UNIVERSITY OF BATH

DEPARTMENT OF MATHEMATICAL SCIENCES

Iterative Solution of Saddle Point Problems Using Divergence-free Finite Elements with Applications to Groundwater Flow

Submitted by

Robert Scheichl

for the degree of Doctor of Philosophy
of the

University of Bath

2002

Iterative Solution of Saddle Point Problems Using Divergence-free Finite Elements with Applications to Groundwater Flow

Submitted by

Robert Scheichl

for the degree of Doctor of Philosophy
of the

University of Bath

2002

Summary

In this thesis we consider the numerical solution of second-order elliptic mixed (Dirichlet / Neumann) boundary value problems using a mixed discretisation by Raviart-Thomas-Nédélec elements on two/three-dimensional domains with applications to groundwater flow. The resulting large, sparse and highly ill-conditioned saddle-point systems are solved iteratively using a decoupling strategy, thus leading to a smaller, symmetric positive definite system for the velocity unknown and a triangular system for the pressure unknown.

The decoupling process makes essential use of a basis for the subspace of divergence-free Raviart-Thomas-Nédélec elements and of a complementary basis for it. The divergence-free basis is constructed from the curls of suitable stream functions and vector potentials, in 2D and 3D respectively. The complementary basis, on the other hand, can be chosen to be a particular subset of the original Raviart-Thomas-Nédélec basis. Note in particular that in 3D the construction involves a spanning tree of the finite element mesh. This tree is shown to be easily obtained.

It is shown theoretically and numerically that in conjunction with an efficient parallel two-level overlapping additive Schwarz preconditioner, the decoupled iterative method is almost independent of mesh refinement and jumping coefficients in 2D. Moreover, it is in theory even asymptotically optimal in 2D, if the diameter of the coarse mesh is proportional to the amount of overlap in the subdomains.

The decoupled symmetric positive definite velocity system in 3D is (unfortunately) much harder to analyse, to precondition and to solve than in 2D. This is carefully explained, and a heuristic criterion on the choice of spanning tree is given, which ensures that the decoupled iterative method outperforms established solvers for the same problem even in 3D.

The method is also proven to be extremely robust when applied to actual groundwater flow problems, even in the extreme case of stochastically determined coefficients. In addition, almost optimal parallel efficiency of our implementation of the method is observed.

Acknowledgements

My thanks go first of all to the person that contributed by far the most to this thesis: my supervisor and friend Ivan Graham. Not only was he a very competent, enthusiastic, and supportive mentor, he was also a great friend. Thank you very much Ivan!

A special thank-you goes also to Andrew Cliffe, my industrial advisor at AEA Technology. Although his diary was always bursting with appointments, he always had time to answer my questions or to come for a meeting. It was his initiative that lead to the fruitful collaboration on the heterogeneous media stuff, for which I also owe Linda Stals a large thank-you.

Great thanks are also offered to all the staff and students of the Department of Mathematical Sciences, in particular to all my office mates, past and present, for putting up with me (you all know what I mean!) and for providing continual interest, entertainment and good advice. Special thanks to the "pure guys", Rich, Alfi and Savvas, for helping me with the algebraic topology, to Jon for helping me with Unix and Latex (even if it usually took a minute for him to answer) and to Steve, for being the most enthusiastic person I ever met. I would also like to thank Mark, Jill, Nada and Mary and all the rest of computing support and Admin for their help, and thank the members of the Numerical Analysis group for all the interesting seminars and discussions – in particular Adrian for making even a mathematics talk worth going to!

Furthermore, I would also like to thank the following colleagues for fruitful discussions and useful references: Tom Russell, Pete Jimack, Daniel Loghin, Dave Silvester, Ronald Hoppe, Ulrich Langer, Boris Khoromskij, Wolfgang Hackbusch, Andy Wood, and in particular Andrew Swann for helping me with the proof to Lemma 3.30.

Special thanks are due to the Department, the EPSRC and AEA Technology plc. for sponsoring this project; to UK Nirex Ltd. for kindly allowing us to use their data for the Sellafield site; and to SIAM for giving me all those dollars and for making me famous. A special thank-you is also offered to Sarah Watson at AEA Technology for helping me with `NAMMU` and for checking my descriptions of the geological models.

I would like to particularly thank my family for all the support and encouragement throughout my undergraduate and postgraduate years. Without their belief in me, none of this would have been possible.

Last but certainly not least, I would like to thank all the friends that I made over the last three years here in Bath, for making my life fun. A special thank-you goes to the ones who put up with me in the last few months of writing up, for keeping me sane and for never ceasing to give me encouragement. In particular, I would like to mention Eamo, Dani, Rebecca, Ibon (for being the best "climbing" mate I ever had) and Vicky (for putting up with all my ups and downs for the whole last year). Thanks mates!

Contents

1 Introduction **1**

1.1 The subject of the thesis 1

1.2 The aims of the thesis 4

1.3 The main achievements of the thesis 5

1.4 The structure of the thesis 6

2 Mixed Second-order Elliptic Problems **8**

2.1 Problem definition 8

2.1.1 Mixed formulations 9

2.1.2 Existence and uniqueness 12

2.1.3 Properties of $H_{0,N}(\mathrm{div},\Omega)$ relative to a partitioning of Ω 14

2.2 Mixed finite element approximation 15

2.2.1 Abstract theory 15

2.2.2 Raviart-Thomas-Nédélec elements 16

2.2.3 Error estimates 22

2.3 The resulting saddle point system 24

2.3.1 Derivation of the matrix form 24

2.3.2 Analysis of the spectrum 26

2.3.3 Iterative solution 31

2.4 Summary 37

3 Divergence-free Elements **38**

3.1 The two-dimensional case 40

3.1.1 The stream function space – C^0-elements in $H^1(\Omega)$ 40

3.1.2 The space $\mathring{\mathcal{V}}$ – an elementary approach for $k = 0$ 43

3.1.3 The space $\mathring{\mathcal{V}}$ – the general case 47

3.1.4 Extension to multiply connected domains 50

3.2 The three-dimensional case 54

3.2.1 Vector potentials – Nédélec's edge elements in $H(\vec{\mathrm{curl}},\Omega)$ 54

3.2.2 The space $\mathring{\mathcal{V}}$ – a graph theoretical approach for $k = 0$ 59

3.2.3 Literature on spanning tree techniques for finite elements 68

3.2.4 The space $\mathring{\mathcal{V}}$ – the general case 68

3.3 The complementary space $\mathcal{V}^c$ . . . 71
3.4 Summary . . . 74

4 A Decoupled Iterative Method **76**
4.1 Decoupling procedure . . . 77
4.1.1 Abstract algebraic process . . . 78
4.1.2 Particular case of mixed finite element system . . . 79
4.1.3 Implementation and analysis of the pressure system . . . 81
4.2 The velocity system in 2D . . . 82
4.2.1 Implementation . . . 82
4.2.2 Solution of bordered systems by block elimination . . . 85
4.2.3 Analysis . . . 86
4.2.4 Parallel iterative solution . . . 89
4.2.5 Extensions . . . 91
4.3 The velocity system in 3D . . . 92
4.3.1 Implementation . . . 92
4.3.2 Analysis . . . 95
4.3.3 Extensions . . . 101
4.4 Non-zero divergence - Static condensation . . . 102
4.4.1 Two-level approach . . . 103
4.4.2 Extension – multi-level approach . . . 106
4.5 Numerical Results . . . 108
4.5.1 The two-dimensional case . . . 108
4.5.2 The three-dimensional case . . . 117
4.6 Summary . . . 125

5 Applications in Groundwater Flow **128**
5.1 Layered media . . . 131
5.1.1 The Harwell site . . . 132
5.1.2 The Sellafield site . . . 134
5.1.3 Parallel Efficiency . . . 138
5.2 Heterogeneous media . . . 141
5.2.1 Stochastic modelling of heterogeneous media . . . 142
5.2.2 Numerical solution of a model problem . . . 145
5.2.3 Selection of a stopping criterion . . . 147
5.2.4 Performance of the iterative method . . . 150
5.3 Summary . . . 155

A Asymptotic Mesh Dependency **157**

B Some Results from Graph Theory **160**

C A Topological Result on Simplicial Triangulations **163**

D List of Notations **166**

References **170**

Chapter 1

Introduction

1.1 The subject of the thesis

Partial differential equations (PDEs) play a major rôle in the macroscopic modelling of many processes in continuum physics and mechanics. For the mathematician, they usually serve as the starting point for his/her investigations. As mathematical models, PDEs are generally grouped into linear and non-linear PDEs, and characterised by their order and type. We distinguish between elliptic, parabolic and hyperbolic PDEs, and this distinction is not an arbitrary one. The behaviour of each type of PDE is fundamentally different, and the techniques for their analysis and solution are designed to capture this individual behaviour.

In this thesis, we will consider the scalar linear second-order elliptic PDE

$$-\mathrm{div}\,\left(D(\vec{x})\vec{\nabla}p(\vec{x})\right) = f(\vec{x}), \tag{1.1}$$

in a bounded region $\Omega \subset \mathbb{R}^d$, with $d = 2, 3$, subject to appropriate conditions on the boundary $\partial\Omega$ of Ω. In many applications, $\vec{u}(\vec{x}) := D(\vec{x})\vec{\nabla}p(\vec{x})$ rather than p is the interesting variable. It is therefore of great interest to investigate the saddle point problem

$$\left.\begin{array}{rcl} \vec{u}(\vec{x}) \;-\; D(\vec{x})\vec{\nabla}p(\vec{x}) & = & 0 \\ -\;\mathrm{div}(\vec{u}(\vec{x})) & = & f(\vec{x}) \end{array}\right\} \tag{1.2}$$

that arises from problem (1.1) when we introduce the auxiliary variable $\vec{u}$.

The application we have in mind is the modelling of single phase flow in saturated porous media. The classical equations governing this application in the steady-state case are Darcy's Law,

$$\vec{q}(\vec{x}) = -\frac{k(\vec{x})}{\mu}\left(\vec{\nabla}p_{R}(\vec{x})\right), \tag{1.3}$$

and the incompressibility constraint,

$$\mathrm{div}(\,\vec{q}(\vec{x})\,) = 0. \tag{1.4}$$

Therefore, in this case, (1.2) is the natural formulation. Here, $\vec{q}$ is the specific discharge (Darcy velocity), p_R is the residual pressure and μ is the dynamic viscosity of the fluid, whereas k denotes the permeability tensor for the porous medium.

Since (1.2) is a first order system of PDEs, the conditions on the boundary of Ω can only involve values of $\vec{u}$ and p, rather than their derivatives. We will assume that $\partial\Omega$ is partitioned into $\partial\Omega_D \cup \partial\Omega_N$, and we will consider the following mixed boundary conditions for (1.2):

$$p = g_D \quad \text{on } \partial\Omega_D \qquad \text{and} \qquad \vec{u} \cdot \vec{\nu} = g_N \quad \text{on } \partial\Omega_N, \tag{1.5}$$

where $\vec{\nu}(\vec{x})$ denotes the outward unit normal from Ω at $\vec{x} \in \partial\Omega_N$. In the framework of (1.1), the first condition in (1.5) is called a Dirichlet condition and the second condition is called a Neumann condition.

The boundary value problem of finding functions $\vec{u}$ and p which satisfy (1.2), (1.5), is a classical problem and has been the subject of investigations for many years. If the region Ω is "simple" (e.g. unit square, unit disk, unit cube, unit ball, etc.) and if the data $D(\vec{x})$, $f(\vec{x})$, $g_D(\vec{x})$, and $g_N(\vec{x})$ are sufficiently smooth, it is possible to give explicit analytical solutions. However, in many applications, and in particular in the case of flow in porous media, Ω is more complicated and the data is not smooth. Different types of porous media can have vastly different properties. Therefore, the permeability tensor k in (1.3) will in general be highly discontinuous throughout Ω, and it is not possible to find explicit solutions. In this case, it is the job of the numerical analyst to provide approximate solutions to (1.2), (1.5) through computational simulation.

The methods to solve (1.2), (1.5) numerically are many, but there are some fundamental properties that they all have to satisfy. The approximate solution has to be, in some sense, "close" to the exact solution of (1.2), (1.5), not only in terms of the local and global error, but also in terms of some qualitative features often related to physical aspects of the underlying problem. In the case of flow in porous media, one important such feature is the local conservation of mass. A method which conserves mass locally, and which approximates the saddle point problem (1.2), (1.5) directly by choosing two approximation spaces for $\vec{u}$ and p, respectively, (without resorting to approximations of the second-order PDE (1.1)) is the mixed finite element method. One characteristic of mixed methods is that not all choices of finite element spaces will lead to convergent approximations. A common choice that leads to convergent approximations for (1.2), (1.5) is to use Raviart-Thomas-Nédélec elements for the vector valued unknown $\vec{u}$ and (discontinuous) piecewise polynomial elements for p. The resulting finite dimensional problem can be written as a system of linear equations

$$\underbrace{\begin{pmatrix} M & B \\ B^T & 0 \end{pmatrix}}_{\mathcal{M}} \begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{g} \\ \mathbf{f} \end{pmatrix} \qquad \text{in } \mathbb{R}^m \times \mathbb{R}^n, \tag{1.6}$$

It is the subject of this thesis to analyse and to solve this saddle point system, and to apply the developed methods to the calculation of groundwater flow problems.

To get acceptable accuracy for the approximation it is necessary to take the dimensions m and n of the approximation spaces to be very large (e.g. up to $O(10^6)$ in our applications), and so the matrix $\mathcal{M}$ in (1.6) is also very large. However, it is symmetric and sparse, i.e. the number of nonzero entries in each row of $\mathcal{M}$ does not depend on m or n. Therefore, only $O(m+n)$ operations are necessary to multiply a vector $\mathbf{x} \in \mathbb{R}^{m+n}$ by $\mathcal{M}$. Ideally we would also like to solve (1.6) in $O(m+n)$ operations, but even using a sophisticated direct method (like the frontal method), at least $O\left((m+n)^{(2d-1)/d} \right)$ operations are necessary to invert $\mathcal{M}$ in general. It is clear that for large m and n the cost of solving (1.6) directly rapidly becomes prohibitively large, particularly in 3D.

The alternative and indeed the only practical way to overcome this problem, is to solve (1.6) iteratively. Most classical iterative methods (like Jacobi, Gauss-Seidel, conjugate gradients, and other Krylov subspace methods) are based on multiplications with $\mathcal{M}$ and/or on row-wise relaxation, so that each iteration requires $O(m+n)$ operations. Unfortunately the number of iterations which are necessary to reach a required accuracy for the approximate solution usually depends on the condition number of the matrix, and we will see that in the case of (1.6), under reasonable mesh regularity assumptions, the condition number $\kappa(\mathcal{M}) = O(\,(m+n)^{1/d}\,)$.

However, we can multiply (1.6) from the left by a matrix $\mathcal{P}^{-1}$, and the solution of the resulting system

$$\mathcal{P}^{-1}\mathcal{M}\begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \mathcal{P}^{-1}\begin{pmatrix} \mathbf{g} \\ \mathbf{f} \end{pmatrix} \tag{1.7}$$

is still a solution of (1.6). This process is called preconditioning and the matrix $\mathcal{P}^{-1}$ is called a preconditioner. The convergence of an iterative method applied to (1.7) now depends on $\kappa(\mathcal{P}^{-1}\mathcal{M})$, and if we choose $\mathcal{P}^{-1}$ in such a way that

$$\kappa(\mathcal{P}^{-1}\mathcal{M}) = O(1), \qquad (\textit{Requirement 1})$$

then the number of iterations will not grow, as m and n get larger, and we have achieved our goal of solving (1.6) in $O(m+n)$ operations. The optimal choice would be to set $\mathcal{P}^{-1} := \mathcal{M}^{-1}$, then $\kappa(\mathcal{P}^{-1}\mathcal{M}) = 1$, and most iterative methods would converge in 1 iteration, but this just leads us back to the problem of finding the inverse of $\mathcal{M}$. Thus, as a second requirement on $\mathcal{P}^{-1}$, it is necessary that the operation

$$\mathbf{y} := \mathcal{P}^{-1}\mathbf{x} \quad \text{is cheap,} \qquad (\textit{Requirement 2})$$

(ideally $O(m+n)$ operations). Indeed, preconditioners $\mathcal{P}^{-1}$ which fulfill both requirements (at least in theory) are available for many finite element systems (e.g. multigrid, overlapping domain decomposition and multilevel methods), but they often rely on the positivity of the spectrum of the system matrix.

An additional difficulty in the case of (1.6) is the saddle point form. Because of this, the matrix is indefinite, i.e. it has negative and positive eigenvalues, and so almost all approaches to solve (1.6) efficiently contain at some point the reduction of the system to a positive definite system. In this thesis we will consider a method which is based on one such approach. It involves the construction of a basis for the subspace of divergence-free Raviart-Thomas-Nédélec elements, and therefore apart from describing the method in detail, we will also devote a large part of the thesis to the construction of such a basis.

The research for this thesis has been funded by a CASE award from the EPSRC and AEA Technology. Our industrial collaborators at AEA Technology have developed and market the computer package `NAMMU` (Numerical Assessment Method for Migration Underground) which can be used to numerically model two and three-dimensional groundwater flow through complicated regions with varying geological properties. It is of great interest to them to improve their code by implementing new, more efficient and robust methods for the solution of the arising linear equation systems. Apart from some simple model problems, which we use to study the asymptotic behaviour of our method, we will therefore also apply our method to two of AEA Technology's case studies from two sites in the UK, as well as to a model problem with heterogeneous permeability k, modelled using a Gaussian random field. A placement within the Environmental Assessment Group of AEA Technology has been invaluable in putting the results of these experiments in context.

1.2 The aims of the thesis

The main aim of this thesis is to provide a fast, efficient and robust iterative method for the numerical solution of saddle point problems of the form (1.2), in particular for applications in groundwater flow where the domain Ω can be complicated and where the coefficient $D(\vec{x})$ is usually highly discontinuous. In order to obtain such a method, it is crucial to gain first of all an understanding of the properties of the underlying continuous problem (1.2), as well as of its finite dimensional approximation (1.6). In particular, it is interesting to obtain bounds on the spectrum and on the condition number of the matrix $\mathcal{M}$ in (1.6).

In the construction of our iterative method, we will exploit particular properties of (1.2), which will lead us to a very interesting subproblem in the area of mixed finite element spaces: the construction of a basis for the subspace of divergence-free Raviart-Thomas-Nédélec elements. This problem has been addressed before in the literature, but with further assumptions on the domain Ω, on the boundary conditions (1.5) and/or on the finite element mesh. In particular, the theory for three-dimensional domains Ω and/or mixed boundary conditions is not complete. Therefore, a further aim of this thesis is to close this gap.

As we shall see, our iterative method uses a decoupling of the velocity part $\mathbf{u}$ from the pressure part $\mathbf{p}$ in (1.6), and thus consists of several components. Only after a careful analysis of the cost of each one of these components, will it be possible to claim that our method is efficient. The core task in our method will turn out to be the solution of a sparse, symmetric positive definite system. We will see that the efficiency of our method hinges on the cost of this process. Therefore, in this thesis we will carefully analyse the properties of this system and try to find a preconditioner $\mathcal{P}^{-1}$ for it that is robust (Requirement 1) and cheap (Requirement 2).

As we mentioned above, from a commercial point of view it is of great interest to our industrial sponsor AEA Technology whether an implementation of our method could improve the efficiency and applicability of their groundwater flow simulation code. Therefore we will also aim in this thesis to confirm the efficiency and robustness of our method for some of AEA Technology's actual environmental case studies.

1.3 The main achievements of the thesis

The main achievements of this thesis can be summarised as follows.

(i) An asymptotic bound for the spectral condition number of the indefinite matrix $\mathcal{M}$ in (1.6) in terms of the mesh diameter has been found. This bound is sharp for quasi-uniform triangulations.

(ii) A basis for the subspace of divergence-free Raviart-Thomas elements of arbitrary order has been constructed from the curls of suitable stream functions in 2D. In particular, the extensions to multiply connected domains and mixed boundary conditions are completely original.

(iii) The construction of an explicit basis for the subspace of divergence-free Raviart-Thomas-Nédélec elements has also been achieved for the lowest order case in 3D, using the curls of suitable vector potentials and a spanning tree of the mesh.

(iv) An efficient iterative method for (1.6) has been developed, which is applicable to two and three-dimensional problems. This method decouples the problem of finding the vector $\mathbf{u}$ in (1.6) from the problem of finding the vector $\mathbf{p}$, and reduces (1.6) to a smaller symmetric positive definite system for the velocity $\mathbf{u}$ and to a triangular system for the pressure $\mathbf{p}$. It makes essential use of the explicit bases in (ii) and (iii) and of a complementary basis in the Raviart-Thomas(-Nédélec) space which has been chosen in a special way.

(v) It has been shown theoretically and numerically that in conjunction with an efficient parallel two-level overlapping additive Schwarz preconditioner, the decoupled method in (iv) is almost independent of mesh refinement and jumping coefficients in 2D. Moreover, it is in theory even asymptotically optimal in 2D, if the diameter of the coarse mesh is proportional to the amount of overlap in the subdomains.

(vi) The decoupled symmetric positive definite velocity system in 3D is (unfortunately) much harder to analyse, to precondition and to solve than in 2D. This has been carefully explained, and a heuristic criterion on the choice of spanning tree in (iii) has been given, which ensures that the decoupled method in (iv) outperforms established solvers for (1.6) even in 3D.

(vii) The method in (iv) has been proven to be extremely robust when applied to actual two-dimensional groundwater flow problems, even in the extreme case when the permeability k in (1.3) is a realisation of a stochastic spatial process. In addition, almost optimal parallel efficiency of our implementation of the method has been observed.

1.4 The structure of the thesis

Before presenting a detailed layout of the chapters of this thesis, we would like to make some general remarks about its structure. The four main chapters of this thesis are supposed to reflect the four main issues which have been addressed. The preamble of each chapter will contain a motivation of the subsequent work and a careful review of related literature. In a short summary at the end of each chapter we will briefly recollect the main results, draw some conclusions and outline perspectives for possible future work.

In Chapter 2 we define a mixed boundary value problem for second-order elliptic partial differential equations of Poisson-type over two and three-dimensional domains, as presented in Section 1.1. We establish existence and uniqueness of solutions of the relevant mixed variational problem, and present a mixed discretisation by Raviart-Thomas(-Nédélec) finite elements, which leads to a convergent approximation. This is followed by an analysis of the spectrum of the resulting finite element stiffness matrix.

In Chapter 3 we investigate the subspace $\mathring{\mathcal{V}}$ of divergence-free Raviart-Thomas(-Nédélec) elements and its complementary space $\mathcal{V}^c$. In particular, we are interested in finding bases for these spaces. The basis for $\mathring{\mathcal{V}}$ is constructed from the curls of suitable stream functions and vector potentials in 2D and 3D, respectively. In order to do this, we need to first review the space of $H^1(\Omega)$-conforming C^0 (Lagrange) elements in 2D and the space of $H(\vec{\text{curl}}, \Omega)$-conforming Nédélec (edge) elements in 3D. These reviews are followed by a series of Propositions, Lemmas, Theorems and Corollaries in which it is actually proved for different types of domains and/or boundary conditions that the constructed sets of Raviart-Thomas(-Nédélec) functions form a basis for $\mathring{\mathcal{V}}$ in each case. The proofs for the lowest order case in 3D involve some fundamental notions and results from Graph Theory and Algebraic Topology, which we will briefly review in Appendices B and C. Finally we present a simple algorithm for the construction of a basis for the complementary space $\mathcal{V}^c$ in the lowest order case and give a proof that it is indeed a basis.

In Chapter 4 we develop, analyse and test an iterative method for solving saddle point systems of the form (1.6), arising from mixed discretisations of second-order elliptic problems as discussed in Chapter 2. The central idea is the decoupling of the velocity unknown $\mathbf{u}$ in (1.6) from the pressure unknown $\mathbf{p}$ by using the bases for $\mathring{\mathcal{V}}$ and $\mathcal{V}^c$ constructed in Chapter 3. We introduce the decoupling process as an abstract algebraic procedure and analyse the resulting decoupled systems for $\mathbf{u}$ and $\mathbf{p}$ in the special case of (1.6). In particular, we focus on the analysis and solution of the symmetric positive definite system for $\mathbf{u}$ and define an efficient and robust parallel domain decomposition preconditioner for it in 2D. In 3D the analysis concentrates on determining the influence of the choice of basis for $\mathring{\mathcal{V}}$ on the conditioning of the velocity system. A long series of numerical experiments on some simple 2D and 3D model problems at the end of the chapter examines the robustness and efficiency of the method.

In Chapter 5 we apply the decoupled iterative method which we constructed in Chapter 4 to realistic two-dimensional groundwater flow problems in actual case studies from two sites in the UK to test their robustness and (parallel) efficiency. This is followed in Section 5.2 by a discussion of a model problem describing flow in heterogeneous media. The results in Section 5.2 have been obtained in collaboration with K. A. Cliffe, I. G. Graham and L. Stals [28, 29], and they concern an application of the decoupled iterative method to groundwater flow problems with a stochastically determined permeability field k. In a sequence of experiments we show again the performance of the method in this case.

Throughout this thesis, the notation we use is fairly standard, but for clarity it is summarised in Appendix D. Care has been taken to ensure that by and large parameters do not take on different meanings in different parts of the thesis. Where this is necessary the specific use of the parameter will be clearly described.

Chapter 2

Mixed Second-order Elliptic Problems

2.1 Problem definition

In this section we will define the type of problem we are going to consider and describe the mathematical setting for it. We will mainly follow Brezzi and Fortin [20] and Brenner and Scott [19, Section 10].

We are interested in the solution of mixed boundary value problems for the following Poisson-type second-order elliptic partial differential equation:

$$\left.\begin{array}{rcll} -\mathrm{div}\,(D(\vec{x})\vec{\nabla} p) & = & f & \text{in } \Omega, \\ p & = & g_D & \text{on } \Gamma_D, \\ D(\vec{x})\vec{\nabla} p \cdot \vec{\nu} & = & g_N & \text{on } \Gamma_N \end{array}\right\} \tag{2.1}$$

where, unless further specified, Ω is a bounded and connected open subset of $\mathbb{R}^d$, $d = 2, 3$, with a polygonal (polyhedral) boundary Γ, which is assumed partitioned into $\Gamma_D \cup \Gamma_N$ with $\Gamma_D \neq \emptyset$. Each of Γ_D and Γ_N is assumed to consist of a finite non-empty union of intervals (polygons) of Γ and each of the intervals (polygons) in Γ_N is assumed to contain its boundary. $\vec{\nu}(\vec{x})$ denotes the outward unit normal from Ω at $\vec{x} \in \Gamma$. The requirements on the data in (2.1) are that

$$f \in L_2(\Omega), \qquad g_D \in H^{1/2}(\Gamma_D) \quad \text{and} \qquad g_N \in L_2(\Gamma_N).$$

Furthermore, for all sets $A \subset \Omega$ of measure zero, $D(\vec{x})$ is assumed to be a $d \times d$ positive definite matrix, uniformly with respect to $\vec{x} \in \Omega \backslash A$, i.e.

$$\theta|\vec{\xi}|^2_{\mathbb{R}^d} \;\leq\; D(\vec{x})\vec{\xi} \cdot \vec{\xi} \;\leq\; \Theta|\vec{\xi}|^2_{\mathbb{R}^d}, \qquad \forall \vec{\xi} \in \mathbb{R}^d, \tag{2.2}$$

with $0 < \theta \leq \Theta$ independent of $\vec{x}$. This implies that D is invertible almost everywhere, and that the components of D and D^{-1} are in $L_\infty(\Omega)$. However, $D(\vec{x})$ can be highly discontinuous. We will now derive weak formulations of problem (2.1).

Let $p^* \in H^1(\Omega)$ be a function that coincides with g_D on Γ_D and let $H^1_{0,D}(\Omega)$ denote the Sobolev space $\{\Phi \in H^1(\Omega) : \Phi|_{\Gamma_D} = 0\}$. We define the bilinear form

$$a_{\mathcal{P}}(p, \Phi) := \int_\Omega D(\vec{x})\vec{\nabla} p \cdot \vec{\nabla}\Phi \, d\vec{x}$$

and the linear functional

$$F_{\mathcal{P}}(\Phi) := \int_\Omega f\Phi \, d\vec{x} + \int_{\Gamma_N} g_N \Phi \, ds - a_{\mathcal{P}}(p^*, \Phi).$$

Then the standard weak (or variational) form of (2.1) is to find $p := p^* + p^0$ with $p^0 \in H^1_{0,D}(\Omega)$ such that

$$a_{\mathcal{P}}(p^0, \Phi) = F_{\mathcal{P}}(\Phi) \qquad \text{for all } \Phi \in H^1_{0,D}(\Omega). \tag{2.3}$$

Since $a_{\mathcal{P}}(\cdot,\cdot)$ is bounded and coercive, and since $F_{\mathcal{P}}(\cdot)$ is bounded, problem (2.3) has a unique solution $p^0 \in H^1_{0,D}(\Omega)$ by virtue of the Lax-Milgram Theorem (see [19, Theorem 2.7.7]). For sufficiently smooth data (D, f, g_D, g_N) the solution $p = p^* + p^0$ of (2.3) is also a solution to the classical problem (2.1). Following Brezzi and Fortin [20], we shall call problem (2.3) the *primal formulation*.

2.1.1 Mixed formulations

In many applications, $\vec{u} := D(\vec{x})\vec{\nabla} p$ rather than p is the interesting variable. In groundwater flow, for instance, $\vec{u}$ describes the velocity field, which is usually more important to know than the pressure p. It is therefore of great interest to investigate the saddle point problem that arises from problem (2.3) when we introduce the auxiliary variable $\vec{u}$ and apply duality methods. With the application in mind, we will often refer to $\vec{u}$ as the velocity and to p as the pressure.

To obtain the dual problem we need to introduce a new functional space

$$H(\mathrm{div}, \Omega) := \{\vec{v} \in (L_2(\Omega))^d : \mathrm{div}\,\vec{v} \in L_2(\Omega)\}, \tag{2.4}$$

and the inner product

$$(\vec{u}, \vec{v})_{H(\mathrm{div},\Omega)} := \int_\Omega (\vec{u}\cdot\vec{v} + \mathrm{div}\,\vec{u}\,\mathrm{div}\,\vec{v}) \, d\vec{x}, \tag{2.5}$$

that makes it a Hilbert space. The norm in $H(\mathrm{div}, \Omega)$ will be defined as

$$\|\vec{v}\|_{H(\mathrm{div},\Omega)} := (\vec{v}, \vec{v})^{1/2}_{H(\mathrm{div},\Omega)}. \tag{2.6}$$

It is possible to define the normal trace $\vec{v}\cdot\vec{\nu}|_\Gamma$ of a function $\vec{v} \in H(\mathrm{div}, \Omega)$ on Γ, in $H^{-1/2}(\Gamma)$.

Lemma 2.1. *For $\vec{v} \in H(\mathrm{div}, \Omega)$, we can define $\vec{v} \cdot \vec{\nu}|_\Gamma \in H^{-1/2}(\Gamma)$ by the following Green's formula*

$$\langle \vec{v} \cdot \vec{\nu}, \Phi \rangle_\Gamma = \int_\Omega \mathrm{div}\vec{v}\, \Phi \, d\vec{x} + \int_\Omega \vec{v} \cdot \vec{\nabla} \Phi \, d\vec{x}, \qquad \text{for all } \Phi \in H^1(\Omega), \tag{2.7}$$

where the bracket $\langle \cdot, \cdot \rangle_\Gamma$ denotes the duality pairing between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$.

Proof. See Brezzi and Fortin [20, Lemma III.1.1]. □

We can use this definition of $\vec{v} \cdot \vec{\nu}|_\Gamma$ to introduce a subspace

$$H_{0,N}(\mathrm{div}, \Omega) := \{\vec{v} \in H(\mathrm{div}, \Omega) : \langle \vec{v} \cdot \vec{\nu}, \Phi \rangle_\Gamma = 0 \ \text{ for all } \Phi \in H^1_{0,D}(\Omega)\} \tag{2.8}$$

of functions $\vec{v} \in H(\mathrm{div}, \Omega)$ whose normal traces vanish on Γ_N.

We are now ready to state the weak form of the dual problem to (2.3). To fulfill non-homogeneous Neumann conditions (i.e. $g_N \not\equiv 0$) we need a classical solution p^* of problem (2.1) with $f \equiv 0$ and $g_D \equiv 0$, and we set $\vec{u}^* := D(\vec{x})\vec{\nabla} p^*$. We introduce the bilinear forms

$$m(\vec{u}, \vec{v}) := \int_\Omega D^{-1}(\vec{x})\vec{u} \cdot \vec{v} \, d\vec{x}, \tag{2.9}$$

and

$$b(\vec{v}, w) := \int_\Omega \mathrm{div}\, \vec{v}\, w \, d\vec{x}, \tag{2.10}$$

and the linear functionals

$$G^0(\vec{v}) := \langle \vec{v} \cdot \vec{\nu}, g_D \rangle_\Gamma - m(\vec{u}^*, \vec{v}), \tag{2.11}$$

and

$$F^0(w) := -\int_\Omega f w \, d\vec{x} - b(\vec{u}^*, w). \tag{2.12}$$

Then the dual problem is to find $(\vec{u} := \vec{u}^* + \vec{u}^0, p)$ with $\vec{u}^0 \in H_{0,N}(\mathrm{div}, \Omega)$ and $p \in L_2(\Omega)$ such that

$$\left.\begin{array}{lcll} m(\vec{u}^0, \vec{v}) + b(\vec{v}, p) & = & G^0(\vec{v}), & \text{for all } \vec{v} \in H_{0,N}(\mathrm{div}, \Omega), \\ b(\vec{u}^0, w) & = & F^0(w), & \text{for all } w \in L_2(\Omega). \end{array}\right\} \tag{2.13}$$

For sufficiently smooth data a solution $(\vec{u}, p)$ of (2.13) yields again a solution to the classical problem (2.1). Following Brezzi and Fortin [20], we shall call problem (2.13) the *mixed formulation*[1]. The existence and uniqueness of solutions of (2.13) is discussed below.

Remark 2.2. We observe that the treatment of non-homogeneous Neumann boundary conditions (i.e. $g_N \not\equiv 0$) is of purely analytical nature and only affects the right hand

[1] Problem (2.13) is the dual problem of (2.3), as can be seen by writing them as minimisation problems (see [20] for details).

side of (2.13). Thus their treatment is no different numerically from the homogeneous case and, since all the applications in Chapter 5 employ homogeneous Neumann boundary conditions (i.e. $g_N \equiv 0$), we simplify the presentation by assuming throughout the thesis that $g_N \equiv 0$ and therefore $\vec{u} = \vec{u}^0$ in (2.13), i.e. from now on we consider the problem

$$\left.\begin{array}{rcll} -\mathrm{div}\,(D(\vec{x})\vec{\nabla}p) &=& f & \text{in } \Omega, \\ p &=& g_D & \text{on } \Gamma_D, \\ D(\vec{x})\vec{\nabla}p\cdot\vec{\nu} &=& 0 & \text{on } \Gamma_N. \end{array}\right\} \qquad (2.14)$$

The mixed formulation (2.13) of (2.14) is then equivalent to: Find $(\vec{u}, p) \in H_{0,N}(\mathrm{div}, \Omega) \times L_2(\Omega)$ such that

$$\left.\begin{array}{lllll} m(\vec{u},\vec{v}) & + \; b(\vec{v},p) &=& G(\vec{v})\;, & \text{for all } \vec{v} \in H_{0,N}(\mathrm{div},\Omega), \\ b(\vec{u},w) & &=& F(w)\;, & \text{for all } w \in L_2(\Omega) \end{array}\right\} \qquad (2.15)$$

with the modified functionals

$$G(\vec{v}) := \langle \vec{v}\cdot\vec{\nu}, g_D\rangle_\Gamma \qquad (2.16)$$

and

$$F(w) := -\int_\Omega f w \, d\vec{x} \qquad (2.17)$$

on the right hand side.

It is important to note that in many cases, in particular for groundwater flow problems, the source term f in problem (2.14) is given in divergence form $f = \mathrm{div}\vec{f}_{\mathcal{D}}$. More generally, we have the following lemma:

Lemma 2.3. *There exists a positive constant C such that for all $f \in L_2(\Omega)$ there is a $\vec{f}_{\mathcal{D}} \in H(\mathrm{div}, \Omega)$ satisfying*

$$\mathrm{div}\,\vec{f}_{\mathcal{D}} = f$$

and

$$\|\vec{f}_{\mathcal{D}}\|_{H(\mathrm{div},\Omega)} \leq C\|f\|_{L_2(\Omega)}.$$

Proof. Let $f \in L_2(\Omega)$. There exists a unique $\Phi \in H^1_{0,D}(\Omega)$ satisfying

$$\begin{array}{rcll} -\Delta\Phi &=& f & \text{in } \Omega \\ \Phi &=& 0 & \text{on } \Gamma_D = \Gamma. \end{array}$$

If we define $\vec{f}_{\mathcal{D}} := -\vec{\nabla}\Phi$, then $\vec{f}_{\mathcal{D}} \in H(\mathrm{div}, \Omega)$ and satisfies the conditions of the Lemma. □

Now, introducing a different auxiliary variable $\vec{u} := (D(\vec{x})\vec{\nabla}p + \vec{f}_{\mathcal{D}})$ we obtain a slightly different and slightly weaker mixed formulation: Find $(\vec{u}, p) \in H_{0,N}(\mathrm{div}, \Omega) \times$

$L_2(\Omega)$ such that

$$\left.\begin{array}{rcll} m(\vec{u},\vec{v}) + b(\vec{v},p) &=& F_{\mathcal{D}}(\vec{v})\ , & \text{for all } \vec{v} \in H_{0,N}(\mathrm{div},\Omega), \\ b(\vec{u},w) &=& 0\ , & \text{for all } w \in L_2(\Omega) \end{array}\right\} \tag{2.18}$$

with

$$F_{\mathcal{D}}(\vec{v}) := m(\vec{f}_{\mathcal{D}},\vec{v}) + \langle \vec{v}\cdot\vec{\nu}, g_D\rangle_\Gamma \tag{2.19}$$

and $\vec{f}_{\mathcal{D}} \in (L_2(\Omega))^d$. For sufficiently smooth data a solution p of (2.18) is a solution of (2.15) and therefore again a solution to the classical problem (2.14). If we want to distinguish problem (2.18) from problem (2.15) above, we shall call it the *mixed formulation in divergence form.* We will see later, why the divergence form (2.18) is often easier to solve than the original mixed formulation (2.15)

2.1.2 Existence and uniqueness

We will finish the section by establishing the existence and uniqueness of solutions of (2.15) and (2.18). This analysis can be found in [20, Section II.1], although the exposition in [19, Section 10.2] is more comprehensive and we will mainly follow their analysis.

Let us first consider (2.18) and define a closed subspace

$$\mathcal{Z} := \{\vec{v} \in H_{0,N}(\mathrm{div},\Omega) : b(\vec{v},w) = 0 \text{ for all } w \in L_2(\Omega)\} \tag{2.20}$$

of $H_{0,N}(\mathrm{div},\Omega)$ and its orthogonal complement $\mathcal{Z}^\perp$ such that

$$\mathcal{Z} \oplus \mathcal{Z}^\perp = H_{0,N}(\mathrm{div},\Omega).$$

Problem (2.18) is equivalent to solving the following decoupled system. Find $(\vec{u},p) \in \mathcal{Z} \times L_2(\Omega)$ such that

$$\left.\begin{array}{rcll} m(\vec{u},\vec{v}) &=& F_{\mathcal{D}}(\vec{v})\ , & \text{for all } \vec{v} \in \mathcal{Z}, \\ b(\vec{v},p) &=& F_{\mathcal{D}}(\vec{v}) - m(\vec{u},\vec{v})\ , & \text{for all } \vec{v} \in \mathcal{Z}^\perp. \end{array}\right\} \tag{2.21}$$

In the Theorem 2.5 below we will use this formulation to show existence and uniqueness of solutions of (2.18). In order to do this we need to first verify the following properties of m and b:

Lemma 2.4.

