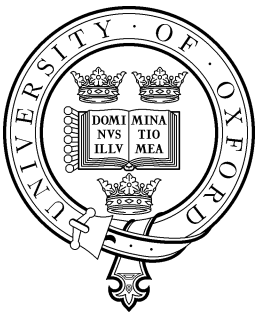
A fictitious domain approach for hybrid simulations of eukaryotic chemotaxis



Jean-Charles Séguis

St Anne's College

University of Oxford

A thesis submitted for

Doctor of Philosophy

Trinity 2013

To my wife, Brigita.

Acknowledgements

First of all, I would like to express my gratitude to all of my three supervisors, David Kay, Radek Erban, and Kevin Burrage, for their continuous support and advice. I think three supervisors were not too much for advising me as I was a stubborn student, and difficult to manoeuvre. I would like to thank them for their patience towards me and I apologise for the frustration I may have caused to them.

During those three and a half years, David Kay had the difficult task to contain my excesses of enthusiasm and to re-boost my confidence after rough patches. I thank him not only for taking on this role but also for his constant optimism. I would also like to thank Radek Erban for his calmness and his ability to bring order to my messy ideas. Last but not least, I thank Kevin Burrage for his wisdom and experience, that have greatly helped me in my DPhil studies.

To finish, I would like to thank my wife and family as well who have constantly encouraged and supported me. The redaction of this dissertation would not have been possible without the regular shipments of *Cantal* and *saucissons* from my family and the help I received from my wife, whose command of the English language was very precious.

Contents

1 Introduction **1**

1.1 A brief introduction to chemotaxis 1

1.1.1 Bacterial chemotaxis 2

1.1.2 Eukaryotic chemotaxis 3

1.2 Mathematical modelling of chemotaxis 7

1.2.1 Continuum models 7

1.2.2 From cell-level dynamics to population-level behaviour 8

1.2.3 Hybrid models 10

1.2.4 Limits of the state-of-the-art computational modelling 14

1.3 Goal of this dissertation 15

1.4 Structure of this dissertation 17

2 Theoretical foundations of the finite element method **19**

2.1 Introduction 19

2.2 Mathematical tools for PDE theory 20

2.2.1 Norms and inner products 20

2.2.2 Cauchy sequences and complete spaces 21

2.2.3 Bilinear forms 23

2.2.4 Lax-Milgram Theorem 24

2.2.5 Lebesgue spaces 25

2.2.6 Weak derivatives 27

2.2.7 Sobolev spaces 29

2.2.8 Trace of Sobolev functions 30

2.3 Variational formulation of elliptic problems 31

2.3.1 Calculus tools 32

2.3.2 Poincaré-Friedrich's Inequality 32
2.3.3 Variational formulation of Poisson's Equation with homogeneous Dirichlet boundary conditions 33
2.3.4 Variational formulation of Poisson's Equation with homogeneous Neumann boundary conditions 34
2.3.5 Mixed variational formulation of the Poisson's equation 35
2.4 Finite element formulation of elliptic problems 37
2.4.1 Definition of a finite element 38
2.4.2 Useful examples of triangular elements 41
2.4.3 Finite element approximation of Poisson's Equation 45
2.4.4 Mixed finite element formulation of Poisson's Equation 47
2.5 Summary 49

3 Fictitious domain methods for elliptic problems 51
3.1 Introduction 51
3.2 Fictitious domain method for Dirichlet elliptic problems 53
3.2.1 Initial problem 53
3.2.2 Classical finite-element approximation of Problem (3.3) 54
3.2.3 A fictitious domain formulation 55
3.2.4 Finite-element approximation of (3.10) 58
3.2.5 Convergence analysis and numerical experiments 59
3.3 A fictitious domain method for Neumann elliptic problems 63
3.3.1 Fictitious domain formulation of (3.25) 64
3.3.2 Finite-element approximation of (3.30) 66
3.3.3 Convergence analysis and numerical experiments 66
3.4 Adaptation of the fictitious domain method in H^{div} to Robin problems . 71
3.4.1 Motivation 71
3.4.2 Definition of the problem 72
3.4.3 A fictitious domain formulation 72
3.4.4 Finite-element approximation 73
3.4.5 Numerical experiments 74
3.5 Summary 78

4 Implementation of the FDM and preconditioning **79**
4.1 Introduction . . . 79
4.2 Implementation of the finite element matrix operators . . . 80
4.2.1 Generation of a uniform triangulation and related maps . . . 80
4.2.2 Building finite element matrices . . . 83
4.2.3 Quadrature rule on the reference triangle . . . 83
4.2.4 Building matrix operators . . . 84
4.3 Efficient implementation of hybrid operators . . . 86
4.3.1 Algorithm . . . 87
4.3.2 Intersections of a regular triangulation and an ellipse . . . 92
4.4 Effective iterative solvers and preconditioners for linear systems arising from the FDM . . . 95
4.4.1 Preconditioned minimal residual iteration algorithm . . . 97
4.4.2 Generalised minimal residual iteration algorithm . . . 99
4.4.3 Preconditioning of saddle-point linear systems arising from the FDM 101
4.5 Summary . . . 112

5 Comparison of two cell models for eukaryotic chemotaxis **115**
5.1 Introduction . . . 115
5.1.1 Two models for finite-sized cells . . . 115
5.1.2 Objective . . . 117
5.2 Case of one immobile cell . . . 118
5.2.1 Non dimensionalisation . . . 118
5.2.2 Matching in 3 dimensions . . . 118
5.2.3 Matching in 2 dimensions . . . 121
5.2.4 Summary . . . 124
5.3 One steadily moving cell . . . 124
5.3.1 Slow cell approximation . . . 124
5.3.2 Formulation of the three dimensional problem . . . 125
5.4 Matching when β is θ-dependent . . . 129
5.5 Discussion . . . 130
5.6 Summary . . . 131

6 A new simulation framework for simulating chemotaxis 133
6.1 Introduction . . . 133
6.2 Mathematical formulation . . . 134
6.2.1 Notation and domain definition . . . 134
6.2.2 Governing strong equations . . . 136
6.2.3 Weak formulations and the FDM . . . 140
6.3 Computational scheme . . . 143
6.3.1 Spatial discretisation . . . 143
6.3.2 Semi-discrete equations . . . 145
6.3.3 Temporal discretisation scheme for (RDE) . . . 146
6.4 Implementation and tests . . . 150
6.4.1 Numerical simulation of the motion of single cell subject to chemotaxis . . . 150
6.4.2 Numerical simulation of the motion of many cells subject to chemotaxis . . . 154
6.5 Qualitative validation of the model: comparison with another computational chemotaxis model . . . 155
6.6 Summary . . . 159

7 Conclusion 161
7.1 Mathematical and biological context . . . 161
7.2 Our contributions . . . 162
7.3 Criticism and issues . . . 164
7.4 Possible future research directions . . . 165

A Solution to the steadily moving hoover cell problem 167

B Avoiding collision of cells 171

References 173

List of Figures

1.1 Starved *D. discoideum* aggregation pattern 4
1.2 Macroscopic patterns of *D. discoideum* chemotaxis 11

2.1 Example of a Lipschitz domain 30
2.2 Two affine-equivalent elements 39
2.3 Common triangular elements 43

3.1 Original domain configuration and a possible triangulation 53
3.2 Fictitious domain and a possible regular triangulation 56
3.3 Convergence results for the Dirichlet FDM 62
3.4 Convergence results for the Neumann FDM 70
3.5 Error plots for the Neumann FDM 70
3.6 Intersection of an interface with a triangulation 75
3.7 Convergence results of the Robin FDM in the H^{div}-norm 76
3.8 Convergence results of the Robin FDM in the L^2-norm 77
3.9 Influence of the choice of ε 77

4.1 Representation of two partitions of an interface 89
4.2 Sparsity pattern of the linear system derived from the Dirichlet FDM . . 96
4.3 Decay of the norm of the residual when using Glowinski's preconditioner 105
4.4 Decay of the norm of the residual when using a preconditioner for Neumann FDM 111
4.5 Decay of the norm of the residual for the Neumann FDM preconditioner in the Robin case. 113

5.1 Relationship of the parameters α and β 124
5.2 Solution of the matched hoover-cell model in $\mathbb{R}^2$ and $\mathbb{R}^3$ 125

5.3 First order approximation to the hoover-cell model in the mobile cell case 128
5.4 Plot of the corrective term for the mobile cell case 129

6.1 Representation of the cavity Ω and cells $C_i(t)$. 135
6.2 Setup of the kinematic parameters of the motion of a cell 136
6.3 Mesh for the simulation framework 143
6.4 Simulation results for one cell 154
6.5 Simulation results for many cells 155
6.6 Migration of NC cells 156
6.7 Comparison of results for simulation of NC cell simulation 158

B.1 Repulsive force for modelled cells 172
B.2 Computation of the repulsive force 172
B.3 Computation of the repulsive moment 173

List of Tables

3.1 Computation times for Dirichlet FDM 63

4.1 Performance of the preconditioner for Dirichlet FDM 105
4.2 Performance of the preconditioner for Neumann FDM 110
4.3 Performance of the Neumann FDM preconditioner in the Robin case . . . 112

6.1 Parameter values for the one-cell simulation 153
6.2 Stability of the one-cell simulation scheme 153
6.3 Parameter values for the NC cell chemotaxis simulation 157

Chapter 1

Introduction

Research in biology has been increasingly relying on theoretical and computational mathematical models for better understanding complex biological processes. Chemotaxis, one of the most important and widespread phenomena in biology, has been intensively studied in the last decades and has been the object of a considerable modelling effort from the mathematical community.

In this dissertation, we contribute to this effort by presenting a numerical framework for simulating the chemotaxis of eukaryotic cells. Its most important feature is to enable its user to bring together models describing biological phenomena taking place at different scales in cell biology. In the process of building up this simulation framework, we make additional contributions in several fields of mathematics, in finite elements, preconditioning and theoretical biology.

In this chapter, we will first provide a description of the phenomenon of chemotaxis and review its most noticeable manifestations in nature, and then discuss its current state-of-the-art mathematical modelling in the literature. This will lead us to identify their advantages and drawbacks and to introduce the challenges that will be tackled in the next chapters.

1.1 A brief introduction to chemotaxis

The phenomenon of chemotaxis describes the directed movement of somatic cells or bacteria in response to gradients of exterior chemical signals. Positive chemotaxis occurs

when cells are attracted to areas with higher concentrations of a certain chemical, which in this case, is called a chemoattractant and conversely, negative chemotaxis occurs when cells are attracted by lower concentrations of chemical, which in that case, is called a chemorepellent.

We will discuss the different manifestations of this phenomenon in nature: first, bacterial chemotaxis [11] and then eukaryotic chemotaxis for somatic cells. Chemotaxis is a key component of a great number of biological processes in complex organisms, such as cancer tumour growth and metastasis [90, 60], wound healing [109, 122] or fertilisation of eggs by sperm [43].

1.1.1 Bacterial chemotaxis

Historically, bacterial chemotaxis was the first type of chemotaxis receiving attention from the scientific community. Engelmann and Pfeffer [49, 110] at the end of the 19th century discovered that bacteria respond to various chemicals and made the first scientific description of the phenomenon.

Later on, in the second part of the 20th century, considerable attention was given to the study of chemotaxis of *Escherichia coli* bacterium. Usually abbreviated *E. coli*, it is a type of bacteria that can commonly be found in the lower intestine of most warm-blooded organisms. It measures about $2\,\mu$m and has the tubular shape. As it can be grown easily, *E. coli* bacteria has been intensively studied in the last decades.

Adler in 1966 [2] showed experimental evidence that *E. coli* bacteria are able to carry out chemotaxis and move to environments providing them with the greatest supplies of energy, whether it is in the form of oxygen or glucose. In 1972, Berg and Brown [12] demonstrated that *E. coli* bacteria move in a 'run and tumble' fashion. *E. coli* bacterium alternates phases of straight line motion and phases of tumble during which the bacterium randomly chooses a new direction of motion. They also showed that the distribution of the length of time intervals between two tumbles is exponentially distributed. In the same year, Macnab and Koshland [88] showed that the frequency of tumbles increases when *E. coli* bacteria move to less favourable areas (climbing down a gradient of chemoattractant) and tend to extend runs in favourable ones. They also

demonstrated that *E. coli* sense gradients of chemoattractant not by measuring concentrations differences along their body length, but by temporal sensing: *E. coli* bacteria compare current concentrations to those measured over the past few seconds along their way [11].

After these discoveries, *E. coli* has continued to be the object of intense research, focused on understanding the signalling pathways involved in the elaboration of the chemotactic response to gradients of chemoattractants by *E. coli*. Transmembrane chemoreceptors of the bacteria have the ability to recognise specific chemicals by chemical binding. This binding then causes a chemical signal to be transmitted in the cytoplasm and be converted in an appropriate response [13, 57]. It has emerged that many intracellular signalling pathways in *E. coli* are monitored by the so-called 'two-component pathways', involving a receptor, histidine kinase and aspartate kinase elements [18].

Due to the amount of research carried out on *E. coli*, chemotaxis of *E. coli* bacterium and its signalling pathways are now well understood and serve as a reference model for analysing more complex chemotactic systems in larger organisms, and especially eukaryotic cells.

1.1.2 Eukaryotic chemotaxis

Unlike prokaryotic cells, eukaryotic cells are cells equipped with a nucleus. They are usually much larger than their prokaryotic counterparts, and comprise most of complex organisms. Chemotaxis has been reported in many types of eukaryotic cells: free-living microorganisms [17], white blood cells [122], endothelial cells in angiogenesis [84], sperm in fertilisation [43], neuronal growth cones in neurogenesis [123] or fibroblasts in wound healing [116]. Chemotaxis in eukaryotic cells has been most extensively studied in cells of the immune system and cellular slime mold. In the following paragraphs, we will review some of the examples of eukaryotic chemotaxis.

1.1.2.1 Amoeboid chemotaxis

In 1947, Bonner [17] showed evidence that chemotaxis is responsible for the aggregation of *Dictyostelium discoideum*, commonly referred to as slime mold, a soil-living unicellular

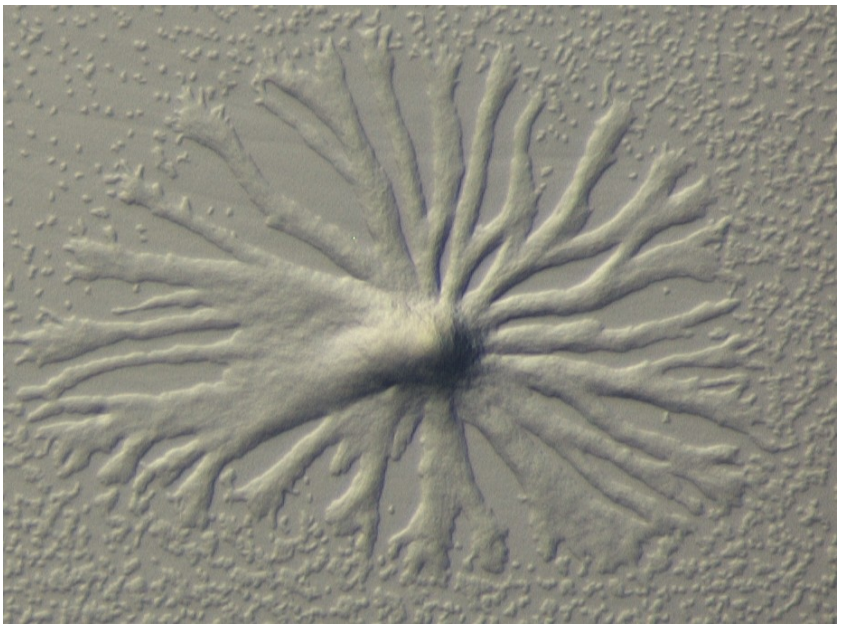


Figure 1.1: Starved *Dictyostelium discoideum* aggregating by chemotaxis. Image from [98], with permission.

eukaryotic amoeba that feeds on bacteria such as *E. coli*. The slime mold amoeba can form large aggregates of up to 100,000 cells, as shown in Figure 1.1 and displays a striking cooperative behaviour, by moving and feeding in coordination. This coordination relies heavily on cell-to-cell signalling.

Later, it was discovered [82, 92, 91] that these organisms release 3'-5'-cyclic adenosine monophosphate, commonly abbreviated cAMP, in periodic waves that propagate between cells over large distances. Raman *et al.* [114] showed that, after about a few hours of starvation, randomly located cells, called pacemakers, are able to release cAMP periodically, then other cells move toward these pacemakers, and eventually the entire population forms a large aggregate, that later transforms in a fruiting body. *D. discoideum* is now used as a model organism for studying chemotaxis of other types of eukaryotic cells, such as somatic cells, due to the fact that they are easy to cultivate in laboratories.

The mechanisms through which *D. discoideum* are able to sense the local gradient of the chemical signal has been postulated to be a combination of temporal and spatial sensing. As Parent and Devreotes [106] explain, larger cells such as amoebae can sense concentration differences across their body length, because of the presence of receptors

equally distributed on the surface of the membrane. Small bacteria such as *E. coli* cannot use this strategy, because at their scale, gradients are shallow and undetectable.

However, the precise chemical and mechanical processes making the measurement of spatial gradients of chemoattractant possible, and leading to the chemotactic response of the *D. discoideum* to cAMP are still largely unclear to the scientific community and several hypotheses have been proposed. Some scientists (Mato *et al.* [92], Zigmond [132]) postulated that cells are able to sense local gradients by measuring spatial differences of occupied receptors on the cell membrane. It is also suggested that cells are able to convert the spatial gradient into a temporal rate of change of chemoattractant, by extending pseudopodia – long filaments that allow the cell to crawl on surfaces – and using them as sensors [62]. It is also believed that spatial external gradients of chemical signal can trigger the formation of an internal gradient, whose direction determines the future orientation of the cell. After adaptation has occurred in the signalling pathways, the internal gradient fades away. Such a process has been experimentally observed in eosinophils – a type of white blood cells – by Brundage *et al.* [25]. This last hypothesis has received more attention recently in Postma *et al.* [112].

1.1.2.2 Chemotaxis and cancer tumour growth

In the field of cancer research, Folkman [58] in 1971 demonstrated that cancer cells in tumours, which are not initially supplied with blood, can secrete growth factor chemical signals such as vascular endothelial growth factor (VEGF). This signalling protein is known for inducing nearby vessels to grow and extend their capillaries, which allows cancer cells in tumour to connect to the capillary network and receive oxygen and essential nutrients.

Angiogenesis, the process through which blood vessels are extended, has been shown to be stimulated by VEGF [74]. This protein triggers a signalling cascade in endothelial cells – cells forming the blood vessels – resulting in their migration due to chemotaxis by VEGF and their local proliferation, leading to the extension of the capillary network.

1.1.2.3 Chemotaxis and the immune system

As Eisenbach explains in [44, p. 308], one of the essential functions attributed to cells in the immune system is their ability to migrate. White blood cells, such as neutrophils, monocytes and lymphocytes migrate from the vascular system through the interstitial tissues to infected areas, by crawling on a substratum or a matrix. This process is very similar to amoeboid chemotaxis.

The first of the cells of the immune system to arrive at sites of injury are neutrophils. Their primary functions are fighting bacterial infections as well as clearing cellular debris. Neutrophils accumulate at the wound site within an hour, and initiate the inflammatory cascade. Monocytes and lymphocytes start agglutinating at the site of injury after 4 or 5 hours. Monocytes contribute to the phagocytic activity, and initiate the regeneration of tissue, by releasing VEGF [24]. These three types of white blood cells are able to release additional chemical signals promoting the anti-inflammatory response.

1.1.2.4 Chemotaxis and fertilisation

Moricard and Bossu [95] in 1951 first interpreted selective aggregation of rat spermatozoa near dissociated follicular cells as the result of a chemotactic response from oocyte. After many similar observations in other mammals, chemotaxis of sperm has been observed consistently in humans [31] and mice. Giojalas and Rovasio [63] proved that mouse spermatozoa can modify their motility parameters in response to chemical changes in the oocyte environment.

However, chemoattractants for spermatozoa have only been clearly identified in the case of non-mammalian species. According to Eisenbach [44, p. 441], one of the most widely studied case of non-mammalian chemotaxis is that of the sea urchins *Arbacia punchulata*. It is shown that the injection of a peptide isolated from the jelly layer of *Arbacia punchulata* eggs leads to the accumulation of sperm near the site of injection [129].

In the case of mammals, despite the fact that sperm chemotaxis has been postulated in many species, the chemical substances driving the migration of sperm towards the egg have not yet been clearly identified. One of the main reasons originates from the

difficulty, without saying the impossibility, of conducting *in vivo* experiments, and all studies so far have been carried out *in vitro*. Most experiments conducted to date have consisted in demonstrating *in vitro* chemotaxis of sperm to follicular fluid, known for containing secretions from the egg.

1.2 Mathematical modelling of chemotaxis

Due to the striking population-level manifestation of cell-level response to chemoattractants, the phenomenon of chemotaxis has quickly sparked interest in the mathematical community. Many models aim at reproducing the patterns and aggregation of cells that are observed experimentally, as well as their collective migration. The first models were continuum models, then discrete models emerged when computing power allowed large scale simulations of discrete agents. We will review the most important of them in this section.

1.2.1 Continuum models

One of the earliest contribution to the mathematical modelling of chemotaxis is due to Keller and Segel [80]. They tackle the problem of bacterial chemotaxis, in particular, that of *E. coli*. Their model is a continuum model where the cells and the molecules of chemoattractant are described by their continuous respective density n and c. The model consists of a system of two coupled partial differential equations (PDEs) satisfied by c and n:

$$\partial_t n = D_n \triangle n - \nabla \cdot [\chi(c)\, n\, \nabla c]\,, \tag{1.1}$$

$$\partial_t c = D_c \triangle c - f(c, n). \tag{1.2}$$

where D_n, D_c are diffusion coefficients, $\chi(c)$ is the chemotactic sensitivity and $f(c, n)$ describes the production, consumption or degradation of the chemotactic signal. This continuum model implicitly assumes that both cells and chemoattractant are present in sufficiently large numbers.

We observe that the change of the density of cells at any point is driven by two terms: a diffusion term $D_n \triangle n$ that reflects the random nature of the motion of the cells,

and a transport term $[\chi(c)\, n\, \nabla c]$ in which the advection field is shaped by the gradient of chemoattractant ∇c. This equation is a particular case of a Fokker-Plank partial differential equation describing the biased random motion of a set of particles, which were derived and studied by Patlak [108] in 1953, by using kinetic theory arguments for expressing the chemotactic velocity in terms of a statistical average of velocities of individual cells.

The second equation, that describes the change of the density of chemoattractant, is a classical reaction-diffusion equation, with $D_c \triangle c$ being the diffusion term and $f(c, n)$ the reaction term encapsulating the description of the interaction of the cells with the chemoattractant.

Keller and Segel [81] showed that the PDE problem above admits travelling wave solutions if the chemotactic sensitivity is singular in c. Physically, the existence of travelling waves solutions validates the PDE model since it implies that the model is able to reproduce the collective motion of bacteria towards the higher concentrations of chemoattractant observed in nature.

One of the downfalls of this model is that the chemotactic sensitivity term is only hypothesised and has been derived by a phenomenological approach. However, from a biological point of view, this term is of great importance, because it carries the information on how the signal transduction mechanisms translate in the macroscopic collective behaviour of the cells, modelled by the Keller–Segel system. Hence, subsequent models, instead of testing different hypotheses on the form of the chemotactic term, have incorporated more details of the signal transduction mechanisms in the cells, by taking advantage of the considerable progress in the understanding of the internal cell dynamics.

1.2.2 From cell-level dynamics to population-level behaviour

The collective properties of cell populations are determined by the individual-level behaviour. If one is able to relate the coefficients of population-level PDEs (1.1)–(1.2) to the parameters of individual-level models, then the population-level PDEs provide an efficient way for determining how the population-level behaviour depends on the individual-level parameters [50]. The connections between individual-level and population-level

models have been established for simplified models of bacterial behaviour where cells are modelled as points [52, 53, 130].

In Ref. [52], Erban and Othmer study the chemotaxis of *E. coli* bacteria by aspartate. Due to the particular nature of the motion of such bacteria, the "run and tumble" strategy, the authors model the change of velocity of the bacteria by a velocity jump process. They write the governing evolution PDE, derived by Othmer *et al.* in [104], of the density of cells $p(x, v, t)$, depending on time, position and velocity:

$$\frac{\partial}{\partial t}p(x, v, t) + v \cdot \nabla p(x, v, t) = -\lambda p(x, v, t) + \lambda \int_V T(v, v')p(x, v', t)\,\mathrm{d}v', \tag{1.3}$$

where λ is the turning rate – the frequency of velocity jumps – and $T(v, v')$ the transition probability for a cell to switch from velocity v' to v. The internal dynamics of the signalling pathways are simplified to a 'cartoon' model, that captures the main characteristic excitation-adaption feature of the chemotactic response. In order to link the macroscopic equation for the density and the intracellular dynamics model, Erban and Othmer assume that the turning rate depends linearly on the adapted signal. The authors derive an expression for the chemotactic sensitivity term, function of the gradient of the external chemical signal and the parameters of the cartoon model for the intracellular dynamics. The comparison of outcomes of Monte-Carlo simulations with the numerical solution of the macroscopic PDE model illustrates the derived chemotactic sensitivity term.

For finite-size cells, the population-level equations have been derived for disk-like cells with a simplified description of signal transduction [54] and for the cellular Potts model [3]. Adding more biological details to individual-level models complicates the derivation of coefficients of macroscopic equations. In some cases, one can estimate them from short bursts of appropriately initialised simulations of the individual-level model [51]. However, these computational approaches are only useful when we know the functional form of the macroscopic equations. For more complicated models of chemotaxis, it is not clear whether the population-level description can be reduced to the classical chemotaxis equation (1.1) with suitably defined coefficients [54]. Hence it appears that the development of efficient simulation techniques for collections of cells is necessary for understanding how individual-based properties influence collective behaviour of unicellular organisms.

1.2.3 Hybrid models

Since the influence of the signal transduction mechanisms at the cell level on the population-level behaviour of cell is still unclear to scientists, the computational model approaches have shifted from purely continuum models to 'partially' discrete models. Indeed, in order to incorporate more details in models for cell movement and signal transduction pathways, Equation (1.1) can be replaced by a discrete description of the cells, modelled as individual agents. This type of method is named 'hybrid' because it describes in the same framework different components of the system using different modelling approaches. Hybrid modelling has become popular and widespread in mathematical modelling applied to biology [59].

1.2.3.1 Point-like cells modelling

One of the first hybrid models for studying chemotaxis was proposed by Parnas and Segel [107]. In order to simulate the chemotactic migration of *D. discodeum* organisms due to cAMP, a computational model is set up. The reaction-diffusion Equation (1.2) of Keller–Segel model is discretised using a finite-difference scheme, and the cells are modelled as point-like discrete agents, constrained to the same mesh as the concentration field, and 'reading' the local value of the continuous concentration of the chemical signal. One advantage of such an approach lies in the fact that different assumptions on the sensing mechanism can be tested when specifying the rules for the movement of the cells. However, since the positions of the cells are constrained to a grid, the simulation results greatly depend on the value of the spatial step size. Using values of the biological parameters from the literature, Parnas and Segel reproduce aggregation patterns that are observed experimentally.

In Ref. [36], Dallon and Othmer present and analyse a hybrid model for the aggregation of *D. discoideum* that encapsulates the signal transduction mechanisms, as well as the production and release of cAMP by pacemaker cells. They introduce a set of differential equations governing the chemical dynamics of intracellular cAMP, the reaction-diffusion PDE describing the evolution of extracellular concentration of cAMP (1.2) from the Keller–Segel model and a set of deterministic movement rules for the cells. The *D. discoideum* are represented as points in the two-dimensional domain of evolution of cAMP,

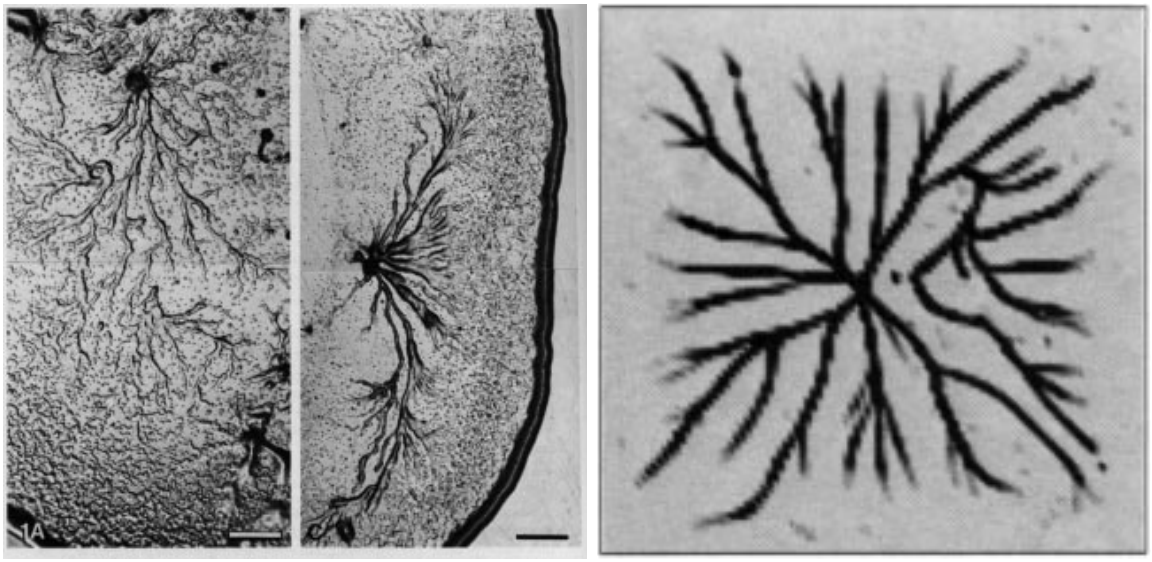


Figure 1.2: The picture on the left shows the aggregation patterns obtained experimentally by Claviez *et al.* [30] of aggregated *D. discoideum*. The picture on the right represents an outcome of Dallon and Othmer's simulation framework for the same phenomenon [36].

and unlike Parnas and Segel's simulation scheme, the positions of cells are not constrained to a grid but are mesh free. The concentration and its mathematical gradient are then interpolated at each cell location from the grid points defined from the spatial discretisation of the solution of the reaction-diffusion equation. In the reaction-diffusion equation (1.2), the reaction term $f(c, n)$ – modelling the local influence of the density of cells on the concentration of chemical signal – is replaced by a juxtaposition of sink terms (Dirac functions), where the position of each sink term is determined by the location of a particular cell in the computational domain. Dallon and Othmer computationally reproduce the aggregation pattern that are observed experimentally by Claviez (see Figure 1.2) and study the influence of some parameters of their model on the shape of the aggregation patterns they produce.

A hybrid model that describes cells as points and external chemicals by PDEs has also been constructed for collections of bacteria in [50]. In this model, two external signal fields are considered (aspartate and succinate) and coupled with the individual-based model of bacteria *E. coli*. Bacteria are chemotactic towards aspartate. They consume succinate, secrete aspartate and proliferate with a doubling time of 2 hours. Succinate-starved cells can also consume aspartate. In this model, the description of

the intracellular dynamics is considerably simplified compared to Dallon and Othmer's approach due to the use of the cartoon model, that we have introduced in Section 1.2.2.

1.2.3.2 Finite-sized cells modelling

In the models we have reviewed so far, the cells, whether they represent *E. coli* bacteria or amoebae such as *D. discoideum*, were always represented as points. The main reason for this is that modelling finite-sized cells leads to solving a free-boundary problem, since the position of the cell is influenced by the chemoattractant concentration field. Numerically, this has proved to be a very challenging problem. Fortunately, several methods have been proposed to tackle such problems. In this section, we review two articles that propose a numerical scheme for simulating a chemotactic system with finite-sized cells.

Bresch *et al.* in [20] present a hybrid simulation of a chemotactic system consisting of *E. coli* bacteria and oxygen as chemoattractant in which the bacteria have a finite size and therefore not considered as points. The main originality of this approach is that the bacteria interact with the fluid environment and the authors mainly focus on the mechanical aspect of the *E. coli* migration. In this simulation framework, Bresch *et al.* account for the hydrodynamic forces acting on the disk-shaped cells and conversely, by numerically solving an unsteady Stokes equation. A penalty method such as developed in [120] is implemented to enforce the rigid-body motion of the cells in the fluid. The propelling action of the flagellum is modelled via a point force, directed radially outward of the cell. Moreover, a non-overlapping condition is imposed on the modelled bacteria, and for that purpose, at each time step, the solution of the Stokes equation is projected onto a set of admissible solutions (satisfying the non-overlapping conditions). Finally, in order to circumvent the difficulty of simulating a reaction-diffusion equation in a moving domain, the consumption of oxygen by the bacteria is modelled by a localised depletion term $-k(\boldsymbol{x})\, c(\boldsymbol{x}, t)$, where the depletion rate k has support only in the regions of the computational domain occupied by the cells. This approach will be studied in more details in Chapter 5 and will be called the 'hoover cell model'.

In [94], McLennan *et al.* study the chemotaxis of Neural crest cells by VEGF during the embryo development. For that purpose, the authors set up a hybrid simulation in which cells are motile finite-sized disk-shaped entities, that evolve in a fast growing domain. The interaction between the chemoattractant concentration field and the cells is modelled via a localised depletion rate, as in Bresch's model, however, the hydrodynamic forces between the fluid carrying VEGF and the cells are neglected. The law of domain growth is fitted from experimental data. Because of the lack of agreement between biological observations and simulation results, the authors then postulate the specialisation of neural crest cells in different categories, followers and leaders, which have distinct roles in the migration process of neural crest cells. They provide biological evidence supporting this conjecture, by showing different gene expressions in two groups of cells they isolated. Finally, the simulation results in this new setting agree with the initial biological observations.

In the two methods discussed above, the cells are modelled as solids thus not deformable. This is a major drawback since it is known that cells are highly plastic objects. Several methods allow to tackle this modelling challenge.

In Ref [75], Graner and Glazier introduce the cellular Potts model, which consists in representing a cell by a set of connected points on a lattice. Each point can take a few number of states, whether there are part of the cell, or outside, or on the boundary. The motion and the deformation of a cell is rendered by the switching of points in the lattice to a new state according to a set of probabilistic rules. Large deformations of the membrane of cells are penalised through the minimisation of an energy function. In computational biology, this method has been primarily used for modelling and simulating the change of the shape of amoebae during their motion [100].

The level-set method, introduced by Osher and Sethian in 1988 [103, 102] is a numerical method designed for simulating the motion of an interface (which can be multi-connected) subject to a velocity field and possibly external forces. It has been successfully applied to the simulation of the deformation of *D. discoideum* by Yang *et al.* [131]. Using a mechanical model calibrated to experimental data, they simulate the evolution of the morphology of the cell when responding to an external chemical gradient. The simulated change in shape of the cell matches that observed experimentally.

Those two numerical method have in common their relatively high computational cost, which restrict the amount of cells that may be fitted in a single simulation, and their inability to simulate objects that are chemically interacting with their environment.

1.2.4 Limits of the state-of-the-art computational modelling

To sum up, when modelling cells as individual agents, we encounter in the literature two types of models: the point-like cell model and the finite-sized cell model. From what we have seen so far, we would like to formulate two remarks regarding the appropriateness of each model, depending on the biological situation they are mimicking.

Firstly, modelling the chemotaxis of small bacteria such as *E. coli* using point-like cells seems to be adequate, because, as we have seen in Section 1.1.1, research has shown that the underlying sensing mechanism allowing the bacteria to find their way is essentially temporal. Moreover, since *E. coli* would be too small to be able to measure gradients of chemicals concentrations across their body-length, the finite-sized cell model in this case would be of little use in capturing more features in the modelling of the signal transduction process.

Due to the spatial nature of the sensing mechanism in amoeba, the point-like cell model cannot be as accurate as it is for *E. coli* bacteria. Using this model leads to the necessity of making a strong assumption on how a cell senses spatial gradients of chemoattractant. For instance, one has to assume that a cell can measure ∇c locally. Moreover, this type of model cannot link the extracellular and the intracellular signalling pathways in an accurate way, because the sensing mechanism can only be modelled phenomenologically.

Lastly, when modelling the chemotaxis of relatively large cells, such as amoebae, the finite-sized cell model seems to be more accurate than the point-like cell model. Indeed, since we know from Section 1.1.2 that in this case, the sensing mechanism of cells is spatial, then more detailed models of the signalling transduction pathways from the extracellular to intracellular levels can be potentially fitted. However, in the simulations presented in [20] and [94], despite the use of disk-shaped cells, the focus is not about chemical modelling, but rather about the effect of confinement of cells in a small area, by enforcing a non-overlapping condition.

Moreover, the volume depletion term accounting for the interaction between the discrete cells and the concentration field is questionable: in 'real' cells, the absorption of molecules of chemoattractant is performed by receptors on the cell membrane. One argument in favour of this modelling is that in two-dimensional simulations, volume absorbing cells accurately represent the projection on a plane of three-dimensional cells able to absorb chemicals on the whole surface of their membrane. However, experiments carried out in Petri dishes can effectively be assimilated to two-dimensional experiments, since the layer of cells cultivated in each dish is generally very thin. Hence, the volume effects of the cells cannot be neglected, that is to say, cells by their volume can delay or block the propagation of a chemical signal. This effect cannot be captured when using volume absorbing cells, which are effectively 'gliding' over the chemical without upsetting its diffusion in the extracellular environment.

To the best of our knowledge, a simulation of a collection of mobile finite-sized cells subject to chemotaxis that links extracellular and intracellular dynamics does not exist. Either the cells are set to be mobile and in this case the response to the environment is modelled in a purely phenomenological way, or the cells are immobile and the attention is focused on intracellular dynamics, while the external signals are imposed. For instance, in [86], Levine *et al.* set up a model for describing the sensing mechanism of cAMP by *D. discoideum*. One immobile disk-shaped cell responds to externally imposed signals of cAMP, through the variations and dynamics of the concentrations of a membrane-bound activator and a cytosolic diffusing inhibitor. They reproduce switch-like responses observed experimentally, *i.e.* the rapid localisation of the activator at the front of the cell.

1.3 Goal of this dissertation

In this dissertation, we present an approach to simulating the chemotaxis of cells that is capable to describe the system at both intracellular and extracellular levels. Its main feature consists in modelling cells as finite-sized non-deformable solid objects while adopting a continuous approach to model the evolution of the concentration of extracellular chemicals. The simulation framework we build also allows the implementation of the modelling of chemical dynamics at three levels in a modular way: intracellular level, at

the interface between the intracellular and the extracellular domains and in the extracellular space with possible feedbacks and effects on the motion of cells. This approach can, as well, accommodate the coupling of spatial extracellular dynamics and spatial intracellular dynamics even if we are not implementing this idea in the dissertation.

The main challenge consists in making this simulation computationally affordable whilst not compromising on the level of details: this includes differentiating the extracellular and intracellular spaces and modelling the cell membrane governing the exchanges between the bulk and the cytosol. Hence, instead of a volume depletion (or production) term $f(c, n)$, the absorption (or emission) of chemoattractant by the cells will appear as a boundary condition at the cells boundary.

Those requirements lead us to solve a reaction-diffusion equation similar to (1.2) on a deforming domain (the extracellular domain) due to the motion of the cells. In order to make this cost acceptable, our computational model takes advantage of the fictitious domain method for finite elements that allows us to avoid using costly re-meshing and interpolations between grids between two time steps. A mixed method is employed that allocates specific degrees of freedom to the flux of chemicals, allowing to better monitor the exchanges between the cytosol and the bulk.

The fictitious domain method is widely used in computational fluid dynamics [32, 70, 68] for the simulation of particulate flows for instance. In spite of the complex geometry of the problem, often posed in a moving domain, the fictitious domain method allows the use of a single computational grid for the whole simulation by extending the original domain to a simpler fixed domain. The boundary conditions are then imposed through Lagrange multipliers defined on the original boundary or are volume-supported [69]. In terms of numerics, the resulting linear problem is a saddle-point system, which requires adequate preconditioning for optimal performance of iterative solvers of linear systems [10].

All the aspects of the simulation mentioned above will be addressed in the subsequent chapters of this dissertation: the fictitious domain method, the imposition of the immersed boundary conditions and the preconditioning of the resulting linear systems.

1.4 Structure of this dissertation

In Chapter 2, we introduce the basic concepts necessary to understand the finite element method for elliptic partial differential equations, which is at the root of the fictitious domain method. We will address the derivation of variational formulation of elliptic problems, then present the core mathematical concepts for studying the convergence of the finite element method. Finally, we will discuss the derivation of two different finite element formulations of elliptic problems, using two different types of elements.

In Chapter 3, we present the fictitious domain method for finite elements, as originally developed by Glowinski, for elliptic Dirichlet problems. We also introduce Joly's fictitious domain method [79] for elliptic Neumann problems. Finally, we will discuss our own extension of the fictitious domain method for elliptic Robin boundary PDEs. It is based on the scheme for Neumann fictitious domain method. We provide numerical evidence for its convergence and we show similar convergence properties as the pure Neumann fictitious domain method.

In Chapter 4, we discuss practical computer implementation issues that one encounters when working with the fictitious domain method. Despite the abundance of the literature in this field, we will provide simple `Matlab` tools for building regular triangular meshes of two-dimensional rectangular domains with adequate maps between nodes, elements and edges, and a routine for constructing common finite element matrix operators. Then we will present our algorithm for efficiently constructing fictitious domain specific matrix operators. Finally, after we have introduced Glowinski's preconditioner for the linear system deriving from the Dirichlet fictitious domain method, we use Glowinski's method to derive an optimal preconditioner for the linear system obtained from the Neumann and Robin fictitious domain methods.

In Chapter 5, we introduce the boundary cell model, at the basis of the modelling we propose in this thesis for eukaryotic cells chemotaxis simulations and we compare it with the hoover cell model, the cell model used by Bresch and McLennan, in simple configurations, by attempting to match their main coefficients; we will see that this is only possible in a basic configuration.

Finally, in Chapter 6, we set up a new framework for simulating chemotactic systems, that allows its user to incorporate detailed linked models of the extracellular and intracellular signalling pathways. We first present the equations of the simulation, then we employ the fictitious domain method for discretising them. At last, we test the simulation in a certain number of realistic situations and compare it with the boundary cell model set up.

For the readers who are familiar with numerical methods such as the finite element method, Chapter 2 may not be of interest to them, so we suggest skipping it and begin their reading at Chapter 3. Readers who are only interested in understanding the simulation framework we design in this dissertation can skip Chapter 5, dedicated to the theoretical comparison of two types of cell modelling for chemotaxis.

Chapter 2

Theoretical foundations of the finite element method

2.1 Introduction

The finite element method (FEM) is a numerical method for approximating the solution of partial differential equations (PDEs). It has emerged during the second world war for approximating the solution of an elasticity problem [33] and has become one of the most popular numerical method in engineering for solving PDEs in complex geometries. It has numerous application in industry, including meteorology for weather forecasting, structural mechanics in the car industry and in aeronautics.

The FEM relies on the theory of PDEs and approximation theory in Sobolev spaces. It consists in approximating the unknown function by polynomials of a given degree on each piece of a decomposition of the domain of computation. Each piece of this computation is called an element and can take various shapes: for instance, when the problem is one-dimensional, elements are segments, when two-dimensional, elements are usually chosen as triangles or quadrilaterals.

In the case of linear PDEs, this discretisation process reduces the original problem to a set of linear equations. The number of degrees of freedom determine the size of this linear system. The greater the number of degrees of freedom, the closer to the true solution of the original problem the approximation will be. One of the mathematical challenges in finite elements is defining a suitable meaning for the word 'closer'. This is one of the most important aspects of the FEM that we will discuss in this chapter.

In Section 2.2, we introduce the basic concepts necessary to understand the theory of finite elements: this will include the definition and basis properties of Lebesgue and Sobolev spaces. In Section 2.3, we discuss the definition and the derivation of a *variational formulation* – also called *weak formulation* – for elliptic PDEs, in particular for the simplest of them, the Poisson equation. This will lead us to formulate existence and uniqueness results for the Poisson problem in several configurations, for Dirichlet and Neumann boundary conditions. Lastly, in Section 2.4, we derive the finite element formulation for the Poisson equation from two different variational formulations and using different discretisation schemes. In both configurations, we will provide some theoretical convergence results.

The proofs of most propositions and theorems will mostly be omitted, however we will indicate to the reader references where those proofs can be found.

2.2 Mathematical tools for PDE theory

2.2.1 Norms and inner products

Let E be a linear space, with scalar field $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$.

Definition 2.1 (Norm)**.** The real valued map $\|\cdot\| : E \to \mathbb{R}^+$ is called a norm on E if and only if it satisfies the following:

(N1) for $x \in E$, $\|x\| = 0 \implies x = 0$;

(N2) $\forall \lambda \in \mathbb{R}$, $\forall x \in E$, $\|\lambda x\| = |\lambda| \, \|x\|$;

(N3) $\forall x, y \in E$, $\|x + y\| \leq \|x\| + \|y\|$ (Triangle Inequality).

Example. For $E = \mathbb{R}^n$, $n \geq 1$, and $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we define the norms:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \,, \qquad \|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}, \qquad \|x\|_\infty = \sup_{1 \leq i \leq n} |x_i| \,.$$

Remark 2.2. When **(N1)** is not satisfied, the map is called a semi-norm on E.

Definition 2.3 (Normed space)**.** A linear space equipped with a norm is called a normed space and is denoted $(E, \|\cdot\|)$.

Definition 2.4 (Inner product)**.** A inner product over the space E is a defined as a map $(\cdot,\cdot) : E \times E \to \mathbb{F}$ satisfying the following axioms, for all elements $x, y, z \in E$ and all scalars $a \in \mathbb{F}$:

1. Conjugate symmetry: $(x, y) = \overline{(y, x)}$;
2. Linearity in the first argument: $(ax, y) = a(x, y)$ and $(x + y, z) = (x, z) + (y, z)$;
3. Positive-definiteness: $(x, x) \geq 0$ with equality only for $x = 0$.

Proposition 2.1 (Cauchy-Schwarz Inequality)**.** *From an inner product, we can define a natural norm:* $\|x\| = \sqrt{(x, x)}$*. From the Triangle Inequality, we deduce the Cauchy-Schwarz inequality:*

$$\forall x, y \in E,\ |(x, y)| \leq \|x\|\|y\|. \tag{CS}$$

2.2.2 Cauchy sequences and complete spaces

We recall here the definition of Banach and Hilbert spaces.

Definition 2.5 (Cauchy sequence)**.** Let $(E, \|\cdot\|)$ be a normed space. The sequence $(x_n) \subset E$ is said to be a Cauchy sequence if and only if $\forall \epsilon > 0$, $\exists N \in \mathbb{N}$ such that $\forall p, q > N$, $\|x_p - x_q\| < \epsilon$.

Any converging sequence is Cauchy, however the reciprocal is false.

Definition 2.6 (Completeness)**.** A normed space $(E, \|\cdot\|)$ is complete if and only if any Cauchy sequence converges.

Definition 2.7 (Banach and Hilbert spaces)**.**

- A complete normed space $(E, \|\cdot\|)$ is called a Banach space.
- Let $(E, (\cdot,\cdot)$ be a linear space equipped with an inner product: $(E, (\cdot,\cdot)$ is a normed space for its natural norm. If E is complete then it is called a Hilbert space.

2.2.2.1 Banach fixed point theorem

Let $(E, \|\cdot\|)$ be a Banach space.

Theorem 2.2. *Let $f : E \to E$ be a contracting map,* i.e. *satisfying:* $\exists M \in [0,1), \forall (x,y) \in E \times E$,

$$\|f(x) - f(y)\| \leq M\|x - y\|.$$

Then $\exists!\, x^\star \in E$ *such that* $f(x^\star) = x^\star$.

Proof.

- Uniqueness: this results from the contracting nature of f. If $x_1^\star$ and $x_2^\star$ are two distinct fixed points of f, then $\|f(x_1^\star) - f(x_2^\star)\| = \|x_1^\star - x_2^\star\| \leq M\|x_1^\star - x_2^\star\|$, which implies $\|x_1^\star - x_2^\star\| = 0$ since $M < 1$. Hence $x_1^\star = x_2^\star$.
- Existence: let $x_0 \in E$ and define $x_{k+1} = f(x_k)$ for $k \geq 1$. Then, since f is contracting:

  $$\|x_{k+1} - x_k\| = \|f(x_k) - f(x_{k-1})\| \leq M\|x_k - x_{k-1}\| \leq \ldots \leq M^k\|x_1 - x_0\|.$$

  Since for any $n, k \in \mathbb{N}$, $x_{n+k} - x_n = \sum_{i=0}^{k-1}(x_{n+i+1} - x_{n+i})$, we have:

  $$\|x_{n+k} - x_n\| \leq \sum_{i=0}^{k-1}\|x_{n+i+1} - x_{n+i}\| \leq \left(\sum_{i=0}^{k-1} M^{n+i}\right)\|x_1 - x_0\| \leq \frac{M^n}{1-M}\|x_1 - x_0\|.$$

  This shows that $(x_n)_n$ is a Cauchy sequence, hence convergent because E is a Banach space. By continuity of f (implied by its contracting nature), $x^\star := \lim x_n$ satisfies

  $$f(x^\star) = f(\lim x_n) = \lim f(x_n) = \lim x_{n+1} = x^\star,$$

  which ends the proof.

□

2.2.2.2 Riesz Representation Theorem

Let $(H, (\cdot,\cdot))$ be a Hilbert space. Given $u \in H$, a continuous linear functional L_u can be defined on H by $L_u(v) = (u, v)$. The Riesz representation theorem proves that the converse is also true. We will present a proof for this theorem, important when proving the existence and uniqueness of solutions of weak problems. This proof can be found in [19, Theorem 2.4.2, p. 55].

Theorem 2.3. *Let L be a linear functional on H, a Hilbert space. Then there exists a unique $u \in H$ such that $\forall v \in H$, $L(v) = (u, v)$. Moreover, $\|u\| = \|L\|_{H'} := \sup_{v \in H} \frac{|L(v)|}{\|v\|}$.*

Proof.

- Uniqueness: if there exists $y, z \in H$ such that $\forall v \in H, L(v) = (y, v) = (z, v)$, then $\forall v \in H, (y - z, v) = 0$. Choosing $v = y - z$ leads to $\|y - z\|^2 = 0$, hence $y = z$, since $\|\cdot\|$ is a norm.
- Existence: consider $M = \ker L := L^{-1}(\{0\})$ a closed subspace of H, by continuity of L. Then $H = M \oplus M^\perp$. If $M = H$ then $L \equiv 0$ so take $u = 0$. Otherwise $M^\perp \neq \{0\}$ and we pick $z \in M^\perp$, $z \neq 0$. For any $v \in H$, let $\lambda = L(v)/L(z)$ and we have

  $$L(v - \lambda z) = L(v) - \lambda L(z) = 0, \ \textit{i.e.} \ v - \lambda z \in M.$$

  Thus by orthogonality of M and $M^\perp$, $0 = (z, v - \lambda z) = (z, v) - \frac{L(v)}{L(z)}\|z\|^2$, so rearranging leads to $L(v) = (\frac{L(z)}{\|z\|^2} z, v)$, hence $u = \frac{L(z)}{\|z\|^2} z$ and this ends the proof.

□

2.2.3 Bilinear forms

Let E be a linear space over a field $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$.

Definition 2.8 (Bilinear form)**.** A map $B : E \times E \to \mathbb{F}$ is a bilinear form if it is linear in each argument.

Definition 2.9 (Symmetry)**.** A bilinear form b on a linear space E is *symmetric* if $\forall u, v \in E$, $b(u, v) = b(v, u)$.

Definition 2.10 (Continuity)**.** We assume E to be a normed linear space. Let b be a bilinear form. It will be said to be *continuous* if and only if $\exists C \in \mathbb{R}_+$ such that $\forall u, v \in E$, $|b(u, v)| \leq C\|u\|\|v\|$.

Definition 2.11 (Coercivity)**.** Let b be a bilinear form on E, a normed linear space. b is said to be *coercive* if and only if $\exists \alpha > 0$ such that $\forall v \in E$, $|b(v, v)| \geq \alpha\|v\|^2$.

Theorem 2.4. *Let H be a Hilbert space, and let a be a* continuous, symmetric *bilinear form, that is* coercive *on a closed subspace of H. Furthermore, let l be a continuous linear functional on H, then $\exists! u \in H$ such that $\forall v \in H,\ a(u,v) = l(v)$.*

Proof. This results from the Riesz representation theorem 2.3, applied to the linear function l on the Hilbert space $(V, a(\cdot,\cdot))$. □

2.2.4 Lax-Milgram Theorem

Let $(H, (\cdot,\cdot))$ be a Hilbert space. This theorem lies at the root of the theory of PDEs. A proof is reproduced here. More details can be found in [19, p. 77].

Theorem 2.5. *Let a be a continuous, coercive bilinear form and $l \in H'$ a continuous linear functional on H. Then $\exists!\ u \in H$ such that $\forall v \in H,\ a(u,v) = l(v)$.*

Proof. The application of the Riesz theorem 2.3 to L yields the existence of a unique l such that $\forall v \in H,\ Lv = (l, v)$. Moreover, by bilinearity and continuity of a, the map $\mathcal{A}_u$ such that $\forall u \in H,\ a(u,v) = \mathcal{A}_u(v)$ is also linear and continuous. Similarly, the map $\mathcal{A} : u \mapsto \mathcal{A}_u$ is also linear and continuous, so by application of the Riesz theorem, there exists a unique $A(u) \in H$ such that $\forall v \in H,\ a(u,v) = \mathcal{A}_u(v) = (Au, v)$.
Since $\|A(u)\| = \|\mathcal{A}_u\|_{H'}$, then A is a linear continuous map on H. We then must find a unique $u \in H$ such that $\forall v \in H,\ (Au, v) = (l, v)$, *i.e.* $A(u) = l$. This amounts to showing that A is a bijection on H.

- Injectivity: this results from the coercivity of a and the Cauchy-Schwarz inequality. For any $u \in H$, $\alpha\|u\|^2 \leq |a(u,u)| = |(A(u), u)| \leq \|A(u)\|\|u\|$, which implies that $\|A(u)\| \geq \alpha\|u\|$.
- Surjectivity: consider $A(H) = \{A(u)\ :\ u \in H\}$. We will show $A(H)$ is closed. Take a Cauchy sequence $(Au_n) \subset A(H)$, then u_n is also Cauchy in H since $\|u_n - u_p\| \leq \alpha^{-1}\|A(u_n - u_p)\| \leq \|Au_n - Au_p\|$, for any $n, p \in \mathbb{N}$. Since H is complete, (u_n) converges and Au_n converges to Au by continuity of A. So, one can write $H = A(H) \oplus A(H)^\perp$, and for any $v \in A(H)^\perp$, $(A(v), v) = 0$, which implies $v = 0$. So $A(H)^\perp = \{0\}$ and this ends the proof.

□

2.2.5 Lebesgue spaces

Here we present the construction of Lebesgue integrable functions and their associated spaces. We commence by recalling the definition of a measure, then defining a measurable function, as well as integrable functions and Lebesgue spaces. More details about the construction of the Lebesgue integral are readily available in [9].

