
$q_{21} = 5.857864376269050e-01$	$\beta_{21} = 1.171572875253810e+00$
$q_{22} = 0.00e+00$	$\beta_{22} = 5.857864376269050e-01$
$\alpha_1 = 0.000000000000000e+00$	$\gamma_1 = 1.707106781186547e+00$
$\alpha_2 = 1.000000000000000e+00$	$\gamma_2 = -1.707106781186547e+00$
$m_1 = 8.786796564403575e-01$	$\hat{m}_1 = 5.857864376269050e-01$
$m_2 = 2.928932188134525e-01$	$\hat{m}_2 = 0.00e+00$

Table 6.4: Ros2 coefficients

6.4 Exponential methods

Exponential solvers are not new in fact, but they have been regarded as impractical. Krylov subspace approximations of the matrix exponential operator have been found useful in the 1980s. Since then a number of classes of exponential integrators have been implemented. Here we recall that the exact solution of problem (5.22) within the analytic semigroup framework in the interval $[t_n, t_{n+1}]$ of time is given by

$$\mathbf{c}_h(\tau) = e^{(\tau \mathbf{U}_h)}\mathbf{c}(t_n) + \int_0^{\tau} e^{(\tau - s)\mathbf{U}_h}\mathbf{f}_h(\mathbf{c}(s), s)ds,\ \tau \in [t_n, t_{n+1}]. \tag{6.19}$$

As we cannot compute the integral term exactly, approximations are needed. Here we present some of them.

6.4.1 Exponential Time Differencing (ETD)

The technique of ETD is to approximate $\mathbf{f}_h(\mathbf{c}_h(s), s)$ in (6.19) by a polynomial in the interval $[t_n, t_{n+1}]$. The simplest ETD, denoted by ETD1, is first-order accurate [28, 29] and is such that $\mathbf{f}_h(\mathbf{c}(s), s)$ is approximated by $\mathbf{f}_h(\mathbf{c}_h^n, t_n) = \mathbf{f}_h^n$. Applied to the problem (5.22), ETD1 yields

$$\begin{aligned} \mathbf{c}_h^{n+1} &= e^{\tau_n \mathbf{U}_h}\mathbf{c}_h^n + \int_0^{\tau_n} e^{(\tau_n - s)\mathbf{U}_h}\mathbf{f}_h^n ds, \\ &= e^{\tau_n \mathbf{U}_h}\mathbf{c}_h^n + \mathbf{U}_h^{-1}\mathbf{f}_h^n(e^{\tau_n \mathbf{U}_h} - I). \end{aligned} \tag{6.20}$$

The scheme (6.20) is commonly put in the form

$$\mathbf{c}_h^{n+1} = \mathbf{c}_h^n + \tau_n \varphi_1(\tau_n \mathbf{U}_h)(\mathbf{f}_h^n + \mathbf{U}_h \mathbf{c}_h^n), \tag{6.21}$$

where φ_1 is defined by

$$\begin{cases} \varphi_1(\mathbf{U}) = \mathbf{U}^{-1}\left(e^{\mathbf{U}} - \mathbf{I}\right), \text{ if } \mathbf{U} \text{ is invertible,} \\ \varphi_1(\mathbf{U}) = \displaystyle\sum_{i=1}^{\infty} \frac{\mathbf{U^{i-1}}}{i!} \text{ otherwise.} \end{cases}$$

To achieve a higher order of ETD, the approximation of $\mathbf{f}_h(\mathbf{c}_h(s), s)$ uses information about $\mathbf{f}_h$ at n^{th} and previous time steps, and hence this leads to a multi-step exponential method. An arbitrary order of ETD can be found in [29]. The problem with ETD methods of high order is that they are multi-step while only one value is available at the initial time. To avoid this problem, exponential Runge-Kutta or exponential Rosenbrock can be used.

6.4.2 Exponential Runge-Kutta methods

Exponential Runge-Kutta methods are similar to explicit-implicit Runge-Kutta methods with internal stages. A general scheme of s-stage of exponential Runge-Kutta for equation (6.19) is given by

$$\begin{cases} \mathbf{c}_h^{n+1} = e^{\tau_n \mathbf{U}_h}\mathbf{c}_h^n + \tau_n \displaystyle\sum_{i=1}^{s} b_i(\tau_n \mathbf{U}_h)\mathbf{f}_h^{ni}, \\ \mathbf{f}_h^{ni} = \mathbf{f}_h(\tau_n + c_i\tau_n, \mathbf{c}_h^{ni}), \\ \mathbf{c}_h^{ni} = e^{c_i\tau_n \mathbf{U}_h}\mathbf{c}_h^n + \tau_n \displaystyle\sum_{j=1}^{s} a_{ij}(\tau_n \mathbf{U}_h)\mathbf{f}_h^{nj}, \end{cases} \tag{6.22}$$

where c_i, b_i and a_{ij}, $1 \leq i, j \leq s$ are the parameters that determine the method. As usual, these parameters are given by the order conditions [31, 30, 29, 24, 25]. The coefficients a_{ij} and b_i are linear combinations of the entire functions φ_k defined by

$$\varphi_k(z) = \int_0^1 e^{(1-\theta)z} \frac{\theta^{k-1}}{k-1}\, d\theta,\ k \geq 1.$$

The functions φ_k satisfy $\varphi_k(0) = \dfrac{1}{k!}$ and the recurrence relation

$$\varphi_{k+1}(z) = \frac{\varphi_k(z) - \varphi_k(0)}{z},\ \varphi_0(z) = e^z.$$

The coefficients of the method have to satisfy the conditions

$$\sum_{i=1}^{s} b_i(z) = \varphi_1(z), \quad \sum_{j=1}^{s} a_{ij}(z) = c_i \varphi_1(c_i z),\ 1 \leq i \leq s. \tag{6.23}$$

A rearrangement of (6.22) yields a simpler scheme [30, 32, 33]

$$\begin{cases} \mathbf{c}_h^{n+1} = \mathbf{c}_h^n + \tau_n \sum_{i=1}^{s} b_i(\tau_n \mathbf{U}_h)(\mathbf{f}_h^{ni} + \mathbf{U}_h \mathbf{c}_h^n), \\ \mathbf{f}_h^{ni} = \mathbf{f}_h(t_n + c_i\tau_n, \mathbf{c}_h^{ni}), \\ \mathbf{c}_h^{ni} = \mathbf{c}_h^n + \tau_n \sum_{j=1}^{s} a_{ij}(\tau_n \mathbf{U}_h)(\mathbf{f}_h^{nj} + \mathbf{U}_h \mathbf{c}_h^n). \end{cases} \tag{6.24}$$

Exponential Runge-Kutta methods can be implicit or explicit [31, 33]. In the case of explicit methods, the scheme (6.23) reduces to

$$\begin{cases} \mathbf{c}_h^{n+1} = \mathbf{c}_h^n + \tau_n \sum_{i=1}^{s} b_i(\tau_n \mathbf{U}_h)(\mathbf{f}_h^{ni} + \mathbf{U}_h \mathbf{c}_h^n), \\ \mathbf{f}_h^{ni} = \mathbf{f}_h(t_n + c_i\tau_n, \mathbf{c}_h^{ni}), \\ \mathbf{c}_h^{ni} = \mathbf{c}_h^n + \tau_n \sum_{j=1}^{i-1} a_{ij}(\tau_n \mathbf{U}_h)(\mathbf{f}_h^{nj} + \mathbf{U}_h \mathbf{c}_h^n). \end{cases} \tag{6.25}$$

Here we give an example of an explicit exponential Runge-Kutta method.

Example 6.4.1.

The simplest example is the method with one stage. From conditions (6.23), we have

$$\begin{cases} b_1 = \varphi_1, \\ c_1 = 0, \\ \mathbf{c}_h^{n1} = \mathbf{c}_h^n, \\ \mathbf{f}_h^{n1} = \mathbf{f}_h^n = \mathbf{f}_h(t_n, \mathbf{c}_h^n). \end{cases}$$

Hence the scheme is given by

$$\mathbf{c}_h^{n+1} = \mathbf{c}_h^n + \tau_n \varphi_1(\tau_n \mathbf{U}_h)(\mathbf{f}_h^n + \mathbf{U}_h \mathbf{c}_h^n). \tag{6.26}$$

Other examples of s-stage with $s > 1$ and high order can be found in [33].