(a) *m is coercive on $\mathcal{Z}$, i.e. there exists $\alpha > 0$ such that*

$$m(\vec{v},\vec{v}) \geq \alpha \|\vec{v}\|^2_{H(\mathrm{div},\Omega)}, \qquad \text{for all } \vec{v} \in \mathcal{Z}. \tag{2.22}$$

(b) *b fulfils the* inf-sup condition *(or Ladyzhenskaya-Babuška-Brezzi condition), i.e.*

there exists $\beta > 0$ such that

$$\sup_{\vec{v} \in H_{0,N}(\mathrm{div},\Omega)} \frac{b(\vec{v}, w)}{\|\vec{v}\|_{H(\mathrm{div},\Omega)}} \geq \beta \, \|w\|_{L_2(\Omega)}, \qquad \text{for all } w \in L_2(\Omega). \tag{2.23}$$

Proof. (a) Let $\vec{v} \in \mathcal{Z}$. Then $\|\mathrm{div}\, \vec{v}\|_{L_2(\Omega)} = b(\vec{v}, \mathrm{div}\, \vec{v}) = 0$, and using (2.9) and (2.2),

$$m(\vec{v}, \vec{v}) \geq \Theta^{-1} \|\vec{v}\|_{(L_2(\Omega))^d} = \Theta^{-1} \|\vec{v}\|_{H(\mathrm{div},\Omega)}.$$

(b) Let $w \in L_2(\Omega)$. In the same way as in the proof to Lemma 2.3 we can solve the auxiliary problem

$$\begin{aligned} -\Delta \Phi &= w && \text{in } \Omega \\ \Phi &= 0 && \text{on } \Gamma_D \\ \vec{\nabla} \Phi \cdot \vec{\nu} &= 0 && \text{on } \Gamma_N \end{aligned}$$

to find $\Phi \in H^1_{0,D}(\Omega)$. If we define $\vec{v}_w := -\vec{\nabla}\Phi$, then $\vec{v}_w \in H_{0,N}(\mathrm{div}, \Omega)$ and fulfils $\|\vec{v}_w\|_{H(\mathrm{div},\Omega)} \leq C \|w\|_{L_2(\Omega)}$ for some constant $C > 0$. Finally, using the fact that $b(\vec{v}_w, w) = \|w\|^2_{L_2(\Omega)}$ we have

$$\sup_{\vec{v} \in H_{0,N}(\mathrm{div},\Omega)} \frac{b(\vec{v}, w)}{\|\vec{v}\|_{H(\mathrm{div},\Omega)}} \geq \frac{b(\vec{v}_w, w)}{\|\vec{v}_w\|_{H(\mathrm{div},\Omega)}} \geq \frac{1}{C} \frac{b(\vec{v}_w, w)}{\|w\|_{L_2(\Omega)}} = \frac{1}{C} \|w\|_{L_2(\Omega)}.$$

□

Theorem 2.5. *Problem (2.18) has a unique solution $(\vec{u}, p) \in H_{0,N}(\mathrm{div}, \Omega) \times L_2(\Omega)$.*

Proof. We will prove this theorem by establishing the existence and uniqueness of a solution for problem (2.21).

The first equation in (2.21) has a unique solution $\vec{u} \in \mathcal{Z}$ by virtue of the Lax-Milgram Theorem (see [19, Thm. 2.7.7]), since $F_{\mathcal{D}}(\vec{v})$ and $m(\vec{u}, \vec{v})$ are bounded on $\mathcal{Z}$ and since m is coercive (cf. Lemma 2.4(a)).

Let us now look at the second equation in (2.21). Uniqueness of a solution $p \in L_2(\Omega)$ is a direct consequence of the inf-sup condition (2.23). Existence also follows from condition (2.23), but this requires a bit more explanation. Recall that $H_{0,N}(\mathrm{div}, \Omega)$ is a Hilbert space, and that therefore $\mathcal{Z}^{\perp}$ is also a Hilbert space with the inner product $(\cdot, \cdot)_{H(\mathrm{div},\Omega)}$ inherited from $H_{0,N}(\mathrm{div}, \Omega)$. Now, let $p \in L_2(\Omega)$. The linear functional $\vec{v} \to b(\vec{v}, p)$ is continuous on $\mathcal{Z}^{\perp}$, so the Riesz Representation Theorem (see [19, Thm. 2.4.2]) guarantees the existence of a linear operator $\vec{T} : L^2(\Omega) \to \mathcal{Z}^{\perp}$ such that

$$(\vec{T}p, \vec{v})_{H(\mathrm{div},\Omega)} = b(\vec{v}, p), \qquad \text{for all } \vec{v} \in \mathcal{Z}^{\perp}. \tag{2.24}$$

Moreover,

$$\|\vec{T}p\|_{H(\mathrm{div},\Omega)} = \sup_{\vec{v} \in \mathcal{Z}^{\perp}} \frac{b(\vec{v}, p)}{\|\vec{v}\|_{H(\mathrm{div},\Omega)}} \leq C \|p\|_{L_2(\Omega)},$$

where the inequality follows form the boundedness of b. Let R denote the image of $\vec{T}$ in

$\mathcal{Z}^{\perp}$. If we can show that $R = \mathcal{Z}^{\perp}$, then another application of the Riesz Representation Theorem completes the proof.

To show that $R = \mathcal{Z}^{\perp}$, we begin by showing that R is closed. Let $\vec{v} \in \mathcal{Z}^{\perp}$ and suppose that $w_j \in L_2(\Omega)$ is a sequence with the property that $\vec{T}w_j \to \vec{v}$ in $\mathcal{Z}^{\perp}$. Then $\{\vec{T}w_j\}$ is a Cauchy sequence in $\mathcal{Z}^{\perp}$. Since the action of $b(\cdot,\cdot)$ is trivial on $\mathcal{Z}$, we can use (2.23) on $\mathcal{Z}^{\perp}$ and the definition (2.24) of $\vec{T}$ to obtain

$$\begin{aligned}
\beta\|w_j - w_k\|_{L_2(\Omega)} &\leq \sup_{\vec{q}\in\mathcal{Z}^{\perp}} \frac{b(\vec{q}, w_j - w_k)}{\|\vec{q}\|_{H(\mathrm{div},\Omega)}} \\
&= \sup_{\vec{q}\in\mathcal{Z}^{\perp}} \frac{(\vec{q}, \vec{T}w_j - \vec{T}w_k)_{H(\mathrm{div},\Omega)}}{\|\vec{q}\|_{H(\mathrm{div},\Omega)}} \quad = \quad \|\vec{T}w_j - \vec{T}w_k\|_{H(\mathrm{div},\Omega)}.
\end{aligned}$$

The last equality is a simple application of the Cauchy-Schwarz Inequality. Therefore $\{w_j\}$ is a Cauchy sequence in $L_2(\Omega)$ and there exists a $w \in L_2(\Omega)$ with $w = \lim_{j\to\infty} w_j$. By the boundedness of $\vec{T}$, we know that $\vec{T}w = \vec{v}$. Therefore $\vec{v} \in R$ and so R is closed. If $R \neq \mathcal{Z}^{\perp}$, let $\vec{v} \neq \vec{0}$ be an element of $R^{\perp}$, the orthogonal complement of R in $\mathcal{Z}^{\perp}$. Then $b(\vec{v}, w) = (\vec{T}w, \vec{v})_{H(\mathrm{div},\Omega)} = 0$ for all $w \in L_2(\Omega)$. But this implies that $\vec{v} \in \mathcal{Z}$, a contradiction. So $R = \mathcal{Z}^{\perp}$ and the proof is complete. ☐

Corollary 2.6. *Problem (2.15) has a unique solution* $(\vec{u}, p) \in H_{0,N}(\mathrm{div},\Omega) \times L_2(\Omega)$.

Proof. It follows directly from Theorem 2.5 using Lemma 2.3. ☐

2.1.3 Properties of $H_{0,N}(\mathrm{div},\Omega)$ relative to a partitioning of Ω

Let Ω be partitioned into a family $\mathcal{T}$ of open subdomains $T \subset \Omega$, with polygonal (polyhedral) boundary $\partial T := \overline{T}\backslash T$, such that

$$\begin{aligned}
&(1) \qquad \overline{\Omega} = \bigcup_{T\in\mathcal{T}} \overline{T} \\
&(2) \qquad T \cap T' = \emptyset, \qquad \text{for all } T \neq T' \in \mathcal{T}.
\end{aligned}$$

Furthermore, we will denote by $\vec{\nu}_T(\vec{x})$ the outward unit normal from T at $\vec{x} \in \partial T$.

We have the following characterisation of functions in $H_{0,N}(\mathrm{div},\Omega)$:

Proposition 2.7. *A function* $\vec{v} \in (L_2(\Omega))^d$ *is in* $H_{0,N}(\mathrm{div},\Omega)$*, if and only if the following two conditions hold true:*

$$\vec{v}|_T \in H(\mathrm{div},T) \qquad \text{for all } T \in \mathcal{T}, \tag{2.25}$$

$$\sum_{T\in\mathcal{T}} \langle \vec{v}\cdot\vec{\nu}_T, \Phi\rangle_{\partial T} = 0 \qquad \text{for all } \Phi \in H^1_{0,D}(\Omega). \tag{2.26}$$

Proof. Let $\vec{v} \in H_{0,N}(\mathrm{div},\Omega)$. Obviously $\vec{v}|_T \in H(\mathrm{div},T)$ for all $T \in \mathcal{T}$. Now, let

$\Phi \in H^1_{0,D}(\Omega)$. Using the Green's formula (2.7) we have

$$0 = \langle \vec{v} \cdot \vec{\nu}, \Phi \rangle_\Gamma = \int_\Omega \mathrm{div}\vec{v}\, \Phi \, d\vec{x} + \int_\Omega \vec{v} \cdot \vec{\nabla}\Phi \, d\vec{x}.$$

We can decompose the integrals on the right hand side and apply the Green's formula (2.7) in each subdomain $T \in \mathcal{T}$ to obtain condition (2.26):

$$0 = \sum_{T \in \mathcal{T}} \left\{ \int_T \mathrm{div}\vec{v}\, \Phi \, d\vec{x} + \int_T \vec{v} \cdot \vec{\nabla}\Phi \, d\vec{x} \right\} = \sum_{T \in \mathcal{T}} \langle \vec{v} \cdot \vec{\nu}_T, \Phi \rangle_{\partial T}.$$

Conversely, let $\vec{v} \in (L_2(\Omega))^d$ and assume that conditions (2.25) and (2.26) hold true. Using the same argument as above we can use Green's formula on each subdomain to show that

$$\int_\Omega \mathrm{div}\vec{v}\, \Phi \, d\vec{x} = -\int_\Omega \vec{v} \cdot \vec{\nabla}\Phi \, d\vec{x}, \qquad \text{for all } \Phi \in H^1_{0,D}(\Omega). \tag{2.27}$$

This implies for all $\Phi \in H^1_{0,D}(\Omega)$ that

$$\left| \int_\Omega \mathrm{div}\vec{v}\, \Phi \, d\vec{x} \right| \leq |\Phi|_{H^1(\Omega)} \|\vec{v}\|_{(L_2(\Omega))^d} \tag{2.28}$$

Now, let $\mathcal{D}(\Omega)$ be the linear space of infinitely differentiable functions with compact support on Ω, then $\mathcal{D}(\Omega) \subset H^1_{0,D}(\Omega)$. Since $\mathcal{D}(\Omega)$ is dense in $L_2(\Omega)$ (see Girault and Raviart [44, Lemma I.1.1]), it follows from (2.28) that $\mathrm{div}\vec{v} \in L_2(\Omega)$ and therefore $\vec{v} \in H(\mathrm{div}, \Omega)$. We can now apply Green's formula (2.7) on the whole of Ω and note that (2.27) is equivalent to $\langle \vec{v} \cdot \vec{\nu}, \Phi \rangle_\Gamma = 0$, for all $\Phi \in H^1_{0,D}(\Omega)$. Hence $\vec{v} \in H_{0,N}(\mathrm{div}, \Omega)$. □

Proposition 2.7 states that the normal traces of functions in $H_{0,N}(\mathrm{div}, \Omega)$ are continuous across any surface $\Gamma_I \subset \Omega$. This will be an essential point for finite element approximations.

2.2 Mixed finite element approximation

We will now turn the attention to the approximation of problems (2.15) and (2.18) by finite elements. Since (as a variational problem) (2.18) is a special case of (2.15) with $G(\vec{V}) = F_{\mathcal{D}}(\vec{V})$ and $F(W) = 0$, we will only look at problem (2.15). A more detailed account of the results in this section can be found in Brezzi & Fortin [20] again.

2.2.1 Abstract theory

To approximate (2.15) we choose finite dimensional subspaces $\mathcal{V} \subset H_{0,N}(\mathrm{div}, \Omega)$ and $\mathcal{W} \subset L_2(\Omega)$ and seek $(\vec{U}, P) \in \mathcal{V} \times \mathcal{W}$ such that

$$\left. \begin{array}{lcl l} m(\vec{U}, \vec{V}) \;+\; b(\vec{V}, P) & = & G(\vec{V}), & \text{for all } \vec{V} \in \mathcal{V}, \\ b(\vec{U}, W) & = & F(W), & \text{for all } W \in \mathcal{W}. \end{array} \right\} \tag{2.29}$$

In the same way as for the continuous problem, we define a closed subspace of $\mathcal{V}$:

$$\mathring{\mathcal{V}} := \{\vec{V} \in \mathcal{V} : b(\vec{V}, W) = 0 \text{ for all } W \in \mathcal{W}\}. \tag{2.30}$$

Lemma 2.8. *If* $\operatorname{div} \mathcal{V} = \mathcal{W}$, *then*

(a) $\mathring{\mathcal{V}} \subset \mathcal{Z}$,

(b) *m is coercive on $\mathring{\mathcal{V}}$, i.e. there exists an $\alpha_0 > 0$ such that*

$$m(\vec{V}, \vec{V}) \geq \alpha_0 \, \|\vec{V}\|^2_{H(\mathrm{div},\Omega)}, \qquad \text{for all } \vec{V} \in \mathring{\mathcal{V}}. \tag{2.31}$$

(c) *b satisfies the inf-sup condition on $\mathcal{V} \times \mathcal{W}$, i.e. there exists $\beta_0 > 0$ such that*

$$\sup_{\vec{V} \in \mathcal{V}} \frac{b(\vec{V}, W)}{\|\vec{V}\|_{H(\mathrm{div},\Omega)}} \geq \beta_0 \, \|W\|_{L_2(\Omega)}, \qquad \text{for all } W \in \mathcal{W}. \tag{2.32}$$

Proof. See Brezzi and Fortin [20, p. 138] □

Existence and uniqueness of a solution $(\vec{U}, P) \in \mathcal{V} \times \mathcal{W}$ of (2.29) follows, as in the continuous problem, directly from the conditions (2.31) and (2.32) in Lemma 2.8. Additionally, we can derive error estimates in terms of approximation properties of the spaces $\mathcal{V}$ and $\mathcal{W}$. We have the following theorem.

Theorem 2.9. *If $\operatorname{div} \mathcal{V} = \mathcal{W}$, then problem (2.29) has a unique solution $(\vec{U}, P) \in \mathcal{V} \times \mathcal{W}$. Moreover, if $(\vec{u}, p) \in H_{0,N}(\mathrm{div}, \Omega) \times L_2(\Omega)$ is the solution of problem (2.15), we have the estimates*

$$\|\vec{u} - \vec{U}\|_{H(\mathrm{div},\Omega)} \leq C \inf_{\vec{V} \in \mathcal{V}} \|\vec{u} - \vec{V}\|_{H(\mathrm{div},\Omega)}, \tag{2.33}$$

$$\|p - P\|_{L_2(\Omega)} \leq C \left(\inf_{W \in \mathcal{W}} \|p - W\|_{L_2(\Omega)} + \inf_{\vec{V} \in \mathcal{V}} \|\vec{u} - \vec{V}\|_{H(\mathrm{div},\Omega)} \right) \tag{2.34}$$

where C is a generic constant that depends on α_0, β_0, $\|m\|$ and $\|b\|$.

Proof. Let $\operatorname{div} \mathcal{V} = \mathcal{W}$. The proof of existence and uniqueness follows directly from Lemma 2.8 together with the general theory presented in the previous section. The derivation of the error estimates can be found in Brezzi & Fortin [20, Prop. II.2.6–7]. □

2.2.2 Raviart-Thomas-Nédélec elements

We will only consider methods based on simplicial elements for problem (2.15), and in fact restrict attention to the (most practically important) case when $\mathcal{V}$ is the space of Raviart-Thomas-Nédélec elements (in 2D they are usually only referred to as Raviart-Thomas elements). The Raviart-Thomas-Nédélec elements were introduced by Raviart and Thomas [77] and later generalised and extended to the three-dimensional case by Nédélec [73]. To define them, we need a triangulation of Ω.

Definition 2.10. (Simplicial triangulation)

(a) A *simplicial triangulation* $\mathcal{T}$ is a partitioning of Ω into open simplices $T \in \mathcal{T}$ of dimension d, i.e. triangles for $d = 2$ and tetrahedra for $d = 3$, such that

$$(1) \quad \overline{\Omega} = \bigcup_{T \in \mathcal{T}} \overline{T}$$

$$(2) \quad \overline{T} \cap \overline{T'} = \left\{ \begin{array}{l} \text{either } \emptyset \\ \text{or a } \textit{node} \text{ of } T \text{ and } T', \\ \text{or an } \textit{edge} \text{ of } T \text{ and } T', \\ \text{or a } \textit{face} \text{ of } T \text{ and } T', \end{array} \right\} \quad \text{for all } T \neq T' \in \mathcal{T}.$$

The $T \in \mathcal{T}$ are called *elements*. Let $h(T)$ denote the diameter of an element T of $\mathcal{T}$ and let $h := \max_{T \in \mathcal{T}} h(T)$. We will write $\mathcal{T}_h$ instead of $\mathcal{T}$ when we want to study asymptotics as $h \to 0$.

(b) A family of triangulations $\{\mathcal{T}_h\}$ is called *shape regular* provided there exists a number $\kappa > 0$, independent of h, such that

$$\rho(T)/h(T) \geq \kappa, \qquad \text{for all } T \in \mathcal{T}_h, \tag{2.35}$$

where $\rho(T)$ denotes the diameter of the largest circle (or ball) contained within T.

(c) A family of triangulations $\{\mathcal{T}_h\}$ is called *quasi-uniform* (or uniformly shape regular) provided there exists a number $\kappa > 0$, independent of h, such that

$$\rho(T)/h \geq \kappa, \qquad \text{for all } T \in \mathcal{T}_h. \tag{2.36}$$

Notation 2.11. In $\mathbb{R}^2$, i.e. for $d = 2$, we will use the term *faces* when we are talking about the edges of the triangles. This is in line with the notation in algebraic topology and will simplify the presentation.

Let $\mathcal{T}$ be a simplicial triangulation of Ω. Throughout the thesis we assume that the *collision points* (*interfaces* between Γ_D and Γ_N) are nodal points (or edges for $d = 3$) of the triangulation. By $\mathcal{N}$ and $\mathcal{F}$ we denote the union of all sets of nodes and faces of the elements $T \in \mathcal{T}$, respectively. Furthermore, let $\mathcal{N}_I$, $\mathcal{N}_D$ and $\mathcal{N}_N$ be subsets of $\mathcal{N}$ containing the nodes which lie in Ω, Γ_D and Γ_N, respectively. Analogously we can define the sets $\mathcal{F}_I$, $\mathcal{F}_D$ and $\mathcal{F}_N$. Finally, we define for each $F \in \mathcal{F}$ a unit normal vector

$\vec{\nu}_F$ to the face F which, for convenience, is assumed orientated so that it lies in

$$R^+ := \begin{cases} \{\vec{x} \in \mathbb{R}^2 : x_1 > 0\} \cup \left\{ \left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix} \right) \right\}, & \text{if } d = 2, \\ \{\vec{x} \in \mathbb{R}^3 : x_1 > 0\} \cup \{\vec{x} \in \mathbb{R}^3 : x_1 = 0,\ x_2 > 0\} \cup \left\{ \left(\begin{smallmatrix} 0 \\ 0 \\ 1 \end{smallmatrix} \right) \right\}, & \text{if } d = 3. \end{cases} \tag{2.37}$$

For $d = 3$ we will additionally need the set $\mathcal{E}$, which denotes the union of all sets of edges of the elements $T \in \mathcal{T}$. In the same way as above we assume that $\mathcal{E}$ can be partitioned into the subsets $\mathcal{E}_I$, $\mathcal{E}_D$ and $\mathcal{E}_N$ containing the edges which lie in Ω, Γ_D and Γ_N, respectively. On each edge $E \in \mathcal{E}$ we define a unit tangent vector $\vec{\tau}_E \in R^+$.

As usual, the Raviart-Thomas-Nédélec elements are constructed as piecewise polynomial functions on each element $T \in \mathcal{T}$. To achieve this, let k be an integer ≥ 0. On each element $T \in \mathcal{T}$ we define

$$P_k(T): \quad \text{the space of polynomials of degree } \leq k. \tag{2.38}$$

The dimension of $P_k(T)$ is $\frac{1}{2}(k+1)(k+2)$ and $\frac{1}{6}(k+1)(k+2)(k+3)$ for $d = 2$ and $d = 3$ respectively. We shall also need polynomial spaces on the faces of the elements. We define

$$R_k(\partial T) := \{\phi \in L_2(\partial T) : \phi|_F \in P_k(F) \text{ for all } F \subset \partial T\} \tag{2.39}$$

where ∂T denotes the boundary of T, and F denotes a face of T as defined above. The dimension of $R_k(\partial T)$ is $3(k+1)$ and $2(k+1)(k+2)$ for $d = 2$ and $d = 3$ respectively.

We can now define the Raviart-Thomas-Nédélec elements. For each $k \geq 0$, let

$$RT_k(T) := \{\vec{\alpha} + \gamma\,\vec{x} : \vec{x} \in T \text{ with } \vec{\alpha} \in (P_k(T))^d \text{ and } \gamma \in P_k(T)\}. \tag{2.40}$$

It can easily be checked that the dimension of $RT_k(T)$ is given by

$$\dim RT_k(T) = \begin{cases} (k+1)(k+3) & \text{for } d = 2, \\ \frac{1}{2}(k+1)(k+2)(k+4) & \text{for } d = 3. \end{cases} \tag{2.41}$$

These spaces satisfy the following properties

Proposition 2.12. *Let $T \in \mathcal{T}$. Then*

$$\operatorname{div} RT_k(T) = P_k(T). \tag{2.42}$$

Moreover, for any $\vec{v} \in RT_k(T)$

$$\vec{v} \cdot \vec{\nu}|_{\partial T} \in R_k(\partial T), \tag{2.43}$$

where $\vec{\nu}$ denotes the outward unit normal from T on ∂T.

Proof. Let $T \in \mathcal{T}$. The first statement follows directly from the definition (2.40) of $RT_k(T)$. For the second statement let $\vec{v} \in RT_k(T)$ and let $\vec{x} \in \partial T$.

$$(\vec{v} \cdot \vec{\nu})(\vec{x}) = \vec{\alpha}(\vec{x}) \cdot \vec{\nu}(\vec{x}) + \gamma(\vec{x})(\vec{x} \cdot \vec{\nu}(\vec{x}))$$

with $\vec{\alpha} \in (P_k(T))^d$ and $\gamma \in P_k(T)$. However, $\vec{\nu}(\vec{x})$ and $\vec{x} \cdot \vec{\nu}(\vec{x})$ are constant on each face $F \subset \partial T$, so that $\vec{v} \cdot \vec{\nu}|_F \in P_k(F)$, and hence $\vec{v} \cdot \vec{\nu}|_{\partial T} \in R_k(\partial T)$. □

We also have,

Proposition 2.13. (Unisolvence)
Let $T \in \mathcal{T}$ and $\vec{v} \in RT_k(T)$. The following relations imply $\vec{v} = \vec{0}$:

$$\left.\begin{aligned} &\int_{\partial T} \vec{v} \cdot \vec{\nu}\, p_k \, ds = 0, && \text{for all } p_k \in R_k(\partial T), \\ &\int_T \vec{v} \cdot \vec{p}_{k-1} \, d\vec{x} = 0, && \text{for all } \vec{p}_{k-1} \in (P_{k-1}(T))^d. \end{aligned}\right\} \tag{2.44}$$

Proof. See Raviart and Thomas [77] for $d = 2$ and Nédélec [73] for $d = 3$. □

Propositions 2.12 and 2.13 imply that we can use the following degrees of freedom to uniquely define a function $\vec{v} \in RT_k(T)$ (see Figure 2.1):

- The moments of order up to k of $\vec{v} \cdot \vec{\nu}_F$ on each face F of T, i.e.

$$\int_F \vec{v} \cdot \vec{\nu}_F \, p_k \, ds, \qquad p_k \in P_k(F).$$

- The moments of order up to $k - 1$ of $\vec{v}$ on T, for $k > 0$, i.e.

$$\int_T \vec{v} \cdot \vec{p}_{k-1} \, d\vec{x}, \qquad \vec{p}_{k-1} \in (P_{k-1}(T))^d.$$

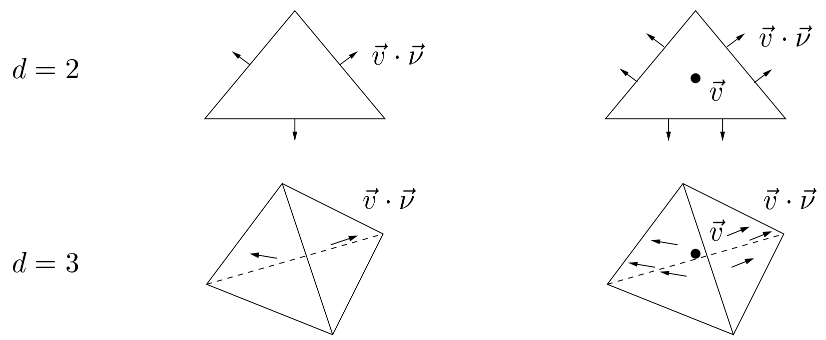


Figure 2.1: Degrees of Freedom for $RT_0(T)$ (left) and $RT_1(T)$ (right)

Let $\vec{v} \in H(\mathrm{div}, T)$. Provided $\vec{v}$ is slightly smoother than merely belonging to $H(\mathrm{div}, T)$, it is possible to define an interpolation operator

$\pi_T : H(\mathrm{div}, T) \cap (L_s(T))^d \to RT_k(T)$ for $s > 2$ such that

$$\left.\begin{aligned} &\int_{\partial T} (\vec{v} - \pi_T \vec{v}) \cdot \vec{\nu}\, p_k \, ds = 0, \qquad &&\text{for all } p_k \in R_k(\partial T), \\ &\int_T (\vec{v} - \pi_T \vec{v}) \cdot \vec{p}_{k-1} \, d\vec{x} = 0, \qquad &&\text{for all } \vec{p}_{k-1} \in (P_{k-1}(T))^d. \end{aligned}\right\} \tag{2.45}$$

Remark 2.14. The increased regularity is necessary to define the boundary integrals in (2.45), since the functions $p_k \in R_k(\partial T)$ do not belong to $H^{1/2}(\partial T)$. (From Lemma 2.1, if $\vec{v} \in H(\mathrm{div}, T)$, then the normal trace $\vec{v} \cdot \vec{\nu}$ can only be expected to lie in $H^{-1/2}(\partial T)$.)

If we furthermore define ρ_T to be the L_2-projection on $P_k(T)$ we have the following proposition, often referred to as the *commuting diagram property* (see Douglas & Roberts [35]).

Proposition 2.15.

$$\begin{array}{ccc} H(\mathrm{div}, T) \cap (L_s(T))^d & \xrightarrow{\mathrm{div}} & L_2(\Omega) \\ {\scriptstyle \pi_T}\big\downarrow & & {\scriptstyle \rho_T}\big\downarrow \\ RT_k(T) & \xrightarrow{\mathrm{div}} & P_k(T) \end{array} \tag{2.46}$$

Proof. Let $\vec{v} \in H(\mathrm{div}, T) \cap (L_s(T))^d$. From Proposition 2.12 we know that $\mathrm{div}\, \pi_T \vec{v} \in P_k(T)$. Also, using a Green's formula similar to (2.7), we have

$$\int_T w \mathrm{div}\, (\vec{v} - \pi_T \vec{v}) \, d\vec{x} = \int_T (\vec{v} - \pi_T \vec{v}) \cdot \vec{\nabla} w \, d\vec{x} - \int_{\partial T} (\vec{v} - \pi_T \vec{v}) \cdot \vec{\nu}\, w \, ds,$$

for all $w \in P_k(T)$, which is 0 by (2.45). Therefore,

$$\int_T \mathrm{div}\, \vec{v} w d\vec{x} = \int_T \mathrm{div}\, \pi_T \vec{v} w d\vec{x}, \qquad \text{for all } w \in P_k(T),$$

which means that $\mathrm{div}\, \pi_T \vec{v}$ is the L_2-projection of $\mathrm{div}\, \vec{v}$ onto $P_k(T)$, or equivalently

$$\rho_T \mathrm{div}\, \vec{v} = \mathrm{div}\, \pi_T \vec{v}.$$

□

The choice of degrees of freedom on each element $T \in \mathcal{T}$ enables us now to use Proposition 2.7 to build a finite dimensional subspace of $H(\mathrm{div}, \Omega)$ from the polynomial spaces $RT_k(T)$. We define

$$\mathcal{RT}_k(\Omega, \mathcal{T}) := \{\vec{v} \in H(\mathrm{div}, \Omega) : \vec{v}|_T \in RT_k(T) \ \text{ for all } T \in \mathcal{T}\}. \tag{2.47}$$

In a similar manner we use the polynomial spaces $P_k(T)$ to define a finite dimensional subspace of $L_2(\Omega)$:

$$\mathcal{P}_k(\Omega, \mathcal{T}) := \{w \in L_2(\Omega) : w|_T \in P_k(T) \ \text{ for all } T \in \mathcal{T}\}. \tag{2.48}$$

By the commuting diagram property (2.46) we have

$$\mathcal{P}_k(\Omega,\mathcal{T}) = \operatorname{div}\mathcal{RT}_k(\Omega,\mathcal{T}). \tag{2.49}$$

Finally, to obtain a finite dimensional subspace $\mathcal{V}$ of $H_{0,N}(\operatorname{div},\Omega)$ we simply set

$$\mathcal{V} := \{\vec{v} \in \mathcal{RT}_k(\Omega,\mathcal{T}) : \vec{v}\cdot\vec{\nu}|_{\Gamma_N} = 0\}. \tag{2.50}$$

The dimension of $\mathcal{V}$ can be calculated easily from the dimensions of the spaces $RT_k(T)$ taking into account the continuity of the normal component in $H(\operatorname{div},\Omega)$ on the interface between two elements. We have

$$n_{\mathcal{V}} := \dim\mathcal{V} = \begin{cases} (k+1)(\#\mathcal{F}_I + \#\mathcal{F}_D + k\,\#\mathcal{T}) & \text{for } d=2, \\ \frac{1}{2}(k+1)(k+2)(\#\mathcal{F}_I + \#\mathcal{F}_D + k\,\#\mathcal{T}) & \text{for } d=3 \end{cases} \tag{2.51}$$

where, throughout, $\#A$ denotes the number of elements of a (finite) set A.

To guarantee existence and uniqueness of a solution $(\vec{U},P) \in \mathcal{V}\times\mathcal{W}$ of (2.29), in view of Theorem 2.9, we choose

$$\mathcal{W} := \operatorname{div}\mathcal{V} = \mathcal{P}_k(\Omega,\mathcal{T}), \tag{2.52}$$

The dimension of $\mathcal{W}$ is

$$n_{\mathcal{W}} := \dim\mathcal{W} = \begin{cases} \frac{1}{2}(k+1)(k+2)\,\#\mathcal{T} & \text{for } d=2, \\ \frac{1}{6}(k+1)(k+2)(k+3)\,\#\mathcal{T} & \text{for } d=3. \end{cases} \tag{2.53}$$

Example 2.16. *The lowest order case:* $k = 0$

The most interesting case from the computational point of view is the lowest order case, i.e. $k = 0$, especially when we can not expect high regularity of the solution $(\vec{u},p)$ of the continuous problem (2.15).

In the case $k = 0$ the functions $\vec{v} \in \mathcal{V}$ have a particularly simple form. For each $T \in \mathcal{T}$, there exist $\vec{\alpha} \in \mathbb{R}^d$ and $\gamma \in \mathbb{R}$ such that

$$\vec{v}(\vec{x}) = \vec{\alpha} + \gamma\vec{x}, \qquad \text{for all } \vec{x} \in T. \tag{2.54}$$

From Proposition 2.12 we know that

$$\vec{v}\cdot\vec{\nu}_F|_F = const, \quad \text{for all faces } F \subset \partial T.$$

In accordance with (2.45), the constant values on the $d+1$ faces of T can be chosen to be the $d+1$ degrees of freedom for $\vec{v}$ on T. This will lead us to introduce the standard basis for $\mathcal{V}$ in Example 2.22 in Section 2.3.

We can define the global space $\mathcal{V}$ now to be the space of all functions $\vec{v} : \Omega \to \mathbb{R}^d$ which satisfy (2.54) for each $T \in \mathcal{T}$, and also

$$\begin{aligned} &\text{(i)} \quad \vec{v} \cdot \vec{\nu}_F|_F \quad \text{is continuous across each face } F \in \mathcal{F}_I \cup \mathcal{F}_D, \\ &\text{(ii)} \quad \vec{v} \cdot \vec{\nu}_F|_F = 0 \quad \text{for all} \quad F \in \mathcal{F}_N. \end{aligned} \tag{2.55}$$

Therefore

$$n_{\mathcal{V}} = \dim \mathcal{V} = \#\mathcal{F}_I + \#\mathcal{F}_D \tag{2.56}$$

in accordance with (2.51).

The functions $w \in \mathcal{W}$, on the other hand, can be uniquely defined by their constant value on each element $T \in \mathcal{T}$. Thus

$$n_{\mathcal{W}} = \dim \mathcal{W} = \#\mathcal{T} \tag{2.57}$$

in accordance with (2.53). □

2.2.3 Error estimates

To establish estimates for the approximation error we will now look at a shape regular family $\{\mathcal{T}_h\}$ of simplicial triangulations, parameterised by the maximum diameter h of the elements $T \in \mathcal{T}_h$.

Proposition 2.17. *Let $\vec{v} \in H(\mathrm{div}, \Omega) \cap (L_s(\Omega))^d$ for some $s > 2$. For each $k \in \mathbb{N} \cup \{0\}$ there exists a generic constant c independent of h such that*

$$\inf_{\vec{v}_h \in \mathcal{RT}_k(\Omega, \mathcal{T}_h)} \|\vec{v} - \vec{v}_h\|_{(L_2(\Omega))^d} \leq c h^m |\vec{v}|_{(H^m(\Omega))^d}, \tag{2.58}$$

$$\inf_{\vec{v}_h \in \mathcal{RT}_k(\Omega, \mathcal{T}_h)} \|\mathrm{div}\,(\vec{v} - \vec{v}_h)\|_{(L_2(\Omega))^d} \leq c h^m |\mathrm{div}\,\vec{v}|_{H^m(\Omega)} \tag{2.59}$$

for $1 \leq m \leq k+1$.

Proof. Let $\vec{v} \in H(\mathrm{div}, \Omega) \cap (L_s(\Omega))^d$. By $\mathbf{\Pi}_h$ we denote the projection operator from $H(\mathrm{div}, \Omega) \cap (L_s(\Omega))^d$ onto $\mathcal{RT}_k(\Omega, \mathcal{T}_h)$ such that on each $T \in \mathcal{T}_h$

$$(\mathbf{\Pi}_h \vec{v})|_T := \pi_T(\vec{v}|_T)$$

where π_T is defined as in (2.45). Now consider $\vec{v}_h := \mathbf{\Pi}_h \vec{v}$ as a candidate for the approximation to $\vec{v}$ in (2.58) and (2.59), and let c denote a generic constant independent of h. Then

$$\|\vec{v} - \vec{v}_h\|_{(L_2(\Omega))^d} = \left\{ \sum_{T \in \mathcal{T}_h} \|\vec{v} - \pi_T \vec{v}\|^2_{L_2(T)} \right\}^{1/2} \tag{2.60}$$

Using the shape regularity (2.35) it can be shown by the use of a reference element $\widehat{T}$

that

$$\|\vec{v} - \pi_T \vec{v}\|_{(L_2(T))^d} \leq ch(T)^m |\vec{v}|_{(H^m(T))^d} \tag{2.61}$$

We refer to Raviart and Thomas [77] and to Nédélec [73] for the proof of (2.61). Combining (2.60) and (2.61) we obtain (2.58).

Similarly we can write

$$\begin{aligned} \|\operatorname{div}(\vec{v} - \vec{v}_h)\|_{(L_2(\Omega))^d} &= \left\{ \sum_{T \in \mathcal{T}_h} \|\operatorname{div}(\vec{v} - \pi_T \vec{v})\|^2_{L_2(T)} \right\}^{1/2} \\ &= \left\{ \sum_{T \in \mathcal{T}_h} \|\operatorname{div} \vec{v} - \rho_T(\operatorname{div} \vec{v})\|^2_{L_2(T)} \right\}^{1/2} \end{aligned} \tag{2.62}$$

where in the last step we used Proposition 2.15. As for π_T above, it can be shown for the L_2-projection ρ_T that

$$\|w - \rho_T w\|_{L_2(T)} \leq ch(T)^m |w|_{H^m(T)} \tag{2.63}$$

for $w \in L_2(T)$ (see Ciarlet [27, Section 3.1]). Now, combining (2.62) and (2.63) we obtain (2.59). □

In the same way we can derive error estimates for $\mathcal{P}_k(\Omega, \mathcal{T}_h)$ and $L_2(\Omega)$:

Proposition 2.18. *Let $w \in L_2(\Omega)$. There exists a constant c independent of h such that*

$$\inf_{w_h \in \mathcal{P}_k(\Omega, \mathcal{T}_h)} \|w - w_h\|_{L_2(T)} \leq ch^m |w|_{H^m(\Omega)}. \tag{2.64}$$

Proof. Exactly as for (2.59), using (2.63). □

Combining the results of Theorem 2.9 and Propositions 2.17 and 2.18 we get the following error estimates for the solution of the approximate problem (2.29):

Theorem 2.19. *Let $k \geq 0$ and let $\mathcal{V}$ and $\mathcal{W}$ be the spaces defined in (2.50) and (2.52). Let $(\vec{u}, p) \in H_{0,N}(\operatorname{div}, \Omega) \times L_2(\Omega)$ be the solution of (2.15) and let $(\vec{U}, P) \in \mathcal{V} \times \mathcal{W}$ be the solution of (2.29). Then there exists a generic constant c independent of h such that*

$$\|\vec{u} - \vec{U}\|_{(L_2(\Omega))^d} \leq ch^m \left(|\vec{u}|_{(H^m(\Omega))^d} + |\operatorname{div} \vec{u}|_{H^m(\Omega)} \right), \tag{2.65}$$

$$\|p - P\|_{L_2(\Omega)} \leq ch^m \left(|\vec{u}|_{(H^m(\Omega))^d} + |\operatorname{div} \vec{u}|_{H^m(\Omega)} + |p|_{H^m(\Omega)} \right) \tag{2.66}$$

for $1 \leq m \leq k+1$. Moreover, we also have

$$\|\operatorname{div} \vec{u} - \operatorname{div} \vec{U}\|_{L_2(\Omega)} \leq ch^m |\operatorname{div} \vec{u}|_{H^m(\Omega)}. \tag{2.67}$$

Proof. The estimates (2.65) and (2.66) are an immediate consequence of Theorem 2.9, Proposition 2.17 and Proposition 2.18.

To show (2.67) we put $w = W$ in the second equation in problem (2.15) and subtract the second equation in (2.29). This yields

$$\int_\Omega (\operatorname{div}\vec{u} - \operatorname{div}\vec{U}) W \, d\vec{x} = 0, \qquad \text{for all } W \in \mathcal{W}.$$

In other words this means that $\operatorname{div}\vec{U}$ is the L_2-projection of $\operatorname{div}\vec{u}$ onto $\mathcal{W}$, i.e. $\rho_T \operatorname{div}\vec{u} = \operatorname{div}\vec{U}$ on each $T \in \mathcal{T}$. Therefore the estimate (2.67) is a consequence of (2.62) and (2.63) as in the proof to Proposition 2.17. □

Example 2.20. *The lowest order case: $k = 0$*

In the lowest order case we have

$$\begin{aligned} \|\vec{u} - \vec{U}\|_{(L_2(\Omega))^d} &\leq ch\left(|\vec{u}|_{(H^1(\Omega))^d} + |\operatorname{div}\vec{u}|_{H^1(\Omega)}\right) \\ \|p - P\|_{L_2(\Omega)} &\leq ch\left(|\vec{u}|_{(H^1(\Omega))^d} + |\operatorname{div}\vec{u}|_{H^1(\Omega)} + |p|_{H^1(\Omega)}\right). \end{aligned}$$

□

Remark 2.21. Following Falk and Osborn [40] and Douglas and Roberts [35], the estimates (2.65) and (2.66) can be improved for Ω convex, so that

$$\begin{aligned} \|\vec{u} - \vec{U}\|_{(L_2(\Omega))^d} &\leq ch^m \, \|\vec{u}\|_{(H^m(\Omega))^d}, \\ \|p - P\|_{L_2(\Omega)} &\leq ch^m \, \|p\|_{H^{m^*}(\Omega)}, \end{aligned}$$

for $1 \leq m \leq k+1$ and $m^* := \max\{2, m\}$.