Definition 2.12 (σ-algebra)**.** Let S be a set. A σ-algebra Σ over S is a subset of 2^S (the set of all subsets of S) that satisfies the following properties:

- $\emptyset \in \Sigma$.
- For any $A \in \Sigma$, $S \setminus A \in \Sigma$.
- For any $A, B \in \Sigma$, $A \cup B \in \Sigma$, and this can be extended for any countable family of sets in Σ.

Definition 2.13 (Measure)**.** Let S be a set and Σ a σ-algebra over S. A function $\mu : \Sigma \to \mathbb{R}^+ \cup \{+\infty\}$ is called a *measure* if it satisfies:

- $\mu(\emptyset) = 0$.
- For any countable collection $\{S_i\}_{i \in I}$ of pairwise disjointed sets of Σ, we have $\mu(\cup_{i \in I} E_i) = \sum_{i \in I} \mu(E_i)$. This property is called countable additivity.

Definition 2.14 (Measurable function)**.** Let $\Omega \subset \mathbb{R}^n$ be measurable with respect to some measure μ. Then a function $f : \Omega \to \overline{\mathbb{R}}$ is measurable if $\forall \alpha \in \overline{\mathbb{R}}$, the set $\{x \in \Omega : f(x) \geq \alpha\}$ is measurable.

Primarily, the Lebesgue integral is defined over the set of *simple* functions *i.e.* piecewise constant functions. Using density arguments, the definition of the Lebesgue integral can then be extended to bounded measurable functions, then to unbounded positive measurable functions, and finally to any unbounded measurable function. It is not the purpose of this dissertation to reproduce the construction of the Lebesgue integral, however, we wish to provide the reader with its main properties, which we will extensively use in the sequel.

Definition 2.15 (Lebesgue integrable function)**.** Let Ω be a measurable set and let f be a real-valued measurable function over Ω. The function f is said to be *Lebesgue*

integrable if and only if $\int_\Omega |f| \ \mathrm{d}x$ exists and is finite. We denote by $L^1(\Omega)$ the set of measurable Lebesgue integrable functions over Ω, that is to say:

$$L^1(\Omega) = \left\{ f \text{ measurable over } \Omega \ : \ \exists \int_\Omega |f| \ \mathrm{d}x < \infty \right\}.$$

The Lebesgue integral of f is then denoted as $\int_\Omega f \ \mathrm{d}x$.

Proposition 2.6 (Properties of the Lebesgue integral)**.** *The Lebesgue integral on Ω defined on the set $L^1(\Omega)$ satisfies the following properties:*

- *Let $f \in L^1(\Omega)$, then $\left|\int_\Omega f \ \mathrm{d}x\right| \leq \int_\Omega |f| \ \mathrm{d}x$.*
- *Linearity: $\forall f, g \in L^1(\Omega)$, $\forall a, b \in \mathbb{R}$, $\int_\Omega af + bg \ \mathrm{d}x = a\int_\Omega f \ \mathrm{d}x + b\int_\Omega g \ \mathrm{d}x$.*
- *Monotonicity: $\forall f, g \in L^1(\Omega)$ such that $f \leq g$, then $\int_\Omega f \ \mathrm{d}x \leq \int_\Omega g \ \mathrm{d}x$.*
- *Dominated convergence theorem: let (f_n) be a sequence of integrable functions converging almost everywhere to some measurable function f. If there exists $g \in L^1(\Omega)$ such that $|f_n| \leq g$ for any n, then f is integrable and $\int_\Omega f \ \mathrm{d}x = \lim_{n\to\infty} \int_\Omega f_n \ \mathrm{d}x$.*

Definition 2.16 (p-norms and Lebesgue spaces)**.** For $1 \leq p < \infty$ and $f : \Omega \to \mathbb{R}$ a measurable function, let

$$\|f\|_{p,\Omega} := \left(\int_\Omega |f(x)|^p \, \mathrm{d}x \right)^{1/p}.$$

In the case of $p = \infty$,

$$\|f\|_{\infty,\Omega} := \operatorname{ess\,sup} \{ |f(x)| \ : \ x \in \Omega \}.$$

In both cases, Lebesgue spaces are defined as:

$$L^p(\Omega) = \{ f \text{ measurable on } \Omega : \ \|f\|_{p,\Omega} < \infty \}.$$

Whenever no confusion is possible, the second subscript Ω will be dropped.

Remark 2.17. The definition of Lebesgue spaces leads us to identify two functions f and g that only differ on a set of measure zero as the same function, by interpreting $L^p(\Omega)$ as the set of *equivalence classes* of functions with respect to the relation: $f \approx g \Leftrightarrow \|f - g\|_p = 0$.

Theorem 2.7. *For $1 \leq p \leq \infty$, $L^p(\Omega)$ is a Banach space.*
In the special case of $p = 2$, $L^2(\Omega)$ equipped with the inner product

$$(f, g)_{L^2(\Omega)} = \int_\Omega f\, g \,\mathrm{d}x, \text{ for any } f, g \in L^2(\Omega),$$

is a Hilbert space.

Proof. A proof of this important result can be found in [1, Theorem 2.16, p. 29]. □

2.2.6 Weak derivatives

Lesbesgue spaces lie at the root of the construction of Sobolev spaces. We will first define a*weak* derivative for locally integrable function, then formally introduce Sobolev spaces. More details about the definition and results presented here are available in [1, 19].

Definition 2.18 (Multi-index notation)**.** A multi-index, α, is an n-tuple of non-negative integers, α_i. The length of α is given by

$$|\alpha| := \sum_{i=1}^{n} \alpha_i.$$

Notation. Given $\phi \in C^\infty$, the expression $D^\alpha \phi$ denotes the classic pointwise partial derivative of order $|\alpha|$ of ϕ, after differentiating ϕ α_i times with respect to the i-th variable, for $1 \leq i \leq n$. We will also denote by $\partial\phi/\partial x_i$ the quantity $D^{(0,\ldots,0,1,0,\ldots,0)}\phi$ where the only '1' is in the i-th coordinate of the multi-index. Finally, we denote $\nabla\phi := (\partial\phi/\partial x_1, \ldots, \partial\phi/\partial x_n)^\top$ and for $\boldsymbol{u}$ a vector-valued function, $\nabla\cdot\boldsymbol{u} := \sum_{i=1}^n \partial u_i/\partial x_i$.

Definition 2.19. Let $\Omega \subset \mathbb{R}^n$. We will denote by $C_0^\infty(\Omega)$ the set of $C^\infty(\Omega)$ functions with compact support in Ω, *i.e.* such that the set

$$\overline{\{x \in \Omega \ : \ f(x) \neq 0\}} \tag{2.1}$$

is compact and a subset of the interior of Ω, for $f \in C^\infty(\Omega)$.

Example. Take $\Omega = \mathbb{R}^n$, and choose ϕ as:

$$\phi(x) := \begin{cases} \exp\left(-\dfrac{1}{1 - \|x\|^2}\right) & \text{if } \|x\| \leq 1, \\ 0 & \text{if } \|x\| > 1. \end{cases}$$

One can show inductively that ϕ is an element of $C^\infty(\mathbb{R}^n)$ and moreover, its support is the unit closed ball of $(\mathbb{R}^n, \|\cdot\|)$ that is compact since it is closed and bounded in a finite dimensional normed space. Therefore $\phi \in C_0^\infty(\mathbb{R}^n)$. The choice of the norm $\|\cdot\|$ here is irrelevant since all norms on a finite-dimensional space are equivalent.

Definition 2.20 (Locally integrable functions)**.** For some subset $\Omega \subset \mathbb{R}^n$, the set of *locally integrable* functions is defined as

$$L^1_{loc}(\Omega) := \left\{ f \text{ measurable} : \ \forall K \text{ compact} \subset \mathring{\Omega}, f \in L^1(K) \right\},$$

where $\mathring{\Omega}$ is the interior of Ω, the largest opened set contained in Ω.

The next definition is the classical definition of a weak derivative for locally integrable functions, as used in standard books on theory of PDEs [56].

Definition 2.21 (Weak derivative of a locally integrable function)**.** Let $f \in L^1_{loc}(\Omega)$ and α be a multi-index. We say that f has a *weak derivative* of order $|\alpha|$, denoted $D^\alpha_w f$, if there exists a function $g \in L^1_{loc}(\Omega)$ such that

$$\int_\Omega g(x)\phi(x)\,\mathrm{d}x = (-1)^{|\alpha|} \int_\Omega f(x) D^\alpha \phi(x)\,\mathrm{d}x, \quad \forall \phi \in C_0^\infty(\Omega).$$

If such a g exists, then we define $D^\alpha_w f := g$. This definition of a weak derivative for locally integrable functions can be extended further for *distributions* and is called a *distributional* derivative [1]. We will not discuss this extension.

In the next proposition, we state a result that identifies the classical and the weak derivatives of a $C^{|\alpha|}(\Omega)$ function.

Proposition 2.8. *Let α be a multi-index and let $\phi \in C^{|\alpha|}(\Omega)$. Then the weak derivative $D^\alpha_w \phi$ exists and is equal to $D^\alpha \phi$. From now on, we will drop the 'w' in the notation for the weak derivative since it is only an extension of the classical derivative.*

Proof. The proof of this proposition results from the integration by parts formula for $C^{|\alpha|}(\Omega)$ functions. □

2.2.7 Sobolev spaces

Definition 2.22 (Sobolev norms)**.** Let $k \in \mathbb{N}$ and $f \in L^1_{loc}(\Omega)$. Let us assume that the weak derivatives $D^\alpha_w f$ exist, for any $|\alpha| \leq k$. We define the Sobolev norm

$$\|f\|_{k,p,\Omega} := \left(\sum_{|\alpha| \leq k} \|D^\alpha f\|^p_{p,\Omega} \right)^{1/p}, \text{ for } 1 \leq p < \infty.$$

For $p = \infty$, we define

$$\|f\|_{k,\infty,\Omega} := \max_{|\alpha| \leq k} \|D^\alpha f\|_{\infty,\Omega}.$$

We also define the Sobolev semi-norm, for $1 \leq p < \infty$,

$$\begin{aligned} |f|_{k,p,\Omega} &= \left(\sum_{|\alpha| = k} \|D^\alpha f\|^p_{p,\Omega} \right)^{1/p}, \quad p < \infty, \\ |f|_{k,\infty,\Omega} &= \max_{|\alpha| = k} \|D^\alpha f\|_{p,\Omega}, \qquad\qquad p = \infty. \end{aligned}$$

Definition 2.23 (Sobolev spaces)**.** In either case, for any $1 \leq p \leq \infty$, we define the Sobolev space $W^k_p(\Omega)$ by

$$W^k_p(\Omega) = \left\{ f \in L^1_{loc}(\Omega) \ : \ \|f\|_{k,p,\Omega} < \infty \right\}.$$

Theorem 2.9. *For any $1 \leq p \leq \infty$, $W^k_p(\Omega)$ is a Banach space.*
Moreover, in the case of $p = 2$, $H^k(\Omega) := W^k_2(\Omega)$, equipped with the inner product defined by

$$(u, v)_{k,\Omega} = \sum_{|\alpha| < k} (D^\alpha u, D^\alpha v)_{L^2(\Omega)}, \text{ for any } u, v \in H^k(\Omega),$$

is a Hilbert space.

Proof. See for instance [1, Theorem 3.3, p. 60]. □

The next result is a consequence of the mutual inclusion of Lebesgue L^p spaces.

Proposition 2.10. *For any $k, m \in \mathbb{N}$, $k \leq m$, given $1 \leq p \leq \infty$, then $W^m_p(\Omega) \subset W^k_p(\Omega)$. This also implies that $H^m\Omega) \subset H^k(\Omega)$.*

2.2.8 Trace of Sobolev functions

Since Lebesgue integrable functions are not defined pointwisely, the trace, that is the restriction of a Lebesgue function on some manifold does not exist *a priori*. This section introduces a brief description of the mathematical tools necessary to handle traces of Sobolev functions. The next definitions can be found in [1, 19, 56].

Definition 2.24 (Lipischitz domain)**.** Let $\Omega \subset \mathbb{R}^n$ be open and bounded. Informally, a Ω is *Lipschitz* if its boundary can be thought as locally being the graph of a Lipschitz continuous function.
In mathematical terms, this can be expressed in the following way: we say that Ω is a bounded open set with Lipschitz boundary (resp. C^k, $k \geq 1$ boundary) if $\forall x \in \partial\Omega$, $\exists U \subset \mathbb{R}^n$ neighbourhood of x and a bijective map $H \ : \ Q \to U$ where Q is the open unit ball of $\mathbb{R}^n$ for the $\|\cdot\|_\infty$-norm *i.e.*

$$Q = \{x \in \mathbb{R}^n \ : \ |x_j| < 1, j = 1, 2, \ldots, n\}\,.$$

Here, H is Lipschitz (resp. C^k) on $\overline{Q}$, H^{-1} Lipschitz (resp. C^k) on $\overline{U}$, and $H(Q_+) = U \cap \Omega$, $H(Q_0) = U \cap \partial\Omega$ with $Q_+ = \{x \in Q \ : \ x_n > 0\}$ and $Q_0 = \{x \in Q \ : \ x_n = 0\}$.

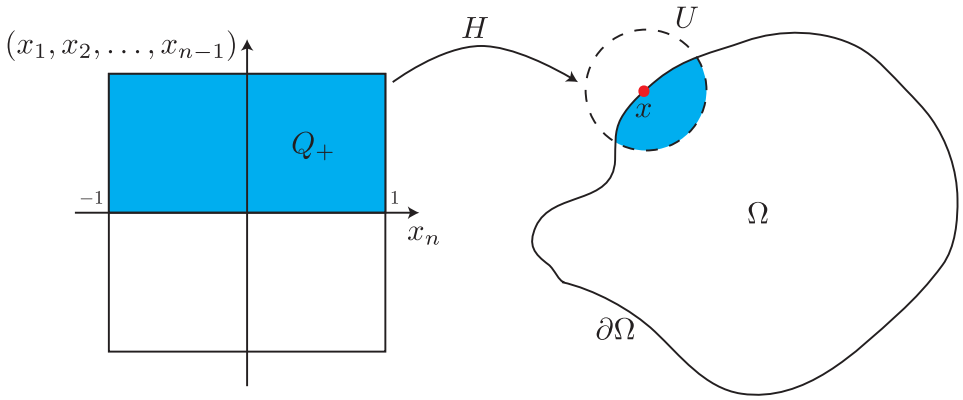


Figure 2.1: A Lipschitz domain with Lipschitz boundary [35, Figure 1.1, p. 34].

Definition 2.25. Let us consider $\Omega \subset \mathbb{R}^n$ to be a Lipschitz domain, then we can define the integral on $\partial\Omega$ of some function $f \in L^1(\mathbb{R}^n)$ as

$$\int_{\partial\Omega} f \ \mathrm{d}s := \int_{\mathbb{R}^{n-1}} f(x, \phi(x)) \sqrt{1 + |\nabla\phi(x)|^2} \ \mathrm{d}x, \tag{2.2}$$

since $\partial\Omega$ is interpreted as the graph of a Lipschitz function ϕ.
In this expression, $\sqrt{1+|\nabla\phi(x)|^2} \in L^\infty$, since ϕ is Lipschitz (recall that $\mathrm{Lip}(\Omega) \subset W^1_\infty(\Omega)$).

Theorem 2.11. *Let $\Omega \subset \mathbb{R}^n$ be a Lipschitz domain, and $1 \leq p \leq \infty$, then $\exists C > 0$ such that $\forall v \in W^1_p(\Omega)$:*

$$\|v\|_{p,\partial\Omega} \leq C\|v\|^{1-1/p}_{p,\Omega}\|v\|^{1/p}_{1,p,\Omega}. \tag{2.3}$$

In particular, for $p = 2$, we can write: $\exists C > 0$ such that $\forall v \in H^1(\Omega)$,

$$\|v\|_{2,\partial\Omega} \leq C\|v\|_{1,2,\Omega}. \tag{2.4}$$

This last inequality implies that the linear map $(\cdot)|_{\partial\Omega} : H^1(\Omega) \to L^2(\partial\Omega)$ is well-defined and continuous on $H^1(\Omega)$. However, this map is not onto, since any function in $L^2(\partial\Omega)$ cannot be written as the trace of a $H^1(\Omega)$ function. We then denote by $H^{1/2}(\partial\Omega)$ the image of $(\cdot)|_{\partial\Omega}$, i.e. *$H^{1/2}(\partial\Omega) = (H^1(\Omega))|_{\partial\Omega}$.*

Proof. See [77]. □

Definition 2.26. We denote:

- $H^1_0(\Omega) = \{v \in H^1(\Omega) \ : \ \gamma_0 v := v|_{\partial\Omega} = 0 \text{ in } L^2(\partial\Omega)\}$.
- $H^k_0(\Omega) = \{v \in H^k(\Omega) \ : \ D^\alpha v|_{\partial\Omega} = 0 \text{ in } L^2(\partial\Omega),\ \forall\, |\alpha| < k\}$.

Furthermore, we will also denote by $H^1_g(\Omega)$ the affine space of $H^1(\Omega)$ functions u such that $u|_{\partial\Omega} = g$.

2.3 Variational formulation of elliptic problems

We are now in the position to prove existence and uniqueness theorems for some important elliptic problems: the Poisson problem equipped with homogeneous Dirichlet boundary conditions and homogeneous Neumann boundary conditions. We will present for each problem two different approaches leading to different variational problems, also qualified as weak in the literature, namely the classical formulation in H^1 and a so-called mixed formulation in $H^{\mathrm{div}} \times L^2$, that can be constructed by introducing a new variable in the problem.

2.3.1 Calculus tools

Let $\Omega \subset \mathbb{R}^n$ an open bounded domain with Lipschitz boundary. We denote by $\boldsymbol{n}$ the outward unit normal vector to $\partial\Omega$, which is by assumption in $L^\infty(\Omega)^n$.

Proposition 2.12 (Divergence theorem for Sobolev functions)**.** *Let $\boldsymbol{u} \in W_1^1(\Omega)^n$, then*

$$\int_\Omega \nabla \cdot \boldsymbol{u} \, \mathrm{d}x = \int_{\partial\Omega} \boldsymbol{u} \cdot \boldsymbol{n} \, \mathrm{d}x. \tag{2.5}$$

Proposition 2.13. *Let $u \in H^2(\Omega)$ and $v \in H^1(\Omega)$, then*

$$\int_\Omega (-\Delta u)\, v \, \mathrm{d}x = \int_\Omega \nabla u : \nabla v \, \mathrm{d}x - \int_{\partial\Omega} (\nabla u \cdot \boldsymbol{n})\, v \, \mathrm{d}s, \tag{2.6}$$

where

$$\nabla u : \nabla v = \sum_{i=1}^{2} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i}.$$

This is the application of the divergence theorem extended to Sobolev functions.

Proof. The proof for this result can be found in [77, p. 62]. □

2.3.2 Poincaré-Friedrich's Inequality

We state the Poincaré-Friedrich's inequality for H^1 Sobolev functions.

Proposition 2.14. *Let Ω be an open bounded Lipschitz domain, then $\exists C > 0$ such that $\forall u \in H^1(\Omega)$,*

$$\|u - \bar{u}\|_{2,\Omega} \leq C \|\nabla u\|_{2,\Omega}, \text{ with } \bar{u} = \frac{1}{|\Omega|} \int_\Omega u \, \mathrm{d}x.$$

Another version can be written as:

Proposition 2.15. *Let Ω be an open bounded Lipschitz domain, then $\exists C > 0$ such that $\forall u \in H_0^1(\Omega)$,*

$$\|u\|_{2,\Omega} \leq C \|\nabla u\|_{2,\Omega}.$$

Proof. A proof for those results is readily available in [56, Theorem 1, p. 275] in the case of a domain Ω with a C^1 boundary. □

2.3.3 Variational formulation of Poisson's Equation with homogeneous Dirichlet boundary conditions

Let us consider the Dirichlet problem posed in $\Omega \subset \mathbb{R}^n$, an open and bounded Lipschitz domain :

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{2.7}$$

where Δ denotes the Laplace operator :

$$\Delta := \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}. \tag{2.8}$$

The problem (2.7) is classically known as a Poisson problem and this PDE arises in numerous physical applications [111]. It describes the distribution of potential of a vector field such as the electric field when the right-hand side represents a charge distribution, or the potential of a gravitational field when the right-hand side is a distribution of masses.

Definition 2.27. Let $u \in H_0^1(\Omega)$, then u is called a *weak* solution of Problem (2.7) if

$$\forall v \in H_0^1(\Omega), \quad a(u,v) := \int_\Omega \nabla u : \nabla v \, \mathrm{d}x = \int_\Omega f \, v \, \mathrm{d}x := l(v). \tag{2.9}$$

This equation is called the variational or weak formulation of the PDE (2.7).

We are now in position to state the existence and uniqueness of the solution of a Poisson problem with homogeneous boundary conditions.

Proposition 2.16. *Let $u \in H^2(\Omega) \cap H_0^1(\Omega)$ and $f \in L^2(\Omega)$. Then the following statements are equivalent:*

1. *u solves Problem* (2.7)*.*
2. *u is a weak solution of Problem* (2.7)*.*

Proof. This result is a direct consequence of Proposition 2.13. □

The following theorem is an important result proving the existence and uniqueness of the solution of the weak formulation of the Poisson problem. We provide a proof for this result, which is only an application of the Lax-Milgram theorem.

Theorem 2.17. *The weak problem posed in $H_0^1(\Omega)$:*

$$\text{find } u \in H_0^1(\Omega) \text{ such that } \forall v \in H_0^1(\Omega), \quad a(u,v) = l(v) \tag{2.10}$$

admits a unique solution.

Proof. The bilinear form a is symmetric. Moreover, using Poincaré's inequality in $H_0^1(\Omega)$:

$$a(v,v) = \|\nabla v\|_{2,\Omega}^2 \geq \frac{1}{2}(\|\nabla v\|_{2,\Omega}^2 + \|\nabla v\|_{2,\Omega}^2) \geq \frac{1}{2}\|\nabla v\|_{2,\Omega}^2 + \frac{1}{2C^2}\|v\|_{2,\Omega}^2 \geq C_0\|v\|_{1,2,\Omega}^2,$$

with $C_0 = \frac{1}{2}\min(1, \frac{1}{C^2}) > 0$. This shows that a is coercive on $H_0^1(\Omega)$.
The continuity of a follows from the Cauchy-Schwarz inequality :

$$\begin{aligned} \forall u, v \in H_0^1(\Omega), \quad |a(u,v)| &\leq \|\nabla u\|_{2,\Omega}\|\nabla v\|_{2,\Omega} && \text{(Cauchy-Schwartz inequality in } L^2(\Omega)), \\ &\leq \|u\|_{1,2,\Omega}\|v\|_{1,2,\Omega} && (\|\nabla u\|_{2,\Omega} \leq \|u\|_{1,2,\Omega}). \end{aligned}$$

The boundedness of the linear map l follows similarly from the Cauchy-Schwarz inequality:

$$\begin{aligned} \forall v \in H_0^1(\Omega), \quad |l(v)| &\leq \|f\|_{2,\Omega}\|v\|_{2,\Omega}, && \text{(Cauchy-Schwartz inequality in } L^2(\Omega)), \\ &\leq \|f\|_{2,\Omega}\|v\|_{1,2,\Omega}, && (\|\nabla v\|_{2,\Omega} \leq \|v\|_{1,2,\Omega}). \end{aligned}$$

So the bilinear form a is coercive and continuous, and the linear map l is continuous as well. Hence the Lax-Milgram Theorem 2.5 can be applied and this yields the existence and uniqueness of the solution of the weak equation in $H_0^1(\Omega)$. □

Remark 2.28. In the case of non-homogeneous Dirichlet boundary conditions, u satisfying $u = g$ on $\partial\Omega$ with $g \in H^{1/2}(\partial\Omega)$, then $\exists w \in H^1(\Omega)$ such that $w|_{\partial\Omega} = g$. Then $\bar{u} = u - w \in H_0^1(\Omega)$ satisfies the homogenous Dirichlet problem (2.7) with data $\bar{f} = f - (-\Delta w)$. In this case, $f \notin L^2(\Omega)$, but $f \in H^{-1}(\Omega) = (H^1(\Omega))'$ instead. The existence and uniqueness result for the Poisson equation derived only for L^2 right-hand sides still holds in this case [56].

2.3.4 Variational formulation of Poisson's Equation with homogeneous Neumann boundary conditions

Let us consider the Neumann problem posed in the open and bounded Lipschitz domain $\Omega \subset \mathbb{R}^n$:

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ \nabla u \cdot \boldsymbol{n} &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{2.11}$$

The homogeneous Neumann boundary condition is also called the *no flux* boundary condition because it models an isolated boundary when u represents a temperature distribution or the concentration of a chemical.

Definition 2.29. Let $L^2_0(\Omega)$ be a subspace of $L^2(\Omega)$ defined by:

$$L^2_0(\Omega) = \left\{ v \in L^2(\Omega) \ : \ \int_\Omega v \, \mathrm{d}x = 0 \right\}.$$

This is the subspace of $L^2(\Omega)$ functions with zero mean.

Proposition 2.18 (Variational formulation in $V(\Omega)$)**.** *Suppose $f \in L^2_0(\Omega)$. Let $u \in H^2(\Omega) \cap L^2_0(\Omega)$ be a solution of Problem* (2.11)*, then u satisfies*

$$\forall v \in V(\Omega) = H^1(\Omega) \cap L^2_0(\Omega), \quad a(u,v) := \int_\Omega \nabla u : \nabla v \, \mathrm{d}x = \int_\Omega f \, v \, \mathrm{d}x := l(v). \quad (2.12)$$

Proof. This results from Proposition 2.13. □

Theorem 2.19. *The weak problem posed in $V(\Omega)$:*

$$\textit{find } u \in V(\Omega) \textit{ such that } \forall v \in V(\Omega), \quad a(u,v) = l(v) \quad (2.13)$$

admits a unique solution.

Proof. The bilinear form a is continuous as we know from Theorem 2.17. The coercivity on $V(\Omega)$ is obtained using the first version of Poincaré-Friedrich's inequality 2.14, with $\bar{u} = 0$. Again, the Lax-Milgram Theorem 2.5 yields the existence and uniqueness of the solution to the weak formulation in the space $V(\Omega)$. □

2.3.5 Mixed variational formulation of the Poisson's equation

Let us again consider the Poisson's equation posed in $\Omega \subset \mathbb{R}^n$, equipped with homogenous Dirichlet or Neumann boundary conditions. We now introduce a new variable, namely $\boldsymbol{\sigma}$, and write the new system of equations:

$$\begin{array}{ll} \boldsymbol{\sigma} + \nabla u = 0 & \text{in } \Omega \\ -\nabla \cdot \sigma = f & \text{in } \Omega \\ u = 0 \text{ or } \boldsymbol{\sigma} \cdot \boldsymbol{n} = 0 & \text{on } \partial\Omega. \end{array} \quad (2.14)$$

The new variable σ represents the gradient of u at any point in Ω and takes values in $\mathbb{R}^n$. The variational formulation that is presented here will then involve not only u but

also $\boldsymbol{\sigma}$. For this purpose, we need to define a Sobolev space tailored to this new vectorial variable. This is precisely the object of the next definition.

Definition 2.30. Let $H^{\mathrm{div}}(\Omega)$ be a subspace of $L^2(\Omega)^n$ defined by

$$H^{\mathrm{div}}(\Omega) = \left\{ \boldsymbol{\tau} \in L^2(\Omega)^n \ : \ \nabla \cdot \boldsymbol{\tau} \in L^2(\Omega) \right\}.$$

This is the Sobolev space of L^2-integrable fields with L^2-integrable divergence. It is equipped with its natural norm given by

$$\|\boldsymbol{\tau}\|^2_{H^{\mathrm{div}}(\Omega)} = \|\boldsymbol{\tau}\|^2_{2,\Omega} + \|\nabla \cdot \boldsymbol{\tau}\|^2_{2,\Omega}. \tag{2.15}$$

We also define

$$H^{\mathrm{div}}_0(\Omega) = \left\{ \boldsymbol{\tau} \in H^{\mathrm{div}}(\Omega) \ : \ \boldsymbol{\tau} \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega \right\}.$$

Proposition 2.20 (Mixed variational formulation of the Poisson's equation with homogeneous Dirichlet boundary conditions)**.** *Let $f \in L^2(\Omega)$ and assume that $\boldsymbol{\sigma} \in H^{div}(\Omega)$ and $u \in H^1_0(\Omega)$ satisfy* (2.14)*, then they are also solutions of :*

$$\forall \boldsymbol{\tau} \in H^{div}(\Omega), \quad a(\boldsymbol{\sigma}, \boldsymbol{\tau}) + b(\boldsymbol{\tau}, u) = 0 \tag{2.16a}$$

$$\forall v \in L^2(\Omega), \qquad\qquad b(\boldsymbol{\sigma}, v) = (f, v)_{2,\Omega}, \tag{2.16b}$$

where the forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are defined as

$$a(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \sum_{i=1}^{n} (\sigma_i, \tau_i)_{2,\Omega},$$
$$b(\boldsymbol{\tau}, v) = (\nabla \cdot \boldsymbol{\tau}, v)_{2,\Omega}.$$

This weak formulation is symmetric and qualified as 'mixed' *because it involves two different spaces $H^{div}(\Omega)$ and $L^2(\Omega)$.*

Proof. This formulation is derived in [22, p. 18]. □

Proposition 2.21 (Mixed variational formulation of the Poisson equation with homogeneous Neumann boundary conditions)**.** *Let $f \in L^2_0(\Omega)$ and assume that $\boldsymbol{\sigma} \in H^{div}_0(\Omega)$ and $u \in H^1(\Omega) \cap L^2_0(\Omega)$ satisfy* (2.14)*, then they are also solutions of*

$$\forall \boldsymbol{\tau} \in H^{div}_0(\Omega), \quad a(\boldsymbol{\sigma}, \boldsymbol{\tau}) + b(\boldsymbol{\tau}, u) = 0 \tag{2.17a}$$

$$\forall v \in L^2_0(\Omega), \qquad\qquad b(\boldsymbol{\sigma}, v) = (f, v)_{2,\Omega}. \tag{2.17b}$$

Proof. This formulation is derived in [22, p. 18]. □

In order to prove the existence and uniqueness of the weak problems above, we need to verify a so-called *'inf-sup condition'* for $b(\cdot,\cdot)$ that arises in any class of saddle-point problems, involving two different spaces, as introduced by Babuska [7] and Brezzi [22]. This leads us to write the following abstract result due to Babuska and Brezzi.

Theorem 2.22 (Babuska-Brezzi theorem)**.** *Let V and Q be two Hilbert spaces. Let a be a continuous bilinear form on $V \times V$ and let b be a continuous bilinear form on $V \times Q$. We assume that the range of $B : v \to b(v,\cdot)$ is closed in Q' – the dual of Q, that is to say, $\exists k > 0$ such that:*

$$\sup_{v \in V} \frac{|b(v,q)|}{\|v\|} \geq k\|q\|_{Q/\ker B^\top}. \tag{2.18}$$

We also assume that a is coercive on $\ker B$, so there exists $\alpha > 0$ such that:

$$\begin{aligned} \inf_{u \in \ker B} \sup_{v \in \ker B} \frac{|a(u,v)|}{\|u\|\|v\|} &\geq \alpha, \\ \inf_{v \in \ker B} \sup_{u \in \ker B} \frac{|a(u,v)|}{\|u\|\|v\|} &\geq \alpha. \end{aligned} \tag{2.19}$$

Then there exists a unique solution (u,p) to the mixed problem:

$$\begin{aligned} &\forall v \in V, \quad a(u,v) + b(v,u) = 0 \\ &\forall q \in Q, \qquad\qquad b(u,q) = (f,q). \end{aligned} \tag{2.20}$$

Proof. A proof can be found in [21]. □

Proposition 2.23. *There exists a unique solution $(u,\boldsymbol{\sigma}) \in (L^2(\Omega) \times H^{div}(\Omega))$ (resp. $(u,\boldsymbol{\sigma}) \in (L^2(\Omega) \times H_0^{div}(\Omega))$) to the weak problem* (2.16a) *(resp.* (2.17a)*).*

Proof. We choose $V = H^{\mathrm{div}}(\Omega)$ and $Q = L^2(\Omega)$ (resp. $V = H_0^{\mathrm{div}}(\Omega)$ and $Q = L_0^2(\Omega)$). A proof can be found in [22, Example 1.2, p. 43] for the 'inf sup' condition. □

2.4 Finite element formulation of elliptic problems

Having introduced the notion of variational formulation for a simple elliptic problem, we present here the basic definitions and results for using the FEM.

Firstly, we define a finite element and provide basic definitions and results in approximation theory for two classes of elements that we will use in further chapters: the Lagrange and the Raviart–Thomas triangular elements.

Secondly, we will derive finite element formulations for a Poisson problem, using the classical variational formulation presented in Sections 2.3.3 and 2.3.4 with Lagrange elements. We will introduce the classical results that allow us to prove the convergence of the finite-element method in these cases. Lastly, we will deal with the finite element approximation of the more complex mixed variational problems seen in 2.3.5 using first order Raviart–Thomas elements.

2.4.1 Definition of a finite element

Definition 2.31 (Finite element)**.** This definition is due to Ciarlet [29]. Let

(i) $K \subset R^n$ be a bounded close set with nonempty interior and piecewise smooth boundary. K is called the *element domain.*

(ii) $\mathcal{P}$ be a finite-dimensional space of functions on K. $\mathcal{P}$ is called the space of shape functions,

(iii) $\mathcal{N} = \{N_1, N_2, \ldots, N_d\}$ be a basis for $\mathcal{P}'$, the dual of $\mathcal{P}$. $\mathcal{N}$ is called the set of *nodal variables.*

Then $(K, \mathcal{P}, \mathcal{N})$ is called a *finite element.*

Definition 2.32 (Nodal basis)**.** Given $(K, \mathcal{P}, \mathcal{N})$ a finite element, the basis $\{\phi_1, \phi_2, \ldots, \phi_k\}$ of $\mathcal{P}$ dual to $\mathcal{N}$ *i.e.* such that $N_i(\phi_j) = \delta_i j$ is called the *nodal basis* of $\mathcal{P}$.

Definition 2.33 (Local interpolant)**.** Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element with nodal basis $\{\phi_1, \phi_2, \ldots, \phi_d\}$, then for a function v, we define the *local interpolant* by

$$\mathcal{I}_K(v) := \sum_{i=1}^{d} N_i(v)\phi_i.$$

The next theorem is a general result that provides a bound for the error of interpolation on a finite element, assuming certain properties on the finite element. More details on polynomial approximation theory in Sobolev spaces can be found in [19].

Theorem 2.24 (Local approximation theorem)**.** *Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element satisfying:*

(i) K is star-shaped with respect to some ball and has a diameter of order h;

(ii) $\mathcal{P}_{m-1} \subset \mathcal{P} \subset W^m_\infty(K)$;

(iii) $\mathcal{N} \subset (C^l(\overline{K}))'$, meaning that the nodal variables in $\mathcal{N}$ involve derivatives up to order l, and that any $N \in \mathcal{N}$ is a bounded linear functional on $C^l(\overline{K})$.

Let us suppose that $1 \leq p \leq \infty$ and either $m - l - n/p > 0$ if $p > 1$ or $m - l - n \geq 0$ if $p = 1$. Then there exists $C > 0$ such that for $0 \leq i \leq m$ and $v \in W^m_p(K)$

$$|v - \mathcal{I}_K(v)|_{i,p,K} \leq C(diam\ K)^{m-i}\ |v|_{m,p,K}\,. \tag{2.21}$$

The constant C depends on m, l, n, p, K but not on v. The diameter of K can be defined as the smallest real number r for which there exists $x \in K$ such that $K \subset B(x, r)$.

Proof. A proof for this theorem can be found in [19, p. 118]. □

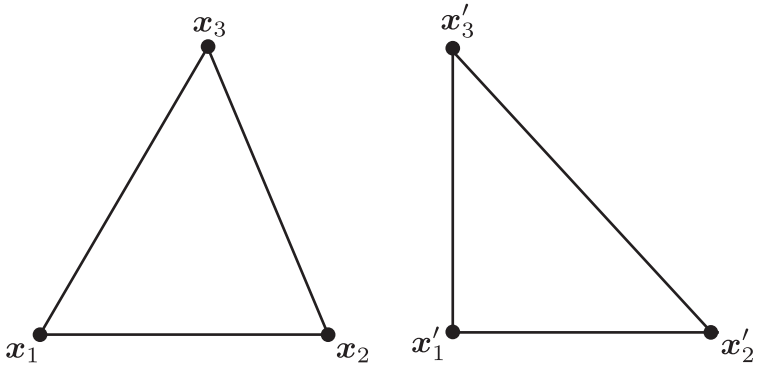


Figure 2.2: Example of two affine-equivalent elements.

Definition 2.34 (Affine-equivalent elements)**.** Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element and $F(= Ax + b)$ a nonsingular affine map. The finite element $(\widehat{K}, \widehat{\mathcal{P}}, \widehat{\mathcal{N}})$ is said *affine-equivalent* to $(K, \mathcal{P}, \mathcal{N})$ if

(i) $F(K) = \widehat{K}$,

(ii) $\mathcal{P} = \widehat{\mathcal{P}} \circ F := \left\{ \hat{f} \circ F \ : \ \hat{f} \in \widehat{\mathcal{P}} \right\}$,

(iii) $\forall \widehat{N} \in \widehat{\mathcal{N}}$, $\exists N \in \mathcal{N}$ such that $\widehat{N}(\hat{f}) = N(\hat{f} \circ F)$.

Definition 2.35 (Subdivision)**.** Let $\Omega \subset \mathbb{R}^n$ be a polygonal domain. A *subdivision* $\mathcal{T}$ of the domain Ω is a finite collection of element domains $\{K_i\}$ such that

(i) $\mathring{K}_i \cap \mathring{K}_j = \emptyset$ if $i \neq j$,

(ii) $\bigcup K_i = \overline{\Omega}$.

Definition 2.36 (Mesh size)**.** When denoting a subdivision by $\mathcal{T}^h$, then the parameter h, a strictly positive real number, refers to the diameter of the largest element domain of $\mathcal{T}^h$:

$$h = \max\{r \ : \ \forall i, \exists x_i \in K_i, \ K_i \subset B(x_i, r)\}. \tag{2.22}$$

When h is large, the subdivision $\mathcal{T}^h$ is then said to be '*coarse*' and when h is small, $\mathcal{T}^h$ is said to be '*fine*'.

Definition 2.37. Assume Ω is a domain to which we assign a subdivision $\mathcal{T}$. We also assume that each element domain K_i is part of a finite element $(K_i, \mathcal{P}_i, \mathcal{N}_i)$. For some function v defined on Ω, we define its *global interpolant* by

$$\mathcal{I}_{\mathcal{T}}|_{K_i}(v) = \mathcal{I}_{K_i}(v) \quad \forall K_i \in \mathcal{T}.$$

Definition 2.38 (C^r elements)**.** A finite element $(K, \mathcal{P}, \mathcal{N})$ is said to be C^r, with $r \geq 0$, if $\mathcal{I}^h(C^l(\overline{\Omega})) \subset C^r(\Omega) \cap W^{r+1}_\infty(\Omega)$. In other words, this element is C^r if its global interpolator $\mathcal{I}^h$ interpolates C^l functions on $\overline{\Omega}$ as at least C^r functions on Ω with all derivatives up to order $r+1$ bounded for the L^2 norm.

The next result is a general theorem that gives a bound, in terms of the mesh size h, for the error of interpolation using a subdivision of Ω of mesh size h.

Theorem 2.25 (Global approximation theorem)**.** *Let $h > 0$ and $\mathcal{T}^h$ be a 'well-behaved'[1] subdivision of a polyhedral domain Ω. Let $(K, \mathcal{P}, \mathcal{N})$ be a reference element satisfying the assumptions of Theorem 2.24 for some l, m and p.*

Suppose further that any element $(T, \mathcal{P}_T, \mathcal{N}_T)$ of $\mathcal{T}^h$ is affine-interpolation equivalent to $(K, \mathcal{P}, \mathcal{N})$. Then there exists $C > 0$ depending only on the reference element, m, l, n

[1] The requirement on the regularity of the subdivision $\mathcal{T}^h$ is rather complex; since in further chapters, we will only deal with regular subdivisions – the most regular type of subdivisions – we do not wish to go in more details about this requirement. In the following, we will refer to 'well-behaved' subdivisions as subdivisions that are regular enough so they satisfy the assumptions of the previous result.

and p such that $\forall s$ satisfying $0 \leq s \leq \min(m, r+1)$,

$$\|v - \mathcal{I}^h(v)\|_{s,p,\Omega} \leq Ch^{m-s} |v|_{m,p,\Omega}, \quad \forall v \in W_p^m(\Omega). \tag{2.23}$$

Proof. A proof for this result is presented in [19, Corollary 4.4.24, p. 123] □

Remark 2.39. Despite the seemingly complexity of its assumptions, the result given by this theorem can be interpreted in the following way: when the chosen subdivision $\mathcal{T}^h$ of Ω is sufficiently 'well-behaved', and when using continuous (or more regular) elements on this subdivision, then the order of approximation of $\mathcal{I}^h(v)$ for v in $H^s(\Omega)$ norm worsens as s increases. When s is equal to m, then the interpolant $\mathcal{I}^h(v)$ only approximates v to some constant with respect to the $H^m(\Omega)$ norm. For instance, if we choose to interpolate a $H^2(\Omega)$ function using continuous elements in $H^1(\overline{\Omega})$, then the order of approximation will be $h^{2-1} = h$.

2.4.2 Useful examples of triangular elements

In this section, we work in dimension $n = 2$. Let K be a triangle with vertices $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{x}_3$, and $\mathcal{P}_k(K)$, $k \geq 0$, the set of polynomials of degree less than k on K. In further chapters, we will make an extensive use of the triangular elements presented here.

2.4.2.1 Lagrange triangular elements

Definition 2.40 (Area coordinates)**.** Let T be a triangle with vertices $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$. The area coordinate $\lambda_i(\boldsymbol{x})$ for $1 \leq i \leq 3$ and $\boldsymbol{x} \in T$ is defined by the area of the triangle $(\boldsymbol{x}, \boldsymbol{x}_j, \boldsymbol{x}_k)$ with $j \neq k$ and $j, k \neq i$, divided by the area of T. Area coordinates add up to 1 and can be used to parameterise the location of any point in the triangle.

Definition 2.41 (Linear Lagrange triangular element)**.** Let $\mathcal{P} = \mathcal{P}_1(K)$ and $\mathcal{N} = \{N_1, N_2, N_3\}$ defined by $N_i(v) = v(\boldsymbol{x}_i)$. Then the triplet $(K, \mathcal{P}, \mathcal{N})$ is called a Lagrange triangular element. A basis of $\mathcal{P}$ is formed by $\{\lambda_1, \lambda_2, \lambda_3\}$ the area coordinates (or the barycentric coordinates) functions, as shown in Figure 2.3. By definition, they have the property that $\lambda_i(\boldsymbol{x}_j) = \delta_{ij}$.

Definition 2.42 (Lagrange element of order k)**.** Let $k \geq 1$ be an integer. Let $\mathcal{P} = \mathcal{P}_k(K)$ and $\mathcal{N} = \{N_{0,0}, \ldots, N_{i,j}, \ldots, N_{k,0}\}_{\substack{0 \leq i \leq k \\ 0 \leq j \leq k-i}}$, defined by

$$N_{i,j}(v) = v(\boldsymbol{x}_{ij}) \text{ where } \boldsymbol{x}_{ij} = \left(1 - \frac{i+j}{k}\right)\boldsymbol{x}_1 + \left(\frac{i}{k}\right)\boldsymbol{x}_2 + \left(\frac{j}{k}\right)\boldsymbol{x}_3.$$

At node $\boldsymbol{x}_{ij}$, the polynomial P_{ij} takes the value 1 and 0 at any other node. The arrangement of the nodes $\boldsymbol{x}_{ij}$ is depicted in Figure 2.3b. The construction of each polynomial P_{ij} can be performed by noticing that for each node $\boldsymbol{x}_{ij}$, we can always find k straight lines passing by all the other nodes at least once. Hence, the polynomial P_{ij} will be defined as the product of the equations of those lines, which are polynomials of degree 1, multiplied by a normalisation constant so that $P_{ij}(\boldsymbol{x}_{ij}) = 1$. For instance, in the case $k = 3$, we can write the basis $\mathcal{P}$ in terms of the area coordinates variables:

$$\begin{array}{lll}
P_{0,0} = \frac{9}{2}\lambda_1(\lambda_1 - \frac{1}{3})(\lambda_1 - \frac{2}{3}), & P_{1,0} = \frac{27}{2}\lambda_1(\lambda_1 - \frac{1}{3})\lambda_2, & P_{2,0} = \frac{27}{2}\lambda_1(\lambda_2 - \frac{2}{3})\lambda_2, \\
P_{3,0} = \frac{9}{2}\lambda_2(\lambda_2 - \frac{1}{3})(\lambda_2 - \frac{2}{3}), & P_{0,1} = \frac{27}{2}\lambda_1(\lambda_1 - \frac{1}{3})\lambda_3, & P_{1,1} = \frac{27}{1}\lambda_1\lambda_2\lambda_3, \\
P_{2,1} = \frac{27}{2}\lambda_2(\lambda_2 - \frac{1}{3})\lambda_3, & P_{0,2} = \frac{27}{2}\lambda_1\lambda_3(\lambda_3 - \frac{1}{3}), & P_{1,2} = \frac{27}{2}\lambda_2\lambda_3(\lambda_3 - \frac{1}{3}), \\
P_{0,3} = \frac{9}{2}\lambda_3(\lambda_3 - \frac{1}{3})(\lambda_3 - \frac{2}{3}). & &
\end{array} \tag{2.24}$$

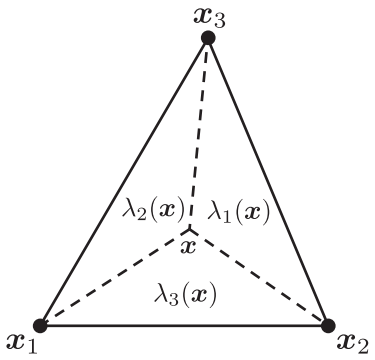
Note that the dimension of $\mathcal{N}$ is $\frac{1}{2}(k+1)(k+2))$ when $\mathcal{P} = \mathcal{P}_k$. Hence, for $k = 1$, $\mathcal{N}$ has dimension 3, for k=2, $\dim \mathcal{N} = 6$ and for $k = 3$, $\dim \mathcal{N} = 10$.

Proposition 2.26 (Approximation properties of Lagrange elements)**.** *Let $u \in H^m(\Omega)$ and let $\mathcal{T}^h$ be a 'well-behaved' triangulation of Ω, and suppose that any element $(K_T, \mathcal{P}_{m-1}(T), \mathcal{N}_T)$ is a (m-1)-order Lagrange triangular element on T. Then the application of Theorem 2.25 yields, for any $0 \leq s \leq m$,*

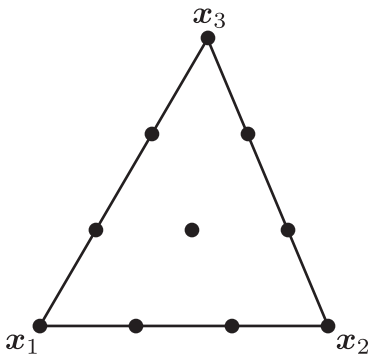
$$\|u - \mathcal{I}^h(u)\|_{s,2,\Omega} \leq Ch^{m-s} |v|_{m,2,\Omega}. \tag{2.25}$$

2.4.2.2 Linear Raviart–Thomas triangular elements

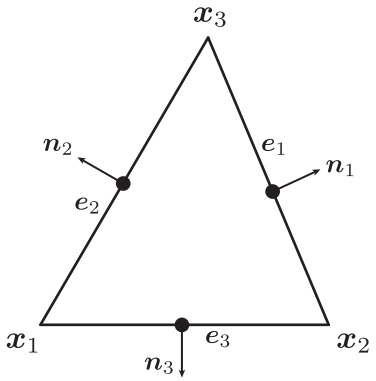
We present here the linear triangular Raviart–Thomas element [115], depicted in Figure 2.3c. We define below the interpolation polynomials and the interpolator. Lastly, we provide a result on the approximation properties of this linear element.



(2.3a) Linear Lagrange triangular element



(2.3b) Lagrange element of order $k = 3$



(2.3c) Raviart–Thomas triangular element of order 0

Figure 2.3: Linear Lagrange triangular element with node-located degrees of freedom and corresponding area coordinates $\lambda_1, \lambda_2, \lambda_3$ in Figure 2.3a, arrangement of local degrees of freedom in the triangular Lagrange element of order $k = 3$ in Figure 2.3b and Raviart–Thomas triangular element in Figure 2.3c with edge-located degrees of freedom.

Definition 2.43 (Linear Raviart–Thomas triangular element)**.** Let $\mathcal{RT}_0(K) = \mathcal{P}_0(K)^2 + \boldsymbol{x}\,\mathcal{P}_0(K) \subset \mathcal{P}_1(K)^2$ and $\mathcal{N} = \{N_1, N_2, N_3\}$ be defined by

$$N_i(\boldsymbol{v}) = \int_{e_i} \boldsymbol{v} \cdot \boldsymbol{n}_i \, \mathrm{d}s, \tag{2.26}$$

where $\boldsymbol{e}_i$ denotes the i-th edge of the triangle K, and $\boldsymbol{n}_i$ its outing normal. Then the triplet $(K, \mathcal{RT}_0(K), \mathcal{N})$ is defined as the first order Raviart–Thomas triangular element. We observe that $\nabla \cdot (\mathcal{RT}_0(K)) \subset \mathcal{P}_0(K)$.

Furthermore, the $\mathbb{R}^2$-valued polynomials defined as

$$\boldsymbol{\psi}_i(\boldsymbol{x}) = \frac{|\boldsymbol{e}_i|}{2\,|K|}(\boldsymbol{x} - \boldsymbol{x}_i), \quad \forall i \in \{1, 2, 3\} \tag{2.27}$$

have the property that $\boldsymbol{\psi}_i(\boldsymbol{e}_j) \cdot \boldsymbol{n}_j = \delta_{ij}$ and form a basis of $\mathcal{RT}_0(K)$.

The expression for the local interpolator $\mathcal{I}_K$ defined on $H^1(K)$ is:

$$\mathcal{I}_K(\boldsymbol{v}) = \sum_{i=1}^{3} N_i(\boldsymbol{v})\, \boldsymbol{\psi}_i. \tag{2.28}$$

Definition 2.44 (Global Raviart–Thomas interpolator)**.** Let $\mathcal{T}^h$ be a triangulation of Ω with mesh size h. Associated with this triangulation, we introduce the global space:

$$\mathcal{RT}_0(\mathcal{T}_h) = \left\{ \boldsymbol{\tau} \in H^{\mathrm{div}}(\Omega) \ : \ \boldsymbol{\tau}|_T \in \mathcal{RT}_0(T), \quad \forall T \in \mathcal{T}_h \right\}. \tag{2.29}$$

When no confusion arises, we will drop the indication of triangulation from the definition. The global interpolator is then defined on $H^{\mathrm{div}}(\Omega) \cap \prod_{T \in \mathcal{T}_h} H^1(T)^n$ by $\mathcal{I}^h(\boldsymbol{\tau})\big|_T = \mathcal{I}_T(\boldsymbol{\tau})$, where $\mathcal{I}_T(\boldsymbol{\tau})$ is the local interpolator defined in Definition 2.43.

Proposition 2.27 (Local approximation property for a linear triangular Raviart–Thomas element)**.** *Let $(\mathcal{K}, \mathcal{RT}_0(T), \mathcal{N}_T)$ be a linear triangular Raviart–Thomas element with support on triangle T. Then, there exists $C > 0$ such that for $\boldsymbol{\tau} \in H^m(T)^2$ with $\nabla \cdot \boldsymbol{\tau} \in H^m(T)$*

$$\|\boldsymbol{\tau} - \mathcal{I}^h(\boldsymbol{\tau})\|_{H^{div}(T)} \leq C h^{m+1} \left(\|\boldsymbol{\tau}\|_{m,2,T} + \|\nabla \cdot \boldsymbol{\tau}\|_{m,2,T} \right). \tag{2.30}$$

Proof. This result was originally proved by Raviart and Thomas in [115, Theorem 3]. □

Proposition 2.28 (Global approximation properties for the linear triangular Raviart–Thomas element)**.** *Let $\mathcal{T}^h$ be a sufficiently well-behaved triangulation of Ω with mesh size h. We assume that for any triangle T of $\mathcal{T}^h$, $(\mathcal{K}_T, \mathcal{RT}_0(T), \mathcal{N}_T)$ is a linear triangular Raviart–Thomas element. Then there exists $C > 0$ such that for any $\boldsymbol{\tau} \in H^m(T)^2$ with $\nabla \cdot \boldsymbol{\tau} \in H^m(T)$,*

$$\|\boldsymbol{\tau} - \mathcal{I}^h(\boldsymbol{\tau})\|_{H^{div}(T)} \leq Ch^{m+1}\|\boldsymbol{\tau}\|_{0,2,\Omega}. \tag{2.31}$$

Proof. This result is also due to Raviart and Thomas [115]. □

2.4.3 Finite element approximation of Poisson's Equation

Let $\Omega \subset \mathbb{R}^n$ a Lipschitz domain. We wish to find an approximation of the weak solution of the Poisson equation, with either homogeneous Dirichlet or Neumann boundary conditions. Recall the weak formulation of the Poisson's equation:

$$\text{find } u \in V \ : \ a(u, v) = l(v), \forall v \in V, \tag{2.32}$$

where V is either chosen as $H_0^1(\Omega)$ in the homogeneous Dirichlet case or $H^1(\Omega) \cap L_0^2(\Omega)$ in the homogenous Neumann case. For those two choices, V is a (closed) subspace of $H^1(\Omega)$. Recall the properties of a and l: a is a bilinear continuous and coercive form and l is a linear continuous map on V. The following definition and results are classical and for more details, we would refer the reader to [19].

Definition 2.45 (Galerkin Approximation problem)**.** Let $V_h \subset V$ be a finite-dimensional subspace of V, then the Galerkin approximation problem reads as follows:

$$\text{find } u_h \in V_h \ : \ a(u_h, v) = l(v), \forall v \in V_h. \tag{2.33}$$

Proposition 2.29. *There exists a unique u_h solving for the Galerkin approximation* (2.33)*.*

Proof. This is a direct consequence of the Riesz Representation Theorem 2.3. □

Proposition 2.30 (Galerkin Orthogonality)**.** *Let u and u_h be, respectively, the (unique) solution of* (2.32) *and* (2.33)*, then*

$$a(u - u_h, v) = 0, \quad \forall v \in V_h. \tag{2.34}$$

Proof. We observe that:

$$\forall v \in V_h, a(u - u_h, v) = a(u, v) - a(u_h - v) = l(v) - l(v) = 0, \tag{2.35}$$

which is exactly the required result. □

Lemma 2.31 (Céa's Lemma)**.** *The unique solutions u and u_h of the variational problem* (2.32) *and* (2.33) *satisfy*

$$\|u - u_h\|_{1,2,\Omega} \leq \frac{C}{\alpha} \min_{v \in V_h} \|u - v\|_{1,2,\Omega}, \tag{2.36}$$

where C is the continuity constant of a and α its coercivity constant, on V.