The explicit exponential Runge-Kutta methods have been shown to be efficient for parabolic problems that have small remainder, or when the remainder is bounded with respect to the operator of the linear part. In our case, the remainder is $\mathbf{f}_h$ and the operator is $\mathbf{U}_h$. However, these methods turn out to be inefficient in some problems. For instance, in the case in which the numerical solution stays near an equilibrium point (e.g, a saddle point) of a long time problem, the integrator is forced to take small time steps due to the stability requirements (see [32] and references therein). Exponential Rosenbrock-type methods overcome this issue.

6.4.3 Exponential Rosenbrock-type methods (EROW)

The EROW methods are based on the continuous linearization on the ODEs along the numerical solutions at each time step. Consider for example the autonomous version of the problem (5.22) which is given by (5.23). The non-autonomous case can be recovered by some transformation that we state later.

The linearization of (5.23) at $\mathbf{c}_h^n$ yields

$$\begin{cases} \dfrac{d\,\mathbf{c}_h}{dt} = \mathbf{J}_h^n \mathbf{c}_h + \mathbf{g}_h(\mathbf{c}_h), \\ \mathbf{c}_h(t_0) = \mathbf{c}_0 \end{cases} \tag{6.27}$$

where $\mathbf{J}_h^n = \mathbf{U}_h + \dfrac{\partial \mathbf{f}_h}{\partial \mathbf{c}_h}(\mathbf{c}_h^n) = \dfrac{\partial \mathbf{r}_h}{\partial \mathbf{c}_h}(\mathbf{c}_h^n)$ and $\mathbf{g}_n(\mathbf{c}_h) = \mathbf{r}_h(\mathbf{c}_h) - \mathbf{J}_h^n \mathbf{c}_h$.

When applied to the system (6.27), the explicit exponential methods yield the so-called EROW given by

$$\begin{cases} \mathbf{c}_h^{n+1} = \mathbf{c}_h^n + \tau_n \displaystyle\sum_{i=1}^{s} b_i(\tau_n \mathbf{J}_h^n)(\mathbf{g}_h^{ni} + \mathbf{J}_h^n \mathbf{c}_h^n), \\ \mathbf{g}_h^{ni} = \mathbf{g}_h(t_n + c_i \tau_n, \mathbf{c}_h^{ni}), \\ \mathbf{c}_h^{ni} = \mathbf{c}_h^n + \tau_n \displaystyle\sum_{j=1}^{i-1} a_{ij}(\tau_n \mathbf{J}_h^n)(\mathbf{g}_h^{nj} + \mathbf{J}_h^n \mathbf{c}_h^n). \end{cases} \tag{6.28}$$

The methods are fully explicit. The simplest of this method class is given in the following example.

Example 6.4.2.

For $s = 1$ we get the simplest of EROW known as the exponential Rosenbrock-Euler method.

Its scheme for the autonomous problem (5.23) is given by

$$\begin{cases} \mathbf{c}_h^{n+1} = \mathbf{c}_h^n + \tau_n \varphi_1(\tau_n \mathbf{J}_h^n)(\mathbf{g}_h^n + \mathbf{J}_h^n \mathbf{c}_h^n), \\ \mathbf{g}_h^n = \mathbf{g}_h(\mathbf{c}_h^n). \end{cases} \tag{6.29}$$

The exponential Rosenbrock-Euler method is computationally attractive. It is second-order accurate and requires only one Jacobian computation per step. More details on EROW can be found in [30, 32] and references therein.

To extend the scheme (6.28) to non-autonomous problems, we add the trivial equation $t' = 1$ to the system (6.27), linearize the resulting system at $(t_n, \mathbf{c}_h^n)$, and then apply the explicit exponential Runge-Kutta again.

6.4.4 Krylov spaces

In the exponential methods developed above, a_{ij} and b_i, $1 \leq i, j \leq s$ are linear combinations of φ_k. The computation of $\varphi_k(\tau_n \mathbf{U}_h)$ for large systems of ODEs is a real challenge that made exponential methods non-practical for a long time. The implementation of Krylov space approximations and other techniques like Léja point [28] boosted the use of exponential methods. Here in this section, we only discuss Krylov space approximations.

For a given non-singular matrix $\mathbf{U}$ (in our case, $\mathbf{U} = \mathbf{U}_h$ or $\mathbf{U} = \mathbf{J}_h^n$) of dimension d and a given vector $\mathbf{v} \in \mathbb{R}^d$, the Krylov subspace method approximates $\varphi_k(\tau_n \mathbf{U})$ by an element in the so-called Krylov space:

$$\begin{aligned} K_m(\tau_{\mathbf{n}} \mathbf{U}, \mathbf{v}) &= \text{span}\{\mathbf{v}, (\tau_n \mathbf{U})\mathbf{v}, .., (\tau_n \mathbf{U})^{m-1}\mathbf{v}\}, \\ &= \text{span}\{\mathbf{v}, \mathbf{U}\mathbf{v}, .., \mathbf{U}^{m-1}\mathbf{v}\}, \\ &= K_m(\mathbf{U}, \mathbf{v}), \\ &= K_m, \end{aligned} \tag{6.30}$$

with an arbitrary $m << d$. The approximation is done by the projection of $\mathbf{U}$ onto an orthogonal basis $V_m = [\mathbf{v}_1, ..., \mathbf{v}_m]$ of the space K_m. The basis V_m is built in general by the Arnoldi algorithm, using a modified Gram-Schmidt process. In the case where the matrix $\mathbf{U}$ is symmetric, the Arnoldi process can be replaced by the Lanczos process for computational savings. Here we present the Arnoldi process; the details of the Lanczos process can be found in [34].

Algorithm: Arnoldi

1: initialize: $\mathbf{v}_1 = \mathbf{v}/||\mathbf{v}||_2$

2: Iterate: for j=1,..,m do

(a) Compute $\mathbf{w} = \mathbf{U}\mathbf{v}_j$

(b) for i=1,...,j do

$h_{ij} = \langle \mathbf{w}, \mathbf{v}_i \rangle$

$\mathbf{w} = \mathbf{w} - h_{ij} - h_{ij}\mathbf{v}_i$

(c) $h_{j+1,j} = ||\mathbf{w}||_2$

$\mathbf{v}_{j+1} = \mathbf{w}/h_{j+1,j}$.

The matrix $\mathbf{H}_m = [h_{ij}]$ from the Arnoldi process is called the upper Huessenberg matrix. It fulfils the relations [28, 34]

$$\mathbf{U}V_m = V_m\mathbf{H}_m + h_{m+1,m}\mathbf{v}_{m+1}(e_1^m)^T, \tag{6.31}$$

$$V_m^T\mathbf{U}V_m = \mathbf{H}_m, \tag{6.32}$$

where $h_{m+1,m}$ and v_{m+1} are recovered from the Arnoldi precess, and $\mathbf{e}_1^m$ is the first vector of $\mathbb{R}^m$ standard vector basis.

The value $\varphi_k(\tau_n\mathbf{U}\mathbf{v})$ is then approximated by

$$\varphi_k(\tau_n\mathbf{U}\mathbf{v}) \simeq ||\mathbf{v}||_2 V_{m+1}\varphi_k(\tau_n\bar{\mathbf{H}}_{m+1})\mathbf{e}_1^{m+1}, \tag{6.33}$$

where

$$\bar{\mathbf{H}}_m = \begin{bmatrix} & \mathbf{H}_m & & 0 \\ 0 & .. \quad 0 & h_{m+1,m} & 0 \end{bmatrix}. \tag{6.34}$$

What remains now is the computation of $\varphi_k(\tau_n\bar{\mathbf{H}}_{m+1})$. Since we have $\varphi_0(z) = e^z$ and $\varphi_{k+1}(z) = \dfrac{\varphi_k(z) - \varphi_k(0)}{z}$ for $z \neq 0$, $\varphi_k(\tau_n\bar{\mathbf{H}}_{m+1})$ can be recursively obtained once we find $e^{\tau_n\bar{\mathbf{H}}_{m+1}}$. The computation of $e^{\tau_n\bar{\mathbf{H}}_{m+1}}$ and the recursion formula can be found in [35].