2.3 The resulting saddle point system

In practice, problem (2.29) is implemented by choosing bases for $\mathcal{V}$ and $\mathcal{W}$ and writing it in matrix form suitable for numerical computations. In this section we will describe the derivation of the matrix form, analyse it, and present some standard iterative techniques for its solution.

2.3.1 Derivation of the matrix form

Consider first (2.29) as an abstract system again, and let $\{\vec{v}_i : i = 1, \ldots, n_\mathcal{V}\}$ and $\{w_j : j = 1, \ldots, n_\mathcal{W}\}$ be bases for $\mathcal{V}$ and $\mathcal{W}$. We define

$$M_{i,i'} := m(\vec{v}_i, \vec{v}_{i'}), \tag{2.68}$$

$$B_{i,j} := b(\vec{v}_i, w_j), \tag{2.69}$$

$$g_i := G(\vec{v}_i), \tag{2.70}$$

$$f_j := F(w_j) \tag{2.71}$$

and denote $M := [M_{i,i'}]_{n_\mathcal{V}\times n_\mathcal{V}}$, $B := [B_{i,j}]_{n_\mathcal{V}\times n_\mathcal{W}}$, $\mathbf{g} := [g_i]_{n_\mathcal{V}}$ and $\mathbf{f} := [f_j]_{n_\mathcal{W}}$. By writing

$$\vec{U} = \sum_{i=1}^{n_\mathcal{V}} u_i \vec{v}_i, \tag{2.72}$$

$$P = \sum_{j=1}^{n_\mathcal{W}} p_j w_j, \tag{2.73}$$

problem (2.29) is reduced to a system of linear equations,

$$\begin{pmatrix} M & B \\ B^T & 0 \end{pmatrix}\begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{g} \\ \mathbf{f} \end{pmatrix} \qquad \text{in} \quad \mathbb{R}^{n_\mathcal{V}} \times \mathbb{R}^{n_\mathcal{W}} \ , \tag{2.74}$$

which is to be solved for $\mathbf{u} := [u_i]_{n_\mathcal{V}}$ and $\mathbf{p} := [p_j]_{n_\mathcal{W}}$. In the following we will often refer to M as the *mass matrix*, to B as the *discrete gradient operator* and to B^T as the *discrete divergence operator*.

Example 2.22. *The lowest order case: $k = 0$*

As we have already shown in Example 2.16, the elements of $\mathcal{V}$ can be fully defined by the constant values of $\vec{v} \cdot \vec{\nu}_F$ on each face $F \in \mathcal{F}_I \cup \mathcal{F}_D$ in the case $k = 0$. Therefore the natural way to construct a basis for $\mathcal{V}$ is to associate with each face $F \in \mathcal{F}_I \cup \mathcal{F}_D$ a function $\vec{v}_F \in \mathcal{V}$ with the property that

$$\vec{v}_F \cdot \vec{\nu}_{F'}|_{F'} = \delta_{F,F'}, \qquad \text{for all } F' \in \mathcal{F} \tag{2.75}$$

with δ denoting the Kronecker delta.

The elements of $\mathcal{W}$ on the other hand, can be fully described by their constant value on each element $T \in \mathcal{T}$. A natural basis for $\mathcal{W}$ is given by the characteristic functions $w_T \in \mathcal{W}$ of each of the elements $T \in \mathcal{T}$, i.e. functions fulfilling the property that

$$w_T|_{T'} = \delta_{T,T'}, \qquad \text{for all } T' \in \mathcal{T}. \tag{2.76}$$

We can write

$$\vec{U} = \sum_{F \in \mathcal{F}_I \cup \mathcal{F}_D} u_F \vec{v}_F, \qquad P = \sum_{T \in \mathcal{T}} p_T w_T, \tag{2.77}$$

and identify the rows and columns of M with the indices $F, F' \in \mathcal{F}_I \cup \mathcal{F}_D$ and we have

$$M_{F,F'} = m(\vec{v}_F, \vec{v}_{F'}) = \int_\Omega D^{-1}(\vec{x}) \vec{v}_F \cdot \vec{v}_{F'} \, d\vec{x}. \tag{2.78}$$

In the same way, the rows of B correspond to indices $F \in \mathcal{F}_I \cup \mathcal{F}_D$, whereas the columns of B can be identified with the indices $T \in \mathcal{T}$. Using (2.76) and the Divergence Theorem on T we get

$$B_{F,T} = b(\vec{v}_F, w_T) = \int_T \operatorname{div} \vec{v}_F \, d\vec{x} = \int_{\partial T} \vec{v}_F \cdot \vec{\nu}_T \, ds \tag{2.79}$$

where $\vec{\nu}_T$ denotes the unit outward normal from T on ∂T as defined in Section 2.1.3. Recalling that on each face F of T we have $\vec{\nu}_F = \pm\vec{\nu}_T$ (see (2.37)), and using (2.75), (2.79) can be simplified to

$$B_{F,T} = \begin{cases} \int_F \vec{v}_F \cdot \vec{\nu}_F \, ds =: \quad |F|, & \text{for } F \subset \partial T \text{ and } \vec{\nu}_F = \vec{\nu}_T, \\ -\int_F \vec{v}_F \cdot \vec{\nu}_F \, ds =: \ -|F|, & \text{for } F \subset \partial T \text{ and } \vec{\nu}_F = -\vec{\nu}_T, \\ 0, & \text{for } F \not\subset \partial T. \end{cases} \tag{2.80}$$

Finally, we also identify the columns of the vectors $\mathbf{g}$ and $\mathbf{f}$ on the right hand side of (2.74) with indices $F \in \mathcal{F}_I \cup \mathcal{F}_D$ and $T \in \mathcal{T}$. Using (2.70) and (2.16) we obtain

$$g_F = G(\vec{v}_F) = \langle \vec{v}_F \cdot \vec{\nu}, g_D \rangle_\Gamma = \begin{cases} \int_F g_D \, ds & \text{for } F \in \mathcal{F}_D \text{ and } \vec{\nu}_F = \vec{\nu}, \\ -\int_F g_D \, ds & \text{for } F \in \mathcal{F}_D \text{ and } \vec{\nu}_F = -\vec{\nu}, \\ 0 & \text{for } F \in \mathcal{F}_I, \end{cases}$$

where $\vec{\nu}$ denotes the unit outward normal from Ω on Γ as defined at the beginning of Section 2.1, and using (2.71) and (2.17) we obtain

$$f_T = -\int_T f \, d\vec{x}.$$

Thus (after specifying an ordering of the faces $F \in \mathcal{F}_I \cup \mathcal{F}_D$ and of the elements $T \in \mathcal{T}$) we have again a system of linear equations (2.74) which needs to be solved for $\mathbf{u} := [u_F]_{F \in \mathcal{F}_I \cup \mathcal{F}_D}$ and $\mathbf{p} := [p_T]_{T \in \mathcal{T}}$.

In the following we will often use this way of indexing (2.74) by $F \in \mathcal{F}_I \cup \mathcal{F}_D$ and $T \in \mathcal{T}$, when $k = 0$. □

2.3.2 Analysis of the spectrum

Let us now analyse the matrix

$$\mathcal{M} := \begin{pmatrix} M & B \\ B^T & 0 \end{pmatrix}$$

in (2.74). Obviously the block M is symmetric. Moreover, M is positive definite because of (2.22). On the other hand, it follows from (2.32) that the block B has full rank. More precisely, if $\mathbf{p} \in \mathbb{R}^{n_{\mathcal{W}}}$ such that $B\mathbf{p} = \mathbf{0}$, it follows from (2.69) and (2.73) that $b(\vec{v}_i, P) = 0$ for all $i = 1, \dots, n_{\mathcal{V}}$. Since the vectors $\{\vec{v}_i\}$ form a basis of $\mathcal{V}$, the inf-sup condition (2.32) implies $P = 0$ and therefore $\mathbf{p} = \mathbf{0}$.

From the above conditions we have that $\mathcal{M}$ is symmetric and non-singular. However, $\mathcal{M}$ is not positive definite as can be seen by choosing any $\mathbf{w} \in \mathbb{R}^{n_{\mathcal{W}}} \backslash \{\mathbf{0}\}$ and setting $\mathbf{v} = \varepsilon B\mathbf{w}$. Then we have

$$(\mathbf{v}^T \ \mathbf{w}^T) \begin{pmatrix} M & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ \mathbf{w} \end{pmatrix} = \mathbf{v}^T M\mathbf{v} + 2\mathbf{v}^T B\mathbf{w} = \varepsilon \mathbf{w}^T C\mathbf{w}. \qquad (2.81)$$

with $C := \varepsilon B^T M B + 2B^T B$. Obviously the matrices $B^T M B$ and $B^T B$ are positive definite and thus the right hand side of (2.81) is positive if $\varepsilon > 0$. However, if we choose $\varepsilon < 0$ with $|\varepsilon|$ sufficiently small, the right hand side of (2.81) is negative, thus showing that $\mathcal{M}$ is indefinite. In fact, we have the following characterisation of the spectrum of $\mathcal{M}$:

Theorem 2.23. *Let $0 < \mu_{min} \leq \mu_{max}$ be the minimum and maximum eigenvalues of M, and let $0 < \sigma_{min} \leq \sigma_{max}$ be the minimum and maximum singular values of B. If we denote the spectrum of $\mathcal{M}$ by $\Lambda(\mathcal{M})$, then*

$$\Lambda(\mathcal{M}) \subset [\alpha^-_{min}, \alpha^-_{max}] \cup [\alpha^+_{min}, \alpha^+_{max}], \qquad (2.82)$$

where

$$\begin{aligned} \alpha^-_{min} &:= \tfrac{1}{2}\left(\mu_{min} - \sqrt{\mu^2_{min} + 4\sigma^2_{max}}\right) && < 0, \\ \alpha^-_{max} &:= \tfrac{1}{2}\left(\mu_{max} - \sqrt{\mu^2_{max} + 4\sigma^2_{min}}\right) && < 0, \\ \alpha^+_{min} &:= \mu_{min} && > 0, \\ \alpha^+_{max} &:= \tfrac{1}{2}\left(\mu_{max} + \sqrt{\mu^2_{max} + 4\sigma^2_{max}}\right) && > 0. \end{aligned}$$

Proof. See Rusten and Winther [82, Lemma 2.1]. □

Remark 2.24. More detail is given in Benbow [14] where in fact it is proved that the positive eigenvalues of $\mathcal{M}$ lie in two intervals:

$$\Lambda(\mathcal{M}) \subset [\alpha^-_{min}, \alpha^-_{max}] \cup [\mu_{min}, \mu_{max}] \cup [\alpha^+_{min}, \alpha^+_{max}],$$

where

$$\alpha^+_{min} := \tfrac{1}{2}\left(\mu_{min} + \sqrt{\mu^2_{min} + 4\sigma^2_{min}}\right) > \mu_{min} > 0.$$

The performance of iterative methods for the solution of linear equations systems depends on the condition number of the system matrix. The *spectral condition number* of a nonsingular symmetric matrix A is defined as

$$\kappa(A) := \frac{\max_{\lambda\in\Lambda(A)} |\lambda|}{\min_{\lambda\in\Lambda(A)} |\lambda|} \tag{2.83}$$

For our purposes it is interesting to bound the spectral condition number of $\mathcal{M}$ in terms of powers of h and

$$h_{min} := \min_{T\in\mathcal{T}_h} h(T).$$

We will now consider the case $k = 0$ in detail. However, the results extend in a straightforward way to $k > 0$, i.e. higher order elements.

Suppose for the rest of this section that $\{\mathcal{T}_h\}$ is a shape regular family of triangulations (see Definition 2.10), and let c and C be generic positive constants independent of h and h_{min}. Furthermore let $|\mathbf{x}| := \{\mathbf{x}^T\mathbf{x}\}^{1/2}$ denote the Euclidean norm of a vector $\mathbf{x} \in \mathbb{R}^n$. We have the following relationship between the L_2-norm of functions $P \in \mathcal{W}$ and $\vec{U} \in \mathcal{V}$ and the Euclidean norm of their coefficient vectors:

Lemma 2.25. *Let $k = 0$ and let $P \in \mathcal{W}$ and $\vec{U} \in \mathcal{V}$. Then*

$$ch^d_{min}|\mathbf{p}|^2 \;\leq\; \|P\|^2_{L_2(\Omega)} \;\leq Ch^d|\mathbf{p}|^2, \tag{2.84}$$

$$ch^d_{min}|\mathbf{u}|^2 \;\leq\; \|\vec{U}\|^2_{(L_2(\Omega))^d} \;\leq Ch^d|\mathbf{u}|^2, \tag{2.85}$$

where $\mathbf{u} := [u_F]_{F\in\mathcal{F}_I\cup\mathcal{F}_D}$ and $\mathbf{p} := [p_T]_{T\in\mathcal{T}_h}$ are the vectors of coefficients in the representation (2.77) of $\vec{U}$ and P.

Proof. The shape regularity condition (2.35) guarantees that the length $h_{min}(T)$ of the smallest edge of T satisfies $h_{min}(T) \geq 2\rho(T) \geq 2\kappa h(T)$ and therefore

$$ch(T)^d \leq |T| \leq Ch(T)^d \tag{2.86}$$

Furthermore, it follows from (2.77) and (2.76) that $\int_T P^2 dx = |T|\, p_T^2$, and combining this with (2.86), we obtain

$$ch(T)^d p_T^2 \leq \int_T P^2 dx \leq Ch(T)^d p_T^2.$$

The proof of (2.84) then follows by summation over $T \in \mathcal{T}_h$.

Similarly using the shape regularity (2.35), it can be shown by the use of a reference element $\widehat{T}$ and a Piola transformation that

$$ch(T)^d \sum_{F\subset\overline{T}} u_F^2 \leq \|\vec{U}\|^2_{(L_2(T))^3} \leq Ch(T)^d \sum_{F\subset\overline{T}} u_F^2, \tag{2.87}$$

and the proof of (2.85) follows again by summation. We refer to Raviart & Thomas [77] for the proof of (2.87) in 2D and to Corollary A.3 in the Appendix for 3D. $\square$

We can now bound the eigenvalues of M and the singular values of B:

Proposition 2.26. *Let $k = 0$. Then*

$$ch^d_{min} \;\leq\; \mu_{min} \;\;\leq\;\; \mu_{max} \;\leq\; Ch^d, \tag{2.88}$$

$$ch^d_{min} \;\leq\; \sigma_{min} \;\;\leq\;\; \sigma_{max} \;\leq\; Ch^{d-1}. \tag{2.89}$$

Proof. Let $\mathbf{u} = [u_F]_{F\in\mathcal{F}_I\cup\mathcal{F}_D}$ be an arbitrary element of $\mathbb{R}^{n_\mathcal{V}}$ and let $\vec{U}$ be the corresponding element of $\mathcal{V}$ defined in (2.77). Using the definition (2.78) of M together with (2.2),

$$\Theta^{-1}\|\vec{U}\|_{(L_2(\Omega))^d} \leq \mathbf{u}^T M\mathbf{u} = m(\vec{U},\vec{U}) \leq \theta^{-1}\|\vec{U}\|_{(L_2(\Omega))^d}.$$

Combining this with (2.85) we have

$$ch^d_{min}|\mathbf{u}|^2 \leq \mathbf{u}^T M\mathbf{u} \leq Ch^d|\mathbf{u}|^2$$

which establishes (2.88).

Now, let $\mathbf{p} = [p_T]_{T\in\mathcal{T}_h}$ be an arbitrary element of $\mathbb{R}^{n_\mathcal{W}}$. Using the notation from Example 2.22 and in particular (2.80) we have

$$|B\mathbf{p}|^2 = \sum_{T,T'\in\mathcal{T}_h} p_T\Bigg(\sum_{F\in\mathcal{F}_I\cup\mathcal{F}_D} B_{F,T}B_{F,T'}\Bigg)p_{T'} = \sum_{F\in\mathcal{F}_I\cup\mathcal{F}_D}|F|^2\sum_{\substack{T,T'\in\mathcal{T}_h \text{ s.t.}\\ F\subset\overline{T}\cap\overline{T}'}} p_Tp_{T'}. \tag{2.90}$$

However, each face $F \in \mathcal{F}_I \cup \mathcal{F}_D$ is contained in at most two elements $T, T' \in \mathcal{T}_h$ and we can eliminate the "cross terms" $p_Tp_{T'}$ in (2.90) by the elementary inequality $p_T^2 + p_Tp_{T'} + p_{T'}^2 \leq 2(p_T^2 + p_{T'}^2)$. Finally, using this in (2.90) together with $|F|^2 \leq Ch^{2(d-1)}$, we have

$$|B\mathbf{p}|^2 \;\leq\; Ch^{2(d-1)}|\mathbf{p}|^2$$

which establishes the upper bound in (2.89). For the lower bound we will need the inf-sup condition (2.32). Let P be the element of $\mathcal{W}$ defined by (2.77). Since $\mathcal{V}$ is finite dimensional, there exists a $\vec{0} \neq \vec{V}_P \in \mathcal{V}$ such that

$$\sup_{\vec{V}\in\mathcal{V}}\frac{b(\vec{V},P)}{\|\vec{V}\|_{H(\mathrm{div},\Omega)}} = \frac{b(\vec{V}_P,P)}{\|\vec{V}_P\|_{H(\mathrm{div},\Omega)}}.$$

Let $\mathbf{v}_P$ be the corresponding vector of coefficients in $\mathbb{R}^{n_\mathcal{W}}$. Then using the definition (2.79) of B and (2.32) we get

$$|\mathbf{v}_P^TB\mathbf{p}| = |b(\vec{V}_P,P)| \geq \beta_0\|\vec{V}_P\|_{H(\mathrm{div},\Omega)}\|P\|_{L_2(\Omega)} \geq \beta_0\|\vec{V}_P\|_{L_2(\Omega)}\|P\|_{L_2(\Omega)}.$$

Thus using Lemma 2.25,

$$|\mathbf{v}_P^TB\mathbf{p}| \geq ch^d_{min}|\mathbf{v}_P||\mathbf{p}|.$$

Using the Cauchy-Schwarz inequality and dividing through by $|\mathbf{v}_P| \neq 0$ we finally have

$$|B\mathbf{p}| \geq ch_{min}^d |\mathbf{p}|$$

which establishes the lower bound in (2.89). $\square$

Combining the results of Theorem 2.23 and Proposition 2.26 we can now bound the spectral condition number of $\mathcal{M}$ in the lowest order case $k = 0$.

Theorem 2.27. *Let $k = 0$. Then*

$$\kappa(\mathcal{M}) \leq C \left(\tfrac{h}{h_{min}}\right)^{2d} h^{-1}. \tag{2.91}$$

Proof. Using the bounds for the negative and positive eigenvalues of $\mathcal{M}$ given in Theorem 2.23 we can first of all write

$$\kappa(\mathcal{M}) \leq \frac{\max(-\alpha_{min}^-, \alpha_{max}^+)}{\min(-\alpha_{max}^-, \alpha_{min}^+)}. \tag{2.92}$$

However, by (2.88) and (2.89) we have

$$\begin{aligned} -\alpha_{min}^- &= \tfrac{1}{2}\left(\sqrt{\mu_{min}^2 + 4\sigma_{max}^2} - \mu_{min}\right) \;\leq\; \sigma_{max} \;\leq\; Ch^{d-1} \qquad \text{and} \\ \alpha_{max}^+ &= \tfrac{1}{2}\left(\sqrt{\mu_{max}^2 + 4\sigma_{max}^2} + \mu_{max}\right) \;\leq\; \mu_{max} + \sigma_{max} \;\leq\; Ch^{d-1} \end{aligned}$$

and therefore

$$\max(-\alpha_{min}^-, \alpha_{max}^+) \;\leq\; Ch^{d-1}. \tag{2.93}$$

To bound the denominator in (2.92), we first need to establish a lower bound for $-\alpha_{max}^-$. Let us distinguish two cases. If $\mu_{max} \leq 2\sigma_{min}$ then

$$-\alpha_{max}^- = \frac{1}{2}\left(\sqrt{\mu_{max}^2 + 4\sigma_{min}^2} - \mu_{max}\right) \geq \tfrac{\sqrt{2}-1}{2}\mu_{max} \;\geq\; ch_{min}^d.$$

If $\mu_{max} \geq 2\sigma_{min}$, on the other hand, then

$$-\alpha_{max}^- = \frac{1}{2}\left(\sqrt{\mu_{max}^2 + 4\sigma_{min}^2} - \mu_{max}\right) = \sigma_{min}\tan(\vartheta/2)$$

where $\tan\vartheta := 2\sigma_{min}/\mu_{max}$. Now, using the fact that $\tan(\vartheta/2) \geq \frac{\pi}{8}\tan\vartheta$ for $0 \leq \tan\vartheta \leq 1$ it follows that

$$-\alpha_{max}^- = \sigma_{min}\tan(\vartheta/2) \geq \frac{\pi}{4}\frac{\sigma_{min}^2}{\mu_{max}} \;\geq\; c\left(\tfrac{h_{min}^2}{h}\right)^d.$$

Since $\alpha_{min}^+ = \mu_{min} \geq ch_{min}^d$ and since $h \geq h_{min}$, we get a lower bound for the denominator in (2.92):

$$\min(-\alpha_{max}^-, \alpha_{min}^+) \;\geq\; c\left(\tfrac{h_{min}^2}{h}\right)^d. \tag{2.94}$$

Finally combining (2.94) and (2.93) with (2.92) establishes (2.91). □

This is almost certainly a very pessimistic upper bound for the spectral condition number of $\mathcal{M}$. However, for a quasi-uniform family of triangulations $\{\mathcal{T}_h\}$ the estimate (2.91) can be improved.

Corollary 2.28. *Let $\{\mathcal{T}_h\}$ be a quasi-uniform family of triangulations. Then*

$$\kappa(\mathcal{M}) \leq Ch^{-1}.$$

Proof. Condition (2.36) guarantees that $h(T) \geq ch$ for each $T \in \mathcal{T}_h$. Therefore $h_{min} \geq ch$ and so the proof follows directly from Theorem 2.27. □

2.3.3 Iterative solution

It is a well known fact that numerical methods for positive definite systems are more efficient and more stable than those for indefinite systems. Classical multigrid, domain decomposition or preconditioned conjugate gradients, all rely for their most powerful theoretical results on the positivity of the spectrum. Therefore, almost all approaches to solve system (2.74) efficiently, contain at some point a reduction of the system to a positive definite system. Some of the most important methods are presented below.

Preconditioned MINRES

The standard approach to solve (2.74), is to use an appropriate Krylov subspace method (see Saad [84] for a discussion of Krylov subspace methods). The most famous member of this family of iterative methods is the conjugate gradient method (CG), but its convergence is only guaranteed for positive definite matrices. It is often claimed that the most efficient method for symmetric, but indefinite systems is the MINRES algorithm by Paige & Saunders [75]. It is a stabilised version of the conjugate residual method (CR), and a special case of the GMRES method for general non-symmetric systems.

Like all Krylov subspace methods it is not robust to mesh refinement or strong discontinuities in $D(\vec{x})$, and it is necessary to precondition (2.74) before applying MINRES. The left-preconditioned version of MINRES is given in Figure 2.2. Instigated by a paper by Rusten & Winther [82], several symmetric positive definite block preconditioners have been developed for $\mathcal{M}$ in recent years – see for example [9, 10, 80, 81, 82, 83, 90]. They are all restricted to 2D, i.e. $d = 2$, and can generally be put into two classes: preconditioners of the form

$$\mathcal{P}_{Schur}^{-1} := \begin{pmatrix} I & 0 \\ 0 & P_{Schur}^{-1} \end{pmatrix}$$

input matrix A, right hand side $\mathbf{b}$, initial guess $\mathbf{u}_0$, preconditioner P_L, tolerance ε

% Initialise

$\tilde{\mathbf{v}}_0 = \mathbf{v}_0 = \tilde{\mathbf{p}}_0 = \tilde{\mathbf{p}}_{-1} = \mathbf{0}, \quad c_0 = c_{-1} = 1, \quad s_0 = s_{-1} = 0$

$\mathbf{r}_0 = \mathbf{b} - A\,\mathbf{u}_0$

$\tilde{\mathbf{r}}_0 = P_L^{-1}\mathbf{r}_0$

$\eta = \tilde{\beta}_0 = \sqrt{\mathbf{r}_0^T\tilde{\mathbf{r}}_0}$

$\tilde{d}_0 = \|\tilde{\mathbf{r}}_0\|$

for $i = 1, 2, \ldots$ **until** $\tilde{d}_{i-1}/\tilde{d}_0 < \varepsilon$ **do**

% Lanczos

$\mathbf{v}_i = \mathbf{r}_{i-1}/\tilde{\beta}_{i-1}$

$\tilde{\mathbf{v}}_i = \tilde{\mathbf{r}}_{i-1}/\tilde{\beta}_{i-1}$

$\alpha = \tilde{\mathbf{v}}_i^T A\tilde{\mathbf{v}}_i$

$\mathbf{r}_i = A\tilde{\mathbf{v}}_i - \alpha\mathbf{v}_i - \tilde{\beta}_{i-1}\mathbf{v}_{i-1}$

$\tilde{\mathbf{r}}_i = P_L^{-1}A\tilde{\mathbf{v}}_i - \alpha\tilde{\mathbf{v}}_i - \tilde{\beta}_{i-1}\tilde{\mathbf{v}}_{i-1}$

$\tilde{\beta}_i = \sqrt{\mathbf{r}_i^T\tilde{\mathbf{r}}_i}$

% QR factorisation

$\rho_0 = c_{i-1}\alpha - c_{i-2}s_{i-1}\tilde{\beta}_{i-1}$

$\rho_1 = \sqrt{\rho_0^2 + \tilde{\beta}_i^2}$

$\rho_2 = s_{i-1}\alpha + c_{i-2}c_{i-1}\tilde{\beta}_{i-1}$

$\rho_3 = s_{i-2}\tilde{\beta}_{i-1}$

% Givens rotation

$c_i = \rho_0/\rho_1$

$s_i = \tilde{\beta}_i/\rho_1$

% Update

$\tilde{\mathbf{p}}_i = (\tilde{\mathbf{v}}_i - \rho_2\tilde{\mathbf{p}}_{i-1} - \rho_3\tilde{\mathbf{p}}_{i-2})/\rho_1$

$\mathbf{u}_i = \mathbf{u}_{i-1} + \eta c_i\tilde{\mathbf{p}}_i$

$\eta = -s_i\eta$

$\tilde{d}_i = |s_i|\tilde{d}_{i-1}$

end

Figure 2.2: Left-preconditioned version of MINRES (Paige & Saunders [75])

(see [82, 83, 80]) and, more recently, preconditioners of the form

$$\mathcal{P}_{H(\mathrm{div})}^{-1} := \begin{pmatrix} P_{H(\mathrm{div})}^{-1} & 0 \\ 0 & \delta I \end{pmatrix}$$

(see [90, 9, 10, 81]), where in the first case P_{Schur}^{-1} is a preconditioner for the Schur complement $-B^T M^{-1} B$ (see also below), and in the second case $P_{H(\mathrm{div})}^{-1}$ is a preconditioner for the matrix $A_{H(\mathrm{div})}$ corresponding to the $H(\mathrm{div}, \Omega)$ inner product in $\mathcal{RT}_k(\Omega, \mathcal{T})$. This preconditioner $P_{H(\mathrm{div})}^{-1}$ has very recently also been extended to 3D, in a paper by Wohlmuth et al. [93], where they construct a substructuring preconditioner for $A_{H(\mathrm{div})}$.

See also Silvester & Wathen [92, 86] for some related work on preconditioned MINRES for the Stokes problem. These papers are particularly interesting because of the careful discussion of the dependency of the convergence on the spectrum of the system.

Block elimination

Since the matrix M is regular, a very natural idea is to eliminate the velocity unknown $\mathbf{u}$ from (2.74) thus arriving at the positive definite Schur-complement system

$$-B^T M^{-1} B \mathbf{p} = \mathbf{f} - B^T M^{-1} \mathbf{g}. \tag{2.95}$$

However, this requires M^{-1}. In our case M is the mass matrix in $\mathcal{RT}_k(\Omega, \mathcal{T})$. By implementing the underlying finite element method using a special quadrature rule (at least for $k = 0$) M is replaced by a diagonal matrix (*mass lumping*), making the application of M^{-1} extremely simple. (It is important to note though that this will change the approximation properties of the solution.)

Nevertheless, we still have to solve problem (2.95), and it is shown in Hiptmair [57, Remark 3.6] that the condition number $\kappa(-B^T M^{-1} B) = O(h^{-2})$, thus making it necessary to precondition (2.95). If we look back to the continuous problem (2.15) underlying (2.74), we observe that (for sufficiently smooth data) by eliminating the velocity unknown we obtain the primal formulation (2.3). Thus, $-B^T M^{-1} B$ corresponds to the bilinear form $a_{\mathcal{P}}(\cdot, \cdot)$ in (2.3), and one might think that it could be tackled by the usual preconditioning strategies (domain decomposition, multilevel) which have been successfully employed to the matrices arising from $a_{\mathcal{P}}(\cdot, \cdot)$. Unfortunately, the finite element space $\mathcal{W} \subset L_2(\Omega)$ associated with the pressure unknown $\mathbf{p}$ here, lacks the kind of regularity required for these ideas to work in a straightforward manner. More sophisticated techniques are necessary. In Baranger et al. [13] it is shown that for a particular choice of quadrature rule in M, the matrix $-B^T M^{-1} B$ corresponds to a cell-centred finite volume discretisation of (2.1), and in Pavarino & Ramé [76] this fact is used to devise an overlapping additive Schwarz preconditioner for (2.95).

A way to avoid the mass lumping, is to use a combined inner and outer iteration. An example of this is the following method.

Augmented Lagrangian Method

The *augmented Lagrangian method* is a combination of the penalty method with an Uzawa-type algorithm (see Fortin & Glowinski [41] or Hiptmair et al. [60] for details).

Uzawa-type methods essentially arise from applying classical iterative methods for positive definite systems, like Richardson's iteration or conjugate gradients (CG), to the Schur-complement system (2.95). The original form applying a Richardson iteration to (2.95), i.e. given $\mathbf{p}^{(n)}$ and $\rho \in \mathbb{R}$ find

$$\mathbf{p}^{(n+1)} = \mathbf{p}^{(n)} + \rho\,(-B^T M^{-1} B\mathbf{p}^{(n)} - \mathbf{f} + B^T M^{-1}\mathbf{g}),$$

is presented in Algorithm 2.29 below.

Algorithm 2.29 (Uzawa's Algorithm).

- *Let* $\mathbf{p}^{(0)}$ *be chosen arbitrarily.*
- *For* $n = 0, 1, \dots,$
  1. *find* $\mathbf{u}^{(n+1)}$ *such that* $M\mathbf{u}^{(n+1)} = \mathbf{g} - B\mathbf{p}^{(n)}$,
  2. *find* $\mathbf{p}^{(n+1)}$ *such that* $\mathbf{p}^{(n+1)} = \mathbf{p}^{(n)} + \rho(B^T\mathbf{u}^{(n+1)} - \mathbf{f})$.
- *End loop over* n.

In each step we now have to solve a system with matrix M, which is done approximately in an inner iteration. Since in our case M is well-conditioned (cf. Proposition 2.26), this can be done very cheaply. However, this advantage is traded in for an ill-conditioned system (2.95) faced by the outer iteration (recall $\kappa(-B^T M^{-1} B) = O(h^{-2})$). Even a more efficient iterative method like CG would not cure this problem. Preconditioning is necessary, and this leads to the same kind of problems as discussed above.

A way to overcome this problem, is to add a penalty term to the matrix in system (2.74) resulting in the equivalent system

$$\begin{pmatrix} M_\varepsilon & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{g} + \frac{1}{\varepsilon} B S^{-1}\mathbf{f} \\ \mathbf{f} \end{pmatrix} \qquad \text{with } M_\varepsilon := M + \frac{1}{\varepsilon} B S^{-1} B^T, \quad (2.96)$$

where S is an arbitrary positive definite matrix, and the real parameter $\varepsilon > 0$ governing the strength of the penalisation is called the *augmented Lagrangian parameter*. Now, if we apply Uzawa's Algorithm to (2.96), we obtain the augmented Lagrangian method.

In [60, Theorem 3.2] it is shown that for a particular choice of s, $\kappa(-B^T M_\varepsilon^{-1} B) \to 1$ for $\varepsilon \to 0$, so the augmented Lagrangian method will converge significantly faster for small values of ε. But again this advantage does not come for free: It is also shown in [60, Theorem 3.1] that for $\varepsilon \to 0$, we have $\kappa(M_\varepsilon) = O(\varepsilon^{-1} h^{-2})$, now leading to an ill-conditioned system faced by the inner iteration. For these reasons, the augmented

Lagrangian approach does not pay off, unless an efficient preconditioner for M_ε is available that is robust with respect to ε and h. For 2D, an optimal multilevel preconditioner for M_ε that is independent of the choice of ε and h, is presented in Hiptmair et al. [60], but in general there are no rigorous rules on the choice of ε.

Hybridisation of the velocity space $\mathcal{V}$

Another way of reducing (2.74) to a symmetric positive definite system is to relax the continuity condition on the normal components in $\mathcal{V}$, thus resulting in a nonconforming discretisation of $H_{0,N}(\mathrm{div},\Omega)$ – see for example [6, 8, 18, 25, 26, 30].

To illustrate the idea, let $\Gamma_N = \emptyset$ and $g_D = 0$ in (2.14). We define

$$\mathcal{RT}_k^{-1}(\Omega,\mathcal{T}) := \{\vec{v} \in (L_2(\Omega))^d : \vec{v}|_T \in RT_k(T) \ \text{ for all } T \in \mathcal{T}\},$$

i.e the functions in $\mathcal{RT}_k^{-1}(\Omega,\mathcal{T})$ are Raviart-Thomas-Nédélec functions of degree k on each element, but the interelement continuity in $\mathcal{RT}_k(\Omega,\mathcal{T})$ has been dropped. We also define

$$\mathcal{P}_k(\Omega,\mathcal{F}) := \Big\{\mu \in L_2\Big(\bigcup_{F\in\mathcal{F}} \overline{F}\Big) : \mu|_F \in P_k(F) \ \text{ for } F \in \mathcal{F}_I, \quad \mu|_F = 0 \ \text{ for } F \in \mathcal{F}_D\Big\},$$

and consider the problem of seeking $(\vec{U},P,\lambda) \in \mathcal{RT}_k^{-1}(\Omega,\mathcal{T}) \times \mathcal{P}_k(\Omega,\mathcal{T}) \times \mathcal{P}_k(\Omega,\mathcal{F})$ such that

$$\left.\begin{array}{lcll} m(\vec{U},\vec{V}) + b(\vec{V},P) + c(\vec{V},\lambda) & = & 0, & \text{for all } \vec{V} \in \mathcal{RT}_k^{-1}(\Omega,\mathcal{T}), \\ b(\vec{U},W) & = & F(W)\ , & \text{for all } W \in \mathcal{P}_k(\Omega,\mathcal{T}), \\ c(\vec{U},\mu) & = & 0\ , & \text{for all } \mu \in \mathcal{P}_k(\Omega,\mathcal{F}). \end{array}\right\} \qquad (2.97)$$

with

$$c(\vec{U},\mu) := \sum_{T\in\mathcal{T}} \int_{\partial T} \mu(\vec{U}\cdot\vec{\nu}_T)\,ds.$$

The function λ is usually called a *Lagrange multiplier*. Thus, the required continuity of the normal component of the velocity $\vec{U} \in \mathcal{RT}_k^{-1}(\Omega,\mathcal{T})$ across inter-element boundaries is enforced through an extra equation involving Lagrange multipliers $\mu \in \mathcal{P}_k(\Omega,\mathcal{F})$. In Arnold & Brezzi [8] and Arbogast & Chen [6] it is shown that this problem (2.97) is equivalent to (2.29).

Without going into any details about how one would choose bases for $\mathcal{RT}_k^{-1}(\Omega,\mathcal{T})$ and $\mathcal{P}_k(\Omega,\mathcal{F})$, this problem can be written as a linear equation system

$$\begin{pmatrix} \tilde{M} & \tilde{B} & C \\ \tilde{B}^T & 0 & 0 \\ C^T & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{p} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{f} \\ \mathbf{0} \end{pmatrix}. \qquad (2.98)$$

Note, that since we relaxed the continuity in the velocity space $\mathcal{RT}_k^{-1}(\Omega,\mathcal{T})$, the first equation in (2.97) holds elementwise, and therefore the matrix $\tilde{M}$ in (2.98) is block-diagonal (with the blocks corresponding to the elements $T \in \mathcal{T}$) and can be inverted easily. Now, block elimination of the velocity unknown $\mathbf{u}$ results in the symmetric positive definite system

$$-\begin{pmatrix} \tilde{B}^T\tilde{M}^{-1}\tilde{B} & \tilde{B}^T\tilde{M}^{-1}C \\ C^T\tilde{M}^{-1}\tilde{B} & C^T\tilde{M}^{-1}C \end{pmatrix}\begin{pmatrix} \mathbf{p} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{0} \end{pmatrix}. \tag{2.99}$$

which is approximately of the same size as the original saddle point system (2.74) and which can be solved by multigrid (see Brenner [18]) or domain decomposition techniques (see Cowsar et al. [30]). Furthermore, it is even possible to eliminate the pressure unknown $\mathbf{p}$ from (2.99) leading to a symmetric positive definite system in the Lagrange multipliers only (see Chen et al. [25, 26]).

Direct elimination of the incompressibility constraint

This final approach to efficiently solve (2.74) is the one we are going to use in this thesis. It differs from the last three approaches in that it aims to eliminate the pressure unknown, rather than the velocity unknown. Furthermore, it does not relax the local mass conservation of the mixed finite element discretisation, as done in the mass lumping or in the Lagrange multiplier case. On the contrary, the incompressibility constraint is enforced a priori on each element by using divergence-free Raviart-Thomas-Nédélec elements. The result is a much smaller, symmetric positive definite system for the velocity unknown $\mathbf{u}$. We will describe this approach at great length in Chapter 4.

This idea has been first proposed for 2D by Chavent et al. [24], and appears at least in the background of papers by Ewing & Wang [38, 39] and Mathew [71, 72], where they develop powerful multilevel methods for (2.74). The implementation of this idea requires the construction of a basis for the space $\mathring{\mathcal{V}}$ of divergence-free Raviart-Thomas-Nédélec elements (cf. (2.30)). This is done in [38, 39, 71, 72] for simply connected domains and pure Neumann boundary conditions in 2D. Since the construction of a basis for 3D is more difficult, the literature is restricted to a paper by Cai et al. [21] for uniform rectangular grids, and a paper by Hiptmair et al. [59], where they avoid the construction of a basis for $\mathring{\mathcal{V}}$ at the expense of a semidefinite velocity system (see Remark 4.6). However, we will show in Chapter 3 that a basis for $\mathring{\mathcal{V}}$ can be constructed even for general mixed boundary conditions on simply connected domains in 2D and 3D, and for multiply connected domains in 2D, making the approach much more interesting from a practical point of view than previously thought.

The pressure unknown $\mathbf{p}$ in (2.74) can be recovered by solving an additional triangular system (see Section 4.1.3 for an extensive discussion). Although this point is mentioned in Chavent et al. [24], it has not been rigorously investigated yet. It involves the construction of a basis for the complement of $\mathring{\mathcal{V}}$ in $\mathcal{V}$, which we will give in Section 3.3.