Proof. For all $v \in V_h$,

$$\begin{aligned}
\alpha\|u - u_h\|^2_{1,2,\Omega} &\leq a(u - u_h, u - u_h) && \text{(by coercivity of } a\text{)} \\
&\leq a(u - u_h), u - v) + a(u - u_h, v - u_h) && \text{(linearity w.r.t the 2nd argument)} \\
&\leq a(u - u_h, u - v) && \text{(using Galerkin orthogonality)} \\
&\leq C\|u - u_h\|_{1,2,\Omega}\|u - v\|_{1,2,\Omega}. && \text{(by continuity of } a\text{)}
\end{aligned}$$

Finally, we obtain:

$$\begin{aligned}
\|u - u_h\|_{1,2,\Omega} &\leq \frac{C}{\alpha}\|u - v\|_{1,2,\Omega}, && \forall v \in V_h, \\
&\leq \frac{C}{\alpha} \inf_{v \in V_h} \|u - v\|_{1,2,\Omega}, \\
&\leq \frac{C}{\alpha} \min_{v \in V_h} \|u - v\|_{1,2,\Omega}, && \text{(since V is closed).}
\end{aligned}$$

□

Theorem 2.32. *We suppose that $\mathcal{T}^h$ is a subdivision of Ω. Let $\mathcal{I}^h$ denote a global interpolator for a set of finite elements based on the element domains of $\mathcal{T}^h$ and V_h be a finite-dimensional subspace. Let us make the following assumptions:*

(i) $\mathcal{I}^h(u)$ is continuous and the interpolant satisfies the following approximation property:

$$\|u - \mathcal{I}^h(u)\|_{1,2,\Omega} \leq Ch^m \left|u\right|_{m,2,\Omega}. \tag{2.37}$$

(ii) $V_h \subset V$.

(iii) $\mathcal{I}^h(V) \subset V_h$.

Then the unique solution of the Galerkin approximation problem satisfies

$$\|u - u_h\|_{1,2,\Omega} \leq Ch^{m-1} \left|u\right|_{m,2,\Omega}. \tag{2.38}$$

Proof. Let us prove this result in the special case of $m = 1$ and $\mathcal{T}^h$ being a triangulation of Ω. If we choose V_h as

$$V_h = \{v_h \in C^0(\overline{\Omega}) \cap H^1(\Omega) \ : \ \forall T \in \mathcal{T}^h, \ v_h|_T \ \in \mathcal{P}_1(T)\}, \tag{2.39}$$

then by definition $V_h \subset V = H^1(\Omega)$ and we know from Proposition 2.26 that first order linear Lagrange elements satisfy Equation (2.37), for $m = 2$. Lastly, by construction, $\mathcal{I}^h(V) \subset V_h$. Then using Lemma 2.31, we obtain the required result for $m = 2$. □

2.4.4 Mixed finite element formulation of Poisson's Equation

Let $\Omega \subset \mathbb{R}^n$ be a Lipschitz domain. In this section, we aim to derive a mixed finite element formulation of Poisson's Equation, equipped with homogeneous Dirichlet or Neumann boundary conditions. Let us recall the weak mixed formulation of Poisson's equation:

$$\text{find } (\boldsymbol{\sigma}, u) \in W \times M \ : \quad \begin{array}{ll} \forall \boldsymbol{\tau} \in W, & a(\boldsymbol{\sigma}, \boldsymbol{\tau}) + b(\boldsymbol{\tau}, u) = 0 \\ \forall v \in M, & b(\boldsymbol{\sigma}, v) = l(v), \end{array} \tag{2.40}$$

where W and M are, respectively, subspaces of $H^{\text{div}}(\Omega)$ and $L^2(\Omega)$. Recall the properties of $a(\cdot,\cdot)$, $b(\cdot,\cdot)$ and $l(\cdot)$:

- $a(\cdot,\cdot)$ is a continuous coercive bilinear form on $W \times W$,
- $b(\cdot,\cdot)$ is a continuous bilinear form on $W \times M$ and satisfies an 'inf-sup' condition,
- $l(\cdot)$ is a continuous form on M.

Definition 2.46 (Mixed approximation problem)**.** Let $W_h \subset W$ and $M_h \subset M$ be finite-dimensional subspaces of W and M. The approximation of the mixed problem reads:

$$\text{find } (\boldsymbol{\sigma}_h, u_h) \in W_h \times M_h \ : \quad \begin{array}{ll} \forall \boldsymbol{\tau} \in W_h, & a(\boldsymbol{\sigma}_h, \boldsymbol{\tau}) + b(\boldsymbol{\tau}, u_h) = 0 \\ \forall v \in M_h, & b(\boldsymbol{\sigma}_h, v) = l(v). \end{array} \tag{2.41}$$

Let us make the following choice for the finite-dimensional subspaces W_h and M_h of W and M:

$$W_h = \mathcal{RT}_0(\Omega) = \left\{ \boldsymbol{\tau}_h \in C^0(\overline{\Omega}) \cap H^{\mathrm{div}}(\Omega) \ : \ \forall T \in \mathcal{T}^h, \ \boldsymbol{\tau}_h|_T \in \mathcal{RT}_0(T) \right\}, \tag{2.42}$$

$$M_h = \left\{ v_h \in L^2(\Omega) \ : \ \forall T \in \mathcal{T}^h, \ v_h|_T \in \mathcal{P}_0(T) \right\}. \tag{2.43}$$

Proposition 2.33 (Discrete Inf-sup condition)**.** *For the above choice of W_h and M_h, there exists $k_0 > 0$ such that*

$$\inf_{v \in M_h} \sup_{\boldsymbol{\tau} \in W_h} \frac{|b(\boldsymbol{\tau}, v)|}{\|\boldsymbol{\tau}\|_W \|v\|_M} \geq k_0. \tag{2.44}$$

Proof. The proof of this result is due to Raviart and Thomas [115]. A crucial point in the proof is to observe that $\nabla \cdot W_h \subset M_h$. □

Brezzi and Fortin [21] give bounds for the error estimate using the previous discrete 'inf-sup' condition. It is analogous to Céa's theorem if we choose the norm $\|(\cdot,\cdot)\| = \|\cdot\|_W + \|\cdot\|_M$ for the product space $W \times M$ in which we seek the solution couple $(\boldsymbol{\sigma}, u)$. We will not go into the details of the proof.

Theorem 2.34. *Let $(\boldsymbol{\sigma}, u) \in W \times M$ and $(\boldsymbol{\sigma}_h, u_h) \in W_h \times M_h$ be, respectively, the solutions of* (2.40) *and* (2.41)*. Let us assume that b satisfies the discrete inf-sup condition* (2.44) *for some choice of discretisation spaces W_h and M_h, the following estimate holds, where c is a constant independent of h:*

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_W + \|u - u_h\|_M \leq c \left(\inf_{\boldsymbol{\tau}_h \in W_h} \|\boldsymbol{\sigma} - \boldsymbol{\tau}_h\|_W + \inf_{v_h \in M_h} \|u - v_h\|_M \right). \tag{2.45}$$

Proof. This general result was proved by Brezzi and Fortin [21]. □

Proposition 2.35 (Bound for the error estimate)**.** *Using the same assumptions as in Theorem 2.34, given the choice of discretisation subspaces W_h and M_h performed in Equations* (2.42)*, we obtain:*

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_W + \|u - u_h\|_M \leq \hat{c} h \left(\|\boldsymbol{\sigma}\|_W + \|u\|_M \right), \tag{2.46}$$

where $\hat{c}$ is a positive constant, independent of h.

Proof. This inequality results from Theorem 2.34 and the approximation properties of the Raviart–Thomas element in Proposition 2.28. □

In summary, the use of linear triangular Raviart–Thomas elements for the approximation of the flux variable $\boldsymbol{\sigma}$ with piecewise constant elements for that of the density variable u leads to $O(h)$ convergence of the numerical scheme. This is comparable with the classical first-order Lagrange finite element formulation, yielding the same convergence rate in the H^1 norm.

2.5 Summary

In this chapter, we have laid down important theoretical mathematical tools for the variational formulation of elliptic problems. In particular, we have reproduced theoretical existence and uniqueness results for the solution of the Poisson equation equipped with Dirichlet or Neumann boundary conditions.

Those mathematical notions are necessary in order to introduce the FEM, a numerical method to approximate the solution of PDEs. We have then given two examples of finite elements, the well-known Lagrange element and the Raviart–Thomas element, both defined on triangles, and given their approximation properties.

Finally, we have used them to produce two distinct finite element formulations of the Poisson problem. We displayed in each case a bound for the error estimate and concluded that both formulations yielded comparable convergence properties.

In the next chapter, we introduce a subset of the FEM, called the fictitious domain method. It will take advantage of the two FEMs presented in this chapter.

Chapter 3

Fictitious domain methods for elliptic problems

3.1 Introduction

The *fictitious domain method* (FDM), also known as 'domain embedding method', is a family of finite-element based numerical methods to compute approximations of solutions of partial differential equations. The main motivation for this, is so that a simple structured domain can be employed. Such structured meshes improve the speed of solution techniques that may be used: for instance, algebraic multigrid methods are known to be very efficient on regular meshes. FDMs are particularly popular for simulating time-dependent problems in which the domain changes in time. This avoids costly remeshing and interpolation steps between time steps. On the other hand, the method introduces new complications. In particular, we have to compute and store information concerning the geometry of the problem, which is no longer encoded in the mesh of the original computational domain. This is achieved by introducing a new variable accounting for the geometry and imposing the boundary condition at the same time.

The FDM was known in the western world under the name of domain embedding method and was originally designed by Buzbee *et al.* in the 1970s for solving elliptic problems posed on irregular domains using a rectangular mesh [28]. The name of FDM is due to numerical analysts from the former Soviet Union. Glowinski, in the early 1990s, introduced the first FDM for the FEM using a Lagrange multiplier for solving Dirichlet

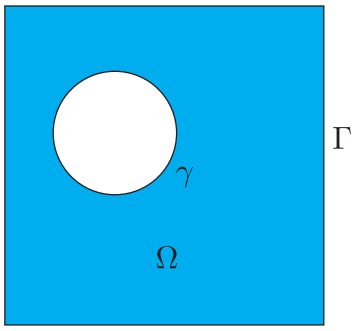
problems [68]. The main motivation of his paper consisted in simulating a nonlinear Navier-Stokes equation in a cylindrical pipe.

However, the most important domain of application for the FDM is the simulation of particulate flows. In 1999, Glowinski *et al.* published an article discussing the simulation of particles falling in a Newtonian fluid, by coupling the movement of the particles with the state of the fluid [71, 67]. Also in 1999, Joly and Rhaouti [79] discussed a new FDM for elliptic problems with Neumann boundary conditions. Their technique relies on Lagrange multipliers as well, however the setup and the discretisation scheme differ, when using Raviart-Thomas finite elements.

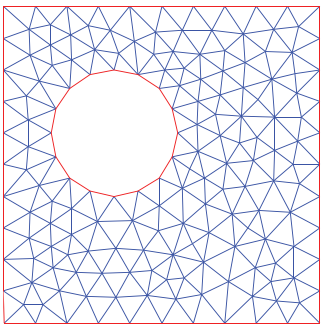
Finally, in 2011, Glowinski and He published a FDM for solving Robin elliptic boundary problems [66]. However, after we have implemented it and reproduced their numerical results, it seemed to us that the convergence properties of this method were not satisfactory. This behaviour was not highlighted in their paper, mainly because their numerical experiments were based on a computational domain in which the elliptic obstacle was small compared to the largest chosen mesh size. As a consequence, the presence of the obstacle was numerically barely noticeable, and did not upset their convergence results.

Hence, we present here our own alternative solution for solving Robin elliptic boundary problems with a fictitious finite element method. The main motivation for developing such a method is linked to the fact that Robin boundary conditions model the chemical interaction between external chemical signals and the cell membrane in cellular biology [133], for instance, the binding of a chemical signal on the receptors distributed on the surface of the cell membrane.

In this chapter, we will introduce in Section 3.2 Glowinski's original FDM for Dirichlet problems, which we apply to a Poisson's equation. In Section 3.3, we present Joly and Rhaouti's FDM for Neumann problems, and finally, in the Section 3.4, we show our contribution to the FDM, by adapting Joly and Rhaouti's method for Robin boundary problems.



(3.1a) Configuration of domain Ω



(3.1b) A possible triangulation of Ω

Figure 3.1: Representation of Ω and a possible triangulation. The shape of the obstacle is here represented as a disk, but no assumption in the method further developed has been made on its shape. We only assume that the boundary γ is smooth enough (at least Lipschtiz).

3.2 Fictitious domain method for Dirichlet elliptic problems

This first example introduces FDM and was elaborated by Glowinski [68]. We will make use of this method later for solving a Stokes problem on a moving domain, in which the flow is created by slowly moving solid objects in a Newtonian viscous fluid. This FDM is commonly used in simulations of particulate flows [67].

3.2.1 Initial problem

Let Ω be an open and bounded domain of $\mathbb{R}^2$ with Lipschitz boundary $\partial\Omega = \Gamma \cup \gamma$ such that $\Gamma \cap \gamma = \emptyset$, as shown in Figure 3.1a. This domain is composed of a square of which the surface of an obstacle has been removed, in other words, Ω is a patched domain. The shape of the obstacle inside the domain Ω in principle has little importance, as long as the immersed boundary γ is regular enough *i.e.* Lispschitz. In Figure 3.1a, the obstacle has been represented as a disk.

Let us consider a generic reaction-diffusion equation, which is encountered countless times in computational biology and particularly in modelling of chemotaxis:

$$\begin{aligned} \partial_t u - \nu \Delta u &= f(u,t) \quad \text{in } \Omega, \\ u &= 0 \qquad\quad \text{on } \gamma. \end{aligned} \tag{3.1}$$

Now, let us partially discretise this equation, with respect to time, and use an implicit Euler scheme with time step $\Delta t > 0$. We obtain:

$$\begin{array}{rl} -\nu\Delta u_n + \frac{1}{\Delta t}u_n = f(u_{n-1}, t_{n-1}) + \frac{1}{\Delta t}u_{n-1} & \text{in } \Omega, \\ u_n = 0 & \text{on } \gamma. \end{array} \tag{3.2}$$

This last equation leads us to consider the following generic homogeneous Dirichlet problem. Let f be a $L^2(\Omega)$ distribution such that u satisfies:

$$\begin{array}{rl} -\nu\Delta u + \alpha u = f & \text{in } \Omega, \\ u = 0 & \text{on } \gamma, \end{array} \tag{3.3}$$

where $\alpha \geq 0$ and $\nu > 0$. The parameter α represents the inverse of the time step, as in Equation (3.2). In the case where the parameter α is non zero, increasing α adds coercivity to the problem since the coercivity constant is equal to $\min(\alpha, \nu)$. One can wonder why no boundary condition was provided on Γ: the reason is because, in our problem, it does not play any specific role. For simplicity we assume an homogeneous Dirichlet boundary condition at Γ, but the method we will discuss does not depend on this.

We can see from previous chapter that Problem (3.3) has a unique weak solution u in $H_0^1(\Omega)$, of the following variational problem:

$$\text{find } u \in H_0^1(\Omega) \ : \ a_\Omega(u, v) = (f, v)_{2,\Omega}, \quad \forall v \in H_0^1(\Omega), \tag{3.4}$$

where the form $a_\Omega : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is defined as

$$a_\Omega(u, v) = \int_\Omega (\alpha uv + \nu\nabla u \cdot \nabla v) \ \mathrm{d}x = (\alpha u, v)_{2,\Omega} + (\nu\nabla u, \nabla v)_{2,\Omega}, \quad \forall u, v \in H_0^1(\Omega) \tag{3.5}$$

which is a symmetric coercive bilinear form on $H_0^1(\Omega)$.

3.2.2 Classical finite-element approximation of Problem (3.3)

We consider a triangulation $\mathcal{T}_h$, $h > 0$ of domain Ω such as shown in Figure 3.1b. We have committed the usual variational crime [19, Chapter 10] by approximating the inside boundary γ by a set of segments, so the computational domain is Ω_h and no longer Ω. Using triangular first-order Lagrange elements, the approximation space $V_h \subset H_0^1(\Omega_h)$ is defined as

$$V_h = \left\{ v_h \in H_0^1(\Omega_h) \cap C^0(\overline{\Omega_h}) \ : \ v_h|_K \in \mathcal{P}_1(K), \forall K \in \mathcal{T}_h \right\}. \tag{3.6}$$

The approximation problem reads:

$$\text{find } u_h \in V_h \ : \ a_{\Omega_h}(u_h, v_h) = (f, v_h)_{2,\Omega_h}, \quad \forall v \in V_h. \tag{3.7}$$

Since V_h is finite-dimensional, we can write (3.7) in the form of a linear system of equations. Let us denote $N = \dim V_h$ and define $U \in \mathbb{R}^N$ the vector of the components of u_h in the nodal basis of V_h, we have in this case:

$$\text{find } U \in \mathbb{R}^N \ : \ AU = F, \text{ with } A = (a_{\Omega_h}(v_i, v_j))_{1\leq i,j\leq N} \text{ and } F = ((f, v_i)_{2,\Omega_h})_{1\leq i\leq N}. \tag{3.8}$$

The matrix A is symmetric positive definite.

3.2.3 A fictitious domain formulation

As performed in [65], let us derive a fictitious domain formulation of Problem (3.4). The computational domain Ω can be drastically simplified if one 'fills' the hole in the square and one would be able to fit a regular mesh on the new square domain. Let us define $\widetilde{\Omega}$ the square domain resulting from filling the hole in Figure 3.1a.

In order to encode the geometry of the problem and enforce the boundary condition at the immersed interface γ, a new unknown is required, that we call λ. It will make its first appearance in the fictitious domain weak formulation. As we will discover later on, this unknown is only defined on the interface curve, not on the whole volume domain Ω.

Now, we change the initial Dirichlet problem (3.3) to: given $\widetilde{f} \in L^2(\widetilde{\Omega})$, find a function u satisfying:

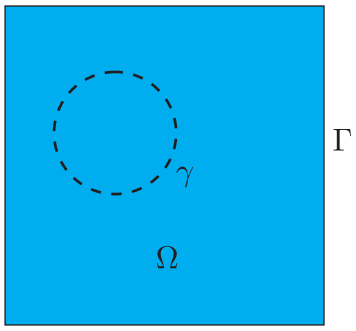
$$\begin{aligned} -\nu\Delta u + \alpha u &= \widetilde{f} \quad \text{in } \widetilde{\Omega}, \\ u &= 0 \quad \text{on } \partial\widetilde{\Omega} \cup \gamma. \end{aligned} \tag{3.9}$$

This equation contains one boundary condition but also an interface condition, since γ is immersed in $\widetilde{\Omega}$. One can easily check that the restriction to Ω of any solution of (3.9) also satisfies (3.3) if $\widetilde{f}\Big|_{\Omega} = f$.

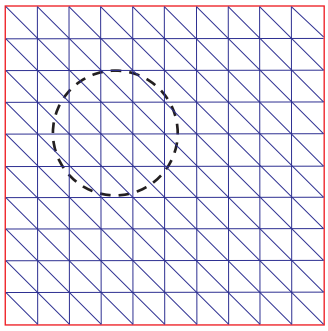
Proposition 3.1. *A fictitious domain variational formulation of* (3.3) *reads:*
find $\widetilde{u} \in H_0^1(\widetilde{\Omega})$ *and* $\lambda \in H^{-1/2}(\gamma)$ *such that*

$$\forall v \in H_0^1(\widetilde{\Omega}), \quad a_{\widetilde{\Omega}}(\widetilde{u}, v) + b(v, \lambda) = (\widetilde{f}, v)_{2,\widetilde{\Omega}} \tag{3.10a}$$

$$\forall \mu \in H^{-1/2}(\gamma), \qquad b(\widetilde{u}, \mu) = 0, \tag{3.10b}$$



(3.2a) Fictitious domain $\widetilde{\Omega}$



(3.2b) Regular triangulation of $\widetilde{\Omega}$

Figure 3.2: Representation of the fictitious domain $\widetilde{\Omega}$ and a possible regular uniform triangulation, that ignores the presence of the obstacle.

where the form $b(\cdot,\cdot) : H_0^1(\widetilde{\Omega}) \times H^{-1/2}(\gamma) \to \mathbb{R}$ *is defined by:* $b(v,\lambda) = \int_\gamma v\,\lambda\,\mathrm{d}s$.

The operator $b(\cdot,\cdot)$ is a visibly non symmetric operator that projects a function defined on the bulk domain $\widetilde{\Omega}$ on the interface γ. It carries the geometrical information about the location and shape of the interface. The unknown λ is associated to it, as well as the new set of test functions, which elements are generically symbolised by μ in Equation (3.10b).

Proof. Let us denote $\omega = \widetilde{\Omega}\backslash\overline{\Omega}$ and write for sufficiently smooth $\widetilde{u}$ and for any $v \in H_0^1(\widetilde{\Omega})$:

$$\begin{aligned}(-\nu\Delta\widetilde{u}, v)_{2,\widetilde{\Omega}} &= (-\nu\Delta\widetilde{u}, v)_{2,\Omega} + (-\nu\Delta\widetilde{u}, v)_{2,\omega} \\ &= (\nu\nabla\widetilde{u}, \nabla v)_{2,\Omega} + (\nu\nabla\widetilde{u}, \nabla v)_{2,\omega} - \int_\gamma (\nu\nabla\widetilde{u}\cdot\boldsymbol{n}_\Omega + \nu\nabla\widetilde{u}\cdot\boldsymbol{n}_\omega)\, v\,\mathrm{d}s \\ &= (\nu\nabla\widetilde{u}, \nabla v)_{2,\widetilde{\Omega}} - \int_\gamma [\nu\nabla\widetilde{u}\cdot\boldsymbol{n}_\gamma]\, v\,\mathrm{d}s.\end{aligned}$$

This is the result of simple algebraic manipulations using the divergence theorem on the two domains Ω and ω. Lastly, we define $\boldsymbol{n}_\gamma \in L^\infty(\gamma)$ with constant Euclidian norm 1, as the outwards normal of γ, and $[\cdot]$ represents the jump across the interface γ of the quantity between brackets.

We then define the new variable $\lambda = -[\nu\nabla\widetilde{u}\cdot\boldsymbol{n}_\gamma]$ living on γ. Since $v|_\gamma \in H^{1/2}(\gamma)$, then a priori $\lambda \in \left(H^{1/2}(\gamma)\right)' = H^{-1/2}(\gamma)$ and $\int_\gamma v\,\lambda\,\mathrm{d}s$ has to be interpreted as the dual pairing between $H^{1/2}(\gamma)$ and $H^{-1/2}(\gamma)$. This yields:

$$(-\nu\Delta\widetilde{u}, v)_{2,\widetilde{\Omega}} = (\nu\nabla\widetilde{u}, \nabla v)_{2,\widetilde{\Omega}} + b(v,\lambda).$$

The condition at the interface γ is then expressed by $b(\widetilde{u},\mu)=0$, $\forall \mu \in H^{-1/2}(\gamma)$. As a result, we obtain (3.10). □

The next result states the 'inf-sup' condition sufficient to obtain existence and uniqueness of Problem (3.10).

Proposition 3.2 (Inf-sup condition)**.** *There exist $C>0$ such that:*

$$\inf_{\mu\in H^{-1/2}(\gamma)} \sup_{v\in H_0^1(\widetilde{\Omega})} \frac{|b(v,\mu)|}{\|v\|_{1,2,\widetilde{\Omega}}\|\mu\|_{-\frac{1}{2},2,\gamma}} \geq C. \tag{3.11}$$

Proof. This proof comes from [64]. The map $\mu \mapsto \sup_{v\in H_0^1(\widetilde{\Omega})} \frac{|b(v,\mu)|}{\|v\|_{1,2,\widetilde{\Omega}}}$ defines a norm on $H^{-1/2}(\gamma)$ that is equivalent to the natural norm of $H^{-1/2}(\gamma)$. Hence there exists $C>0$ such that:

$$C\|\mu\|_{-\frac{1}{2},2,\gamma} \leq \sup_{v\in H_0^1(\widetilde{\Omega})} \frac{|b(v,\mu)|}{\|v\|_{1,2,\widetilde{\Omega}}}, \quad \forall \mu \in H^{-1/2}(\gamma). \tag{3.12}$$

Hence the result. □

Theorem 3.3 (Existence and uniqueness of the solution of the fictitious problem.)**.** *Let $\widetilde{f}$ be an extension in $L^2(\widetilde{\Omega})$ of $f\in L^2(\Omega)$. For this choice of $\widetilde{f}$, there exists a unique solution $(\widetilde{u},\lambda)\in H_0^1(\widetilde{\Omega})\times H^{-1/2}(\gamma)$ to the fictitious problem* (3.10) *and $\widetilde{u}$, restricted to Ω, satisfies* (3.4)*.*

Proof. Because a is continuous and coercive on $H_0^1(\widetilde{\Omega})$ and b satisfies the above inf-sup condition, then the existence and uniqueness of the solution to this mixed problem results from the application of the Babuska-Brezzi Theorem 2.22.

Now, let us consider the extension map $\pi_{\widetilde{\Omega}} : H_0^1(\Omega) \to H_0^1(\widetilde{\Omega})$ extending any $v\in H_0^1(\Omega)$ to an element $\widetilde{v}\in H_0^1(\widetilde{\Omega})$ such that $\widetilde{v}|_\Omega = v$ and $\widetilde{v}|_\omega = 0$. This map is well defined and taking any $v\in \pi_{\widetilde{\Omega}}(H_0^1(\Omega)) \subset H_0^1(\widetilde{\Omega})$, then we obtain $b(v,\lambda)=0$ and moreover:

$$a_{\widetilde{\Omega}}(\widetilde{u},v) = a_\Omega(\widetilde{u},v) = (\widetilde{f}\Big|_\Omega, v) = (f,v). \tag{3.13}$$

Finally, since $\forall \mu \in H^{-1/2}(\gamma)$, $b(\widetilde{u},\mu)=0$, hence $\widetilde{u}|_\Omega \in H_0^1(\Omega)$ and satisfies the homogeneous Dirichlet boundary condition at γ, therefore $\widetilde{u}|_\Omega$ is the unique solution of (3.4). □

Remark 3.1. The variational problem (3.10) is called a hybrid formulation because it involves two test spaces $H_0^1(\widetilde{\Omega})$ and $H^{-1/2}(\gamma)$ that are defined on domains of different topological nature: $\widetilde{\Omega}$ in this example is a 2D domain whereas γ is only a 1D curve. The operator $b(\cdot,\cdot)$ associates the two spaces. This is the operator carrying the information about the geometry of the problem. The new unknown λ is a Lagrange multiplier that enforces the Dirichlet boundary condition at the interface γ.

This new formulation will require us to introduce two different approximation spaces, conforming to each test space. This makes the implementation of the FDM somewhat more complex than the classical finite-element discretisation.

However, we could argue that the FDM only consists in transferring the geometric complexity from the triangulation to the variational formulation itself. Indeed, in classical finite elements, the triangulation is adapted to the geometry of the problem as shown in Figure 3.1b whereas in the fictitious domain finite element formulation, the mesh as shown in Figure 3.2b covers the whole square. The information relating to the actual geometry was transferred in the operator b, associating the volume test functions in $H_0^1(\widetilde{\Omega})$ and the interface test functions in $H^{-1/2}(\gamma)$.

This is the very characteristic that makes the FDM so useful for simulating partial differential equations defined on time-dependent domains. At each time step, instead of changing the triangular mesh fitted to some time-dependent domain $\Omega(t)$, one has to compute the new discretised operator b adapted to the new geometry.

Remark 3.2. We see that $\widetilde{f}$ has been defined such that:

$$\widetilde{f}\Big|_{\Omega} = f. \tag{3.14}$$

We then have some freedom to choose how $\widetilde{f}$ is defined in $\omega = \widetilde{\Omega} \setminus \overline{\Omega}$, even if one usually extends f by 0.

3.2.4 Finite-element approximation of (3.10)

Let $\widetilde{\mathcal{T}}_h$, $h > 0$ be a regular triangulation of the extended domain $\widetilde{\Omega}$ such as shown in 3.2b and $\mathcal{L}_\eta$ a set of M curved segments of length $\eta > 0$ such that $\mathcal{L}_\eta$ partitions γ. A priori,

the mesh sizes h and η are completely decoupled. We then define the approximation spaces:

$$\widetilde{V}_h = \left\{ v \in C^0(\bar{\widetilde{\Omega}}) \cap H_0^1(\widetilde{\Omega}) \ : \ v|_K \in \mathcal{P}_1(K), \forall K \in \widetilde{\mathcal{T}}_h \right\}, \tag{3.15}$$

$$\Lambda_\eta = \left\{ \ v \in H^{-1/2}(\gamma) \ : \ v|_S \in \mathcal{P}_0(S), \forall S \in \mathcal{L}_\eta \right\}. \tag{3.16}$$

The approximation space $\widetilde{V}_h$ is the space of piecewise linear polynomials on $\widetilde{\Omega}$, *i.e.* the set of first order Lagrange triangular elements. The approximation of (3.10) hence is written as: find $(\widetilde{u}_h, \lambda_\eta) \in \widetilde{V}_h \times \Lambda_\eta$ such that:

$$\forall v_h \in \widetilde{V}_h, \quad a_{\widetilde{\Omega}}(\widetilde{u}_h, v_h) + b(v_h, \lambda_\eta) = (\widetilde{f}, v_h)_{2,\widetilde{\Omega}} \tag{3.17a}$$

$$\forall \mu_\eta \in \Lambda_\eta, \qquad b(\widetilde{u}_h, \mu_\eta) = 0. \tag{3.17b}$$

Using the similar notations as in Section 3.2.2, $\widetilde{A} = \left(a_{\widetilde{\Omega}}(v_i, v_j)\right)_{1 \le i,j \le N}$ and $B = \left(b(v_i, \mu_j)\right)_{\substack{1 \le i \le N \\ 1 \le j \le M}}$, the approximation problem (3.17) can be written in the form of a linear system:

$$\text{find } (\widetilde{U}, L) \in \mathbb{R}^N \times \mathbb{R}^M \ : \ \begin{pmatrix} \widetilde{A} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} \widetilde{U} \\ L \end{pmatrix} = \begin{pmatrix} \widetilde{F} \\ 0 \end{pmatrix}. \tag{3.18}$$

We will now present the convergence properties of the FDM for Dirichlet problems elaborated by Glowinski.

3.2.5 Convergence analysis and numerical experiments

In order to ensure the existence and uniqueness of the solution to the approximated problem, we need to check a discrete inf-sup condition. This condition is not usually straight-forward to show and its proof depends on the discretisation spaces chosen. Under the hypothesis of uniform boundedness of η with respect to h, Girault and Glowinski [64] prove the following discrete version of the inf-sup condition. This discrete condition was proven only for V_h chosen as the space of piecewise linear polynomials on $\widetilde{\Omega}$ and μ_η the space of piecewise constant functions on γ.

Proposition 3.4 (Discrete Inf-sup condition)**.** *Let us define V_h as the space of piecewise linear polynomials for a regular triangular discretisation $\widetilde{\mathcal{T}}_h$ of $\widetilde{\Omega}$ and Λ_η, the space of piecewise constant polynomials on the partition $\mathcal{L}_\eta$, as described in* (3.17)*. Let us further*

assume that bilinear form b satisfies the 'inf-sup' condition (3.11). *If we choose η, the mesh size for the interface γ such that $3h \leq \eta$, then there exists $k_0 > 0$ such that:*

$$\inf_{\mu_H \in \Lambda_\eta} \sup_{v_h \in \widetilde{V}_h} \frac{b(v_h, \mu_\eta)}{\|v_h\|_{H_0^1(\widetilde{\Omega})} \|\mu_\eta\|_{H^{-1/2}(\gamma)}} \geq k_0. \tag{3.19}$$

Proof. This proof is to be credited to Girault and Glowinski [64]. It was primarily carried out in the case of γ being a polygonal line, then extended to curvilinear lines. □

Remark 3.3. We have noticed, through our own numerical experiments, that the condition $\eta \geq 3h$ can be relaxed slightly, although we do not know the lower bound for the number κ in $\eta \geq \kappa h$ such that the discrete Inf-sup condition is still verified. This could be the object of future work.

Brezzi and Fortin [21] provide bounds for the error estimate using the discrete 'inf-sup' condition (3.19).

Theorem 3.5 (Bound for the error estimate)**.** *Let $(\widetilde{u}, \lambda) \in H_0^1(\widetilde{\Omega}) \times H^{-1/2}(\gamma)$ and $(\widetilde{u}_h, \lambda_\eta) \in \widetilde{V}_h \times \Lambda_\eta$ be, respectively, the solutions of* (3.10) *and* (3.17). *Then there exists a constant $c > 0$ independent of h and η such that:*

$$\|\widetilde{u} - \widetilde{u}_h\|_{H_0^1(\widetilde{\Omega})} + \|\lambda - \lambda_\eta\|_{H^{-1/2}(\gamma)} \leq c \left(\inf_{v_h \in \widetilde{V}_h} \|\widetilde{u} - v_h\|_{H_0^1(\Omega)} + \inf_{\mu_\eta \in \Lambda_\eta} \|\lambda - \mu_\eta\|_{H^{-1/2}(\gamma)} \right). \tag{3.20}$$

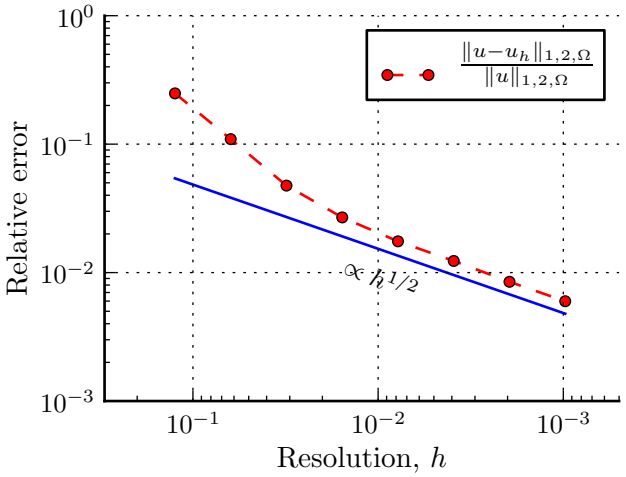
The quality of the approximation of $\widetilde{u}$ and λ by $\widetilde{u}_h$ and Λ_η greatly depends on their regularity. As the normal flux of $\widetilde{u}$ is *a priori* not continuous across the interface γ (unless $\lambda \equiv 0$ on γ) then $\nabla \widetilde{u} \in (H^{1/2-\varepsilon}(\widetilde{\Omega}))^2$, for some ε positive, which implies that u belongs to $H^{3/2-\varepsilon}(\widetilde{\Omega})$. The conclusion is that in this case, the error of approximation of $\widetilde{u}$ by $\widetilde{u}_h$ can only be expected to shrink as fast as $h^{\frac{1}{2}-\varepsilon}$. In mathematical terms, this translates into:

$$\exists K > 0 \text{ such that } \|\widetilde{u} - \widetilde{u}_h\|_{H_0^1(\widetilde{\Omega})} \leq K h^{\frac{1}{2}-\varepsilon} \|\widetilde{u}\|_{H_0^1(\widetilde{\Omega})}. \tag{3.21}$$

Moreover, in the worst case, λ will lie in $L^2(\gamma)$, so there exists $K > 0$ such that:

$$\|\lambda - \lambda_\eta\|_{-\frac{1}{2},2,\gamma} \leq K \eta^{1/2} \|\lambda\|_{\frac{1}{2},2,\gamma}. \tag{3.22}$$

These two inequalities allow us to derive in the next proposition a bound for the convergence rate of the numerical approximation $\widetilde{u}_h$ towards $\widetilde{u}$ when h shrinks to 0.



(3.3a) Convergence plot

h	$\frac{\|u-u_h\|_{1,2}}{\|u\|_{1,2}}$	Rate
1/8	0.245583	–
1/16	0.110577	1.19
1/32	0.052350	1.19
1/64	0.030087	0.81
1/128	0.019978	0.61
1/256	0.014503	0.50
1/512	0.009976	0.54
1/1024	0.006960	0.50

(3.3b) Relative error values

Proposition 3.6. *Assuming that* $\widetilde{u} \in H^{3/2-\varepsilon}(\widetilde{\Omega})$ *and* $\lambda \in L^2(\gamma)$*, then a bound for the error estimate for the approximation* (3.10) *of the fictitious problem* (3.9) *reads :*

$$\|\widetilde{u} - \widetilde{u}_h\|_{H_0^1(\widetilde{\Omega})} + \|\lambda - \lambda_\eta\|_{H^{-1/2}(\gamma)} \leq \widetilde{c}\left(h^{1/2-\varepsilon}\|\widetilde{u}\|_{\frac{3}{2}-\varepsilon,2,\widetilde{\Omega}} + \eta^{1/2}\|\lambda\|_{\frac{1}{2},2,\gamma}\right). \tag{3.23}$$

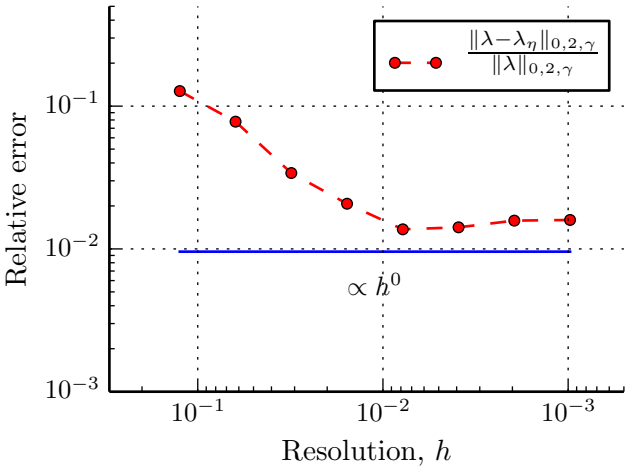
Example. Let $\Omega = (0,1)^2 \setminus \{\boldsymbol{x} \ : \ \|\boldsymbol{x} - (0.4, 0.6)^\top\| \leq 0.2\}$. We define, for $(x, y) \in \Omega$

$$u(x, y) = xy(1-x)(1-y)((x-0.4)^2 + (y-0.6)^2 - 0.2^2), \tag{3.24}$$

a polynomial function positive on Ω such that $u = 0$ on $\partial\Omega$.
We wish to solve Problem (3.3) using the FDM, with known solution u. We extend $f = -\nu\Delta u + \alpha u$ by 0 on $\widetilde{\Omega}$, and set $\nu = \alpha = 1$ for this numerical experiment. Since the solution $\widetilde{u}$ is chosen as 0 in ω, then $\lambda = D\nabla u \cdot \boldsymbol{n}$.

For different values of the resolution h, we solve the fictitious domain formulation of the approximation problem (3.17) and compare the numerical solution u_h with the exact known solution u, with respect to the $H^1(\Omega)$-norm. As h shrinks, we choose $\eta = |\gamma| / \lfloor \frac{|\gamma|}{4h} \rfloor$, so that $\eta \approx 4h$ and $|\gamma| / \eta$ is an integer. Results are given in Table 3.3b and Figure 3.3a plots those values against the line representing the theoretical convergence behaviour.

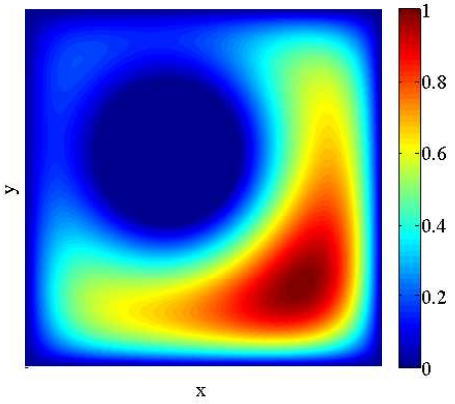
We observe a good agreement between our numerical experiments and the theoretical bounds for the error estimate since we see an asymptotic convergence rate of order $\mathcal{O}(h^{1/2})$ for the $H^1(\Omega)$-norm of the error as predicted by [64]. We see that in our example, the $L^2(\gamma)$-norm of $\lambda - \lambda_\eta$ does not converge to zero as observed in Figure 3.3c. The reason is that λ is not regular enough (only $L^2(\gamma)$).



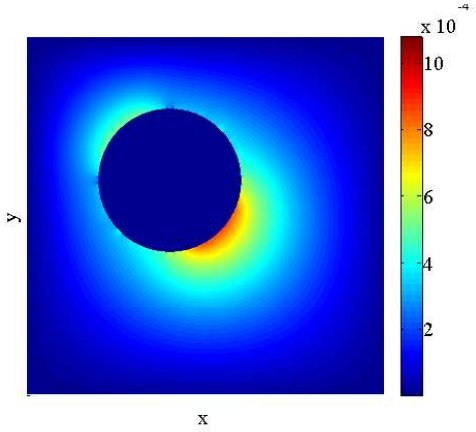
(3.3c) Convergence plot

h	$\frac{\|\lambda-\lambda_\eta\|_{0,2,\gamma}}{\|\lambda\|_{0,2,\gamma}}$	Rate
1/8	0.127448	–
1/16	0.077842	0.71
1/32	0.033998	1.19
1/64	0.020721	0.71
1/128	0.013712	0.59
1/256	0.014165	-0.04
1/512	0.015779	-0.15
1/1024	0.015947	-0.01

(3.3d) Relative error values



(3.3e) Approximation for $h = \frac{1}{1024}$



(3.3f) Distribution of the error for $h = \frac{1}{1024}$

Figure 3.3: Convergence results obtained for the FDM in H^1 for Dirichlet elliptic problems. In Figures 3.3a and 3.3c, we present convergence results in the $H^1(\Omega)$-norm and the $L^2(\gamma)$-norm. The tables on the right-hand side of each plot provided the numerical figures plotted on the corresponding left-hand side plot. In Figures 3.3e-3.3f, we plotted the numerical approximation when $h = 1/1024$ and the plot of the distribution of the absolute error for this particular resolution.

h	Time for solving (3.18)	Time for computing B
1/8	0.047501 s	0.613519 s
1/16	0.000729 s	0.057041 s
1/32	0.006671 s	0.108482 s
1/64	0.014570 s	0.214192 s
1/128	0.094214 s	0.421473 s
1/256	0.387436 s	0.861185 s
1/512	2.212996 s	1.669986 s
1/1024	8.728376 s	3.211874 s

Table 3.1: Times needed for solving the linear system (3.18) (first column) and for forming the fictitious domain operator (second column). The algorithm used to compute B is described in Chapter 4. We can observe that the computing cost for building B scales linearly with the resolution.

3.3 A fictitious domain method for Neumann elliptic problems

The FDM used in previous section is only able to address the problem of the numerical approximation of the solution of Dirichlet elliptic partial differential equations. The main reason is that a Lagrange multiplier would not appear naturally. A way around this issue is to try to transform the initial Neumann-type problem in a Dirichlet one by a change of unknown function. A first choice consists in seeing that a Neumann boundary condition for the initial unknown function is a Dirichlet boundary condition for its gradient. The following method makes full use of this idea.

A multitude of methods have been developed to apply FDM to Neumann elliptic problems, see [79, 39, 61] for examples. The method about to be presented here is close to that introduced in [79]: a mixed weak formulation will be derived from the initial PDE, where $H^{\mathrm{div}}(\widetilde{\Omega})$ will appear naturally. This formulation is an hybrid version of Raviart's in [115].

Let Ω be the domain of $\mathbb{R}^2$ drawn in Figure 3.1a. We are given the following Neumann problem : given $f \in L^2(\Omega)$ and $g \in H^{-1/2}(\gamma)$, find some function u satisfying:

$$\begin{aligned} -\nu\Delta u + \alpha u &= f \quad \text{in } \Omega, \\ -\nu\nabla u \cdot \boldsymbol{n} &= g \quad \text{on } \gamma. \end{aligned} \tag{3.25}$$

3.3.1 Fictitious domain formulation of (3.25)

Let $\widetilde{\Omega}$ be the extension of Ω as shown in Figure 3.2a. Let us define a fictitious problem on the extended domain. Let us first denote by u_i the solution of (3.25) and secondly u_e the solution of a complementary Neumann problem on $\omega := \widetilde{\Omega} \setminus \overline{\Omega}$:

$$\begin{aligned} -\nu\Delta u_e + \alpha u_e &= \widetilde{f} \quad \text{in } \omega \\ -\nabla u_e \cdot \boldsymbol{n} + \nabla u_i \cdot \boldsymbol{n} &= 0 \quad \text{on } \gamma, \end{aligned} \tag{3.26}$$

where $\boldsymbol{n}$ is the local outwards normal to γ. The functions u_i and u_e do not match at the interface γ, however, their normal derivative does since by definition $-\nabla u_e \cdot \boldsymbol{n} + \nabla u_i \cdot \boldsymbol{n} = 0$. Let us define $\boldsymbol{\sigma}$ on $\widetilde{\Omega}$ by :

$$\boldsymbol{\sigma} = \begin{cases} -\nu\nabla u_i & \text{in } \Omega, \\ -\nu\nabla u_e & \text{in } \omega. \end{cases} \tag{3.27}$$

At the interface γ, by definition of u_e, the normal component of the flux $\boldsymbol{\sigma}$ is continuous so $\boldsymbol{\sigma}$ belongs to $H^{\mathrm{div}}(\widetilde{\Omega})$. However, the jump of the tangential component of $\boldsymbol{\sigma}$, $[\boldsymbol{\sigma} \cdot \boldsymbol{t}]$, *a priori* is not zero, so one cannot expect $\boldsymbol{\sigma}$ to belong to $(H^1(\widetilde{\Omega}))^2$. Now, let $u \in L^2(\widetilde{\Omega})$ such that

$$u = \begin{cases} u_i & \text{in } \Omega, \\ u_e & \text{in } \omega. \end{cases} \tag{3.28}$$

Then $\boldsymbol{\sigma}$ and u satisfy the following system

$$\begin{aligned} \boldsymbol{\sigma} + \nu\nabla u &= 0 \quad \text{in } \widetilde{\Omega}, \\ \nabla \cdot \boldsymbol{\sigma} + \alpha u &= \widetilde{f} \quad \text{in } \widetilde{\Omega}, \\ \boldsymbol{\sigma} \cdot \boldsymbol{n} &= g \quad \text{on } \gamma. \end{aligned} \tag{3.29}$$

This system of PDEs constitutes the fictitious domain problem that we propose to solve, instead of (3.25).

Proposition 3.7. *A fictitious domain variational formulation of* (3.25) *reads :*
find $(\boldsymbol{\sigma}, u, \lambda) \in H^{div}(\widetilde{\Omega}) \times L^2(\widetilde{\Omega}) \times H^{1/2}(\gamma)$ *such that*

$$\forall \boldsymbol{\tau} \in H^{div}(\widetilde{\Omega}), \quad (\nu^{-1}\boldsymbol{\sigma}, \boldsymbol{\tau}) - (\nabla \cdot \boldsymbol{\tau}, u) + \langle \boldsymbol{\tau} \cdot \boldsymbol{n}, \lambda \rangle = 0 \tag{3.30a}$$

$$\forall v \in L^2(\widetilde{\Omega}), \quad -(\nabla \cdot \boldsymbol{\sigma}, v) - (\alpha u, v) = (-f, v) \tag{3.30b}$$

$$\forall \mu \in H^{1/2}(\gamma), \quad \langle \boldsymbol{\sigma} \cdot \boldsymbol{n}, \mu \rangle = \langle g, \mu \rangle, \tag{3.30c}$$

where $\langle \cdot, \cdot \rangle$ *denotes the dual pairing in* $H^{-\frac{1}{2}}(\gamma) \times H^{\frac{1}{2}}(\gamma)$.

The Lagrange multiplier λ represents the jump of u across the interface γ and is *a priori* an element of $H^{1/2}(\gamma)$, since u restricted to one of the subdomains Ω or ω would be a H^1 function. If u was continuous across the interface, then λ would be zero.

Proof. This is a direct consequence of the application of the Divergence Theorem. □

Remark 3.4. In the case $\alpha > 0$, one could eliminate the unknown u from the system (3.30) by choosing $v = \nabla \cdot \boldsymbol{\tau}$ with $\boldsymbol{\tau} \in H^{\mathrm{div}}(\widetilde{\Omega})$ and adding the first two equations together. One would then obtain a system very similar in the form to (3.10) :

$$\forall \boldsymbol{\tau} \in H^{\mathrm{div}}(\widetilde{\Omega}), \quad (\alpha\nu^{-1}\boldsymbol{\sigma}, \boldsymbol{\tau}) + (\nabla \cdot \boldsymbol{\sigma}, \nabla \cdot \boldsymbol{\tau}) + \langle \boldsymbol{\tau} \cdot \boldsymbol{n}, \lambda \rangle = (f, \nabla \cdot \boldsymbol{\tau}) \tag{3.31a}$$

$$\forall \mu \in H^{1/2}(\gamma), \qquad \langle \boldsymbol{\sigma} \cdot \boldsymbol{n}, \mu \rangle = \langle g, \mu \rangle. \tag{3.31b}$$

Proposition 3.8 (Inf-sup condition)**.** *Let $b(\cdot,\cdot)$ be the bilinear map defined on $H^{div}(\widetilde{\Omega}) \times H^{\frac{1}{2}}(\gamma)$ such that $b(\boldsymbol{\tau}, \mu) = \langle \boldsymbol{\tau} \cdot \boldsymbol{n}, \mu \rangle$. Then $\exists C > 0$ such that*

$$\inf_{\mu \in H^{1/2}(\gamma)} \sup_{\boldsymbol{\tau} \in H^{div}(\widetilde{\Omega})} \frac{|b(\boldsymbol{\tau}, \mu|)}{\|\boldsymbol{\tau}\|_{H^{div}(\widetilde{\Omega})} \|\mu\|_{\frac{1}{2},2,\gamma}} \geq C. \tag{3.32}$$

Proof. This result was proved by Joly et al. [79]. The proof requires a proof of the existence in $H^{-1/2}(\gamma)$ of the trace of a H^{div} function. □

Theorem 3.9 (Well-posedness of the fictitious variational problem)**.** *Problem* (3.30) *admits a unique solution in $H^{div}(\widetilde{\Omega}) \times L^2(\widetilde{\Omega}) \times H^{1/2}(\gamma)$. Moreover, the pair $(\boldsymbol{\sigma}, u)$ restricted to Ω satisfies* (3.25)*.*

Proof. This is a consequence of the Babuska-Brezzi Theorem 2.22. □

Remark 3.5 (Regularity of the solution)**.** If the data and the interface are smooth enough, one can expect more regularity on the solution and one can hope at least that $\boldsymbol{\sigma} \in H^{\mathrm{div},\frac{1}{2}-\varepsilon}(\widetilde{\Omega})$ and $u \in H^{\frac{1}{2}-\varepsilon}(\widetilde{\Omega})$, for some $\varepsilon > 0$, where the Sobolev space $H^{\mathrm{div},s}$, with s a positive real number is naturally defined as the set:

$$H^{\mathrm{div},s}(\Omega) = \left\{ \boldsymbol{\tau} \in H^s(\Omega)^2 \text{ such that } \nabla \cdot \boldsymbol{\tau} \in H^s(\Omega) \right\}. \tag{3.33}$$

3.3.2 Finite-element approximation of (3.30)

Let $\mathcal{T}_h$, $h > 0$, be a regular triangulation of $\widetilde{\Omega}$ such as shown in Figure 3.2b and $\mathcal{L}_\eta$, $\eta > 0$, a set of segments of equal length η partitioning γ. Again, a priori, the two mesh sizes h and η are decoupled. The following approximation spaces are introduced:

$$W_h = \left\{ \boldsymbol{\tau} \in H^{\mathrm{div}}(\widetilde{\Omega}) \ : \ \boldsymbol{\tau}|_K \in \mathcal{RT}_0(K), \forall K \in \mathcal{T}_h \right\}, \tag{3.34}$$

$$M_h = \left\{ v \in L^2(\widetilde{\Omega}) \ : \ v|_K \in \mathcal{P}_0(K), \forall K \in \mathcal{T}_h \right\}, \tag{3.35}$$

$$\Lambda_\eta = \left\{ \mu \in H^{1/2}(\gamma) \ : \ \mu|_S \in \mathcal{P}_0(S), \forall S \in \mathcal{L}_\eta \right\}. \tag{3.36}$$

We denote $N_W = \dim W_h$, $N_M = \dim M_h$ and $N_\Lambda = \dim \Lambda_\eta$.

The approximation space Λ_η has been chosen as the space of piecewise constant polynomials thus conforming with $H^{1/2}(\gamma)$. Thus, the finite-element approximation problem of (3.30) reads: find $(\boldsymbol{\sigma}_h, u_h, \lambda_\eta) \in W_h \times M_h \times \Lambda_\eta$ such that

$$\forall \boldsymbol{\tau}_h \in W_h, \quad (\nu^{-1}\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) - (\nabla \cdot \boldsymbol{\tau}_h, u_h) + \langle \boldsymbol{\tau}_h \cdot \boldsymbol{n}, \lambda_\eta \rangle = 0 \tag{3.37a}$$

$$\forall v_h \in M_h, \quad - (\nabla \cdot \boldsymbol{\sigma}_h, v_h) - (\alpha u_h, v_h) = (-f, v_h) \tag{3.37b}$$

$$\forall \mu_\eta \in \Lambda_\eta, \quad \langle \boldsymbol{\sigma}_h \cdot \boldsymbol{n}, \mu_\eta \rangle = \langle g, \mu_\eta \rangle. \tag{3.37c}$$

One can also write this symmetric linear problem in a matrix form, as follows:

$$\text{find } (\Sigma_h, U_h, \Lambda_\eta) \in \mathbb{R}^{N_W} \times \mathbb{R}^{N_M} \times \mathbb{R}^{N_\Lambda} \ : \ \begin{pmatrix} A_{\boldsymbol{\sigma}} & D & B \\ D^\top & -A_u & 0 \\ B^\top & 0 & 0 \end{pmatrix} \begin{pmatrix} \Sigma_h \\ U_h \\ \Lambda_\eta \end{pmatrix} = \begin{pmatrix} 0 \\ F \\ G \end{pmatrix}, \tag{3.38}$$

where $A_{\boldsymbol{\sigma}} = ((\nu^{-1}\boldsymbol{\tau}_i, \boldsymbol{\tau}_j))_{1 \le i,j \le N_W}$, $D = (-(\nabla \cdot \boldsymbol{\tau}_i, v_j))_{\substack{1 \le i \le N_W \\ 1 \le j \le N_M}}$, $B = ((\boldsymbol{\tau}_i, \mu_j))_{\substack{1 \le i \le N_W \\ 1 \le j \le N_\Lambda}}$ and $A_u = ((\alpha v_i, v_j))_{1 \le i,j \le N_M}$.

3.3.3 Convergence analysis and numerical experiments

In order to prove the convergence of the discrete approximation towards the true solution for decreasing values of h, the following inf-sup condition provides a sufficient condition.