Chapter 7

Numerical Simulations

After the development of the theories in the previous chapters, we present in this chapter a series of simulations on two main problems. All the problems are two-dimensional. We use FMV for spacial discretization for all the simulations and we use the implicit, semi-implicit, Ros2, Ros3p and θ-methods with $\theta = 0.5$ for temporal discretization. We also present the relative errors of the temporal discretization. However, we do not run any simulation with the exponential methods. A comparative work on the exponential methods and the methods we use in our simulations can be found in [28].

The Chapter is structured as follows. We will first present two simulations. The first simulation is a diffusion dominated advective and reactive transport problem in a heterogeneous porous medium, and the second simulation is the corresponding convection-dominated. Finally simulations on convection-dominated reactive transport are performed in homogeneous and isotropic medium for both the permeability and the diffusion tensor, and in anisotropic medium with heterogeneous permeability and constant diffusion tensor.

7.1 Transport in a heterogeneous porous medium

7.1.1 Problem setting

We consider the general cases set out in (2.15a)-(2.15d) and (4.20a)-(4.20d). Then we simplify these general cases to some specific cases as follows. For the fluid velocity, we neglect gravity and we assume that no fluid is released or consumed in the domain. Hence we get the resulting system from (2.15a)-(2.15d),

$$-\nabla.(\mathbf{K}(\nabla p) = 0, \tag{7.1a}$$

$$\mathbf{u} = -\mathbf{K}\nabla p, \tag{7.1b}$$

$$p = f_1 \text{ on } \Gamma_D, \tag{7.1c}$$

$$\mathbf{K}\nabla p \cdot \nu = 0 \text{ on } \Gamma_N, . \tag{7.1d}$$

We assume that there is no contaminant at the initial time in the domain. The reaction function is set to $Q(\mathcal{C}) = \dfrac{\mathcal{C}}{1+\mathcal{C}}$. Hence the model problem to be solved is given by

$$\partial_t \mathcal{C} - \nabla.(\mathbf{D}\nabla\mathcal{C} - \mathbf{u}\mathcal{C}) = \frac{\mathcal{C}}{1+\mathcal{C}}, \tag{7.2a}$$

$$\mathcal{C} = g_1 \text{ on } (0,T] \times \Gamma_D \tag{7.2b}$$

$$\mathbf{D}\nabla\mathcal{C}.\nu = 0 \text{ on } (0,T] \times \Gamma_N, \tag{7.2c}$$

$$\mathcal{C}(0,x) = 0, \quad x \in \bar{\Omega}. \tag{7.2d}$$

As shown in the Figure 7.1, we scale the functions f_1 and g_1 to 0 and 1. For the sake of simplicity, we take the permeability and the diffusion tensors to be diagonal in each control volume. More precisely we take

$$\mathbf{K} = d_1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{7.3}$$

$$\mathbf{D} = d_2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{7.4}$$

where d_1 and d_2 are positive real numbers.

7.1.2 Domain

In this section we consider the rectangular domain $\Omega = ABCD$ (Figure 7.1) with $AB = 10$, $BC = 1$. We set Neuman boundary conditions at Γ_2 and Γ_4, and Dirichlet boundary conditions at Γ_1 and Γ_3. We use a structured mesh of 2400 rectangular control volumes. Each control volume has as dimensions $\triangle_x \times \triangle_y$ with $\triangle_x = 0.25$ and $\triangle_y = 0.016$. We denote by ν_2 and ν_4 the outward normal vectors to Γ_2 and Γ_4 respectively.

The domain is heterogeneous for the permeability tensor (Figure 7.2a) and homogeneous for the diffusion tensor (Fig.7.2b) . In the Figure 7.2a, the permeability in the white regions is one thousand times the permeability in the blue region.

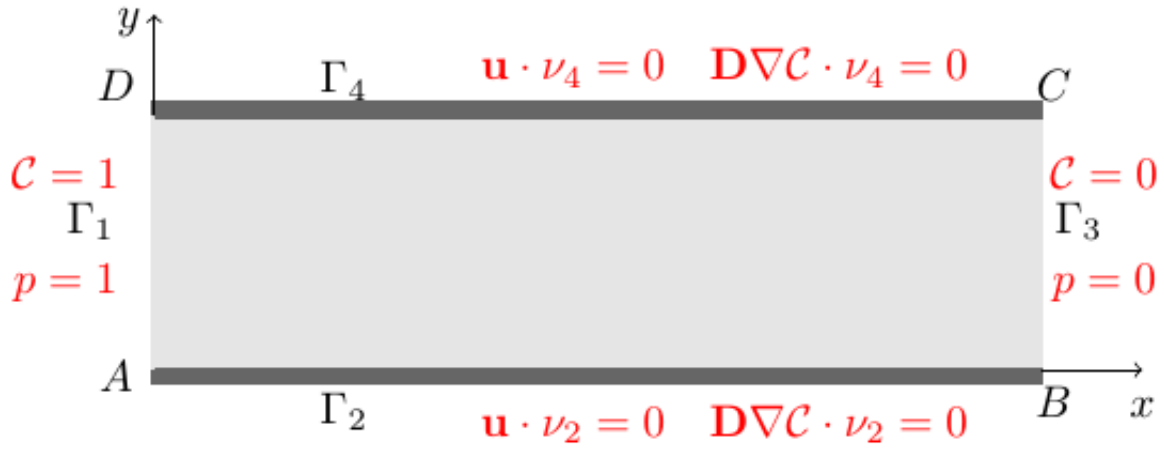
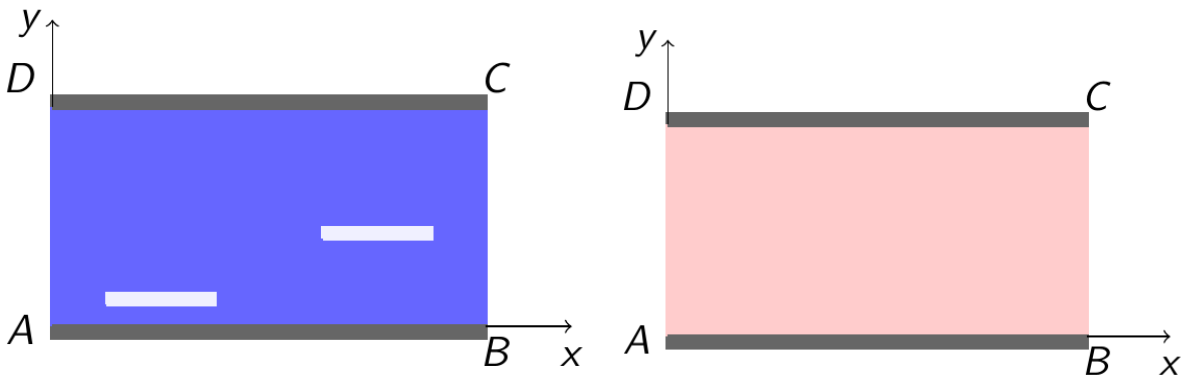


Figure 7.1: Domain

For the convection dominated simulation, $d_1 = 100$ in the blue region and $d_1 = 100000$ in the white blocks, while $d_2 = 0.1$ in the entire domain. For the diffusion dominated simulation, $d_1 = 0.1$ in the blue region and $d_1 = 100$ in the white blocks, and $d_2 = 100$ in the entire domain.