2.4 Summary

The main point of this chapter was to define the type of problem we are going to consider in this thesis, and to describe the mathematical setting for it. We considered the mixed boundary value problem (2.1) for second-order elliptic problems of Poisson-type over two/three-dimensional domains with polygonal/polyhedral boundaries, and their mixed variational formulation (2.15). We established existence and uniqueness of solutions of (2.15) in $H_{0,N}(\mathrm{div}, \Omega) \times L_2(\Omega)$, and discussed the interesting property that the normal traces of functions in $H_{0,N}(\mathrm{div}, \Omega)$ are continuous across any surface in Ω. This is useful for the approximation of $H_{0,N}(\mathrm{div}, \Omega)$ by conforming finite elements, which we then defined in Section 2.2 for simplicial triangulations of Ω. The elements that we introduced are called the Raviart-Thomas-Nédélec elements (often just called Raviart-Thomas elements in 2D). We showed that they are conforming in $H_{0,N}(\mathrm{div}, \Omega)$, and in the lowest order case they can be fully described by the value of their normal trace on each face of the triangulation, accounting for the widely used term *face elements*. We also defined an appropriate finite element space for $L_2(\Omega)$ of discontinuous piecewise polynomial functions, to ensure that the discrete inf-sup condition (2.32) is satisfied and that the discrete problem (2.29) has a unique solution (cf. Theorem 2.9). The usual error estimates hold (cf. Theorem 2.19).

The implementation of the finite element method required the introduction of bases for these finite element spaces, resulting in large, sparse systems of linear equations of saddle-point form (2.74). In the remainder of the chapter we analysed the spectrum of (2.74) and the most common approaches to solve this system. The system is indefinite, making a direct application of all the powerful classical methods for positive definite systems difficult. Therefore, almost all the approaches to solve the system efficiently contain at some point a reduction of the system to a positive definite system. We will see in Chapter 4 how a direct elimination of the pressure unknown results in a decoupling of (2.74) into such a positive definite system for the velocity unknown and a triangular system for the pressure, and how this can be exploited to solve the system efficiently.

The chapter is meant to set the ground for the next chapters and does not contain any new results apart (as far as we are aware) from Theorem 2.27, where we used the results in Rusten & Winther [82] to bound the spectral condition number $\kappa(\mathcal{M})$ of the matrix $\mathcal{M}$ in (2.74) in terms of the minimum and maximum mesh diameters h_{min} and h, assuming only shape regularity of the mesh. In the special case of a quasi-uniform mesh, we obtained the expected h^{-1} dependency of $\kappa(\mathcal{M})$ (cf. Corollary 2.28) which is observed in our practical applications.

Chapter 3

Divergence-free Elements

In this chapter we will investigate the subspace

$$\mathring{\mathcal{V}} := \{\vec{V} \in \mathcal{V} : b(\vec{V}, W) = 0 \text{ for all } W \in \mathcal{W}\} \tag{3.1}$$

of divergence-free Raviart-Thomas-Nédélec elements as defined in (2.30), where

$$\mathcal{V} := \{\vec{v} \in \mathcal{RT}_k(\Omega, \mathcal{T}) : \vec{v} \cdot \vec{\nu}|_{\Gamma_N} = 0\} \quad \text{and} \quad \mathcal{W} := \mathcal{P}_k(\Omega, \mathcal{T}). \tag{3.2}$$

Moreover, we will also devise a *complementary space* $\mathcal{V}^c$ with the property that

$$\mathcal{V} = \mathring{\mathcal{V}} + \mathcal{V}^c \qquad \text{and} \qquad \mathring{\mathcal{V}} \cap \mathcal{V}^c = \{\vec{0}\}, \tag{3.3}$$

where the sum of two spaces is defined in the usual way, i.e. $\mathring{\mathcal{V}} + \mathcal{V}^c := \{\vec{v}^{\,\circ} + \vec{v}^{\,c} : \vec{v}^{\,\circ} \in \mathring{\mathcal{V}},\ \vec{v}^{\,c} \in \mathcal{V}^c\}$. Note that this space is obviously not unique.

In particular, we are interested in this chapter in finding bases for the spaces $\mathring{\mathcal{V}}$ and $\mathcal{V}^c$. This is motivated by the fact that we will use $\mathring{\mathcal{V}}$ and $\mathcal{V}^c$ in Chapter 4 to decouple the discrete problem (2.29) in the same way as we used the spaces $\mathcal{Z}$, $\mathcal{Z}^\perp \subset H_{0,N}(\mathrm{div}, \Omega)$ in Section 2.1.2 to decouple the continuous problem (2.18). The bases for $\mathring{\mathcal{V}}$ and $\mathcal{V}^c$ are then needed to implement the resulting decoupled problems as linear equation systems, leading to a very efficient iterative method for (2.74).

Let us first determine the dimension of $\mathring{\mathcal{V}}$. Since problem (2.29) has a unique solution, we know that the condition

$$b(\vec{V}, W) = 0, \qquad \text{for all } W \in \mathcal{W},$$

imposes exactly $n_{\mathcal{W}} = \dim \mathcal{W}$ independent constraints on the function $\vec{V} \in \mathcal{V}$. The spaces $\mathcal{V}$ and $\mathcal{W}$ are both finite dimensional and therefore

$$\mathring{n} := \dim \mathring{\mathcal{V}} = \dim \mathcal{V} - \dim \mathcal{W} = n_{\mathcal{V}} - n_{\mathcal{W}}. \tag{3.4}$$

Divergence-free finite elements have been first developed in the related but different context of the classical Stokes problem describing slow viscous incompressible flow, and there is a large literature on it - see, for example, [32, 37, 43, 47, 48, 54, 55, 56, 67, 74, 87, 88, 94, 95]. The mixed formulation of the Stokes problem takes a similar form as (2.18), but with the fundamental difference that here

$$m(\vec{u}, \vec{v}) := \nu \sum_{i=1}^{d} \int_{\Omega} \vec{\nabla} u_i \cdot \vec{\nabla} v_i \, d\vec{x} \ ,$$

which is corresponding to a second-order differential operator. Therefore the velocity $\vec{u}$ is sought in the space $(H^1(\Omega))^d$, with ν denoting the kinematic viscosity of the fluid.

The first papers on a divergence-free finite element basis for the Stokes problem, by Crouzeix [32], Thomasset [87, 88] and Hecht [54, 55], concentrated on the non-conforming triangular P1-P0 element. Hecht even extended the construction to 3D. However, since it is not possible to approximate the divergence-free subspace of $(H^1(\Omega))^d$ with conforming finite elements of degree one, the rest of the work concentrates on higher order elements in 2D (e.g. Griffiths [47], Gustafson & Hartman [48], Ye et al. [95, 94], Mack [67]). An interesting recent paper by Ainsworth & Sherwin [3] returns to this approach to construct what they call a *natural preconditioner* for p and hp finite element approximations of the Stokes problem.

Divergence-free Raviart-Thomas-Nédélec elements for $H(\mathrm{div}, \Omega)$ were first discussed in relation to the vector-potential vorticity formulation of the Stokes problem by Nédélec [74], but he does not construct a basis for $\mathring{\mathcal{V}}$. The first construction of a basis is given in the context of two-dimensional groundwater flow problems of the form (2.18) by Chavent et al. [24], but it contains no proof. This is given much later in the development of preconditioning strategies for the saddle point system (2.74) [38, 39, 71, 72, 60].

The literature in 3D is much sparser. In an unpublished manuscript [56] that deals again with the solution of the Stokes problem, Hecht extends his results on the P1-P0 element [54, 55] to a wider family of finite elements in $(H^1(\Omega))^d$ including the (non-conforming) Raviart-Thomas-Nédélec elements, and constructs a basis for $\mathring{\mathcal{V}}$. However, the published literature is restricted to the pure Neumann case (i.e. $\Gamma_D = \emptyset$): in a paper by Dubois [37], where he uses it to solve model incompressible flow problems with prescribed vorticity; and more recently in a paper by Cai et al. [21] on preconditioning strategies for the saddle point system (2.74), although in that paper the method is developed only for the case of uniform rectangular grids.

In this chapter we give for the first time an explicit basis for $\mathring{\mathcal{V}}$ and $\mathcal{V}^c$ in the case of general mixed boundary conditions for the lowest order case $k = 0$ in 2D and 3D and an extension to higher order elements in 2D. We also discuss the higher order case in 3D. Most of the results in this chapter can also be found in the joint papers Cliffe et al. [28, 29] for the 2D case, and in Scheichl [85] for the 3D case.

3.1 The two-dimensional case

Let $\Omega \subset \mathbb{R}^2$ (i.e. $d = 2$). To construct a basis for $\mathring{\mathcal{V}}$ we will follow Ewing & Wang [38] and write the divergence-free Raviart-Thomas elements as the curls of suitable *stream functions* introduced in the following Section 3.1.1.

In Sections 3.1.2 and 3.1.3, we will prove that our construction of a basis for general mixed boundary conditions works for any simply connected domain Ω. Finally, in Section 3.1.4 we will show how these results can be extended to multiply connected domains, provided the Dirichlet boundary Γ_D is contained within one connected component of the boundary. The general, multiply connected case requires the introduction of a small number of additional basis functions, and we will present their construction for a special example.

3.1.1 The stream function space – C^0-elements in $H^1(\Omega)$

To construct the stream function space for $\mathring{\mathcal{V}}$, let us first look at the continuous problem, and recall (2.20) that

$$\mathcal{Z} := \{\vec{v} \in H_{0,N}(\mathrm{div}, \Omega) : b(\vec{v}, w) = 0 \text{ for all } w \in L_2(\Omega)\}. \tag{3.5}$$

We have the following fundamental result.

Proposition 3.1. *Let Ω be simply connected and $\Gamma_N = \emptyset$. A function $\vec{v} \in (L_2(\Omega))^2$ is in $\mathcal{Z}$, if and only if there exists a stream function $\Phi \in H^1(\Omega)$ such that:*

$$\vec{v} = \vec{\mathrm{curl}}\,\Phi := (\partial\Phi/\partial x_2, -\partial\Phi/\partial x_1)^T.$$

Proof. See Girault & Raviart [44, Theorem I.3.1]. □

Remark 3.2. Throughout, if $\vec{x}, \vec{y}$ are vectors in $\mathbb{R}^2$, then $\vec{x} \times \vec{y}$ denotes the x_3 component of the cross product of the vectors $(x_1, x_2, 0)^T$ and $(y_1, y_2, 0)^T$, and for any $\Phi : \mathbb{R}^2 \to \mathbb{R}$, $\vec{\mathrm{curl}}\,\Phi$ denotes the x_1 and x_2 components of the 3D curl of the function $(0, 0, \Phi)^T$ which can be expressed $\vec{\mathrm{curl}}\,\Phi = (\partial\Phi/\partial x_2, -\partial\Phi/\partial x_1)^T$.

Note that, formally the 2D curl of a function $\Phi \in H^1(\Omega)$ has to be understood in the sense of distributions, i.e.

$$\int_\Omega \vec{\mathrm{curl}}\,\Phi \cdot \vec{\xi}\, d\vec{x} = \int_\Omega \Phi \left(\tfrac{\partial \xi_2}{\partial x_1} - \tfrac{\partial \xi_1}{\partial x_2}\right) d\vec{x}, \qquad \text{for all } \vec{\xi} \in (\mathcal{D}(\Omega))^2.$$

We now come back to the discrete problem and the space $\mathring{\mathcal{V}} \subset \mathcal{Z}$ (cf. Lemma 2.8 (a)). Proposition 3.1 motivates us to seek the divergence-free Raviart-Thomas elements as the 2D curls of suitable finite elements in $H^1(\Omega)$.

Similar to Proposition 2.7 we can characterise functions in $H^1(\Omega)$ in the following way.

Proposition 3.3. *A function $\Phi \in L_2(\Omega)$ is in $H^1(\Omega)$, if*

$$\Phi \in C^0(\overline{\Omega}) \qquad \text{and} \qquad \Phi|_T \in H^1(T), \quad \text{for all } T \in \mathcal{T}.$$

Proof. See Ciarlet [27, Theorem 2.1.1]. $\square$

Let $\mathcal{T}$ be a simplicial triangulation of Ω as defined in Definition 2.10. Furthermore, let k be a non-negative integer and recall (2.38), that on each element $T \in \mathcal{T}$

$$P_{k+1}(T) := \quad \text{the space of polynomials of total degree } \leq k+1 \tag{3.6}$$

and that

$$\dim P_{k+1}(T) = \frac{(k+2)(k+3)}{2}. \tag{3.7}$$

Following Ciarlet [27, Theorem 2.2.1], any polynomial $p \in P_{k+1}(T)$ is uniquely determined by its values on the set

$$\Sigma_{k+1}(T) := \left\{ \vec{x} := \sum_{i=1}^{3} \lambda_i \vec{a}_i \ : \ \sum_{i=1}^{3} \lambda_i = 1 \text{ and } \lambda_i \in \{0, \tfrac{1}{k+1}, \ldots, \tfrac{k}{k+1}, 1\} \right\} \tag{3.8}$$

where $\vec{a}_1, \vec{a}_2, \vec{a}_3$ are the vertices of T and $\lambda_1, \lambda_2, \lambda_3$ are the *barycentric coordinates.* For example

$$\Sigma_1(T) = \{\vec{a}_1, \vec{a}_2, \vec{a}_3\} \quad \text{and} \quad \Sigma_2(T) = \{\vec{a}_1, \vec{a}_2, \vec{a}_3, \tfrac{\vec{a}_1+\vec{a}_2}{2}, \tfrac{\vec{a}_1+\vec{a}_3}{2}, \tfrac{\vec{a}_2+\vec{a}_3}{2}\}$$

as depicted in Figure 3.1.

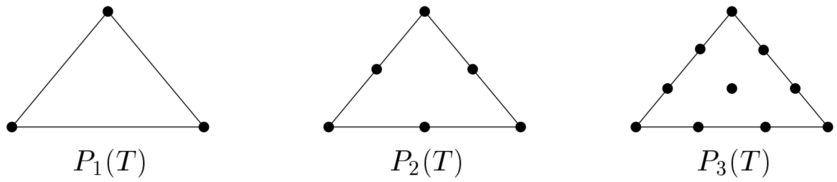


Figure 3.1: Degrees of Freedom for the spaces $P_{k+1}(T)$ for $k = 0, 1, 2$.

We can use the polynomial spaces $P_{k+1}(T)$, for each $T \in \mathcal{T}$, in a natural way now to define the following family of finite element spaces

$$\mathcal{S}_{k+1}(\Omega, \mathcal{T}) := \{\Phi \in C^0(\overline{\Omega}) : \Phi|_T \in P_{k+1}(T) \ \text{ for all } T \in \mathcal{T}\} \tag{3.9}$$

called the C^0-elements (or *Lagrange elements*). Proposition 3.3 guarantees that the spaces $\mathcal{S}_{k+1}(\Omega, \mathcal{T})$ are conforming in $H^1(\Omega)$, i.e.

$$\mathcal{S}_{k+1}(\Omega, \mathcal{T}) \subset H^1(\Omega). \tag{3.10}$$

The dimension of $\mathcal{S}_{k+1}(\Omega, \mathcal{T})$ can be calculated easily from the dimensions of the spaces

$P_{k+1}(T)$ taking into account the continuity of the functions $\Phi \in \mathcal{S}_{k+1}(\Omega, \mathcal{T})$ across any interface of two elements. We have

$$\dim \mathcal{S}_{k+1}(\Omega, \mathcal{T}) = \#\mathcal{N} + k\#\mathcal{F} + \frac{k(k-1)}{2}\#\mathcal{T} \tag{3.11}$$

and the global set of degrees of freedom

$$\Sigma_{k+1} := \bigcup_{T \in \mathcal{T}} \Sigma_{k+1}(T)$$

therefore contains all the nodes $P \in \mathcal{N}$, k equidistributed nodes in the interiors of each face $F \in \mathcal{F}$ (for $k > 0$), and $\frac{k(k-1)}{2}$ equidistributed nodes in the interior of each $T \in \mathcal{T}$ (for $k > 1$). Thus, the canonical basis for $\mathcal{S}_{k+1}(\Omega, \mathcal{T})$ is given by the functions $\Phi_P \in \mathcal{S}_{k+1}(\Omega, \mathcal{T})$, $P \in \Sigma_{k+1}$, such that

$$\Phi_P(P') = \delta_{P,P'}, \qquad \text{for all } P' \in \Sigma_{k+1}. \tag{3.12}$$

Example 3.4. *The lowest order case:* $k = 0$

In the case $k = 0$ we have $\Sigma_1 = \mathcal{N}$, and the space $\mathcal{S}_1(\Omega, \mathcal{T})$ consists of continuous piecewise linear functions with the canonical basis

$$\Big\{ \Phi_P \in \mathcal{S}_1(\Omega, \mathcal{T}) \ : \ P \in \mathcal{N} \ \text{ such that } \ \Phi_P(P') = \delta_{P,P'}, \ \text{ for all } P' \in \mathcal{N} \Big\}, \tag{3.13}$$

usually called the *hat functions* because of their particular form (see Figure 3.2). □

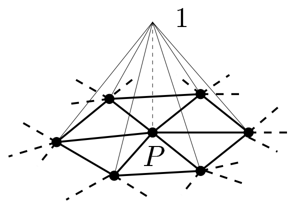


Figure 3.2: A typical basis function for $\mathcal{S}_1(\Omega, \mathcal{T})$ associated with node P.

In the next two sections we will use these basis functions of the spaces $\mathcal{S}_{k+1}(\Omega, \mathcal{T})$ to construct a basis for $\mathring{\mathcal{V}}$ for general mixed boundary conditions on any simply connected domain Ω. First, in Section 3.1.2 we will present an elementary approach for the lowest order case $k = 0$, mainly to motivate the construction of a basis for the lowest order case in 3D. Then, in Section 3.1.3 we will give a more general proof that extends the results to higher order elements in 2D. Therefore for the moment, unless otherwise specified, let Ω be simply connected.

3.1.2 The space $\mathring{\mathcal{V}}$ – an elementary approach for $k = 0$

Recall that by $\mathcal{F} = \mathcal{F}_I \cup \mathcal{F}_D \cup \mathcal{F}_N$ we denoted the set of all faces of the mesh $\mathcal{T}$, assumed to be open intervals and partitioned into the faces $\mathcal{F}_I$ in the interior of Ω, the faces $\mathcal{F}_D$ on the Dirichlet boundary Γ_D and the faces $\mathcal{F}_N$ on the Neumann boundary Γ_N (cf. Section 2.2.2). Analogously, we denoted by $\mathcal{N} = \mathcal{N}_I \cup \mathcal{N}_D \cup \mathcal{N}_N$ the set of all nodes of $\mathcal{T}$. When we defined the problem at the beginning of Section 2.1, we assumed that the endpoints of each of the components of Γ_N belong to Γ_N, and since the collision points between Neumann and Dirichlet boundaries are mesh points, these endpoints lie in $\mathcal{N}_N$.

Now let $k = 0$, and let

$$\{\Phi_P : P \in \mathcal{N}\}$$

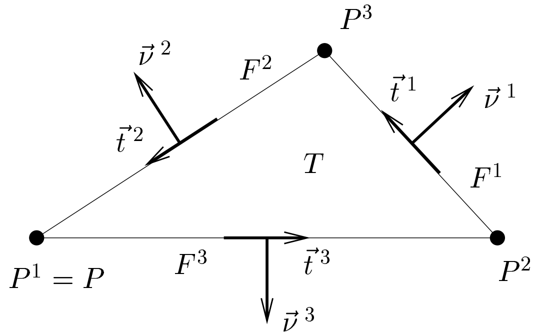
be the canonical basis of $\mathcal{S}_1(\Omega, \mathcal{T})$ as defined in (3.13). The basis for $\mathring{\mathcal{V}}$ will be constructed from the fundamental functions:

$$\vec{\Psi}_P := \vec{\operatorname{curl}}\, \Phi_P = (\partial\Phi_P/\partial x_2, -\partial\Phi_P/\partial x_1)^T, \qquad P \in \mathcal{N}. \tag{3.14}$$

This means Φ_P is the *stream function* for $\vec{\Psi}_P$ (cf. Proposition 3.1). The functions (3.14) clearly satisfy $\operatorname{div}\vec{\Psi}_P = 0$ on each triangle of the mesh, and a subset of them lie in $\mathring{\mathcal{V}}$ as the following proposition shows.

Proposition 3.5. *For each $P \in \mathcal{N}_I \cup \mathcal{N}_D$, $\quad \vec{\Psi}_P \in \mathring{\mathcal{V}}$.*

Proof. First consider the general case of $P \in \mathcal{N}$. Then clearly supp $\vec{\Psi}_P$ consists only of the triangles touching node P. A typical such triangle T with nodes $P^1 := P$, P^2, and P^3 is depicted below,



with $\vec{t}^{\,\alpha}$ and $\vec{\nu}^{\,\alpha}$ denoting unit vectors in the directions shown, and $P^\alpha = (p_1^\alpha, p_2^\alpha)^T$, $\alpha = 1, 2, 3$. Then

$$\Phi_P(\vec{x}) := \frac{1}{2|T|} \begin{vmatrix} 1 & x_1 & x_2 \\ 1 & p_1^2 & p_2^2 \\ 1 & p_1^3 & p_2^3 \end{vmatrix}, \quad \vec{x} \in T$$

and it follows easily that

$$\vec{\Psi}_P(\vec{x}) = \frac{|F^1|}{2|T|}\vec{t}^{\,1}, \quad \vec{x} \in T,$$

which is easily seen to be of the form (2.54) (in fact with $\gamma = 0$). Furthermore

$$\left.\begin{array}{rcl}\vec{\Psi}_P(\vec{x}) \cdot \vec{\nu}^{\,1} & = & \dfrac{|F^1|}{2|T|}\,\vec{t}^{\,1} \cdot \vec{\nu}^{\,1} = 0 \\ \vec{\Psi}_P(\vec{x}) \cdot \vec{\nu}^{\,2} & = & \dfrac{|F^1|}{2|T|}\,\vec{t}^{\,1} \cdot \vec{\nu}^{\,2} = \dfrac{|F^1|}{2|T|}\,\vec{t}^{\,1} \times \vec{t}^{\,2} = \dfrac{1}{|F^2|} \\ \vec{\Psi}_P(\vec{x}) \cdot \vec{\nu}^{\,3} & = & \dfrac{|F^1|}{2|T|}\,\vec{t}^{\,1} \cdot \vec{\nu}^{\,3} = \dfrac{|F^1|}{2|T|}\,\vec{t}^{\,1} \times \vec{t}^{\,3} = -\dfrac{1}{|F^3|}\end{array}\right\}\ \vec{x} \in T\,. \tag{3.15}$$

Now to obtain the result observe that, since $\mathrm{div}\vec{\Psi}_P = 0$ trianglewise, it is sufficient to show that

$$\vec{\Psi}_P \in \mathcal{V}, \quad \text{for all } P \in \mathcal{N}_I \cup \mathcal{N}_D\,. \tag{3.16}$$

To obtain (3.16), consider first $P \in \mathcal{N}_I$. Let $F \in \mathcal{F}_I$. If $F \not\subset \mathrm{supp}\,\vec{\Psi}_P$ we have trivially

$$\vec{\Psi}_P \cdot \vec{\nu}_F \quad \text{is continuous across } F. \tag{3.17}$$

Now take a general triangle $T \subset \mathrm{supp}\,\vec{\Psi}_P$, as pictured above. If $F = F^3$, then performing the computation (3.15) in the other triangle adjoining F and combining with (3.15) establishes (3.17). Similarly (3.17) holds if $F = F^2$. On the other hand, when $F = F^1$, (3.17) also holds since $\vec{\Psi}_P \cdot \vec{\nu}_F|_T = 0$ and since the other triangle adjoining F lies outside $\mathrm{supp}\,\vec{\Psi}_P$. Altogether we have established that $\vec{\Psi}_P$ satisfies criterion (2.55)(i).

To establish (2.55)(ii), let $F \in \mathcal{F}_N$. If $F \not\subset \mathrm{supp}\,\vec{\Psi}_P$, then $\vec{\Psi}_P \cdot \vec{\nu}_F = 0$ trivially. If $F \subset T \subset \mathrm{supp}\,\vec{\Psi}_P$, then with the above notation F has to be F^1 and again $\vec{\Psi}_P \cdot \vec{\nu}_F = 0$, proving (2.55)(ii).

Thus we have shown that $\vec{\Psi}_P \in \mathcal{V}$ for all $P \in \mathcal{N}_I$. Similar arguments establish that $\vec{\Psi}_P \in \mathcal{V}$ for all $P \in \mathcal{N}_D$, proving (3.16). □

Note that each $\vec{\Psi}_P$ can be expressed as a local linear combination of the basis functions $\vec{v}_F$ of $\mathcal{V}$ satisfying (2.75); in fact only those $\vec{v}_F$ corresponding to faces F touching node P appear in the expansion of $\vec{\Psi}_P$ (see Figure 3.3 (left)).

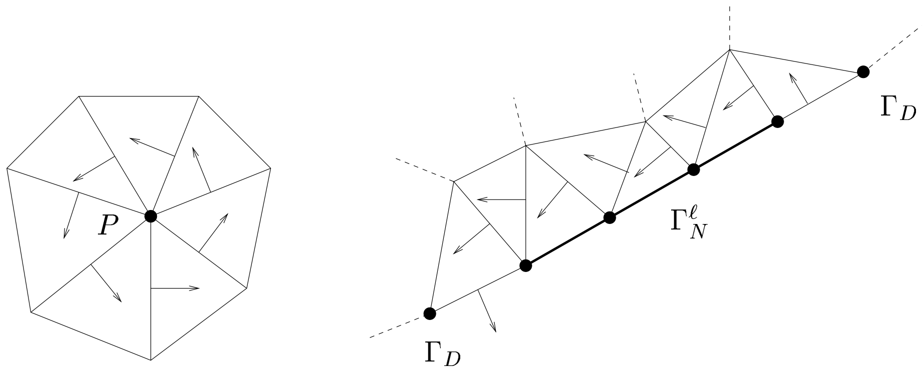


Figure 3.3: Divergence-free basis functions $\vec{\Psi}_P$ (left) and $\sum_{P \in \mathcal{N}_N^\ell} \vec{\Psi}_P$ (right)

The functions introduced in Proposition 3.5 span a subset of $\mathring{\mathcal{V}}$, but for general boundary conditions, there are not enough of them to constitute a basis for $\mathring{\mathcal{V}}$. A small number of additional basis functions may need to be added. These are introduced in Proposition 3.6. In this we let s_N denote the number of connected components in Γ_N and write

$$\Gamma_N = \Gamma_N^1 \cup \Gamma_N^2 \cup \ldots \cup \Gamma_N^{s_N} \ , \qquad \Gamma_N^{\ell} \cap \Gamma_N^{\ell'} = \emptyset \quad \text{for all } 1 \leq \ell \neq \ell' \leq s_N \ . \tag{3.18}$$

For $\ell = 1, \ldots, s_N$, let $\mathcal{N}_N^{\ell} \subset \mathcal{N}$ denote the set of mesh nodes on Γ_N^{ℓ}.

Proposition 3.6. *For each $\ell = 1, \ldots, s_N$,*

$$\sum_{P \in \mathcal{N}_N^{\ell}} \vec{\Psi}_P \in \mathring{\mathcal{V}} \ . \tag{3.19}$$

Proof. Let $\ell \in \{1, \ldots, s_N\}$. Analogously to the proof of Proposition 3.5 it is sufficient to show that $\left(\sum_{P \in \mathcal{N}_N^{\ell}} \vec{\Psi}_P\right) \in \mathcal{V}$. Criterion (2.55)(i) is a mere consequence of (3.15) again, and to establish (2.55)(ii) observe first that the components $\Gamma_N^{\ell'}$, $\ell' = 1, \ldots, s_N$, are disjoint, and therefore $\left(\sum_{P \in \mathcal{N}_N^{\ell}} \vec{\Psi}_P \cdot \vec{\nu}_F\right)\Big|_F = 0$ for all $F \subset \Gamma_N^{\ell'}$, $\ell' \neq \ell$, by definition. Now let $F \subset \Gamma_N^{\ell}$, and let $P^1, P^2 \in \mathcal{N}_N^{\ell}$ be the two endpoints of F. Then, using (3.15),

$$\Big| \sum_{P \in \mathcal{N}_N^{\ell}} \vec{\Psi}_P \cdot \vec{\nu}_F \Big| = \Big| \vec{\Psi}_{P^1} \cdot \vec{\nu}_F + \vec{\Psi}_{P^2} \cdot \vec{\nu}_F \Big| = \left| \frac{1}{|F|} - \frac{1}{|F|} \right| = 0 \qquad \text{on } F. \tag{3.20}$$

Thus $\left(\sum\limits_{P \in \mathcal{N}_N^{\ell}} \vec{\Psi}_P \cdot \vec{\nu}_F\right)\Big|_F = 0$ for all $F \in \mathcal{F}_N$, and the proof is complete. ☐

In contrast to the functions in Proposition 3.5, the functions introduced in Proposition 3.6 are non-local linear combinations of the functions $\vec{v}_F$; however, the non-locality of $\sum_{P \in \mathcal{N}_N^{\ell}} \vec{\Psi}_P$ is confined to the vicinity of Γ_N^{ℓ} (see Figure 3.3 (right)). The number of such triangles is typically $O((\#\mathcal{T})^{1/2})$, so they are only modestly non-local.

From these elementary results we have our first theorem. It shows that when $\Gamma_N \neq \emptyset$, by combining all the functions found in Proposition 3.5 with all but one of the functions in Proposition 3.6 we have the required basis.

Theorem 3.7. *Suppose $s_N \neq 0$. Then the functions*

$$\{\vec{\Psi}_P : P \in \mathcal{N}_I \cup \mathcal{N}_D\} \cup \{ \sum_{P \in \mathcal{N}_N^{\ell}} \vec{\Psi}_P : \ell = 1, \ldots, s_N - 1\} \tag{3.21}$$

are a basis for $\mathring{\mathcal{V}}$.

Before we prove the theorem, let us define the *Euler characteristic* of a compact, two-dimensional surface in $\mathbb{R}^3$.

Definition 3.8. Let M be a compact, two-dimensional surface in $\mathbb{R}^3$ which is subdivided into polygons. Then

$$\chi(M) := \#\text{ nodes} - \#\text{ edges} + \#\text{ polygons} \tag{3.22}$$

is called the *Euler characteristic* of M.

Lemma 3.9. *The Euler characteristic of an orientable*[1]*, compact, two-dimensional surface M with $s(M)$ disjoint boundary components is*

$$\chi(M) = 2 - s(M). \tag{3.23}$$

Proof. See Massey [70, Page 44]. □

Proof of Theorem 3.7. Suppose $s_N \neq 0$. First we shall check that the number of basis functions in (3.21) coincides with $\mathring{n} = \dim \mathring{\mathcal{V}}$. Consider a typical Neumann boundary segment Γ_N^ℓ. This contains $\#\mathcal{N}_N^\ell$ nodes, two of which are the end points of Γ_N^ℓ, and so the number of edges in Γ_N^ℓ is $(\#\mathcal{N}_N^\ell - 1)$. Summing over $\ell = 1, \ldots, s_N$, we obtain

$$\#\mathcal{F}_N = \#\mathcal{N}_N - s_N \ . \tag{3.24}$$

Now the number of functions in (3.21) is $(\#\mathcal{N}_I + \#\mathcal{N}_D + s_N - 1)$ and, using the fact that $\mathcal{N}_I$, $\mathcal{N}_D$ and $\mathcal{N}_N$ partition $\mathcal{N}$, together with (3.24), we have

$$(\#\mathcal{N}_I + \#\mathcal{N}_D + s_N - 1) = (\#\mathcal{N} - \#\mathcal{N}_N + s_N - 1) = (\#\mathcal{N} - \#\mathcal{F}_N - 1) \ . \tag{3.25}$$

Furthermore, using the fact that $\mathcal{F}_I$, $\mathcal{F}_D$ and $\mathcal{F}_N$ partition $\mathcal{F}$, we have

$$(\#\mathcal{N} - \#\mathcal{F}_N - 1) = (\#\mathcal{F}_I + \#\mathcal{F}_D) - (\#\mathcal{F} - \#\mathcal{N} + 1) = (\#\mathcal{F}_I + \#\mathcal{F}_D) - \#\mathcal{T} \ , \tag{3.26}$$

where in the last step we have used Lemma 3.9 and the assumption that Ω is simply connected (and has therefore a connected boundary). Now recalling (3.4), that $\mathring{n} = n_{\mathcal{V}} - n_{\mathcal{W}}$ (and from Example 2.16 we have $n_{\mathcal{V}} = (\#\mathcal{F}_I + \#\mathcal{F}_D)$ and $n_{\mathcal{W}} = \#\mathcal{T}$), it follows from (3.25) and (3.26) that the number of functions in (3.21) is $\mathring{n}$, as required.

To complete the proof we merely need to show that the functions in (3.21) are linearly independent. So suppose $\{\alpha_P : P \in \mathcal{N}_I \cup \mathcal{N}_D\}$ and $\{\beta_\ell : \ell = 1, \ldots, s_N - 1\}$ are scalars such that

$$\vec{0} = \sum_{P \in \mathcal{N}_I \cup \mathcal{N}_D} \alpha_P \vec{\Psi}_P + \sum_{\ell=1}^{s_N - 1} \beta_\ell \sum_{P \in \mathcal{N}_N^\ell} \vec{\Psi}_P \ . \tag{3.27}$$

This may be rewritten

$$\vec{0} = \sum_{P \in \mathcal{N}} \alpha_P \vec{\Psi}_P \ , \tag{3.28}$$

[1] The simplest example of a non-orientable, compact, two-dimensional surface is the well-known *Möbius strip*. For an exact definition of orientable see Massey [70, Section I.3]

where $\alpha_P := \beta_\ell$ when $P \in \mathcal{N}_N^\ell$ for $\ell = 1, \ldots, s_N - 1$ and $\alpha_P := 0$ for $P \in \mathcal{N}_N^{s_N}$. Now consider any face $F \in \mathcal{F}$ with end points P' and P''.

If $F \in \mathcal{F}_I \cup \mathcal{F}_D$, combining (3.28) with (3.15) we get

$$0 = \Big| \sum_{P \in \mathcal{N}} \alpha_P \vec{\Psi}_P \cdot \vec{\nu}_F \Big| = \left| \alpha_{P'} \frac{1}{|F|} - \alpha_{P''} \frac{1}{|F|} \right| = \frac{1}{|F|} |\alpha_{P'} - \alpha_{P''}|$$

which shows $\alpha_{P'} = \alpha_{P''}$.

On the other hand if $F \in \mathcal{F}_N$, $\alpha_{P'} = \alpha_{P''}$ by definition. Since every two vertices can be connected by a series of edges, it follows that α_P is independent of P. Recalling that $\mathcal{N}_N^{s_N} \neq \emptyset$, we have $\alpha_P = 0$ for all $P \in \mathcal{N}$. □

Remark 3.10. In the pure Dirichlet case (i.e. $\Gamma_N = \emptyset$), a suitable basis is $\{\vec{\Psi}_P : P \in \mathcal{N},\ P \neq P_0\}$ for any choice of $P_0 \in \mathcal{N}$. The proof follows exactly the same steps with a few changes in notation.

3.1.3 The space $\mathring{\mathcal{V}}$ – the general case

Now let $k \geq 0$, and let

$$\{\Phi_P : P \in \Sigma_{k+1}\}$$

be the canonical basis of $\mathcal{S}_{k+1}(\Omega, \mathcal{T})$, as defined in (3.12). The basis for $\mathring{\mathcal{V}}$ will again be constructed from the fundamental functions:

$$\vec{\Psi}_P := \vec{\operatorname{curl}}\, \Phi_P, \qquad P \in \Sigma_{k+1} \tag{3.29}$$

This means Φ_P is the *stream function* for $\vec{\Psi}_P$. The functions (3.29) clearly satisfy $\operatorname{div}\vec{\Psi}_P = 0$ on each triangle of the mesh, and a subset of them lie in $\mathring{\mathcal{V}}$ as the following proposition shows.

Proposition 3.11. *Let $P \in \Sigma_{k+1}$. If $P \notin \Gamma_N$, then*

$$\vec{\Psi}_P \in \mathring{\mathcal{V}}.$$

Proof. Consider any $P \in \Sigma_{k+1}$. For each $T \in \mathcal{T}$, we have by definition that $\Phi_P|_T \in P_{k+1}(T)$, and therefore $\vec{\Psi}_P|_T \in (P_k(T))^2 \subset RT_k(T)$. Furthermore, using (3.10) we have $\Phi_P \in H^1(\Omega)$ and we can apply Proposition 3.1 to obtain $\vec{\Psi}_P = \vec{\operatorname{curl}}\, \Phi_P \in H(\operatorname{div}, \Omega)$. Thus it follows from (2.47) that $\vec{\Psi}_P \in \mathcal{RT}_k(\Omega, \mathcal{T})$.

Next we will show that $\vec{\Psi}_P \in \mathcal{V}$, and thus by (2.50) we only have to verify that

$$\vec{\Psi}_P \cdot \vec{\nu}|_{\Gamma_N} = 0, \qquad \text{for all } P \in \Sigma_{k+1} \text{ with } P \notin \Gamma_N, \tag{3.30}$$

where $\vec{\nu}$ denotes the unit outward normal from Ω on Γ_N as defined at the beginning of Section 2.1. To obtain (3.30) consider an arbitrary node $P \in \Sigma_{k+1}$ with $P \notin \Gamma_N$. Let $F \in \mathcal{F}_N$. If $F \not\subset \operatorname{supp}\vec{\Psi}_P$, then $\vec{\Psi}_P \cdot \vec{\nu}|_F = 0$ trivially. Let $F \subset T$ where $T \subset \operatorname{supp}\vec{\Psi}_P$.

By definition, $\Phi_P|_T \in P_{k+1}(T)$ and therefore the restriction $\Phi_P|_F \in P_{k+1}(F)$ Now we know from (3.12) that $\Phi_P|_F$ has $k+2$ roots on $\overline{F}$, implying $\Phi_P|_F = 0$. Thus, if $S := \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ and if $\vec{\tau}(\vec{x}) := S^T \vec{\nu}(\vec{x})$ denotes the unit tangent vector at $\vec{x} \in \Gamma_N$, then $\vec{\nabla}\, \Phi_P \cdot \vec{\tau}|_F = 0$, and it follows from (3.29) that

$$\vec{\Psi}_P \cdot \vec{\nu}|_F = \vec{\operatorname{curl}}\, \Phi_P \cdot \vec{\nu}|_F = S\vec{\nabla}\, \Phi_P \cdot \vec{\nu}|_F = \vec{\nabla}\, \Phi_P \cdot \vec{\tau}|_F = 0,$$

which proves (3.30).

The result then follows directly from the definition (3.1) of $\mathring{\mathcal{V}}$, since $\operatorname{div}\vec{\Psi}_P = 0$ trianglewise, and therefore by definition $b(\vec{\Psi}_P, W) = 0$, for all $W \in \mathcal{W}$ (see (2.10)). $\square$

As in Section 3.1.2, the functions introduced in Proposition 3.11 span a subset of $\mathring{\mathcal{V}}$, but for general boundary conditions, there are not enough of them to constitute a basis of $\mathring{\mathcal{V}}$ and a small number of additional basis functions may need to be added. These are introduced in Proposition 3.12. Recall the partitioning (3.18) of Γ_N into s_N connected components Γ_N^ℓ, $\ell = 1, \ldots, s_N$. For each $\ell = 1, \ldots, s_N$, let $\Sigma_{k+1}^\ell \subset \Sigma_{k+1}$ denote the set of degrees of freedom on Γ_N^ℓ.

Proposition 3.12. *For each $\ell = 1, \ldots, s_N$,*

$$\sum_{P \in \Sigma_{k+1}^\ell} \vec{\Psi}_P \in \mathring{\mathcal{V}}. \tag{3.31}$$

Proof. We know from the proof to Proposition 3.11 that $\vec{\Psi}_P \in \mathcal{RT}_k(\Omega, \mathcal{T})$ for all $P \in \Sigma_{k+1}$ and that $\operatorname{div}\vec{\Psi}_P = 0$ trianglewise. Thus it suffices to show that

$$\Big(\sum_{P \in \Sigma_{k+1}^\ell} \vec{\Psi}_P \cdot \vec{\nu} \Big)\Big|_{\Gamma_N} = 0, \qquad \text{for all } \ell = 1, \ldots, s_N. \tag{3.32}$$

Let $\ell \in \{1, \ldots, s_N\}$. Observe first that the components $\Gamma_N^{\ell'}$, $\ell' = 1, \ldots, s_N$, are disjoint, and therefore $\Big(\sum_{P \in \Sigma_{k+1}^\ell} \vec{\Psi}_P \cdot \vec{\nu} \Big)\Big|_{\Gamma_N^{\ell'}} = 0$, for all $\ell' \neq \ell$, by definition. Now let $F \subset \Gamma_N^\ell$. As in the proof to Proposition 3.11 it follows from (3.12) that there are $k+2$ distinct points $P' \in \overline{F}$ such that $\Big(\sum_{P \in \Sigma_{k+1}^\ell} \Phi_P \Big)(P') = 1$, which implies that $\Big(\sum_{P \in \Sigma_{k+1}^\ell} \Phi_P \Big)\Big|_F = 1$. Finally, using the same arguments as in that proof, we deduce

$$\Big(\sum_{P \in \Sigma_{k+1}^\ell} \vec{\Psi}_P \cdot \vec{\nu} \Big)\Big|_F = \Big(\sum_{P \in \Sigma_{k+1}^\ell} \vec{\nabla}\, \Phi_P \cdot \vec{\tau} \Big)\Big|_F = 0.$$

$\square$

By combining all the functions found in Proposition 3.11 with all but one of the functions in Proposition 3.12 we have the required basis of $\mathring{\mathcal{V}}$ again, as the following theorem shows.