Proposition 3.10 (Discrete inf-sup condition)**.** *Let us assume that b satisfies the inf-sup condition* (3.32)*. Assume $\exists a > 0$ such that $ah \leq \eta$, then there exists $C > 0$ independent of h and η such that :*

$$\inf_{\mu_\eta \in \Lambda_\eta} \sup_{\boldsymbol{\tau}_h \in W_h} \frac{|\langle \boldsymbol{\tau}_h, \mu_\eta \rangle|}{\|\boldsymbol{\tau}_h\|_{H^{div}(\widetilde{\Omega})} \|\mu\|_{\frac{1}{2},2,\gamma}} \geq C. \tag{3.39}$$

Proof. This is the subject of Joly's paper [79]. □

Again, Brezzi's theory yields the following bound for the error estimate.

Proposition 3.11 (Bound for the error estimate)**.** *Assuming that the bilinear form b satisfies the discrete inf-sup condition, we obtain the following bound for the error estimate:*

$$\begin{aligned} &\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{H^{div}(\widetilde{\Omega})} + \|u - u_h\|_{2,\widetilde{\Omega}} + \|\lambda - \lambda_\eta\|_{\frac{1}{2},2,\gamma} \\ &\quad \leq \widetilde{C} \left(\inf_{\boldsymbol{\tau}_h \in W_h} \|\boldsymbol{\sigma} - \boldsymbol{\tau}_h\|_{H^{div}(\widetilde{\Omega})} + \inf_{v_h \in M_h} \|u - v_h\|_{2,\widetilde{\Omega}} + \inf_{\mu_\eta \in \Lambda_\eta} \|\lambda - \mu_\eta\|_{\frac{1}{2},2,\gamma} \right). \end{aligned} \tag{3.40}$$

Each term in the right-hand side of the previous inequality can be explicitly bounded by a function of the mesh size h or η. The order of approximation depends on the regularity of the true solution and the approximation properties of the finite elements employed in the discretised problem. In our case, we have, for some $\varepsilon > 0$,

$$\inf_{\boldsymbol{\tau}_h \in W_h} \|\boldsymbol{\sigma} - \boldsymbol{\tau}_h\|_{H^{\mathrm{div}}(\widetilde{\Omega})} \leq \hat{C} h^{1/2-\varepsilon} \|\boldsymbol{\sigma}\|_{H^{\mathrm{div},1/2-\varepsilon}(\widetilde{\Omega})}, \tag{3.41a}$$

$$\inf_{v_h \in M_h} \|u - v_h\|_{2,\widetilde{\Omega}} \leq \hat{C} h^{1/2-\varepsilon} \|u\|_{\frac{1}{2}-\varepsilon,2,\widetilde{\Omega}}, \tag{3.41b}$$

$$\inf_{\mu_\eta \in \Lambda_\eta} \|\lambda - \mu_\eta\|_{\frac{1}{2},2,\gamma} \leq \hat{C} \eta^{\frac{1}{2}-\varepsilon} \|\lambda\|_{1-\varepsilon,2,\gamma}, \tag{3.41c}$$

where the $\hat{C}$ are constants independent of h or η.

Theorem 3.12 (Bound for the error estimate and convergence rate)**.** *Let us suppose that Equation* (3.40) *is satisfied. Assuming that there exists $\varepsilon > 0$ such that $\boldsymbol{\sigma} \in H^{div,1/2-\varepsilon}(\widetilde{\Omega})$, $u \in H^{\frac{1}{2}-\varepsilon}(\widetilde{\Omega})$ and $\lambda \in H^{1/2-\varepsilon}(\gamma)$ so that Equations* (3.41) *are satisfied, then we obtain a lower bound for the convergence rate of the FDM for Neumann problems:*

$$\begin{aligned} &\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{H^{div}(\widetilde{\Omega})} + \|u - u_h\|_{2,\widetilde{\Omega}} + \|\lambda - \lambda_\eta\|_{\frac{1}{2},2,\gamma} \\ &\qquad \leq \widetilde{C} \left(h^{1/2-\varepsilon} \left(\|\boldsymbol{\sigma}\|_{H^{div,\frac{1}{2}-\varepsilon}(\widetilde{\Omega})} + \|u\|_{\frac{1}{2}-\varepsilon,2,\widetilde{\Omega}} \right) + \eta^{\frac{1}{2}-\varepsilon} \|\lambda\|_{1-\varepsilon,\gamma} \right). \end{aligned} \tag{3.42}$$

Example. Let $\Omega = (0,1)^2 \setminus \mathrm{B}((0.4,0.6)^\top, 0.2)$ as defined earlier on. We define, for $(x,y) \in \Omega$,

$$u(x,y) = \left(\frac{1}{3}((x-0.4)^2 + (y-0.6)^2) - 0.2^2\right)(x+y-1) + \frac{2}{15}, \tag{3.43}$$

The diffusive flux $\boldsymbol{\sigma} = -D\nabla u$ reads:

$$\boldsymbol{\sigma}(x,y) = -\frac{2}{3}D(x+y-1)\begin{pmatrix} x-0.4 \\ y-0.6 \end{pmatrix} - D\left(\frac{1}{3}((x-0.4)^2 + (y-0.6)^2) - 0.2^2\right)\begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{3.44}$$

This function has the property of having zero normal component on γ since in polar coordinates centered at $(0.4,0.6)^\top$, we have:

$$\frac{\partial u}{\partial r} = (r^2 - R^2)(\cos\theta + \sin\theta). \tag{3.45}$$

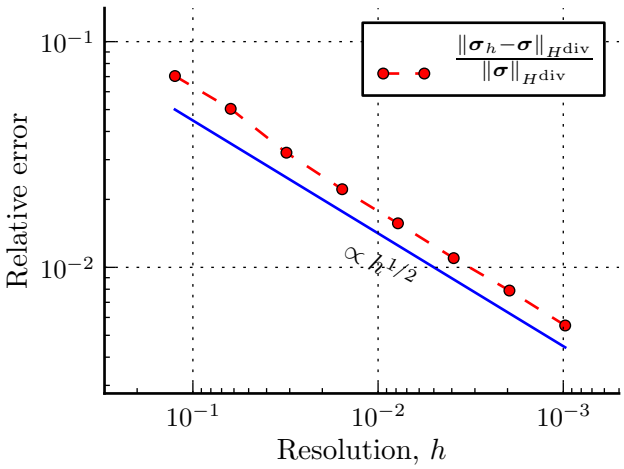
The function u is the unique solution of the elliptic problem (3.25) with:

$$f(x,y) = -\frac{8}{3}D(x+y-1) + \frac{\alpha}{3}\left(\left((x-0.4)^2 + (y-0.6)^2 - 3\times 0.2^2\right)(x+y-1) + \frac{2}{15}\right).$$

and $g \equiv 0$. Moreover, since the flux through γ is zero, we can expect the fictitious extension of u to $\widetilde{\Omega}$ be also zero in $\widetilde{\Omega} \setminus \overline{\Omega}$ if we choose to extend f by 0 as well. As a consequence, we should obtain $\lambda \equiv -u|_\gamma$.

We wish to solve this problem using the FDM presented in this section. For different values of the resolution h, we solve the fictitious domain formulation of the approximation problem (3.37) and compare the numerical solution $(\boldsymbol{\sigma}_h, u_h)$ with the exact known solution $(\boldsymbol{\sigma}, u)$, with respect to the H^{div} and L^2 norms. We choose for this convergence study $\nu = 1$ and $\alpha = 1$. The right-hand side f is extended by zero in $\widetilde{\Omega} \setminus \overline{\Omega}$. As in the Dirichlet case, as we shrink h, we choose $\eta = |\gamma| / \lfloor \frac{|\gamma|}{4h} \rfloor$, so that $\eta \approx 4h$ and $|\gamma|/\eta$ is an integer.

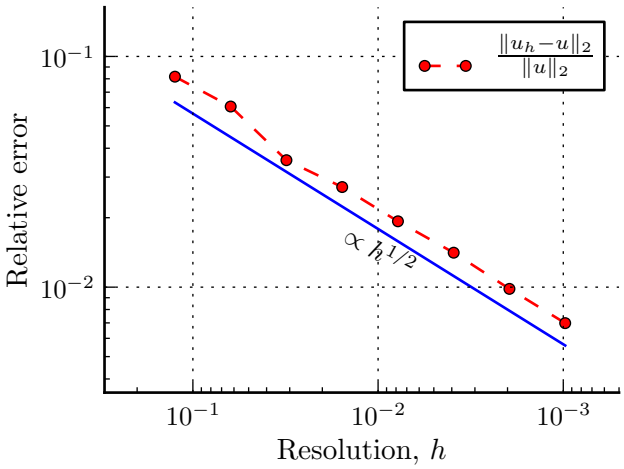
Results are given in Table 3.4b and Figure 3.4a plots those values against the line representing the theoretical convergence behaviour. We observe a good agreement between our numerical experiments and the theoretical bounds for the error estimate since we see an asymptotic convergence rate of order $\mathcal{O}(h^{1/2})$ for $\boldsymbol{\sigma}_h$ towards $\boldsymbol{\sigma}$ (Figure 3.4a) and u_h towards u (Figure 3.4c). The rate of convergence of λ_η towards λ (Figure 3.4e) agrees with the theory, since it was measured in the in the L^2-norm.



(3.4a) Convergence plot

h	$\frac{\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{H^{\mathrm{div}}(\Omega)}}{\|\boldsymbol{\sigma}\|_{H^{\mathrm{div}}(\Omega)}}$	Rate
1/8	0.070405	–
1/16	0.050389	0.48
1/32	0.032211	0.64
1/64	0.022188	0.54
1/128	0.015670	0.50
1/256	0.010986	0.51
1/512	0.007892	0.48
1/1024	0.005526	0.51

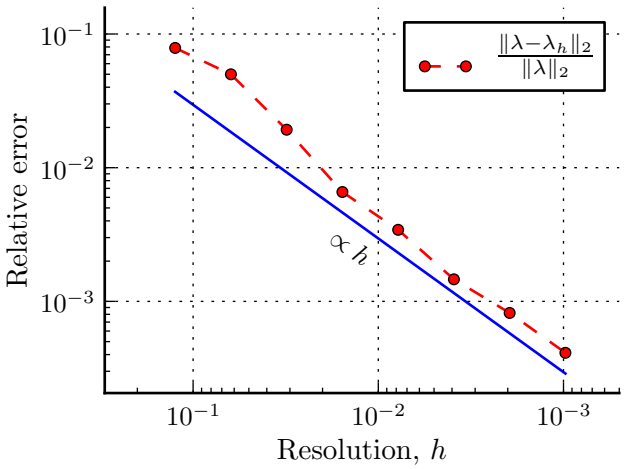
(3.4b) Relative error values



(3.4c) Convergence plot

h	$\frac{\|u-u_h\|_{2,\Omega}}{\|u\|_{2,\Omega}}$	Rate
1/8	0.081709	–
1/16	0.060656	0.43
1/32	0.035505	0.77
1/64	0.027156	0.39
1/128	0.019292	0.49
1/256	0.014104	0.45
1/512	0.009829	0.52
1/1024	0.006983	0.49

(3.4d) Relative error values

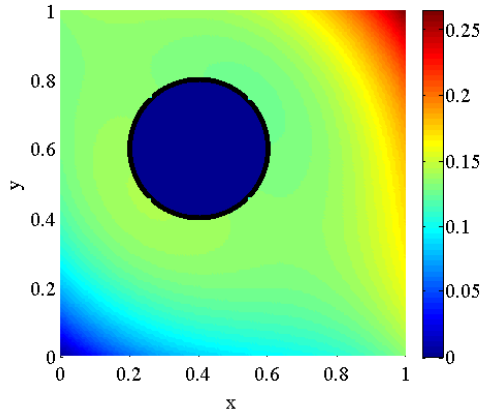


(3.4e) Convergence plot

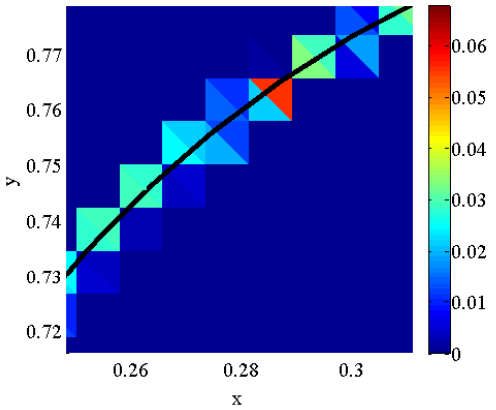
h	$\frac{\|\lambda-\Lambda_\eta\|_{2,\Omega}}{\|\lambda\|_{2,\Omega}}$	Rate
1/8	0.07866172	–
1/16	0.04995506	0.66
1/32	0.01923094	1.37
1/64	0.00657768	1.54
1/128	0.00342942	0.94
1/256	0.00146207	1.23
1/512	0.00081955	0.84
1/1024	0.00041234	0.99

(3.4f) Relative error values

Figure 3.4: Convergence results obtained for the FDM in H^{div} for Neumann elliptic problems. In Figures 3.4a, 3.4c and 3.4e, we present convergence results for the approximations $\boldsymbol{\sigma}_h$, u_h and λ_η in the $H^{\text{div}}(\Omega)$-norm, $L^2(\Omega)$-norm and $L^2(\gamma)$-norm, respectively. The tables on the right-hand side of each plot provided the numerical figures plotted on the corresponding left-hand side plot.



(3.5a) Plot of u_h for $h = 1/128$.



(3.5b) Plot of the error $|u - u_h|$; zoom on the interface.

Figure 3.5: Figure 3.5a is the plot of the approximation u_h of u for $h = 1/128$. On the right-hand side, Figure 3.5b is a plot of the absolute error $|u - u_h|$ zoomed on a tiny portion of $\widetilde{\Omega}$ near the interface γ represented by a black line.

Remark 3.6. in Figure 3.5b, we have plotted the absolute error $|u - u_h|$ for $h = 1/128$ (see the corresponding plot of the approximation u_h in Figure 3.5a). We observe that on all triangles intersected by the interface γ (represented by a black line), the approximation u_h does not converge to u. The reason for this is that those triangles are neither entirely included of Ω nor $\widetilde{\Omega} \setminus \Omega$ so one cannot make sense of the restriction of $u_h|_T$ to Ω or $\widetilde{\Omega} \setminus \Omega$ for any triangle T crossed by γ, as an approximation to $u|_\Omega$ or $u|_{\widetilde{\Omega}\setminus\Omega}$. In short, $u_h|_T$, for any triangle T intersecting with γ, does not converge to $u|_{T\cap\Omega}$ nor to $u|_{T\cap(\widetilde{\Omega}\setminus\Omega)}$.

From this observation, we conclude that when using the FDM for Neumann problems, we will only dispose of an approximation of $u_i = u|_\Omega$ on all the triangles that are strictly subset of Ω. Fortunately, as h shrinks, the size of the layer of triangles intersected by γ shrinks as well by the same proportion. So in the limit of $h \to 0$, we obtain a valid approximation on the whole of Ω.

3.4 Adaptation of the fictitious domain method in H^{div} to Robin problems

3.4.1 Motivation

Robin boundary condition for reaction-diffusion-equation are widely used in the literature for modelling the diffusion of chemicals interacting with the membrane of the cell. Examples of interactions include binding with membrane receptors [133], signal transduction mechanisms [34].

Since Robin boundary conditions are a natural way of modelling interactions between external signals and membrane of cells, it seems critical to us to be able to implement Robin boundary conditions in a FDM, in order to reach our goal of simulating in the same framework the processes that lead a cell to respond to external chemical signals.

The available literature in this field [66] could not provide us with any acceptable solution, so we adapted the previously discussed method to fit the resolution of Robin boundary type elliptic problems.

3.4.2 Definition of the problem

Let us define the mathematical problem. Let $\Omega \subset \mathbb{R}^2$ be a Lipschitz domain, such as shown in Figure 3.1a. Using the same notations as in previous sections:

$$\begin{aligned} -\nu\Delta u_i + \alpha u_i &= f \quad \text{in } \Omega \\ -\nu\nabla u_i \cdot \boldsymbol{n} + \beta u_i &= g \quad \text{on } \gamma, \end{aligned} \tag{3.46}$$

where $f \in L^2(\Omega)$, $g \in H^{3/2}(\gamma)$, $\nu > 0$, and $\alpha, \beta \geq 0$. This elliptic problem is well-posed in $H^1(\Omega)$ and one can show that the solution u belongs to $H^2(\Omega)$. Let us introduce, as in the previous section, the flux variable :

$$\boldsymbol{\sigma}_i = -\nu\nabla u_i \quad \text{on } \Omega. \tag{3.47}$$

The Robin boundary condition becomes : $\boldsymbol{\sigma}_i \cdot \boldsymbol{n} + \beta u_i = g$. One can write a fictitious formulation of (3.46) by extending u and $\boldsymbol{\sigma}$ to $\widetilde{\Omega}$. To simplify notations, the extended functions will be denoted in the same way as the original functions.

3.4.3 A fictitious domain formulation

Let u_e be the solution of the following problem posed in $\omega = \widetilde{\Omega}\backslash\Omega$:

$$\begin{aligned} -\nu\Delta u_e + \alpha u_e &= 0 \quad \text{in } \omega \\ -\nu\nabla u_e \cdot \boldsymbol{n} + \nu\nabla u_i \cdot \boldsymbol{n} &= 0 \quad \text{on } \gamma. \end{aligned} \tag{3.48}$$

Then defining u and $\boldsymbol{\sigma}$ as the concatenation of u_i and u_e on the one hand, $\boldsymbol{\sigma}_i$ and $\boldsymbol{\sigma_e}$ on the other hand, the boundary condition in (3.46) is equivalent to:

$$\boldsymbol{\sigma} \cdot \boldsymbol{n} + \beta\, u|_\Omega = g. \tag{3.49}$$

Indeed, although $\boldsymbol{\sigma} \cdot \boldsymbol{n}$ is continuous across the interface γ by its very definition, *a priori* u is not. Hence the two functions u and $\boldsymbol{\sigma}$ satisfy the following PDE system :

$$\begin{aligned} \boldsymbol{\sigma} + \nu\nabla u &= 0 \quad \text{in } \widetilde{\Omega}, \\ \nabla \cdot \boldsymbol{\sigma} + \alpha u &= \widetilde{f} \quad \text{in } \widetilde{\Omega}, \\ \boldsymbol{\sigma} \cdot \boldsymbol{n} + \beta\, u|_\Omega &= g \quad \text{on } \gamma, \end{aligned} \tag{3.50}$$

where $\widetilde{f}$ represents a L^2-extension of f. The problem (3.50) is the fictitious problem we propose to solve instead of (3.46).

Proposition 3.13. *A fictitious domain variational formulation of* (3.46) *reads :*
find $(\boldsymbol{\sigma}, u, \lambda) \in H^{div}(\widetilde{\Omega}) \times M \times H^{1/2}(\gamma)$ *such that*

$$\forall \boldsymbol{\tau} \in H^{div}(\widetilde{\Omega}), \quad (\nu^{-1}\boldsymbol{\sigma}, \boldsymbol{\tau}) - (\nabla \cdot \boldsymbol{\tau}, u) + \langle \boldsymbol{\tau} \cdot \boldsymbol{n}, \lambda \rangle = 0 \tag{3.51a}$$

$$\forall v \in L^2(\widetilde{\Omega}), \quad - (\nabla \cdot \boldsymbol{\sigma}, v) - (\alpha u, v) = (-f, v) \tag{3.51b}$$

$$\forall \mu \in H^{1/2}(\gamma), \quad (\beta\, u|_{\Omega}, \mu) + \langle \boldsymbol{\sigma} \cdot \boldsymbol{n}, \mu \rangle = \langle g, \mu \rangle, \tag{3.51c}$$

where $\langle \cdot, \cdot \rangle$ *denotes the dual pairing in* $H^{-\frac{1}{2}}(\gamma) \times H^{\frac{1}{2}}(\gamma)$ *and*

$$M = \left\{ u \in L^2(\widetilde{\Omega}) \ : \ u|_{\Omega} \in H^1(\Omega) \right\}. \tag{3.52}$$

Unfortunately, we have not been able to prove the well-posedness of this problem, mainly because of its lack of symmetry and the presence of a restriction operator in Equation (3.51c), which is non standard in this type of PDE problem. Nevertheless, we will proceed in the next section by providing a finite-element formulation of this problem.

3.4.4 Finite-element approximation

Let $\mathcal{T}^h$ with $h > 0$ be a regular and uniform triangulation of domain $\widetilde{\Omega}$ and $\mathcal{L}_\eta$, $\eta > 0$, a set of segments S forming a partition of γ. Using the same approximation spaces introduced in section 3.3.2, we obtain:
find $(\boldsymbol{\sigma}_h, u_h, \lambda_\eta) \in W_h \times M_h \times \Lambda_\eta$ such that:

$$\forall \boldsymbol{\tau}_h \in W_h, \quad (\nu^{-1}\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) - (\nabla \cdot \boldsymbol{\tau}_h, u_h) + \langle \boldsymbol{\tau}_h \cdot \boldsymbol{n}, \lambda_\eta \rangle = 0 \tag{3.53a}$$

$$\forall v_h \in M_h, \quad - (\nabla \cdot \boldsymbol{\sigma}_h, v_h) - (\alpha u_h, v_h) = (-f, v_h) \tag{3.53b}$$

$$\forall \mu_\eta \in \Lambda_\eta, \quad (\beta\, u_h|_{\Omega}, \mu_\eta) + \langle \boldsymbol{\sigma}_h \cdot \boldsymbol{n}, \mu_\eta \rangle = \langle g, \mu_\eta \rangle. \tag{3.53c}$$

In this system, one can observe that the computation of $(\beta\, v_h|_{\Omega}, \mu_\eta)$ will be problematic, because on each triangle intersecting with the interface γ, the degree of freedom associated to this triangle is shared by the two domains. This implies that, for any triangle T such that $T \cap \gamma \neq \emptyset$, for any $v_h \in M_h$ with support on T, $v_h|_{T \cap \Omega}$ will not be an approximation of $u|_{T \cap \Omega}$.

In order to overcome this issue, let us approximate the Robin boundary condition by making a strong assumption : the solution u of (3.46) is C^1 in a neighbourhood of the

boundary γ. Then we are allowed to expand u in Taylor series near the interface, such that, for $\varepsilon > 0$, $\forall \boldsymbol{x} \in \gamma$,

$$\begin{aligned} u(\boldsymbol{x}) &= u(\boldsymbol{x} - \varepsilon \boldsymbol{n}(\boldsymbol{x})) + \nu^{-1} \boldsymbol{\sigma}(\boldsymbol{x} - \varepsilon \boldsymbol{n}(\boldsymbol{x})) \cdot \varepsilon \boldsymbol{n}(\boldsymbol{x}) + \mathcal{O}(\varepsilon^2) \\ &= u(\boldsymbol{x} - \varepsilon \boldsymbol{n}(\boldsymbol{x})) + \varepsilon \nu^{-1} \boldsymbol{\sigma}(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x}) + \mathcal{O}(\varepsilon^2), \end{aligned} \tag{3.54}$$

this is obtained by only keeping first order terms in ε and using the relationship (3.47). Then choosing $\varepsilon = \varepsilon(h)$ such that $\lim_{h \to 0} \varepsilon(h) = 0$, Equation (3.53c) becomes

$$\forall \mu_\eta \in \Lambda_\eta, \left(\frac{\beta}{1 + \varepsilon(h)\, \beta \nu^{-1}} u_h(\cdot - \varepsilon(h)\, \boldsymbol{n}(\cdot)), \mu_\eta \right) + \langle \boldsymbol{\sigma}_h \cdot \boldsymbol{n}, \mu_\eta \rangle = \left\langle \frac{g}{1 + \varepsilon(h)\, \beta \nu^{-1}}, \mu_\eta \right\rangle. \tag{3.55}$$

Figure 3.6 shows the triangles crossed by $\gamma - \varepsilon(h)\, \boldsymbol{n}(\gamma)$, shaded of blue. Here $\varepsilon(h)$ has been chosen so that the degrees of freedom associated to the triangles passed through by $\gamma - \varepsilon(h)\, \boldsymbol{n}(\gamma)$ have their entire support in Ω. The smallest value of ε satisfying this requirement is $\varepsilon(h) = \sqrt{2}h$, using a purely geometrical argument: each triangle has a $\sqrt{2}h$ long hypotenuse.

However, we will see in the next numerical experiment section that the convergence rate of the numerical scheme is not affected by the choice of $\varepsilon(h)$, as long as it is larger than $\sqrt{2}h$.

Using the same notations as in the previous section, the fictitious approximation problem can be written in a matrix form:

$$\text{find } (\Sigma_h, U_h, \Lambda_\eta) \in \mathbb{R}^{N_W} \times \mathbb{R}^{N_M} \times \mathbb{R}^{N_\Lambda} \ : \ \begin{pmatrix} A_{\boldsymbol{\sigma}} & D & B \\ D^\top & -A_u & 0 \\ B^\top & P_h & 0 \end{pmatrix} \begin{pmatrix} \Sigma_h \\ U_h \\ \Lambda_\eta \end{pmatrix} = \begin{pmatrix} 0 \\ F \\ G \end{pmatrix}, \tag{3.56}$$

where P_h is the matrix operator $\left(\left(\frac{\beta}{1+\varepsilon(h)\, \beta \nu^{-1}} v_i(\cdot - h\, \boldsymbol{n}(\cdot)), \mu_j \right) \right)_{\substack{1 \le i \le N_M \\ 1 \le j \le N_\Lambda}}$. We observe that the presence of this new projection operator breaks the symmetry of the linear system.

3.4.5 Numerical experiments

3.4.5.1 Estimation of the convergence rate

As we are not able to prove both the well-posedness of the fictitious problem and the convergence of the numerical approximation problem, we will content ourselves with providing numerical evidence for the convergence of the scheme.

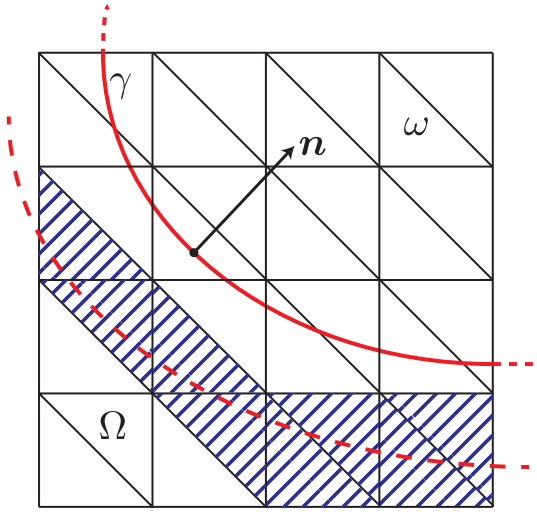


Figure 3.6: Intersection of the interface with the triangulation $\mathcal{T}^h$: the shaded triangles on this picture are all the triangles part of Ω, having a non-empty intersection with the curve $\gamma - \varepsilon(h)\,\boldsymbol{n}(\gamma)$, represented by the red dashed line on the picture.

Let Ω be defined as in previous sections, and

$$u(x,y) = x(1-x)y(1-y)\exp(-((x-0.4)^2 + (y-0.6)^2)). \tag{3.57}$$

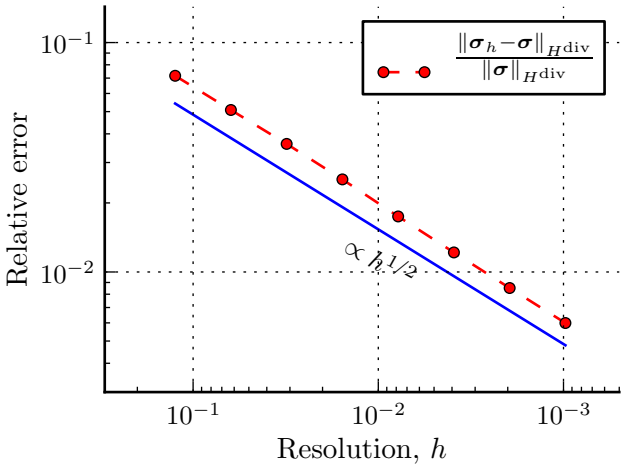
We introduce a regular triangulation of the fictitious domain $\widetilde{\Omega}$ with mesh size h and a partition of γ with mesh size η, composed of segments of equal length.

For different mesh sizes h and $\eta \approx 2h$ and choosing $\varepsilon(h) = \sqrt{2}h$, we solve the fictitious problem (3.53) and compute the error of approximation using the true solution u. The results are summarised and displayed in Figure 3.7 and Figure 3.8.

We observe that the rate of convergence of the approximation towards the solution has not been upset by the projection operator of $u|_\Omega$ on γ. This rate of convergence is estimated to be roughly $h^{1/2}$, hence comparable to the FDM for Neumann problems.

3.4.5.2 Influence of the choice of $\varepsilon(h)$

We wish to assess whether the choice of a particular $\varepsilon(h)$ upsets the convergence of the FDM for Robin boundary conditions. Ideally, the choice of such an $\varepsilon(h)$ should not



(3.7a) Convergence plot

h	$\frac{\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{H^{\mathrm{div}}(\Omega)}}{\|\boldsymbol{\sigma}\|_{H^{\mathrm{div}}(\Omega)}}$
1/8	0.0785013
1/16	0.0556918
1/32	0.0396258
1/64	0.0277496
1/128	0.0191329
1/256	0.0133314
1/512	0.0093277
1/1024	0.0065645

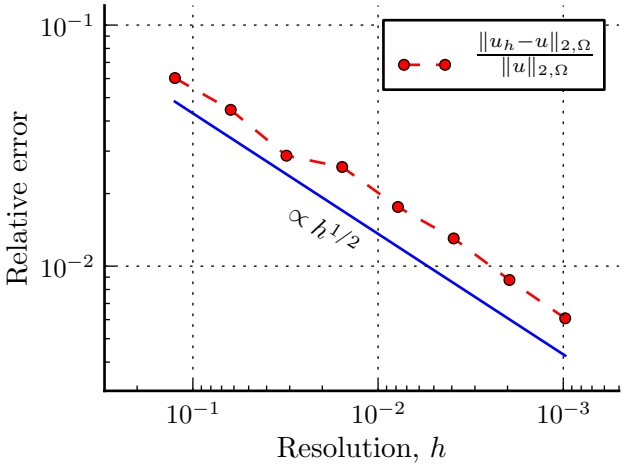
(3.7b) Relative error values

Figure 3.7: Convergence results in the H^{div}-norm obtained for the FDM for Robin boundary elliptic problems.

influence the convergence of this numerical scheme; however, we can *a priori* expect the convergence to be altered in two cases:

- when ε is too small: in this case, the curve $\gamma - \varepsilon(h)\boldsymbol{n}(\gamma)$ will cross parts of the triangulation that are not exclusively part of Ω. A lower bound for this epsilon is then estimated to be $\sqrt{2}h$. This bound does not depend on the actual solution of the original equation.
- When ε is too large, the first order approximation (3.54) of u on the boundary will lose of its accuracy.

Hence we can expect that the optimal convergence rate will be observed only for a range of values of ε. In order to numerically provide evidence that the convergence of the scheme does not depend too much on the choice of $\varepsilon(h)$, we have measured the mean convergence rate of the numerical scheme for different values of $\frac{\varepsilon}{h}$. We plotted in Figure 3.9 the value of the experimental convergence rate for different values of $\frac{\varepsilon}{h}$. We observe that below a critical value of $\frac{\varepsilon}{h}$, around 0.5, the numerical approximation fails to converge, as expected.



(3.8a) Convergence plot

h	$\frac{\|u-u_h\|_{2,\Omega}}{\|u\|_{2,\Omega}}$
1/8	0.0602787
1/16	0.0445033
1/32	0.0287175
1/64	0.0257835
1/128	0.0175949
1/256	0.0130296
1/512	0.0087594
1/1024	0.0060762

(3.8b) Relative error values

Figure 3.8: Convergence results in the L^2-norm obtained for the FDM for Robin boundary elliptic problems.

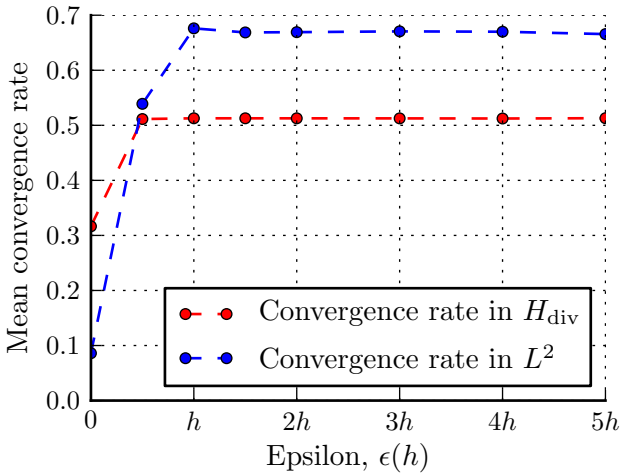


Figure 3.9: Study of the influence of the choice of $\varepsilon(h)$ on the convergence rate of the numerical scheme.

3.5 Summary

In this chapter, we have presented three FDMs for solving Dirichlet, Neumann and Robin boundary elliptic problems. All three present a common feature: the boundary condition at the immersed interface is imposed through the use of a Lagrange multiplier that is defined only on the interface. The operator projecting the volume unknown on the interface carries the geometry of the problem instead of the triangulation as in classical finite element methods.

Although FDMs for pure Dirichlet or Neumann problems are readily available in the literature, an acceptable solution for fictitiously solving Robin boundary problems, very useful when simulating cell-level biological phenomena, did not exist so we developed and proposed a robust numerical scheme fulfiling this need. It lies on the same discretisation scheme as the Neumann problem and seems to possess the same convergence properties, as demonstrated by our numerical experiments. Unfortunately, we were not able to prove the theoretical convergence of this numerical scheme.

In the next chapter, the details of the actual implementation of the FDM will be dealt with and we will provide algorithms and `Matlab` code specific to the implementation. We will then study how efficiently solve the linear systems arising from the implementation of the FDM and the different solutions proposed in the literature.

Chapter 4

Implementation of the FDM and preconditioning

4.1 Introduction

A wide range of textbooks discuss the computer implementation of the FEM [55, 72, 6], and many implementations of the FEM are available in the internet in different programming languages.

However, since this dissertation uses a special branch of the FEM, the FDM, it seems prudent to provide a basic implementation of the algorithms for constructing a rectangular mesh, as well as the finite element operators supported by the triangulation. This implementation is not original in any way, however it will lay sound bases for the algorithms we will discuss further in this chapter, dealing with the efficient construction of the matrix operators specific to the FDM.

This chapter also addresses the issues relating to the computation of the numerical solution of the linear systems derived from the FDM. We will first introduce two of the most widely used iterative algorithms for solving symmetric and non symmetric linear systems, namely Minres and Gmres. Then we will discuss possible preconditioners for the different Schur complements of the linear systems we have derived in Chapter 3. Following Glowinski's method for deriving a preconditioner for the linear system arising from the FDM for Dirichlet problems in Section 3.3, we derive a preconditioner for the Schur complement of the linear system arising from the Neumann and Robin FDM.

4.2 Implementation of the finite element matrix operators

In this section, we provide `Matlab` routines for computing the discretised finite element operators supported on the triangulation of the extended domain (usually a rectangle). There is no particular difficulty to report on those routines and numerous books, for instance [72], provide ample information about the implementation of classical finite elements. The implementation we provide here is of course not optimal in terms of computational cost or numerical accuracy, but we desire to provide a comprehensive set of routines implementing a fictitious domain FEM.

4.2.1 Generation of a uniform triangulation and related maps

The first routine that we present, called `build_uniform_triangulation`, is shown in Listing 4.1 and builds a uniform triangulation of a rectangular domain of size $L_1 \times L_2$, with origin at 0. The triangulation is built by creating $n \times m$ squares of size h and partitioning each of them into two triangles according to their descending diagonal. This routine requires three parameters as inputs: `r` and `s`, the number of nodes in each direction, and `L=[L1 L2]` the dimensions of the rectangle. Most of the implementation is classical, however the computation of the edge-related data structures in this routine were inspired by [8] and improved to take advantage of `Matlab`'s fast operations on vectors. The methods return a structure `triangulation` containing some useful information described below:

`nodes:` this is a matrix of size $2 \times$ `rs` containing the coordinates of each node of the triangulation. The way the nodes are arranged in this matrix induces their ordering.

`bndry_nodes:` this is a vector of size $2 \times$ (`r` + `s` − 2) containing the number of all the nodes on the boundary of the rectangle.

`tri2nodes:` this is a $3 \times$ (`r` − 1)(`s` − 1) matrix that lists the number of the three vertices of each triangle.

`edges:` this matrix contains the information needed to locate the edges of the triangulation and the orientation of their normal. In each row, say row `i`, the first two

```
function triangulation = build_uniform_triangulation(...
    r, ...                                % Nb of points in the x direction
    s, ...                                % Nb of points in the y direction
    L)                                    % Dimensions of the rectangle
m = r-1; n = s-1;                         % Number of intervals in x & y.
n_p = r*s;n_t = 2*m*n;n_e=3*m*n+(m+n);% Number of nodes, triangles, edges
tn = zeros(n_t,3);                        % Initialisation of tri->nodes map
te = zeros(n_t,3);                        % Initialisation of tri->edges map
x = linspace(0,L(1),r);                   % Def of points on x axis
y = linspace(0,L(2),s);                   % Def of points on y axis
p = [reshape(x'*ones(1,s),1,n_p) ; ...    % Def of nodes coordinates vector
     reshape(ones(r,1)*y, 1,n_p)];
e = unique([find(mod(p(1,:),L(1))==0)...% Def of bdnry nodes
            find(mod(p(2,:),L(2))==0)]);

mt2v = @(m) m(:);                         % Transforms a matrix in vector
cart2lin = @(k,l,n) k + (l-1)*n;          % Func transforming cartesian coor-
a = -1;                                   % dinates in linear coordinates.
for j=1:n
    for i=1:m                             % Building triangles to nodes map :
        a = a+2;                          % numbering starts at right angles
        tn(a,:) = cart2lin([i i+1 i]   , [j j j+1]   , r);
        tn(a+1,:)=cart2lin([i+1 i i+1], [j+1 j+1 j], r);
    end
end
node2tri = sparse(mt2v(tn'),...           % This matrix is such that if pair
  mt2v(tn(:,[2, 3, 1])'),....             % (i,j) are an oriented edge of T
  mt2v(ones(3,1)*(1:n_t)),....            % then nodes2tri(i,j) is nonzero &
  n_p,n_p);                               % is equal to the index of T.
                                          % Construction of nodes->edge map:
[I,J] = find(triu(node2tri+node2tri'));   % ne(i,j)=e if edge e = (i,j)
et = [I J diag(node2tri(I,J)) ...         % Creation of edge map:
          diag(node2tri(J,I))];           % - col 1,2 : start and end nodes
to_swap = find(et(:,3)==0);               % - col 3,4 : triangles defining
et(to_swap,:) = et(to_swap,[2 1 4 3]);    % orientation of edge: 3 -> 4.
                                          % if col 4 is zero -> bndry edge.
ne=sparse(I,J,1:n_e,n_p,n_p);ne=ne+ne';   % Creation of numbering of edges
[I,J,E]=find(ne);                         % Creation of triangle->edge map :
id_j = [0; find(diff(J)) ; length(J)];    %
for j=1:n_t                               % We travel all triangles indices
    st_n = tn(j,[2 3 1]);                 % find all starting nodes of edges
    nd_n = tn(j,[3 1 2]);                 % find all end nodes of edges
    for i=1:3                             % efficiently search through ne to
        col = (id_j(nd_n(i))+1):(id_j(nd_n(i)+1)); % find corresponding
        te(j,i) = E(col(ismembc(I(col),st_n(i)))); % edges.
    end
end
triangulation = struct(...                % Store all the maps in a structure
    'nodes', p', ...
    'edges', full(et), ...
    'bndry_nodes', e', ...
    'bndry_edges', find(et(:,4)==0), ...
    'tri2nodes',tn , ...
    'tri2edges',te , ...
    'nonodes', n_p ,...
    'notriangles',n_t,...
    'noedges', n_e, ...
    'h', L./[m n]);
```

Listing 4.1: Routine building a uniform triangulation on a rectangle.

columns give the starting and the end node of the edge `i`, and in the last two columns, the two triangles T_+ and T_- sharing the edge `i`. The normal of this edge is directed outwards T_+ (hence inwards T_-).

The construction of the map `edges` is worth being carefully explained: we firstly form the graph `node2tri` that associates, for each pair of node (`i`, `j`) either 0 if (`i`, `j`) is not an (oriented) edge of a triangle, or the index of the triangle hosting this edge. We notice that if `node2tri`(`i`, `j`) is non zero, then it implies that `node2tri`(`j`, `i`) is also non zero or `i` and `j` are boundary nodes. As a result, each non oriented edge is in general counted twice (except boundary edges).

The second step consists in forming the matrix `node2tri` + `node2tri`$'$: in this graph, the two oriented edges sharing the same starting and end nodes (in any order) are identified in a single non oriented edge. We can now select those edges, performed on line `32` of Listing 4.1 by extracting the upper triangular entries of the non oriented edge graph, the indices of which are `I` and `J`. The numbering of the edges is then induced by this extraction.

The last step of this construction follows straightforwardly. The array `I` contains the starting node of each edge, `J` the end node, and for each edge of number `e`, the call `node2tri`(`I`(`e`), `J`(`e`)) returns T_+ and the call `node2tri`(`J`(`e`), `I`(`e`)) returns T_-.

`tri2edges:` This matrix lists on each row, corresponding to a triangle, the three non oriented edges forming this triangle. It is built by going through each triangle with nodes (i, j, k) and by taking each consecutive pair of nodes (i, j), (j, k) and (k, i), we find the corresponding edge number in the replica of `node2tri` + `node2tri`$'$ in which the values in the graph correspond to the edge numbering.

Let us note that the two matrices `tri2nodes` and `tri2edges` can be used as connectivity matrices when dealing with nodal first order Lagrange or edge-located Raviart-Thomas degrees of freedom. When implementing second order triangular Lagrange elements, a way to construct the connectivity matrix associated to the rectangular triangulation is to concatenate `tri2nodes` and `tri2edges`, since the second order Lagrange element possesses three more nodes compared to the linear triangular element, all of which are located on each edge of each triangle.

4.2.2 Building finite element matrices

The second step in the implementation of the FEM is building the matrices resulting from the original finite element formulation. The basic routine consists in travelling through all the elements and for each pair of degrees of freedom with support on a given triangular element, computing the contribution of a pair of degrees of freedom towards the global finite element operator on the current triangle. We obtain a local matrix operator that is later integrated in the global finite element operator, spanning all the elements of the triangulation. This integration is performed using the connectivity matrix, which relates the local numbering of the degrees of freedom in a given triangle with their global numbering. This is due to the fact that, generally, a degree of freedom has support on several triangular elements.

Due to the regularity of the triangulation, it is then possible to compute only one local matrix. The numbering of the degrees of freedom in the connectivity matrix becomes crucial since one must ensure that the local to global numbering is done in the same manner for each triangle. It is then left to the user to choose between hard coding the local matrix for each operator, as such matrices are of small size for low order finite elements, or using a numerical integration quadrature rule to compute each of its coefficients. We engage in the latter option for reasons of flexibility.

4.2.3 Quadrature rule on the reference triangle

Let us recall that a Gaussian quadrature rule on a domain D is a set of n weights w_i and evaluation points $x_i \in D$ such that the quantity $\sum_{i=1}^{n} w_i f(x_i)$ approximates the value of the integral $\int_D f(x)\,\mathrm{d}x$, where f is some integrable function on D. The weights and the evaluation points are chosen to minimise the error of approximation [27]. The order of a quadrature rule p is defined as the integer such that the error of approximation $E(f) = \left|\int_D f(x)\,\mathrm{d}x - \sum_{i=1}^{n} w_i f(x_i)\right|$ can be bounded by $D^n f$. Hence, by definition, a quadrature rule of order p integrates exactly polynomials of degree less than p.

The method presented in Listing 4.2 allows the exact integration of any polynomial of degree equal or less than 5, on the reference triangle with vertices $(0,0)^\top, (1,0)^\top$, and $(0,1)^\top$. It is a 7-point Gaussian quadrature rule, with weights and evaluation

```
function I = gauss_quadrature_triangle(f)
% Numerical integration of a function f on the reference TRIANGLE.
% This quadrature is a 7-point Gauss rule, O(h^6) accurate.
% f is a function of the three barycentric coordinates.
% Define the weights :
w    = [0.225000000000000 ...
        0.125939180544827 0.125939180544827 0.125939180544827  ...
        0.132394152788506 0.132394152788506 0.132394152788506]';
% Define the quadrature points using barycentric coordinates :
zeta = [0.333333333333333 0.333333333333333 0.333333333333333; ...
        0.797426985353087 0.101286507323456 0.101286507323456; ...
        0.101286507323456 0.797426985353087 0.101286507323456; ...
        0.101286507323456 0.101286507323456 0.797426985353087; ...
        0.059715871789770 0.470142064105115 0.470142064105115; ...
        0.470142064105115 0.059715871789770 0.470142064105115; ...
        0.470142064105115 0.470142064105115 0.059715871789770];
% Computation of the desired integral :
I = 1/2*sum(w.*f(zeta));
```

Listing 4.2: 5th order Gauss quadrature rule

points derived by Dunavant [41]. The evaluation points are provided in terms of their barycentric coordinates, that are well suited for triangular finite elements.

4.2.4 Building matrix operators

The routine in Listing 4.3 creates a finite element matrix operator on a regular triangulation. This method takes in three inputs: an integrand function `fun` = `fun`(`i`, `j`, `x`), the connectivity matrix linking the local numbering of the degrees of freedom (abridged as `dof` in the code) to their global numbering in the triangulation, and finally `a` = `a`(`localK`, `i`), an optional function that can transform the local matrix according to the location of the current element `i`. It first computes the coefficients of the local matrix by integrating the function `fun` on the reference triangle for all pairs (i, j) of indices in the local numbering of degrees of freedom in the reference triangle. Then the routine parses every triangle and allocates to the global nodes forming the currently visited triangle the contribution contained in the local matrix `a(localK,i)`.

For instance, let us assume we wish to create the matrix operator K such that:

$$K_{ij} = \int_{\Omega} w(x)\phi_i(x)\phi_j(x)\,\mathrm{d}x, \tag{4.1}$$

where the function ϕ_k is the basis function associated to the degree of freedom k. Then we would define: `fun`$(\mathtt{i}, \mathtt{j}, [\lambda_1\ \lambda_2\ \lambda_3]) = h^2\phi_i(x(\lambda_1, \lambda_2, \lambda_3))\,\phi_j(x(\lambda_1, \lambda_2, \lambda_3))$ with

```
function A = create_fe_operator(...
    fun, ...                                % Integrand generator
    T2dof, ...                              % Connectivity matrix
    a)                                      % Optional argument
if (nargin==2), a = @(k,i) k;  end          % Default value for a.
if isempty(a),  a = @(k,i) k;  end
m2v = @(x) x(:);                            % Transforms a matrix in a vector
p = size(T2dof,2);                          % Define the number of dof/element
n_elements = size(T2dof,1);                 % Define the number of elements
n_dof = max(max(T2dof));                    % Define the total number of dof
idi = zeros(p*p,n_elements); idj = idi;     % Initialise index for the matrix
values = idi;                               % Initialise values vector
loc_A = zeros(p);                           % Initialise the local matrix
for i=1:p                                   % Compute each coefficient of the
    for j=1:p                               % local matrix.
        loc_A(i,j)=gauss_quadrature_triangle(@(x) fun(i,j,x));
    end
end
for i=1:n_elements                          % Main loop
    dof = T2dof(i,:)'*ones(1,p);            % Retrieve dof of element i.
    idi(:,i)=m2v(dof');                     % Feed in the global indices
    idj(:,i)=m2v(dof );                     %
    m = a(loc_A,i);                         % Optional treatment on local_K
    values(:,i) = m(:);                     % Feed local matrix in global matrx
end
A = sparse(idi(:),idj(:),values(:),n_dof,n_dof); % Construct global matrix
end
```

Listing 4.3: Routine building global operator provided the element matrix and the connectivity matrix.

$x(\lambda_1, \lambda_2, \lambda_3) = \sum_{i=1}^{3} \lambda_i x_i$ and $\texttt{a(localK, i)} = \frac{1}{|T_i|} \int_{T_i} w(x)\,\mathrm{d}x \times \texttt{localK}$. We have made the first order approximation:

$$\int_T w(x)\phi_i(x)\phi_j(x)\,\mathrm{d}x = \left(\frac{1}{|T|}\int_T w(x)\,\mathrm{d}x + O(h)\right)\int_T \phi_i(x)\phi_j(x)\,\mathrm{d}x. \tag{4.2}$$

Example. Let $w \equiv 1$, and assume we are using linear Lagrange triangular elements, as described in Chapter 2. Since we are operating in barycentric coordinates, the expression of the basis function is:

```
basis = @(i,x) (i==1)*(1-x(:,2)-x(:,3)) + (i==2)*x(:,2) + (i==3)*x(:,3);
```

where it is assumed that the variable `x` has three components, namely the three barycentric coordinates of point x in the reference triangle. The computation of the mass matrix, namely, $M = \left(\int_\Omega \phi_i\phi_j\,\mathrm{d}x\right)$ is given by:

```
triangulation = build_uniform_triangulation(10,10,[1 1]);
M =create_fe_operator(@(i,j,x) h*h*basis(i,x).*basis(j,x), triangulation.tri2nodes,[]);
```

In the first line, we built a triangulation of the unit square $\Omega = [0,1]^2$ of 10 nodes by 10 nodes. In the second line, we use the connectivity matrix `tri2nodes` to build the mass matrix associated to this mesh using first order Lagrange degrees of freedom. In this particular example, the optional function a has not been specified since $w \equiv 1$.

4.3 Efficient implementation of hybrid operators

One of the main advantages of the FDM over fitted formulations consists in reducing the computational cost of mesh generation. This costly operation requires the fitting of a triangular mesh on the original domain. One of the most popular algorithm, Delaunay's algorithm [40], generates a triangulation from a set of given vertices, that maximises the smallest angle of the three angles of each formed triangle. The computational cost of this routine amounts to $\mathcal{O}(n \log n)$ where n represents the number of mesh vertices.

By contrast, in all the FDMs we have discussed in this dissertation, the triangulation used is regular. However, both the geometry of the domain and the boundary condition at the interface are enforced by the operator associating the Lagrange multiplier space and the volume-supported unknown space. This operator in some sense interpolates the rectangular mesh onto the interface. The computation of the resulting discretised operator must be handled with care and carried out efficiently since not only does it hold essential information about the original PDE, but its computation is non trivial. Moreover, in time dependent problems, the computation of this operator has to be carried out at each time step.

The literature does not provide any detailed solution on the efficient implementation of the FDM and that is why we will tackle this issue in the next section and propose an algorithm that allows the computation of the discretised operator at a reasonable computational cost.

Let us assume that the inner interface γ on which the Lagrange multiplier is supported can be parametrized by $\boldsymbol{f}$, at least Lipschitz, such that:

$$\forall s \in [0,1], \quad \boldsymbol{f}(s) = \begin{pmatrix} x(s) \\ y(s) \end{pmatrix} \in \gamma. \tag{4.3}$$

We further assume that $\boldsymbol{f}$ is periodic, so $\boldsymbol{f}(0) = \boldsymbol{f}(1)$ and bounded since γ is closed and bounded. We suppose that there exists a rectangle such that $\gamma \subset [a,b] \times [c,d]$. Our goal consists in computing integrals such as $\langle v_h, \mu_\eta \rangle$ in (3.10) or (3.30) where v_h is volume-supported and μ_η is an element of the interface approximation space.

4.3.1 Algorithm

We are given a regular triangulation $\mathcal{T}^h$ of the fictitious domain $\widetilde{\Omega}$ assumed to be a rectangle, with mesh size h, and a partition $\mathcal{L}^\eta$ of γ composed of curved segments of equal length η. Furthermore, we suppose that the triangulation results from dividing the rectangle $\widetilde{\Omega}$ in $N \times M$ squares of size h and dividing each of them in two rectangle triangles according to the descending diagonal. Let us consider the following integral that we wish to compute numerically:

$$L(T, S) = \int_\gamma v_T \, \mu_S \, \mathrm{d}s, \tag{4.4}$$

where v_T is a polynomial function with support on a triangle $T \in \mathcal{T}^h$ and μ_S a polynomial defined on γ, with support on $S \in \mathcal{L}^\eta$, a curved segment of γ. Most of the times, the integral $L(T, S)$ will be zero, since the support of each polynomial will not intersect. Hence one has to consider two cases : $T \cap S = \emptyset$ and $T \cap S \neq \emptyset$: in the first case, the integral $L(T, S)$ will be zero and no computational effort should be granted to the calculation of $L(T, S)$ and in the second case, $L(T, S)$ is worth concentrating on.

The problem we would like to consider in this section is not the actual computation of the integral, which is tackled in many numerical analysis books such as [38], but the efficient determination of which integrals such as (4.4) are worthy to be fed to a numerical integration routine. In other words, we propose ourselves to find the couples (T, S) of elements such that (4.4) is *a priori* non zero, *i.e.* when $T \cap S \neq \emptyset$, in order to minimise the number of calls to the numerical integration routine.

As T and S are convex sets, it appears that a critical step of the routine to compute the integral $L(T, S)$, will consist in finding the intersection of the interface γ with the edges of the triangulation. This problem amounts to a root-finding problem. If γ is represented by an explicit parametrisation, one has to solve with respect to s:

$$\begin{aligned} &x(s) = ih, \text{ for } 0 \leq i \leq N, \\ &y(s) = jh, \text{ for } 0 \leq j \leq M, \\ &x(s) + y(s) = kh, \text{ for } 0 \leq k \leq M + N - 1. \end{aligned} \tag{4.5}$$

If the previous equations had no solution, for any i, j or k, this would imply that the curve γ is entirely contained in a single triangle. This would mean that the triangulation

is too coarse and does not allow us to resolve the interface. Rather let us examine the case when the intersection between the mesh and the interface is non empty.

Leaving aside the pathological curve cases, such as $(x, x^3 \sin 1/x)$ for instance, we should account for two cases: either we obtain a finite number of intersections, or an uncountable infinite set of intersections. This latter case is encountered when dealing with (partly) polygonal lines, when one or several edges partly or totally coincide with the edges of the triangulation. We will set this case aside as well because it would only complicate the presentation of our argument. However, at the end of this section, we will come back to this case.

For our purpose, a reasonably behaved curve should possess a radius of curvature much larger than the mesh size h, for instance a circle with radius much larger than h. A similar assumption on the curvature of the immersed boundary was also made by Girault and Glowinski to prove the discrete inf-sup condition (3.19) seen in Chapter 3 for the FDM for elliptic Dirichlet problems.