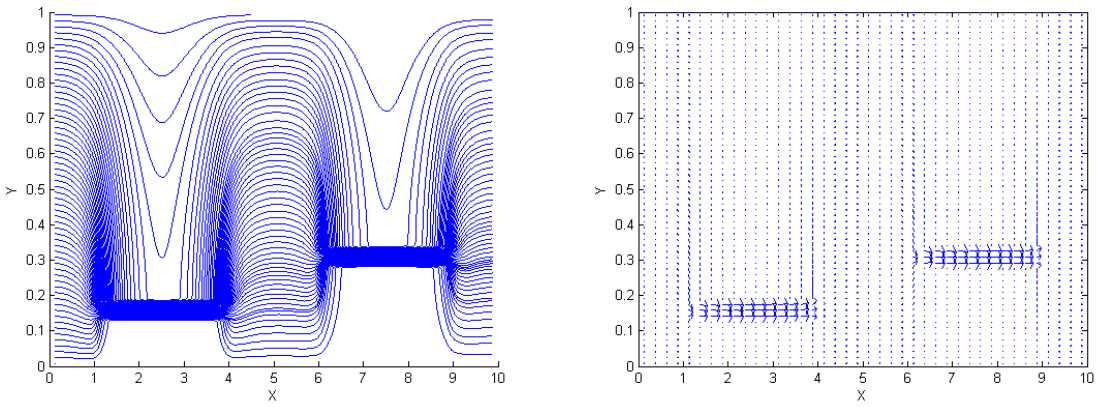


(a) Heterogeneous permeability (b) Homogeneous diffusion

Figure 7.2: Permeability and Diffusion tensors

7.1.3 Simulations

We run the simulations for the advection dominated and diffusion dominated cases. The streamlines and the velocity field are shown in the Figures (7.3a, 7.3b) in the both cases. The fluid mostly passes through the high permeable region (two white blocks in the Figure 7.2a). But this does not mean that the fluid does not passe through the other regions. As shown by the velocity field in the Figure 7.3b, the fluid flows through any region of the medium but at a different velocity.

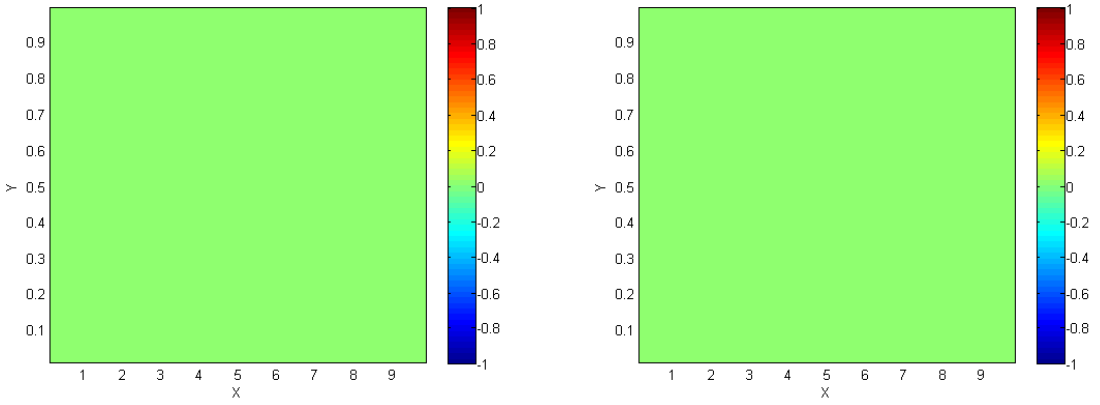


(a) streamlines (b) Velocity field

Figure 7.3: Streamlines & Velocity field

For the propagation of the contaminant across the region, we run the simulations on the interval of time $[0, 1]$, and we compare the concentrations of the contaminant at times $t = 0$, $t = 0.25$, $t = 0.5$ and $t = 1$.

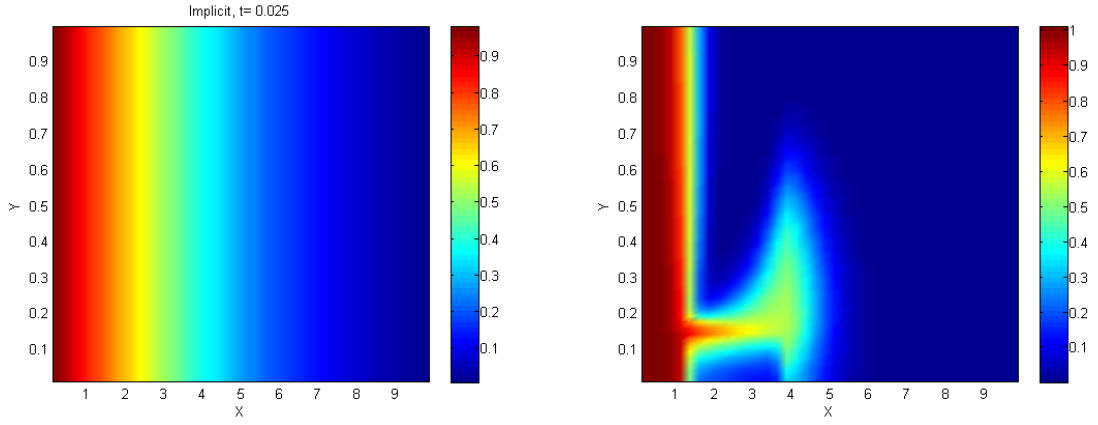
For the diffusion dominated case, the propagation of the contaminant is uniform. This is due to the fact that the diffusion tensor is uniformly distributed in the domain. For the advection dominated cases, the contaminant moves rapidly through the region with high permeability. This is due to the fact that the contaminant is driven by the fluid in motion.



(a) Diffusion dominated (b) Advection dominated

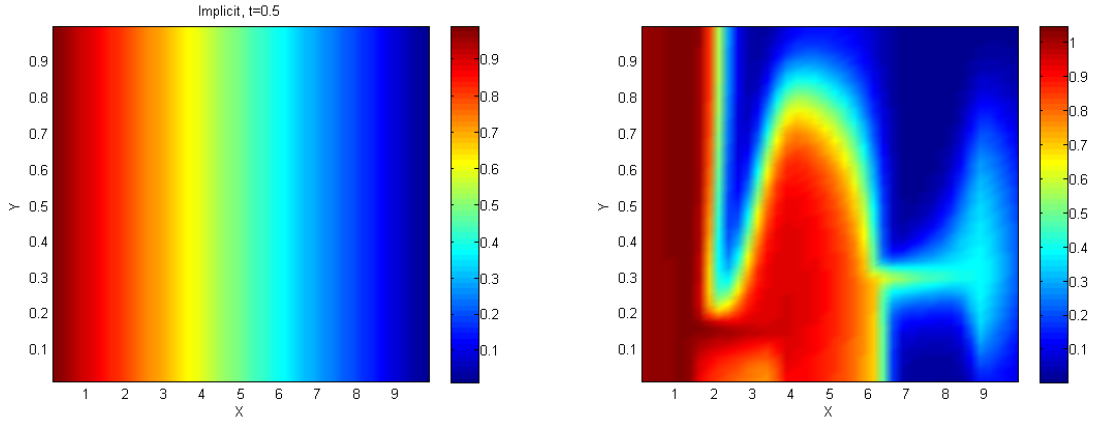
Figure 7.4: Concentration at time $t = 0$

We take as exact solutions for both simulations (diffusion dominated and advection dominated), the solutions with constant time step $\tau = 1/4000$ for all the time integrators. We then run the simulations with different constant time steps and we compute the relative L^2 error at the final time $t = 1$. We plot the rate of convergence of each time integrator in the Figure



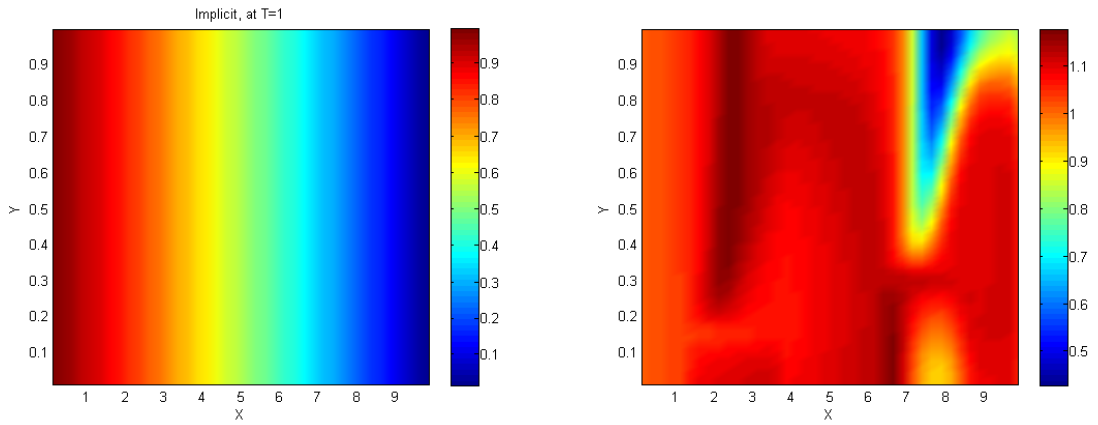
(a) Diffusion dominated (b) Advection dominated