Theorem 3.13. *Suppose $s_N \neq 0$. Then the functions*

$$\{\vec{\Psi}_P : P \in \Sigma_{k+1} \text{ with } P \notin \Gamma_N\} \cup \{\sum_{P\in\Sigma_{k+1}^{\ell}} \vec{\Psi}_P : \ell = 1, \ldots, s_N - 1\} \tag{3.33}$$

are a basis for $\mathring{\mathcal{V}}$.

Proof. Suppose $s_N \neq 0$. First we shall check that the number of basis functions in (3.33) coincides with $\mathring{n} = \dim \mathring{\mathcal{V}}$, if $k > 0$ (see the proof to Theorem 3.7 for $k = 0$). Using (3.11) we know that the number of degrees of freedom $P \in \Sigma_{k+1}$ with $P \notin \Gamma_N$ is

$$(\#\mathcal{N} - \#\mathcal{N}_N) + k(\#\mathcal{F} - \#\mathcal{F}_N) + \tfrac{k(k-1)}{2}\#\mathcal{T}.$$

Therefore, the number of functions in (3.33) is

$$(\#\mathcal{N} - \#\mathcal{N}_N) + k(\#\mathcal{F} - \#\mathcal{F}_N) + \tfrac{k(k-1)}{2}\#\mathcal{T} + s_N - 1$$

which is the same as

$$(k+1)(\#\mathcal{F} - \#\mathcal{F}_N) + \tfrac{(k+1)(k-2)}{2}\#\mathcal{T} + R \tag{3.34}$$

with $R = (\#\mathcal{N} - \#\mathcal{F} + \#\mathcal{T} - 1) - (\#\mathcal{N}_N - \#\mathcal{F}_N - s_N)$. Since Ω is simply connected, we can apply Lemma 3.9 and use (3.24) to obtain $R = 0$. Substituting this into (3.34) and using the fact that $\mathcal{F}_I$, $\mathcal{F}_D$ and $\mathcal{F}_N$ partition $\mathcal{F}$, we have that the number of functions in (3.33) is

$$(k+1)(\#\mathcal{F}_I + \#\mathcal{F}_D + k\#\mathcal{T}) - \tfrac{(k+1)(k+2)}{2}\#\mathcal{T} = n_{\mathcal{V}} - n_{\mathcal{W}}$$

where in the last step we used (2.51) and (2.53). Now recalling (3.4), this is equal to $\mathring{n}$, as required.

To complete the proof we merely need to show that the functions in (3.33) are linearly independent. So suppose $\{\alpha_P : P \in \Sigma_{k+1} \text{ with } P \notin \Gamma_N\}$ and $\{\beta_\ell : \ell = 1, \ldots, s_N - 1\}$ are scalars such that

$$\vec{0} = \sum_{\substack{P\in\Sigma_{k+1} \text{ s.t.} \\ P\notin\Gamma_N}} \alpha_P \vec{\Psi}_P + \sum_{\ell=1}^{s_N-1} \beta_\ell \sum_{P\in\mathcal{N}_N^{\ell}} \vec{\Psi}_P \; .$$

Using the linearity of $\vec{\text{curl}}$, this may be rewritten

$$\vec{0} = \sum_{P\in\Sigma_{k+1}} \alpha_P \vec{\Psi}_P = \sum_{P\in\Sigma_{k+1}} \alpha_P \,\vec{\text{curl}}\, \Phi_P = \vec{\text{curl}} \Big(\sum_{P\in\Sigma_{k+1}} \alpha_P \Phi_P \Big),$$

where $\alpha_P := \beta_\ell$ when $P \in \Sigma_{k+1}^{\ell}$ for $\ell = 1, \ldots, s_N - 1$ and $\alpha_P := 0$ for $P \in \Sigma_{k+1}^{s_N}$. But this implies that $\sum_{P\in\Sigma_{k+1}} \alpha_P \Phi_P$ is constant on Ω and hence (since $\Sigma_{k+1}^{s_N} \neq \emptyset$), $\alpha_P = 0$

for all $P \in \Sigma_{k+1}$. $\square$

Remark 3.14.

(a) In the pure Dirichlet case (i.e. $\Gamma_N = \emptyset$), a suitable basis is $\{\vec{\Psi}_P : P \in \Sigma_{k+1},\ P \neq P_0\}$ for any choice of $P_0 \in \Sigma_{k+1}$. The proof follows exactly the same steps with a few changes in notation.

(b) Propositions 3.11, 3.12 and Theorem 3.13 can also be easily extended to functions $\vec{\Psi}_P$ constructed from other bases of $\mathcal{S}_{k+1}(\Omega, \mathcal{T})$. For example, in the case of a hierarchical basis, Proposition 3.11 holds true without modifications, whereas Proposition 3.12 and Theorem 3.13 hold true, if the sum (3.31) is composed only of degrees of freedom from the coarsest grid level. The proofs follow again exactly the same steps with a few changes in notation.

3.1.4 Extension to multiply connected domains

The results in the two previous sections extend in a straightforward manner to the case of a multiply connected domain Ω. This extension is a simple Corollary to Theorem 3.13, if the Dirichlet boundary Γ_D is contained in one connected component of the boundary (see Figure 3.4), an important special case in the context of groundwater

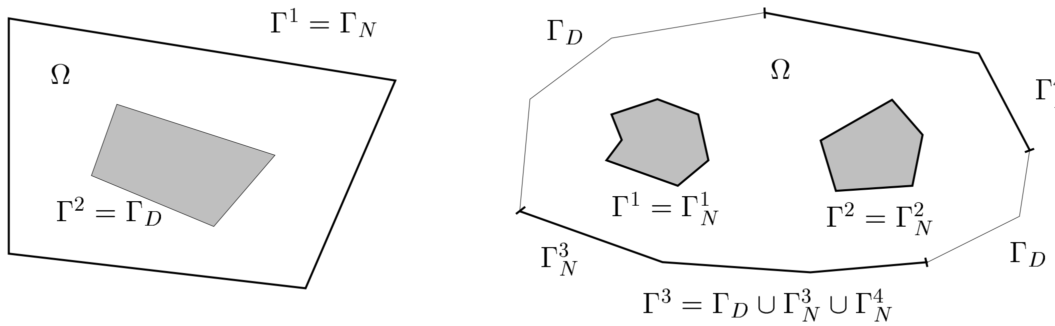


Figure 3.4: Two examples of multiply connected domains Ω.

flow, where the holes in Ω correspond to wells with a prescribed inflow or outflow. If the Dirichlet boundary Γ_D is not contained in one connected component of the boundary, on the other hand, we still obtain a basis for $\mathring{\mathcal{V}}$, if we introduce a small number of additional basis functions.

To make it precise, let Ω be a general, not necessarily simply connected domain, let s be the number of connected components in Γ and write

$$\Gamma = \Gamma^1 \cup \Gamma^2 \cup \ldots \cup \Gamma^s, \qquad \Gamma^\ell \cap \Gamma^{\ell'} = \emptyset \quad \text{for all } 1 \leq \ell \neq \ell' \leq s.$$

We have the following extension to Theorem 3.13.

Corollary 3.15. *Let $s_N \neq 0$ and let $\Gamma_D \subset \Gamma^s$. Then the functions*

$$\{\vec{\Psi}_P : P \in \Sigma_{k+1} \text{ with } P \notin \Gamma_N\} \cup \{\sum_{P\in\Sigma^\ell_{k+1}} \vec{\Psi}_P : \ell = 1, \ldots, s_N - 1\} \tag{3.35}$$

are a basis for $\mathring{\mathcal{V}}$.

Proof. Linear independence of the functions in (3.35) is a mere consequence of Theorem 3.13.

It remains to check whether the number of functions in (3.35) is $\mathring{n}$ as required. As in the proof to Theorem 3.13 (cf. (3.34)), the number of functions in (3.35) is

$$(k+1)(\#\mathcal{F} - \#\mathcal{F}_N) + \tfrac{(k+1)(k-2)}{2}\#\mathcal{T} + R \tag{3.36}$$

where now $R = (\#\mathcal{N} - \#\mathcal{F} + \#\mathcal{T} + s - 2) - (\#\mathcal{N}_N - \#\mathcal{F}_N - (s_N - s + 1))$.

It remains to show that $R = 0$. Since we assumed that Ω has s disjoint boundary components, we have

$$(\#\mathcal{N} - \#\mathcal{F} + \#\mathcal{T} + s - 2) = 0 \tag{3.37}$$

by Lemma 3.9 again. To show that the second term in R is also 0, let us first consider any $\ell \in \{1, \ldots, s-1\}$. Since we assumed that $\Gamma_D \subset \Gamma^s$ we know that $\Gamma^\ell \subset \Gamma_N$. Now let the connected components of Γ_N (as defined in (3.18)) be numbered such that

$$\Gamma_N^{\ell'} = \Gamma^{\ell'}, \qquad \text{for all } \ell' = 1, \ldots, s-1.$$

Then Γ_N^ℓ is a closed curve and contains $\#\mathcal{N}_N^\ell$ nodes and $\#\mathcal{N}_N^\ell$ faces. The remaining $(s_N - s + 1)$ components Γ_N^ℓ, $s \leq \ell \leq s_N$, on the other hand, are subsets of Γ^s and (since $\Gamma_D \neq \emptyset$) contain $\#\mathcal{N}_N^\ell$ nodes and $(\#\mathcal{N}_N^\ell - 1)$ faces. Summing over $\ell = 1, \ldots, s_N$, we obtain

$$\#\mathcal{F}_N = (\#\mathcal{N}_N - (s_N - s + 1)) \tag{3.38}$$

Combining (3.37) and (3.38) we have $R = 0$, as in the proof to Theorem 3.13, and the number of functions in (3.35) is $\mathring{n}$ as required. □

Finally, let us consider the most general situation of $s > 1$ and $\Gamma_D \cap \Gamma^\ell \neq \emptyset$ for at least two different $\ell = 1, \ldots, s$. This requires the introduction of additional basis functions in order to construct a basis for $\mathring{\mathcal{V}}$. Since the proof is very technical, we will restrict to a special example to illustrate this case. Thus, for the remainder of this section, unless otherwise specified, let us consider lowest order elements, i.e. $k = 0$, and let Ω have two holes, i.e. $s = 3$, and let the outer boundary be denoted by Γ^1. Furthermore, let $\Gamma_N = \Gamma^1$ and $\Gamma_D = \Gamma^2 \cup \Gamma^3$, as presented in Figure 3.5.

In order to identify the extra basis function needed in this case, it is useful to introduce some graph theory (see Appendix B). Now let $\mathbf{G} := (\mathcal{N}, \mathcal{F})$ denote the graph formed by the nodes and faces (i.e. edges in 2D) of the triangulation $\mathcal{T}$ and

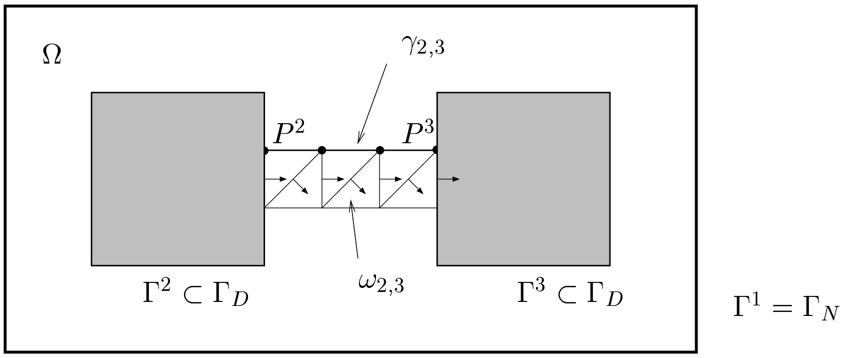


Figure 3.5: Additional basis function $\vec{\Psi}_{2,3}$ for a multiply connected domain Ω.

let $P^2, P^3 \in \mathcal{N}$ such that $P^2 \in \Gamma^2$ and $P^3 \in \Gamma^3$. By Definition B.1 (d), since $\mathbf{G}$ is connected, there exists an elementary chain $\gamma_{2,3}$ in $\mathbf{G}$ that connects P^2 and P^3 (as depicted in Figure 3.5). Let $\mathcal{N}_{2,3} \subset \mathcal{N}$ be the set of all nodes on $\gamma_{2,3}$. Since Γ^2 and Γ^3 are disjoint, we can assume that $\gamma_{2,3}$ uses no edge on Γ^2 or Γ^3, i.e. $\gamma_{2,3} \cap \Gamma_D = \{P^2, P^3\}$. Furthermore, we denote by $\omega_{2,3}$ the union of all triangles $T \in \mathcal{T}$ that fulfil $\gamma_{2,3} \cap \overline{T} \neq \emptyset$ and that lie to the right of $\gamma_{2,3}$ (when looking from P^2, see Figure 3.5 again).

Proposition 3.16. *Let*

$$\vec{\Psi}_{2,3}(\vec{x}) := \begin{cases} \sum\limits_{P \in \mathcal{N}_{2,3}} \vec{\Psi}_P(\vec{x}) & \textit{for all } \vec{x} \in \omega_{2,3}, \\ 0 & \textit{otherwise}. \end{cases} \tag{3.39}$$

Then $\vec{\Psi}_{2,3} \in \mathring{\mathcal{V}}$.

Proof. Obviously $\operatorname{div} \vec{\Psi}_{2,3} = 0$ trianglewise and it is sufficient again to show that $\vec{\Psi}_{2,3} \in \mathcal{V}$. Since $\operatorname{supp}\vec{\Psi}_{2,3} \cap \Gamma_N = \emptyset$, we only need to check that

$$\vec{\Psi}_{2,3} \cdot \vec{\nu}_F \quad \text{is continuous across } F, \tag{3.40}$$

for each face $F \in \mathcal{F}_I$. If F is in the interior of $\omega_{2,3}$, (3.40) is a mere consequence of (3.15). If $F \subset \gamma_{2,3}$, on the other hand, then there exists an element $T \subset \omega_{2,3}$ such that $F \subset T$, and performing a similar computation on T as in (3.20), we can show that $\vec{\Psi}_{2,3} \cdot \vec{\nu}_F|_T = 0$. Since the other triangle adjoining F lies outside $\operatorname{supp}\vec{\Psi}_{2,3}$, (3.40) holds also for $F \subset \gamma_{2,3}$. Similar arguments establish $\vec{\Psi}_{2,3} \cdot \vec{\nu}_F = 0$ for the remaining faces $F \in \mathcal{F}_I$, thus proving (3.40) for all $F \in \mathcal{F}_I$. Altogether we have established Criterion (2.55) and the proof is complete. $\square$

Note that $\vec{\Psi}_{2,3}$ can again be expressed as a linear combination of the basis functions $\vec{v}_F$ of $\mathcal{V}$ satisfying (2.75); in fact only those $\vec{v}_F$ corresponding to faces F marked with an arrow in Figure 3.5 appear in the expansion of $\vec{\Psi}_{2,3}$. Similar to the functions defined in Proposition 3.6, $\vec{\Psi}_{2,3}$ is non-local, but again the non-locality is confined to (typically)

$O((\#\mathcal{T})^{1/2})$ triangles.

Now, using this additional function $\vec{\Psi}_{2,3}$ together with the functions found in Proposition 3.5, we have the required basis for Ω as defined above.

Theorem 3.17. *Let $k = 0$, let Ω be defined as above and let $\vec{\Psi}_{2,3}$ be the function found in Proposition 3.16. Then the functions*

$$\{\vec{\Psi}_P : P \in \mathcal{N}_I \cup \mathcal{N}_D\} \cup \{\vec{\Psi}_{2,3}\} \tag{3.41}$$

are a basis for $\mathring{\mathcal{V}}$.

Proof. First we shall compare dimensions again. Since the number of disjoint boundary components of Ω is 3, by Lemma 3.9 we have

$$(\#\mathcal{N} - \#\mathcal{F} + \#\mathcal{T} + 1) = 0, \tag{3.42}$$

as in the proof to Corollary 3.15. Now the number of functions in (3.41) is $(\#\mathcal{N}_I + \#\mathcal{N}_D + 1)$ and, using the fact that $\mathcal{N}_I$, $\mathcal{N}_D$ and $\mathcal{N}_N$ partition $\mathcal{N}$, together with (3.42),

$$(\#\mathcal{N}_I + \#\mathcal{N}_D + 1) = (\#\mathcal{N} - \#\mathcal{N}_N + 1) = (\#\mathcal{F} - \#\mathcal{T} - \#\mathcal{N}_N). \tag{3.43}$$

Furthermore, using the fact that $\mathcal{F}_I$, $\mathcal{F}_D$ and $\mathcal{F}_N$ partition $\mathcal{F}$, and that $\#\mathcal{F}_N = \#\mathcal{N}_N$ (since Γ_N is a closed curve) we have

$$(\#\mathcal{F} - \#\mathcal{T} - \#\mathcal{N}_N) = (\#\mathcal{F}_I + \#\mathcal{F}_D - \#\mathcal{T}) = n_{\mathcal{V}} - n_{\mathcal{W}}, \tag{3.44}$$

where in the last step we used (2.56) and (2.57). Finally recalling (3.4), it follows from (3.43) and (3.44) that the number of functions in (3.41) is $\mathring{n}$ as required.

To show linear independency let $\{\alpha_P : P \in \mathcal{N}_I \cup \mathcal{N}_D\}$ and β be scalars such that

$$\vec{0} = \sum_{P \in \mathcal{N}_I \cup \mathcal{N}_D} \alpha_P \vec{\Psi}_P + \beta \vec{\Psi}_{2,3}.$$

This may be rewritten

$$\vec{0} = \sum_{P \in \mathcal{N}} \alpha_P \vec{\Psi}_P + \beta \vec{\Psi}_{2,3}, \tag{3.45}$$

with $\alpha_P := 0$ for $P \in \mathcal{N}_N$. Let $\tilde{\mathcal{F}} \subset \mathcal{F}$ denote the set of all faces $F \not\subset \overline{\omega}_{2,3}$. In the same way as in the proof to Theorem 3.7 we can show that for each face $F \in \tilde{\mathcal{F}}$, with end points P' and P'', we have $\alpha_{P'} = \alpha_{P''}$. Now, since Ω is multiply connected, the partial graph $\tilde{\mathbf{G}} := (\mathcal{N}, \tilde{\mathcal{F}})$ of $\mathbf{G}$ is still connected (see also Figure 3.5), and it follows that α_P is independent of P, as before. Recalling that $\mathcal{N}_N \neq \emptyset$, we have $\alpha_P = 0$ for all $P \in \mathcal{N}$. Finally, substituting this into (3.45) we get $\beta = 0$, which completes the proof. $\square$

Remark 3.18. In general, if m boundary components Γ^ℓ contain part of the Dirichlet boundary Γ_D, i.e. $\Gamma^\ell \cap \Gamma_D \neq \emptyset$, then we need exactly $m - 1$ functions $\vec{\Psi}_{i,j}$, as defined in Proposition 3.16, which link two such components Γ^i and Γ^j. This can be proved exactly in the same way as Theorem 3.17.

It is also straightforward to include more than one Neumann boundary component and therefore basis functions introduced in Proposition 3.6, or to employ higher order elements, i.e. $k > 0$. Here, the additional basis function $\vec{\Psi}_{2,3}$ is defined as

$$\vec{\Psi}_{2,3}(\vec{x}) := \begin{cases} \sum_{P \in \Sigma_{k+1}^{(2,3)}} \vec{\Psi}_P(\vec{x}) & \text{for all } \vec{x} \in \omega_{2,3}, \\ 0 & \text{otherwise,} \end{cases}$$

where $\Sigma_{k+1}^{(2,3)}$ denotes the set of all degrees of freedom of $\mathcal{S}_{k+1}(\Omega, \mathcal{T})$ that lie on $\gamma_{2,3}$.

3.2 The three-dimensional case

Now let $\Omega \subset \mathbb{R}^3$ (i.e. $d = 3$). The situation in 3D is vastly different to that of 2D. Instead of stream functions we need *vector potentials* of the functions $\vec{v} \in \mathring{\mathcal{V}}$ to construct a basis for $\mathring{\mathcal{V}}$, and in general these vector potentials are not in $(H^1(\Omega))^3$. Another fundamental difference is, that the kernel of the 2D curl consists only of the constant functions on Ω and can therefore be eliminated by essential boundary conditions, while the kernel of the 3D curl is much larger, making it necessary to eliminate some degrees of freedom from the vector potential space in a not so obvious way. (In the context of computational electromagnetics this is called *gauging*).

First, in Section 3.2.1, we will introduce Nédélec's edge elements which are conforming in the space $H(\vec{\text{curl}}, \Omega)$. These will turn out to be suitable vector potentials for the functions $\vec{v} \in \mathring{\mathcal{V}}$. Then, in Section 3.2.2, we will show, for $k = 0$, how these vector potentials can be used to construct a basis for $\mathring{\mathcal{V}}$ again. We will need fundamental results from Graph Theory and Algebraic Topology to extract a linear independent set of degrees of freedom from the basis of the vector potential space. Finally, in Section 3.2.4, we will briefly discuss the situation for higher order elements, i.e. $k > 0$.

We will restrict to simply connected domains without cavities. So, for the remainder of this section, let Ω be simply connected with a connected boundary Γ.

3.2.1 Vector potentials – Nédélec's edge elements in $H(\vec{\text{curl}}, \Omega)$

To construct the vector potentials for $\mathring{\mathcal{V}}$, let us first look at the continuous problem again. As in the 2D case, let $\Gamma_N = \emptyset$ and recall (3.5) that

$$\mathcal{Z} := \{\vec{v} \in H_{0,N}(\text{div}, \Omega) : b(\vec{v}, w) = 0 \text{ for all } w \in L_2(\Omega)\}.$$

The following fundamental result is taken from Girault & Raviart [44].

Proposition 3.19. *Let $\Gamma_N = \emptyset$. A function $\vec{v} \in (L_2(\Omega))^3$ is in $\mathcal{Z}$, if and only if there exists a vector potential $\vec{\Phi} \in (H^1(\Omega))^3$ such that:*

$$\vec{v} = \vec{\operatorname{curl}}\, \vec{\Phi} := \left(\frac{\partial \Phi_3}{\partial x_2} - \frac{\partial \Phi_2}{\partial x_3} \,, \, \frac{\partial \Phi_1}{\partial x_3} - \frac{\partial \Phi_3}{\partial x_1} \,, \, \frac{\partial \Phi_2}{\partial x_1} - \frac{\partial \Phi_1}{\partial x_2} \right)^T .$$

Proof. See Girault & Raviart [44, Theorem I.3.4]. □

Remark 3.20. Note that, formally the 3D curl of a function $\vec{\Phi} \in (H^1(\Omega))^3$ has to be understood in the sense of distributions again, i.e.

$$\int_\Omega \vec{\operatorname{curl}}\, \vec{\Phi} \cdot \vec{\xi}\, d\vec{x} = \int_\Omega \vec{\Phi} \cdot \vec{\operatorname{curl}}\, \vec{\xi}\, d\vec{x}, \qquad \text{for all } \vec{\xi} \in (\mathcal{D}(\Omega))^3.$$

However, if $\Gamma_N \neq \emptyset$, these vector potentials are in general not in $(H^1(\Omega))^3$. Let us verify this statement in the special case, when $\Gamma = \Gamma_N$. We need to introduce the following functional space

$$H(\vec{\operatorname{curl}}, \Omega) := \{\vec{\Phi} \in (L_2(\Omega))^3 : \vec{\operatorname{curl}}\, \vec{\Phi} \in (L_2(\Omega))^3\}. \tag{3.46}$$

By analogy to Lemma 2.1, it is possible to define a tangential[2] trace $\vec{\Phi} \times \vec{\nu}|_\Gamma$ of each function $\vec{\Phi} \in H(\vec{\operatorname{curl}}, \Omega)$ on Γ, as follows:

Lemma 3.21. *For $\vec{\Phi} \in H(\vec{\operatorname{curl}}, \Omega)$, we can define $\vec{\Phi} \times \vec{\nu}|_\Gamma \in (H^{-1/2}(\Gamma))^3$ by the following Green's formula*

$$\left\langle \vec{\Phi} \times \vec{\nu}, \vec{\xi} \right\rangle_\Gamma = \int_\Omega \vec{\operatorname{curl}}\, \vec{\Phi} \cdot \vec{\xi}\, d\vec{x} + \int_\Omega \vec{\Phi} \cdot \vec{\operatorname{curl}}\, \vec{\xi}\, d\vec{x}, \qquad \text{for all } \vec{\xi} \in (H^1(\Omega))^3, \tag{3.47}$$

where the bracket $\langle \cdot, \cdot \rangle_\Gamma$ denotes the duality between $(H^{-1/2}(\Gamma))^3$ and $(H^{1/2}(\Gamma))^3$.

Proof. See Girault & Raviart [44, Theorem I.2.11]. □

We can use this definition of $\vec{\Phi} \times \vec{\nu}|_\Gamma$ to introduce the subspace

$$H_0(\vec{\operatorname{curl}}, \Omega) := \{\vec{\Phi} \in H(\vec{\operatorname{curl}}, \Omega) : \left\langle \vec{\Phi} \times \vec{\nu}, \vec{\xi} \right\rangle_\Gamma = 0 \ \text{ for all } \vec{\xi} \in (H^1(\Omega))^3\} \tag{3.48}$$

of functions $\vec{\Phi} \in H(\vec{\operatorname{curl}}, \Omega)$ whose tangential traces vanish on Γ.

For $\Gamma = \Gamma_N$, we can only expect the vector potential to be in $H_0(\vec{\operatorname{curl}}, \Omega)$, as the following proposition shows.

Proposition 3.22. *Let $\Gamma = \Gamma_N$. For each $\vec{v} \in \mathcal{Z}$, there exists a vector potential $\vec{\Phi} \in H_0(\vec{\operatorname{curl}}, \Omega)$ such that:*

$$\vec{v} = \vec{\operatorname{curl}}\, \vec{\Phi}.$$

[2] The tangential component of a vector $\vec{\Phi}$ on the boundary Γ is defined by $\vec{\Phi}_\tau := \vec{\Phi} - (\vec{\Phi} \cdot \vec{\nu})\vec{\nu} = (\vec{\nu} \times \vec{\Phi}) \times \vec{\nu}$. Since $|\vec{\Phi}_\tau| = |\vec{\nu} \times \vec{\Phi}|$, the vector $\vec{\Phi}$ has vanishing tangential component if and only if $\vec{\Phi} \times \vec{\nu} = \vec{0}$. With an abuse of terminology, we will refer to $\vec{\Phi} \times \vec{\nu}$ as the tangential component.

Furthermore, if

$$\int_\Omega \operatorname{div}\vec{\Phi}\, p\, d\vec{x} = 0, \qquad \text{for all } p \in L_2(\Omega) \tag{3.49}$$

then $\vec{\Phi}$ is unique.

Proof. See Girault & Raviart [44, Theorem I.3.6]. $\square$

Remark 3.23. Only if, in addition, Ω is convex in Proposition 3.22, will $\vec{\Phi} \in (H^1(\Omega))^3$ (see Girault [43, Theorem 2.2]).

Condition (3.49) in Proposition 3.22 is called the *Coulomb gauge*, and it is essential for the uniqueness of the vector potential. There are other possibilities to choose the *gauge condition*, and in the next section we will use graph theoretical ideas to construct such a condition for discrete vector potentials. The reason, why we need this condition, is the large kernel of the 3D curl, as illustrated in the following proposition.

Proposition 3.24. *A function $\vec{\Phi} \in H(\vec{\operatorname{curl}}, \Omega)$ satisfies*

$$\int_\Omega \vec{\operatorname{curl}}\, \vec{\Phi} \cdot \vec{\xi} = 0, \qquad \text{for all } \vec{\xi} \in (L_2(\Omega))^3,$$

if and only if there exists a unique function $q \in H^1(\Omega)\backslash\mathbb{R}$ such that

$$\vec{\Phi} = \vec{\nabla} q.$$

Proof. See Girault & Raviart [44, Theorem I.2.9]. $\square$

We conclude the discussion of the continuous problem and come back to the discrete problem and the space $\mathring{\mathcal{V}}$. It follows from Lemma 2.8(a) that $\mathring{\mathcal{V}} \subset \mathcal{Z}$, and so Proposition 3.22 motivates us to seek the divergence-free Raviart-Thomas-Nédélec elements as the 3D curls of suitable finite elements in $H(\vec{\operatorname{curl}}, \Omega)$.

Let $\mathcal{T}$ be a simplicial triangulation of Ω as defined in Definition 2.10, and recall that by $\mathcal{E} = \mathcal{E}_I \cup \mathcal{E}_D \cup \mathcal{E}_N$ we denoted the union of all sets of edges of the the elements $T \in \mathcal{T}$, partitioned into the subsets $\mathcal{E}_I$, $\mathcal{E}_D$ and $\mathcal{E}_N$ containing the edges which lie in Ω, Γ_D and Γ_N, respectively. Furthermore, recall that with each of these edges $E \in \mathcal{E}$ we associated a unit tangent vector $\vec{\tau}_E \in R^+$, and since we assume Γ_N to be closed, all the edges connecting Neumann and Dirichlet boundaries must lie in $\mathcal{E}_N$. Analogously, we can write $\mathcal{F} = \mathcal{F}_I \cup \mathcal{F}_D \cup \mathcal{F}_N$ and $\mathcal{N} = \mathcal{N}_I \cup \mathcal{N}_D \cup \mathcal{N}_N$ to denote the sets of all faces (assumed to be open triangles) and nodes of $\mathcal{T}$.

For each $T \in \mathcal{T}$, we denote by $P_k(T)$ the space of polynomials of degree $\leq k$ on T (cf. (2.38)), and we will also need

$$\tilde{P}_{k+1}(T) := \quad \text{the space of } \textit{homogeneous} \text{ polynomials of degree } \leq k+1, \tag{3.50}$$

where *homogeneous* means that $p(\vec{0}) = 0$, for all $p \in \tilde{P}_{k+1}(T)$.

We can now define the Nédélec elements (or *edge elements*). These elements have been introduced by Nédélec in [73]. We also refer to Girault & Raviart [44, Section III.5.3] for details. Let $T \in \mathcal{T}$. For each $k \geq 0$, we consider the following subspace of $(P_{k+1}(T))^3$:

$$ND_{k+1}(T) := \{\vec{p}_k + \vec{q}_{k+1} \ : \ \vec{p}_k \in (P_k(T))^3 \text{ and } \vec{q}_{k+1} \in Q_{k+1}(T)\}, \tag{3.51}$$

where

$$Q_{k+1}(T) := \{\vec{q}_{k+1} \in (\tilde{P}_{k+1}(T))^3 \ : \ \vec{q}_{k+1}(\vec{x}) \cdot \vec{x} = 0 \text{ for all } \vec{x} \in T\}.$$

It is shown in [73, Lemma 4] that

$$\dim ND_{k+1}(T) = \frac{(k+1)(k+3)(k+4)}{2}. \tag{3.52}$$

Furthermore, Nédélec shows, see [73, Theorem 1], that any function $\vec{\Phi} \in ND_{k+1}(T)$ is uniquely defined by the following degrees of freedom (see also Figure 3.6):

Definition 3.25. (Degrees of freedom for $ND_{k+1}(T)$)

- The moments of order up to k of $\vec{\Phi} \cdot \vec{\tau}_E$ on each edge E of T, i.e.

$$\int_E \vec{\Phi} \cdot \vec{\tau}_E \, p_k \, ds, \qquad p_k \in P_k(E).$$

- The moments of order up to $k-1$ of $\vec{\Phi} \times \vec{\nu}_F$ on each face F of T, for $k > 0$, i.e.

$$\int_F [\vec{\Phi} \times \vec{\nu}_F]_\tau \cdot \vec{p}_{k-1} \, ds \, dt, \qquad \vec{p}_{k-1} \in (P_{k-1}(F))^2,$$

  where $[\vec{\Phi} \times \vec{\nu}_F]_\tau \in \mathbb{R}^2$ denotes the tangential component of $\vec{\Phi} \times \vec{\nu}_F$ on F parameterised by s and t. Note that by definition the normal component $(\vec{\Phi} \times \vec{\nu}_F) \cdot \vec{\nu}_F = 0$.

- The moments of order up to $k-2$ of $\vec{\Phi}$ on T, for $k > 1$, i.e.

$$\int_T \vec{\Phi} \cdot \vec{p}_{k-2} \, d\vec{x}, \qquad \vec{p}_{k-2} \in (P_{k-2}(T))^3.$$

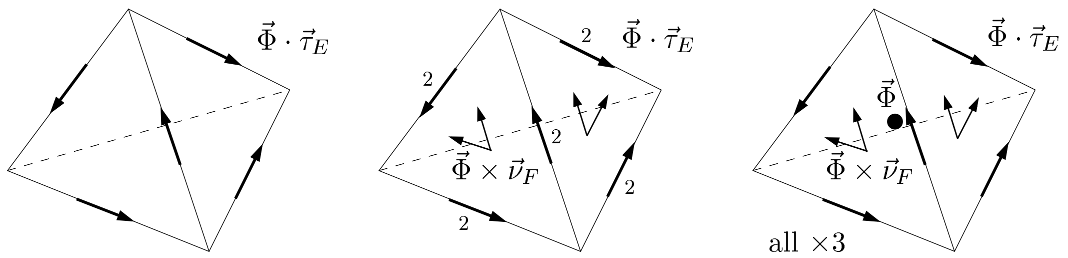


Figure 3.6: Degrees of Freedom for $ND_{k+1}(T)$, for $k = 0, 1, 2$.

In complete analogy to Proposition 2.7 (using the Green's formula (3.47) instead of (2.7)), we can show the following proposition (see Nédélec [73, Lemma 6]).

Proposition 3.26. *Let* $(H_0^1(\Omega))^3 := \{\vec{\xi} \in (H^1(\Omega))^3 : \vec{\xi}|_\Gamma = \vec{0}\}$. *A function* $\vec{\Phi} \in (L_2(\Omega))^3$ *is in* $H(\vec{\text{curl}}, \Omega)$, *if and only if the following two conditions hold true:*

$$\vec{\Phi}|_T \in H(\vec{\text{curl}}, T) \qquad \text{for all } T \in \mathcal{T}, \tag{3.53}$$

$$\sum_{T\in\mathcal{T}} \left\langle \vec{\Phi} \times \vec{\nu}_T, \vec{\xi} \right\rangle_{\partial T} = 0 \qquad \text{for all } \vec{\xi} \in (H_0^1(\Omega))^3, \tag{3.54}$$

where $\vec{\nu}_T$ *denotes the outward unit normal from* T *on* ∂T.[3]

Proposition 3.26 states that the tangential trace $\vec{\Phi}\times\vec{\nu}_F$ of a function $\vec{\Phi} \in H(\vec{\text{curl}}, \Omega)$ is continuous across each face $F \in \mathcal{F}_I$. Therefore, the choice of degrees of freedom on each element $T \in \mathcal{T}$ in Definition 3.25 above, enables us to build a finite dimensional subspace of $H(\vec{\text{curl}}, \Omega)$ from the polynomial spaces $ND_{k+1}(T)$ (We refer again to [73] for the details.) We define

$$\mathcal{ND}_{k+1}(\Omega, \mathcal{T}) := \{\vec{\Phi} \in H(\vec{\text{curl}}, \Omega) : \vec{\Phi}|_T \in ND_{k+1}(T) \quad \text{for all } T \in \mathcal{T}\}. \tag{3.55}$$

Taking into account the continuity of $\vec{\Phi} \times \vec{\nu}_F$ across any interface F of two elements, we get

$$\dim \mathcal{ND}_{k+1}(\Omega, \mathcal{T}) = (k+1)\left(\#\mathcal{E} + k\,\#\mathcal{F} + \frac{k(k-1)}{2}\,\#\mathcal{T}\right) \tag{3.56}$$

Example 3.27. *The lowest order case:* $k = 0$

All homogeneous polynomials $\vec{q}$ of order one that satisfy $\vec{q}(\vec{x})\cdot\vec{x} = 0$, i.e. all elements $\vec{q} \in Q_1(T)$, must necessarily be of the form

$$\vec{q}(\vec{x}) = \vec{c} \times \vec{x}, \qquad \text{for some } \vec{c} \in \mathbb{R}^3.$$

Thus

$$ND_1(T) = \left\{\vec{a} + \vec{c} \times \vec{x} \;:\; \vec{a}, \vec{c} \in \mathbb{R}^3\right\}. \tag{3.57}$$

The canonical basis of $\mathcal{ND}_1(\Omega, \mathcal{T})$ is given by the functions $\{\vec{\Phi}_E \in \mathcal{ND}_1(\Omega, \mathcal{T}) : E \in \mathcal{E}\}$ which are required to have the property

$$\int_{E'} \vec{\Phi}_E \cdot \vec{\tau}_{E'}\, ds = \delta_{E,E'}, \qquad \text{for all } E' \in \mathcal{E}. \tag{3.58}$$

This choice of basis functions accounts for the widely used term *edge elements.* □

As in the 2D-case, we will use the basis functions of the space $\mathcal{ND}_1(\Omega, \mathcal{T})$ in the next section to construct a basis for $\mathring{\mathcal{V}}$ for the lowest order case $k = 0$. As a motivating idea

[3] Note that this result is not restricted to simplicial triangulations, but also holds true for more general partitionings $\mathcal{T}$, as defined at the beginning of section 2.1.3.

we would like to note that this construction basically uses a part of the *de Rham diagram* (see for example Hiptmair [57, Theorem 2.36], Hiptmair & Hoppe [59, Theorem 1] or Bossavit [17, Chapter 5]).

3.2.2 The space $\mathring{\mathcal{V}}$ – a graph theoretical approach for $k = 0$

Let $k = 0$, and let

$$\{\vec{\Phi}_E : E \in \mathcal{E}\} \tag{3.59}$$

be the canonical basis of the Nédélec space $\mathcal{ND}_1(\Omega, \mathcal{T})$, as defined in (3.58). The basis for $\mathring{\mathcal{V}}$ will now be constructed from the fundamental functions $\vec{\Psi}_E$ defined by:

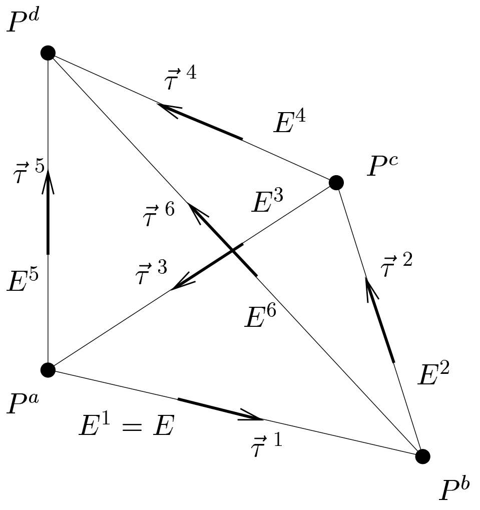
$$\vec{\Psi}_E := \vec{\mathrm{curl}}\, \vec{\Phi}_E, \qquad E \in \mathcal{E}. \tag{3.60}$$

This means that $\vec{\Phi}_E$ is the *vector potential* of $\vec{\Psi}_E$. The functions (3.60) clearly satisfy $\mathrm{div}\vec{\Psi}_E = 0$ on each tetrahedron of the mesh, and a subset of them lie in $\mathring{\mathcal{V}}$ as the following proposition shows.

Proposition 3.28. *For each* $E \in \mathcal{E}_I \cup \mathcal{E}_D$, $\quad \vec{\Psi}_E \in \mathring{\mathcal{V}}$.

Proof. This fact is already observed in the original paper of Nédélec [73], and proved at various points in the literature for pure Dirichlet or Neumann boundary conditions (e.g. [57, Theorem 2.36]). We give here a general proof.