Let us sort by ascending order of curvilinear abscissa the points of intersection of the edges of the triangulation with the interface. We now have a set of points $s_1, s_2, \ldots, s_l$, $l \geq 1$ such that $\{\boldsymbol{f}([s_i, s_{i+1}))\}_{1 \leq i \leq l}$ partitions γ. Therefore, for triangle T supporting v_T, there exists i_{TS} such $S \cap T = \cup_{i \in i_{TS}} \boldsymbol{f}([s_i, s_{i+1}))$. Hence the initial integral now becomes

$$L(T, S) = \sum_{i \in i_{TS}} \int_{[s_i, s_{i+1})} v_T(x(s), y(s)) \mu_S(s) \|\boldsymbol{f}'(s)\| \, \mathrm{d}s. \tag{4.6}$$

Now let us denote by $p_0, p_1, p_2, \ldots, p_k$, with $k \geq 1$, the abscissa of the end points of each segment $S \in \mathcal{L}^\eta$, such that $S = [p_j, p_{j+1})$, for $0 \leq i < k$. We adopt the convention $p_0 = 0 = p_k$. Hence μ_S has support on $[p_j, p_{j+1})$ and the expression of $L(T, S)$ becomes:

$$L(T, S) = \sum_{i \in i_{TS}} \int_{[s_i, s_{i+1}) \cap [p_j, p_{j+1})} v_T(x(s), y(s)) \mu_S(s) \|\boldsymbol{f}'(s)\| \, \mathrm{d}s. \tag{4.7}$$

The two partitions of γ play a different role:

1. the partition of the intersections $\{s_i\}_{1 \leq i \leq l}$ is strongly linked to the triangulation, so we associate to each s_i, the triangle T_i such that $\boldsymbol{f}(s_i^+) \in \mathring{T}_i$ $(:= T_i \setminus \partial T_i)$;

2. the partition $\{p_j\}_{1 \leq j \leq k}$, by definition, relates to the finite elements on the interface, so we associate to each p_j the element S_j such that $\boldsymbol{f}(p_j^+) \in \mathring{S}_j$.

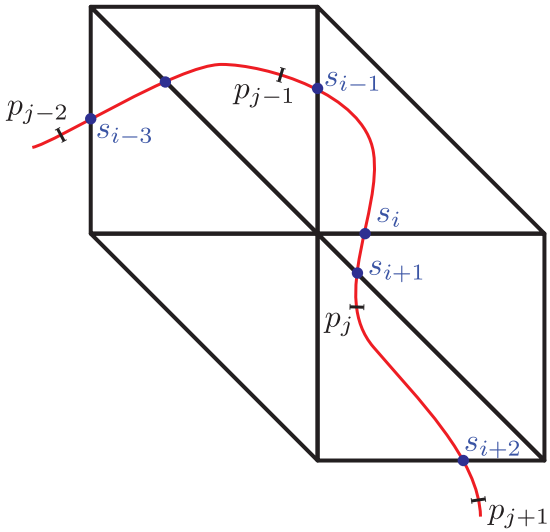


Figure 4.1: Representation of the two partitions of the interface γ: the finite-element partition with points $p_1, p_2, \ldots, p_k$ and the partition $s_1, s_2, \ldots, s_l$ of intersections between the triangulation edges and the interface.

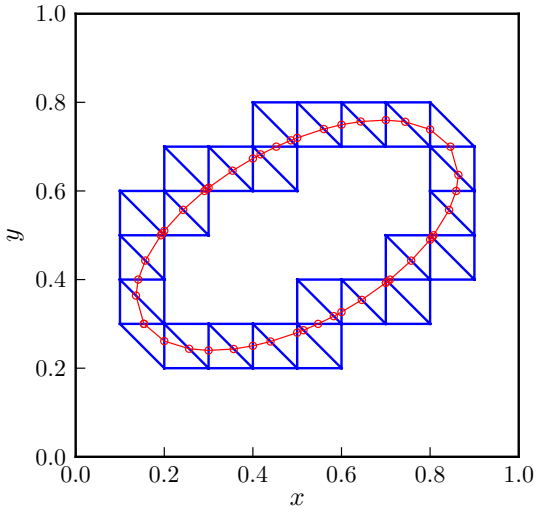
In order to compute the fictitious operator, we visit each segment $[t_i, t_{i+1}]$, where the sequence $\{t_i\}_{1 \leq i \leq k+l}$ results from the merging and the sorting of the two partitions, forming a finer partition of γ. After getting rid of possible duplicates, the sequence $\{t_i\}$ has at most $k + l$ elements. By definition, each of the segments $[t_i, t_{i+1}]$ is entirely contained in a single triangle of $\mathcal{T}^h$ and a single segment of $\mathcal{L}^\eta$, so the only work that remains consists in computing the integral $L(T, S)$ for each pair of basis function (v_T, μ_S) supported on T and S.

The corresponding routine is provided in Listing 4.4. It requires the sequences $\{s_i\}_{1 \leq i \leq l}$ and $\{T_i\}_{1 \leq i \leq l}$ on the one hand, and the sequences $\{p_j\}_{1 \leq j \leq k}$ and $\{S_j\}_{1 \leq j \leq k}$ on the other hand. The routine returns A, the fictitious operator matrix.

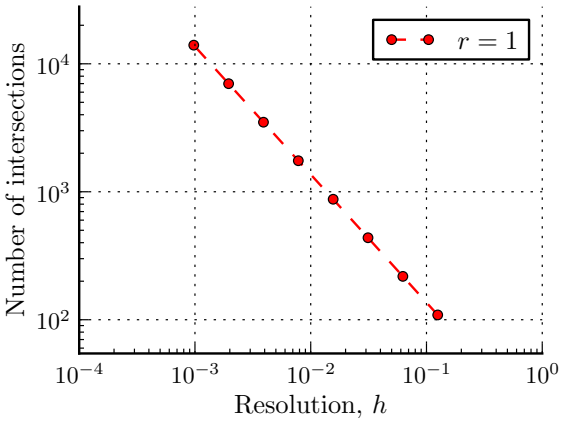
The computational cost of this routine is linear with respect to the number of intersections between the triangulation plus the number of segments, that is to say, $k + l$. The main cost in the loop is the computation of the value of the integral $L(T, S)$. In practice, for common simple curves such as circles or ellipses, the number of intersections grows linearly with the inverse of h, see Figure 4.2b. In the same way, the number of segments grows as the inverse of the interface resolution η, which is generally chosen proportional to h (see Chapter 3 for more details). Overall, the cost of this routine grows linearly

```
function A = create_fict_fe_operator(...
    s_i,...                          % List of intersections abscissa
    T_i,...                          % List of intersecting triangles
    p_j,...                          % List of end points of segments
    S_j,...                          % List of segments
    T2dof,...                        % List of dof with support on a given T
    S2dof,...                        % List of dof with support on a given S
    value)                           % Handle generating an integrand from a
                                     % given v_T and mu_S.
n = length(s_i); m = length(p_j);    % Give names to the length of each list
n_T = size(T2dof,2);                 % Number of dof per triangle
n_S = size(S2dof,2);                 % Number of dof per segment
A = zeros((n+m)*n_T*n_S,3); x=1;     % Initialise the matrix A
s_i(s_i<0)=2*pi+s_i(s_i<0);          % (sparse structure) and a counter x
next = @(i) mod(i,n+m)+1;            % Compute the integer next to i (n+m)
[t_l,idx] = sort([s_i p_j]);         % Merge and sort the two lists s and p
i=n+m; while (idx(i)>n), i=i-1; end % Find 1st triangle thro which s_0 goes
T = T_i(idx(i)); S = S_j(1);         % Initialise current triangle & segment
for p=1:(n+m)                        % Computation of fictitious operator
    a=t_l(p); b=t_l(next(p));        % Set integration bounds
    a=(a>b)*(a-2*pi)+(a<=b)*a;
    if (idx(p)>n), S=S_j(idx(p)-n); % Update current triangle or segment
    else T = T_i(idx(p)); end
    for q=1:n_T                      % Compute the contribution on S and T
        v_T = T2dof(T,q);            % Set current volumic dof as v_T
        for r=1:n_S
            mu_S = S2dof(S,r);       % Set current interface fof as mu_S
            A(x,:) = ...             % Adds A(v_T, mu_S) to the matrix index
            [v_T,mu_S,value(T,S,q,r,a,b)];
            x=x+1;                   % Update counter
        end
    end
end                                  % Final assembly of A (sparse matrix)
A = sparse(A(:,1),A(:,2),A(:,3),max(max(T2dof)),max(S2dof));
end
```

Listing 4.4: Routine building a fictitious domain operator from a list of intersections between the triangulation $\mathcal{T}^h$ and the partition $\mathcal{L}^\eta$.



(4.2a) Plot of the output of the method 4.5: in red circles, we represented the intersections of the triangulation and the interface, and in blue, their associated triangles.



(4.2b) Plot of the number of intersections between a regular triangulation of $\mathbb{R}^2$ of mesh size h with a unit radius circle ($r = 1$) centered randomly in the plane. The results given have been averaged over 10 trials for each value of the resolution.

with the inverse of the resolution.

Remark 4.1. Dealing with polygonal lines seems, in principle, quite straightforward, as computing the intersection of two lines is a trivial exercise. However, one should not also forget to keep track of curved segments or parts of them, coinciding with edges (or parts of them) of the triangulation. The algorithm presented in this section aims at building a map between couples of curvilinear abscissa of γ, (t_i, t_j), and couples of elements T, S such that $\boldsymbol{f}([t_i, t_j]) \subset T \cap S$. Coinciding edges will not fit in to that description and therefore, they must be treated differently. Fortunately, this does not make the resulting integrals $L(T, S)$ harder to compute, as long as those edges are correctly identified.

Remark 4.2. Figure 4.2a displays examples of very short segments $[s_i, s_{i+1}]$. In fact, our algorithm can produce subdivisions where some of the segments can be arbitrarily small. This leads to adding arbitrarily small (in absolute value) entries in the discretised fictitious matrix operator. We have not investigated how this affects the conditioning of the global discretised operator. This could be the object of future works. However, our numerical experiments did not point towards any significant worsening of the conditioning of the linear systems derived from the FDM.

4.3.2 Intersections of a regular triangulation and an ellipse

In this section, we have not yet dealt with the problem of finding the intersections of the interface with the triangulation. When dealing with a general closed curve γ parameterised by $\boldsymbol{f}$, the only solution at our disposal requires the use of a root-finding algorithm. In `Matlab`, some very efficient methods are available, such as `roots` in the `Chebfun` package [126]. To overcome the cost of using this type of solution, it seems preferable to tailor the intersection-finding method to the equation of the curve.

As we will consider only elliptic objects in further chapters, we will provide a routine that returns the intersection of an ellipse with a regular triangulation of $\mathbb{R}^2$, of mesh size h. This problem amounts to solving a series of quadratic equations, for which one obviously can provide closed form solutions.

In cartesian coordinates, for a ellipse $\mathcal{E}$ centered in 0, with axes a and b, aligned with the x-axis, then any point $\boldsymbol{X} = (X, Y)^\top$ on the ellipse satisfies:

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 1. \tag{4.8}$$

Then the line $\alpha X + \beta Y + \delta = 0$ intersects the ellipse $\mathcal{E}$ if and only if $|\delta| \leq \sqrt{\alpha^2 a^2 + \beta^2 b^2}$ and in this case, the line intersects the ellipse in one or two places, whose coordinates are given by:

$$\boldsymbol{X} = \begin{pmatrix} \alpha\delta \dfrac{a^2}{\alpha^2 a^2 + \beta^2 b^2} \mp \beta \dfrac{ab}{\sqrt{\alpha^2 a^2 + \beta^2 b^2}} \sqrt{1 - \dfrac{\delta^2}{\alpha^2 a^2 + \beta^2 b^2}} \\ \beta\delta \dfrac{b^2}{\alpha^2 a^2 + \beta^2 b^2} \pm \alpha \dfrac{ab}{\sqrt{\alpha^2 a^2 + \beta^2 b^2}} \sqrt{1 - \dfrac{\delta^2}{\alpha^2 a^2 + \beta^2 b^2}} \end{pmatrix}. \tag{4.9}$$

Now, an ellipse centred on $\boldsymbol{x}_c$, with dimensions a and b and angle to the x-axis θ, intersects the lines $x = ih$, $y = jh$, and $x + y = kh$ in the points $\boldsymbol{X}$ with coordinates

defined in (4.9) for parameters α, β and δ defined by:

$$x = ih : \begin{cases} \alpha = \cos\theta \\ \beta = -\sin\theta \\ \delta = ih - x_c \end{cases} \tag{4.10a}$$

$$y = jh : \begin{cases} \alpha = \sin\theta \\ \beta = \cos\theta \\ \delta = jh - y_c \end{cases} \tag{4.10b}$$

$$x + y = kh : \begin{cases} \alpha = \cos\theta + \sin\theta \\ \beta = \cos\theta - \sin\theta \\ \delta = kh - (x_c + y_c). \end{cases} \tag{4.10c}$$

This is because in the frame of reference of the ellipse, any point $\boldsymbol{x}$ of the original frame transforms into:

$$\boldsymbol{X} = \begin{pmatrix} a\cos\theta & b\sin\theta \\ -a\sin\theta & b\cos\theta \end{pmatrix} (\boldsymbol{x} - \boldsymbol{x_c}), \tag{4.11}$$

where $\boldsymbol{X}$ denotes the coordinates in the frame of reference of the ellipse. Finally, to obtain the curvilinear abscissa of those points on the interface, we compute:

$$t = \texttt{atan2}(Y/b, X/a) = \arctan\left(\frac{Y/b}{1 + X/a}\right), \tag{4.12}$$

where `atan2` is the four-quadrant atan function.

After we have completed this task, we deduce the numbers of the triangles that intersect the ellipse. Any triangle in the triangulation is included in a small square with bottom left corner with coordinates $h(i,j)^\top$ and top right corner $h(i+1, j+1)^\top$. This triangle has numbering $T = 2(i + jN) + 1$ or $2(i + jN) + 2$. Hence, for $\boldsymbol{x}$ any intersection point, we find T by computing:

$$T = 2\left(\lfloor xN \rfloor + \lfloor yM \rfloor N\right) + 1 \pm 1, \tag{4.13}$$

where $\boldsymbol{x} = (x, y)^\top$ and $\lfloor \cdot \rfloor$ denotes the function 'floor' associating to any real its largest previous interger. To determine whether T should be the lower or the upper triangle, we measure the distance of $\boldsymbol{x}$ to the centre of each of them.

The routine shown in Listing 4.5 implements this methodology and in Figure 4.2a, we have plotted a possible output by showing the points of intersections (circled in red) of an elliptic interface with a regular triangulation of mesh size $h = 1/10$. The triangles (in blue) that have been represented are those intersecting the interface.

```
function [s,T] = find_intersections_ellipse_triangulation(...
    x, ...                                % Centre of the ellipse
    a, b, ...                             % Length of axes of the ellipse
    theta, ...                            % Angle to the x-axis
    h, L)                                 % mesh size and dimensions of grid

eps = 1e-7;
c2l=@(k,l) (k-1)+(l-1)*round(L(1)/h);  % Cartesian -> Linear coordinates
next = @(I) [I(2:end) I(1)];
% We first gather all the intersections with the triangulation:
X = [elpse_line_intrsectn(a,b,cos(theta),-sin(theta),(0:h:L(1))-x(1)), ...
     elpse_line_intrsectn(a,b,sin(theta), cos(theta),(0:h:L(2))-x(2)), ...
     elpse_line_intrsectn(a,b,cos(theta) +sin(theta), ...
                              cos(theta) -sin(theta),(0:h:sum(L))-sum(x))];
% We then convert them in terms of curvilinear coordinates by using
% the four-quadrant atan to find the corresponding curvilinear coords.
[s,I] = sort(atan2(X(2,:)/b,X(1,:)/a));% Sort the results
X = 1/h*([cos(theta) -sin(theta); ...  % Transforms middle of each segment
          sin(theta)  cos(theta)] ...  % [X_i X_i+1] back in the original
    *(1/2*(X(:,I)+X(:,next(I))))...    % coordinates system.
    +x*ones(size(s)));
I = sum((X-(floor(X)+1/3)).^2,1) - ... % Find in which triangle they are
    sum((X-(ceil(X) -1/3)).^2,1) < 0;  % located : above or below diagonal
T=2*c2l(ceil(X(1,:)),ceil(X(2,:)))+1;  % of each square [i,i+1]x[j,j+1] ?
T(~I)= T(~I)+1;
I=abs([1 diff(s)])>eps; s=s(I); T=T(I);% Eliminate the doublons (up to eps)
end

% This sub-function computes the intersection between a line and an ellipse
function y = elpse_line_intrsectn(a,b,alpha,beta,gamma)
                                       % Note: gamma can be a line vector!
r = sqrt((alpha*a)^2+(beta*b)^2);      % Def of r, a useful quantity.
g = gamma(abs(gamma)<=r);              % Selection of gammas for which a
c = a*b/r*sqrt(1-g.^2/r^2);            % a solution exists.
y = [alpha*g*a^2/r^2 - beta*c  ...     % Computation of solutions:
     alpha*g*a^2/r^2 + beta*c ;...     % y(1,:) contains the x component
     beta*g*b^2/r^2 + alpha*c  ...     % y(2,:) contains the y component of
     beta*g*b^2/r^2 - alpha*c];        % each intersection.
end
```

Listing 4.5: The routine `find_intersections_ellipse_triangulation` computes the curvilinear coordinates of the intersection of an ellipse and the triangulation.

Example. Let us use the `Matlab` routines presented in this section in a simple situation. We suppose we have built a uniform regular triangulation of the unit square in two dimensions, as performed in Section 4.2.4. Now, let us assume that in the unit square, there is a circular obstacle centred in $\boldsymbol{x}_0 = (0.5, 0.5)^\top$ and of radius $r = 0.2$. Our goal is to build the matrix associated to the discrete fictitious domain operator b in Equation (3.17). We therefore write the following commands:

```
x0 = [1 1]/2; r = 0.2; N = 5;
mesh_interface = struct('nodes', 2*pi*(0:1/N:(1-1/N)), ...
                'segments', (1:N)', ...
                'segments2nodes', [(1:N)' [(2:(N-1) 1]']));
g = @(t) [r*cos(t)+x0(1); r*sin(t)+x0(2)];
[s,T] =  find_intersections_ellipse_triangulation(x0,r,r,0,h,[1 1]);
B=build_fictitious_operator(s,T, mesh_interface.segments2nodes(:,2),mesh_interface.
    segments, triangulation.tri2nodes, mesh_interface.segments, ...
@(T,S,i,j,a,b) (@(t) ...
r*basis(i,triangulation.nodes(triangulation.tri2nodes(T,:),:)',g(a*t+b*(1-t)))));
```

The first line of this code defines the location, the radius of the obstacle, as well as the number of degrees of freedom for the Lagrange multiplier on the circle. In line 2, we create a structure containing the information relative to the meshing of the interface. In line 5, we define a parametrisation of the circle. In line 6, we compute the intersection of the triangulation with the obstacle. Finally, in line 7, we call the routine for building the discrete fictitious domain operator.

4.4 Effective iterative solvers and preconditioners for linear systems arising from the FDM

One of the differences between the classical FEM and the FDM lies in the nature of the linear systems that both yield at the end of the discretisation process. Even in the case of a symmetric elliptic problem, the linear system obtained from the FDM is not positive definite. It is called a saddle-point problem because its eigenvalues are situated on both sides of the real axis.

This type of linear systems not only arises in FDM but in many other applications, such as fluid dynamics modelling via Navier-Stokes equations [46, 47]. The general form of a saddle-point linear system is:

$$\underbrace{\begin{pmatrix} A & B_2 \\ B_1^\top & 0 \end{pmatrix}}_{\mathcal{A}} \underbrace{\begin{pmatrix} u \\ \lambda \end{pmatrix}}_{x} = \underbrace{\begin{pmatrix} f \\ g \end{pmatrix}}_{b}. \tag{4.14}$$

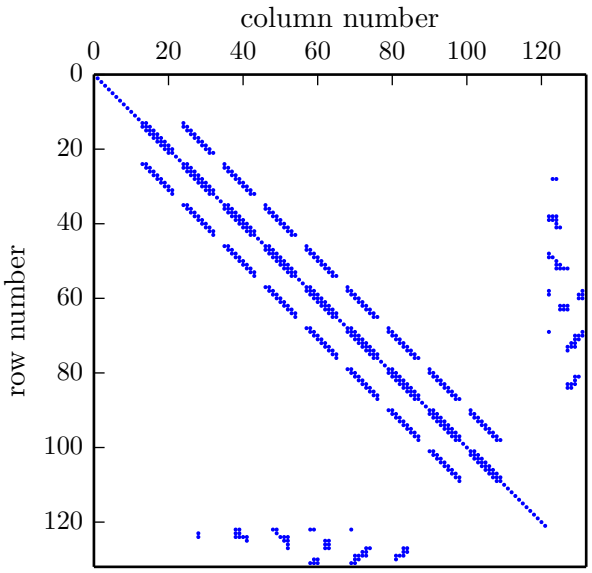


Figure 4.2: Sparsity pattern obtained from the linear system resulting from the discretisation of a fictitious domain formulation of the Dirichlet elliptic problem given in (3.18). The space resolution is $h = 1/10$.

The matrix A is assumed to be nonsingular and B_1 and B_2 are of full (column) rank, to ensure the uniqueness of the solution. Unlike fluid dynamics, where the unknown λ represents the pressure and where its support spans the whole computational domain, in the FDM, the dimension of the space supporting λ is generally much smaller than that of the volumic unknown u. The ratio between those dimensions shrinks proportionally with h. For fine resolutions, say $h = 1/100$, then A is approximately $100^2 \times 100^2$ while the mass matrix on the interface would have size 100×100.

While the dimension of the Lagrange multiplier space is relatively small, the projection matrices B_1 and B_2 remains unstructured. This decreases the effectiveness of direct methods for solving linear systems, often relying on the limited bandwidth of the linear system. Hence, direct methods such as Gauss substitution are not optimal and other techniques must be considered, such as iterative methods. These usually require only matrix to vector multiplication, therefore they are well suited to the sparse structure of the matrices involved [118].

In the next sections, we will discuss efficient iterative methods and preconditioners for the resolution of large sparse saddle-point linear systems arising in the FDM. The first iterative method developed specifically for solving indefinite symmetric linear system is

the Uzawa iteration [5]. It takes advantage of the structure of the matrix $\mathcal{A}$ by splitting it and using separately each block A and B in a two-step iteration.

The classical Uzawa iteration starts with an arbitrary λ_0 and computes the sequence, for $k \geq 0$,

$$\begin{aligned} u_{k+1} &= A^{-1}(f - B^\top \lambda_k) \\ \lambda_{k+1} &= \lambda_k + \omega\,(Bu_{k+1} - g)\,, \end{aligned} \tag{4.15}$$

until the convergence of the iteration, that is $\frac{\|u_{k+1}-u_k\|}{\|u_k\|} \leq \varepsilon$, where ε represents the tolerance set by the user. The parameter ω must be chosen before the start of the iteration.

The convergence rate of the scheme amounts to $\frac{\kappa-1}{\kappa+1}$, where κ is the condition number of the matrix $\mathcal{A}$, defined as the ratio between the largest and the smallest eigenvalue in modulus. In practice, some issues can be encountered at the implementation stage [45]:

1. At each iteration, we have to apply A^{-1} once even if we need not form A^{-1} explicitly. This is in general a costly procedure even when A has good properties.
2. Since $\kappa \to \infty$ as $h \to 0$, the convergence rate $\frac{\kappa-1}{\kappa+1}$ decreases dramatically as it approaches 1. This convergence rate is in any case not as advantageous as that of the conjugate gradient algorithm, which is $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ [73].
3. There is not an *a priori* explicit value for ω, and this makes the implementation of the scheme less useful.

4.4.1 Preconditioned minimal residual iteration algorithm

Modern algorithms for solving large saddle-point linear systems usually do not split the system. This is the case of Krylov subspaces based methods such as the Minimum Residual algorithm, MINRES, the first iteration algorithm designed by Paige and Saunders [105] specifically for symmetric indefinite linear systems. We will briefly explain the underlying theory but more substantial information can be found in [10].

Let x_0 be an initial guess of the solution of (4.14). Then we can define the initial residual *i.e.* the error, by $r_0 = b - \mathcal{A}x_0$. The k-th iterate of methods based on Krylov subspaces

satisfies the following property: for any $j \geq 1$,

$$x_j \in x_0 + \mathcal{K}_j(\mathcal{A}, r_0) = \text{span}\left(r_0, \mathcal{A}r_0, \mathcal{A}^2 r_0, \ldots, \mathcal{A}^{j-1} r_0\right). \tag{4.16}$$

It is well known that $\mathcal{K}_1(\mathcal{A}, r_0) \subset \mathcal{K}_2(\mathcal{A}, r_0) \subset \ldots \subset \mathcal{K}_j(\mathcal{A}, r_0)$ since $\mathcal{A}$ is nonsingular, and $\dim \mathcal{K}_j(\mathcal{A}, r_0) \leq \min(j, N) \leq N$, where N is the dimension of the matrix $\mathcal{A}$. Choosing the j-th residual $r_j = b - \mathcal{A}x_j$ to be orthogonal to a constraint space $\mathcal{C}_j$ of dimension j fulfils the requirement of uniqueness of the definition of x_j, providing j independent constraints on x_j. Different choices of the constraint spaces lead to different iteration algorithms. The MINRES algorithm sets $\mathcal{C}_j = \mathcal{A}\mathcal{K}_j(\mathcal{A}, r_0)$.

At each step, we search for x_j in the form $x_j = x_{j-1} + \alpha_j r_{j-1}$, which fulfils the requirement of x_j being an element of $x_0 + \mathcal{K}_j(\mathcal{A}, r_0)$. This can be shown by a simple mathematical induction:

$$r_j = b - \mathcal{A}x_{j-1} - \alpha_j \mathcal{A}r_{j-1} = r_{j-1} - \alpha_j \mathcal{A}r_{j-1}. \tag{4.17}$$

We deduce the value of α_j by requiring: $r_j \perp \mathcal{A}\mathcal{K}_j(\mathcal{A}, r_0)$. To enforce this condition, we remember that $r_{j-1} \perp \mathcal{A}\mathcal{K}_{j-1}(\mathcal{A}, r_0)$. So we understand that it suffices to subtract the contribution of the projection of r_{j-1} on $\text{span}(\mathcal{A}r_{j-1}) \subset \mathcal{A}\mathcal{K}_j(\mathcal{A}, r_0)$ from r_{j-1}. This yields the value of α_j:

$$\alpha_j = \frac{r_{j-1}^\top \mathcal{A} r_{j-1}}{\|\mathcal{A}r_{j-1}\|_2^2}. \tag{4.18}$$

The iteration algorithm is then:

1. Choose x_0 and set $r_0 = b - \mathcal{A}x_0$.
2. For $j \geq 1$, while $\|r_{j-1}\| > \varepsilon \|b\|$, compute:
   (a) α_j as per Equation (4.18),
   (b) $x_j = x_{j-1} + \alpha_j r_{j-1}$,
   (c) $r_j = r_{j-1} - \alpha_j r_{j-1}$.
   (d) $j := j + 1$.

This iteration only requires the matrix-vector multiplication $\mathcal{A}x$ and the input matrix $\mathcal{A}$ is not split. We notice that this iteration algorithm constructs an orthogonal basis $(r_0, r_1, \ldots, r_n)$ for the Krylov subspaces $\mathcal{K}_n$, even if in a computer implementation it

is never explicitly stored (only the last vector of the basis r_n is available) in order to contain the memory costs.

Theorem 4.1. *Let $\mathcal{A}$ be a nonsingular symmetric (indefinite) matrix of size $N \times N$ and let $b \in \mathbb{R}^N$ be a vector. We assume that the spectrum of $\mathcal{A}$ has the form: $\sigma(\mathcal{A}) \subset [\lambda_{\min}, \lambda_s] \cup [\lambda_{s+1}, \lambda_{\max}]$, where $\lambda_{\min} < \lambda_s < 0 < \lambda_{s+1} < \lambda_{\max}$. We also let $\mathcal{A} = Q\mathcal{D}Q^{-1}$, where $\mathcal{D}$ is a diagonal matrix with the eigenvalues of $\mathcal{A}$ as diagonal entries. In this case, the residual r_j satisfies the following bound:*

$$\|r_j\|_2 \leq 2\kappa(Q) \left(\frac{\sqrt{|\lambda_{\min}\lambda_{\max}|} - \sqrt{|\lambda_s\lambda_{s+1}|}}{\sqrt{|\lambda_{\min}\lambda_{\max}|} + \sqrt{|\lambda_s\lambda_{s+1}|}} \right)^{[j/2]} \|r_0\|_2, \tag{4.19}$$

where $\kappa(Q)$ is the condition number of Q.

Proof. This result was proved by Greenbaum [76]. □

Benzi *et al.* [10] provide a useful illustration of the previous bound. If we suppose that $\lambda_{\max} = |\lambda_{\min}| = 1$ and $\lambda_{s+1} = |\lambda_s|$, then $\kappa(\mathcal{A}) = |\lambda_s|^{-1}$ and we obtain another bound, involving the condition number of $\mathcal{A}$:

$$\|r_j\|_2 \leq 2\kappa(Q) \left(\frac{\kappa(\mathcal{A}) - 1}{\kappa(\mathcal{A}) + 1} \right)^{[j/2]} \|r_0\|_2. \tag{4.20}$$

In practice, we could try to improve the convergence rate by transforming the linear system to another with more attractive spectral properties. This operation is called 'preconditioning'.

4.4.2 Generalised minimal residual iteration algorithm

The algorithm presented in the previous section, MINRES, has been designed for symmetric nonsingular linear systems. In the case of a nonsingular but non symmetric linear system, MINRES will not always converge. To address this issue, Saad and Schultz [119] proposed a generalisation of the MINRES iteration algorithm to general linear systems, named generalised minimal residual method, or GMRES.

The GMRES iteration algorithm also relies on building an orthogonal basis of the Krylov subspaces, and minimising the norm of the residual $r_n = \mathcal{A}x_n - b$ over the Krylov subspace of order n.

The construction of the orthonormal basis for $\mathcal{K}_n(\mathcal{A}, r_0)$ is performed using a specialisation of the Gram-Schmidt method, called Arnoldi's method [4]. This iteration algorithm builds two matrices: V_n of size $N \times n$ an orthonormal basis for $\mathcal{K}_n(\mathcal{A}, r_0)$ with first vector $v_0 = r_0/\|r_0\|$, expressed in the canonical basis and H_n of size $(n+1) \times n$ such that:

$$\mathcal{A}V_n = V_{n+1}H_n. \tag{4.21}$$

The n-th iterate x_n approximating the exact solution is no longer computed as an update of x_{n-1}, but is sought in the form $x_n = x_0 - z_n$, where $z_n \in \mathcal{K}_n(\mathcal{A}, r_0)$, such that z_n is the solution of the following least-squares problem:

$$\min_{z \in \mathcal{K}_n(\mathcal{A}, r_0)} \|\mathcal{A}(x_0 - z) - b\| = \min_{z \in \mathcal{K}_n(\mathcal{A}, r_0)} \|\mathcal{A}z - r_0\|. \tag{4.22}$$

Now, if we choose: $z_n = V_n y_n$ then we have:

$$r_n = \mathcal{A}x_n - b = V_{n+1}\left(H_n y_n - \|r_0\| e_1\right). \tag{4.23}$$

Hence the least squares problem (4.22) becomes:

$$\min_{y_n \in \mathbb{R}^n} \|H_n y_n - \|r_0\| e_1\|. \tag{4.24}$$

We summarise below the steps of the GMRES iteration algorithm [118, Algorithm 6.9, p. 172]:

1. Choose x_0 and set $r_0 = b - \mathcal{A}x_0$.
2. $v_0 = r_0/\|r_0\|$.
3. While $\|r_{n-1}\| > \varepsilon$:
   (a) Perform Arnoldi's iteration: for $1 \leq j \leq n$, compute:
      i. $w_j = \mathcal{A}v_j$.
      ii. $h_{i,j} = v_i^\top w_j$, for $0 \leq i \leq j$.
      iii. $\hat{v}_{j+1} = w_j - \sum_{i=1}^{j} h_{i,j} v_i$.
      iv. $h_{j+1,j} = \|\hat{v}_{j+1}\|$.
      v. $v_{j+1} = \hat{v}_{j+1}/h_{j+1,j}$.

(b) Compute y_n the solution of the least squares problem (4.24) and new residual r_n.

(c) $n := n + 1$.

This iteration algorithm performs two matrix-vector multiplications at every step and uses $O(n^2)$ storage, where n is the number of iterations required to achieve the desired tolerance ε. Therefore we understand that it will be crucial to use efficient preconditioners to minimise n, as the cost of each iteration increases quadratically with the iteration number.

4.4.3 Preconditioning of saddle-point linear systems arising from the FDM

4.4.3.1 Introduction to preconditioning

Let $\mathcal{A}$ be a nonsingular matrix operator, for instance derived from the discretisation of a differential operator, with the structure:

$$\mathcal{A} = \begin{pmatrix} A & B_1^\top \\ B_2 & 0 \end{pmatrix}, \tag{4.25}$$

where A is a symmetric positive definite matrix and B is rectangular with full rank. For any right-hand side $(F, G)^\top$, we wish to solve:

$$\begin{pmatrix} A & B_1^\top \\ B_2 & 0 \end{pmatrix} \begin{pmatrix} U \\ \Lambda \end{pmatrix} = \begin{pmatrix} F \\ G \end{pmatrix}. \tag{4.26}$$

When solving for Λ first, we obtain:

$$\begin{pmatrix} A & B_1^\top \\ 0 & -B_2 A^{-1} B_1^\top \end{pmatrix} \begin{pmatrix} U \\ \Lambda \end{pmatrix} = \begin{pmatrix} F \\ G - B_2 A^{-1} F \end{pmatrix}. \tag{4.27}$$

Definition 4.3 (Schur complement)**.** We define the Schur complement as the matrix:

$$S = -B_2 A^{-1} B_1^\top. \tag{4.28}$$

Hence Equation (4.27) becomes:

$$\begin{pmatrix} A & B_1^\top \\ 0 & S \end{pmatrix} \begin{pmatrix} U \\ \Lambda \end{pmatrix} = \begin{pmatrix} F \\ G - B_2 A^{-1} F \end{pmatrix}. \tag{4.29}$$

Inverting a saddle-point linear system of the form (4.25) amounts to inverting the block A and the Schur complement.

Definition 4.4 ((Left) Preconditioner)**.** A preconditioner is defined as a matrix $\mathcal{P}$ chosen carefully so that it has eigenvalues close to that of $\mathcal{A}$ but with a simpler structure, so the matrix $\mathcal{P}^{-1}\mathcal{A}$ has a lower condition number than $\mathcal{A}$. Given a right-hand side b, solving for x such that $\mathcal{A}x = b$ is equivalent to solving:

$$\begin{aligned} \mathcal{P}c &= b, \\ \left[\mathcal{P}^{-1}\mathcal{A}\right] x &= c, \end{aligned} \tag{4.30}$$

Remark 4.5. For example, if we choose $\mathcal{P} = \mathcal{A}$, then $P^{-1}\mathcal{A} = I$, so the condition number of the new system is optimal since equal to 1. However, the computational effort is transferred to computing $c = \mathcal{P}^{-1}b$. On the other hand, if we choose $\mathcal{P} = I$, then computing c becomes straightforward, however $P^{-1}\mathcal{A} = \mathcal{A}$.

Remark 4.6. In general, we never need to form the product $\mathcal{P}^{-1}\mathcal{A}$, but rather we apply each operator independently, one after the other.

Remark 4.7. Ideally, we would choose a preconditioner such that the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ are clustered around 1, and such that its condition number would be independent of the characteristic mesh size of the discretisation scheme. In contrast, the discretisation of the two-dimensional Laplacian operator using a finite-difference scheme leads to a matrix operator whose condition number grows as $1/h^2$.

In the case of saddle-point linear systems, we generally look for block preconditioners. That is to say:

$$\mathcal{P} = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}. \tag{4.31}$$

When choosing $A_1 = A$ and $A_2 = -S = B_2A^{-1}B_1^\top$, the opposite of the Schur complement, then Murphy *et al.* [96] have shown that the preconditioned system $\mathcal{P}^{-1}\mathcal{A}$ satisfies:

$$\left(\mathcal{P}^{-1}\mathcal{A} - I\right)\left(\mathcal{P}^{-1}\mathcal{A} - \frac{1+\sqrt{5}}{2}I\right)\left(\mathcal{P}^{-1}\mathcal{A} - \frac{1-\sqrt{5}}{2}I\right) = 0, \tag{4.32}$$

so $\mathcal{P}^{-1}\mathcal{A}$ is diagonalisable and its spectrum is a subset of $\left\{\frac{1-\sqrt{5}}{2}, 1, \frac{1+\sqrt{5}}{2}\right\}$. Hence, the dimension of the Krylov subspace for any initial guess x_0 does not exceed 3, so the MINRES or GMRES algorithm will converge in at most 3 steps. In practice, forming this

preconditioner $\mathcal{P}$ is not computationally adequate, because it comes down to computing the inverse of the symmetric positive definite block A. The conclusion is that in general, we cannot use exact preconditioners, and we must replace them with an approximation. The approximation chosen is specfic to the linear system we are dealing with, so there exists no general rule for forming an efficient preconditioner for a given matrix operator.

Many solutions have been proposed for approximating the block A: among them, let us quote the incomplete LU factorisation [118] for general nonsingular square sparse matrices, the incomplete Cholesky factorisation for square positive definite matrices [89, 73], and the algebraic multigrid iteration method [117]. We focus on the second block of the preconditioner.

In the next sections, we will discuss possible choices of preconditioners adapted to linear systems arising from the discretisation of fictitious domain formulated problems discussed in Chapter 3.

4.4.3.2 Preconditioning for linear systems arising from Dirichlet FDM

Let us recall the form of the linear system obtained in Chapter 3 from the discretisation of the fictitious formulation of an elliptic problem with an immersed Dirichlet boundary, *i.e.* Equation (3.18):

$$\text{find } (\widetilde{U}, L) \in \mathbb{R}^N \times \mathbb{R}^M : \begin{pmatrix} A & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} \widetilde{U} \\ L \end{pmatrix} = \begin{pmatrix} \widetilde{F} \\ 0 \end{pmatrix}. \tag{3.18}$$

In this equation, A represents the first order discretisation of the second order elliptic operator $u \mapsto -\nu\nabla^2 u + \alpha u$ on a rectangle with arbitrary boundary conditions and is a symmetric positive definite matrix, and B is the fictitious domain specific matrix operator, carrying the information relative to the embedded Dirichlet boundary condition. As we know from the previous section, preconditioners for this kind of linear systems are chosen with a block structure:

$$\mathcal{P} = \begin{pmatrix} \widetilde{A} & 0 \\ 0 & \widetilde{S} \end{pmatrix}, \tag{4.33}$$

where $\widetilde{A}$ is an approximation to A, such that solving $\widetilde{A}x = b$ is a cheap computational task and such that the eigenvalues of $\widetilde{A}$ are similar to those of A. In the same expression, $\widetilde{S}$ is an approximation to the Schur complement.

For the approximation of the Schur complement, Glowinski [68] advocates the use of the square-root of the discretised one-dimensional equivalent of our differential operator, that is to say:

$$\widetilde{S}^{-1} = \left(\frac{\alpha}{\nu} I_\eta - \frac{\Delta_\eta}{|\gamma|^2} \right)^{1/2}, \tag{4.34}$$

where

$$\Delta_\eta = -\frac{1}{\eta^2} \begin{pmatrix} 2 & -1 & 0 & \cdot & -1 \\ -1 & 2 & -1 & \cdot & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -1 & 2 & -1 \\ -1 & \cdots & 0 & -1 & 2 \end{pmatrix}, \tag{4.35}$$

and $|\gamma|$ is the perimeter of the interface. The (positive) square-root of the matrix between brackets is well-defined and unique, since it is symmetric positive definite. The matrix Δ_η represents the 3-point discretisation of the one-dimensional Laplacian operator, with periodic boundary conditions.

The method of derivation of this preconditioner will be shown in Section 4.4.3.3, when addressing the preconditioning of the linear systems arising in Neumann FDM. However, since the fictitious domain operator B represents a projection operator, it seems reasonable to approximate the Schur complement by the projection of the volume elliptic operator on the one-dimensional immersed boundary.

In order to form the preconditioner $\widetilde{S}$, it is necessary to compute the square root of the matrix between brackets on the right-hand side of (4.34). This matrix is symmetric positive definite, so its eigenvalues are real and strictly positive, which ensures the existence and uniqueness of the square-root. The formation of the square root involves the diagonalisation of $\frac{\alpha}{\nu} I_\eta - \frac{\Delta_\eta}{|\gamma|^2}$. Given the reasonable size of the Lagrange multiplier space, the cost of such operation remains reasonable.

We show in Table 4.1 how Glowinski's preconditioner (4.34) performs. We also provide the number of iterations in MINRES when no preconditioner is used. In order to ensure that we are solely evaluating the performance of the block $\widetilde{T}$ as a preconditioner, we use exact preconditioning for block $\widetilde{A}$, hence we choose $\widetilde{A} = A$. We will not further discuss

	Number of iterations in MINRES	
h	$\widetilde{S} = \left(\frac{\alpha}{\nu} I - \frac{\Delta_\eta}{\lvert\gamma\rvert^2}\right)^{-\frac{1}{2}}$	$\widetilde{S} = I$
1/8	13	13
1/16	19	21
1/32	19	33
1/64	21	43
1/128	21	55
1/256	19	71

Table 4.1: Performance in terms of iteration count in MINRES of the preconditioner $\widetilde{S}$ provided in (4.34) for solving the linear system (3.18), given for several grid resolutions.

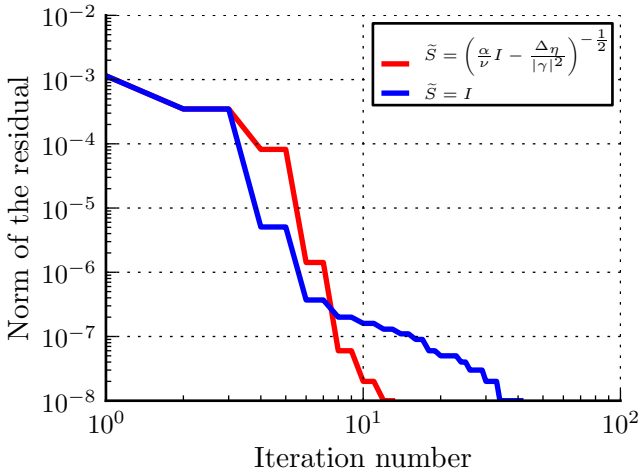


Figure 4.3: Plot of the norm of the residual after the n-th iteration of the MINRES iteration algorithm, for preconditioner $\widetilde{S}$ derived in Equation (4.34). The plot of the norm of the residual when using no preconditioner is plotted for comparison. The mesh size chosen here is $h = \frac{1}{256}$.

choices for the preconditioning of block A, however typically, when regular meshes are used, algebraic multigrid methods prove to be very efficient (see [23]).

From Table 4.1, we observe that, as the resolution h becomes finer, the number of iterations required to achieve the convergence of the MINRES algorithm, for a tolerance set to $\varepsilon = 10^{-8}$, does not vary. This is the sign of optimal preconditioning. As an illustration, we have plotted in Figure 4.3 the norm of the residual after the n-th iteration of the MINRES algorithm when the linear system is preconditioned either with $\widetilde{S}$ as in (4.34) or without any preconditioning.

4.4.3.3 Preconditioning for linear systems arising from the Neumann FDM

Let us recall the linear system we have obtained in Section 3.3 after discretising the elliptic Neumann boundary problem (3.25):

$$\begin{pmatrix} A_{\boldsymbol{\sigma}} & D & B \\ D^\top & -A_u & 0 \\ B^\top & 0 & 0 \end{pmatrix} \begin{pmatrix} \Sigma_h \\ U_h \\ \Lambda_\eta \end{pmatrix} = \begin{pmatrix} 0 \\ F \\ G \end{pmatrix}. \tag{3.38}$$

The literature does not provide us with an efficient preconditioner when using an iterative solver on this linear system. Our goal in this section is to derive and test a preconditioner improving the efficiency of the MINRES iteration algorithm when solving the linear system (3.38). To simplify this problem, let us first notice that for a nonsingular A_u, then (3.38) reduces to:

$$\begin{pmatrix} A_{\boldsymbol{\sigma}} + DA_u^{-1}D^\top & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} \Sigma_h \\ \Lambda_\eta \end{pmatrix} = \begin{pmatrix} -A_u^{-1}F \\ G \end{pmatrix}, \tag{4.36}$$

and

$$A_u U_h = D^\top \Sigma_h - F. \tag{4.37}$$

When using piecewise constant triangular elements as performed in Section 3.3.2, A_u is a diagonal matrix, so inverting A_u becomes straightforward. We hence concentrate on (4.36), and investigate the possibility of using a block preconditioner, that is to say, in the form:

$$\mathcal{P} = \begin{pmatrix} \widetilde{C} & 0 \\ 0 & \widetilde{T} \end{pmatrix}, \tag{4.38}$$

where $\widetilde{C}$ is an approximation of $C = A_{\boldsymbol{\sigma}} + DA_u^{-1}D^\top$ and $\widetilde{T}$ an approximation of the Schur complement of the linear system. We focus on finding an efficient block $\widetilde{T}$. Let us adapt Glowinski's Fourier approach in [68] that derived a preconditioner for the Schur complement in the Dirichlet FDM.

Let $\gamma = \{(0, y) : \ y \in \mathbb{R}\}$. Let us define the following operator F, mapping $H^{1/2}(\gamma)$ to $H^{-1/2}(\gamma)$ such that $\forall \mu \in H^{1/2}(\gamma)$, $F\mu = (\boldsymbol{\sigma}_\mu \cdot \boldsymbol{n})|_\gamma$, where $\boldsymbol{n_\gamma} = (1, 0)^\top$ is the outwards normal of γ and $\boldsymbol{\sigma}_\mu$ the solution of the weak problem: find $\boldsymbol{\sigma}_\mu \in H^{\mathrm{div}}(\mathbb{R}^2)$ such that $\forall \boldsymbol{\tau} \in H^{\mathrm{div}}(\mathbb{R}^2)$,

$$\left(\nu^{-1}\boldsymbol{\sigma}_\mu, \boldsymbol{\tau}\right)_{(L^2(\mathbb{R}^2))^2} + (\alpha^{-1}\nabla \cdot \boldsymbol{\sigma}_\mu, \nabla \cdot \boldsymbol{\tau})_{L^2(\mathbb{R}^2)} = \langle \boldsymbol{\tau} \cdot \boldsymbol{n}, \mu \rangle_{H^{-1/2} \times H^{1/2}}. \tag{4.39}$$

Equivalently, the couple $(\boldsymbol{\sigma}_\mu, u_\mu)$ is the strong solution of the problem:

$$\boldsymbol{\sigma}_\mu = -\nabla u_\mu \quad \text{and} \quad \nabla \cdot \boldsymbol{\sigma}_\mu + \frac{\alpha}{\nu} u_\mu = 0 \quad \text{in } \mathbb{R}^2, \tag{4.40}$$

with boundary and interface conditions:

$$[\boldsymbol{\sigma}_\mu \cdot \boldsymbol{n}_\gamma]_\gamma = 0 \quad \text{and} \quad [u_\mu]_\gamma = \mu \quad \text{on } \gamma, \tag{4.41}$$

$$u \to 0 \quad \text{as } |x| \to \infty, \tag{4.42}$$

where $[u]_\gamma$ denotes the jump of u across γ. We observe that F is well-defined, by existence and uniqueness of the solution of (4.39) for a given $\mu \in H^{1/2}(\gamma)$, and is also linear. We can interpret the operator F as the continuous equivalent of the Schur complement $S = -BC^{-1}B^\top$.

In this simple setting, it is actually possible to obtain an explicit form for the expression of F. Let $\mu \in H^{1/2}(\gamma)$ and let $\hat{\mu}$ be the Fourier transform of μ, then:

$$\forall y \in \mathbb{R}, \quad \mu(y) = \int_{\mathbb{R}} \exp(2i\pi s\, y)\hat{\mu}(s)\, \mathrm{d}s. \tag{4.43}$$

Then, using the linearity of the integral, we compose by F, which leads to:

$$\forall y \in \mathbb{R}, \quad (F\mu)(y) = \int_{\mathbb{R}} [F \exp(2i\pi s\, \cdot)](y)\hat{\mu}(s)\, \mathrm{d}s. \tag{4.44}$$

We can see that it is sufficient to compute the action of F on $\exp(2i\pi s\, \cdot)$ in order to determine F completely. This amounts to solving (4.40) with $\mu(y) = \exp(2i\pi s\, y)$, for any $s \in \mathbb{R}$.

Because of the symmetry of the problem, let us seek u in the form: $u(x,y) = u_1(x)u_2(y)$. By separation of variables, we obtain two ordinary differential equations:

$$u_1'' - K^2 u_1 = 0 \quad \text{in } \mathbb{R}, \tag{4.45a}$$

$$u_2'' - \left(\frac{\alpha}{\nu} - K^2\right) u_2 = 0 \quad \text{in } \mathbb{R}, \tag{4.45b}$$

with boundary conditions: $u \to 0$ as $|x| \to \infty$, $[u]_\gamma = \exp(2i\pi s\, y)$ and $[\nabla u \cdot \boldsymbol{n}_\gamma]_\gamma = 0$. The constant K is unknown and has to be determined.

The general solution of (4.45a) reads:

$$u_1(x) = \begin{cases} A^+ \exp(-Kx) + B^+ \exp(Kx), & \text{for } x > 0 \\ A^- \exp(Kx) + B^- \exp(-Kx), & \text{for } x < 0, \end{cases} \tag{4.46}$$

from which we can immediately deduce that $B^+ = B^- = 0$, in order to preserve the boundedness of u. Using the jump conditions on u, we obtain:

$$\begin{aligned} A^+ &= \frac{1}{2}, \\ A^- &= -\frac{1}{2}. \end{aligned} \tag{4.47}$$

Hence the expression for u_1 reads:

$$u_1(x) = \frac{1}{2}\left(\exp(-Kx)\mathbb{1}_{\mathbb{R}^+}(x) - \exp(Kx)\mathbb{1}_{\mathbb{R}^-}(x)\right). \tag{4.48}$$

Moreover, Equation (4.45b) yields:

$$K = \sqrt{\frac{\alpha}{\nu} + 4\pi^2 s^2}, \tag{4.49}$$

and the second part of the solution u_2 is:

$$u_2(y) = \exp(2i\pi s\, y). \tag{4.50}$$

We finally obtain the expression for the solution $\boldsymbol{\sigma}_{\exp(2i\pi sy)}$:

$$\boldsymbol{\sigma}_{\exp(2i\pi sy)} = \frac{1}{2}\sqrt{\frac{\alpha}{\nu} + 4\pi^2 s^2}\, \exp\left(2i\pi s\, y - \sqrt{\frac{\alpha}{\nu} + 4\pi^2 s^2}\, |x|\right) (1,0)^\top. \tag{4.51}$$

This last expression allows us to write the following expression for F:

$$\begin{aligned} \forall y \in \mathbb{R}, \quad (F\mu)(y) &= \int_{\mathbb{R}} \left(\boldsymbol{\sigma}_{\exp(2i\pi sy)} \cdot \boldsymbol{n}\right)\big|_\gamma \hat{\mu}\, \mathrm{d}s \\ &= \frac{1}{2}\int_{\mathbb{R}} \exp(2i\pi s\, y) \sqrt{\frac{\alpha}{\nu} + 4\pi^2 s^2}\, \hat{\mu}(s)\, \mathrm{d}s. \end{aligned} \tag{4.52}$$

So we find from the expression above that the Fourier transform of $F\mu$ denoted by $\widehat{F\mu}$, satisfies

$$\forall s \in \mathbb{R}, \quad \widehat{F\mu}(s) = \frac{1}{2}\sqrt{\frac{\alpha}{\nu} + 4\pi^2 s^2}\,\hat{\mu}(s). \tag{4.53}$$

Therefore, by composing F with itself once, we obtain

$$\widehat{(F \circ F)\mu} = \frac{1}{4}\left(\frac{\alpha}{\nu} + 4\pi^2 s^2\right)\hat{\mu}. \tag{4.54}$$

We deduce that $(F \circ F)(\mu)$ satisfies

$$(F \circ F)(\mu) = \frac{1}{4}\left(\frac{\alpha}{\nu}\mu - \mu''\right), \tag{4.55}$$

and finally

$$F\mu = \frac{1}{2}\left(\frac{\alpha}{\nu}I - \frac{\mathrm{d}^2}{\mathrm{d}y^2}\right)^{\frac{1}{2}}\mu, \tag{4.56}$$

where the power $\frac{1}{2}$ is to be understood in terms of operator composition: F is the operator such that, when composed with itself yields the operator (4.55). Let us observe that, in the case of an infinite domain, the operator F can be determined exactly and is the 'square-root' of a second order differential operator, and defined on $H^{1/2}(\gamma)$ onto $H^{-1/2}(\gamma)$. Of course, in the case of a bounded domain, it might not be possible to derive an explicit expression for F, however, we can take advantage of this simple unbounded case by guessing what an adequate preconditioner for the linear system (4.36) would be.

Therefore we conclude this argument by taking as an approximation to the Schur complement of (4.36) the following matrix operator:

$$\widetilde{T} = \left(\frac{\alpha}{\nu}I_\eta - \frac{\Delta_\eta}{|\gamma|^2}\right)^{1/2}. \tag{4.57}$$

The (positive) square root exists since the interior matrix is symmetric positive definite.

Remark 4.8. In order to form this preconditioner numerically using `Matlab`, we first compute the matrix $U = \frac{\alpha}{\nu}I_\eta - \frac{\Delta_\eta}{|\gamma|^2}$, and we diagonalise it numerically using `eig` command in `Matlab` and obtain $U = VDV^{-1}$, where D is the diagonal matrix of eigenvalues of U. Given the size of the Lagrange multiplier space, this operation is not too computationally intensive. Since we expect all the eigenvalues to be strictly positive, we can take the

	Number of iterations in MINRES	
h	$\widetilde{T} = \left(\frac{\alpha}{\nu} I - \frac{\Delta_\eta}{\lvert\gamma\rvert^2}\right)^{\frac{1}{2}}$	$\widetilde{T} = I$
1/8	13	13
1/16	17	19
1/32	17	27
1/64	19	37
1/128	19	47
1/256	19	65

Table 4.2: Performance in terms of iteration count in MINRES of the preconditioner $\widetilde{T}$ derived in (4.57) for solving the linear system (3.38), and for several grid resolutions h.

square root of each one of them and so we form the matrix $\widetilde{T} = VD^{1/2}V^{-1}$. Furthermore, since in each iteration of the MINRES algorithm a solution $\widetilde{T}x = b$ is computed, it is reasonable to compute $\widetilde{T}^{-1}$ directly, by computing $D^{-1/2}$ instead of $D^{1/2}$.

We show in Table 4.2 how the preconditioner we have derived in this section performs on linear systems arising from the FDM for Neumann problems. We provide the number of iterations in MINRES when no preconditioner is used for comparison. In order to ensure that we are only evaluating the performance of the block $\widetilde{T}$ as a preconditioner, we use exact preconditioning for block $\widetilde{C}$.