Figure 7.5: Concentration at time $t = 0.1$



(a) Diffusion dominated (b) Advection dominated

Figure 7.6: Concentration at time $t = 0.5$

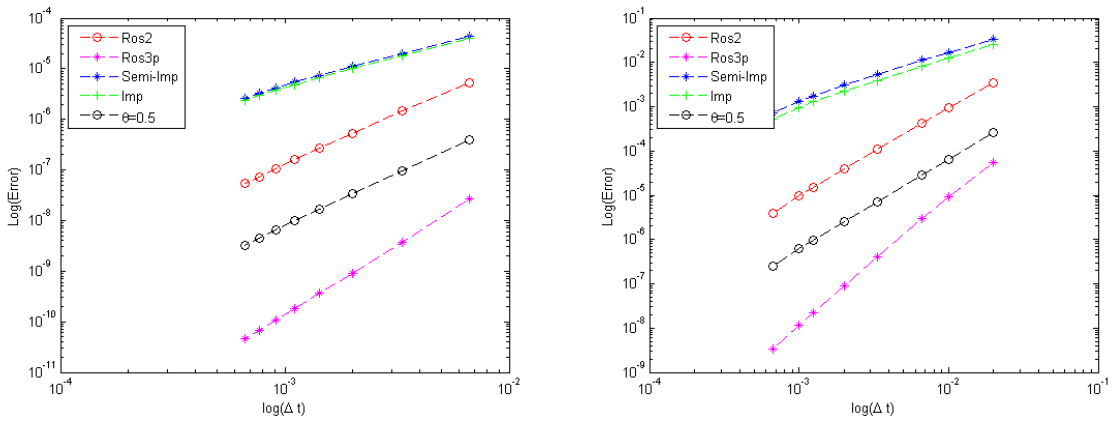


(a) Diffusion dominated (b) Advection dominated

Figure 7.7: Concentration at time $t = 1$

7.8. We notice that the orders of convergence in both simulations for each time integrator are roughly the same.

We observe an order of convergence of 1.21 for both implicit and semi-implicit methods, an order of convergence of 2.00 for Ros2, 2.06 for the $\theta-$method with $\theta = 0.5$ and an order of 2.87 for Ros3p.



(a) Diffusion dominated (b) Advection dominated

Figure 7.8: Convergence accuracy

7.2 Transport in anisotropic and isotropic media

7.2.1 Problem setting

We consider again the problems (7.1a)-(7.1d) and (7.2a)-(7.2d) with some changes. For an isotropic medium, the permeability tensor on each control volume is of the form

$$\mathbf{K} = d_1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

while for an anisotropic medium, the permeability tensor on each control volume is given by

$$\mathbf{K} = d_2 \begin{bmatrix} 1000 & 0 \\ 0 & 1 \end{bmatrix},$$

where d_1 and d_2 are positive real numbers. That is the permeability tensor in the x-direction is one thousand times the permeability in the y-direction for the anisotropic case. The diffusion tensor has the same form as the permeability tensor in the isotropic medium, that is

$$\mathbf{D} = d_3 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

For isotropic flow, we set $d_1 = 1000$ and $d_3 = 1$, while for anisotropic flow, we set $d_2 = 10$ and $d_3 = 1$.

7.2.2 Domain

In this section we consider the square domain $\Omega = ABCD$ (Figure 7.1). The domain has dimensions $AB = 5$, $BC = 5$. We set Neuman boundary conditions at Γ_2 and Γ_4, and some parts of Γ_1 and Γ_3. Only some parts of the boundary around the vertices A and C are set to Dirichlet boundary conditions as shown in the figure. We use a total number of 2500 rectangular control volumes. Each control volume has dimension $\triangle_x \times \triangle_y$ with $\triangle_x = 0.02$ and $\triangle_y = 0.02$.

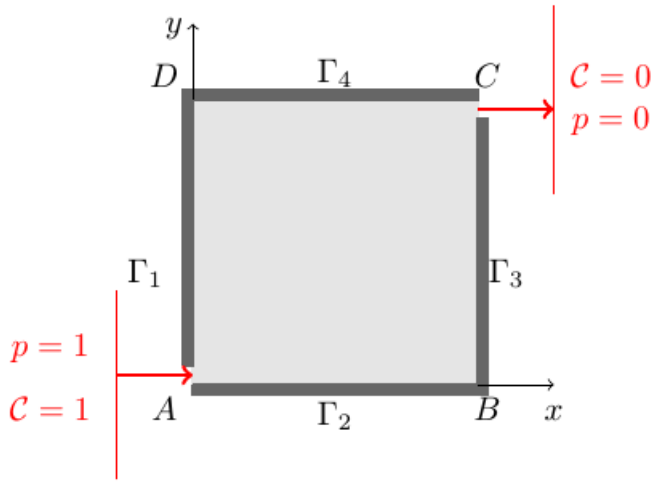


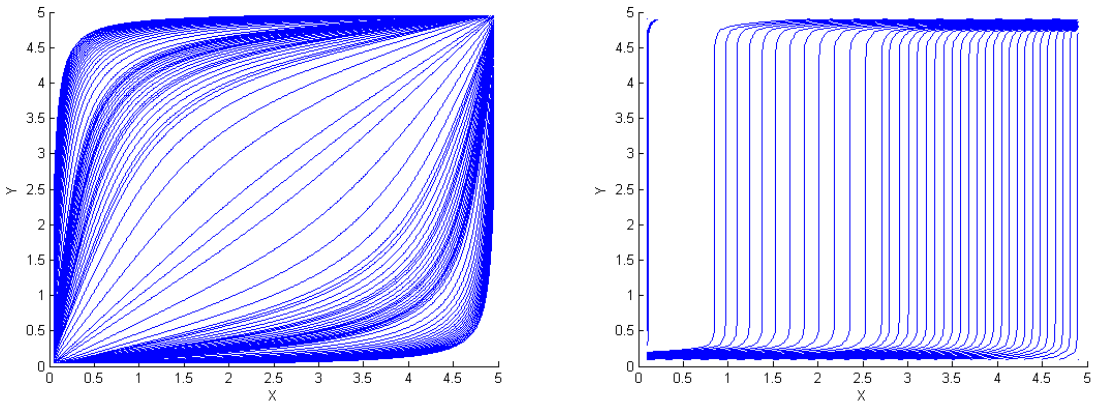
Figure 7.9: Domain

7.2.3 Simulations

In this section we run simulations for the isotropic and anisotropic media. The streamlines are shown in the Figure 7.10. For the isotropic medium, the streamlines are symmetric about the diagonal line AC. These properties are also shown in Figure 7.11. However, the velocity field varies in the different region of the domain.

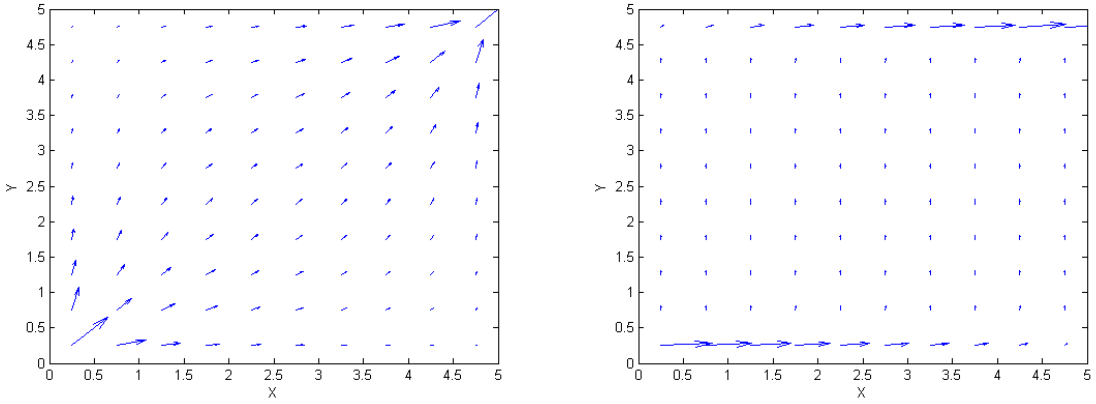
The spread of the contaminant follows mainly the path of the flows since the problems are advection dominated. We run the simulations over the interval $[0, 1]$, and we compare the concentrations of the contaminant at times $t = 0.0$, $t = 0.1$, $t = 0.5$ and $t = 1$.