Consider a general edge $E \in \mathcal{E}$. Conditions (3.57) and (3.58) clearly imply that $\mathrm{supp}\, \vec{\Psi}_E$ consists only of the tetrahedra touching edge E. A typical such tetrahedron T with edges $E := E^1, E^2, \ldots, E^6$, and nodes $P^a, \ldots, P^d$ is depicted below,



with $\vec{\tau}^{\,\alpha}$, $\alpha = 1, \ldots, 6$ denoting unit tangent vectors in the directions shown, and $\vec{r}^{\,\beta}$ denoting the position vector of P^β, $\beta = a, b, c, d$. The faces are denoted by $F^a, \ldots, F^d$, where F^β is opposite P^β, $\beta = a, b, c, d$, and the unit outward normal on each face is denoted by $\vec{\nu}^{\,\beta}$.

Note first that, using (3.57), for all $\vec{x} \in T$

$$\vec{\Psi}_E(\vec{x}) = \vec{\mathrm{curl}}\, \vec{\Phi}_E(\vec{x}) = \vec{\nabla} \times (\vec{c} \times \vec{x}) = 2\vec{c}. \tag{3.61}$$

In the next paragraph we will therefore consider the restriction of $\vec{\Phi}_E$ to T and show that

$$\vec{\Psi}_E(\vec{x}) = \frac{|E^4|}{3|T|}\vec{\tau}^{\,4}, \qquad \text{for all } \vec{x} \in T, \tag{3.62}$$

which is easily seen to be of the form (2.54) (in fact with $\gamma = 0$).

On T, $\vec{\Phi}_E$ can be written in form (3.57) whence it is easy to show that $\vec{\Phi}_E(\vec{x}) \cdot \vec{\tau}^{\,\alpha}$ is constant on edge E^α, $\alpha = 1, \ldots, 6$. In fact, take for example edge E^3, and let $\vec{x} \in E^3$. We can write $\vec{x} = \vec{r}^{\,c} + \lambda\vec{\tau}^{\,3}$, for some $\lambda \in \mathbb{R}$, and therefore using (3.57)

$$\vec{\Phi}_E(\vec{x}) \cdot \vec{\tau}^{\,3} = \left(\vec{a} + \vec{c} \times (\vec{r}^{\,c} + \lambda\vec{\tau}^{\,3})\right) \cdot \vec{\tau}^{\,3} = (\vec{a} + \vec{c} \times \vec{r}^{\,c}) \cdot \vec{\tau}^{\,3}$$

independent of λ (observe that the choice of the position vector $\vec{r}^{\,c}$ rather than $\vec{r}^{\,a}$ is arbitrary). Therefore writing (3.58) for each edge E' of $\overline{T}$ in turn, we have

$$\begin{array}{llcl@{\qquad}llcl}
(i) & |E^1|\vec{\tau}^{\,1} \cdot \left(\vec{a} + \vec{c} \times \vec{r}^{\,b}\right) & = & 1 & (iv) & |E^4|\vec{\tau}^{\,4} \cdot \left(\vec{a} + \vec{c} \times \vec{r}^{\,d}\right) & = & 0 \\
(ii) & |E^2|\vec{\tau}^{\,2} \cdot (\vec{a} + \vec{c} \times \vec{r}^{\,c}) & = & 0 & (v) & |E^5|\vec{\tau}^{\,5} \cdot \left(\vec{a} + \vec{c} \times \vec{r}^{\,d}\right) & = & 0 \\
(iii) & |E^3|\vec{\tau}^{\,3} \cdot (\vec{a} + \vec{c} \times \vec{r}^{\,c}) & = & 0 & (vi) & |E^6|\vec{\tau}^{\,6} \cdot \left(\vec{a} + \vec{c} \times \vec{r}^{\,b}\right) & = & 0
\end{array}$$

By taking linear combinations of these equations we can eliminate $\vec{a}$ and obtain the following system for $\vec{c}$:

$$\begin{array}{lrcl}
-(i) - (ii) - (iii) & |E^1|\vec{\tau}^{\,1} \cdot \left(\vec{c} \times |E^2|\vec{\tau}^{\,2}\right) & = & -1 \\
-(i) + (v) - (vi) & |E^5|\vec{\tau}^{\,5} \cdot \left(\vec{c} \times |E^6|\vec{\tau}^{\,6}\right) & = & -1 \\
-(iii) + (iv) - (v) & |E^3|\vec{\tau}^{\,3} \cdot \left(\vec{c} \times |E^4|\vec{\tau}^{\,4}\right) & = & 0
\end{array}$$

The unique solution of this system is

$$\vec{c} = \frac{|E^4|}{6|T|}\vec{\tau}^{\,4}.$$

Substituting into (3.61) we have established (3.62).

Furthermore note that

$$|E^4|\vec{\tau}^{\,4} \cdot |F^\beta|\vec{\nu}^{\,\beta} = \left\{ \begin{array}{rl} 3|T| & \text{for } \beta = c \\ -3|T| & \text{for } \beta = d \\ 0 & \text{otherwise} \end{array} \right. ,$$

and therefore

$$\left.\begin{array}{rcl}\vec{\Psi}_E(\vec{x})\cdot\vec{\nu}^{\,a} &=& \dfrac{|E^4|}{3|T|}\,\vec{\tau}^{\,4}\cdot\vec{\nu}^{\,a}=0\\ \vec{\Psi}_E(\vec{x})\cdot\vec{\nu}^{\,b} &=& \dfrac{|E^4|}{3|T|}\,\vec{\tau}^{\,4}\cdot\vec{\nu}^{\,b}=0\\ \vec{\Psi}_E(\vec{x})\cdot\vec{\nu}^{\,c} &=& \dfrac{|E^4|}{3|T|}\,\vec{\tau}^{\,4}\cdot\vec{\nu}^{\,c}=\dfrac{|E^4|\vec{\tau}^{\,4}\cdot|F^c|\vec{\nu}^{\,c}}{3|T||F^c|}=\dfrac{1}{|F^c|}\\ \vec{\Psi}_E(\vec{x})\cdot\vec{\nu}^{\,d} &=& \dfrac{|E^4|}{3|T|}\,\vec{\tau}^{\,4}\cdot\vec{\nu}^{\,d}=\dfrac{|E^4|\vec{\tau}^{\,4}\cdot|F^d|\vec{\nu}^{\,d}}{3|T||F^d|}=-\dfrac{1}{|F^d|}\end{array}\right\}\ \vec{x}\in T\ . \tag{3.63}$$

Now to obtain the result observe that, since $\mathrm{div}\vec{\Psi}_E = 0$ on each tetrahedron, it is sufficient to show that

$$\vec{\Psi}_E \in \mathcal{V}, \quad \text{for all } E \in \mathcal{E}_I \cup \mathcal{E}_D\ . \tag{3.64}$$

To show this we shall verify criterion (2.55).

Consider first $E \in \mathcal{E}_I$. Let $F \in \mathcal{F}_I$. If $F \not\subset \mathrm{supp}\,\vec{\Psi}_E$ we have trivially

$$\vec{\Psi}_E \cdot \vec{\nu}_F \quad \text{is continuous across } F. \tag{3.65}$$

Now take a general tetrahedron $T \subset \mathrm{supp}\,\vec{\Psi}_E$, as pictured above, and note $E^1 = E$. If $F = F^c$ or F^d, then performing the computation (3.63) in the other tetrahedron adjoining F and combining with (3.63) establishes (3.65). On the other hand, when $F = F^a$ or F^b, (3.65) also holds, since $\vec{\Psi}_E \cdot \vec{\nu}_F|_T = 0$ and since the other tetrahedron adjoining F lies outside $\mathrm{supp}\,\vec{\Psi}_E$. Altogether we have established that $\vec{\Psi}_E$ satisfies criterion (2.55)(i).

To establish (2.55)(ii), let $F \in \mathcal{F}_N$. If $F \not\subset \mathrm{supp}\,\vec{\Psi}_E$, then $\vec{\Psi}_E \cdot \vec{\nu}_F = 0$ trivially. If $F \subset \overline{T} \subset \mathrm{supp}\,\vec{\Psi}_E$, then (since $E \in \mathcal{E}_I$) with the local notation specified at the beginning of the proof, F has to be either F^a or F^b and again $\vec{\Psi}_E \cdot \vec{\nu}_F = 0$, proving (2.55)(ii).

Thus we have shown that $\vec{\Psi}_E \in \mathcal{V}$ for all $E \in \mathcal{E}_I$. Similar arguments establish that $\vec{\Psi}_E \in \mathcal{V}$ for all $E \in \mathcal{E}_D$, proving (3.64). □

Note that each $\vec{\Psi}_E$ can be expressed as a local linear combination of the basis functions $\vec{v}_F$ of $\mathcal{V}$ satisfying (2.75); in fact only those $\vec{v}_F$ corresponding to faces F that contain edge E appear in the expansion of $\vec{\Psi}_E$ (see Figure 3.7).

To find a basis for $\mathring{\mathcal{V}}$, let us first look at the pure Dirichlet case, $\Gamma_N = \emptyset$. Similar to the situation in 2D the functions introduced in Proposition 3.28 are sufficient to span $\mathring{\mathcal{V}}$, but because of the large kernel of the 3D curl these functions are not linearly independent. The following theorem identifies a linearly independent subset of the functions in Proposition 3.28 that constitutes a basis of $\mathring{\mathcal{V}}$. A similar statement for the pure Neumann case, $\Gamma_D = \emptyset$, has already been proved by Dubois [37].

The proof involves some fundamental notions and results from Graph Theory and Algebraic Topology (see Appendices B and C for a brief introduction). In particular

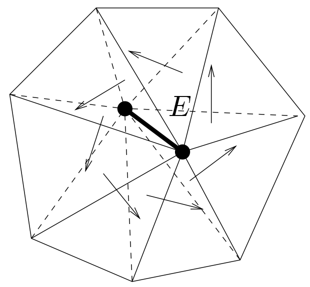


Figure 3.7: Divergence-free basis function $\vec{\Psi}_E$

we need the notion of *spanning tree* of a graph (see Theorem B.4). Let $\mathbf{G} := (\mathcal{N}, \mathcal{E})$ be the graph formed by the nodes and (orientated) edges of the triangulation $\mathcal{T}$.

Theorem 3.29. *Let $\Gamma_N = \emptyset$ and let $\mathcal{H} \subset \mathcal{E}$ be such that $\mathbf{H} := (\mathcal{N}, \mathcal{H})$ is a spanning tree of $\mathbf{G}$, then*

$$\{\vec{\Psi}_E : E \in \mathcal{E} \backslash \mathcal{H}\} \quad \textit{is a basis of} \quad \mathring{\mathcal{V}}. \tag{3.66}$$

Before proving Theorem 3.29, we will first prove two lemmas. Let $\mathcal{V}(\mathbf{G})$ denote the vector space over $\mathbb{Z}$ generated by the cycles of $\mathbf{G}$ as defined in Definition B.1(e). Furthermore, for each face $F \in \mathcal{F}$ let μ^F be the *elementary cycle* of $\mathbf{G}$ formed by the edges E of F. We fix the orientation of this cycle with respect to $\vec{\nu}_F$ by applying the right-hand rule. The associated vector $\boldsymbol{\mu}^F := \left[\mu_E^F\right]_{E \in \mathcal{E}} \in \mathcal{V}(\mathbf{G})$ is given by

$$\mu_E^F = \begin{cases} 1 & \textit{if } E \textit{ is an edge of } F \textit{ and } \vec{\tau}_E \textit{ is positively orientated wrt. } \vec{\nu}_F \\ -1 & \textit{if } E \textit{ is an edge of } F \textit{ and } \vec{\tau}_E \textit{ is negatively orientated wrt. } \vec{\nu}_F \\ 0 & \textit{otherwise.} \end{cases} \tag{3.67}$$

Lemma 3.30. *Let $\boldsymbol{\mu} := [\mu_E]_{E \in \mathcal{E}} \in \mathcal{V}(\mathbf{G})$. Then there exist $\{\alpha_F \in \mathbb{Z} : F \in \mathcal{F}\}$ such that*

$$\boldsymbol{\mu} := \sum_{F \in \mathcal{F}} \alpha_F \boldsymbol{\mu}^F. \tag{3.68}$$

Proof. Let K be the simplicial complex underlying our simplicial triangulation $\mathcal{T}$. In the notation of algebraic topology (see Appendix C) the vector $\boldsymbol{\mu} \in \mathcal{V}(\mathbf{G})$ can be identified with the vector of coefficients of a cycle μ of K (with orientation of its edges defined by the tangent vectors $\vec{\tau}_E$).

Since $|K| = \overline{\Omega}$ is simply connected, we know from Corollary C.4 that each cycle of K is a bounding cycle and can therefore be written as a linear combination of the boundaries of the orientated triangles of K. In particular, there exist $\{\tilde{\alpha}_F \in \mathbb{Z} : F \in \mathcal{F}\}$ such that

$$\mu = \sum_{F \in \mathcal{F}} \tilde{\alpha}_F \partial F. \tag{3.69}$$

The boundary ∂F of an orientated triangle F of K is a special cycle $\tilde{\mu}^F$ of K. As above it can therefore be identified with a vector $\tilde{\boldsymbol{\mu}}^F \in \mathcal{V}(\mathbf{G})$. Depending on the orientation of $\vec{\nu}_F$ we either have $\tilde{\boldsymbol{\mu}}^F = \boldsymbol{\mu}^F$ or $\tilde{\boldsymbol{\mu}}^F = -\boldsymbol{\mu}^F$ and we can write (3.69) in vector notation

$$\boldsymbol{\mu} = \sum_{F\in\mathcal{F}} \alpha_F \boldsymbol{\mu}^F \qquad \text{with} \quad \alpha_F = \begin{cases} \tilde{\alpha}_F & \text{if} \quad \tilde{\boldsymbol{\mu}}^F = \boldsymbol{\mu}^F, \\ -\tilde{\alpha}_F & \text{if} \quad \tilde{\boldsymbol{\mu}}^F = -\boldsymbol{\mu}^F. \end{cases}$$

□

Lemma 3.31. *Let $\boldsymbol{\mu} \in \mathcal{V}(\mathbf{G})$ and let $\{\alpha_F \in \mathbb{Z} : F \in \mathcal{F}\}$ be such that $\boldsymbol{\mu} := \sum_{F\in\mathcal{F}} \alpha_F \boldsymbol{\mu}^F$. Then*

$$\sum_{F\in\mathcal{F}} \alpha_F \int_F \vec{\Psi}_E \cdot \vec{\nu}_F \, ds = \mu_E, \qquad \text{for all } E \in \mathcal{E}. \tag{3.70}$$

Proof. Let $F \in \mathcal{F}$. Using (3.63) we get

$$\int_F \vec{\Psi}_E \cdot \vec{\nu}_F \, ds = \begin{cases} 1 & \text{if } E \subset \overline{F} \text{ and } \vec{\tau}_E \text{ positively orientated wrt. } \vec{\nu}_F \\ -1 & \text{if } E \subset \overline{F} \text{ and } \vec{\tau}_E \text{ negatively orientated wrt. } \vec{\nu}_F \\ 0 & \text{otherwise} \end{cases}$$

and therefore, recalling the definition (3.67), we have $\int_F \vec{\Psi}_E \cdot \vec{\nu}_F \, ds = \mu_E^F$. Multiplying this by α_F and summing over $F \in \mathcal{F}$ we obtain (3.70). □

We can now prove Theorem 3.29:

Proof of Theorem 3.29. Let us first check that the the number of basis functions in (3.66) coincides with $\mathring{n} = \dim \mathring{\mathcal{V}}$. Since Ω is simply connected without cavities, we can apply *Euler's Polyhedron Theorem*

$$(\#\mathcal{N} - \#\mathcal{E} + \#\mathcal{F} - \#\mathcal{T}) \; = \; 1 \tag{3.71}$$

to the triangulation $\mathcal{T}$ (or more exactly Cauchy's generalisation of Euler's result [22, 16]). Recall that $\mathbf{H}$ is a tree, and therefore $\#\mathcal{H} = \#\mathcal{N} - 1$ (cf. Theorem B.3(iii)). Using this fact together with (3.71) we get

$$\#(\mathcal{E}\backslash\mathcal{H}) = (\#\mathcal{E} - \#\mathcal{N} + 1) = (\#\mathcal{F} - \#\mathcal{T}). \tag{3.72}$$

Now recalling that $\mathring{n} = n_{\mathcal{V}} - n_{\mathcal{W}} = (\#\mathcal{F} - \#\mathcal{T})$, it follows from (3.72) that the number of functions in (3.66) is $\mathring{n}$, as required.

To establish linear independency of the functions in (3.66), suppose $\{\beta_{E'} : E' \in \mathcal{E}\backslash\mathcal{H}\}$ are scalars such that

$$\vec{0} = \sum_{E'\in\mathcal{E}\backslash\mathcal{H}} \beta_{E'} \vec{\Psi}_{E'}.$$

Now let $E \in \mathcal{E}\backslash\mathcal{H}$ and let $\boldsymbol{\mu}^E$ denote the vector associated to the unique cycle μ^E generated by taking edge E into the tree $\mathbf{H}$, which has the property that $\mu^E_{E'} := \delta_{E,E'}$, for all $E' \in \mathcal{E}\backslash\mathcal{H}$ (cf. Theorem B.6). Then using Lemma 3.30 we can find $\{\alpha_F \in \mathbb{Z} :$

$F \in \mathcal{F}\}$ such that $\boldsymbol{\mu}^E := \sum_{F\in\mathcal{F}} \alpha_F \boldsymbol{\mu}^F$, and so by Lemma 3.31

$$0 = \sum_{F\in\mathcal{F}} \alpha_F \int_F \Bigg(\sum_{E'\in\mathcal{E}\backslash\mathcal{H}} \beta_{E'} \vec{\Psi}_{E'} \Bigg) \cdot \vec{\nu}_F \, ds = \sum_{E'\in\mathcal{E}\backslash\mathcal{H}} \beta_{E'} \mu^E_{E'} = \beta_E \,,$$

which establishes the linear independency of the functions in (3.66). $\square$

Now let us look at mixed boundary conditions, $\Gamma_N \neq \emptyset$. In the following corollary we will see that the results of Theorem 3.29 extend to this case, provided each component of Γ_N is simply connected, or equivalently provided Γ_D is connected. Our proof of this result makes use of the methods of Hecht [55] developed for the non-conforming P1-P0 elements for the approximation of solenoidal vector fields in $(H^1(\Omega))^3$.

So, let s_N denote the number of connected components in Γ_N, and write

$$\Gamma_N = \Gamma_N^1 \cup \Gamma_N^2 \cup \ldots \cup \Gamma_N^{s_N} \,, \qquad \Gamma_N^\ell \cap \Gamma_N^{\ell'} = \emptyset \quad \text{for all } 1 \le \ell \neq \ell' \le s_N.$$

For $\ell = 1, \ldots, s_N$, let $\mathcal{N}_N^\ell \subset \mathcal{N}$, $\mathcal{E}_N^\ell \subset \mathcal{E}$, and $\mathcal{F}_N^\ell \subset \mathcal{F}$ denote the set of mesh nodes, edges, and faces on Γ_N^ℓ, respectively.

Corollary 3.32. *Suppose $s_N \neq 0$, and suppose that Γ_N^ℓ is simply connected for each $\ell = 1, \ldots, s_N$. Let $\mathcal{H} \subset \mathcal{E}$ such that $\mathbf{H} = (\mathcal{N}, \mathcal{H})$ is a spanning tree of $\mathbf{G}$, and such that for each $\ell = 1, \ldots, s_N$, the restriction $\mathbf{H}_N^\ell := (\mathcal{N}_N^\ell, \mathcal{H} \cap \mathcal{E}_N^\ell)$ of $\mathbf{H}$ to Γ_N^ℓ, is also a tree. Then*

$$\{\vec{\Psi}_E : E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}\} \quad \textit{is a basis of} \quad \mathring{\mathcal{V}}. \tag{3.73}$$

Proof. Since $(\mathcal{E}_I \cup \mathcal{E}_D) \subset \mathcal{E}$, we also have $(\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H} \subset \mathcal{E}\backslash\mathcal{H}$, and therefore following the proof of Theorem 3.29 the functions $\vec{\Psi}_E$ in (3.73) are linearly independent.

We only have to check that the number of basis functions in (3.73) coincides with $\mathring{n} = \dim \mathring{\mathcal{V}}$. To do this, we need *Euler's Polyhedron Theorem* (3.71) and the *Euler characteristic* (3.22) again. Consider a typical Neumann boundary segment Γ_N^ℓ. Since Γ_N^ℓ is simply connected, using Lemma 3.9, we have

$$(\#\mathcal{N}_N^\ell - \#\mathcal{E}_N^\ell + \#\mathcal{F}_N^\ell) = 1. \tag{3.74}$$

Now observe that $\mathbf{H}_N^\ell$ is a tree, and therefore (again by virtue of Theorem B.3(iii)) $\#(\mathcal{H} \cap \mathcal{E}_N^\ell) = (\#\mathcal{N}_N^\ell - 1)$. Using (3.74) and summing over $\ell = 1, \ldots, s_N$, we obtain

$$\#(\mathcal{H} \cap \mathcal{E}_N) = \sum_{\ell=1}^{s_N} (\#\mathcal{N}_N^\ell - 1) = \sum_{\ell=1}^{s_N} (\#\mathcal{E}_N^\ell - \#\mathcal{F}_N^\ell) = (\#\mathcal{E}_N - \#\mathcal{F}_N). \tag{3.75}$$

Since the sets $\mathcal{E}_I$, $\mathcal{E}_D$ and $\mathcal{E}_N$ partition $\mathcal{E}$, we also have

$$(\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H} = (\mathcal{E}\backslash\mathcal{E}_N)\backslash\mathcal{H} = \mathcal{E}\backslash(\mathcal{E}_N \cup \mathcal{H})$$

and therefore the number of functions in (3.73) is $\#\mathcal{E} - (\#\mathcal{E}_N + \#\mathcal{H} - \#(\mathcal{H} \cap \mathcal{E}_N))$.

Combining this with (3.75), and using the fact that $\mathbf{H}$ is a tree (and therefore $\#\mathcal{H} = (\#\mathcal{N} - 1)$), we finally get

$$\#((\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}) = (\#\mathcal{E} - \#\mathcal{N} + 1) - \#\mathcal{F}_N = (\#\mathcal{F} - \#\mathcal{T}) - \#\mathcal{F}_N \ , \tag{3.76}$$

where in the last step we have used Euler's Polyhedron Theorem (3.71). Now recalling that $\mathring{n} = n_{\mathcal{V}} - n_{\mathcal{W}}$ (and from Example 2.16 we have $n_{\mathcal{V}} = (\#\mathcal{F}_I + \#\mathcal{F}_D) = (\#\mathcal{F} - \#\mathcal{F}_N)$ and $n_{\mathcal{W}} = \#\mathcal{T}$), it follows from (3.76) that the number of functions in (3.73) is $\mathring{n}$, as required. □

The general case, when Γ_D is not connected, involves the introduction of a small number of additional basis functions. To simplify the presentation, let us assume that Γ_D has two connected components that are disjoint and simply connected, i.e.

$$\Gamma_D = \Gamma_D^1 \cup \Gamma_D^2, \qquad \text{such that} \ \ \overline{\Gamma}_D^1 \cap \overline{\Gamma}_D^2 = \emptyset. \tag{3.77}$$

Note that this implies that Γ_N is connected, i.e. $s_N = 1$. We will come back to the general case in Remark 3.35.

Now let $E^{in} \in \mathcal{E}_N$ be an edge on the interface between Γ_D^1 and Γ_N, and let $E^{out} \in \mathcal{E}_N$ be an edge on the interface between Γ_D^2 and Γ_N. Since Γ_N is connected there exists a sequence of distinct edges

$$\mathcal{E}^{1,2} := \{E_1, \ldots, E_m : m \geq 3\} \tag{3.78}$$

(see also Figure 3.8), such that

$$\left.\begin{array}{rl} \text{(i)} & E_1 = E^{in} \text{ and } E_m = E^{out}, \\[1ex] \text{(ii)} & E_i \in \mathcal{E}_N, \text{ for } i = 2, \ldots, m-1, \\[1ex] \text{(iii)} & \text{there exists } F \in \mathcal{F}_N, \text{ such that } E_i, E_{i+1} \subset \overline{F}, \text{ for each } i = 1, \ldots, m-1. \end{array}\right\} \tag{3.79}$$

Proposition 3.33.

$$\sum_{E \in \mathcal{E}^{1,2}} \vec{\Psi}_E \in \mathring{\mathcal{V}}.$$

Proof. The proof is similar to the proof of Proposition 3.16. As in the proof to Proposition 3.16 or 3.28, it is sufficient to show that $\left(\sum_{E \in \mathcal{E}^{1,2}} \vec{\Psi}_E\right) \in \mathcal{V}$. Criterion (2.55)(i) is a mere consequence of (3.63) again. To establish (2.55)(ii) let $F \in \mathcal{F}_N$. If there is no $E \in \mathcal{E}^{1,2}$ such that $E \subset \overline{F}$, then we have $\left(\sum_{E \in \mathcal{E}^{1,2}} \vec{\Psi}_E\right) \cdot \vec{\nu}_F = 0$ by definition. Otherwise, it follows from (3.79)(iii) that there have to be two edges $E_i, E_{i+1} \in \mathcal{E}^{1,2}$

such that $E_i, E_{i+1} \subset \overline{F}$, and using (3.63) we get

$$\Big|\sum_{E\in\mathcal{E}^{1,2}} \vec{\Psi}_E \cdot \vec{\nu}_F\Big| = \Big|\vec{\Psi}_{E_i} \cdot \vec{\nu}_F + \vec{\Psi}_{E_{i+1}} \cdot \vec{\nu}_F\Big| = \left|\frac{1}{|F|} - \frac{1}{|F|}\right| = 0\,.$$

Thus $\left(\sum_{E\in\mathcal{E}^{1,2}} \vec{\Psi}_E\right) \cdot \vec{\nu}_F = 0$ for all $F \in \mathcal{F}_N$, and the proof is complete. □

In contrast to the functions in Proposition 3.28, the function introduced in Proposition 3.33 is a non-local linear combination of the functions $\vec{v}_F$; however, the non-locality of $\sum_{E\in\mathcal{E}^{1,2}} \vec{\Psi}_E$ is confined to the vicinity of the edges $E \in \mathcal{E}^{1,2}$ (see Figure 3.8). The

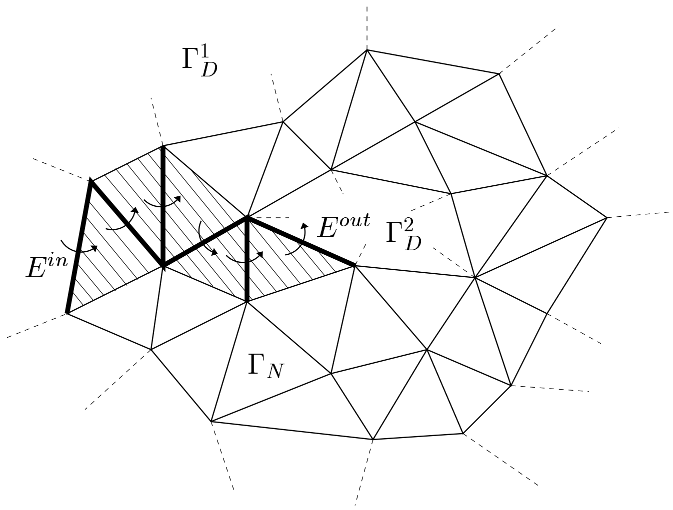


Figure 3.8: Divergence-free basis function $\sum\limits_{E\in\mathcal{E}^{1,2}} \vec{\Psi}_E$

number of such tetrahedra is typically $O((\#\mathcal{T})^{1/3})$, so they are only modestly non-local. Compare this to the 2D situation in Section 3.1.4.

In Theorem 3.34 below, we prove that by adding the additional function found in Proposition 3.33 to the functions in (3.73) results in a basis for $\mathring{\mathcal{V}}$, in the special case when Γ_N and Γ_D are given as above.

Theorem 3.34. *Suppose Γ_N is connected and $\Gamma_D = \Gamma_D^1 \cup \Gamma_D^2$, as defined in (3.77). There exists a set $\mathcal{H} \subset \mathcal{E}\backslash\mathcal{E}^{1,2}$ such that $\mathbf{H} = (\mathcal{N}, \mathcal{H})$ is a (spanning) tree of $\mathbf{G} = (\mathcal{N}, \mathcal{E})$, and such that the restriction $\mathbf{H}_N := (\mathcal{N}_N, \mathcal{H} \cap \mathcal{E}_N)$ of $\mathbf{H}$ to Γ_N, is also a tree. Furthermore the functions*

$$\{\vec{\Psi}_E : E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}\} \;\cup\; \{\sum_{E\in\mathcal{E}^{1,2}} \vec{\Psi}_E\} \tag{3.80}$$

are a basis for $\mathring{\mathcal{V}}$.

Proof. First of all, let us prove the existence of such a set $\mathcal{H} \subset \mathcal{E}$ with $\mathcal{H} \cap \mathcal{E}^{1,2} = \emptyset$. Let $\mathbf{G}_N = (\mathcal{N}_N, \mathcal{E}_N)$ be the (connected) restriction of the graph $\mathbf{G}$ to Γ_N. We need to prove the existence of a spanning tree $\mathbf{H}_N$ of $\mathbf{G}_N$ that does not use any of the edges $E \in \mathcal{E}^{1,2}$. Now, since Γ_N is multiply connected, the partial graph

$$\tilde{\mathbf{G}}_N = (\mathcal{N}_N, \mathcal{E}_N \backslash \mathcal{E}^{1,2}) \quad \text{is still connected} \tag{3.81}$$

(see also Figure 3.8) and the existence of a spanning tree $\mathbf{H}_N = (\mathcal{N}_N, \mathcal{H}_N)$ that does not use any of the edges $E \in \mathcal{E}^{1,2}$ is guaranteed by virtue of Theorem B.4. This implies that the graph $(\mathcal{N}, \mathcal{H}_N)$ is without cycles, and therefore there exists an $\mathcal{H} \supset \mathcal{H}_N$ such that $\mathbf{H} = (\mathcal{N}, \mathcal{H})$ is a spanning tree of $\mathbf{G}$ as claimed.

Obviously $(\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H} \subset \mathcal{E}\backslash\mathcal{H}$, but we also assumed that $\mathcal{E}^{1,2} \subset \mathcal{E}\backslash\mathcal{H}$. Therefore the linear independence of the functions in (3.80) is a mere consequence of Theorem 3.29 again.

All that remains, is to check that the number of basis functions in (3.80) is $\mathring{n} = \dim \mathring{\mathcal{V}}$, as required. Since Γ_N has 2 disjoint boundary components, using Lemma 3.9, we have

$$(\#\mathcal{N}_N - \#\mathcal{E}_N + \#\mathcal{F}_N) = 0. \tag{3.82}$$

Using (3.82) instead of (3.74), the remainder of the proof follows exactly the same lines as the proof to Corollary 3.32. $\square$

Remark 3.35. Theorem 3.34 extends in a straightforward way to the general case, where Γ_D consists of s_D connected components that are disjoint, i.e.

$$\Gamma_D = \Gamma_D^1 \cup \Gamma_D^2 \cup \ldots \cup \Gamma_D^{s_D}, \qquad \overline{\Gamma}_D^j \cap \overline{\Gamma}_D^{j'} = \emptyset \quad \text{for all } 1 \le j \ne j' \le s_D,$$

and Γ_N consists of s_N connected components that are disjoint, i.e.

$$\Gamma_N = \Gamma_N^1 \cup \Gamma_N^2 \cup \ldots \cup \Gamma_N^{s_N}, \qquad \overline{\Gamma}_N^\ell \cap \overline{\Gamma}_N^{\ell'} = \emptyset \quad \text{for all } 1 \le \ell \ne \ell' \le s_N.$$

For each index $\ell = 1, \ldots, s_N$, let $\mathcal{J}^\ell \subset \{1, \ldots, s_D\}$ be such that Γ_D^j is adjacent to Γ_N^ℓ, for all indices $j \in \mathcal{J}^\ell$. The additional functions are now constructed as in Proposition 3.33, by choosing for each pair $j, j' \in \mathcal{J}^\ell$ with $j \ne j'$, a sequence $\mathcal{E}^{j,j'}$ of distinct edges, which satisfies conditions similar to (3.79) and connects Γ_D^j and $\Gamma_D^{j'}$. As above it can be shown that then

$$\sum_{E \in \mathcal{E}^{j,j'}} \vec{\Psi}_E \in \mathring{\mathcal{V}} \tag{3.83}$$

(cf. Proposition 3.33 and Figure 3.8). Since the connected components of Γ_D are pairwise disjoint, the supports of two such functions must be disjoint as well, and so they are linearly independent.

To obtain a basis for $\mathring{\mathcal{V}}$ in this case, we need to add exactly $(\#\mathcal{J}^\ell - 1)$ functions of the form (3.83) to (3.73), for each $\ell = 1, \ldots, s_N$. Some simple considerations show that the total number of functions that are added in this way is $s_D - 1$. However, to rigorously prove that this choice gives a basis for $\mathring{\mathcal{V}}$ is very technical and we will omit it.

3.2.3 Literature on spanning tree techniques for finite elements

The idea of spanning trees in the context of finite element methods appears first in the context of the Stokes problem in a paper by Hecht [55], where it is used in the same way as here to find a basis for the space of divergence-free non-conforming P1-P0 elements for the approximation of divergence-free vector fields in $(H^1(\Omega))^3$.

In an unpublished manuscript [56], Hecht extends these results to a wider family of finite elements in $(H^1(\Omega))^3$ including the (non-conforming) Raviart-Thomas-Nédélec elements. The published literature on divergence-free Raviart-Thomas-Nédélec elements in $H(\mathrm{div}, \Omega)$ considered here is restricted to the pure Neumann case, $\Gamma_D = \emptyset$, in a paper by Dubois [37], where he uses it to solve model incompressible flow problems with prescribed vorticity.

In the context of the decoupled method for problem (2.1) presented in Chapter 4, the only other work in 3D which we are aware of is the recent paper by Cai et al. [21], but this is restricted to uniform rectangular meshes and a special spanning tree which can be constructed *a priori*. Alotto & Perugia [5] also use tree-cotree decompositions for the solution of (2.1) in 2D, but their decoupling strategy is completely different.

Independently, spanning trees also appear as a technique for computing a discrete gauge condition for $\mathcal{ND}_1(\Omega, \mathcal{T})$ in eddy-current calculations in computational electromagnetics (e.g. in Albanese & Rubinacci [4]). A thorough presentation of the theoretical foundation of those techniques using homology theory (related to our Appendix C) can be found in Bossavit [17, Ch. 5].

3.2.4 The space $\mathring{\mathcal{V}}$ – the general case

Now let $k \geq 0$, i.e. include higher order elements. We will see in this section that even in the general case, the vector potentials for divergence-free Raviart-Thomas-Nédélec elements $\mathring{\mathcal{V}}$ of order k (i.e. $\mathring{\mathcal{V}} \subset \mathcal{RT}_k(\Omega, \mathcal{T})$) can still be taken to be the Nédélec elements $\mathcal{ND}_{k+1}(\Omega, \mathcal{T})$ of order $k+1$. Moreover, the 3D curls of these vector potentials span $\mathring{\mathcal{V}}$ again. However, our method of extracting an explicit basis from this spanning set by a graph theoretical technique as presented in Section 3.2.2, seems to be confined to the lowest order case and has not yet been extended to $k > 0$. In this section we will outline why this extension seems difficult and what the alternatives are that have been investigated. To simplify the presentation we will restrict attention to the case $\Gamma_N = \emptyset$.

Let $n_{\mathcal{ND}} := \dim \mathcal{ND}_{k+1}(\Omega, \mathcal{T})$ and let

$$\Xi_{k+1} := \{\alpha_i : i = 1, \ldots, n_{\mathcal{ND}}\}$$

be the set of degrees of freedom of $\mathcal{ND}_{k+1}(\Omega, \mathcal{T})$ as defined in Definition 3.25. Naturally, the canonical basis

$$\{\vec{\Phi}_i : i = 1, \ldots, n_{\mathcal{ND}}\} \tag{3.84}$$

of $ND_{k+1}(T)$ is then given by the functions $\vec{\Phi}_i$, $i = 1, \ldots, n_{\mathcal{ND}}$ whose degrees of freedom

$\alpha_j(\vec{\Phi}_i)$, $j = 1, \ldots, n_{\mathcal{ND}}$, satisfy

$$\alpha_j(\vec{\Phi}_i) = \delta_{i,j}. \tag{3.85}$$

Once again, we define

$$\vec{\Psi}_i := \vec{\text{curl}}\, \vec{\Phi}_i, \tag{3.86}$$

for each $i = 1, \ldots, n_{\mathcal{ND}}$, so that $\vec{\Phi}_i$ is the vector potential of $\vec{\Psi}_i$. The functions (3.86) clearly satisfy $\text{div}\vec{\Psi}_i$ on each tetrahedron of the mesh again. Moreover, if $\Gamma_N = \emptyset$, it can also be shown that they lie in $\mathcal{V}$ and span the entire subset $\mathring{\mathcal{V}}$. This is made precise in the following result:

Proposition 3.36. *Let* $\Gamma_N = \emptyset$. *Then*

$$\mathring{\mathcal{V}} = \text{span}\,\{\vec{\Psi}_i : i = 1, \ldots, n_{\mathcal{ND}}\}.$$

Proof. See Hiptmair [57, Theorem 2.36]. □

However, because of the large kernel of the 3D curl (see Proposition 3.24) the representation

$$\vec{v} = \sum_{i=1}^{n_{\mathcal{ND}}} v_i \vec{\Psi}_i$$

for $\vec{v} \in \mathring{\mathcal{V}}$ is not unique. To obtain a basis for $\mathring{\mathcal{V}}$ from the set $\{\vec{\Psi}_i : i = 1, \ldots, n_{\mathcal{ND}}\}$ it would be necessary to eliminate

$$n_{\mathcal{H}} := n_{\mathcal{ND}} - \mathring{n} = \dim \mathcal{ND}_{k+1}(\Omega, \mathcal{T}) - \dim \mathring{\mathcal{V}}$$

functions.

Let us calculate $n_{\mathcal{H}}$. We know from (3.4) that $\mathring{n} = n_{\mathcal{V}} - n_{\mathcal{W}}$. Now, using (2.51) and (2.53) and the fact that $\#\mathcal{F}_N = 0$, we have

$$\begin{aligned} \mathring{n} &= \left(\tfrac{(k+1)(k+2)}{2}\#\mathcal{F} + \tfrac{k(k+1)(k+2)}{2}\#\mathcal{T}\right) - \left(\tfrac{(k+1)(k+2)(k+3)}{6}\#\mathcal{T}\right) \\ &= \tfrac{(k+1)(k+2)}{2}\#\mathcal{F} + \tfrac{(k+1)(k+2)(2k-3)}{6}\#\mathcal{T}, \end{aligned}$$

and combining this with (3.56), we finally have

$$\begin{aligned} n_{\mathcal{H}} &= (k+1)\left(\#\mathcal{E} + k\,\#\mathcal{F} + \tfrac{(k-1)k}{2}\#\mathcal{T}\right) - \left(\tfrac{(k+1)(k+2)}{2}\#\mathcal{F} + \tfrac{(k+1)(k+2)(2k-3)}{6}\#\mathcal{T}\right) \\ &= (\#\mathcal{N} - 1) + \left(k\,\#\mathcal{E} + \tfrac{(k-1)k}{2}\#\mathcal{F} + \tfrac{(k-2)(k-1)k}{6}\#\mathcal{T}\right) \end{aligned} \tag{3.87}$$

where in the last step we used Euler's Polyhedron Theorem (3.71) and the assumption that Ω is simply connected. For $k = 0$, this formula reduces to $n_{\mathcal{H}} = (\#\mathcal{N} - 1)$, in correspondence with Theorem 3.29, and we saw that eliminating the $(\#\mathcal{N} - 1)$ degrees of freedom defined on the edges of a spanning tree of the graph associated with the triangulation, leads to a basis for $\mathring{\mathcal{V}}$. However, if $k > 0$, then $n_{\mathcal{H}}$ is much larger, and it is not obvious at all, how the redundant degrees of freedom could be eliminated using graph theoretical techniques.