We observe that, as the resolution h becomes finer, the number of iterations required to achieve the convergence of the MINRES algorithm, for a tolerance set to $\varepsilon = 10^{-8}$, does not vary. This is the sign of optimal preconditioning, since the number of iterations does not depend on the size of the linear system. As an illustration, we have plotted in Figure 4.4 the norm of the residual after the n-th iteration of the MINRES algorithm, when either the linear system is preconditioned with $\widetilde{S}$ as in (4.57) or without any preconditioning.

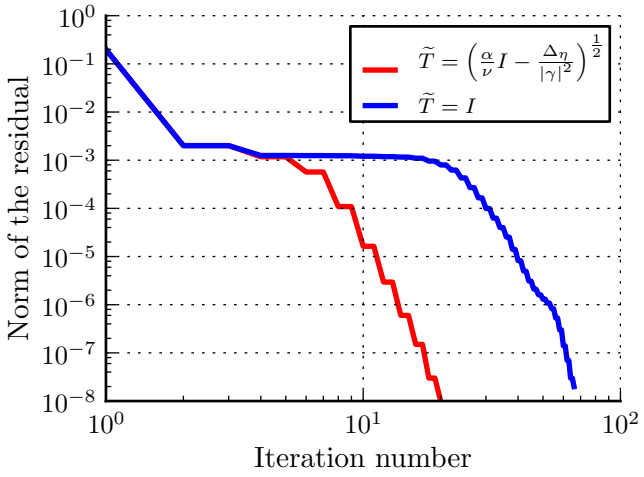


Figure 4.4: Plot of the norm of the residual after the n-th iteration of the MINRES iteration algorithm, for preconditioner $\widetilde{T}$ derived in Equation (4.57). The plot of the norm of the residual when using no preconditioner is plotted for comparison. The mesh size chosen here is $h = \frac{1}{256}$.

4.4.3.4 Preconditioning for linear systems arising from the Robin FDM

Let us now recall from Chapter 3 the form of the linear system obtained in Section (3.4) after having discretised the elliptic Robin boundary problem (3.46):

$$\begin{pmatrix} A_{\boldsymbol{\sigma}} & D & B \\ D^\top & -A_u & 0 \\ B^\top & P_h & 0 \end{pmatrix} \begin{pmatrix} \Sigma_h \\ U_h \\ \Lambda_\eta \end{pmatrix} = \begin{pmatrix} 0 \\ F \\ G \end{pmatrix}. \tag{3.56}$$

As performed in (4.36), let us simplify the previous linear system and eliminate the unknown U_h. We obtain the following non symmetric linear system:

$$\begin{pmatrix} A_{\boldsymbol{\sigma}} + DA_u^{-1}D^\top & B \\ B^\top + P_h A_u^{-1} D^\top & 0 \end{pmatrix} \begin{pmatrix} \Sigma_h \\ \Lambda_\eta \end{pmatrix} = \begin{pmatrix} -DA_u^{-1}F \\ G - P_h A_u^{-1} F \end{pmatrix}. \tag{4.58}$$

The literature does not provide us with an efficient preconditioner when using the GMRES solver on this linear system. Our goal in this section is to find a suitable a preconditioner for linear system (3.56).

This linear system, although non symmetric, has a form similar to (4.36). This observation encourages us to investigate whether the preconditioner derived and tested for the Neumann FDM would also be suitable in the Robin case.

Using the same procedure as in the previous section, we obtain the results showed in Table 4.3. We observe that the preconditioner $\widetilde{T}$ presented in Equation (4.57) appears

	Number of iterations in GMRES	
h	$\widetilde{T} = \left(\frac{\alpha}{\nu} I - \frac{\Delta_\eta}{\lvert\gamma\rvert^2}\right)^{\frac{1}{2}}$	$\widetilde{T} = I$
1/8	9	9
1/16	15	15
1/32	19	29
1/64	19	43
1/128	21	61
1/256	19	67

Table 4.3: Performance in terms of iteration count in GMRES of the preconditioner $\widetilde{T}$ derived in (4.57) for solving the linear system (3.56), and for several grid resolutions h.

also to be optimal for the non symmetric linear system (4.58). The number of iterations to achieve convergence of the GMRES algorithm for the preconditioned system seems to be independent of the size of the linear system, which is the sign of optimal preconditioning. In Figure 4.5, we have plotted for mesh size $h = 1/256$ the norm of the residual after the n-th iteration of the GMRES iteration algorithm either when using $\widetilde{T}$ as a preconditioner or without preconditioning for comparison.

4.5 Summary

In this chapter, we have presented the basic algorithms and `Matlab` code that allow us to build a uniform triangular mesh of a rectangular domain and the classic finite element operators associated with it. We have also provided and coded in `Matlab` an efficient algorithm for computing the hybrid operators arising from the discretisation of problems formulated using the FDM. The algorithm has two main steps. In the first step, it consists in computing the intersections of the immersed interface denoted γ with the main uniform triangulation, in order to create a new finer partition of γ. This partition has the property that each segment is included entirely in one triangle and in one element of γ. The second step of the algorithm consists in visiting each segment of this new partition and computing the corresponding contribution of the fictitious projection operator for the current segment associated to a unique pair of elements from the triangulation of the rectangle and the immersed interface.

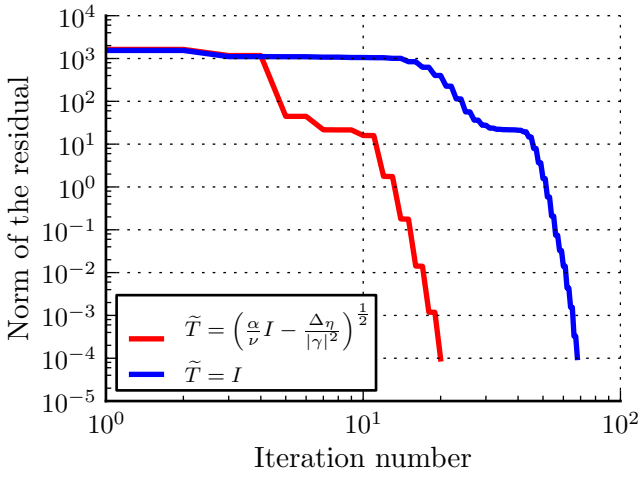


Figure 4.5: Plot of the norm of the residual after the n-th iteration of the MINRES iteration algorithm, for preconditioner $\widetilde{T}$ derived in Equation (4.57). The plot of the norm of the residual when using no preconditioner is plotted for comparison. The mesh size chosen here is $h = \frac{1}{256}$.

On an other note, we have presented the widely used iterative algorithms MINRES and GMRES for solving large sparse linear systems. Following the method suggested by Glowinski for deriving a suitable preconditioner for the saddle-point linear system arising from the Dirichlet FDM, we have derived a suitable preconditioner approximating the Schur complement in the Neumann case. We have tested this preconditioner and results show that it is optimal, since the number of iterations required to achieve the specified tolerance by the user does not depend on the size of the system. Furthermore, it turned out that the same preconditioner also appears to be optimal for the linear system deriving from the Robin FDM.

In the next chapter, we introduce the biological background of cell chemotaxis and review the literature. We will then compare two possible mathematical models for finite-sized cells – as opposed to point-like cells – undergoing chemotaxis in two simple settings.

Chapter 5

Comparison of two cell models for eukaryotic chemotaxis

5.1 Introduction

In the first chapter of this dissertation, we presented the phenomenon of chemotaxis in cell biology. We also discussed important contributions to the mathematical and computational modelling of bacterial and eukaryotic chemotaxis.

In this chapter, we will present two mathematical models for the absorption of chemicals by a cell that has a finite size. We will show that these models, even though having different topological properties, can be matched in the simple situation where we only consider the absorption of one chemical species by one immobile spherical cell. In order to extend this result, we will examine the situation of a steadily moving cell in a pool of signalling chemical. After performing this analysis, we will be in a position to draw conclusions on the advantages and drawbacks of the use of each model.

5.1.1 Two models for finite-sized cells

Let u be the concentration field of a chemical undergoing diffusion in the space $\mathbb{R}^d$, with $d = 2$ or 3. We suppose that a circular cell sits at the centre of the domain, at $\boldsymbol{x} = 0$. We further assume that the chemical with concentration u reacts when in contact with the membrane of the cell: for instance, this can represent the binding of a chemical signal

on the membrane of a cell, activating a membrane protein receptor. This is one of the crucial steps in the signalling transduction mechanism.

Several modelling approaches can be considered to account for the phenomenon of binding with membrane receptors. We would like to compare two of those in this chapter: we have named them the 'boundary cell model' and the 'hoover cell model'. Those approaches are similar as they are based on modelling the cell as a finite-sized object, as opposed to other possible approaches in which the cell is assumed to be dimensionless, which we call 'point-like cell' models. Let us introduce those two signal detection models for finite-sized cells.

5.1.1.1 Boundary cell model

In this approach, the domain of definition for the chemical concentration is defined as the extracellular space *i.e.* the space outside the cell, a sphere (if $d = 3$) or a disk (if $d = 2$) of radius R. Inside the cell, the concentration of signalling chemical is *a priori* undefined. Since the chemical is assumed to be diffusing in the extracellular space, then at its concentration satisfies the following PDE:

$$-D\triangle u = 0 \quad \text{in } \mathbb{R}^d \setminus B(0,R), \tag{5.1}$$

where D is the diffusion constant of the chemical. At the membrane of the cell, the reaction of the chemical with the membrane receptors is accounted for in the boundary conditions. Let us suppose that the function g encapsulates the chemical reactions of the signalling chemical with the receptors on the membrane of the cell, then we use Robin boundary conditions to model the reactions that the chemical undergoes on the membrane of the cell. We assume that the receptors are continuously distributed on the membrane of the cell. Hence, the boundary condition modelling the interaction of the cell with the chemical reads:

$$-D\nabla u \cdot \boldsymbol{n} = g(u). \tag{5.2}$$

5.1.1.2 Hoover cell model

We could argue that the previous mathematical modelling approach for the system formed by a cell and a chemical signal is complicated. One reason is that the mathe-

matical domain of evolution for the chemical is quite complex. Another reason is that the reactive term is contained in the boundary condition, although the literature about reaction diffusion equations (RDE) is abundant when the reaction term is located in the right-hand side of the diffusion equation [97]. One way to simplify the boundary cell model would consist in transferring the reaction term in the boundary condition (5.2) to the equation itself by assuming that the chemical signal can also freely diffuse in the interior of the cell and that the reactions take place in the intracellular space. The PDE satisfied by the concentration of chemical signal is then:

$$-D\triangle u = f(u)\mathbb{1}_{B(0,R)} \quad \text{in } \mathbb{R}^d, \tag{5.3}$$

where f is a reactive term 'corresponding' to g in (5.2). The meaning of this denomination will be discussed later in this chapter.

Working with this last equation seems significantly more convenient than with Equations (5.1)-(5.2) because the domain of definition has been greatly simplified, however we have lost the requirement of flux through the membrane of the cell, in other words, we do not control the flux of chemical through the boundary of the cell anymore.

5.1.2 Objective

Our objective in this chapter is to determine in which way the two previous models for spherical (or discoid, if $d = 2$) cells in chemotaxis are similar. In the simple case where the chemical signal binds on the membrane and is depleted in the process, the chemical signal symbolised by A undergoes the following reaction:

$$A \xrightarrow{k} \emptyset, \tag{5.4}$$

where k is the reaction rate. The symbol $\emptyset$ only means that we are not interested in the products of the reaction. Implicitly, it is assumed that the number of receptors on the membrane is in excess, compared to the number of signalling molecules, since in the previous equation, the depletion of A is not limited by the concentration of receptors on the membrane. We also make the assumption that k is constant on the whole length of the membrane, which means that the receptors are uniformly distributed on the membrane of the cell. Although this is not a biologically realistic assumption, this will conveniently allow us to work analytically.

The boundary reactive term g becomes $g(u) = -ku$ and the volume reaction term $f(u) = -Ku$, where K is the 'corresponding' reaction rate. We assume that both parameters k and K are positive. The goal is to determine explicitly the relationship between k and K in various settings.

We will first examine in Section 5.2 the case of $d = 3$, then we will deal with the case $d = 2$. In both cases, we will be able to find an exact matching between the two parameters k and K. In Section 5.3, we will discuss the same problem, with making the assumption that the cell is now steadily moving in a known direction.

5.2 Case of one immobile cell

5.2.1 Non dimensionalisation

Before comparing those two models, we first carry out a non-dimensionalisation step by setting $r := \|\boldsymbol{x}\|/R$ and $u_0 = 1$, this leads to defining two dimensionless parameters:

$$\alpha = \left(\frac{KR^2}{D}\right)^{1/2} \qquad \text{and} \qquad \beta = \frac{kR}{D}. \tag{5.5}$$

Rewritting (5.1)-(5.2) and (5.3) leads to the two following non-dimensionalised PDEs. Droping the dashes for non-dimensionalised radius, we obtain for the boundary cell model:

$$\begin{aligned} -\nabla^2 u = 0, \quad &\text{in } \mathbb{R}^d \setminus B(0,1), \\ -\nabla u \cdot n + \beta u = 0, \quad &\text{on } r = 1; \end{aligned} \tag{5.6}$$

and for the hoover cell model:

$$-\nabla^2 u + \alpha^2 \mathbb{1}_{B(0,1)}\, u = 0, \quad \text{in } \mathbb{R}^d. \tag{5.7}$$

5.2.2 Matching in 3 dimensions

In this section, we treat the case $d = 3$. Because of the spherical symmetries in both problems (5.6) and (5.7), the solution only depends on the distance to the centre of the cell r. We denote by u_h the solution of the hoover cell model and u_b the solution of the boundary cell model.

5.2.2.1 Solution of (5.6)

Since we seek a bounded solution to (5.6), we obtain the following problem to solve:

$$\begin{aligned} -\nabla^2 u &= 0, \quad \text{in } \mathbb{R}^3 \setminus B(0,1), \\ u &\to 1, \quad \text{as } r \to +\infty, \\ -\nabla u \cdot n + \beta u &= 0, \quad \text{on } r = 1. \end{aligned} \tag{5.6}$$

Owing to the spherical symmetry of the problem, the solution only depends on the radius r. The general solution of the Poisson equation on the domain $\mathbb{R}^3 \setminus B(0,1)$ is:

$$u_b(r) = C_1 + \frac{C_2}{r}. \tag{5.8}$$

The constants of integration are determined using the boundary conditions: taking the limit $r \to 1$, $C_1 = 1$ and at $r = 1$, $C_2 + \beta(1 + C_2) = 0$, so $C_2 = -\frac{\beta}{1+\beta}$.

Hence, the solution of the stationary diffusion equation for the boundary cell model is:

$$u_b(r) = 1 - \frac{\beta}{1+\beta}\frac{1}{r}, \quad r > 1. \tag{5.9}$$

5.2.2.2 Solution of (5.7)

Looking only for bounded solution in $\mathbb{R}^3$, Equation (5.7) can be rewritten as:

$$u''(r) + \frac{2}{r}u'(r) + \alpha^2 \mathbb{1}_{r\leq 1}\, u(r) = 0 \quad \text{and} \quad u \to 1 \text{ as } r \to +\infty. \tag{5.10}$$

We will also impose the continuity of the solution and of its first derivative at $r = 1$:

$$[u] = 0 \quad \text{and } [u'] = 0 \quad \text{at } r = 1. \tag{5.11}$$

For $r > 1$, the reaction disappears and the solution is expressed as:

$$u_h^{r>1} = A_1 + \frac{A_2}{r}. \tag{5.12}$$

The two constants of integration can be determined using the boundary conditions and the continuity condition of the solution and its first derivative at $r = 1$. Taking the limit $r \to +\infty$, we immediately find that $A_1 = 1$.

For $r \leq 1$, by performing the change of unknown function $u(r) = v(r)/r$, we can write the general solution:

$$u_h^{r\leq 1} = \frac{1}{r}\left(B_1 \cosh(\alpha r) + B_2 \sinh(\alpha r)\right). \tag{5.13}$$

Since we are interested only in bounded solutions to (5.7), we set $B_1 = 0$. Now matching both parts of the solution and their first derivatives at $r = 1$ yields the two following linear equations, with unknowns A_2 and B_2:

$$\begin{array}{rcrcr} A_2 & - & \sinh(\alpha)B_2 & = & -1, \\ A_2 & + & (\alpha\cosh(\alpha) - \sinh(\alpha))B_2 & = & 0. \end{array} \tag{5.14}$$

Solving this systems yields:

$$A_2 = \frac{\tanh(\alpha)}{\alpha} - 1 \quad \text{and} \quad B_2 = \frac{1}{\alpha\cosh(\alpha)}. \tag{5.15}$$

Finally, we deduce the solution of (5.7):

$$u_h(r) = \left[1 - \left(1 - \frac{\tanh(\alpha)}{\alpha}\right)\frac{1}{r}\right]\mathbb{1}_{r>1} + \left[\frac{\sinh(\alpha r)}{\alpha\cosh(\alpha)r}\right]\mathbb{1}_{r\leq 1}. \tag{5.16}$$

Let us note that u as well as its derivative are continuous at $r = 1$, as required.

5.2.2.3 Parameter matching

As we explained in the introduction, we wish both solutions to match such that both u_h and u_b satisfy the same Robin 'boundary' condition at $r = 1$:

$$-\nabla u \cdot \boldsymbol{n} + \beta u = 0. \tag{5.17}$$

By construction, u_b in (5.9) satisfies the condition mentioned above. Let us find the relationship between α and β such that the condition in (5.17) holds for u_h in (5.16). We write:

$$-u_h'(1) + \beta u_h(1) = 0, \tag{5.18}$$

with β the only unknown of the equation, which yields:

$$\beta = \frac{\alpha}{\tanh(\alpha)} - 1 = h_3(\alpha). \tag{5.19}$$

The purpose of the subscript '3' in $h_3(\alpha)$ is to remember that this relationship has been derived for a three-dimensional problem. This relationship is one-to-one for any positive α and β, so we can write:

$$\alpha = h_3^{-1}(\beta). \tag{5.20}$$

We see that as $\alpha \to 0$, $\beta \sim \alpha^2/3$ and then $\beta \sim \alpha$ when $\alpha \to \infty$.

Because of the small number of parameters in this problem, the matching we have performed is equivalent to matching the two solutions in the extracellular domain, since they would satisfy the same PDE with the same boundary conditions.

5.2.3 Matching in 2 dimensions

In two dimensions ($d = 2$), if we wish to work with bounded densities of chemoattractant, it is necessary to bound the extracellular space. To maintain the spherical symmetry of the problem, we choose as our domain a disk of radius ρR, where R is the radius of the cell and ρ a dimensionless parameter strictly larger than 1. Again, our objective consists in matching two models in such a way that the densities solving for each PDE model satisfy the same condition at $r = 1$. We present here the two PDEs satisfied by u_h, the density profile for the hoover cell model and u_b, the density for the boundary cell model.

In the hoover cell model, the domain of u_h is $B(0, \rho) \subset \mathbb{R}^2$. We obtain the following PDE problem for u_h:

$$\begin{aligned} -\nabla^2 u + \alpha^2 \mathbb{1}_{B(0,1)}\, u = 0, &\quad \text{in } B(0,\rho), \\ u = 1, &\quad \text{at } r = \rho. \end{aligned} \tag{5.21}$$

Lastly, in the boundary cell model, the domain is $B(0, \rho) \setminus B(0, R)$ and u_b satisfies the following nondimensionalised PDE:

$$\begin{aligned} -\nabla^2 u = 0, &\quad \text{in } B(0,\rho) \setminus B(0,1), \\ u = 1, &\quad \text{at } r = \rho, \\ -\nabla u \cdot \boldsymbol{n} + \beta u = 0, &\quad \text{at } r = 1. \end{aligned} \tag{5.22}$$

Both (5.21) and (5.22) admit a closed form solution. The details of the actual derivation are addressed in the following paragraphs.

5.2.3.1 Solution of (5.21)

Recall the equation:

$$\begin{aligned} -\nabla^2 u + \alpha^2 \mathbb{1}_{B(0,1)}\, u = 0, \quad \text{in } B(0,\rho), \\ u = 1, \quad \text{at } r = \rho. \end{aligned} \tag{5.21}$$

As performed in the case $d = 3$, we will also impose a continuity condition on the solution and its first derivative at $r = 1$. Using the same symmetry argument as before, the equation reduces to:

$$u''(r) + \frac{1}{r} u'(r) - \alpha^2 \mathbb{1}_{B(0,1)}\, u = 0. \tag{5.23}$$

For $r > 1$, the general solution is:

$$u_h^{r>1}(r) = D_1 + D_2 \log(r). \tag{5.24}$$

where D_1 and D_2 are integration constants to be determined using the boundary condition and the continuity condition of the solution and its first derivative at $r = 1$. For $r \leq 1$, the general solution is:

$$u_h^{r\leq 1}(r) = E_1 I_0(\alpha r) + E_2 K_0(\alpha r), \tag{5.25}$$

where I_0 and K_0 are the zero-th order modified Bessel functions of the first and second kind, respectively. Immediately, using the boundary condition at $r = \rho$, we deduce: $D_1 = 1 - D_2 \log(\rho)$, and using the boundedness of the solution at $r = 0$, $E_2 = 0$.

Now, matching $u_h^{r>1}$ and $u_h^{r\leq 1}$ at $r = 1$ and their first derivatives yields the following linear system:

$$\begin{array}{rcrcl} \log(\rho) D_2 & + & I_0(\alpha) E_1 & = & 1 \\ D_2 & - & \alpha I_0'(\alpha) E_1 & = & 0. \end{array} \tag{5.26}$$

Solving this linear system, we compute the values of D_2 and E_1:

$$D_2 = \frac{1}{\log(\rho) + \frac{I_0(\alpha)}{\alpha I_0'(\alpha)}} \quad \text{and} \quad E_1 = \frac{1}{I_0(\alpha) + \alpha \log(\rho) I_0'(\alpha)}. \tag{5.27}$$

Finally, we find the solution u_h of the initial equation:

$$u_h(r) = \begin{cases} 1 + \dfrac{\log(r/\rho)}{\log(\rho) + \frac{I_0(\alpha)}{\alpha I_0'(\alpha)}}, & r > 1, \\ \dfrac{I_0(\alpha r)}{I_0(\alpha) + \alpha \log(\rho) I_0'(\alpha)}, & 0 \leq r \leq 1. \end{cases} \tag{5.28}$$

Let us note that the solution u, as well as its derivative, are continuous at $r = 1$, as required.

5.2.3.2 Solution of (5.22)

Let us reproduce here Equation (5.22):

$$\begin{aligned} -\nabla^2 u &= 0, \quad \text{in } B(0,\rho) \setminus B(0,1), \\ u &= 1, \quad \text{at } r = \rho, \\ -\nabla u \cdot n + \beta u &= 0, \quad \text{at } r = 1. \end{aligned} \tag{5.22}$$

The general solution of this equation is:

$$u_b(r) = F_1 + F_2 \log(r). \tag{5.29}$$

The boundary conditions will again allow us to determine the constants of integration. At $r = \rho'$, we deduce $F_1 = 1 - F_2 \log(\rho)$ and at $r = 1$, we find: $-F_2 + \beta(1 - F_2 \log(\rho)) = 0$. This yields:

$$F_1 = \frac{1}{1 + \beta \log(\rho)} \quad \text{and} \quad F_2 = \frac{\beta}{1 + \beta \log(\rho)}. \tag{5.30}$$

The solution to (5.22) is then:

$$u_b(r) = 1 + \frac{\log(r/\rho)}{\log(\rho) + \frac{1}{\beta}} \qquad r > 1. \tag{5.31}$$

5.2.3.3 Parameter matching

We are now able to express β in terms of α, by finding for which value of β, u_h satisfies the same Robin boundary condition as u_b at $r = 1$. We obtain:

$$\beta = \frac{\alpha I_0'(\alpha)}{I_0(\alpha)} = h_2(\alpha). \tag{5.32}$$

Even though we have chosen a bounded domain to carry out this matching, we notice that the relationship we obtain between α and β does not depend either on R or ρ, or their ratio. The map h_2 is one-to-one for any positive α and β, so we are able to write:

$$\alpha = h_2^{-1}(\beta). \tag{5.33}$$

This allows us to answer positively the question whether it is possible to find a value of α such that for any β, the hoover cell model can match the boundary cell model. When $\alpha \to 0$, then $\beta \sim \alpha^2/2$ and as $\alpha \to +\infty$, $\beta \sim \alpha$.

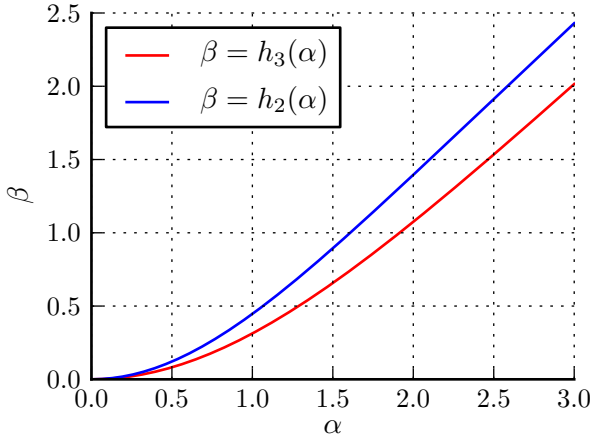


Figure 5.1: Plot of the relationship between the parameters α and β when the hoover cell model is matched with the boundary cell model.

5.2.4 Summary

We have been able to match the two models in the simple setting where one immobile cell sits at the centre of a domain with radial symmetry. The two relationships we obtain between the parameters α and β are plotted in Figure 5.1 for comparison. The functions h_2 and h_3 display the same behaviour in the limits of α near 0 and ∞. For illustration, we have plotted in Figure 5.2 the matching solution in the $d = 2$ and $d = 3$ cases for $\alpha = 3$.

5.3 One steadily moving cell

5.3.1 Slow cell approximation

Now, we wish to examine the case of a steadily moving cell, excited by the chemical signal. We suppose that this signal is carried by a fluid whose properties are similar to that of water. We suppose that the direction of the motion of the cell is constant in time.

Owing to the relative motion of the cell to the fluid carrying the chemical signal, then if we place ourselves in the reference frame of the cell, then the distribution of the concentration of the signal is not only influenced by the diffusion of the chemical in the

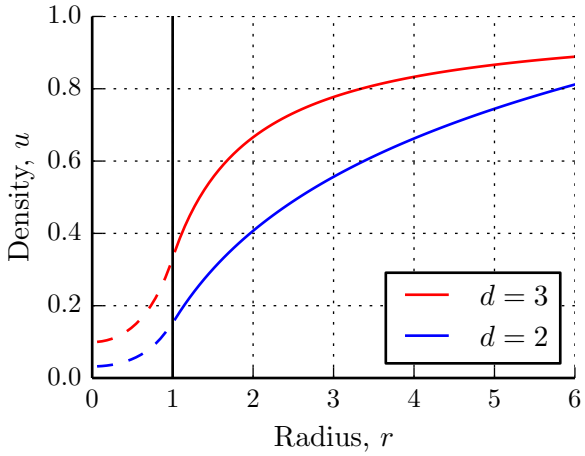


Figure 5.2: For $\alpha = 3$, we have plotted the density of chemoattractant for the two-dimensional and three-dimensional hoover cell models. For a matching $\beta = h_3(\alpha)$ or $\beta = h_2(\alpha)$, the density for the boundary cell model would exactly match that of the hoover cell model in the range $r \geq 1$, *i.e.* in the extracellular domain.

fluid, but also by the relative convection of the signal with respect to the cell. That is why we must investigate the relative importance of those two phenomena on the concentration field.

We consider the case of a *D. discoideum* cell excited by cyclic adenosine monophosphate (cAMP). *D. discoideum* have a typical radius $R = 10\mu$m [16], and a typical velocity of order $\bar{s} = 10^2 \mu\text{m.h}^{-1}$ [124]. Furthermore, the diffusion coefficient of cAMP is evaluated as $D = 10^5 \mu\text{m}^2\text{.h}^{-1}$ [42]. Those values lead us to compute the ratio between the magnitude of the diffusion and the transport of the chemical, induced by the motion of the cell. The Péclet number for this problem amounts to $\varepsilon := Pe = 10^{-2} \ll 1$.

5.3.2 Formulation of the three dimensional problem

5.3.2.1 Boundary cell model

In this model, the cell and the fluid bathing the chemical signal interact. The cell acts as an obstacle for the fluid. The velocity field is then non uniform and further assumptions must be made about the interaction of the fluid and the cell at the boundary of the cell.

We denote by $\varepsilon \boldsymbol{v}$ the velocity field created by the steady motion of the cell, where $\boldsymbol{v}$ is of magnitude 1. We suppose further that $\boldsymbol{v}$ is zero at the membrane of the cell. As performed in the previous section, we will only consider bounded solutions for the concentration field. However, as opposed to the previous case, we do not know *a priori* the concentration of the chemical signal at infinity. However, we can still impose that this concentration should be 1 at the 'inlet', which, in this situation, would be the point at infinity in the direction of the motion of the cell.

Assuming that the cell moves with constant velocity $\varepsilon(1,0,0)^\top$, we impose $u = 1$ when $r \to \infty$ and $\theta = 0$. The non dimensionalised PDE equation satisfied by the density of chemoattractant is a advection-diffusion equation equipped with Robin boundary conditions:

$$\begin{aligned} -\nabla^2 u + \varepsilon v \cdot \nabla u = 0, &\quad \text{in } \mathbb{R}^3 \\ -\nabla u \cdot n + \beta u = 0, &\quad \text{at } r = 1, \\ u \to 1, &\quad \text{as } r \to \infty \text{ and } \theta = 0. \end{aligned} \tag{5.34}$$

Due to the fact that further assumptions on the shape of the velocity-field are needed, we will not attempt to solve this equation, because the solution we would obtain would depend on the shape of the velocity field chosen. We will rather address the case of one steadily moving cell with the hoover cell model.

5.3.2.2 Hoover cell model

We assume that the motion of the cell does not affect the rest of the fluid, which implies that, in a stationary situation, the relative velocity of the fluid with respect to the cell (immobile in its own frame of reference) is uniform.

Similarly, we assume that the cell steadily travels at speed $\boldsymbol{v} = \varepsilon(1,0,0)^\top$, where $\varepsilon = Pe$, the Péclet number, therefore the relative velocity field convecting the concentration field is $\varepsilon(-1,0)^\top$. The concentration of the chemical signal hence satisfies the following (non dimensionalised) advection-diffusion-reaction equation:

$$\begin{aligned} -\nabla^2 u - \varepsilon \frac{\partial u}{\partial x} + \alpha^2 \mathbb{1}_{B(0,1)}\, u = 0, &\quad \text{in } \mathbb{R}^3 \\ u \to 1, &\quad \text{as } r \to \infty \text{ and } \theta = 0. \end{aligned} \tag{5.35}$$

We have made the same assumption on the concentration at infinity as for the boundary cell model.

5.3.2.3 Matching protocol

From the immobile cell case, we notice that the matching that has been performed does not really depend on the solution in the boundary cell model. Therefore, we will only develop a solution of the PDE satisfied by the concentration in the hoover cell model and later on, attempt to find a relationship between the binding rate β and the volume reaction rate α when the cell is moving steadily.

5.3.2.4 Solution of the Poisson problem (5.35)

In Appendix A, we provide the details of the derivation of the solution of the PDE problem (5.35). The first order correction u_1 finally is then written:

$$u_1(r,\theta) = -\begin{cases} \left(1-\dfrac{\tanh\alpha}{\alpha}\right)\dfrac{\sinh(\alpha r)}{2\alpha r\cosh(\alpha)}+ \\ \left(\dfrac{\sinh(\alpha r)}{2\alpha\cosh(\alpha)}+\dfrac{3}{2\sinh(\alpha)}\left(\dfrac{\sinh(\alpha r)}{(\alpha r)^2}-\dfrac{\cosh(\alpha r)}{\alpha r}\right)\right)\cos\theta, & r\leq 1 \\ \dfrac{1}{2}\left(1-\dfrac{\tanh\alpha}{\alpha}\right)\left(1-\left(1-\dfrac{\tanh\alpha}{\alpha}\right)\dfrac{1}{r}\right)+ \\ \left[\dfrac{1}{2}\dfrac{\tanh\alpha}{\alpha}-\dfrac{1}{2}+\left(\dfrac{1}{2}+\dfrac{3}{2\alpha^2}\left(1-\dfrac{\alpha}{\tanh(\alpha)}\right)\right)\dfrac{1}{r^2}\right]\cos\theta. & r>1 \end{cases} \tag{A.22}$$

This correction term is always negative. As a sanity check, we can observe that if we take α close to 0, then the correction term also tends to 0 uniformly with respect to α and decays as fast as α^2. We have plotted in Figure 5.3 the first order approximation to the solution of the steadily moving hoover cell problem.

5.3.2.5 Parameter matching

Now, this corrective term u_1 helps us to find a corrective β_1 such that:

$$-(u_0+\varepsilon u_1)'(1)+(\beta_0+\varepsilon\beta_1)(u_0+\varepsilon u_1)(1)=0, \tag{5.36}$$

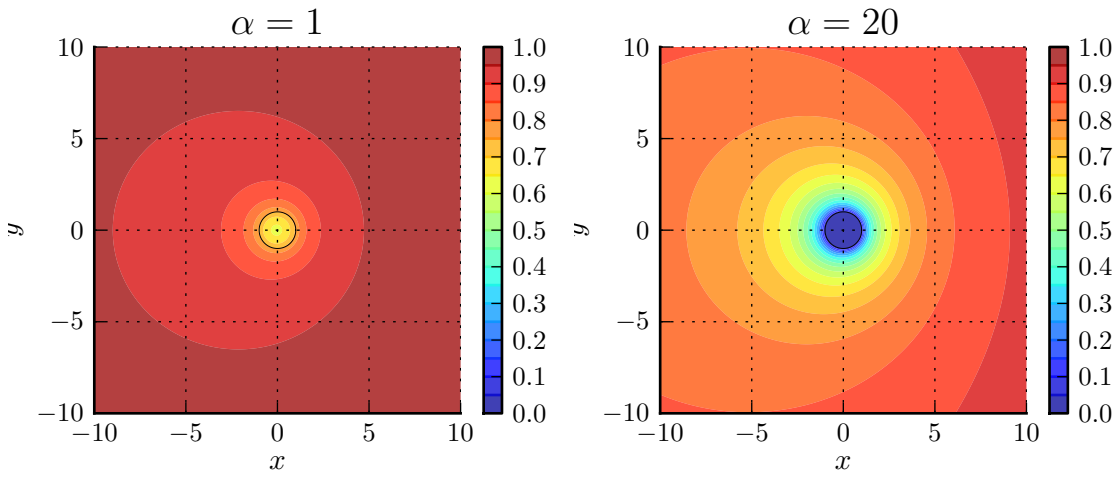


Figure 5.3: Plot of the first order approximation to the solution of the mobile hoover cell model for $\alpha = 1$ and $\alpha = 20$, with $\varepsilon = 0.1$. This is a projection in $d = 2$ of the $d = 3$ solution, which possesses a rotational symmetry around the x-axis.

where β_0 is the matching parameter computed in the immobile cell case. The only unknown in this equation being β_1, we find:

$$\beta_1 = \frac{u_1'(1) - \beta_0 u_1(1)}{u_0(1)}. \tag{5.37}$$

The expression for the first order correction is:

$$\beta_1 = -\frac{1}{2}\left(1 - \frac{\tanh(\alpha)}{\alpha}\right) - \frac{\alpha}{\tanh(\alpha)}\left(\frac{\tanh(\alpha)}{2\alpha} + \frac{3}{2}\left(\frac{1}{\tanh(\alpha)^2} - \frac{1}{\alpha^2} - 1\right)\right)\cos\theta. \tag{5.38}$$

First, we notice in this expression that all terms are bounded in α and θ. Moreover, the limit of β_1 when α tends to 0 is also 0 and we have $\beta_1 \sim -\frac{\alpha^2}{3}(\frac{1}{2} - \frac{1}{5}\cos\theta)$ near 0. When $\alpha \to +\infty$, β_1 is bounded and satisfies $\beta_1 \to -\frac{1}{2}(1 + \cos\theta)$. We have plotted in Figure 5.4 the correction term β_1 function of α and θ.

This expression shows that, since β_1 depends on the angle θ, the relationship between α and $\beta = \beta_0 + \varepsilon\beta_1$ is not independent of θ, so it cannot be inverted. This implies that, in the steadily moving cell case, the hoover cell model is no longer suitable for matching the boundary cell model. However, given that the dependence in θ only appears in the first order term, this dependence has a small impact on β as we can observe in Figure 5.4.

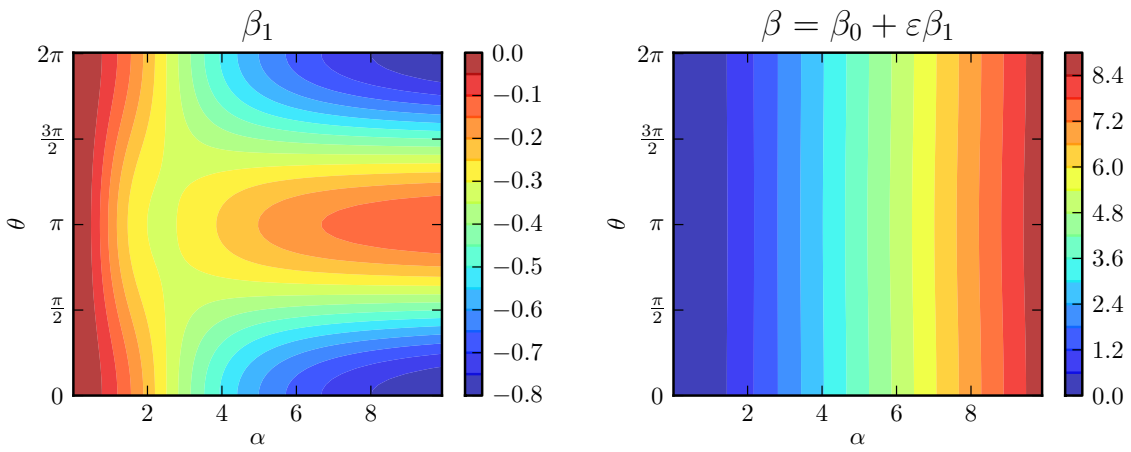


Figure 5.4: Plot of the corrective term β_1 function of α and the angle θ and of the first order approximation of β for the steady moving hoover cell case, for $\varepsilon = 0.1$.

5.4 Matching when β is θ-dependent

Let us return to the immobile cell setting examined in Section 5.2. Let us now suppose $\beta = \beta(\theta)$. In this situation, we are modelling a cell whose sensitivity to the external chemical signals is non uniform, *i.e.* whose receptors are not distributed evenly on its membrane. We are looking for a function $\alpha = \alpha(x, y)$ such the solution of the boundary cell model and that of the hoover model satisfy the same Robin boundary condition at the interface $r = 1$. For the boundary cell model, the density u should satisfy:

$$\begin{aligned} -\nabla^2 u &= 0, \quad \text{in } \mathbb{R}^d \setminus B(0,1), \\ -\nabla u \cdot \boldsymbol{n} + \beta(\theta) u &= 0, \quad \text{on } r = 1. \end{aligned} \tag{5.39}$$

And for the hoover cell model:

$$-\nabla^2 u + \alpha^2(x, y) \mathbb{1}_{B(0,1)}\, u = 0, \quad \text{in } \mathbb{R}^d. \tag{5.40}$$

In this setting, we notice that if a match can be found, then the solutions of each model will be identical for $r \geq 1$. Matching those two problems consists in determining $\alpha(x, y)$ as a function of $\beta(\theta)$, which is a much more complex inverse problem than those we have dealt with in the previous sections. In this case, we do not know whether a solution exists or not.

However, we point out that in the case that such a solution exists, it would only be valid for immobile cells, which would greatly limit the significance of that result. This is why

we have not investigated this problem, however it could be the object of future work. While not having a solution at our disposal, the boundary cell model becomes more valuable, because the hoover cell model cannot cope with a non uniform distribution of receptors on the membrane.

5.5 Discussion

In Section 5.2, we have shown that it is possible to match the boundary cell model and the hoover cell model, when one immobile cell is present, and assuming binding on the membrane of the cell with rate k.

In Section 5.3, we have seen that, since the velocity of a large cell (such as that of an eukaryotic cell) is relatively low, the Péclet number comparing the magnitude of the advection and diffusion terms in a reaction-advection-diffusion PDE is small. This allows the use of an asymptotic method for deriving an approximate solution of the hoover cell model, when the cell is moving steadily in a fixed direction and compute a correction term for the relationship between α and β.

Since we have demonstrated that the correction term β_1 is bounded, it seems reasonable to conclude that, for small values of ε, using only the zero-th order matched rate β_0, the hoover cell model decently approximates the boundary cell model. In other words, in the absence of more modelling details, the hoover cell model can be sufficient for approximating the behaviour of the boundary cell model, even when cells are mobile in the computational domain. Moreover, the relationship we have obtained between α and β is one-to-one, and allows the modeller to scale the effective volume rate α useful for a simulation to the 'real' rate, β, that can potentially be measured *in vitro*. The hoover cell model, due to the simplicity of its computer implementation and its low computational cost, can be considered as a good choice.

However, the assumptions made in Section 5.2 are quite restrictive. It has been established, as we have seen in Chapter 1, that the differential of bound receptors concentration on the cell membrane is a key component in the process of gradient detection in eukaryotic cells. Thus we realise that the simple binding reaction $g(u) = -ku$ becomes biologically inaccurate for modelling the binding on membrane receptors, because it is

implicitly assumed that the receptors are uniformly distributed on the membrane of the cell.

In order to be able to incorporate more precise biological details, the use of the boundary cell model seems to be unavoidable, as it can consistently link the extracellular and intracellular dynamics of the signalling transduction process. Due to the relatively complex geometry of the boundary cell model, we can expect the computational cost of the simulation of such a model to be more substantial than that of the hoover cell model.

5.6 Summary

In this chapter, we have presented two models for modelling finite-sized cells undergoing chemotaxis: the boundary cell model and the hoover cell model. The boundary cell approach models the absorption of chemical signals by cells more accurately than the hoover cell model, because in the boundary cell approach, the membrane of the cell is modelled, whereas in the hoover cell model, the chemical signal can diffuse freely in the volume occupied by the cell and its absorption is modelled by a volume depletion term. The accuracy of the boundary cell model is at the expense of the mathematical complexity of the domain of definition of the density of the chemical signal.

We have found that it is possible in a simple setting (one immobile cell) to exactly match the boundary cell and hoover cell models. However, when the cell becomes mobile, the matching becomes only one-way: we managed to express β as a function of α and the angle θ. This relationship has a limited use because it does not allow the matching of the most complex model *i.e.* the boundary cell model, to the simplest one, but only the opposite.

In the next chapter, a numerical method for simulating the movement of unicellular organisms that respond to chemical signals is presented. Cells are modelled as objects of finite size using the boundary cell model presented in this chapter, while the extracellular space is described by reaction-diffusion PDEs. This modular simulation allows the implementation of different models at the different scales encountered in cell biology and couples them in one single framework. The global computational cost is contained

thanks to the use of the FDM that we have presented in Chapter 3, allowing the efficient solve of PDEs in moving domains.

Chapter 6

A new simulation framework for simulating chemotaxis

6.1 Introduction

In this chapter, we present an approach to simulating the chemotaxis of cells that is capable of describing the system at both intracellular and extracellular levels. The main feature consists in modelling cells as finite-sized non-deformable[1] solid objects while adopting a continuous approach to model the evolution of the concentration of extracellular chemicals. The simulation framework we present here also allows us to implement the modelling of chemical dynamics at three levels in a modular way: intracellular level, at the interface between the intracellular and the extracellular domains and in the extracellular space with possible feedbacks and effects on the motion of cells.

The main challenge consists in making this simulation computationally affordable whilst not compromising on the level of details: this includes differentiating the extracellular and intracellular spaces and modelling the cell membrane governing the exchanges between the bulk and the cytosol. Hence, instead of a volume depletion (or production) term $f(c, n)$, the absorption (or emission) of chemoattractant by the cells will appear as a boundary condition at the cells boundary.

Those requirements lead us to solve a reaction-diffusion equation similar to (1.2) on a

[1] This is not a requirement, since the framework we present would also work with deformable cells keeping their area constant. However, considering only non-deformable cells simplifies the presentation of our scheme.

deforming domain (the extracellular domain) due to the motion of the cells. In order to make this cost acceptable, our computational model takes advantage of the FDM for finite elements that allow us to avoid using costly remeshing and interpolations between grids between two time steps. A mixed method is employed that allocates specific degrees of freedom to the flux of chemicals, allowing us to better monitor the exchanges between the cytosol and the bulk. The advantages of adopting our approach can be summarised as follows:

- The modelled cells have a finite size, and are not point particles.
- The membrane of cells is represented mathematically; this allows the accurate modelling of exchanges between the bulk and the cytosol.
- Our simulation framework is highly modular, allowing us to couple easily different models of different nature.
- The computational cost of the simulation is contained due to the FDM. Furthermore, fast numerical techniques can be applied for solving linear systems resulting from the discretisation of PDEs on regular grids.

This chapter is organised in the following way: in Section 6.2, we introduce the various objects of the modelling and write the mathematical formulation of their evolution. In Section 6.3, we will discuss the computation of fictitious domain specific matrices. Finally in Section 6.4, we will actually implement the scheme developed in the earlier sections and show some results when simulating a simplified chemotactic system.

6.2 Mathematical formulation

6.2.1 Notation and domain definition

Let us define a cavity Ω, a bounded subset of $\mathbb{R}^2$, with a smooth boundary $\partial\Omega$. This domain Ω will be our computational domain, including both the interior and the exterior of the modelled cells. For ease, Ω is taken as a rectangle with dimension $L_1 \times L_2$.

Let $C_i(t)$ be the interior of the i-th modelled cell, for $i = 1, \ldots, N$, where N is the total number of modelled cells (see Figure 6.1). Hence, at any instant t, the cellular domain

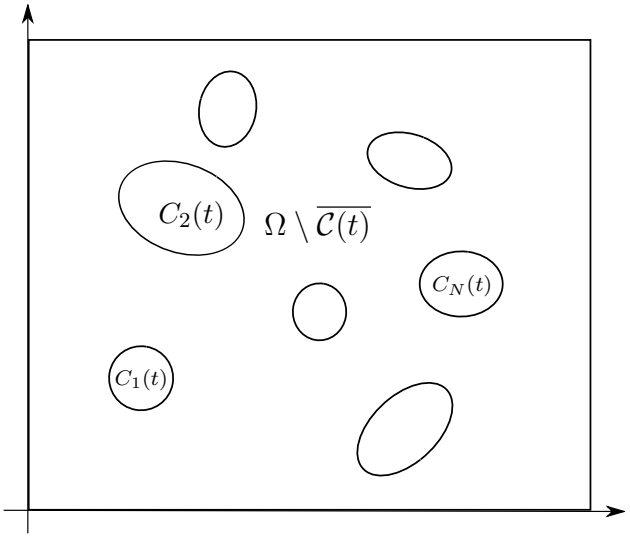


Figure 6.1: Representation of the cavity Ω and cells $C_i(t)$.

is $\mathcal{C}(t) = \bigcup_{j=1}^{N} C_j(t)$ and the extracellular space $\Omega \setminus \overline{\mathcal{C}(t)} = \Omega \setminus \bigcup_{j=1}^{N} \overline{C_j(t)}$. The interface $\mathcal{I}(t) = \bigcup_{j=1}^{N} \partial C_j(t)$ is the frontier between these two. We also denote $\mathcal{I}_j(t) = \partial C_j(t)$. Furthermore, we assume that the modelled cells are bounded and convex[2].

Let $\boldsymbol{u} = (u_1, u_2, \ldots, u_p)^\top$ with $p \geq 1$ be the vector of densities of p signalling chemicals. The vector $\boldsymbol{u}$ is defined in the spatio-temporal domain $\mathcal{G}$ formed by the successive shapes of the extracellular domain in time, that is to say

$$\mathcal{G} = \{(\boldsymbol{x}, t) \in \Omega \times [0, T] \ : \ \boldsymbol{x} \notin \overline{\mathcal{C}(t)}\}. \tag{6.1}$$

Let $\boldsymbol{x}_i(t)$, for $i = 1, \ldots, N$ be the position of the centre of mass of cell i at time t, and $\theta_i(t)$ its angle with respect to the horizontal axis (see Figure 6.2). We also denote $\boldsymbol{v}_i(t)$ its velocity and by $\boldsymbol{\omega}_i(t)$ its angular speed. Given that two cells cannot overlap at any instant, the set of positions of the centres of mass of the cells can be written as $\{\{\boldsymbol{x}_i\} \subset \Omega^N$ such that $\text{dist}(C_i(t), C_j(t)) > 0$, $\forall (i,j)$, $i \neq j\}$. The distance between two cells is defined as follows:

$$\text{dist}(C_i(t), C_j(t)) = \min_{(x,y) \in C_i(t) \times C_j(t)} \|x - y\|. \tag{6.2}$$

Enforcing this non-overlapping condition is non trivial. We opted for an artificial short range force, that tends to keep cell away from each other, see Appendix B.

[2]This is not a strict requirement, as long as the local curvature of the boundary of the cell is not too large.

Finally, we introduce the variable $\boldsymbol{y}_i = (y_1, y_2, \ldots, y_q)^\top \in \mathbb{R}^q$ describing the internal processes undergone in each cell [53]. They can be of different nature, for instance chemical or mechanical.

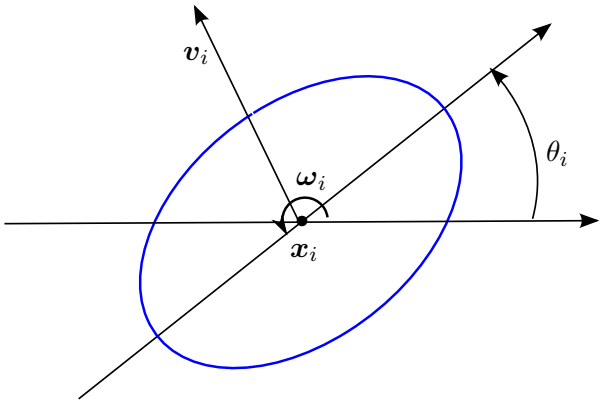


Figure 6.2: Kinematic parameters characterising the movement of a cell of a given shape (here elliptic).

6.2.2 Governing strong equations

6.2.2.1 Signalling chemicals

We describe the evolution in time and space of the density of signalling chemicals $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x}, t)$ by a system of p coupled reaction-advection-diffusion equations. The density of the i-th signalling chemical satisfies the following PDE:

$$\begin{aligned} \partial_t u_i + \boldsymbol{v} \cdot \nabla u_i = \;& d_i \triangle u_i + f_i(\boldsymbol{u}), && \text{in } \mathcal{G}, \\ -d_i \nabla u_i \cdot \boldsymbol{n} = \;& g_i(\boldsymbol{u}), && \text{on } \mathcal{I}(t), \forall t \in [0, T], \\ u_i(\cdot, 0) = \;& u_{i0}, && \text{in } \Omega \setminus \mathcal{C}(0), \end{aligned} \tag{RDE}$$

where the symbol d_i represents the diffusion constant of the i-th chemical.

The functions $f_i : \mathbb{R}^p \to \mathbb{R}$ and $g_i : \mathbb{R}^p \to \mathbb{R}$, respectively model the chemical interactions between the p species in the extracellular space and on the membrane of the cells. *A priori*, they could be both nonlinear in $\boldsymbol{u}$.

This reaction-diffusion equation is very similar to (1.2), except that the influence of the cells does not appear as a volume depletion term, but as a reaction term in the boundary conditions, at the interface between the cells and the extracellular space. Moreover, we

notice the presence of an advection term, which takes care of the conservation of mass in the system.

The rate of change of the density u_i is written in a Eulerian frame of reference and the velocity field $\boldsymbol{v}$ characterises the deformation of the extracellular domain. Additional information about the evolution of $\boldsymbol{v}$ is necessary, since the velocity field is only known at the boundary of the cells, that is $\boldsymbol{v}_i + \boldsymbol{\omega}_i \times (\cdot - \boldsymbol{x}_i)$, $\boldsymbol{v}_i$ being the velocity of the centre of gravity of cell i, $\boldsymbol{\omega}_i$ its angular speed. When decreasing the size of the cells to zero, we expect that the range over which the modulus of velocity field is significant tends to zero as well, so the advection term completely disappears when the cells become point particles.

In Equation (RDE), no boundary condition on $\partial\Omega$ has been provided. This is because the boundary conditions are imposed by the biological situation to model.

6.2.2.2 Velocity field

We assume the extracellular environment can be considered as a Newtonian fluid with viscosity ν.

The extracellular domain deforms in time, however, since the area of each cell remains constant in time, the area of the extracellular space is also constant. Therefore the velocity field $\boldsymbol{v}$ is divergence-free.

We assume that the velocity field $\boldsymbol{v}$ satisfies a Stokes equation, with Dirichlet boundary conditions at the boundary of each cell and at the boundary of the cavity Ω:

$$\begin{aligned} \nu\triangle\boldsymbol{v} - \nabla p = &\ 0 && \text{in } \mathcal{G}, \\ \nabla\cdot\boldsymbol{v} = &\ 0 && \text{in } \mathcal{G}, \\ \boldsymbol{v} = &\ \boldsymbol{v}_j + \boldsymbol{\omega}_j \times (\cdot - \boldsymbol{x}_j) && \text{on } \partial C_j(t), \forall t \in [0,T] \\ & && \forall j \in \{1,\dots,N\}. \end{aligned} \tag{STO}$$

Equation (STO) is defined on the same spatio-temporal domain as Equation (RDE). Computationally, this will prove to be of importance later on when discretising the domain $\mathcal{G}$. Given the dynamic viscosity of water at 25°C at the micrometer scale, the viscosity term would be dominant in an unsteady Stokes equation, over time scales of more than a second. That justifies neglecting the material derivative term and therefore

the use of a steady Stokes equation. Moreover, since we have made no assumption on the regularity in time of the forcing at the interface with cells, we could not guarantee the well-posedness of the unsteady Stokes equations in this case.

Note that the Dirichlet data at the boundary of each cell satisfies the compatibility condition:

$$\int_{\partial C_j(t)} \boldsymbol{v} \cdot \boldsymbol{n} \, \mathrm{d}s = 0, \quad \forall j \in \{1, \dots, N\}, \tag{6.3}$$

where the normal $\boldsymbol{n}$ points outwards of $\mathcal{G}$, *i.e.* inwards each cell.