In the isotropic medium, the contaminant moves uniformly in x and $y-$directions, while in the anisotropic medium, the contaminant moves rapidly in the $x-$direction and roughly uniformly in the $y-$direction.



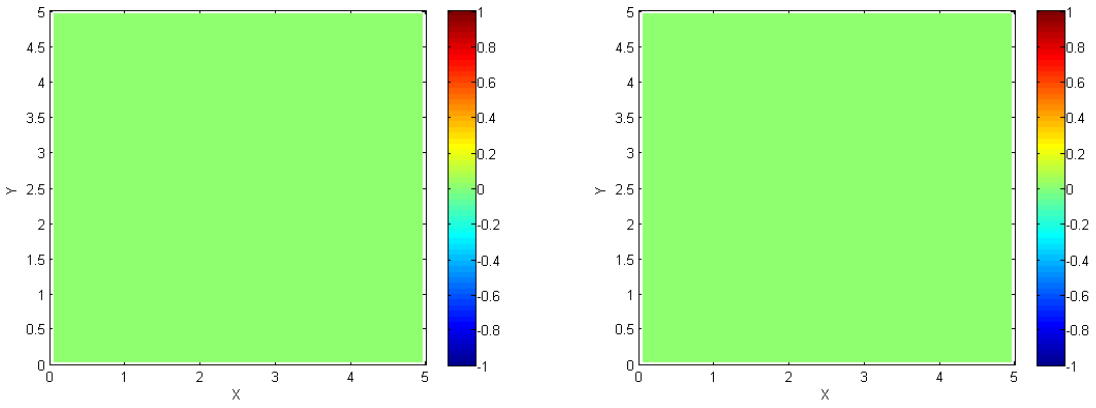
(a) Isotropic medium (b) Anisotropic medium

Figure 7.10: Streamlines



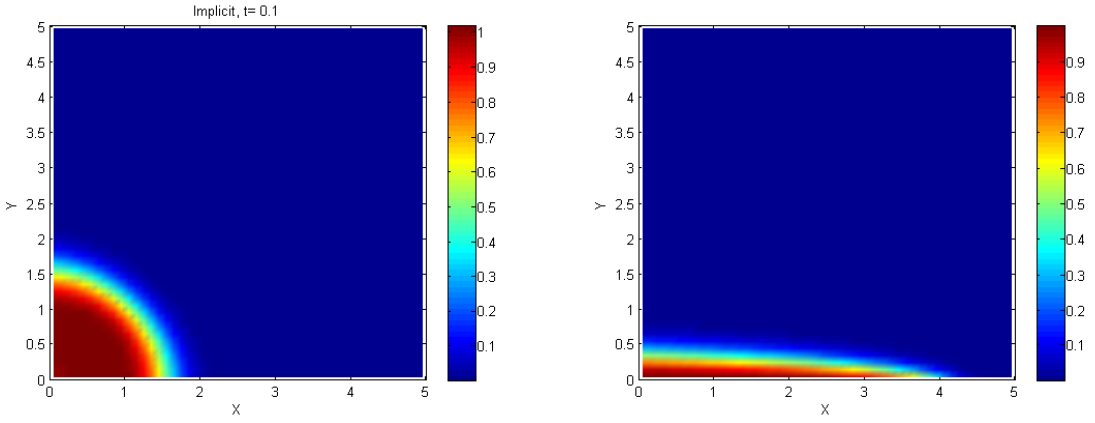
(a) Isotropic medium (b) Velocity field anisotropic medium

Figure 7.11: Velocity fields



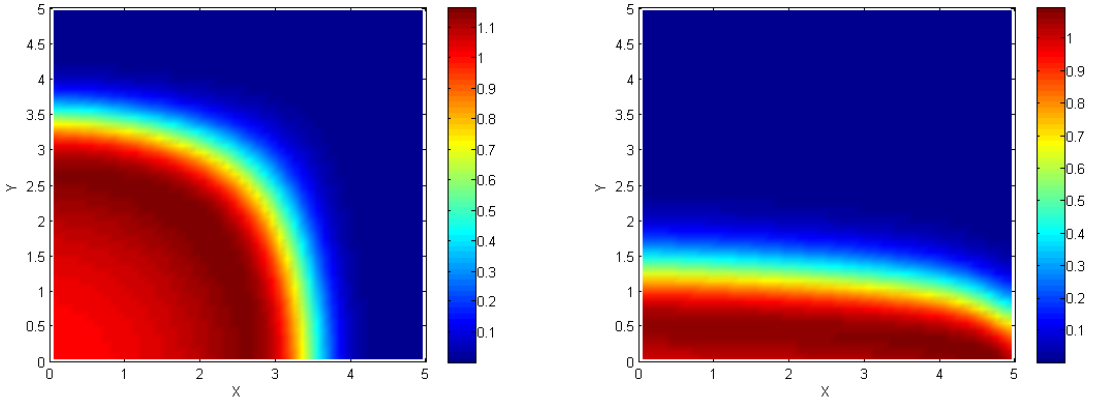
(a) Isotropic medium (b) Anisotropic medium

Figure 7.12: Concentrations at $t = 0$



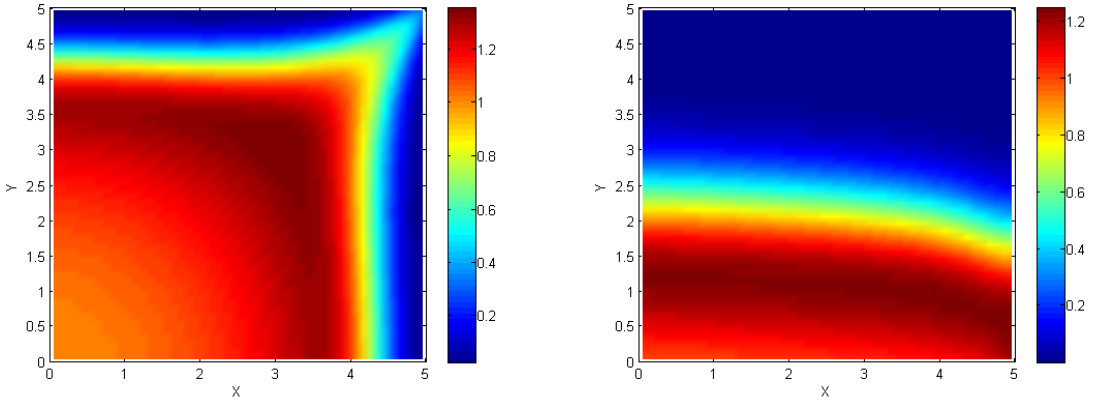
(a) Isotropic medium (b) Anisotropic

Figure 7.13: Concentration at $t = 0.1$



(a) Isotropic (b) Anisotropic

Figure 7.14: Concentration at $t = 0.5$

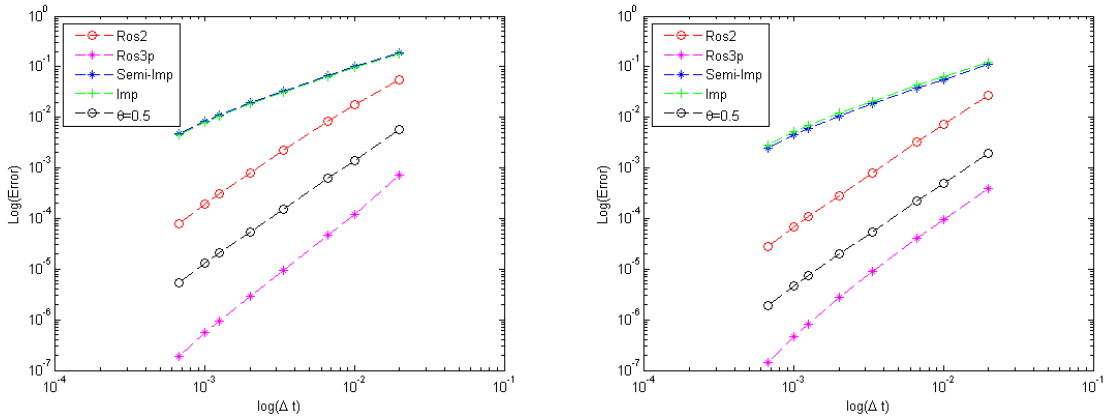


(a) Isotropic (b) Anisotropic

Figure 7.15: Concentrations at $t = 1$

Here again as in the previous Section, we take as exact solutions, the solutions with constant time step $t = 1/4000$. We then run the simulation for different times and we compute the L^2 relative errors at the final time $t = 1$. We also plot the relative errors for each time integrator in Figure 7.15. We note that the orders of convergence in both simulations for each time integrator are nearly the same.