An alternative approach to graph theoretical techniques, is to enforce a discrete version of the Coulomb gauge condition (3.49). This approach has been investigated by Nédélec in [74]. To present this, we need to introduce the space $\mathcal{S}_{k+1}(\Omega, \mathcal{T})$ of $H^1(\Omega)$-conforming finite element functions, which was presented in (3.9) in Section 3.1.1 for $d = 2$, for $d = 3$ as well (see Ciarlet [27]). The only difference to the 2D case is that instead of (3.8), the degrees of freedom on an element $T \in \mathcal{T}$ in 3D are given by the nodes

$$\Sigma_{k+1}(T) := \left\{ \vec{x} := \sum_{i=1}^{4} \lambda_i \vec{a}_i \ : \ \sum_{i=1}^{4} \lambda_i = 1 \text{ and } \lambda_i \in \{0, \tfrac{1}{k+1}, \ldots, \tfrac{k}{k+1}, 1\} \right\}. \tag{3.88}$$

This means that the global set $\Sigma_{k+1} := \bigcup_{T \in \mathcal{T}} \Sigma_{k+1}(T)$ of degrees of freedom contains all the nodes $P \in \mathcal{N}$, k equidistributed nodes in the interior of each edge $E \in \mathcal{E}$ (for $k > 0$), $\frac{(k-1)k}{2}$ equidistributed nodes in the interior of each face $F \in \mathcal{F}$ (for $k > 1$), and $\frac{(k-2)(k-1)k}{6}$ equidistributed nodes in the interior of each element $T \in \mathcal{T}$ (for $k > 2$). Therefore,

$$\dim \mathcal{S}_{k+1}(\Omega, \mathcal{T}) = \#\mathcal{N} + k\#\mathcal{E} + \frac{(k-1)k}{2}\#\mathcal{F} + \frac{(k-2)(k-1)k}{6}\#\mathcal{T}. \tag{3.89}$$

Note that $\dim \mathcal{S}_{k+1}(\Omega, \mathcal{T}) = n_{\mathcal{H}} + 1$. Thus, the elements $\xi \in \mathcal{S}_{k+1}(\Omega, \mathcal{T})$ would provide us with (almost) the right amount of conditions to obtain a unique discrete vector potential. And indeed, we have the following characterisation of an element $\vec{v} \in \mathring{\mathcal{V}}$.

Proposition 3.37. *For each element $\vec{v} \in \mathring{\mathcal{V}}$, there exists one and only one element $\vec{\Phi} \in \mathcal{ND}_{k+1}(\Omega, \mathcal{T})$ such that*

$$\vec{v} = \vec{\mathrm{curl}}\vec{\Phi}, \tag{3.90}$$

and

$$\int_\Omega \vec{\Phi} \cdot \vec{\nabla}\xi d\vec{x} = 0, \quad \text{for all } \xi \in \mathcal{S}_{k+1}(\Omega, \mathcal{T}). \tag{3.91}$$

Proof. See Nédélec [74, Theorem 1]. □

Condition (3.91) is the discrete equivalent of the Coulomb gauge condition (3.49). It gives good theoretical results, but an explicit basis of the corresponding finite dimensional linear space is not natural.

Remark 3.38. Another way of avoiding the construction of an explicit basis is presented by Hiptmair [57] and Hiptmair & Hoppe [59]. They use the entire spanning set in Proposition 3.36, and eliminate the kernel of the $\vec{\mathrm{curl}}$ operator in a multilevel fashion by relaxing the orthogonality with respect to $\ker(\vec{\mathrm{curl}})$. This does not lead to a basis for $\mathring{\mathcal{V}}$, but it leads to a stable splitting of $\mathring{\mathcal{V}}$, which (under some assumptions on the domain and the boundary conditions) is sufficient for the construction of an optimal preconditioner for problem (2.1) (see also Remark 4.6).

3.3 The complementary space $\mathcal{V}^c$

First recall that $\mathcal{V}^c$ denotes a complementary space of $\mathring{\mathcal{V}}$ in $\mathcal{V}$ (see (3.3)). Note that this space is obviously not unique. However, it follows from (3.3) that

$$\dim\,\mathcal{V}^c = \dim\,\mathcal{V} - \dim\,\mathring{\mathcal{V}} = n_{\mathcal{V}} - \mathring{n} = n_{\mathcal{W}} = \#\mathcal{T}.$$

In this section, we present a procedure for the construction of such a space $\mathcal{V}^c$ in the lowest order case $k = 0$ (for $d = 2$ and 3). This can be done by seeking a distinguished subset of faces

$$\mathcal{F}^c \subset \mathcal{F}_I \cup \mathcal{F}_D \tag{3.92}$$

such that the corresponding subset of Raviart-Thomas-Nédélec basis functions, $\{\vec{v}_F : F \in \mathcal{F}^c\}$, constitutes a basis for $\mathcal{V}^c$. Note that this set must contain $n_{\mathcal{W}}$ elements, and that it has to be linearly independent from the basis of $\mathring{\mathcal{V}}$ found above. The following simple algorithm chooses $n_{\mathcal{W}}$ appropriate faces.

Algorithm 3.39.

1. *Choose $T_1 \in \mathcal{T}$ to be any element with a face $F_1 \in \mathcal{F}_D$ and set $\mathcal{F}^c = \{F_1\}$.*
2. *For $j = 2, \ldots, n_{\mathcal{W}}$,*
   - *choose $T_j \in \mathcal{T} \backslash \{T_\ell : \ell = 1, \ldots, j-1\}$ with the property that there exists $F_j \in \mathcal{F}_I$ such that*
   $$F_j \subset \overline{T}_j \cap \left\{ \bigcup_{\ell=1}^{j-1} \overline{T}_\ell \right\} \tag{3.93}$$
   - *update $\mathcal{F}^c = \mathcal{F}^c \cup \{F_j\}$.*

Proposition 3.40. *Algorithm 3.39 is well defined.*

Proof. Since $\Gamma_D \neq \emptyset$, there exists a $T \in \mathcal{T}$ with a face $F \in \mathcal{F}_D$. Choose $T_1 = T$ and $F_1 = F$ in Step 1.

Now assume we have found $j - 1 < n_{\mathcal{W}} = \#\mathcal{T}$ tetrahedra T_ℓ in Step 2 that fulfill property (3.93). Since Ω is connected, there exists a tetrahedron $T \in \mathcal{T}$ that has a face F in common with $\bigcup_{\ell=1}^{j-1} \overline{T}_\ell$. Choose $T_j = T$ and $F_j = F$. The existence of a set $\mathcal{F}^c$ therefore follows by an inductive argument. □

The subset of Raviart-Thomas-Nédélec basis functions corresponding to the faces found in Algorithm 3.39 spans a complementary space $\mathcal{V}^c$ of $\mathring{\mathcal{V}}$ as the following theorem shows. We will only state the theorem for simply connected domains, and furthermore assume that in 3D each component Γ_N^ℓ of Γ_N is simply connected. However, the extension to the multiply connected domains discussed in Corollary 3.15 and Theorem 3.17 for 2D and to three-dimensional domains of the form discussed in Theorem 3.34 is a simple corollary to this theorem.

Theorem 3.41. *Let Ω be simply connected with a connected boundary Γ. If $d = 3$, we furthermore assume that Γ_N^ℓ is simply connected for each $\ell = 1, \ldots, s_N$. Then the functions*

$$\{\vec{v}_F : F \in \mathcal{F}^c\} \tag{3.94}$$

are linearly independent and $\mathcal{V}^c := \operatorname{span}\{\vec{v}_F : F \in \mathcal{F}^c\}$ is a complementary space of $\mathring{\mathcal{V}}$. Thus, (3.94) is a basis of $\mathcal{V}^c$.

We will only give the proof for 3D at the end of this section. The two-dimensional case can be proved analogously. So, in the following let $d = 3$.

Before proving this theorem we will first prove two lemmas again. Recall, that by $\mathbf{G} := (\mathcal{N}, \mathcal{E})$ we denoted the graph formed by the nodes and edges of the triangulation $\mathcal{T}$, and that by $\mathcal{V}(\mathbf{G})$ we denoted the vector space over $\mathbb{Z}$ generated by the cycles of $\mathbf{G}$ as defined in Definition B.1(e). Furthermore, for each face $F \in \mathcal{F}$, recall the definition of the vector $\boldsymbol{\mu}^F \in \mathcal{V}(\mathbf{G})$, which is associated with the elementary cycle μ^F of $\mathbf{G}$ formed by the edges E of F and orientated with respect to $\vec{\nu}_F$ (see Section 3.2.2 after Theorem 3.29).

Lemma 3.42. *Let $\boldsymbol{\mu} := [\mu_E]_{E\in\mathcal{E}} \in \mathcal{V}(\mathbf{G})$. Then there exist $\{\tilde{\alpha}_F \in \mathbb{Z} : F \in \mathcal{F}\backslash\mathcal{F}^c\}$ such that*

$$\boldsymbol{\mu} := \sum_{F\in\mathcal{F}\backslash\mathcal{F}^c} \tilde{\alpha}_F \boldsymbol{\mu}^F. \tag{3.95}$$

Proof. Let $F \in \mathcal{F}^c$. We will first show that there exist $\{\alpha^F_{F'} \in \mathbb{Z} : F' \in \mathcal{F}\backslash\mathcal{F}^c\}$ such that

$$\boldsymbol{\mu}^F = \sum_{F'\in\mathcal{F}\backslash\mathcal{F}^c} \alpha^F_{F'} \boldsymbol{\mu}^{F'} . \tag{3.96}$$

Let $j \in \{1, \ldots, n_{\mathcal{W}}\}$ be such that $F = F_j$ in Algorithm 3.39, and let F', F'', and F''' be the other faces of T_j, then there exist $\alpha^F_{F'}, \alpha^F_{F''}, \alpha^F_{F'''} \in \{-1, 1\}$ such that

$$\boldsymbol{\mu}^F = \alpha^F_{F'}\boldsymbol{\mu}^{F'} + \alpha^F_{F''}\boldsymbol{\mu}^{F''} + \alpha^F_{F'''}\boldsymbol{\mu}^{F'''}$$

as depicted in Figure 3.9.

If $F', F'', F''' \in \mathcal{F}\backslash\mathcal{F}^c$, the proof of (3.96) is complete. Otherwise assume, without loss of generality, that $F' \in \mathcal{F}^c$. By construction there has to be a $j' \in \{j+1, \ldots, n_{\mathcal{W}}\}$ such that $F' = F_{j'}$. Let $\tilde{F}'$, $\tilde{F}''$, and $\tilde{F}'''$ be the other faces of $T_{j'}$. As before there exist $\alpha^{F'}_{\tilde{F}'}, \alpha^{F'}_{\tilde{F}''}, \alpha^{F'}_{\tilde{F}'''} \in \{-1, 1\}$ such that $\boldsymbol{\mu}^{F'} = \alpha^{F'}_{\tilde{F}'}\boldsymbol{\mu}^{\tilde{F}'} + \alpha^{F'}_{\tilde{F}''}\boldsymbol{\mu}^{\tilde{F}''} + \alpha^{F'}_{\tilde{F}'''}\boldsymbol{\mu}^{\tilde{F}'''}$ and therefore

$$\boldsymbol{\mu}^F = \alpha^F_{F'}(\alpha^{F'}_{\tilde{F}'}\boldsymbol{\mu}^{\tilde{F}'} + \alpha^{F'}_{\tilde{F}''}\boldsymbol{\mu}^{\tilde{F}''} + \alpha^{F'}_{\tilde{F}'''}\boldsymbol{\mu}^{\tilde{F}'''}) + \alpha^F_{F''}\boldsymbol{\mu}^{F''} + \alpha^F_{F'''}\boldsymbol{\mu}^{F'''}.$$

If $F'', F''', \tilde{F}', \tilde{F}'', \tilde{F}''' \in \mathcal{F}\backslash\mathcal{F}^c$, the proof of (3.96) is complete. Otherwise, we can repeat

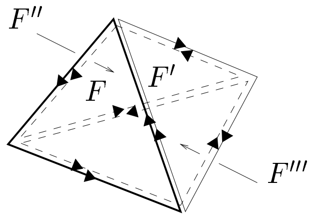


Figure 3.9: The cycle $\mu^F = \mu^{F'} + \mu^{F''} + \mu^{F'''}$ with the orientations as indicated

the above procedure for the faces $F'', F''', \tilde{F}', \tilde{F}''$, and $\tilde{F}'''$, and since the set $\{1, \ldots, n_{\mathcal{W}}\}$ is finite, the procedure will terminate in a finite number of steps. Altogether we have established that there exist $\{\alpha^F_{F'} \in \mathbb{Z} : F' \in \mathcal{F}\backslash\mathcal{F}^c\}$ such that (3.96) holds.

Now let $\boldsymbol{\mu} \in \mathcal{V}(\mathbf{G})$. Substituting (3.96) into (3.68), we find that there exist $\{\tilde{\alpha}_F \in \mathbb{Z} : F \in \mathcal{F}\backslash\mathcal{F}^c\}$ such that $\boldsymbol{\mu} = \sum_{F\in\mathcal{F}\backslash\mathcal{F}^c} \tilde{\alpha}_F \boldsymbol{\mu}^F$. □

Lemma 3.43. *Let $\boldsymbol{\mu} \in \mathcal{V}(\mathbf{G})$ and let $\{\tilde{\alpha}_F \in \mathbb{Z} : F \in \mathcal{F}\backslash\mathcal{F}^c\}$ be such that $\boldsymbol{\mu} := \sum_{F\in\mathcal{F}\backslash\mathcal{F}^c} \tilde{\alpha}_F \boldsymbol{\mu}^F$. Then*

$$\sum_{F\in\mathcal{F}\backslash\mathcal{F}^c} \tilde{\alpha}_F \int_F \vec{v}_{F'} \cdot \vec{\nu}_F \, ds = 0, \qquad \textit{for all } F' \in \mathcal{F}^c. \tag{3.97}$$

Proof. Let $F' \in \mathcal{F}^c$, then $\vec{v}_{F'} \cdot \vec{\nu}_F = 0$ for all $F \in \mathcal{F}\backslash\mathcal{F}^c$ which implies (3.97). □

Proof of Theorem 3.41. Note first that the functions in (3.94) form a subset of the Raviart-Thomas-Nédélec basis functions, so they have to be linearly independent by definition. Furthermore, $\#\mathcal{F}^c = n_{\mathcal{W}}$ which is by (3.4) equal to $\dim \mathcal{V} - \dim \mathring{\mathcal{V}}$. Therefore it only remains to show that, the union of sets of functions (3.94) and (3.73) is a linearly independent set.

To prove this, let $\mathcal{H} \subset \mathcal{E}$ be defined as in Corollary 3.32, and suppose $\{\beta_{E'} : E' \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}\}$ and $\{\gamma_F : F \in \mathcal{F}^c\}$ are scalars such that

$$\vec{0} = \sum_{E'\in(\mathcal{E}_I\cup\mathcal{E}_D)\backslash\mathcal{H}} \beta_{E'} \vec{\Psi}_{E'} + \sum_{F\in\mathcal{F}^c} \gamma_F \vec{v}_F.$$

Let $E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}$ and let $\boldsymbol{\mu}^E$ denote the vector associated to the unique cycle μ^E generated by taking edge E into the tree $\mathbf{H} = (\mathcal{N}, \mathcal{H})$, which has the property that $\mu^E_{E'} := \delta_{E,E'}$, for all $E' \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}$ (cf. Theorem B.6 and the proof to Theorem 3.29). Now, using Lemma 3.42 we can find $\{\tilde{\alpha}_F \in \mathbb{Z} : F \in \mathcal{F}\backslash\mathcal{F}^c\}$ such that $\boldsymbol{\mu} := \sum_{F\in\mathcal{F}\backslash\mathcal{F}^c} \tilde{\alpha}_F \boldsymbol{\mu}^F$, and so by Lemmas 3.31 and 3.43

$$
\begin{aligned}
0 &= \sum_{F\in\mathcal{F}\setminus\mathcal{F}^c} \tilde{\alpha}_F \int_F \Bigg(\sum_{E\in(\mathcal{E}_I\cup\mathcal{E}_D)\setminus\mathcal{H}} \beta_E\,\vec{\Psi}_E + \sum_{F'\in\mathcal{F}^c} \gamma_{F'}\,\vec{v}_{F'} \Bigg) \cdot \vec{\nu}_F\, ds \\
&= \sum_{E'\in(\mathcal{E}_I\cup\mathcal{E}_D)\setminus\mathcal{H}} \beta_{E'} \sum_{F\in\mathcal{F}\setminus\mathcal{F}^c} \tilde{\alpha}_F \int_F \vec{\Psi}_E \cdot \vec{\nu}_F\, ds \;+\; \sum_{F'\in\mathcal{F}^c} \gamma_{F'} \sum_{F\in\mathcal{F}\setminus\mathcal{F}^c} \tilde{\alpha}_F \int_F \vec{v}_{F'} \cdot \vec{\nu}_F\, ds \\
&= \sum_{E'\in(\mathcal{E}_I\cup\mathcal{E}_D)\setminus\mathcal{H}} \beta_{E'}\mu^E_{E'} \;=\; \beta_E\,.
\end{aligned}
$$

Since the functions $\{\vec{v}_F : F \in \mathcal{F}^c\}$ are linearly independent, we also have $\gamma_F = 0$, for all $F \in \mathcal{F}^c$, which establishes the linear independence of the functions in (3.94) and (3.73). □

3.4 Summary

In this chapter we investigated the subspace $\mathring{\mathcal{V}} \subset \mathcal{RT}_k(\Omega,\mathcal{T})$ of divergence-free Raviart-Thomas-Nédélec elements and its complement $\mathcal{V}^c$ in $\mathcal{V}$. In particular, we were interested in finding a basis for $\mathring{\mathcal{V}}$. This basis was constructed from the curls of suitable stream functions and vector potentials, in 2D and 3D respectively.

We found that in 2D the stream functions are given by the C^0-elements $\mathcal{S}_{k+1}(\Omega,\mathcal{T})$ of order $k+1$, which are conforming in $H^1(\Omega)$. We defined the canonical (nodal) basis $\{\Phi_P\}$ for them and discussed some important properties. In a series of Propositions, Theorems and Corollaries we then established that for any simply or multiply connected domain in 2D, if the Dirichlet boundary Γ_D is connected, then the curls of the basis functions Φ_P that vanish on the Neumann boundary Γ_N form a basis for $\mathring{\mathcal{V}}$. If the Dirichlet boundary is not connected, a small number of additional basis functions (introduced in Propositions 3.6, 3.12, 3.16) is needed, which connect the disjoint components of the Dirichlet boundary. They are constructed as linear combinations of the curls of the basis functions Φ_P and are therefore (modestly) non-local.

The vector potentials in 3D, on the other hand, are given by Nédélec's edge elements $\mathcal{ND}_{k+1}(\Omega,\mathcal{T})$ of order $k+1$, which are conforming in $H(\vec{\mathrm{curl}},\Omega)$. To begin with, we discussed some important properties of the space $H(\vec{\mathrm{curl}},\Omega)$. In particular we noted in Proposition 3.24 that the kernel of the 3D curl in $H(\vec{\mathrm{curl}},\Omega)$ is given by $\vec{\nabla} H^1(\Omega)$ and is therefore non-trivial. We also saw (cf. Proposition 3.26) that the tangential trace $\vec{\Phi} \times \vec{\nu}_F$ of a function $\vec{\Phi} \in H(\vec{\mathrm{curl}},\Omega)$ is continuous across each face F of the triangulation. Then we defined $\mathcal{ND}_{k+1}(\Omega,\mathcal{T})$. In the lowest order case, a function $\vec{\Phi} \in \mathcal{ND}_1(\Omega,\mathcal{T})$ can be fully described by the integral of its tangential component along each edge of the triangulation, leading to the standard basis $\{\vec{\Phi}_E : E \in \mathcal{E}\}$ of $\mathcal{ND}_1(\Omega,\mathcal{T})$ and accounting for the widely used term "edge elements".

To construct a basis for $\mathring{\mathcal{V}}$ in 3D, we restricted ourselves to the lowest order case $k = 0$ and to simply connected domains Ω with connected boundary Γ. We saw that the 3D curls of the basis functions $\vec{\Phi}_E$ whose tangential trace $\vec{\Phi}_E \times \vec{\nu}$ vanish on the Neumann boundary Γ_N are sufficient to span the set $\mathring{\mathcal{V}}$ again, if the Dirichlet boundary Γ_D is connected. However, unlike in 2D, in 3D there are too many of them (this is a result of the large kernel of the 3D curl operator in $\mathcal{ND}_1(\Omega, \mathcal{T})$), and to construct a basis for $\mathring{\mathcal{V}}$ it was necessary to eliminate a subset $\{\vec{\Phi}_E : E \in \mathcal{H} \subset \mathcal{E}\}$ of them. Using some fundamental results from graph theory and algebraic topology, we established in Theorem 3.29 and Corollary 3.32 that $\mathcal{H}$ has to be chosen as the set of edges corresponding to a spanning tree in the graph $\mathbf{G} := (\mathcal{N}, \mathcal{E})$ associated with the mesh. Moreover, if $\Gamma_N \neq \emptyset$, this spanning tree has got to reduce to a spanning tree on each connected component of Γ_N. An efficient algorithm for the construction of such a spanning tree is given in Algorithm B.7 in the Appendix. If the Dirichlet boundary Γ_D is not connected, we need, as in 2D, a small number of additional basis functions, which connect the disjoint components of the Dirichlet boundary (c.f. Proposition 3.33, Theorem 3.34 and Remark 3.35). As in 2D, they are constructed as linear combinations of the curls of the basis functions $\vec{\Phi}_E$ and are (modestly) non-local.

For higher order elements in 3D, i.e. $k > 0$, we saw that it is still true that $\mathring{\mathcal{V}} = \vec{\text{curl}}\,\mathcal{ND}_{k+1}(\Omega, \mathcal{T})$ (cf. Proposition 3.36 for $\Gamma_N = \emptyset$), but we were not able to construct an explicit basis. Whether this is possible is an open question.

Finally, in Section 3.3 we presented a simple algorithm for the construction of a basis for the complementary space $\mathcal{V}^c$ in the lowest order case $k = 0$ (in 2D and 3D). By finding a distinguished subset of faces $\mathcal{F}^c \subset \mathcal{F}$ we obtained a basis of $\mathcal{V}^c$ consisting of the corresponding Raviart-Thomas-Nédélec basis functions $\{\vec{v}_F : F \in \mathcal{F}^c\}$.

The results in Sections 3.1 and 3.2 are related to corresponding ideas for the classical Stokes problem. Some of the results for pure Dirichlet or Neumann boundary conditions on simply connected domains can already be found in other places in the literature, but here we present for the first time an extensive discussion of divergence-free Raviart-Thomas-Nédélec elements for mixed boundary conditions, where additional non-local basis functions are needed. In particular, the construction of an explicit basis in 3D and the extension to multiply connected domains and mixed boundary conditions in 2D are novel ideas and will be used in later chapters not only in a theoretical way, but in fact to construct explicit solvers. The results in Section 3.3 on the construction of a basis for the complementary space, are completely original, and we will see in Section 4.1.3, why they are so important.

Chapter 4

A Decoupled Iterative Method

In this chapter we formulate an iterative method for solving indefinite saddle point systems of the form

$$\begin{pmatrix} M & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{g} \\ \mathbf{f} \end{pmatrix} \qquad \text{in} \quad \mathbb{R}^{n_\mathcal{V}} \times \mathbb{R}^{n_\mathcal{W}} , \tag{4.1}$$

as defined in (2.68–2.74) in Chapter 2, arising from mixed finite element approximations of the second-order elliptic boundary value problem (2.1). For an overview of other iterative methods for (4.1) see Section 2.3.3.

To present the principal idea of our method, let us first assume that the underlying continuous problem for (4.1) is given in divergence form (2.18), as is often the case in practical applications (e.g. in groundwater flow, cf. Chapter 5). Then $\mathbf{f} = \mathbf{0}$, and (4.1) reduces to

$$\begin{pmatrix} M & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{g} \\ \mathbf{0} \end{pmatrix} \qquad \text{in} \quad \mathbb{R}^{n_\mathcal{V}} \times \mathbb{R}^{n_\mathcal{W}} , \tag{4.2}$$

with $\mathbf{g} := [g_i]_{n_\mathcal{V}}$ and $g_i := F_\mathcal{D}(\vec{v}_i)$. Using the bases for $\mathring{\mathcal{V}}$ and $\mathcal{V}^c$ that we found in Chapter 3, we can decouple (4.2) into a symmetric positive definite system for the velocity unknown $\mathbf{u}$ and into a triangular system for the pressure unknown $\mathbf{p}$. The bulk of the numerical work lies in the solution of the symmetric positive definite velocity system, which we solve by preconditioned conjugate gradients. We will see that this system is much smaller than the original system, and better suited to efficient preconditioning techniques.

If the underlying continuous problem is not in divergence form and $\mathbf{f} \neq \mathbf{0}$, we first need to find a particular solution $\mathbf{u}^*$ to the constraint

$$B^T \mathbf{u}^* = \mathbf{f}$$

in a preprocessing step. This particular solution $\mathbf{u}^*$ is the discrete analogue to $-\vec{f}_\mathcal{D}$ in the continuous problem (cf. Lemma 2.3), and can be found for example by a certain domain decomposition technique (also known as *static condensation*). Now, by setting

$\mathbf{u} := \mathbf{u}^* + \mathbf{u}^0$, we see that (4.1) is equivalent to solving

$$\begin{pmatrix} M & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}^0 \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{g} - M\mathbf{u}^* \\ \mathbf{0} \end{pmatrix},$$

which is of the same form as (4.2), with $\mathbf{f} = \mathbf{0}$.

The method for decoupling (4.1) was first mentioned in a paper by Chavent et al. [24], and related analysis can be found in a series of papers by Ewing & Wang [38, 39] and Mathew [71, 72]. However, these analyses are restricted to simply connected two-dimensional domains and pure Neumann boundary conditions, and do not provide a general basis for $\mathring{\mathcal{V}}$. A partial extension of the method to 3D by Cai et al. [21] resorts to uniform rectangular grids and pure Neumann boundary conditions to simplify the construction of a basis for $\mathring{\mathcal{V}}$. Related work by Hiptmair et al. [59] treats more general three-dimensional domains and non-uniform grids, but avoids the construction of a basis for $\mathring{\mathcal{V}}$ at the expense of solving a semidefinite velocity system (see Remark 4.6).

With the results of Chapter 3 in hand, we are able to apply the decoupling method to a much larger variety of problems than previously possible; in particular, to three-dimensional problems on non-uniform grids with general mixed boundary conditions, and to multiply connected domains in 2D. We would also like to point out that the recovery of the vector of pressures $\mathbf{p}$ has previously only been investigated by Mathew [71], and because of the lack of an appropriate basis for $\mathcal{V}^c$, they have to resolve to a variation of domain decomposition to obtain $\mathbf{p}$. Here, we are able to use the basis for $\mathcal{V}^c$ found in Section 3.3, and the recovery of $\mathbf{p}$ reduces to a triangular system that can be solved by simple back substitutions. The numerical results at the end of this chapter confirm the predicted excellent behaviour of the method even in the presence of highly discontinuous coefficients.

The chapter is arranged as follows. First in Section 4.1 we present the decoupling procedure and analyse the pressure system. Then, in Sections 4.2 and 4.3, we discuss the solution of the velocity system for 2D and 3D, respectively. In particular, we present a very efficient and robust parallel solver for 2D. In Section 4.4 we deal with the case of non-zero divergence, and finally in Section 4.5 we finish the discussion with a series of numerical results.

A lot of the results in this chapter have already been presented in the joint papers Cliffe et al. [28, 29] for the 2D case, and in Scheichl [85] for the 3D case.

4.1 Decoupling procedure

In this section we formulate our method for decoupling the vector of velocities $\mathbf{u}$ from the vector of pressures $\mathbf{p}$ in systems (4.1) and (4.2). Let us first consider only (4.2). We will come back to (4.1) in Section 4.3.

As a motivation for our procedure, recall the decoupling of the continuous problem

(2.18), which we employed in Section 2.1.2 to show existence and uniqueness of the solution. We used the divergence-free subspace $\mathcal{Z}$ of $H_{0,N}(\mathrm{div},\Omega)$ and its orthogonal complement $\mathcal{Z}^{\perp}$ to write (2.18) as a decoupled system (2.21) for the continuous velocity $\vec{u} \in \mathcal{Z}$ and pressure $p \in L_2(\Omega)$. In the same way, we can use the finite dimensional subspaces $\mathring{\mathcal{V}}$ and $\mathcal{V}^c$ of $\mathcal{V}$ to decouple (2.29) into the equivalent problem of finding $(\vec{U}, P) \in \mathring{\mathcal{V}} \times \mathcal{W}$ such that

$$\left.\begin{array}{rcll} m(\vec{U},\vec{V}) & = & F_{\mathcal{D}}(\vec{V}) \ , & \text{for all } \vec{V} \in \mathring{\mathcal{V}}, \\ b(\vec{V},P) & = & F_{\mathcal{D}}(\vec{V}) - m(\vec{U},\vec{V}) \ , & \text{for all } \vec{V} \in \mathcal{V}^c. \end{array}\right\} . \tag{4.3}$$

However, it is important to note right away that this procedure (in the finite dimensional case) is of strictly algebraic nature and does not change the approximation properties of the finite element solution. On the contrary, the incompressibility constraint is enforced a priori on each element by restricting explicitly to the divergence-free part of the approximation space.

To stress the algebraic nature of the procedure, we begin by treating (4.2) as an abstract algebraic system and only assume that M is symmetric positive definite and that (4.2) has a unique solution.

4.1.1 Abstract algebraic process

If $\mathbf{f} = \mathbf{0}$, then clearly $\mathbf{u}$ is in $\ker B^T$. Furthermore, since (4.2) has a unique solution, it follows that B must have full rank, and

$$\mathring{n} := \dim\left(\ker B^T\right) = n_{\mathcal{V}} - n_{\mathcal{W}}. \tag{4.4}$$

The decoupling of $\mathbf{u}$ from $\mathbf{p}$, can be achieved by finding

$$\text{a basis } \ \{\mathbf{z}_1, \ldots, \mathbf{z}_{\mathring{n}}\} \ \text{ of } \ \ker B^T. \tag{4.5}$$

If we have such a basis, then the solution $\mathbf{u}$ of (4.2) can be written

$$\mathbf{u} = \sum_{j=1}^{\mathring{n}} \mathring{u}_j \mathbf{z}_j = Z^T \mathring{\mathbf{u}}, \tag{4.6}$$

for some $\mathring{\mathbf{u}} \in \mathbb{R}^{\mathring{n}}$, where Z denotes the $\mathring{n} \times n_{\mathcal{V}}$ matrix with rows $\mathbf{z}_1^T, \ldots, \mathbf{z}_{\mathring{n}}^T$. Also, since $ZB = (B^T Z^T)^T = 0$, multiplying the first (block) row of (4.2) by Z shows that $\mathring{\mathbf{u}}$ is a solution of the linear system

$$\mathring{A}\mathring{\mathbf{u}} = \mathring{\mathbf{g}} \qquad \text{in} \quad \mathbb{R}^{\mathring{n}} \tag{4.7}$$

where

$$\mathring{A} = ZMZ^T \quad \text{and} \quad \mathring{\mathbf{g}} = Z\mathbf{g}. \tag{4.8}$$

Since M is symmetric positive definite, so is $\mathring{A}$, and $\mathring{\mathbf{u}}$ is the unique solution of (4.7). Thus if the basis (4.5) can be found, then the vector $\mathbf{u}$ in (4.2) can be computed by solving the decoupled positive definite system (4.7) rather than the indefinite coupled system (4.2).

The vector $\mathbf{p}$ can also be recovered, provided we have

$$\text{a } \textit{complementary basis} \quad \{\mathbf{z}_{\mathring{n}+1}, \ldots, \mathbf{z}_{n_\mathcal{V}}\}, \tag{4.9}$$

with the property that

$$\text{span}\{\mathbf{z}_1, \ldots, \mathbf{z}_{\mathring{n}}, \mathbf{z}_{\mathring{n}+1}, \ldots, \mathbf{z}_{n_\mathcal{V}}\} = \mathbb{R}^{n_\mathcal{V}}. \tag{4.10}$$

If this is known and if Z^c denotes the $n_\mathcal{W} \times n_\mathcal{V}$ matrix with rows $\mathbf{z}^T_{\mathring{n}+1}, \ldots, \mathbf{z}^T_{n_\mathcal{V}}$, then multiplying the first (block) row of (4.2) by Z^c shows that $\mathbf{p}$ is the solution of the $n_\mathcal{W} \times n_\mathcal{W}$ system

$$A^c\mathbf{p} = \mathbf{g}^c \qquad \text{in} \quad \mathbb{R}^{n_\mathcal{W}}, \tag{4.11}$$

where

$$A^c = Z^c B \quad \text{and} \quad \mathbf{g}^c = Z^c(\mathbf{g} - M\mathbf{u}). \tag{4.12}$$

An elementary result from linear algebra states that for any non-singular matrix $P \in \mathbb{R}^{m\times m}$ and for any matrix $Q \in \mathbb{R}^{m\times n}$ we have $\text{rank}(PQ) = \text{rank}(Q)$. Using this result together with (4.10) we have

$$\text{rank}(A^c) = \text{rank}\left(\begin{pmatrix} 0 \\ Z^c B \end{pmatrix}\right) = \text{rank}\left(\begin{pmatrix} Z \\ Z^c \end{pmatrix} B\right) = \text{rank}(B) = n_\mathcal{W}, \tag{4.13}$$

where in the last step we used the fact that (4.2) is non-singular again. Therefore A^c has full rank and so the unique solution $\mathbf{p}$ of (4.11) also determines the vector $\mathbf{p}$ in (4.2) once $\mathbf{u}$ is known.

4.1.2 Particular case of mixed finite element system

We show in the following that in the particular case of the mixed finite element system (4.2):

(i) It is always easy to find the basis (4.5).

(ii) In 2D, the matrix $\mathring{A}$ can be obtained by simple algebraic techniques from the stiffness matrix of an associated H^1-elliptic problem.

Additionally we will show that for lowest order elements (i.e. $k = 0$):

(iii) In 3D, the matrix $\mathring{A}$ can be obtained by simple algebraic techniques from the stiffness matrix of a symmetric positive semidefinite problem in the space $H(\vec{\text{curl}}, \Omega)$.

(iv) The resulting symmetric positive definite matrix $\mathring{A}$ in the reduced problem (4.7) is about 5 times smaller than (4.2) in 2D and about 3 times smaller in 3D.

(v) A simple choice of complimentary basis (4.9) can be made so that the coefficient matrix A^c in the system (4.11) is lower triangular.

To establish conclusions (i) – (v) we need to exploit the particular properties of (4.2), and link our abstract procedure in Section 4.1.1 with the decoupling (4.3) of the finite element system given at the beginning of Section 4.1. In particular note that finding the basis $\{\mathbf{z}_1, \ldots, \mathbf{z}_{\mathring{n}}\}$ in (4.5) is equivalent to finding a basis $\{\vec{\Psi}_1, \ldots, \vec{\Psi}_{\mathring{n}}\}$ of $\mathring{\mathcal{V}}$. To see why, suppose $\mathbf{z}_1, \ldots, \mathbf{z}_{\mathring{n}}$ are known and let $Z = (Z_{i,j})_{\mathring{n}\times n_{\mathcal{V}}}$ be the matrix with rows $\mathbf{z}_1^T, \ldots, \mathbf{z}_{\mathring{n}}^T$. Then the formulae

$$\vec{\Psi}_i = \sum_{j=1}^{n_{\mathcal{V}}} Z_{i,j}\vec{v}_j \ , \quad i = 1, \ldots, \mathring{n}, \tag{4.14}$$

(where $\{\vec{v}_j\}$ is the basis of $\mathcal{V}$) determine the basis $\{\vec{\Psi}_i\}$, because

$$b(\vec{\Psi}_i, w_k) = \sum_{j=1}^{n_{\mathcal{V}}} Z_{i,j} b(\vec{v}_j, w_k) = (ZB)_{i,k} = 0, \qquad \text{for all } k = 1, \ldots, n_{\mathcal{W}}.$$

Conversely if the basis $\{\vec{\Psi}_i\}$ of $\mathring{\mathcal{V}}$ is known, then the matrix Z (and hence the basis $\{\mathbf{z}_1, \ldots, \mathbf{z}_{\mathring{n}}\}$ of $\ker B^T$) is determined by (4.14). We have established in Chapter 3 how we can easily obtain a basis for $\mathring{\mathcal{V}}$, which establishes conclusion (i).

In complete analogy, the complimentary basis $\{\mathbf{z}_{\mathring{n}+1}, \ldots, \mathbf{z}_{n_{\mathcal{V}}}\}$ in (4.9) is uniquely determined by a basis $\{\vec{\Psi}_{\mathring{n}+1}, \ldots, \vec{\Psi}_{n_{\mathcal{V}}}\}$ of $\mathcal{V}^c$, through the formulae

$$\vec{\Psi}_{\mathring{n}+k} = \sum_{j=1}^{n_{\mathcal{V}}} Z^c_{k,j}\vec{v}_j \ , \quad k = 1, \ldots, n_{\mathcal{W}}. \tag{4.15}$$

Using the representation (4.14), together with (2.68)–(2.71) we can now link the abstract system (4.7) with the first part of the decoupled finite element system in (4.3). We have

$$\mathring{A}_{i,i'} = \sum_{j,j'=1}^{n_{\mathcal{V}}} Z_{i,j} M_{j,j'} Z_{i',j'} = m\Big(\sum_{j=1}^{n_{\mathcal{V}}} Z_{i,j}\vec{v}_j \ , \sum_{j'=1}^{n_{\mathcal{V}}} Z_{i',j'}\vec{v}_{j'}\Big) = m(\vec{\Psi}_i \ , \ \vec{\Psi}_{i'}), \tag{4.16}$$

and

$$\mathring{g}_i = \sum_{j=1}^{n_{\mathcal{V}}} Z_{i,j} g_j = F_{\mathcal{D}}\Big(\sum_{j=1}^{n_{\mathcal{V}}} Z_{i,j}\vec{v}_j\Big) = F_{\mathcal{D}}(\vec{\Psi}_i), \tag{4.17}$$

for $i, i' = 1, \ldots, \mathring{n}$. Equivalently, using (4.15), we have

$$A^c_{k,k'} = \sum_{j=1}^{n_{\mathcal{V}}} Z^c_{k,j} B_{j,k'} = b\Big(\sum_{j=1}^{n_{\mathcal{V}}} Z^c_{k,j}\vec{v}_j \ , w_{k'}\Big) = b(\vec{\Psi}_{\mathring{n}+k} \ , \ w_{k'}), \tag{4.18}$$

and

$$g_k^c = \sum_{j=1}^{n_{\mathcal{V}}} Z_{k,j}^c g_j - \sum_{j,j'=1}^{n_{\mathcal{V}}} Z_{k,j}^c M_{j,j'} u_{j'} = F_{\mathcal{D}}(\vec{\Psi}_{\mathring{n}+k}) - m(\vec{U}, \vec{\Psi}_{\mathring{n}+k}). \tag{4.19}$$

for $k, k' = 1, \ldots, n_{\mathcal{W}}$, which links (4.11) with the second part of the decoupled finite element system (4.3).

Thus, since the functions $\{\vec{\Psi}_i\}$ and $\{\vec{\Psi}_{\mathring{n}+k}\}$ that appear in (4.16)–(4.19) form bases of $\mathring{\mathcal{V}}$ and $\mathcal{V}^c$, we have shown that (4.7), (4.11) is an implementation of the decoupled finite element system (4.3). Hence, in Sections 4.1.3, 4.2, and 4.3 below, we will use the bases for $\mathring{\mathcal{V}}$ and $\mathcal{V}^c$ that we found in Chapter 3, to implement and analyse the decoupled velocity and pressure systems (4.7) and (4.11), in the particular case of the mixed finite element system (4.2). Since the treatment is much simpler, we will first consider (4.11).

4.1.3 Implementation and analysis of the pressure system

Let us assume we know $\mathbf{u}$. Then, to implement the decoupled system (4.11) for recovering $\mathbf{p}$, we must work with the matrix A^c and right hand side $\mathbf{g}^c$ specified in (4.12). We observe that these are formally defined in terms of multiplications with the matrix Z^c which, through (4.15), represents the basis $\{\vec{\Psi}_{\mathring{n}+k}\}$ of $\mathcal{V}^c$ in terms of the basis $\{\vec{v}_j\}$ of $\mathcal{V}$. In view of Section 3.3, we consider only lowest order elements $k = 0$.