6.2.2.3 Description of the motion of cells

To close the description of the whole system {cells & chemicals}, it is necessary to provide equations of motion for each cell. Let Δt be some time step, then we write following set of Langevin equations [128, 26], $\forall j \in \{1, \dots, N\}$:

$$\boldsymbol{x}_j(t + \Delta t) = \boldsymbol{x}_j(t) + \boldsymbol{v}_j(t)\Delta t, \tag{6.4a}$$

$$\boldsymbol{v}_j(t + \Delta t) = \boldsymbol{v}_j(t) + \beta(\boldsymbol{F}_j - \boldsymbol{v}_j(t))\,\Delta t + \beta\sqrt{2D_i\Delta t}\;\xi, \tag{6.4b}$$

$$\boldsymbol{\omega}_j(t + \Delta t) = \boldsymbol{\omega}_j(t) + \boldsymbol{G}_j\,\Delta t, \tag{6.4c}$$

where

$$\boldsymbol{F}_j = \boldsymbol{F}_j(t, \{\boldsymbol{x}_k(t), \theta_k(t)\}_j, \{\boldsymbol{v}_k(t), \boldsymbol{\omega}_k(t)\}_k, \boldsymbol{y}_j)$$

and

$$\boldsymbol{G}_j = \boldsymbol{G}_j(t, \{\boldsymbol{x}_k(t), \theta_k(t)\}_k, \{\boldsymbol{v}_k(t), \boldsymbol{\omega}_k(t)\}_k, \boldsymbol{y}_j)$$

are, respectively, the force and torque terms acting on each cell. They can include, as well, repulsive forces allowing us to enforce the non-overlapping condition of cells [93] (see Appendix B). The parameter D_j represents the diffusion constant of cell j, and β is a scaling parameter. Finally, ξ is a sample from the normal distribution $N(0, 1)$.

Remark 6.1. As $\beta \to \infty$, then Equation (6.4a) becomes:

$$\boldsymbol{x}_j(t + \Delta t) = \boldsymbol{x}_j(t) + \boldsymbol{F}_j\,\Delta t + \sqrt{2D_j\Delta t}\;\xi, \tag{6.4a'}$$

which can be recognised as a stochastic differential equation (Langevin equation), where $\boldsymbol{F}_j$ is the drift and $\sqrt{2D_j\Delta t}$ the additive noise term.

When simultaneously taking the number N of cells to ∞ and their size to 0, one can recover the partial differential equation satisfied by the density of cells in the Keller-Segel model (1.1), by choosing $\boldsymbol{F}_j = \boldsymbol{F} = \chi(u)\nabla u$.

Remark 6.2. When $D_j = 0$ and as $\Delta t \to 0$, then Equation (6.4b) simply turns into Newton's second law in classical mechanics.

$$\frac{\mathrm{d}\boldsymbol{x}_j}{\mathrm{d}t}(t) = \boldsymbol{v}_j(t), \tag{6.4a"}$$

$$\frac{\mathrm{d}\boldsymbol{v}_j}{\mathrm{d}t}(t) = \boldsymbol{F}_j. \tag{6.4b"}$$

The framework introduced in this section for describing the motion of the cells is very general and allows a wide range of configurations. Other models are also possible, that do not assume a Gaussian behaviour of the random term. In [53, 20], the cells velocity changes by turns randomly, following a Poisson jump process.

6.2.2.4 Internal dynamics

Following [52], we present a general *black-box* model for the evolution of the vector of internal variables $\boldsymbol{y}$:

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t}(t) = \boldsymbol{I}(t, \boldsymbol{u}, \boldsymbol{y}) \text{ with } \boldsymbol{y}(0) = \boldsymbol{y}_0. \tag{6.5}$$

The function $\boldsymbol{I}$ encapsulates all the causes of variations for $\boldsymbol{y}(t)$. This function *a priori* depends in a nonlinear fashion on the density of chemicals $\boldsymbol{u}$ in the extracellular space and on the current state of internal variables $\boldsymbol{y}(t)$. Spatial effects are not included in the description of the evolution of the internal states.

Conservation of mass ; Intracellular to extracellular exchanges: The vector of internal variables $\boldsymbol{y}$ can also include the same chemical species as in the extracellular space. In that case, the function $\boldsymbol{I}$ can be written as: $\boldsymbol{I}(t, \boldsymbol{u}, \boldsymbol{y}) = \int_{\mathcal{I}(t)} \boldsymbol{g}(\boldsymbol{u})\, \mathrm{d}\gamma + \boldsymbol{I}^\star(t, \boldsymbol{y})$, where $\boldsymbol{I}^\star$ is a purely reactive term. This ensures that mass is conserved during transfers between the intracellular and the extracellular spaces.

6.2.3 Weak formulations and the FDM

6.2.3.1 Choice of a fictitious domain

As recalled in the introduction, the FDM is well adapted to the derivation of weak formulations of time-dependent PDEs that are posed on domains changing shape in time. The approach consists in extending the original domain to a larger simple-shaped domain on which regular meshes can be fitted. The boundary conditions are then imposed through the use of Lagrange multipliers, as we have seen in Chapter 3.

In the next paragraphs, we will derive convenient weak formulations taking advantage of the FDM in order to eliminate of the time-dependent domain $\Omega \setminus \overline{C(t)}$.

A natural choice for the simple-shaped extended domain of the FDM is the whole rectangular cavity Ω. To avoid complicated notation, the extensions to Ω of all unknowns will be denoted by the same symbol as their original counterpart.

6.2.3.2 Weak form for the reaction-advection-diffusion equation (RDE)

The approach we are adopting is the FDM for Robin problems, discussed in Chapter 3. Indeed, the boundary conditions at the boundary of the cells is of Robin type, therefore we introduce the new variable $\boldsymbol{\sigma}_i$ defined by:

$$\boldsymbol{\sigma}_i = -d_i \nabla u_i \quad \text{in } \mathcal{G}. \tag{6.6}$$

The following weak formulation is derived for $\boldsymbol{\sigma}_i \in H^{\text{div}}(\Omega)$ by applying the method discussed in 3.4.

In the original domain, the solution of the reaction-advection-diffusion equation should lie in $H^1(\mathcal{G}(t))$. However, this property cannot be conserved when extending the original domain to Ω, because the extended solution u_i will have a jump at the interface, so this extended solution will only lie in a broken space $H^1_B(\Omega)$ defined as follows:

$$H^1_B(\Omega) = \left\{ u \in L^2(\Omega) : \ u|_{\Omega\setminus\overline{\mathcal{C}}} \in H^1(\Omega \setminus \overline{\mathcal{C}}) \ \& \ u|_{\mathcal{C}} \in H^1(\mathcal{C}) \right\}. \tag{6.7}$$

However, in this formulation, the normal component of the flux, $\boldsymbol{\sigma} \cdot \boldsymbol{n}$, across the interfaces $\mathcal{I}(t) = \bigcup_{j=1}^{N} \partial C_j(t)$ will remain continuous.

This fact shows that this formulation allows us to simultaneously control the flux of chemicals through the interface (representing the membrane of the cells) and conserve mass of chemicals in the system { bulk & cells }. This would not be possible using a FDM based on Lagrange elements, which does not provide separate degrees of freedom for the local density and gradient.

The fictitious weak formulation of Equation (6.6) is:

$$\int_\Omega d_i^{-1}\boldsymbol{\sigma}_i \cdot \boldsymbol{\tau}\ \mathrm{d}\Omega + \int_\Omega u_i\, \nabla\cdot\boldsymbol{\tau}\ \mathrm{d}\Omega + \sum_{j=1}^{N}\langle \lambda_i, \boldsymbol{\tau}\cdot\boldsymbol{n}\rangle_{\partial C_j(t)} = 0, \quad \forall \tau \in H^{\mathrm{div}}(\Omega), \tag{6.8}$$

where $\lambda_i \in H^{1/2}(\mathcal{I}(t))$ represents the jump of u_i across the interface $\mathcal{I}(t) = \bigcup_{j=1}^{N} \partial C_j(t)$, that is to say:

$$\lambda_i(\cdot) = \lim_{\epsilon\to 0}\left[u_i(\cdot + \epsilon\boldsymbol{n}(\cdot)) - u_i(\cdot - \epsilon\boldsymbol{n}(\cdot))\right]. \tag{6.9}$$

Then the weak formulation of Problem (RDE) is written:
find $(u_i(\cdot,t), \boldsymbol{\sigma}_i(\cdot,t), \lambda_i(\cdot,t)) \in L^2(\Omega)\times H^{\mathrm{div}}(\Omega)\times H^{1/2}(\mathcal{I}(t))$ such that $\forall (w,\boldsymbol{\tau},\mu) \in L^2(\Omega)\times H^{\mathrm{div}}(\Omega)\times H^{1/2}(\mathcal{I}(t))$,

$$\int_\Omega d_i^{-1}\boldsymbol{\sigma}_i \cdot \boldsymbol{\tau}\ \mathrm{d}\Omega + \int_\Omega u_i\, \nabla\cdot\boldsymbol{\tau}\ \mathrm{d}\Omega + \sum_{j=1}^{N}\langle \lambda_i, \boldsymbol{\tau}\cdot\boldsymbol{n}\rangle_{\partial C_j(t)} = 0, \tag{6.10a}$$

$$\int_\Omega d_i^{-1}\boldsymbol{\sigma}_i \cdot (\boldsymbol{v}w) + \nabla\cdot\boldsymbol{\sigma}_i\, w\ \mathrm{d}\Omega + \int_\Omega \partial_t u_i\, w\ \mathrm{d}\Omega = \int_\Omega f(\boldsymbol{u})\, w\ \mathrm{d}\Omega, \tag{6.10b}$$

$$\sum_{j=1}^{N}\langle \mu, \boldsymbol{\sigma}_i\cdot\boldsymbol{n} - g(\boldsymbol{u})|_{\mathcal{G}(t)}\rangle_{\partial C_j(t)} = 0. \tag{6.10c}$$

In Equation (6.10b), $\boldsymbol{v}$ has been extended to Ω so this motivates the usage of the fictitious domain formulation for Problem (STO) in Ω.

6.2.3.3 Weak form for the Stokes equation (STO)

The approach we adopt here applies the FDM for elliptic Dirichlet problems by Glowinski [68], as we have described it in Section 3.2. We extend the velocity field $\boldsymbol{v}$ to the whole of Ω, by applying the FDM to every component of $\boldsymbol{v}$, each of which satisfying an elliptic Dirichlet problem. The Dirichlet boundary conditions at the boundary of each cell are then imposed using a Lagrange multiplier $\boldsymbol{\xi}$.

Other methods are possible, for instance, in [71, 70], by imposing rigid-body motion to the fluid 'filling' the objects modelling the cells. The Lagrange multiplier imposing this motion is then distributed in the volume occupied by the cells. This is convenient when one wants to implement a two-way coupling between the solid objects and the fluid. In our case, since the cells are not inert objects but self-propelled, the action of the fluid on the cells has no relevance. We will therefore assume that the motion of cells is **not** influenced by viscous drag due to the fluid.

The fictitious weak formulation of Problem (STO) is written:
find $(\boldsymbol{v}(\cdot,t), p, \boldsymbol{\xi}(\cdot,t)) \in H^1_0(\Omega)^2 \times L^2_0(\Omega) \times H^{-1/2}(\mathcal{I}(t))^2$ such that $(\boldsymbol{w}, q, \boldsymbol{\zeta}) \in H^1_0(\Omega)^2 \times L^2_0(\Omega) \times H^{-1/2}(\mathcal{I}(t))^2$,

$$\int_\Omega \nu \boldsymbol{\nabla v} : \boldsymbol{\nabla w} \, \mathrm{d}\Omega - \int_\Omega p \, \nabla \cdot \boldsymbol{w} \, \mathrm{d}\Omega + \sum_{i=1}^{N} \langle \boldsymbol{w}, \boldsymbol{\xi} \rangle_{\partial C_i(t)} = 0, \tag{6.11a}$$

$$\int_\Omega q \, \nabla \cdot \boldsymbol{v} \, \mathrm{d}\Omega = 0, \tag{6.11b}$$

$$\sum_{i=1}^{N} \langle \boldsymbol{v} - (\boldsymbol{v}_i + \boldsymbol{\omega}_i \times (\cdot - \boldsymbol{x}_i)), \boldsymbol{\zeta} \rangle_{\partial C_i(t)} = 0, \tag{6.11c}$$

where ':' represents the double-dot product. In Equation (6.11a), since the velocity field $\boldsymbol{v}$ is divergence-free and in $H^1_0(\Omega)^2$, then we can take advantage of the fact that:

$$\int_\Omega \nu \boldsymbol{\nabla v} : \boldsymbol{\nabla w} \, \mathrm{d}\Omega = \int_\Omega 2\nu \boldsymbol{D}[\boldsymbol{v}] : \boldsymbol{D}[\boldsymbol{w}] \, \mathrm{d}\Omega, \quad \forall \boldsymbol{w} \in H^1_0(\Omega)^2, \tag{6.12}$$

where $\boldsymbol{D}[\boldsymbol{v}] = \frac{1}{2}(\boldsymbol{\nabla v} + \boldsymbol{\nabla v}^\top)$ is the rate-of-strain tensor.

The Lagrange multiplier $\boldsymbol{\xi}$ can be interpreted as the jump of the stress tensor $-p\boldsymbol{n} + 2\nu \boldsymbol{D}[\boldsymbol{v}](\boldsymbol{n})$ across the interface $\partial C_i(t)$, where $\boldsymbol{n}(\cdot)$ is the local outward normal at the interface $\mathcal{I}(t)$.

In this formulation, we can see that the extended solution $\boldsymbol{v}$ is expected to be continuous across the interfaces $\partial C_i(t)$, for all i. It will therefore be continuous on $\overline{\Omega}$, hence it will be an element of $L^\infty(\Omega)$. This ensures that the advection term of Equation (6.10b) is well-defined.

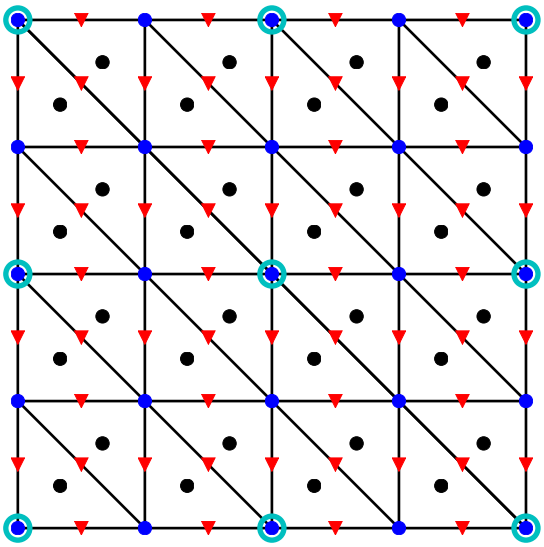


Figure 6.3: Representation of the location of the degrees of freedom for the variables involved in the simulation framework. The degrees of freedom for the velocity field are shown in blue dots, the pressure nodes are represented by cyan circles, that for the densities by black dots, and that for the density fluxes by red triangles.

6.3 Computational scheme

6.3.1 Spatial discretisation

The main advantage of the FDM is to allow the use of regular meshes. We opt for a triangulation $\mathcal{T}_h$ obtained from a rectangular mesh of $(2k_1 - 1) \times (2k_2 - 1)$ vertices, by subdividing each square into two triangles with their diagonal of negative slope. We also assume that the mesh size satisfies $h = \frac{L_1}{2k_1-1} = \frac{L_2}{2k_2-1}$.

We define a self-similar triangulation with mesh size $2h$, labelled $\mathcal{T}_{2h}$. Figure 6.3 shows the two triangulations. This triangulation will be used to host the degrees of freedom for the velocity and pressure variables. The motivation for using a coarser triangulation for the discretisation of the Stokes problem is that we would like to keep in mind the purpose of the Stokes flow: generating a velocity field enforcing the conservation of mass condition in (RDE). If we used the same triangulation as for the discretisation of the reaction-diffusion equation, then we believe that the induced computational cost would

be disproportionate compared to the benefits we would gain from obtaining a more precise velocity field.

Furthermore, the domain formed by the union of the interfaces $\mathcal{I}(t) = \bigcup_{j=1}^{N} \mathcal{I}_j(t)$ must also be discretised. Maintaining the notation consistent with that of previous chapters, let η the mesh size of the discretised interface. The boundary of each modelled cell i is divided into n_i segments of length η_i such that $\frac{1}{2}\eta \leq \eta_i \leq \eta$, for $1 \leq i \leq N$. If all the modelled cells have the same dimensions, then all η_i will be equal to η. The parameter η must be chosen coarser than the resolution chosen for the discretisation of the cavity, as we have seen in Chapter 3.

When refining the main mesh *i.e.* reducing h, we will have to reduce η by the same factor, therefore making η dependent of h. For that reason, the approximation spaces for the Lagrange multipliers instead of bearing a η subscript will bear a h subscript instead. We denote by $\mathcal{I}_h$ the set of segments forming the discretised interface. We can then define the approximation spaces linked to these triangulations.

6.3.1.1 Approximation spaces for Problem (6.10)

As we have seen in Chapter 3, for Problem (6.10), we use linear Raviart-Thomas triangular elements $\mathcal{RT}_0(\Omega)$ for approximating the flux variable, piecewise constant elements for the concentration variable and piecewise constant elements for the Lagrange multiplier space:

$$\boldsymbol{V}_h = \left\{ \boldsymbol{\tau}_h \in H^{\mathrm{div}}(\Omega) \cap C(\overline{\Omega})^2 \mid \forall K \in \mathcal{T}_h,\ \boldsymbol{\tau}_h|_K \in \mathcal{RT}_0(K) \right\}, \tag{6.13a}$$

$$M_h = \left\{ q \in L^2(\Omega) \cap C(\overline{\Omega}) \mid \forall T \in \mathcal{T}_h,\ q|_T \in \mathcal{P}_0(T) \right\}, \tag{6.13b}$$

$$\Lambda_h = \left\{ \lambda \in H^{1/2}(\mathcal{I}(t)) \cap C(\overline{\mathcal{I}(t)}) \mid \forall I \in \mathcal{I}_h,\ \lambda|_I \in \mathcal{P}_0(I) \right\}. \tag{6.13c}$$

6.3.1.2 Approximation spaces for Problem (6.11)

For the Stokes problem (6.11), we use the well-known Hood-Taylor discretisation [125], consisting of quadratic degrees of freedom for the velocity and continuous piecewise linear degrees of freedom for the pressure. This finite element formulation is well-known

for being stable [15] satisfying the required discrete inf-sup condition [14]. The Lagrange multiplier space is discretised in piecewise (discontinuous) constant elements, as performed in Section 3.2:

$$\boldsymbol{W}_h = \left\{ \boldsymbol{w}_h \in H_0^1(\Omega)^2 \cap C(\overline{\Omega})^2 \mid \forall T \in \mathcal{T}_h,\ \boldsymbol{w}_h|_T \in \mathcal{P}_2(T) \times \mathcal{P}_2(T) \right\}, \tag{6.14a}$$

$$Q_h = \left\{ q \in L_0^2(\Omega) \cap C(\overline{\Omega}) \mid \forall T \in \mathcal{T}_{2h},\ q|_T \in \mathcal{P}_1(T) \right\}, \tag{6.14b}$$

$$\boldsymbol{\Xi}_h = \left\{ \boldsymbol{\zeta} \in H^{-1/2}(\mathcal{I}(t))^2 \mid \forall I \in \mathcal{I}_h,\ \boldsymbol{\zeta}|_I \in \mathcal{P}_0(I) \times \mathcal{P}_0(I) \right\}. \tag{6.14c}$$

6.3.2 Semi-discrete equations

Using the definition of the finite element approximation spaces in the previous section, we write the finite element formulation of (RDE) and (STO).

6.3.2.1 FE formulation of the reaction-advection-diffusion equation (RDE)

Find $(u_{i,h}(\cdot,t), \boldsymbol{\sigma}_{i,h}(\cdot,t), \lambda_{i,h}(\cdot,t)) \in M_h \times \boldsymbol{V}_h \times \Lambda_h$ such that $\forall (w_h, \boldsymbol{\tau}_h, \mu_h) \in M_h \times \boldsymbol{V}_h \times \Lambda_h$,

$$\int_\Omega d_i^{-1} \boldsymbol{\sigma}_{i,h} \cdot \boldsymbol{\tau}_h \,\mathrm{d}\Omega + \int_\Omega u_{i,h} \nabla \cdot \boldsymbol{\tau}_h \,\mathrm{d}\Omega + \sum_{j=1}^{N} \int_{\mathcal{I}_j(t)} \lambda_{i,h} \left(\boldsymbol{\tau}_h \cdot \boldsymbol{n} \right) \mathrm{d}\mathcal{I}_j(t) = 0, \tag{6.15a}$$

$$\int_\Omega d_i^{-1} \boldsymbol{\sigma}_{i,h} \cdot (\boldsymbol{v} w_h) + \nabla \cdot \boldsymbol{\sigma}_{i,h}\, w_h \,\mathrm{d}\Omega + \int_\Omega \partial_t u_{i,h}\, w_h \,\mathrm{d}\Omega = \int_\Omega f_i(\boldsymbol{u}_h)\, w_h \,\mathrm{d}\Omega, \tag{6.15b}$$

$$\sum_{j=1}^{N} \int_{\mathrm{d}\mathcal{I}_j(t)} \mu_h \left(\boldsymbol{\sigma}_{i,h} \cdot \boldsymbol{n} - g_i(\boldsymbol{u}_h) \right) \mathrm{d}\mathcal{I}_j(t) = 0. \tag{6.15c}$$

6.3.2.2 FE formulation of the Stokes equation (STO)

Find $(\boldsymbol{v}_h(\cdot,t), p_h(\cdot,t), \boldsymbol{\xi}_h(\cdot,t)) \in \boldsymbol{W}_h \times Q_h \times \boldsymbol{\Xi}_h$ such that $(\boldsymbol{w}_h, q_h, \boldsymbol{\zeta}_h) \in \boldsymbol{W}_h \times Q_h \times \boldsymbol{\Xi}_h$,

$$\int_\Omega \nu \boldsymbol{\nabla} \boldsymbol{v}_h : \boldsymbol{\nabla} \boldsymbol{w}_h \,\mathrm{d}\Omega - \int_\Omega p_h \nabla \cdot \boldsymbol{w}_h \,\mathrm{d}\Omega + \sum_{j=1}^{N} \int_{\mathcal{I}_j(t)} \boldsymbol{w}_h \cdot \boldsymbol{\xi}_h \,\mathrm{d}\mathcal{I}_j(t) = 0, \tag{6.16a}$$

$$\int_\Omega q_h \nabla \cdot \boldsymbol{v}_h \,\mathrm{d}\Omega = 0, \tag{6.16b}$$

$$\sum_{j=1}^{N} \int_{\mathcal{I}_j(t)} \left(\boldsymbol{v}_h - \left(\boldsymbol{v}_j + \boldsymbol{\omega}_j \times (\cdot - \boldsymbol{x}_j) \right) \right) \cdot \boldsymbol{\zeta}_h \,\mathrm{d}\mathcal{I}_j(t) = 0. \tag{6.16c}$$

Here, the numerical solution of (6.16), $\boldsymbol{v}_h$, is an element of $H_0^1(\Omega)^2$. However, when looking carefully at Equation (6.15), it would be more convenient to work with the projection of the flow in $H^{\text{div}}(\Omega)$, so a flow in $\boldsymbol{W}_h = \mathcal{RT}_0(\Omega)$. In this case, we can use the classical projection operator [115] from $H^1(\Omega)^2$ to $\boldsymbol{W}_h$ defined as follows:

$$\pi_h(\boldsymbol{v}_h) = \sum_{e \in \mathcal{E}} \left(\int_e \boldsymbol{v}_h \cdot \boldsymbol{\tau}_e \, \mathrm{d}s \right) \boldsymbol{\tau}_e, \tag{6.17}$$

where $\boldsymbol{\tau}_e \in \boldsymbol{W_h}$ satisfies $\boldsymbol{\tau}_e \cdot \boldsymbol{n}_e = 1$ on e.

6.3.3 Temporal discretisation scheme for (RDE)

Let $\Delta t > 0$ be the constant time step used for the temporal discretisation of Problems (RDE), (STO) and (6.4). We denote the value at time $t_k = k\Delta t$ of any quantity q_h as $q_h(k\Delta t) = q_h^k$. Since we have discretised the initial problem (RDE) in space only, this implies we are adopting a method of lines approach [121]. We will integrate Equation (6.15) with respect to time; since the computational domain Ω does not depend on time, the integrals in time and space commute: this is the main benefit we gain from using the FDM.

Our choice of the temporal discretisation scheme must be motivated by taking into account the two usual constraints: the requirements of stability and accuracy of the scheme. As a conservative choice, we will choose a first order implicit scheme in time, which is known to improve the stability of the numerical scheme, following Glowinski [68]. We thus write, for any quantity $q(t)$ and $k \geq 1$,

$$\int_{t_{k-1}}^{t_k} q(t) \, \mathrm{d}x = \Delta t \; q(t_{k+1}) + O(\Delta t^2). \tag{6.18}$$

We also consider the use of an operator splitting method to decouple the enforcement of the conservation of mass constraint (advection term) and the enforcement of the boundary conditions at the boundary of modelled cells. Operator splitting methods are widely used to solve reaction-advection-diffusion problems [78] because they are able to cope with various difficulties such as advection dominated problems or stiff reaction term, by splitting the problem into simpler components and addressing each of them individually.

We propose a simple first-order accurate two-step operator-splitting scheme to compute the update $u_{i,h}^{k+1}$ from the previously computed $u_{i,h}^{k}$. First, we compute the update $u_{i,h}^{k+1/2}$ from an advection step and then $u_{i,h}^{k+1}$ using the intermediary update $u_{i,h}^{k+1/2}$ from the diffusion step.

Given $u_{i,h}^{k}$, find $(u_{i,h}^{k+1/2}, \boldsymbol{\sigma}_{i,h}^{k+1/2}) \in M_h \times \boldsymbol{V}_h$ such that $\forall (w_h, \boldsymbol{\tau}_h) \in M_h \times \boldsymbol{V}_h$,

$$\int_\Omega \epsilon^{-1} \boldsymbol{\sigma}_{i,h}^{k+1/2} \cdot \boldsymbol{\tau}_h \, \mathrm{d}\Omega + \int_\Omega u_{i,h}^{k+1/2} \nabla \cdot \boldsymbol{\tau}_h \, \mathrm{d}\Omega = 0, \tag{6.19a}$$

$$\int_\Omega \epsilon^{-1} \boldsymbol{\sigma}_{i,h}^{k+1/2} \cdot (\boldsymbol{v} w_h) + \nabla \cdot \boldsymbol{\sigma}_{i,h}^{k+1/2} \, w_h \, \mathrm{d}\Omega + \int_\Omega \frac{1}{\Delta t} u_{i,h}^{k+1/2} \, w_h \, \mathrm{d}\Omega = \int_\Omega \frac{1}{\Delta t} u_{i,h}^{k} \, w_h \, \mathrm{d}\Omega, \tag{6.19b}$$

where ϵ is small penalty parameter to stabilise the scheme [83, 85]. In the first step, no boundary condition need to be imposed since there is no flow in or out of the cavity Ω since, by definition, we have set $\boldsymbol{v} = 0$ on $\partial\Omega$. Moreover, since there is no convective flux in or out the cells, no boundary condition is imposed at the interface between cells and the extracellular space. Hence the advection step is taken in the whole of Ω.

The second step of the operator splitting scheme is:

$$\int_\Omega d_i^{-1} \boldsymbol{\sigma}_{i,h}^{k+1} \cdot \boldsymbol{\tau}_h \, \mathrm{d}\Omega + \int_\Omega u_{i,h}^{k+1} \nabla \cdot \boldsymbol{\tau}_h \, \mathrm{d}\Omega + \sum_{j=1}^{N} \int_{\mathcal{I}_j(t)} \lambda_{i,h}^{k+1} \, (\boldsymbol{\tau}_h \cdot \boldsymbol{n}) \, \mathrm{d}\mathcal{I}_j(t) = 0, \tag{6.20a}$$

$$\int_\Omega \nabla \cdot \boldsymbol{\sigma}_{i,h}^{k+1} \, w_h \, \mathrm{d}\Omega + \int_\Omega \frac{1}{\Delta t} u_{i,h}^{k+1} \, w_h \, \mathrm{d}\Omega = \int_\Omega \left(\frac{1}{\Delta t} u_{i,h}^{k+1/2} + f_i(\boldsymbol{u}_h^{k+1}) \right) w_h \, \mathrm{d}\Omega, \tag{6.20b}$$

$$\sum_{j=1}^{N} \int_{\mathrm{d}\mathcal{I}_j(t)} \mu_h \left(\boldsymbol{\sigma}_{i,h}^{k+1} \cdot \boldsymbol{n} - g_i(\boldsymbol{u}_h^{k+1}) \right) \mathrm{d}\mathcal{I}_j(t) = 0. \tag{6.20c}$$

This is the reaction-diffusion step, taken in the fictitious domain Ω, imposing the reactive boundary conditions at the interface between the cells and the extracellular domain.

6.3.3.1 Matrix form of the discretised equations

We adopt the following notations for the matrix operators:

$$A_\sigma = \left(\int_\Omega \boldsymbol{\tau}_i \cdot \boldsymbol{\tau}_j \, \mathrm{d}\Omega \right), \tag{6.21a}$$

$$D = \left(\int_\Omega \nabla \cdot \boldsymbol{\tau}_i \, w_j \, \mathrm{d}\Omega \right), \tag{6.21b}$$

$$L^k = \left(\sum_{j=1}^{N} \int_{\mathcal{I}_j^k} (\boldsymbol{\tau}_i \cdot \boldsymbol{n}) \, \mu_j \, \mathrm{d}\mathcal{I}_j^k \right), \tag{6.21c}$$

$$A_u = \left(\int_\Omega w_i \, w_j \, \mathrm{d}\Omega \right) \tag{6.21d}$$

$$C^k = \left(\int_\Omega \boldsymbol{\tau_i} \cdot (\boldsymbol{v}^k w_j) \, \mathrm{d}\Omega \right). \tag{6.21e}$$

We notice that only two of those matrix operators are time-dependent: the discretised advection operator C^k and the fictitious operator L^k. We also observe that since we are using piecewise constant elements for the approximation of the concentration variable u_i, then the matrix A_u is diagonal and its coefficient on the diagonal are the area of each triangle, that is to say $h^2/2$.

The matrix form of the linear systems of the two-step operator splitting scheme (6.19)-(6.20) can be formulated as:

$$\begin{bmatrix} \epsilon^{-1} A_\sigma & D \\ D^\top + \epsilon^{-1} C^{k+1^\top} & -\frac{1}{\Delta t} A_u \end{bmatrix} \begin{bmatrix} \Sigma_i^{k+1/2} \\ U_i^{k+1/2} \end{bmatrix} = \begin{bmatrix} 0 \\ -\frac{1}{\Delta t} A_u U_i^k \end{bmatrix}, \tag{6.22a}$$

$$\begin{bmatrix} d_i^{-1} A_\sigma & D & (L^k) \\ D^\top & -\frac{1}{\Delta t} A_u & 0 \\ L^{k^\top} & 0 & 0 \end{bmatrix} \begin{bmatrix} \Sigma_i^{k+1} \\ U_i^{k+1} \\ \Lambda_i^{k+1} \end{bmatrix} = \begin{bmatrix} 0 \\ -\frac{1}{\Delta t} A_u U_i^{k+1/2} + A_u F(U_i^{k+1}) \\ P_h^\top G(U_i^{k+1}) \end{bmatrix}, \tag{6.22b}$$

and equivalently:

$$\left(\epsilon^{-1} (A_\sigma + \frac{2\Delta t}{h^2} D^\top C^{k+1^\top}) + \frac{2\Delta t}{h^2} D D^\top \right) \Sigma_i^{k+1/2} = D^\top U_i^k, \tag{6.23a}$$

$$U_i^{k+1/2} = U_i^k + \frac{2\Delta t}{h^2} (D^\top + \epsilon^{-1} C^{k+1^\top}) \Sigma_i^{k+1/2}, \tag{6.23b}$$

$$\begin{bmatrix} d_i^{-1} A_\sigma + \frac{2\Delta t}{h^2} D D^\top & L^k \\ L^{k^\top} & 0 \end{bmatrix} \begin{bmatrix} \Sigma_i^{k+1} \\ \Lambda_i^{k+1} \end{bmatrix} = \begin{bmatrix} D^\top (U_i^{k+1/2} + \Delta t F(U_i^{k+1})) \\ P_h^\top G(U_i^{k+1}) \end{bmatrix}, \tag{6.23c}$$

$$U_i^{k+1} = U_i^{k+1/2} + \frac{2\Delta t}{h^2} D^\top \Sigma_i^{k+1}. \tag{6.23d}$$

Each element of this system is either explicit (Equations (6.23b) and (6.23d)) or easier to invert than the original system (Equations (6.23a) and (6.23c)). In particular, for Equation (6.23c), we have derived a suitable preconditioner in Chapter 4. In the case of general nonlinear reaction terms F, classical methods for the numerical approximation of solutions of nonlinear ODEs can be used such as a Runge-Kutta method [127], or further splitting can be used, in order to isolate the reaction term.

The linearised system (6.16) is also formulated as:

$$\begin{bmatrix} \nu A_v & B & X^{k+1} \\ B^\top & 0 & 0 \\ {X^{k+1}}^\top & 0 & 0 \end{bmatrix} \begin{bmatrix} V^{k+1} \\ P^{k+1} \\ \Xi^{k+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ W^{k+1} \end{bmatrix}, \tag{6.24}$$

where the matrix blocks A_v, B and X^k are defined by:

$$A_v = \left(\int_\Omega \boldsymbol{\nabla} \boldsymbol{w}_i : \boldsymbol{\nabla} \boldsymbol{w}_j \, \mathrm{d}\Omega \right), \tag{6.25a}$$

$$B = \left(\int_\Omega q_j \, \nabla \cdot \boldsymbol{w}_j \, \mathrm{d}\Omega \right), \tag{6.25b}$$

$$X^k = \left(\sum_{j=1}^{N} \int_{\mathcal{I}_j^k} \boldsymbol{\mu}_i \cdot \boldsymbol{w}_j \, \mathrm{d}\mathcal{I}_j^k \right). \tag{6.25c}$$

Remark 6.3. Since the boundaries of all the modelled cells are discretised using a distinct set of degrees of freedom for each, the preconditioning of the blocks L^k and X^k is straightforward. The Schur complement of the linear system (6.23c) is a block diagonal matrix formed of the approximation of the Schur complements for the linear system derived from the discretisation of a Robin elliptic PDE by the FDM with only one obstacle, as seen in Chapter 4. The form of the preconditioner is:

$$P_{\mathrm{RDE}} = \begin{bmatrix} \tilde{A} & & & & \\ & \widetilde{T} & & & \\ & & \widetilde{T} & & \\ & & & \ddots & \\ & & & & \widetilde{T} \end{bmatrix}, \tag{6.26}$$

where $\widetilde{T}$ is constructed as in Equation (4.57). The approximation of the block $d_i^{-1} A_\sigma + \frac{2\Delta t}{h^2} D D^\top$ is usually performed by algebraic multigrid methods, as discussed in [113]. Moreover, the construction of the approximation of the Schur complement needs only to be computed once, because it only depends on the boundary mesh of one cell, independently of the triangular mesh.

The same principle applies for the preconditioning of the linear system derived from the Stokes equation: we use the approximation of the Schur complement for the Dirichlet FDM discussed in Chapter 4 and the preconditioner for the linear system of the discretised Stokes flow equation is:

$$P_{\text{Stokes}} = \begin{bmatrix} \tilde{A}_\nu & & & & & \\ & M_p & & & & \\ & & \widetilde{V} & & & \\ & & & \widetilde{V} & & \\ & & & & \ddots & \\ & & & & & \widetilde{V} \end{bmatrix}. \tag{6.27}$$

The choice of suitable blocks $\widetilde{A}_\nu$ and M_p has been extensively discussed in the literature and [48, 45] by Elman *et al.* provide optimal block preconditioners for linear systems derived from the Stokes equation.

6.4 Implementation and tests

In this section, we are going to apply the computational method discussed in the previous section to a simplified chemotactic system, first with only one cell and later on, with 14 cells. We will show some numerical results that give empirical proof of the stability of the computational method presented here.

6.4.1 Numerical simulation of the motion of single cell subject to chemotaxis

The objective of this test problem is to simulate the motion of a single amoeboid *D. discoideum* cell C with boundary ∂C, undergoing chemotaxis in a periodic domain along the x-axis, by a single signalling chemical, cAMP, subject only to diffusion.

This chemoattractant is also absorbed by the cell, which determines the direction of the motion by measuring the concentration of chemoattractant on its membrane. Arbitrarily, the shape of the cell is chosen to be elliptic with major and minor axis a and $b = a/2$, respectively.

The concentration of cAMP, $u := u_1$, satisfies (RDE) with $f(u) = 0$ and $g(u) = -ku$, where k is the rate of absorption of the chemoattractant by the cell membrane.

6.4.1.1 Cartoon model for internal variables

The model we adopt for the internal variables of the cell is the 'cartoon model' from Erban and Othmer [52, 54], capturing the main dynamics of the signal transduction process for amoeboid cell organisms such as *Dictyostelium discoideum*. This linear models comprises two variables y_1 and y_2 depending on the location on the membrane of the cell $s \in [0, 2\pi)$ and t, the time. They satisfy:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} t_e\, y_1 \\ t_a\, y_2 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + u(s,t)\begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{6.28}$$

The parameter t_a described the period of adaptation of the cell to the signal u. Roughly speaking, the variable $y_2(s,t)$ follows the signal $u(s,t)$ with a lag of t_a, whereas in the limit of $t_e \ll t_a$, $y_1(s,t)$ follows the difference between $u(s,t)$ and y_2, that is to say: $y_1(s,t) = u(s,t) - y_2(s,t)$.

Erban and Othmer then introduce the quantity $\boldsymbol{q}$ defined by:

$$\boldsymbol{q} = \frac{1}{|\partial C|}\int_0^{2\pi} y_2(s,t)\,\boldsymbol{n}(s)\,\mathrm{d}s, \tag{6.29}$$

representing a directional average of the interval variable y_2. Vector $\boldsymbol{n}$ represents the outwards normal of the cell. Furthermore, they show that $\boldsymbol{q}$ satisfies:

$$t_a\frac{\mathrm{d}\boldsymbol{q}}{\mathrm{d}t} = \frac{1}{|\partial C|}\int_0^{2\pi} u(s,t)\,\boldsymbol{n}(s)\,\mathrm{d}s - \boldsymbol{q}, \tag{6.30}$$

meaning that $\boldsymbol{q}$ follows the direction of the signal $\frac{1}{|\partial C|}\int_0^{2\pi} u(s,t)\,\boldsymbol{n}(s)\,\mathrm{d}s$ with a lag of t_a. The directional information $\boldsymbol{q}$ on the gradient of the signal detected by the cell is used to model the force propelling the cell:

$$\boldsymbol{F} = \frac{\bar{s}\boldsymbol{q}}{\varepsilon_s + \|\boldsymbol{q}\|}, \tag{6.31}$$

where $\bar{s}$ represents the typical speed of the cell and ε_s the sensitivity of the cell to the detected gradient.

6.4.1.2 Equations of motion of the cell

The cell motion is governed by the simplest version of Equations (6.4), that is to say, without random term. The velocity and the axis angle of the cell satisfy:

$$t_d \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = \boldsymbol{F} - \boldsymbol{v}, \tag{6.32a}$$

$$t_d \frac{\mathrm{d}\theta}{\mathrm{d}t} = \arg[\boldsymbol{q}] - \theta(t), \tag{6.32b}$$

where t_d is a relaxation time. The first equation has been proposed by Erban and Othmer in [54]. The second equation translates in mathematical terms the fact that the cell has a tendency to align its axis in the direction of the motion.

This system models the trend for a cell to move towards area of higher concentration of the chemoattractant cAMP. It is almost identical to the Keller-Segel sensitivity term, but has no singularity when there is no detected signal by the cell.

The position of the cell and its angle at time $t + \Delta t$ are then given by:

$$\boldsymbol{x}(t + \Delta t) = \boldsymbol{x}(t) + \Delta t\, \boldsymbol{v}(t, \boldsymbol{y}), \tag{6.33a}$$

$$\theta(t + \Delta t) = \theta(t) + \frac{\Delta t}{t_d} \left(\arg[\boldsymbol{q}] - \theta(t)\right). \tag{6.33b}$$

The model employed for the angle here is minimal. It simply describes the trend to the cell to align with the direction of the movement. This phenomenon has been observed experimentally [94].

6.4.1.3 Results

Having described entirely the system with evolution equations, we now list in Table 6.1 below the values of the parameters used in the simulation.

Since we have no exact solution for comparison with our simulation results, we compute the L^2-distance between the concentration field obtained for a given resolution and the one obtained with the finest resolution in our numerical experiments ($h_{\mathrm{ref}} = 1/512\,\mathrm{mm}$ and $\Delta t_{\mathrm{ref}} = 0.05\,\mathrm{min}$). This can give us empirical proof about the stability of the method. The distance between two approximations is computed in the following way:

$$\|u_h - u_{h_{\mathrm{ref}}}\|_{L^2(L^2(\Omega);[0,T])} = \left(\sum_{k=0}^{T} \|u_h(k\Delta t_{\mathrm{ref}}) - u_{h_{\mathrm{ref}}}(k\Delta t_{\mathrm{ref}})\|_{0,2,\Omega}^2 \right)^{1/2}. \tag{6.34}$$

Parameter	Symbol	Value/Magnitude	Source
Computational domain	Ω	$(0, 1.5)\,\text{mm} \times (0, 0.25)\,\text{mm}$	–
Number of chemical species	p	1	–
Number of cells	N	1	–
Simulation time	T	$10\,\text{min}$	–
Signal diffusivity	$d = d_1$	$1.6 \times 10^{-2}\,\text{mm}^2.\text{min}^{-1}$	[42]
Cell dimension	$2a$	$1.5 \times 10^{-2}\,\text{mm}$	[16]
Typical cell speed	$\bar{s}$	$1 - 10 \times 10^{-2}\,\text{mm.min}^{-1}$	[124, 101]
Sensitivity of the cell to gradients	ε_s	$10^{-4}\,\text{mm}^{-1}$	[54]
Excitation time by signal	t_e	$0\,\text{sec}$	[54]
Adaptation time to signal	t_a	$0.5\,\text{min}$	[54]
Relaxation time	t_d	$2\,\text{min}$	[54]

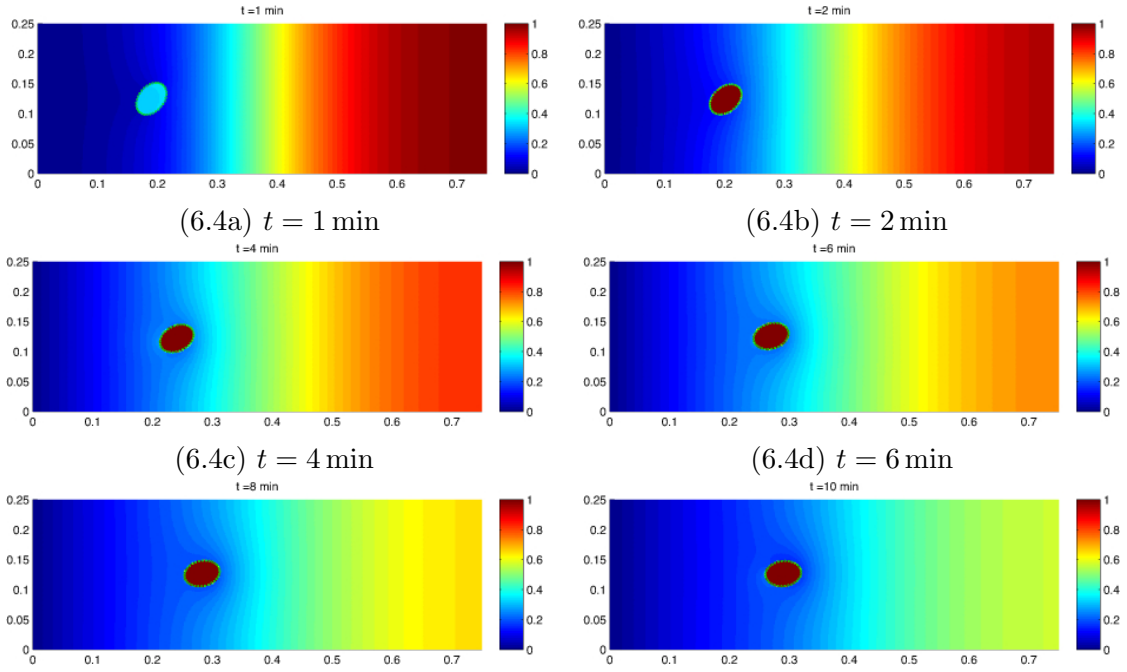
Table 6.1: Table showing the values for the parameters of the problem. The first three parameters were arbitrarily chosen for the purpose of the test. The remaining values are taken from the literature (last column).

L^2-distance	$\Delta t = 0.4$	$\Delta t = 0.2$	$\Delta t = 0.1$
$h = 1/64$	0.91	0.53	0.31
$h = 1/128$	0.60	0.39	0.23
$h = 1/256$	0.28	0.18	0.12

Table 6.2: Evaluation of the L^2 stability of the numerical scheme. The resolutions are expressed in millimetres and the time steps in minutes.

Pictures showing a sample run of the simulation can be found in Figure 6.4 at different times.

The results shown in Table 6.2 display the stability of the method, since the measured distance between the reference concentration field and the concentration fields computed with coarser time and space resolutions decrease by at least a factor of 2 when both the space and time resolution are divided by 2. This suggests that the simulation framework is not only stable, but also that the global scheme is first order accurate in $h + \Delta t$, as expected from the chosen numerical schemes.



(6.4a) $t = 1\,\text{min}$ (6.4b) $t = 2\,\text{min}$

(6.4c) $t = 4\,\text{min}$ (6.4d) $t = 6\,\text{min}$

(6.4e) $t = 8\,\text{min}$ (6.4f) $t = 10\,\text{min}$

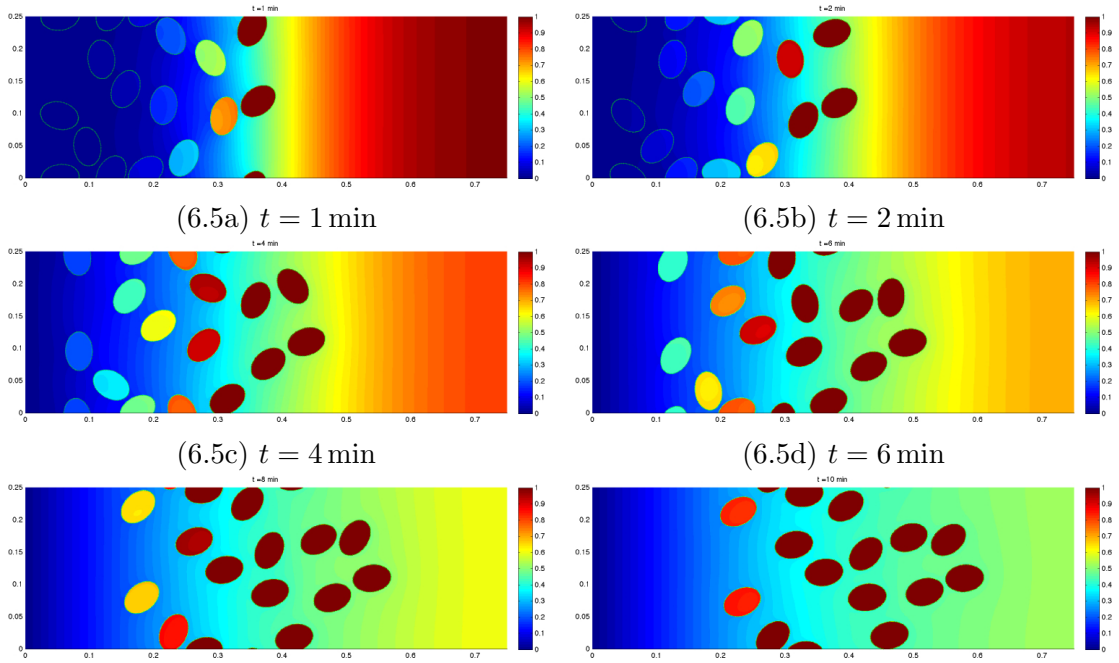
Figure 6.4: Simulation of the motion of one cell undergoing chemotaxis by a single chemoattractant. The colour scheme represent the concentration of chemoattractant.

6.4.2 Numerical simulation of the motion of many cells subject to chemotaxis

The objective of this test problem is to assess the stability of the numerical method we have established when many cells are interacting with the chemoattractant and with other cells.

We adopt here the same setup as that in the previous paragraph, except we now place initially 14 cells in a domain extended in size to accommodate them all. The concentration of chemoattractant satisfies the same PDE as before, with identical boundary conditions.

In order to avoid the collisions of cells, we use a short-ranged repelling force [70]. Details on the actual implementation are provided in Appendix B. The range of the force is of order $\mathcal{O}(h)$ which ensures that it shrinks to zero as $h \to 0$. In practice, we chose $\rho = 2\sqrt{2}h$, in order to keep a minimal distance between two cells corresponding to a line of four triangular elements. Figure 6.5 shows a realisation of this simulation.



(6.5a) $t = 1\,\text{min}$ (6.5b) $t = 2\,\text{min}$

(6.5c) $t = 4\,\text{min}$ (6.5d) $t = 6\,\text{min}$

(6.5e) $t = 8\,\text{min}$ (6.5f) $t = 10\,\text{min}$

Figure 6.5: Simulation of the motion of 14 cells undergoing chemotaxis by a single chemoattractant. The colour indicates the local concentration of chemoattractant (red means 'high' and blue means 'low'). The gradual change in colour of the cells in time shows the absorption of the chemoattractant by the cells. The initial position of the cells is random in the left-hand part of the domain.

6.5 Qualitative validation of the model: comparison with another computational chemotaxis model

Recently, McLennan *et al.* [94] have set up a simulation of the chemotaxis of neural crest (NC) cells by the vascular endothelial growth factor (VEGF) and computationally reproduced qualitative results they obtained from *in vivo* experiments, showing long-range migration of cells responding to long-range chemical signals in a rapidly growing extracellular matrix (see Figure 6.6). The cells were modelled using the hoover cell model introduced in Chapter 5. The signalling pathways that, from the binding of VEGF on receptors on the cell membrane, lead to the response of the cell are considerably simplified and reduced to the sole absorption (or internalisation) of VEGF by a cell. The VEGF is assumed to undergo diffusion in the growing domain and to be produced uniformly following a logistic growth model. Since the production rate of VEGF is set arbitrarily,

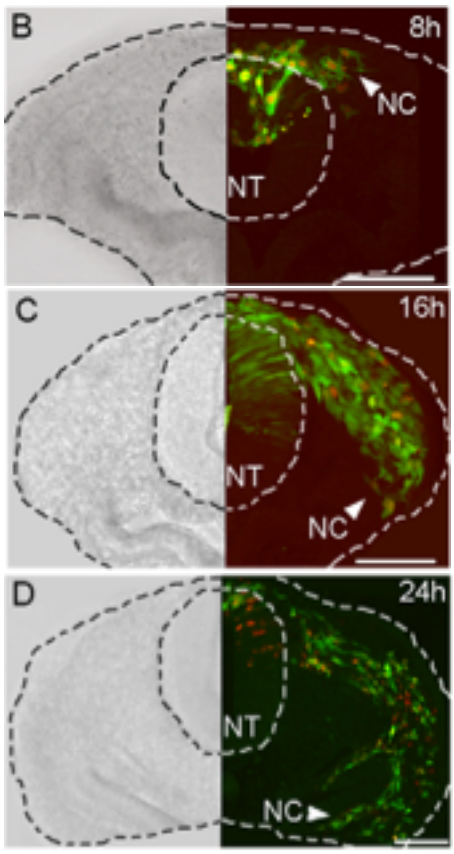


Figure 6.6: Migration of neural crest cells at different times ($+8h$, $+16h$ and $+24h$) after electroporation, courtesy of McLennan *et al.* [94] with permission.

then the internalisation rate can also be set arbitrarily, in such a way that both terms balance. The PDE satisfied by the concentration u of the VEGF is:

$$\frac{\mathrm{d}u}{\mathrm{d}t} = D\triangle u - K\mathbb{1}_{\bigcup C_i}u + f(u), \tag{6.35}$$

where f represents the logistic growth of the chemoattractant and K the internalisation rate of VEGF by the NC cells. The growth of the domain follows a nonlinear law that is described in [94, Supplementary material A].

In order to qualitatively validate the simulation framework presented in this chapter, we propose to reproduce the model proposed by McLennan *et al.* [94] and given a set of parameters, qualitatively compare the outcomes of their framework, considered as the reference framework, and ours.

We recall that, in the hoover cell model, the absorption of the chemoattractant VEGF by each cell is modelled by a depletion term written $-K\mathbb{1}_{\bigcup C_i}u$. We have seen in Chapter

Parameter	Symbol	Value/Magnitude	Source
Number of cells	N	33	–
Simulation time	T	24 h	–
VEGF internalisation rate	K	0.1	–
VEGF diffusivity	D	$3.8 \times 10^{-1}\,\mathrm{mm}^2.\mathrm{h}^{-1}$	[87]
NC cell radius	r	$1.5 \times 10^{-2}\,\mathrm{mm}$	[94]
Typical cell speed	$\bar{s}$	$5.0 \times 10^{-2}\,\mathrm{mm.h}^{-1}$	[94]

Table 6.3: Table showing the values for the parameters of the problem. The first three parameters were arbitrarily chosen for the purpose of the test.

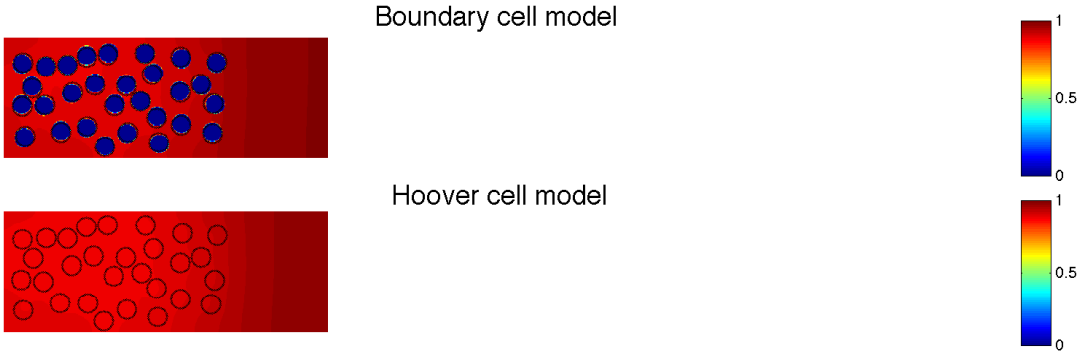
5 that in a simple steady state situation, the parameter K can be matched with the internalisation rate k of the chemoattractant in the boundary cell model. The relationship between those two parameter is written:

$$\frac{k}{D} = \sqrt{\frac{K}{D}} \frac{I_0'(\sqrt{Kr^2/D})}{I_0(\sqrt{Kr^2/D})}, \tag{6.36}$$

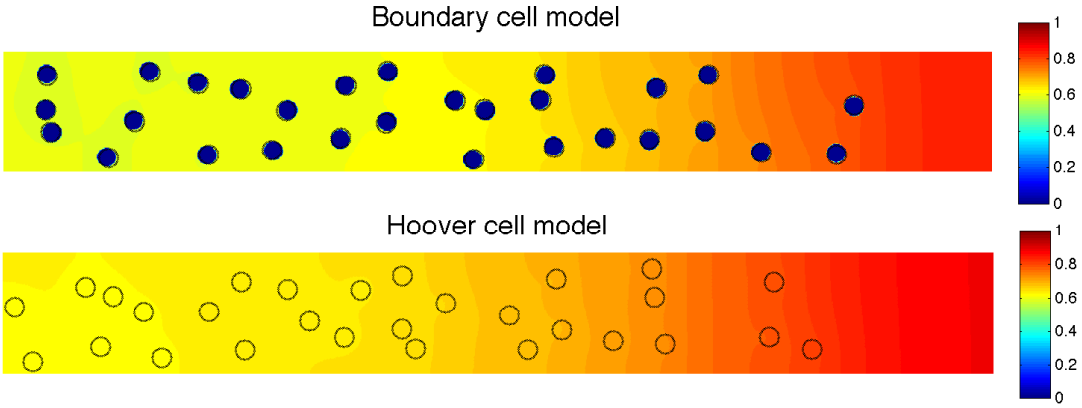
where r is the radius of a cell and D the diffusion constant of VEGF. We will use this relationship to compare the two modelling approaches and given a reference depletion rate K, compute a matching reaction rate k for the boundary cell model.