We observe an order of convergence of 1.09 for both semi implicit and implicit methods, 2.03 for the $\theta-$method and 2.02 for Ros2. But unexpectedly, we observe an order reduction for Ros3p, which is 2.38.



(a) Error anisotropic medium (b) Error isotropic medium

Figure 7.16: Error

For the parabolic problems, the fully discretized equations bear two type of errors, the error from space discretization and the error from time discretization. We made some brief comments on those errors in the previous chapters.

Several authors worked on the combination of the two errors. Recently it has been proven in [41] that the error of the fully discretized parabolic problems using FVM with two-point flux approximation and the ETD1 with constant time step is of the form

$$\| \mathbf{c}(t_n) - \mathbf{c}_h^n \|_{0,h} \leq \beta(\| \mathbf{c}_0 - \mathbf{c}_{0,h} \|_{0,h} + \tau + h), \tag{7.5}$$

where β is a constant depending on the parameters of the problem and $\| \cdot \|_{0,h}$ is a norm defined in [41]. A similar work with FEM and Rosenbrock type integrators can be found in [40], and with FVM and the $\theta-$methods can be found in [4].

Chapter 8

Conclusion

In this thesis we have considered some simple model of fluid flow and transport in porous media. The model is formulated using the laws of conservation of momentum and mass as the total energy is assumed to be constant. More precisely, we have established one phase flow of incompressible fluid which is an elliptic PDE, and transport equation of a contaminant due by advection, diffusion and reaction processes, which is a parabolic PDE. Standard variational formulation and mixed formulation of the elliptic PDE have been used to ensure the existence and uniqueness of the pressure and Darcy 's velocity, while the semigroup theory has been used to ensure the existence and uniqueness of the transport equation.

Afterward we have presented some numerical schemes to approximate the pressure/ Darcy's velocity, and the solute concentraction in transport equation. More precisely, we have presented finite volume method with two-point approximation and mixed finite element for spatial discretization of the elliptic PDE, and the finite volume method combined with implicit methods, Rosenbrock methods and exponential methods for temporal discretization of the transport equation.

Numerical simulations with diffusion-dominated and advection-dominated problems in heterogeneous media have been performed. We have compared the order of convergence of different time integrators. We have found that standard semi-implicit scheme and implicit method are order one in time, Ros2 and the $\theta-$methods with $\theta = 1/2$, are order two in time while Ros3p is order three in time.

We also have ran advection-dominated problem in both an anisotropic and an isotropic media. In both simulations, we found similar order of convergence for standard implicit and semi-implicit methods, the $\theta-$methods with $\theta = 1/2$ and Ros2 as in the previous two simulations. Nevertheless, unexpectedly the order for Ros3p dropped to two in both simulations.

We do not know the causes of this reduction of the order, but we remark that there is a great difference between the boundary conditions for simulations in Section 7.1 and Section 7.2.

This work can be extended in many ways. Here we mention very few. Firstly, we can investigate whether the boundary conditions participate in the reduction of the order for Ros3p, and if so check whether they are the sufficient conditions for this phenomena as we know that boundary conditions are very important when dealing with PDEs.

Another possible extension of this work is to deal with the case where the reaction term is a discontinuous function. In this case, the discretization leads to a system of discontinuous ordinary differential equations . We may now use adequate time integrators or schemes designed for this type of problem.

Finally, the time integrators we used in our simulations including the exponential methods are mainly used for transport problems in porous media. There are also well designed Rosenbrock method for Partial Algebraic Equations (PAEs), especially when the matrix $\partial_{\mathcal{C}} h(\mathcal{C})$ is degenerated; this type of problem occurs sometimes in subsurface reservoir simulation. However, well designed exponential methods for PAEs are not yet available. This also can be on of our future work.

Appendix

General notations

$\mathbb{N} = \{1, 2, 3...\}$	natural numbers set
$\mathbb{N}_0 = \{0, 1, 2, 3...\}$	
$\Omega \subset \mathbb{R}^d$	an open bounded domain
$\partial\Omega = \Gamma$	boundary of Ω
$\partial\Omega_D = \Gamma_D$	Dirchlet boundary
$\partial\Omega_N = \Gamma_N$	Neumann boundary
ν	unit outward normal to Γ
$d \in \mathbb{N}$	dimension of the domain Ω
$x = (x_1, .., x_q)$	
$\alpha = (\alpha_1, ..., \alpha_d) \in \mathbb{N}_0^d$	$\lvert\alpha\rvert = \alpha_1 + ... + \alpha_d$
$\partial_{x_i} f = \frac{\partial f}{\partial_{x_i}}$	$i = 1, 2..., d$
$\partial_x^\alpha f = \frac{\partial^{\lvert\alpha\rvert} f}{\partial_{x_1}^{\alpha_1} ... \partial_{x_d}^{\alpha_d}}$	
$\partial_\nu u$	outward normal derivative of u
$\boldsymbol{supp}\, u$	support of u
$\nabla u = (\partial_{x_1} u, ..., \partial_{x_d} u)$	
$\nabla . q = \mathrm{div}(q) = \partial_{x_1} q_1 + ... + \partial_{x_d} q_d$	
$v.w = v_1 w_1 + .. + v_d w_d$	
$\mathcal{D}(\Omega)$	space of test functions on Ω
μ	generic viscosity
$\mathbf{K}$	generic permeability tensor
$\mathcal{K} = \frac{\mathbf{K}}{\mu}$	
$\mathbf{D}$	generic diffusion tensor
$\mathbf{u}$	Darcy's velocity
$\mathbf{O}$	generic operator
$\mathcal{D}_m(\mathbf{O})$	the domain of operator $\mathbf{O}$

Notation on spaces

$L^p(\Omega)$	$= \{u : \Omega \longrightarrow \mathbb{R} \Big\vert \int_\Omega \vert u\vert^p < \infty, \quad 1 \leq u < \infty\}$	
$L^\infty(\Omega)$	$= \{u : \Omega \longrightarrow \mathbb{R} \Big\vert \vert u\vert < \infty \quad \text{a.e. in } \Omega\}$	
$\mathbf{V}$	generic vector space, Banach space or Hilbert space	
$\mathbf{V}'$	topological dual of $\mathbf{V}$	
$\mathbf{V}^\perp$	orthogonal of $\mathbf{V}$	
H_0	$= \{u \in H^1(\Omega) \,\vert\, u = 0 \text{ on } \Gamma\}$	
$H_{0,D}$	$= \{u \in H^1(\Omega) \,\vert\, u = 0 \text{ on } \Gamma_D\}$	
$H_{0,N}$	$= \{u \in H^1(\Omega) \,\vert\, \partial_\nu u = 0 \text{ on } \Gamma_N\}$	
$H(\text{ div}, \Omega)$	$= \{q \in (L^2(\Omega))^d \,\vert\, \text{ div}(q) \in (L^2(\Omega))^d\}$	
$H_{0,N}(\text{div}, \Omega)$	$= \{q \in H(\text{ div}, \Omega) \,\vert\, q \cdot \nu = 0 \text{ on } \Gamma_N\}$	
$\vert q\vert$	$= (q_1^2 + .. + q_m^2)^{\frac{1}{2}}$	$q = (q_1, ..., q_m)$, $m \in \mathbb{N}$, $q_i : \Omega \to \mathbb{R}$
$\vert u\vert_{m,p}$	$= \sum_{\vert\alpha\vert=m} \Big(\int_\Omega \vert \partial^\alpha u\vert^p dx \Big)^{\frac{1}{p}}$	semi-norm on $L^p(\Omega)$
$\vert .\vert_m$	$= \vert .\vert_{m,2}$	
$\vert\vert u\vert\vert_{m,p}$	$= \sum_{\vert\alpha\vert\leq m} \Big(\int_\Omega \vert \partial^\alpha u\vert^p dx \Big)^{\frac{1}{p}}$	norm on $H^m(\Omega)$
$\vert\vert .\vert\vert_{\mathbf{V}}$		norm on the space $\mathbf{V}$
$\langle , \rangle_0$		scalar product on $L^2(\Omega)', \times L^2(\Omega)$
$\langle , \rangle_1$		scalar product on $H^{1'} \times H^1$
$\vert\vert .\vert\vert_0$	$= \vert\vert .\vert\vert_{0,2}$	norm on $L^2(\Omega)$ induced by $\langle , \rangle$
$\vert\vert .\vert\vert_1$	$= \vert\vert .\vert\vert_{1,2}$	norm on H^1 induced by $\langle , \rangle_1$
$\vert\vert q\vert\vert_{\text{div}}$	$= \Big(\vert q\vert_{0,2}^2 + \vert \text{div}(q)\vert_{0,2}^2 \Big)^{\frac{1}{2}}$	norm on $H(\text{div}, \Omega)$