Concerning the motion of cells, we adopt a model in which the norm of the velocity of the cells is assumed to be constant, only its direction can change, in a run-and-tumble fashion [53]. If the motion of a cell at some time step would provoke a collision with a neighbouring cell, this motion does not take place and the current cell is set to be immobile until the next time step, at which another movement direction will be attempted. In order to determine the direction of the motion, a random direction is chosen, sampled from the angular distribution of the density of chemoattractant at the boundary of the cell. This approach is the one chosen by McLennan *et al.* for their own simulation framework that we are reproducing.

In Figure 6.7, we show a realisation of both simulation models on a 24-hour period. For the boundary cell model, we observe a behaviour comparable to that given by the hoover cell model, as they both qualitatively reproduce the behaviour of cells observed *in vitro*, as shown in Figure 6.6. We can see in particular that the boundary cell models captures the spread of the NC cells and their migration, responding to the chemoattractant.



(6.7a) $t = 0\,h$



(6.7b) $t = 24\,h$

Figure 6.7: Simulation results obtained from two different models for cells: the hoover cell model and the boundary cell model, at $t = 0\,\text{h}$ and $t = 24\,\text{h}$, with a time step $\Delta t = 0.04\,\text{h}$. The space resolution chosen was $h = 1\ \mu\text{m}$, so the main grid was formed of 1024×128 points. The colour scheme represents the value of the concentration of chemoattractant (here VEGF). The simulation of the boundary cell model was about 20 times slower to run than that of the hoover cell model. The total simulation time was approximately 10 hours.

This result also qualitatively validates our numerical framework and the matching relationship (6.36) between the hoover cell and the boundary cell models. At a greater numerical and computational cost and complexity, the boundary cell model used in conjunction with the FDM enables the modelling of non uniform distribution of receptors on the membrane of the cell, which cannot be achieved using a simple hoover cell model.

6.6 Summary

We recall the main characteristics of the modelling approach we have developed in this chapter:

- the cells are discrete non deformable solid entities and no longer modelled as point particles;
- the evolution of the concentration of extracellular chemicals satisfies an advection-reaction-diffusion equation; a mixed finite element method is adopted to better monitor fluxes of chemicals, especially through the membrane of cells;
- the membrane of cells is modelled mathematically and effectively delimits the bulk and the cytosol.
- Internal variables allow the management of the interface, of the fluxes of chemicals through the membrane of cells, and allow the coupling of the motion of cells with the external stimuli received by the cells through its membrane;
- our simulation framework is highly modular, allowing us to couple easily different models of different nature, which other simpler models such as a point-like cell model or a hoover cell model cannot allow;
- the computational cost of the simulation is contained thanks to the FDM. This framework can fit many cells: indeed, the cost in terms of degrees of freedom for the boundary of cells grows linearly with the cumulated perimeter of the cells.

In Section 6.4, we have provided two possible realisations of this simulation set up in a classical chemotaxis example: first with one cell and finally with several cells in the computational domain. To avoid collisions between cells, we implemented a short-range

repulsive force acting between cells. In Section 6.5, we have validated our modelling approach by showing that the behaviour of neural crests cells could be reproduced using our own modelling approach.

Chapter 7

Conclusion

In this dissertation, we were interested in developing a new multiscale simulation framework for the study of the phenomenon of chemotaxis, over length scales of order a few millimetres and time scales of several hours. The main motivation for this resulted from the observation that current computational models failed to provide a satisfying way to link the extracellular and intracellular description of the signal transduction mechanisms, crucial when it comes to studying modelling chemotaxis.

7.1 Mathematical and biological context

The topics that have been discussed in this dissertation have spanned several areas of mathematics.

The main topic discussed in this thesis was the finite element method in general and the fictitious domain method in particular. The FDM, created in the early 1990s, has been extensively used in the literature in computational fluid dynamics for simulating fluids with a high number of solid moving obstacles. We recall that the main issue with the FDM lies in the difficulty in imposing the boundary conditions at the interface between the solid obstacles and the computational domain, that is why the FDM must be tailored to the type of boundary conditions imposed at the interfaces of the moving obstacles. FDMs for Dirichlet and Neumann problems have been proposed by Glowinski and Joly respectively, using H^1 and $H^{\text{div}} \times L^2$ formulations. However, the literature could not provide us with a satisfactory FDM for Robin boundary problems.

Our work also involved numerical algebra and numerical methods for solving large linear saddle-point systems. Such linear systems arise in a number of areas, such as computational fluid dynamics and the fictitious domain method. When solving large linear systems, iterative methods are usually preferred because the cost of direct methods in terms of CPU time and storage becomes prohibitive. The efficiency of iterative methods can be improved by designing appropriate preconditioners, and in particular a preconditioner for the linear system arising from the Dirichlet FDM has been proposed by Glowinski.

The work presented in this dissertation is also related to cell modelling in mathematical biology for chemotaxis. Cells can be represented either using a statistical approach by a density, or by discrete agents independent of each other. When adopting a discrete approach, in the literature, point-like cell model is very popular for studying the aggregation patterns of very large amounts of cells, due to the ease of implementation of the model and its very moderate computational cost.

However, it cannot be used to simulate the signalling pathways that lead to the orientation of the cells. The same applies to the hoover cell model because in this last modelling approach, the cells are not equipped with a membrane and therefore exchanges between the cytoplasm and the extracellular space cannot be accounted for.

7.2 Our contributions

Our main achievement is the production of a stable finite-element numerical scheme that allows its user to simulate a chemotaxis system comprising a reaction-advection-diffusion PDE coupled to a Stokes equation on a moving domain and cells modelled as finite-sized individual agents and characterised by their state variables. The most important issue we had to cope with was containing the computational cost by selecting an appropriate numerical method for discretising the PDEs.

The containment of the computational cost of solving coupled PDEs on a moving domain was performed using the FDM, which we introduced in Chapter 3. After presenting FDMs for Dirichlet and Neumann elliptic problems, we extended the range of applications of the FDM to that of elliptic PDEs equipped with Robin boundary conditions.

The reason why we needed such a method is motivated by our biological application. Robin boundary conditions are useful for modelling the binding on membrane receptors of signalling molecules. Finally, the elaboration of a FDM for Robin-boundary elliptic PDEs has paved the way to solve more complex elliptic problems, for instance PDEs with nonlinear boundary conditions, giving us the ability to describe bimolecular chemical reactions.

Due to the fact that the enforcement of the geometry and the boundary conditions is operated by FDM specific operators, we have also focused on the efficient computation of such operators and provided an algorithm minimising its cost in Chapter 4. In the same chapter, we also dealt with the preconditioning of the linear systems arising from FDM. After having presented Glowinski's optimal preconditioner for the saddle-point matrix derived from the Dirichlet FDM, we derived an optimal preconditioner for the linear system derived from the Neumann and Robin fictitious domain methods.

In Chapter 5, we examined two possible approaches for modelling the binding of a chemotactic signal on the membrane of a cell. The first being called the boundary cell model is the approach we adopted in our simulation framework, and the second is the hoover cell, an approach that is quite popular in the literature because of its ease of implementation and its small computational cost. We established a match between the two models by giving an expression for the non dimensionalised membrane binding rate β of the boundary cell model in terms of the volume depletion rate α of the hoover cell model. This was performed in the simple setting where one cell is immobile and isolated in an infinite domain. We then extended this result by giving the cell a constant speed in a fixed direction. An asymptotic analysis showed that the relationship between β and α was now dependent on the position on the boundary of the disk-shaped cell. We concluded Chapter 5 by a discussion on the appropriateness of the use of each model.

Lastly, in Chapter 6, we presented our simulation framework for eukaryotic chemotaxis. We wrote down the reaction-advection-diffusion PDE satisfied by the signalling chemicals, the Stokes equation for the fluid carrying the chemicals, and the equations of motion for each cell. The two PDEs were discretised using the fictitious domain method: the Robin FDM is employed for the discretisation of the reaction-advection-diffusion equa-

tion and the FDM for Dirichlet elliptic problem is implemented for the Stokes flow, since each cell drags locally the fluid surrounding it. At the implementation stage, we showed the numerical stability of the scheme when one cell subjected to chemotaxis by one signalling chemical species is placed in the computational domain. We also compared the simulation outcomes of our scheme and the simulation of the hoover cell model when both models are matched according to the relationship derived in Chapter 5. This comparison allowed us to confirm the stability of our numerical FDM scheme and the accuracy of the matching between the hoover cell and boundary cell models of Chapter 5.

7.3 Criticism and issues

One of the drawbacks of the use of the FDM is that it requires that the boundary conditions imposed at all the immersed objects in the computation domain should be of the same type. The reason for this is that the FDM is usually tailored to the PDE problem its user is dealing with. The simulation framework we presented in Chapter 6 heavily relies on the fact that the boundary conditions imposed at the boundary of modelled cells are identical. Hence, our framework is not flexible enough in this regard: it cannot accommodate cells with different types of boundary conditions.

This issue is the direct consequence of the numerical method chosen for discretising the reaction-diffusion equation governing the concentration of external chemical signals. If we wished to implement a simulation with a greater diversity of boundary conditions, we should then change of numerical method, as, for now, a FDM accommodating all types of boundary conditions does not exist.

The choice of the FDM also impacts the size and the shapes of the cells, as well as the scalability. Firstly, as the number of cells increases or more generally, their global perimeter, the number of degrees of freedom needed for the discretisation of the interfaces increases, and therefore the global computational cost. Lastly, the shape of the cell influences greatly the cost of the fictitious domain method: when the cells have more complex shapes than simple ellipses, we can expect to have to choose the resolution

of the volume mesh finer. This increases the cost of the computation of the fictitious domain operators and that of solving the resulting linear system.

The simulation framework presented in Chapter 6 is based on the assumption of cells being solid objects, which is largely unrealistic because cells change of shape as they conform to their environment, and we have not discussed the case of cells with dynamically varying shapes. The implementation of such feature would be feasible in the presented setting as long as the area of each cell is conserved, however the computational cost of the whole simulation would increase.

7.4 Possible future research directions

Most of the aspects discussed in this dissertation were either numerical or computational. We presented all the necessary mathematical and computational tools for understanding and implementing the simulation framework presented in Chapter 6. A natural extension of this work would consist in applying it to a concrete biological situation. A significant challenge would be the display of patterns when associating in the same simulation experiment the model discussed by Levine *et al.* [86] for intracellular signalling pathways of cAMP and the Dallon & Othmer model [37] for extracellular dynamics of cAMP. Both models aim at exploring the chemotaxis of *D. discoideum* by cAMP.

More generally, the simulation framework we have developed will allow us to test the different hypotheses that have been formulated in the recent years on the mechanisms of detection of gradients in chemoattractant concentrations by eukaryotic cells such as *D. discoideum*: is it performed by spatial gradient sensing by detecting spatial differences in the number of occupied sensors as postulated by Mato [92], spatio-temporal threshold mechanism where the orientation of the cell is determined by the formation of an internal gradient created by the external signal, as originally suggested by Brundage *et al.* [25] and more recently by Postma *et al.* [112]?

On an other note, significant improvements can be brought to the simulation framework presented in this dissertation. The stability of the method could be improved by introducing a full coupling between the law of motion of cells and the fluid, as performed in [20] for instance. A possibility would be to account for the Cauchy stress that is

exerted on the cells by the fluid in the law of motion of cells. This could potentially help enforcing the non overlapping condition on the cells. The technical difficulty arising in such a setting would be the efficient solve of the resulting linear system and would require further investigations in terms of preconditioning.

As we have seen in Chapter 5, the numerical implementation of the simulation framework developed in this dissertation could be dramatically simplified if we managed to replace the boundary reaction term on the membrane $\beta(\theta)$ of the cell by an equivalent volume reaction term $\alpha(x, y)$. A possible future project would be to investigate whether such an equivalent $\alpha(x, y)$ exists for any $\beta(\theta)$ and a derivation method for it would greatly help in reducing the cost of the chemotaxis simulation for finite-sized cell.

Due to the significant increase of available computational power in modern machines, another numerical challenge that could be tackled in the future would be transposing the numerical scheme of Chapter 6 from 2D to 3D. Since 'real' cells are three-dimensional objects, this new modelling would better mimic reality, at the cost of an additional dimension to cope with. This would imply using three-dimensional elements such as tetrahedra instead of triangles and adapting the algorithm proposed in Chapter 4 to interfaces that are no longer curves but surfaces, for instance ellipsoids, intersecting with the faces of the tetrahedra that form the mesh of the three-dimensional computational domain.

Lastly, another possible project for future research would consist in relaxing the non deformable constraint that we imposed on the cells in Chapter 6 and incorporate dynamics of the shape of the membrane. An example of such simulation model is presented by Neilson [99]. Reaction-diffusion equations describing the evolution of the concentration of an activator-inhibitor system are solved on a deformable contour that represents the membrane of a cell. The dynamics of the change of the shape of the contour are governed by the concentration of an activator. This model would help in joining the chemical processes of chemotaxis together with the mechanical response of a cell that leads to its migration.

Appendix A

Solution to the steadily moving hoover cell problem

Let us recall the PDE problem we are interested in solving:

$$\begin{aligned} -\nabla^2 u - \varepsilon \frac{\partial u}{\partial x} + \alpha^2 \mathbb{1}_{B(0,1)}\, u = 0, \quad & \text{in } \mathbb{R}^3 \\ u \to 1, \quad & \text{as } r \to \infty \text{ and } \theta = 0. \end{aligned} \tag{5.35}$$

In order to provide an approximated solution to this equation, we attempt to find a solution written as a regular expansion in powers of the small parameter ε:

$$u_h = u_0 + \varepsilon^1 u_1 + \varepsilon^2 u_2 + \dots \tag{A.1}$$

At zero-th order in ε, the function u_h is solution of the immobile hoover cell model. At first order, we obtain a Poisson equation satisfied by u_1:

$$\begin{aligned} -\nabla^2 u_1 + \alpha^2 \mathbb{1}_{B(0,1)}\, u_1 = \frac{\partial u_0}{\partial x}, \quad & \text{in } \mathbb{R}^3 \\ u_1 \to 0, \quad & \text{as } r \to \infty \text{ and } \theta = 0. \end{aligned} \tag{A.2}$$

Because of the break in the spherical symmetry and the nonzero right hand side, the derivation of the solution of this PDE is slightly more complex that in the previous section. Recall the expression of u_0. It is the solution of the immobile cell problem:

$$u_0 = 1 - \left(1 - \frac{\tanh \alpha}{\alpha}\right) \frac{1}{r}\, \mathbb{1}_{r>1} + \frac{\sinh(\alpha r)}{\cosh(\alpha)\alpha r}\, \mathbb{1}_{r\leq 1}. \tag{A.3}$$

Then

$$\frac{\partial u_0}{\partial x} = \begin{cases} \left(1 - \dfrac{\tanh \alpha}{\alpha}\right) \dfrac{\cos\theta}{r^2}, & r > 1, \\ \dfrac{r\cosh(\alpha r)}{\cosh(\alpha)} \left(1 - \dfrac{\tanh(\alpha r)}{\alpha r}\right) \dfrac{\cos\theta}{r^2}, & r \leq 1, \end{cases} \tag{A.4}$$

with $\cos\theta = \dfrac{x}{r}$.

Because of the rotational symmetry of the problem along the x-axis, the solution u_1 only depends on r, θ and not on the spherical angle ϕ. Therefore we can rewrite (A.2) as:

$$\frac{\partial}{\partial r}\left(r^2\frac{\partial u_1}{\partial r}\right) + \frac{\partial}{\partial \mu}\left((1-\mu^2)\frac{\partial u_1}{\partial \mu}\right) - \alpha^2 \mathbb{1}_{\{r\leq 1\}}\, r^2 u_1 = \begin{cases} -\left(1 - \dfrac{\tanh\alpha}{\alpha}\right)\mu, & r > 1, \\ -\dfrac{1}{\alpha\cosh(\alpha)}\left((\alpha r)\cosh(\alpha r) - \sinh(\alpha r)\right)\mu, & r \leq 1, \end{cases} \tag{A.5}$$

after we have expressed the Laplacian in spherical coordinates and used the classical change of variable $\mu = \cos\theta$.

Let us start by the easiest case: when $r > 1$, Equation (A.5) becomes:

$$\frac{\partial}{\partial r}\left(r^2\frac{\partial u_1}{\partial r}\right) + \frac{\partial}{\partial \mu}\left((1-\mu^2)\frac{\partial u_1}{\partial \mu}\right) = -\left(1 - \frac{\tanh\alpha}{\alpha}\right)\mu. \tag{A.6}$$

The solution of this equation is well-known and is written as a series of Lagrange polynomials P_n:

$$u_1^{r>1}(r,\mu) = \sum_{n=0}^{\infty}\left(G_n\, r^n + J_n\, r^{-(n+1)}\right) P_n(\mu) + \frac{1}{2}\left(1 - \frac{\tanh\alpha}{\alpha}\right)\mu. \tag{A.7}$$

The second term corresponds to a particular solution of the Poisson equation (A.6). Since we are interested only in bounded solutions, we deduce $G_n = 0, \forall n > 0$. The coefficients J_n will be determined using the matching condition at $r = 1$. Lastly, since we require that $u_1^{r>1}(r,\theta) \to 0$ as $r \to \infty$ and $\theta = 0$, we find:

$$G_0 = -\frac{1}{2}\left(1 - \frac{\tanh\alpha}{\alpha}\right). \tag{A.8}$$

We obtain:

$$u_1^{r>1}(r,\mu) = \sum_{n=0}^{\infty} J_n\, r^{-(n+1)} P_n(\mu) + \frac{1}{2}\left(1 - \frac{\tanh\alpha}{\alpha}\right)(\mu - 1). \tag{A.9}$$

In the case of $r \leq 1$, again, we look for a solution in the form a series:

$$u_1^{r\leq 1} = \sum_{n=0}^{\infty} a_n(r)\, P_n(\mu), \tag{A.10}$$

where P_n are the Legendre polynomials. Equation (A.5) becomes:

$$\sum_{n=0}^{\infty}\left(\frac{\partial}{\partial r}\left(r^2\frac{\partial a_n(r)}{\partial r}\right)-\left(n(n+1)+\alpha^2\,r^2\right)a_n(r)\right)P_n(\mu)=\frac{\sinh(\alpha r)-(\alpha r)\cosh(\alpha r)}{\alpha\cosh(\alpha)}\mu. \tag{A.11}$$

If we write $\alpha^2=-(i\alpha)^2$, we recognise that a_n must be a linear combination of the spherical Bessel functions of order n. Hence, the general solution of this last equation reads:

$$u_1^{r\leq 1}=\sum_{n=0}^{\infty}\left(H_n\,i^n j_n(i\alpha r)+L_n\,y_n(i\alpha r)\right)P_n(\mu)+u_{10}^{r\leq 1}, \tag{A.12}$$

where H_n and L_n are constants to be determined using the boundary conditions and $u_{10}(r,\theta)$ a particular solution satisfying (A.5). The functions j_n and y_n are the spherical Bessel functions. In this equation, they have a purely imaginary argument. Each function $i^n j_n(i\alpha r)$ or $i^n y_n(i\alpha r)$ takes values in $\mathbb{R}$ only since:

$$\begin{aligned} i^n j_n(ix)&=(-x)^n\left(\frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x}\right)^n\frac{\sinh x}{x},\\ y_n(ix)&=(-x)^n\left(\frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x}\right)^n\frac{\cosh x}{x}. \end{aligned} \tag{A.13}$$

Firstly, we notice that

$$u_{10}^{r\leq 1}(r,\mu)=-\frac{\sinh(\alpha r)}{2\alpha\cosh(\alpha)}\mu \tag{A.14}$$

is a particular solution of the equation. The general solution $u_1^{r\leq 1}$ of (A.5) is:

$$u_1^{r\leq 1}(r,\mu)=\sum_{n=0}^{\infty}\left(H_n\,i^n j_n(i\alpha r)+L_n\,y_n(i\alpha r)\right)P_n(\mu)-\frac{\sinh(\alpha r)}{2\alpha\cosh(\alpha)}\mu. \tag{A.15}$$

Since we are only interested in bounded solutions, all L_n must be zero. The solution $u_1^{r\leq 1}(r,\mu)$ then takes the form:

$$u_1^{r\leq 1}(r,\mu)=\sum_{n=0}^{\infty}H_n\,i^n j_n(i\alpha r)P_n(\mu)-\frac{\sinh(\alpha r)}{2\alpha\cosh(\alpha)}\mu. \tag{A.16}$$

Let us match the two pieces $u_1^{r>1}$ and $u_1^{r\leq 1}$ at the interface $r=1$: by orthogonality of the Legendre polynomials, we are allowed to perform a term by term matching. For $n\geq 2$, from matching $H_n\,i^n j_n(i\alpha r)$ and $J_n r^{-(n+1)}$ at $r=1$ and their first derivatives, we obtain:

$$\begin{aligned} J_n&=i^n j_n(i\alpha)H_n\\ -(n+1)J_n&=\alpha i^{n+1}j_n'(i\alpha)H_n \end{aligned} \tag{A.17}$$

As the determinant of this system is nonzero for any real α, then $J_n = H_n = 0$ for $n \geq 2$.

For $n = 1$, we match the terms in $P_1(\mu) = \mu$, we obtain two equations:

$$\begin{aligned} ij_1(i\alpha)H_1 - J_1 &= \frac{1}{2}, \\ -\alpha\, j_1'(i\alpha)H_1 + 2J_1 &= \frac{1}{2}. \end{aligned} \tag{A.18}$$

After simplifications, the solution of this system yields H_1 and J_1:

$$\begin{aligned} H_1 &= -\frac{3}{2\sinh(\alpha)}, \\ J_1 &= -\frac{1}{2} - \frac{3}{2\alpha^2}\left(1 - \frac{\alpha}{\tanh(\alpha)}\right). \end{aligned} \tag{A.19}$$

Now we compare the terms in P_0:

$$\begin{aligned} j_0(i\alpha)H_0 - J_0 &= -\frac{1}{2}\left(1 - \frac{\tanh\alpha}{\alpha}\right), \\ i\alpha j_0'(i\alpha)H_0 + J_0 &= 0. \end{aligned} \tag{A.20}$$

We immediately recover:

$$\begin{aligned} H_0 &= -\frac{1}{2\cosh(\alpha)}\left(1 - \frac{\tanh\alpha}{\alpha}\right), \\ J_0 &= \frac{1}{2}\left(1 - \frac{\tanh\alpha}{\alpha}\right)^2. \end{aligned} \tag{A.21}$$

The first order correction u_1 finally reads:

$$u_1(r,\theta) = -\begin{cases} \left(1 - \frac{\tanh\alpha}{\alpha}\right)\frac{\sinh(\alpha r)}{2\alpha r\cosh(\alpha)} + \\ \left(\frac{\sinh(\alpha r)}{2\alpha\cosh(\alpha)} + \frac{3}{2\sinh(\alpha)}\left(\frac{\sinh(\alpha r)}{(\alpha r)^2} - \frac{\cosh(\alpha r)}{\alpha r}\right)\right)\cos\theta, & r \leq 1 \\ \frac{1}{2}\left(1 - \frac{\tanh\alpha}{\alpha}\right)\left(1 - \left(1 - \frac{\tanh\alpha}{\alpha}\right)\frac{1}{r}\right) + \\ \left[\frac{1}{2}\frac{\tanh\alpha}{\alpha} - \frac{1}{2} + \left(\frac{1}{2} + \frac{3}{2\alpha^2}\left(1 - \frac{\alpha}{\tanh(\alpha)}\right)\right)\frac{1}{r^2}\right]\cos\theta. & r > 1 \end{cases} \tag{A.22}$$

This correction term is always negative. As a sanity check, we can observe that if we take α close to 0, then the correction term also tends to 0 uniformly with respect to α and decays as fast as α^2.

Appendix B

Avoiding collision of cells

To prevent the overlap of two cells whilst minimising the bias introduced in the simulation, we use a short-ranged repulsive force which model is described in [70]. In the case of two circular cells of radius r_i and r_j repelling each other, the expression of this (artificial) force is:

$$\boldsymbol{F}_{i\to j} = \frac{F_0}{\varepsilon_c}\left[\min\left(\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|-r_i-r_j-\rho}{\rho},0\right)\right]^2 \frac{(\boldsymbol{x}_i-\boldsymbol{x}_j)}{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|} \tag{B.1}$$

where F_0 is the typical value of the forces acting on the cells and ρ represents the range of the repulsive force. When the two spherical cells see their distance shrinking to ρ or less, the value of the repulsive force between the two cells becomes nonzero. ε_c is a dimensionless parameter of order $\varepsilon_c = \mathcal{O}\left((\delta/\rho)^2\right)$, where δ is the penetration length.

In the case of arbitrary shaped cells, a moment of force also arises. The force applied by one cell on another is:

$$\boldsymbol{F}_{i\to j} = \int_{C_i}\int_{C_j} \frac{F_0}{\varepsilon_c}\left[\min\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}_2\|-\rho}{\rho},0\right)\right]^2 \frac{(\boldsymbol{x}_1-\boldsymbol{x}_2)}{\|\boldsymbol{x}_1-\boldsymbol{x}_2\|}\,\mathrm{d}\boldsymbol{x_1}\,\mathrm{d}\boldsymbol{x_2} \tag{B.2}$$

and the moment of force:

$$\begin{aligned}\boldsymbol{M}_{i\to j} = \int_{C_i}\int_{C_j} \frac{F_0}{\varepsilon_c}\left[\min\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}_2\|-\rho}{\rho},0\right)\right]^2 \\ \frac{(\boldsymbol{x}_1-\boldsymbol{x}_2)}{\|\boldsymbol{x}_1-\boldsymbol{x}_2\|}\times(\boldsymbol{x}_1-\boldsymbol{x}_i)\,\mathrm{d}\boldsymbol{x_1}\,\mathrm{d}\boldsymbol{x_2}.\end{aligned} \tag{B.3}$$

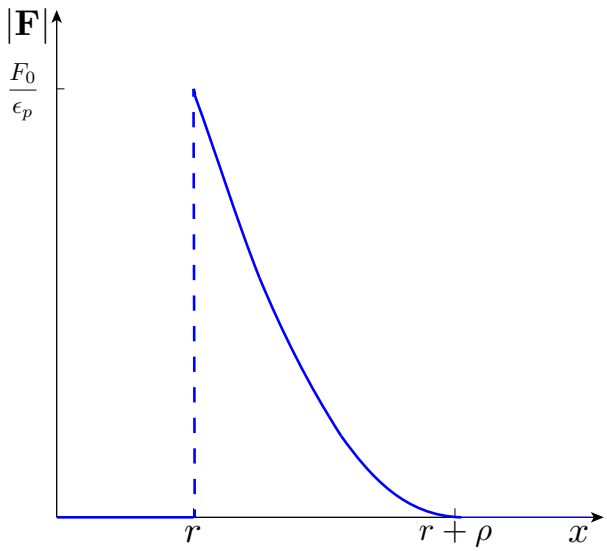


Figure B.1: Sketch of the modulus of the repulsive force in 1D for a cell centred in 0 and of radius r. The range of the force is ρ.

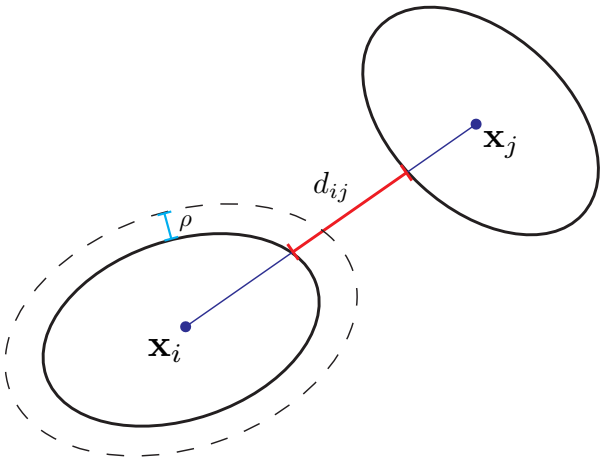


Figure B.2: This figure shows how the force $\boldsymbol{F}_{i\to j}$ is computed: the length d_{ij} is the distance between the points of intersection of the line $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with the boundary of each cell.

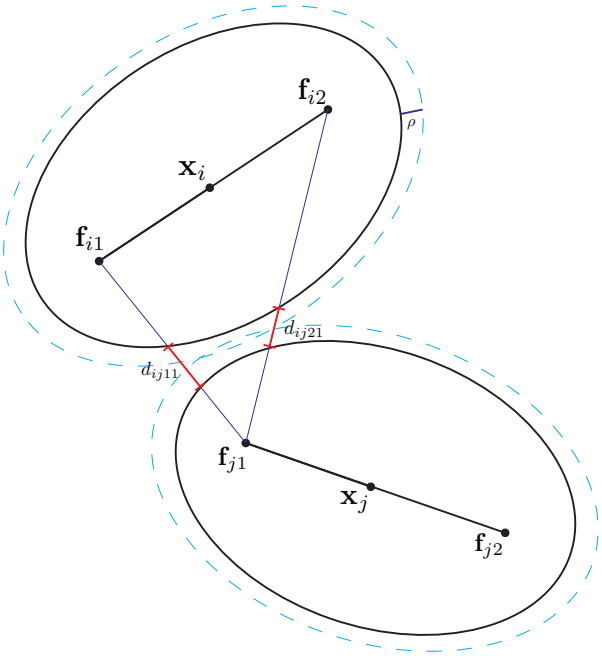


Figure B.3: This figure shows how the moment $\boldsymbol{M}_{i\to j}$ is computed: using the foci of each ellipse, a elementary moment is calculated with respect to the centre of each ellipse.

Numerically, this is an expensive work to compute those two integrals accurately, however, such a precision is not needed for the purpose we are after and we could be contented with a approximation that would still prevent the cells from colliding. We opt for a one-point integration approximation as shown on Figure B.2 and write:

$$\boldsymbol{F}_{i\to j} = \frac{F_0}{\varepsilon_c}\left[\min\left(\frac{d_{ij}-\rho}{\rho}, 0\right)\right]^2 \frac{(\boldsymbol{x}_i - \boldsymbol{x}_j)}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}. \tag{B.4}$$

To compute a reasonable approximation of the moment of force, a two-point approximation scheme is used as shown on Figure B.3 in the case of elliptic cells. The expression for the moment is then written:

$$\boldsymbol{M}_{i\to j} = \sum_{k,l} \frac{F_0}{\varepsilon_c}\left[\min\left(\frac{d_{ijkl}-\rho}{\rho}, 0\right)\right]^2 \frac{(\boldsymbol{f}_{ik} - \boldsymbol{f}_{jl})}{\|\boldsymbol{f}_{ik} - \boldsymbol{f}_{jl}\|} \times (\boldsymbol{f}_{ik} - \boldsymbol{x}_i). \tag{B.5}$$

When computing this approximation, four distances must be computed and they are the main cost in the computation of the moment.

References

[1] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*, **65**. Academic press New York, 1975.

[2] J. Adler. Chemotaxis in bacteria. *Science*, **153**(3737):708–716, 1966.

[3] M. Alber, N. Chen, P.M. Lushnikov, and S.A. Newman. Continuous macroscopic limit of a discrete stochastic model for interaction of living cells. *Physical review letters*, **99**(16):168102, 2007.

[4] W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math*, **9**(1):17–29, 1951.

[5] K. J. Arrow, L. Hurwicz, H. Uzawa, et al. *Studies in linear and non-linear programming*, **2**. Stanford University Press, 1958.

[6] H. Baaser. *Development and Application of the Finite Element Method based on MatLab*. Springer, 2010.

[7] I. Babuska. The finite element method with Lagrange multipliers. *Numer. Math*, **20**:179–192, 1973.

[8] C. Bahriawati and C. Carstensen. Three Matlab implementations of the lowest-order Raviart-Thomas MFEM with a posteriori error control. *Computational Methods in Applied Mathematics*, **5**(4):333, 2005.

[9] R. G. Bartle. *The elements of integration and Lebesgue measure*. Wiley Online Library, 1995.

[10] M. Benzi, G.H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta numerica*, **14**(1):1–137, 2005.

[11] H. C. Berg. *E. coli in Motion*. Springer, 2003.

[12] H. C. Berg, D. A. Brown, et al. Chemotaxis in escherichia coli analysed by three-dimensional tracking. *Nature*, **239**(5374):500–504, 1972.

[13] D. F. Blair. How bacteria sense and swim. *Annual Reviews in Microbiology*, **49**(1):489–520, 1995.

[14] D. Boffi. Stability of higher order triangular Hood-Taylor methods for the stationary Stokes equations. *Mathematical Models and Methods in Applied Sciences*, **4**(02):223–235, 1994.

[15] D. Boffi, F. Brezzi, and M. Fortin. Finite elements for the Stokes problem. *Mixed Finite Elements, Compatibility Conditions, and Applications*, pages 45–100, 2008.

[16] J. T. Bonner and E. B. Frascella. Variations in cell size during the development of the slime mold, *Dictyostelium discoideum*. *The Biological Bulletin*, **104**(3):297–300, 1953.

[17] J. T. Bonner and L. J. Savage. Evidence for the formation of cell aggregates by chemotaxis in the development of the slime mold *Dictyostelium discoideum*. *Journal of Experimental Zoology*, **106**(1):1–26, 1947.

[18] R. B. Bourret, K. A. Borkovich, and M. I. Simon. Signal transduction pathways involving protein phosphorylation in prokaryotes. *Annual review of biochemistry*, **60**(1):401–441, 1991.

[19] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, **15**. Springer Verlag, 2008.

[20] D. Bresch, V. Calvez, E. Grenier, P. Vigneaux, JF Gerbeau, A. Decoene, A. Lorz, S. Martin, B. Maury, and M. Tang. Simulation of self-propelled chemotactic bacteria in a Stokes flow. In *ESAIM: Proceedings*, **30**, pages 104–123. EDP Sciences, 2010.

[21] F. Brezzi. On existence, uniqueness and approximation of saddle-point arising from lagrange multipliers. *RAIRO Anal Numer*, **44**:129–151, 1974.

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

[23] W. Briggs, V. Henson, and S. McCormick. *A Multigrid Tutorial 2nd edition*. SIAM, Philadelphia, 2000.

[24] E. J. Brown. Phagocytosis. *Bioessays*, **17**(2):109–117, 1995.

[25] R. A. Brundage, K. E. Fogarty, R. A. Tuft, and F. S. Fay. Calcium gradients underlying polarization and chemotaxis of eosinophils. *Science*, **254**(5032):703–706, 1991.

[26] J. Buhl, D. J. T. Sumpter, I. D. Couzin, J. J. Hale, E. Despland, E. R. Miller, and S. J. Simpson. From disorder to order in marching locusts. *Science*, **312**(5778):1402–1406, 2006.

[27] R. L. Burden and J. D. Faires. *Numerical analysis, 9th edition*. Kent Publishing Co. Boston, 1989.

[28] B. L. Buzbee, F. W. Dorr, J. A. George, and G. H. Golub. The direct solution of the discrete Poisson equation on irregular regions. *SIAM Journal on Numerical Analysis*, **8**(4):722–736, 1971.

[29] P. G. Ciarlet. *The finite element method for elliptic problems*, **4**. North Holland, 1978.

[30] M. Claviez, M. Brink, and G. Gerisch. Cytoskeletons from a mutant of *Dictyostelium discoideum* with flattened cells. *Journal of Cell Science*, **86**(1):69–82, 1986.

[31] A. Cohen-Dayag, D. Ralt, I. Tur-Kaspa, M. Manor, A. Makler, J. Dor, S. Mashiach, and M. Eisenbach. Sequential acquisition of chemotactic responsiveness by human spermatozoa. *Biology of reproduction*, **50**(4):786–790, 1994.

[32] F. Collino, P. Joly, and F. Millot. Fictitious domain method for unsteady problems. *Journal of Computational Physics*, **138**(2):907–938, 1997.

[33] R. Courant. Variational methods for the solution of problems of equilibrium and vibrations. *Bull. Amer. Math. Soc*, **49**(1):23, 1943.

[34] E. C. M. Crooks, T. Lipniacki, and B. Kazmierczak. Spatially extended model of kinase-receptor interaction. *SIAM Journal on Applied Mathematics*, **73**(1):374–400, 2013.

[35] B. Dacorogna. *Introduction to the Calculus of Variations*. Imperial College Pr, 2004.

[36] J. C. Dallon and H. G. Othmer. A discrete cell model with adaptive signalling for aggregation of *Dictyostelium discoideum*. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **352**(1351):391, 1997.

[37] J. C. Dallon and H. G. Othmer. A continuum analysis of the chemotactic signal seen by *dictyostelium discoideum*. *Journal of Theoretical Biology*, **194**(4):461–483, 1998.

[38] P. J. Davis and P. Rabinowitz. *Methods of numerical integration*. Courier Dover Publications, 2007.

[39] E. J. Dean, Q. V. Dinh, R. Glowinski, J. He, T. W. Pan, and J. Périaux. Least squares/domain imbedding methods for neumann problems: Applications to fluid dynamics. In *Fifth Internat. Symposium on Domain Decomposition Methods for Partial Differential Equations*, pages 451–475, 1992.

[40] B. Delaunay. Sur la sphère vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk*, **7**(793-800):1–2, 1934.

[41] D. A. Dunavant. High degree efficient symmetrical Gaussian quadrature rules for the triangle. *International Journal for Numerical Methods in Engineering*, **21**(6):1129–1148, 1985.

[42] M. Dworkin and K. H. Keller. Solubility and diffusion coefficient of adenosine 3': 5'-monophosphate. *Journal of Biological Chemistry*, **252**(3):864–865, 1977.

[43] M. Eisenbach. Sperm chemotaxis. *Reviews of reproduction*, **4**(1):56–66, 1999.

[44] M. Eisenbach. *Chemotaxis*. Wiley Online Library, 2004.

[45] H. C. Elman. Preconditioners for saddle point problems arising in computational fluid dynamics. *Applied Numerical Mathematics*, **43**(1):75–89, 2002.

[46] H. C. Elman and D. J. Silvester. Fast nonsymmetric iterations and preconditioning for Navier-Stokes equations. *SIAM Journal on Scientific Computing*, **17**(1):33–46, 1996.

[47] H. C. Elman, D. J. Silvester, and A. J. Wathen. Iterative methods for problems in computational fluid dynamics. Technical report, UM Computer Science Department, 1998.

[48] H. C. Elman, D. J. Silvester, and A. J. Wathen. Block preconditioners for the discrete incompressible Navier–Stokes equations. *International Journal for Numerical Methods in Fluids*, **40**(3-4):333–344, 2002.

[49] T. W. Engelmann. Ueber licht-und farbenperception niederster organismen. *Pflügers Archiv European Journal of Physiology*, **29**(1):387–400, 1882.

[50] R. Erban. *From individual to collective behaviour in biological systems.* PhD thesis, University of Minnesota, 2005.

[51] R. Erban, I. G. Kevrekidis, and H. G. Othmer. An equation-free computational approach for extracting population-level behavior from individual-based models of biological dispersal. *Physica D: Nonlinear Phenomena*, **215**(1):1–24, 2006.

[52] R. Erban and H. G. Othmer. From individual to collective behavior in bacterial chemotaxis. *SIAM Journal on Applied Mathematics*, pages 361–391, 2004.

[53] R. Erban and H. G. Othmer. From signal transduction to spatial pattern formation in E. coli: a paradigm for multiscale modeling in biology. *Multiscale Modeling and Simulation*, **3**(2):362–394, 2005.

[54] R. Erban and H. G. Othmer. Taxis equations for amoeboid cells. *Journal of Mathematical Biology*, **54**(6):847–885, 2007.

[55] Alexandre Ern. *Aide-mémoire des éléments finis.* Dunod, 2005.

[56] L.C. Evans. Partial Differential Equations (Graduate Studies in Mathematics, Vol. 19). *Instructor*, 2009.

[57] J. J Falke, Randal B. Bass, S. L. Butler, S. A. Chervitz, and M. A. Danielson. The two-component signaling pathway of bacterial chemotaxis: a molecular view of signal transduction by receptors, kinases, and adaptation enzymes. *Annual review of cell and developmental biology*, **13**:457, 1997.

[58] J. Folkman. Tumor angiogenesis: therapeutic implications. *New England Journal of Medecine*, **285**:1182–1186, 1971.

[59] B. Franz and R. Erban. Hybrid modelling of individual movement and collective behaviour. In *Dispersal, individual movement and spatial ecology: A mathematical perspective*. Springer-Verlag, 2012.

[60] P. Friedl, Y. Hegerfeldt, M. Tusch, et al. Collective cell migration in morphogenesis and cancer. *International Journal of Developmental Biology*, **48**:441–450, 2004.

[61] H. Fujita, H. Kawahara, and H. Kawarada. Distribution theoretic approach to fictitious domain method for Neumann problems. *East West Journal of Numerical Mathematics*, **3**:111–126, 1995.

[62] G. Gerisch, D. Hulser, D. Malchow, and U. Wick. Cell communication by periodic cyclic-AMP pulses. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, **272**(915):181–192, 1975.

[63] L.C. Giojalas and R.A. Rovasio. Mouse spermatozoa modify their motility parameters and chemotactic response to factors from the oocyte microenvironment. *International Journal of Andrology*, **21**(4):201–206, 1998.

[64] V. Girault and R. Glowinski. Error analysis of a fictitious domain method applied to a dirichlet problem. *Japan Journal of Industrial and Applied Mathematics*, **12**(3):487–514, 1995.

[65] R. Glowinski, J. He, V. Kwock, T Pan, and J. Périaux. Lagrange multiplier approach to fictitious domain methods: application to fluid dynamics and electro-magnetics. In *Fifth International Symposium on Domain Decomposition*

Methods for Partial Differential Equations, page 151. Society for Industrial & Applied, 1991.

[66] R. Glowinski and Q. He. Numerical solution of linear elliptic problems with Robin boundary conditions by a least-squares/fictitious domain method. *Domain Decomposition Methods in Science and Engineering XIX*, pages 375–382, 2011.

[67] R. Glowinski, T. W. Pan, T. I. Hesla, and D. D. Joseph. A distributed Lagrange multiplier/fictitious domain method for particulate flows. *International Journal of Multiphase Flow*, **25**(5):755–794, 1999.

[68] R. Glowinski, T. W. Pan, and J. Periaux. A fictitious domain method for Dirichlet problem and applications. *Computer Methods in Applied Mechanics and Engineering*, **111**(3-4):283–303, 1994.

[69] R. Glowinski, T. W. Pan, R. O. Wells, and X. Zhou. Wavelet and finite element solutions for the Neumann problem using fictitious domains. *Journal of Computational Physics*, **126**(1):40–51, 1996.

[70] R. Glowinski, TW Pan, TI Hesla, DD Joseph, and J. Periaux. A fictitious domain approach to the direct numerical simulation of incompressible viscous flow past moving rigid bodies: application to particulate flow. *Journal of Computational Physics*, **169**(2):363–426, 2001.

[71] R. Glowinski, TW Pan, and J. Periaux. Distributed lagrange multiplier methods for incompressible viscous flow around moving rigid bodies. *Computer methods in applied mechanics and engineering*, **151**(1):181–194, 1998.

[72] M. S. Gockenbach. *Understanding and implementing the finite element method*. SIAM, 2006.

[73] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins University Press, 1996.

[74] F. Goto, K. Goto, K. Weindel, J. Folkman, et al. Synergistic effects of vascular endothelial growth factor and basic fibroblast growth factor on the proliferation and cord formation of bovine capillary endothelial cells within collagen gels. *Laboratory Investigation*, **69**(5):508, 1993.

[75] F. Graner and J. A. Glazier. Simulation of biological cell sorting using a two-dimensional extended potts model. *Physical review letters*, **69**(13):2013, 1992.

[76] A. Greenbaum. Iterative methods for solving linear systems. Technical Report ISBN 0-89871-396-X. xiii+ 220 pp. Griewank: 1985: SNE, SIAM, Philadelphia, PA, USA,, 1997.

[77] P. Grisvard. *Elliptic problems in nonsmooth domains*. Pitman Advanced Pub. Program, 1985.

[78] W. Hundsdorfer and J. Verwer. *Numerical solution of time-dependent advection-diffusion-reaction equations*, **33**. Springer, 2003.

[79] P. Joly and L. Rhaouti. Domaines fictifs, éléments finis Hdiv et condition de Neumann: le probleme de la condition inf-sup. *Comptes Rendus de l'Académie des Sciences*, **328**(12):1225–1230, 1999.

[80] E. F. Keller and L. A. Segel. Model for chemotaxis. *Journal of Theoretical Biology*, **30**(2):225–234, 1971.

[81] E. F. Keller and L. A. Segel. Traveling bands of chemotactic bacteria: a theoretical analysis. *Journal of Theoretical Biology*, **30**(2):235–248, 1971.

[82] T. M. Konijn, D. Barkley, Y. Y. Chang, and J. T. Bonner. Cyclic AMP: a naturally occurring acrasin in the cellular slime molds. *American Naturalist*, pages 225–233, 1968.

[83] D. Kuzmin. A guide to numerical methods for transport equations. *University of Nürnberg*, 2010.

[84] L. Lamalice, F. Le Boeuf, and J. Huot. Endothelial cell migration during angiogenesis. *Circulation research*, **100**(6):782–794, 2007.

[85] R. D. Lazarov, L. Tobiska, and P. Vassilevski. Streamline diffusion least-squares mixed finite element methods for convection-diffusion problems. *East West Journal of Numerical Mathematics*, **5**:249–264, 1997.

[86] H. Levine, D. A. Kessler, and W. J. Rappel. Directional sensing in eukaryotic chemotaxis: a balanced inactivation model. *Proceedings of the National Academy of Sciences*, **103**(26):9761, 2006.

[87] F. Mac Gabhann, J. W Ji, and A. S. Popel. Computational model of vascular endothelial growth factor spatial distribution in muscle and pro-angiogenic cell therapy. *PLoS Computational Biology*, **2**(9):127, 2006.

[88] R. M. Macnab and D. E. Koshland Jr. The gradient-sensing mechanism in bacterial chemotaxis. *Proceedings of the National Academy of Sciences*, **69**(9):2509–2512, 1972.

[89] T. A. Manteuffel. An incomplete factorization technique for positive definite linear systems. *Mathematics of computation*, **34**(150):473–497, 1980.

[90] N.V. Mantzaris, S. Webb, and H.G. Othmer. Mathematical modeling of tumor-induced angiogenesis. *Journal of Mathematical Biology*, **49**(2):111–187, 2004.

[91] J. M. Mato, P. J. Van Haastert, F. A. Krens, E. H. Rhijnsburger, F. C. Dobbe, and T. M. Konijn. Cyclic AMP and folic acid mediated cyclic GMP accumulation in *Dictyostelium discoideum*. *FEBS letters*, **79**(2):331, 1977.

[92] J.M. Mato, A. Losada, V. Nanjundiah, and T.M. Konijn. Signal input for a chemotactic response in the cellular slime mold *dictyostelium discoideum*. *Proceedings of the National Academy of Sciences*, **72**(12):4991, 1975.

[93] B. Maury. Direct simulations of 2d fluid-particle flows in biperiodic domains. *Journal of computational physics*, **156**(2):325–351, 1999.

[94] R. McLennan, L. Dyson, K.W. Prather, J.A. Morrison, R.E. Baker, P.K. Maini, and P.M. Kulesa. Multiscale mechanisms of cell migration during development: theory and experiment. *Development*, **139**(16):2935–2944, 2012.

[95] R. Moricard and J. Bossu. Arrival of fertilizing sperm at the follicular cell of the secondary oocyte; a study of the rat. *Fertility and sterility*, **2**(3):260, 1951.

[96] M. F. Murphy, G. H. Golub, and A. J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM Journal on Scientific Computing*, **21**(6):1969–1972, 2000.

[97] J. D. Murray. *Mathematical biology: I. An introduction*, **2**. Springer, 2002.

[98] Freie Universität Berlin Mutzel Group, Department of biology. Slime mold aggregates.

[99] M. P. Neilson, D. M. Veltman, P. J. M. van Haastert, S. D. Webb, J. A. Mackenzie, and R. H. Insall. Chemotaxis: a feedback-based computational model robustly predicts multiple aspects of real cell behaviour. *PLoS Biology*, **9**(5):e1000618, 2011.

[100] S. I. Nishimura and M. Sasai. Modulation of the reaction rate of regulating protein induces large morphological and motional change of amoebic cell. *Journal of Theoretical Biology*, **245**(2):230–237, 2007.

[101] J. M. Oliver, J. R. King, K. J. McKinlay, P. D. Brown, D. M. Grant, C. A. Scotchford, and J. V. Wood. Thin-film theories for two-phase reactive flow models of active cell motion. *Mathematical Medicine and Biology*, **22**(1):53–98, 2005.

[102] S. Osher and R. P. Fedkiw. Level set methods: an overview and some recent results. *Journal of Computational physics*, **169**(2):463–502, 2001.

[103] S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations. *Journal of Computational Physics*, **79**(1):12–49, 1988.

[104] H. G. Othmer, S. R. Dunbar, and W. Alt. Models of dispersal in biological systems. *Journal of mathematical biology*, **26**(3):263–298, 1988.

[105] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, **12**(4):617–629, 1975.

[106] C. A. Parent and P. N. Devreotes. A cell's sense of direction. *Science*, **284**(5415):765–770, 1999.

[107] H. Parnas and L. A. Segel. Computer evidence concerning the chemotactic signal in *dictyostelium discoideum*. *Journal Of Cell Science*, **25**(1):191–204, 1977.

[108] C. S. Patlak. Random walk with persistence and external bias. *The Bulletin of mathematical biophysics*, **15**(3):311–338, 1953.

[109] G. J. Pettet, H. M. Byrne, D. L. S. McElwain, and J. Norbury. A model of wound-healing angiogenesis in soft tissue. *Mathematical biosciences*, **136**(1):35–63, 1996.

[110] W. F. P. Pfeffer. *Locomotorische Richtungsbewegungen durch chemische Reize*. 1884.

[111] A. D. Polyanin. *Handbook of linear partial differential equations for engineers and scientists*. Chapman & Hall/CRC, 2001.

[112] M. Postma, L. Bosgraaf, H.M. Loovers, and P.J.M. Van Haastert. Chemotaxis: signalling modules join hands at front and tail. *EMBO reports*, **5**(1):35–40, 2004.

[113] C. E. Powell and D. Silvester. Optimal preconditioning for Raviart–Thomas mixed formulation of second-order elliptic problems. *SIAM Journal on Matrix Analysis and Applications*, **25**:718, 2003.

[114] R. K. Raman, Y. Hashimoto, M. H. Cohen, and A. Robertson. Differentiation for aggregation in the cellular slime molds: the emergence of autonomously signalling cells in *dictyostelium discoideum*. *Journal of cell science*, **21**(2):243–259, 1976.

[115] P. Raviart and J. Thomas. A mixed finite element method for 2-nd order elliptic problems. *Mathematical aspects of finite element methods*, pages 292–315, 1977.

[116] R. Ross. The fibroblast and wound repair. *Biological Reviews*, **43**(1):51–91, 1968.

[117] J. W. Ruge and K. Stüben. Algebraic multigrid. *Multigrid methods*, **3**:73–130, 1987.

[118] Y. Saad. *Iterative methods for sparse linear systems*, **620**. SIAM, Philadelphia, PA, USA, 1996.

[119] Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, **7**(3):856–869, 1986.

[120] J. San Martín, J-F. Scheid, T. Takahashi, and M. Tucsnak. Convergence of the Lagrange–Galerkin method for the equations modelling the motion of a fluid-rigid system. *SIAM Journal on Numerical Analysis*, **43**(4):1536–1571, 2005.

[121] W. E. Schiesser. *The numerical method of lines: integration of partial differential equations*, **29**. Academic Press San Diego, 1991.

[122] A. J. Singer, R. A. Clark, et al. Cutaneous wound healing. *New England Journal of Medecine*, **341**(10):738–746, 1999.

[123] H-J. Song, G. Ming, and M. Poo. cAMP-induced switching in turning direction of nerve growth cones. *Nature*, **388**(6639):275–279, 1997.

[124] T. M. Svitkina, A. B. Verkhovsky, K. M. McQuade, and G. G. Borisy. Analysis of the actin–myosin ii system in fish epidermal keratocytes: mechanism of cell body translocation. *The Journal of Cell Biology*, **139**(2):397–415, 1997.

[125] C. Taylor and P. Hood. A numerical solution of the Navier-Stokes equations using the finite element technique. *Computers & Fluids*, **1**(1):73–100, 1973.

[126] L. N. Trefethen et al. *Chebfun Version 4.2*. The Chebfun Development Team, `http://www.maths.ox.ac.uk/chebfun/`, 2011.

[127] J. Verwer, B. P. Sommeijer, and W. Hundsdorfer. Rkc time-stepping for advection–diffusion–reaction problems. *Journal of Computational Physics*, **201**(1):61–79, 2004.

[128] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet. Novel type of phase transition in a system of self-driven particles. *Physical Review Letters*, **75**(6):1226–1229, 1995.

[129] G. E. Ward, C. J. Brokaw, D. L. Garbers, and V. D. Vacquier. Chemotaxis of *arbacia punctulata* spermatozoa to resact, a peptide from the egg jelly layer. *The Journal of Cell Biology*, **101**(6):2324–2329, 1985.

[130] C. Xue and H. Othmer. Multiscale models of taxis-driven patterning in bacterial populations. *SIAM Journal on Applied Mathematics*, **70**(1):133–167, 2009.

[131] L. Yang, J. C. Effler, B. L Kutscher, S. E. Sullivan, D. N. Robinson, and P. A. Iglesias. Modeling cellular deformations using the level set formalism. *BMC systems biology*, **2**(1):68, 2008.

[132] S. H. Zigmond. Chemotaxis by polymorphonuclear leukocytes. *The Journal of Cell Biology*, **77**(2):269–287, 1978.

[133] R. Zwanzig. Diffusion-controlled ligand binding to spheres partially covered by receptors: an effective medium treatment. *Proceedings of the National Academy of Sciences*, **87**(15):5856–5857, 1990.