Bibliography

[1] Tambue, A. (2010). Efficient Numerical Schemes for Porous Media Flow. PhD Thesis, Heriot-Watt University.

[2] Bear,J., & Cheng, AH-D. (2010). Modeling Groundwater Flow and Contaminant Transport. Theory and Applications of Transport in Porous Media. Springer, Dordrecht Heidelberg London, New York.

[3] Angelini, O., Brenner, K., & Hilhorst, D. (2013). A Finite Volume Method on General Meshes for a Degenerate Parabolic Convection-Reaction-Diffusion Equation. Numerische Mathematik 123(2), 219-257.

[4] Knabner, P., & Angermann, L.(2003). Numerical Methods for Elliptic and Parabolic Partial Differential Equations. Texts in Applied Mathematics, Springer-Verlag, New York.

[5] Isham, CS.(1989). Lectures on groups and vectors spaces for physicists. World scientific, Singapore.

[6] Reddy, BD. (1998). Introductory Functional Analysis With Applications to Boundary Value Problems and Finite Elements. Texts in Applied Mathematics. Springer-Verlag, New York.

[7] Zuily, C. (1988). Problems in Distributions and Partial Differential Equations. North-Holland Mathematics Studies, Elsevier.

[8] Friedlander, FG., & Joshi, MS. (1998). Introduction to Theory of Distributions. Cambridge University Press, Cambridge.

[9] Strichartz, R. (1994). A Guide to Distribution Theory and Fourier Transforms, Studies in Advanced Mathematics, World Scientific Publishing Co Inc.

[10] Brezis, H. (2010). Functional Analysis, Sobolev Spaces and Partial Differential Equations, Springer Science & Business Media.

[11] Tartar, L. (2007). An Introduction to Sobolev Spaces and Interpolation Spaces. Lecture Notes of the Unione Matematica Italiana, Springer Science & Business

[12] Haroske, D., & Triebel, H.(2008). Distributions, Sobolev Spaces, Elliptic Equations. Textbooks in Mathematics. European Mathematical Society.

[13] Ziemer, WP.(1989) Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation.Graduate texts in mathematics, Springer-Verlag, New york.

[14] Brezzi, F., & Fortin, M. (1991). Mixed and Hybrid Finite Element Methods. Springer Series in Computational Mathematics, Springer-Verlag, New york.

[15] Gatica, GN. (2014). A Simple Introduction to the Mixed Finite Element Method Theory and Applications. Springer Briefs in Mathematics, Springer, London.

[16] Brezzi, F. (1986). New Applications of Mixed Finite Element Methods. In Proceedings of the International Congress of Mathematicians, Berkeley, California, USA.

[17] Giraul, V., & Raviart, PA. (1986). Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms. Springer Science & Business Media.

[18] Pazzi, A. (1983). Semigroups of Linear Operators, and Applications to Partial Differential Equations. Applied Mathematical Sciences. Springer-Verlag, New York.

[19] Henry, D. (1981). Geometric Theory of Semilinear Parabolic Equations. Lecture Notes in Mathematics. Springer-Verlag, New York.

[20] Scott, RL., & Zhang, S. (1990). Finite Element Interpolation of Nonsmooth Functions Satisfying Boundary Conditions. Mathematics of Computation, 54(190), 483-493.

[21] Farhloul, M., & Fortin, M. (2002). Review and Complement on Mixed-hybrid Finite Element Methods for Fluid Flows. Journal of Computational and Applied Mathematics. 140 (1), 301-313.

[22] Ciarlet, PG. (1978). The Finite Element Method for Elliptic Problems. Studies in Mathematics and its Applications. North-Holland, Amsterdam.

[23] Eymard, Y., Gallouet, T., & Herbin, R. (2000). Finite Volume Methods. In Handbook of Numerical Analysis, North Holland, Amsterdam.

[24] Hairer, E., Norsett, S.P., & Wanner, G. (1993). Solving Ordinary Differential Equations I. Nonstiff Problems. Springer Series in Computational Mathematics, Springer.

[25] Wanner, G., & Hairer, E. (1991). Solving Ordinary Differential Equations II . Springer-Verlag, Berlin.

[26] Lang, J., & Verwer, J. (2001). ROS3P-An Accurate Third-order Rosenbrock Solver Designed for Parabolic Problems. BIT 41(4) , 731-738.

[27] Rantrop, P., & Kaps, P. (1979). Generalized Runge-Kutta Methods of Order Four with Step size Control for Stiff Ordinary Differential Equations. Numerische Mathematik, 33(1), 55-68.

[28] Tambue, A., Berre, I., & Nordbotten, JM. (2013). Efficient of Geothermal Processes in Heterogeneous Porous Media Based on the Exponential Rosenbrock-Euler and Rosenbrock-type Methods. Advances in Water Resources, 53, 253-262.

[29] Cox, SM., & Matthews. PC. (2002). Exponential Time Differencing for Stiff Systems. Journal of Computational Physics, 176(2), 430-455.

[30] Caliari, M., & Ostermann, A. (2009). Implementation of Exponential Rosenbrock-type Integrators. Applied Numerical Mathematics, 59(3-4), 568-581.

[31] Hochbruck, M., & Ostermann, A. (2005) Exponential Runge -Kutta methods for Parabolic Problems. Applied Numerical. Mathematics. 53(2-3), 323-339.

[32] Hochbruck, M., Ostermann, A., & Schweitzer, J. (2009). Exponential Rosenbrock-Type Methods. SIAM Journal on Numerical Analysis. 47(1), 786-803.

[33] Hochbruck, M., & Ostermann, A. (2010). Exponential Integrators. Acta Numerica, 19, 209-286.

[34] Saad, Y. (1992). Analysis of Some Krylov Subspace Approximations to the Matrix Exponential Operator. SIAM Journal on Numerical Analysis 29(1), 208-227.

[35] Sidje, RB. (1998). Expokit: A Software Package for Computing Matrix Exponentials. ACM transactions on Mathematical Software, 24(1), 130-156.

[36] Schippmann, B. (2008). Comparison of Rosenbrock Methods with Modified Patankar Schemes used in Biogeochemical modeling. Diploma Thesis in Mathematics, University of Rostock, Germany.

[37] Atkinson, K., Han, W., & Stewart, DE. (2011). Numerical solution of Ordinary Differential Equations. (Vol. 108). John Wiley & Sons.

[38] Dahlquist, GG. (1963). A Special Stability Problem for Multistep Methods. BIT Numerical Analysis 3(1), 27-43.

[39] Butcher, JC. (2009). Order and Stability of Generalized Padé Approximations, Applied Numerical Mathematics. 59(3),558-567.

[40] Jens, L. (2001). Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems; Theory, Algorithm, and Applications. Lecture Notes in Computational Science and Engineering, Springer.

[41] Tambue, A. (2016). An Exponential Integrator for Finite Volume Discretization of a Reaction-Advection-Diffusion Equation. Computers & Mathematics with Applications, 71(9),1875-1897.