However, we saw in Section 3.3 that for $k = 0$ the basis $\{\vec{\Psi}_{\mathring{n}+k}\}$ of $\mathcal{V}^c$ is a subset $\{\vec{v}_F : F \in \mathcal{F}^c\}$ of the basis $\{\vec{v}_F : F \in \mathcal{F}_I \cup \mathcal{F}_D\}$ of $\mathcal{V}$, and we can identify the rows of Z^c with the indices $F \in \mathcal{F}^c$, whereas the columns of Z^c correspond to $F \in \mathcal{F}_I \cup \mathcal{F}_D$. It is easy to see that for this particular choice of bases, (4.15) becomes trivial, i.e.

$$\vec{v}_F = \sum_{F' \in \mathcal{F}_I \cup \mathcal{F}_D} Z_{F,F'}^c \vec{v}_{F'} = \vec{v}_F, \qquad \text{for all} \quad F \in \mathcal{F}^c, \tag{4.20}$$

where $Z_{F,F'}^c = \delta_{F,F'}$. Thus, recalling from (2.79) that the rows of B correspond to indices $F \in \mathcal{F}_I \cup \mathcal{F}_D$, whereas the columns of B can be identified with the indices $T \in \mathcal{T}$, the matrix $A^c = Z^c B$ in (4.12) is nothing more than the minor of B obtained by restricting to rows corresponding to $F \in \mathcal{F}^c$. More precisely it follows from (4.18) and (4.20) that

$$A_{F,T}^c = \sum_{F' \in \mathcal{F}_I \cup \mathcal{F}_D} Z_{F,F'}^c B_{F',T} = B_{F,T}, \qquad \text{for all} \quad F \in \mathcal{F}^c \quad \text{and} \quad T \in \mathcal{T}.$$

In the same way the vector $\mathbf{g}^c = Z^c(\mathbf{g} - M\mathbf{u})$, is nothing more than the subvector of $(\mathbf{g} - M\mathbf{u})$ obtained by restricting to rows corresponding to $F \in \mathcal{F}^c$.

Moreover, we can employ an ordering of the faces $F \in \mathcal{F}^c$ and of the elements $T \in \mathcal{T}$ such that the matrix A^c is lower triangular. Recall that the construction of $\mathcal{F}^c$ in Algorithm 3.39 was iterative, and generated a natural ordering of the faces $F \in \mathcal{F}^c$ and the elements $T \in \mathcal{T}$, namely $\mathcal{F}^c := \{F_1, \ldots, F_{n_{\mathcal{W}}}\}$ and $\mathcal{T} := \{T_1, \ldots, T_{n_{\mathcal{W}}}\}$ such that $F_1 \subset \overline{T}_1 \cap \Gamma_D$ and

$$F_k \subset \overline{T}_k \cap \left\{ \bigcup_{\ell=1}^{k-1} \overline{T}_\ell \right\}, \qquad \text{for all } k = 2, \ldots, n_{\mathcal{W}}. \tag{4.21}$$

Employing this ordering of the faces and elements the matrix A^c is lower triangular, as the following result shows.

Proposition 4.1. *The matrix* $A^c = [A^c_{k,k'}]_{n_{\mathcal{W}} \times n_{\mathcal{W}}}$ *with*

$$A^c_{k,k'} = B_{F_k, T_{k'}}, \qquad \text{for all } k, k' = 1, \ldots, n_{\mathcal{W}},$$

is lower triangular.

Proof. Let $k, k' = 1, \ldots, n_{\mathcal{W}}$ with $k < k'$. Then we know from (4.21) that $F_k \not\subset \overline{T}_{k'}$, and therefore it follows from (2.80) that $A^c_{k,k'} = B_{F_k, T_{k'}} = 0$, and the matrix A^c is lower triangular. □

Therefore, once the velocity $\mathbf{u}$ is known, we can calculate the pressure $\mathbf{p}$ in (4.2) by solving the lower triangular system (4.11) by *simple back substitution*. This concludes our discussion of the pressure system (4.11).

4.2 The velocity system in 2D

The bulk of the computational work in the decoupled method for (4.2) lies in the solution of the symmetric positive definite velocity system (4.7). We will show in this section (for 2D) and in the following section (for 3D) that (4.7) can be solved very efficiently by preconditioned conjugate gradients.

4.2.1 Implementation

Let $\Omega \subset \mathbb{R}^2$ (i.e. $d = 2$). To solve the decoupled system (4.7) for determining $\mathring{\mathbf{u}}$ (and hence $\mathbf{u}$) we must work with the matrix $\mathring{A}$ and right hand side $\mathring{\mathbf{g}}$ specified in (4.8). We observe that these are formally defined in terms of multiplications with the matrix Z which, through (4.14), represents the basis $\{\vec{\Psi}_i\}$ of $\mathring{\mathcal{V}}$ in terms of the basis $\{\vec{v}_j\}$ of $\mathcal{V}$. We will first only consider the case of Ω simply connected, $\Gamma_N = \Gamma_N^1 \cup \ldots \cup \Gamma_N^{s_N} \neq \emptyset$ and lowest order elements (i.e. $k = 0$), and we refer to Section 4.2.5 for possible extensions.

In the specific system (4.2) in 2D, the $\{\vec{v}_j\}$ are the standard Raviart-Thomas velocity basis functions. In the lowest order case $k = 0$, the $\{\vec{v}_j\}$ can be conveniently denoted by $\{\vec{v}_F : F \in \mathcal{F}_I \cup \mathcal{F}_D\}$ (as presented in Example 2.22), whereas the divergence-free basis functions $\{\vec{\Psi}_i\}$ are, as specified in Theorem 3.7,

$$\{\vec{\Psi}_P : P \in \mathcal{N}_I \cup \mathcal{N}_D\} \cup \{ \sum_{P \in \mathcal{N}_N^\ell} \vec{\Psi}_P : \ell = 1, \ldots, s_N - 1\}. \tag{4.22}$$

Thus we can identify the rows of Z with the indices $P \in \mathcal{N}_I \cup \mathcal{N}_D$ and $\ell = 1, \ldots, s_N - 1$, whereas the columns of Z correspond to $F \in \mathcal{F}_I \cup \mathcal{F}_D$.

Using this identification we can rewrite (4.14) as

$$\vec{\Psi}_P = \sum_{F \in \mathcal{F}_I \cup \mathcal{F}_D} Z_{P,F} \vec{v}_F, \qquad P \in \mathcal{N}_I \cup \mathcal{N}_D \tag{4.23}$$

and

$$\sum_{P \in \mathcal{N}_N^\ell} \vec{\Psi}_P = \sum_{F \in \mathcal{F}_I \cup \mathcal{F}_D} Z_{\ell,F} \vec{v}_F, \qquad \ell \in 1, \ldots, s_N - 1\,. \tag{4.24}$$

Note that the matrix Z is sparse, in fact $Z_{P,F} \neq 0$ only when node P is contained in face F (see Figure 3.3, left), whereas $Z_{\ell,F} \neq 0$ only when the Neumann boundary segment Γ_N^ℓ touches the face F (see Figure 3.3, right). To be precise, using (3.15), we have for all $P \in \mathcal{N}_I \cup \mathcal{N}_D$ that

$$Z_{P,F} := \begin{cases} \pm \frac{1}{|F|} & \text{for all } F \in \mathcal{F}_I \cup \mathcal{F}_D \text{ such that } P \in \overline{F}, \\ 0 & \text{otherwise,} \end{cases} \tag{4.25}$$

and for all $\ell = 1, \ldots, s_N - 1$ that

$$Z_{\ell,F} := \begin{cases} \pm \frac{1}{|F|} & \text{for all } F \in \mathcal{F}_I \cup \mathcal{F}_D \text{ such that } \overline{F} \cap \Gamma_N^\ell \neq \emptyset, \\ 0 & \text{otherwise,} \end{cases} \tag{4.26}$$

where the sign depends on the orientation of the normal $\vec{\nu}_F$ associated with F.

With these observations, it is simple to write $\mathring{A}$ as a sum of element matrices.

Elementwise representation

Recall (2.78), that in the lowest order case the elements $M_{F,F'}$ of the matrix M are identified with faces $F, F' \in \mathcal{F}_I \cup \mathcal{F}_D$. It is standard procedure in the application of finite element techniques, to write M as a sum of element matrices M_T, i.e.

$$M = \sum_{T \in \mathcal{T}} M_T, \qquad \text{where } (M_T)_{F,F'} = \int_T D^{-1}(\vec{x}) \vec{v}_F \cdot \vec{v}_{F'}\, d\vec{x}.$$

By Z_T we denote the matrix whose entries equal the entries of Z for columns and rows corresponding to T (i.e. faces $F \subset T$, nodes $P \in T$, and indices ℓ such that Γ_N^ℓ touches T) and are zero elsewhere. Then

$$\mathring{A} = \sum_{T \in \mathcal{T}} \mathring{A}_T, \qquad \text{with} \quad \mathring{A}_T := Z_T M_T Z_T^T. \tag{4.27}$$

The representation (4.27) is important, if one wishes to implement iterative methods for (4.7) using stiffness matrices defined only elementwise, and it means that the work which is necessary to calculate the reduced matrix $\mathring{A}$ from M is proportional to the number of elements in $\mathcal{T}$. A similar elementwise representation can be given for the computation of the load vector $\mathring{\mathbf{g}}$ in (4.8).

Alternatively, $\mathring{A}$ can be determined (elementwise or globally) from a standard piecewise linear approximation of a related bilinear form, without the assembly of any Raviart-Thomas stiffness matrix entries, as the following calculation shows.

Associated H^1-elliptic problem

First recall from (4.16), that by identifying the entries of $\mathring{A}$ with the indices $P, P' \in \mathcal{N}_I \cup \mathcal{N}_D$ and $\ell, \ell' = 1, \ldots, s_N - 1$, we have

$$\begin{aligned}
\mathring{A}_{P,P'} &= m(\vec{\Psi}_P\,,\,\vec{\Psi}_{P'}), \\
\mathring{A}_{P,\ell} &= m\Big(\vec{\Psi}_P\,, \textstyle\sum_{P'\in\mathcal{N}_N^\ell} \vec{\Psi}_{P'}\Big), \\
\mathring{A}_{\ell,\ell'} &= m\Big(\textstyle\sum_{P\in\mathcal{N}_N^\ell} \vec{\Psi}_P\,, \sum_{P'\in\mathcal{N}_N^{\ell'}} \vec{\Psi}_{P'}\Big).
\end{aligned} \tag{4.28}$$

Now, introduce the bilinear form

$$a(\Phi, \Phi') := \int_\Omega \mathcal{D}^{-1}(\vec{x})\vec{\nabla}\Phi \cdot \vec{\nabla}\Phi'\, d\vec{x}, \qquad \Phi, \Phi' \in H^1(\Omega), \tag{4.29}$$

where

$$\mathcal{D}(\vec{x}) := S^T D(\vec{x}) S, \qquad \text{and} \quad S := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{4.30}$$

(Note that $\mathcal{D}(\vec{x}) = D(\vec{x})$, when $D(\vec{x})$ is a scalar multiple of the identity.) Then, for $P, P' \in \mathcal{N}$, set

$$\mathcal{A}_{P,P'} := a(\Phi_P, \Phi_{P'}),$$

where $\{\Phi_P\}$ are the piecewise linear hat functions introduced in (3.13). Thus (after specifying an ordering of the nodes in $\mathcal{N}$), $\mathcal{A}$ is a standard finite element stiffness matrix corresponding to the bilinear form $a(\cdot,\cdot)$ with a natural boundary condition on all of Γ. Let A denote the minor of this matrix obtained by restricting to $P, P' \in \mathcal{N}_I \cup \mathcal{N}_D$, i.e.

$$A_{P,P'} := \mathcal{A}_{P,P'}, \qquad \text{for all} \quad P, P' \in \mathcal{N}_I \cup \mathcal{N}_D. \tag{4.31}$$

(This corresponds to imposing an essential boundary condition on Γ_N.) Moreover, define the matrices

$$C_{P,\ell} := \sum_{P'\in\mathcal{N}_N^\ell} \mathcal{A}_{P,P'}, \qquad P \in \mathcal{N}_I \cup \mathcal{N}_D, \quad \ell = 1, \ldots, s_N - 1$$

and

$$R_{\ell,\ell'} := \sum_{P\in\mathcal{N}_N^\ell} \sum_{P'\in\mathcal{N}_N^{\ell'}} \mathcal{A}_{P,P'}, \qquad \ell, \ell' = 1, \ldots, s_N - 1\ .$$

The following result identifies a simple formula for the matrix $\mathring{A}$. It shows that $\mathring{A}$

can be obtained from $\mathcal{A}$ using a small number of elementary operations on rows and columns corresponding to certain boundary nodes $P \in \mathcal{N}_N$.

Proposition 4.2. *Let Ω be simply connected and $\Gamma_N \neq \emptyset$. Then*

$$\mathring{A} = \begin{pmatrix} A & C \\ C^T & R \end{pmatrix}.$$

Proof. First observe that for all $P, P' \in \mathcal{N}$ we have by definition that

$$m(\vec{\Psi}_P, \vec{\Psi}_{P'}) = \int_\Omega D^{-1}(\vec{x})\vec{\Psi}_P \cdot \vec{\Psi}_{P'} d\vec{x} = \int_\Omega D^{-1}(\vec{x})\vec{\mathrm{curl}}\Phi_P \cdot \vec{\mathrm{curl}}\Phi_{P'} d\vec{x}, \tag{4.32}$$

where in the last step we used (3.14). Since $\vec{\mathrm{curl}}\Phi_P = S\vec{\nabla}\Phi_P$, and since $S^T D^{-1}(\vec{x})S = \mathcal{D}^{-1}(\vec{x})$, we have

$$m(\vec{\Psi}_P, \vec{\Psi}_{P'}) = \int_\Omega D^{-1}(\vec{x})S\vec{\nabla}\Phi_P \cdot S\vec{\nabla}\Phi_{P'} d\vec{x} = a(\Phi_P, \Phi_{P'}) = \mathcal{A}_{P,P'}. \tag{4.33}$$

The result then follows directly from (4.28) and from the definition of the matrices A, C, and R. □

Remark 4.3. Another interpretation of Proposition 4.2 is that $\mathring{A}$ is a piecewise linear finite element approximation of the bilinear form (4.29) with a natural boundary condition on Γ_D and a special type of essential boundary condition on Γ_N. The boundary condition on Γ_N is such that, for each $\ell = 1, \ldots, s_N - 1$, all the degrees of freedom on Γ_N^ℓ are constrained to be equal to a single freedom (to be found), whereas the boundary condition on $\Gamma_N^{s_N}$ forces all freedoms there to vanish. The test functions are chosen to be

$$\{\Phi_P : P \in \mathcal{N}_I \cup \mathcal{N}_D\} \cup \{\sum_{P \in \mathcal{N}_N^\ell} \Phi_P : \ell = 1, \ldots, s_N - 1\}.$$

4.2.2 Solution of bordered systems by block elimination

For general mixed boundary conditions, the coefficient matrix $\mathring{A}$ in (4.7) is a bordered matrix

$$\mathring{A} = \begin{pmatrix} A & C \\ C^T & R \end{pmatrix}. \tag{4.34}$$

with a sparse major block A, as shown in Proposition 4.2. The width n_C of the border (i.e. the number of columns in C) is problem dependent. On a simply connected, two-dimensional domain $\Omega \subset \mathbb{R}^2$ for example, $n_C := s_N - 1$ where s_N is the number of disconnected components in the Neumann boundary Γ_N (see Section 4.2.1).

For most applications n_C will be small and in all applications it does not increase as the mesh is refined (i.e. $n_C \neq n_C(h)$). Thus, it is reasonable to consider solving

(4.7) by block elimination. More precisely, if we write (4.7) as

$$\begin{pmatrix} A & C \\ C^T & R \end{pmatrix} \begin{pmatrix} \mathring{\mathbf{u}}^{(A)} \\ \mathring{\mathbf{u}}^{(C)} \end{pmatrix} = \begin{pmatrix} \mathring{\mathbf{g}}^{(A)} \\ \mathring{\mathbf{g}}^{(C)} \end{pmatrix}, \tag{4.35}$$

we first compute a matrix X and a vector $\mathbf{y}$ satisfying

$$AX = C, \qquad\qquad A\mathbf{y} = \mathring{\mathbf{g}}^{(A)}. \tag{4.36}$$

Then $\mathring{\mathbf{u}}^{(C)}$ is the solution of

$$(R - C^T X)\, \mathring{\mathbf{u}}^{(C)} = (\mathring{\mathbf{g}}^{(C)} - C^T \mathbf{y}), \tag{4.37}$$

from which we obtain

$$\mathring{\mathbf{u}}^{(A)} = \mathbf{y} - X\mathring{\mathbf{u}}^{(C)}. \tag{4.38}$$

When C has a small number of columns, (4.36) constitutes a small number of systems ($n_C + 1$ to be precise) with coefficient matrix A defined in (4.31), which can be efficiently solved iteratively, as we will see in the analysis in Section 4.2.3. The system (4.37), on the other hand, is a small system of dimension $n_C \times n_C$ and can be solved directly.

4.2.3 Analysis

Let us now analyse the complexity and conditioning of the decoupled velocity system (4.7). Here we make use of the formula for $\mathring{A}$ given in Proposition 4.2. Let $\{\mathcal{T}_h\}$ be a shape regular family of triangulations of Ω. We consider only the case of Ω simply connected and $\Gamma_N \neq \emptyset$ in detail and restrict attention to lowest order elements (i.e. $k = 0$), but we refer to Section 4.2.5 for possible extensions.

The coefficient matrix $\mathring{A}$ in (4.7) is a bordered matrix with major block consisting of the standard piecewise linear finite element stiffness matrix A defined in (4.31), and with the width of the border $s_N - 1$ where s_N is the number of disconnected components in the Neumann boundary Γ_N. If $s_N = 1$, then $\mathring{A} = A$. In general, systems of this form can be solved by standard block elimination algorithms using s_N solves with A, as presented in Section 4.2.2. Thus, the complexity and conditioning analysis of $\mathring{A}$ can be reduced to an analysis of its major block A. Since A is probably the most extensively studied finite element matrix, almost any standard finite element book contains an analysis of it (see for example [50, 63]). Here we focus on comparing the properties of A to the properties of the matrix $\mathcal{M} = \begin{pmatrix} M & B \\ B^T & 0 \end{pmatrix}$ in the original coupled system (4.2).

Complexity

First of all, observe that the matrix $\mathring{A}$ and therefore the major block A in the decoupled system (4.7) is about 5 times smaller than the matrix $\mathcal{M}$ in the original coupled system

(4.2). More precisely, the dimension of (4.7) is smaller than that of (4.2) by a factor

$$\mathcal{C} := \frac{\#\mathcal{F}_I + \#\mathcal{F}_D + \#\mathcal{T}_h}{\#\mathcal{F}_I + \#\mathcal{F}_D - \#\mathcal{T}_h}.$$

Since $3\#\mathcal{T}_h = 2\#\mathcal{F}_I + \#\mathcal{F}_D + \#\mathcal{F}_N$, we have

$$\mathcal{C} = 5\left\{\frac{\#\mathcal{F}_I + \frac{4}{5}\#\mathcal{F}_D + \frac{1}{5}\#\mathcal{F}_N}{\#\mathcal{F}_I + 2\#\mathcal{F}_D - \#\mathcal{F}_N}\right\}.$$

Because of the assumed shape regularity of the triangulation, $\#\mathcal{F}_I$ is the dominant part of $\#\mathcal{F}$ as $h \to 0$, and so $\mathcal{C} \to 5$ as $h \to 0$.

Next, let us look at the sparsity of A and $\mathcal{M}$. We shall show that the average number $NZE(A)$ of nonzero entries of A per row is about 7. More precisely, let $P \in \mathcal{N}$. Without loss of generality we assume that $\Gamma_N = \emptyset$. Then $A_{P,P'} \neq 0$, only if $P' = P$ or if there exists a face $F \in \mathcal{F}$ such that P and P' are the end points of F (see the definition of Φ_P in (3.13)). Therefore the total number of nonzero entries of A is $2\#\mathcal{F} + \#\mathcal{N}$, and since the number of rows in A is $\#\mathcal{N}$, we have

$$NZE(A) = \frac{2\#\mathcal{F}}{\#\mathcal{N}} + 1, \tag{4.39}$$

Now, using Proposition 3.9 we have

$$\#\mathcal{N} = \#\mathcal{F} - \#\mathcal{T}_h + 1 = \frac{1}{3}\mathcal{F}_I + \frac{2}{3}\#\mathcal{F}_D + 1, \tag{4.40}$$

where we have used the fact that $3\#\mathcal{T}_h = 2\#\mathcal{F}_I + \#\mathcal{F}_D$. Then, substituting (4.40) into (4.39) we get

$$NZE(A) = \frac{2(\#\mathcal{F}_I + \#\mathcal{F}_D)}{\frac{1}{3}\#\mathcal{F}_I + \frac{2}{3}\#\mathcal{F}_D + 1} + 1 = 6\left\{\frac{\#\mathcal{F}_I + \#\mathcal{F}_D}{\#\mathcal{F}_I + 2\#\mathcal{F}_D + 3}\right\} + 1.$$

Again, because of the assumed shape regularity of the triangulation, $\#\mathcal{F}_I$ is the dominant part of $\#\mathcal{F}$ as $h \to 0$, and so $NZE(A) \to 7$ as $h \to 0$. Note that this result does not require a quasi-uniform family of triangulations. Even if we do not assume shape regularity of $\{\mathcal{T}_h\}$, and if $\#\mathcal{F}_I$ is not the dominant part of $\#\mathcal{F}$ as $h \to 0$, we have still $NZE(A) \leq 7$.

In comparison, the average number $NZE(\mathcal{M})$ of nonzero entries of $\mathcal{M}$ per row is about 5.4. More precisely, let $F \in \mathcal{F}$. Without loss of generality we assume that $\Gamma_N = \emptyset$. Then $M_{F,F'} \neq 0$, only if there exists an element $T \in \mathcal{T}_h$ such that $F \subset \overline{T}$ and $F' \subset \overline{T}$ (see the definition of $\vec{v}_F$ in (2.75)). If $F \in \mathcal{F}_I$, there are 5 such faces F'. If $F \in \mathcal{F}_D$, on the other hand, there are 3. Therefore the total number of nonzero entries of M is $5\#\mathcal{F}_I + 3\#\mathcal{F}_D$. Now let $T \in \mathcal{T}_h$. Then $B_{F,T} \neq 0$, if $F \subset \overline{T}$, and the total number of nonzero entries of B is $3\#\mathcal{T}_h$. Finally, since the number of rows in $\mathcal{M}$ is

$\#\mathcal{F}_I + \#\mathcal{F}_D + \#\mathcal{T}_h$, we have

$$NZE(\mathcal{M}) = \frac{5\#\mathcal{F}_I + 3\#\mathcal{F}_D + 6\#\mathcal{T}_h}{\#\mathcal{F}_I + \#\mathcal{F}_D + \#\mathcal{T}_h} = \frac{27}{5}\left\{\frac{\#\mathcal{F}_I + \frac{1}{3}\#\mathcal{F}_D}{\#\mathcal{F}_I + \frac{4}{5}\#\mathcal{F}_D}\right\}.$$

Therefore, $NZE(\mathcal{M}) \to 5.4$ as $h \to 0$, and we see that the average number of nonzero elements of $\mathcal{M}$ per row is only slightly lower than for A (recall $NZE(A) \to 7$).

Conditioning

Now, let us compare the condition numbers of A and $\mathcal{M}$. We will consider only quasi-uniform families of triangulations $\{\mathcal{T}_h\}$ in detail, where h denotes the maximum diameter of the elements $T \in \mathcal{T}_h$. In this case we know from Corollary 2.28 that the spectral condition number of $\mathcal{M}$ satisfies

$$\kappa(\mathcal{M}) \leq Ch^{-1}. \tag{4.41}$$

Estimates for the spectral condition number of A are well known. In Johnson [63, Section 7.7] it is shown that under the same regularity assumptions on the triangulation, we have

$$\kappa(A) \leq Ch^{-2}. \tag{4.42}$$

Both estimates are sharp, and they show that asymptotically the coupled matrix $\mathcal{M}$ is better conditioned than A. However, since A is symmetric positive definite, we can apply preconditioned conjugate gradients, and a range of optimal preconditioners are available which ensure in theory that the number of iterations does not grow as $h \to 0$, and which are very robust with respect to jumps in $\mathcal{D}^{-1}(\vec{x})$. In Section 4.2.4 we present a parallel implementation where the number of iterations grows with $O(h^{-1/3})$ and logarithmically with the largest jump in $\mathcal{D}^{-1}(\vec{x})$. To solve the coupled system (4.2) on the other hand, we would have to fall back on the methods presented in Section 2.3.3, such as MINRES (see Figure 2.2). Here (in the unpreconditioned case), the number of iterations grows with the condition number of the matrix (i.e. $O(h^{-1})$) and optimal preconditioners often require further restrictions on the domain, the triangulation or the boundary conditions.

Remark 4.4. In the general case of a shape regular family of triangulations $\{\mathcal{T}_h\}$, which is not necessarily quasi-uniform (this might be the case when adaptive refinement is used), the estimate (4.42) has to be modified. In Bank & Scott [12] it is shown that in this case

$$\kappa(A) \leq C\, n\Big(1 + |\log(n\, h_{min}^2)|\Big),$$

where h_{min} is the minimum mesh diameter and n is the size of A, i.e. $n = \#\mathcal{N}_I + \#\mathcal{N}_D$.

So from several points of view the reduction to the decoupled symmetric positive definite velocity system (4.7) makes practical sense.

4.2.4 Parallel iterative solution

In this section we briefly describe our parallel solver for the velocity systems (4.36) arising in Section 4.2.2 with coefficient matrix A (see Sections 4.2.1 and 4.2.3 for the definition and analysis of A). Our method is based on the conjugate gradient algorithm with additive Schwarz preconditioner and uses the implementation provided by the `DOUG` package (Hagger [51], Hagger & Stals [52]) for general unstructured systems.

Let $\{\mathcal{T}_h\}$ be a shape regular family of simplicial triangulations of Ω. The first step in our parallelisation involves the partition of the domain Ω (in this case using the mesh partitioning software `METIS` - Karypis & Kumar [65]) into non-overlapping connected subdomains $\Omega_i,\ i = 1, \ldots, S$, each consisting of a union of elements $T \in \mathcal{T}_h$. The `METIS` software strives to ensure that the Ω_i are of comparable size ("load-balancing") and the interfaces between them contain as few faces as possible (to minimise communication). These subdomains are then used for parallelisation of the vector-vector and matrix-vector operations required in the conjugate gradient algorithm. Good parallel efficiency is achieved for matrix-vector products by ensuring that the necessary communication of boundary data between neighbouring subdomains is overlapped with computations in the (independent) subdomain interiors.

For preconditioning we use the unstructured version of the classical two-level additive Schwarz method (e.g. Chan et al. [23]) which has the general form

$$\mathcal{P}_{AS(2)}^{-1} := R_H^T A_H^{-1} R_H + \sum_{i=1}^{S} R_i^T A_i^{-1} R_i. \tag{4.43}$$

In (4.43) the matrices A_i^{-1} represent local solves of the underlying PDE on overlapping extensions $\tilde{\Omega}_i$ of the Ω_i with a homogeneous Dirichlet condition imposed on the parts of $\partial\tilde{\Omega}_i$ which do not intersect with the boundary Γ. The restriction operator R_i is taken to be the simple injection operator.

In our particular implementation of (4.43), $\tilde{\Omega}_i$ is constructed by adding to each Ω_i all the elements $T \in \mathcal{T}_h$ which touch its boundary $\partial\Omega_i$. The resulting extended subdomains $\tilde{\Omega}_i$ then have overlap δ, say, with δ bounded above and below by the maximum and minimum diameter (respectively h and h_{min}) of all the elements $T \in \mathcal{T}_h$. This choice of overlap represents a compromise between the competing demands of condition number optimality and efficiency of the parallelisation (the former requiring, at least in theory, a reasonable overlap and the latter requiring that the overlap should be as small as possible). This choice also means that A_i is simply the minor of A obtained by removing all the rows and columns corresponding to nodes not on $\Omega_i \cup \partial\Omega_i$.

In the present version of the `DOUG` package the subdomain solves A_i^{-1} are done using a direct frontal solver and so, to achieve good efficiency, the underlying subdomains should not become too large. In `DOUG` the default size is 1000 degrees of freedom (and this is what we use in the numerical experiments later on). Since the package is designed to run on any number of processors, we allow the possibility that each processor will handle several subdomains.

The preconditioner (4.43) also contains a coarse grid solve, A_H^{-1}, which handles the global interaction of the subdomains. This distinguishes (4.43) from block-Jacobi-like methods and is essential for the construction of optimal preconditioners (see Dryja & Widlund [36]). There is no need for the coarse mesh $\mathcal{T}_H$ to be related directly to the fine mesh, but in principle it should be capable of representing the solution of the underlying PDE with appropriate accuracy. What this means in practice is that, if one has constructed a fine mesh which provides a sufficiently good resolution of the underlying problem, then one requires also a coarse mesh with the same qualitative properties at the coarser level. Such a coarsening may sometimes be available (e.g. from an earlier stage of a refinement process) but, since this is not always the case, the `DOUG` package produces a coarsening automatically. For this, an adaptive piecewise uniform strategy is used, the efficiency of which is discussed in detail in Hagger [51]. In our implementation of (4.43) the operator R_H^T denotes piecewise linear interpolation from coarse to fine mesh, R_H denotes its transpose and A_H is the Galerkin product $A_H = R_H^T A R_H$.

In the present version of `DOUG` the coarse mesh problem is assembled and solved directly using the frontal method on a master processor. In order to maintain efficient parallelisation, the time for this should not exceed the time which is being taken by the processors which are working on the subdomain solves. If n denotes the total number of degrees of freedom in the problem and n_P is the number of processors then (assuming load balancing) each processor has to solve $n/(1000 * n_P)$ problems, each with 1000 unknowns. The cost of a frontal solve for a finite element problem with N degrees of freedom (in 2D) is about $8N^{3/2}$ (see the references in Hagger [51]). Thus for parallel efficiency the dimension of the coarse grid problem n_H is chosen in `DOUG` to satisfy

$$n_H^{3/2} = \left(\frac{n}{(1000 * n_p)} \right) * 1000^{3/2}, \tag{4.44}$$

i.e. the cost of solving the coarse grid problem equals the cost of solving the subproblems on each processor. Note that for a fixed n_P this implies that $n_H = O(n^{2/3})$.

The asymptotic performance of the preconditioner (4.43) is analysed in Chan et al. [23], where it is shown that for general symmetric positive definite problems

$$\kappa(\mathcal{P}_{AS(2)}^{-1} A) = O\left(\left(\frac{H}{\delta} \right)^2 \right), \qquad \text{as } H, h \to 0 \tag{4.45}$$

where κ denotes the spectral condition number as defined in (2.83) again, h, H denote the fine and coarse mesh diameters, and δ denotes the overlap in the subdomains $\tilde{\Omega}_i$.

Then with the `DOUG` code as described above applied to a problem on a quasi-uniform family $\{\mathcal{T}_h\}$ of triangulations $n = O(h^{-2})$, and therefore the overlap is $\delta = O(h) = O(n^{-1/2})$. The family $\{\mathcal{T}_H\}$ of coarse triangulations produced by `DOUG` will also be quasi-uniform and will therefore have $n_H = O(n^{2/3})$ degrees of freedom and diameter $H = O(n_H^{-1/2}) = O(n^{-1/3})$. The estimate (4.45) then reduces to

$$\kappa(\mathcal{P}_{AS(2)}^{-1} A) = O((n^{1/2} n^{-1/3})^2) = O(n^{1/3})$$

and the number of iterations of the conjugate gradient method will grow no faster than $O(n^{1/6})$. We examine numerically in Section 4.5.1 the sharpness of this estimate (see also Section 5.2.4).

We shall also discuss the performance of this method in the presence of very rough coefficients. A lot is known about this case provided the jumps occur on a coarser scale than the fine mesh being used to compute the solution. In the case of certain two-level domain decomposition methods on structured meshes, for example, the effect of the jumps can be removed completely provided the coarse mesh resolves the jumping regions. In the unstructured case this is no longer true, indeed the preconditioned problem may be just as ill-conditioned as the original matrix as the jumps worsen. An example showing this was given in Graham & Hagger [45, 46], where it is also shown that the condition number is not a very good guide in this case to the behaviour of the preconditioned conjugate gradient (PCG) method, since the preconditioned problem has only a small cluster of eigenvalues near the origin with the others lying in a bounded region away from the origin as the jumps get worse. The general proof of this phenomenon led in Graham & Hagger [45, 46] to the proof that the corresponding PCG method in fact is very resilient to the existence of jumping coefficients even in the unstructured case. Roughly speaking [46] shows that in the case of a piecewise constant coefficient $\mathcal{D}^{-1}(\vec{x})$ with respect to a fixed number of regions of the domain, the number of PCG iterations will grow only logarithmically in the quantity $\frac{\max|\mathcal{D}^{-1}(\vec{x})|}{\min|\mathcal{D}^{-1}(\vec{x})|}$, whereas the condition number $\kappa(\mathcal{P}_{AS(2)}^{-1} A)$ itself generally grows linearly in $\frac{\max|\mathcal{D}^{-1}(\vec{x})|}{\min|\mathcal{D}^{-1}(\vec{x})|}$. The numerical results in Section 4.5.1 will also confirm the sharpness of this estimate.

It is important to note though, that the results in Graham & Hagger [45, 46] apply when the jumping coefficient varies on a coarser scale than the fine mesh and so they do not strictly apply to the case of the heterogeneous media considered at the end of this thesis in Section 5.2.4, where the coefficient varies on the fine mesh scale. However, interestingly, the numerical results given there indicate that in some sense the results of [45, 46] hold true even in this extreme case, although at the time of writing we know of no proof of this.

4.2.5 Extensions

The results in this section carry over in a straightforward way to higher order elements, i.e. $k > 0$. In the same way as for $k = 0$, we can write $\mathring{A}$ as a sum of element matrices, and by using the functions $\vec{\Psi}_P \in \mathcal{V}$, with $P \in \Sigma_{k+1}$, defined in (3.29), we see as in the proof to Proposition 4.2 that

$$m(\vec{\Psi}_P, \vec{\Psi}_{P'}) = a(\Phi_P, \Phi_{P'}), \qquad \text{for all } P, P' \in \Sigma_{k+1},$$

where $\{\Phi_P : P \in \Sigma_{k+1}\}$ is the canonical basis of the finite element space $\mathcal{S}_{k+1}(\Omega, \mathcal{T})$ in $H^1(\Omega)$. Therefore, using Theorem 3.13 instead of Theorem 3.7, $\mathring{A}$ can again be written as in Proposition 4.2, as a bordered matrix with the width of the border $s_N - 1$. The major block now consists of the stiffness matrix A corresponding to a higher order approximation of the H^1-elliptic bilinear form $a(\cdot, \cdot)$ in (4.29) by C^0-elements (with an essential boundary condition on Γ_N and a natural boundary condition on Γ_D).

In comparison, the decoupled system (4.7) is now about $(3k+5)/(k+1)$ times smaller than the original system (4.2). This can be shown as for $k = 0$ by using the dimensions (2.51) and (2.53) of $\mathcal{V}$ and $\mathcal{W}$ and the fact that $3\#\mathcal{T} = 2\#\mathcal{F}_I + \#\mathcal{F}_D + \#\mathcal{F}_N$. So even if k gets large, the decoupled system is going to be at least 3 times smaller than the original system. The condition number of the major block A of $\mathring{A}$ is still $\kappa(A) = O(h^{-2})$ for quasi-uniform families of triangulations, and the coupled matrix $\mathcal{M}$ still does have a better condition number ($O(h^{-1})$ in fact), but as before this disadvantage is more than made up for by the positivity of the spectrum and the availability of efficient preconditioners.

The results in this section also extend to the case of multiply connected domains, if we use the basis for $\mathring{\mathcal{V}}$ found in Section 3.1.4. As discussed before, this might in general involve the introduction of further non-local basis functions like $\vec{\Psi}_{2,3}$ in Theorem 3.17, and therefore additional borders in the representation given in Proposition 4.2. However, the number of such additional functions is small, and they can be dealt with in the same way as before.

Finally, the results also extend to the case $s_N = 0$, i.e. the pure Dirichlet case, if we use the basis $\{\vec{\Psi}_P : P \in \Sigma_{k+1},\ P \neq P_0\}$ for $\mathring{\mathcal{V}}$ given in Remark 3.14(a). The matrix $\mathring{A}$ is then obtained from $\mathcal{A}$ by deleting the row and column of $\mathcal{A}$ corresponding to P_0. This corresponds to imposing an artificial essential boundary condition at P_0, to eliminate the singularity of $\mathcal{A}$.

4.3 The velocity system in 3D

Now let $\Omega \subset \mathbb{R}^3$ (i.e. $d = 3$). Many of the issues discussed in the previous section for $d = 2$, in particular the structure of (4.7), is very similar to the two-dimensional case. However, the resulting system is different and its solution and analysis are much harder.

4.3.1 Implementation

To identify the structure of the decoupled system (4.7) we must again work with the matrix Z which, through (4.14), represents the basis $\{\vec{\Psi}_i\}$ of $\mathring{\mathcal{V}}$ in terms of the basis

$\{\vec{v}_j\}$ of $\mathcal{V}$. In view of Section 3.2, we consider only lowest order elements $k = 0$ and simply connected domains Ω with connected boundary Γ. The $\{\vec{v}_j\}$ are again the lowest order Raviart-Thomas-Nédélec velocity basis functions $\{\vec{v}_F : F \in \mathcal{F}_I \cup \mathcal{F}_D\}$ as presented in Example 2.22, whereas the $\{\vec{\Psi}_i\}$ are now the basis functions constructed for 3D in Section 3.2.

If we assume (for the moment) that each component of Γ_N is simply connected, then the basis $\{\vec{\Psi}_i\}$ is given by $\{\vec{\Psi}_E : E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}\}$ as specified in Corollary 3.32. We will come back to the general case when one of the components of Γ_N is not simply connected in Section 4.3.3. Thus, if we assume for the moment that the set $\mathcal{H}$ is known, we can identify the columns of Z with the indices $F \in \mathcal{F}_I \cup \mathcal{F}_D$, whereas the rows of Z correspond to $E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}$.

Using this identification of Z we can now rewrite (4.14) as

$$\vec{\Psi}_E = \sum_{F \in \mathcal{F}_I \cup \mathcal{F}_D} Z_{E,F}\, \vec{v}_F, \qquad \text{for } E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}. \tag{4.46}$$

Note that the matrix Z is sparse, in fact $Z_{E,F} \neq 0$ only when edge E is an edge of the face F. To be exact, using (3.63), we have for all $E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}$ that

$$Z_{E,F} := \begin{cases} \pm\frac{1}{|F|} & \text{for all } F \in \mathcal{F}_I \cup \mathcal{F}_D \text{ such that } E \subset \overline{F}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.47}$$

where the sign depends on the orientation of the normal $\vec{\nu}_F$ associated with F and the tangent $\vec{\tau}_E$ associated with E.

With these observations, it is simple to write $\mathring{A}$ as a sum of element matrices, in the same way as presented for $d = 2$ in the previous section. In fact

$$\mathring{A} = \sum_{T \in \mathcal{T}} \mathring{A}_T, \qquad \text{with} \quad \mathring{A}_T := Z_T M_T Z_T^T, \tag{4.48}$$

where Z_T and M_T are as before the element matrices on $T \in \mathcal{T}$. Thus, the work which is necessary to calculate the reduced matrix $\mathring{A}$ from M is proportional to the number of elements in $\mathcal{T}$. A similar elementwise representation can be given for the computation of the load vector $\mathring{\mathbf{g}}$ in (4.8).

Alternatively, $\mathring{A}$ can be determined (element-wise or globally) from an approximation of a related bilinear form by Nédélec's edge elements, without the assembly of any Raviart-Thomas-Nédélec stiffness matrix entries, as the following calculation shows.

Associated bilinear form in $H(\vec{\text{curl}}, \Omega)$

First recall that from (4.8) we have $\mathring{A} = ZMZ^T$, and that we can therefore identify the entries of $\mathring{A}$ with the indices $E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}$, and rewrite (4.16) as

$$\mathring{A}_{E,E'} = m(\vec{\Psi}_E, \vec{\Psi}_{E'}), \qquad E, E' \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}. \tag{4.49}$$

Now, introduce the bilinear form

$$a(\vec{\Phi}, \vec{\Phi}') := \int_\Omega D^{-1}(\vec{x}) \vec{\mathrm{curl}} \vec{\Phi} \cdot \vec{\mathrm{curl}} \vec{\Phi}' \, d\vec{x}, \qquad \text{for all } \vec{\Phi}, \vec{\Phi}' \in H(\vec{\mathrm{curl}}, \Omega), \tag{4.50}$$

and, for $E, E' \in \mathcal{E}$, set

$$\mathcal{A}_{E,E'} := a(\vec{\Phi}_E, \vec{\Phi}_{E'}), \tag{4.51}$$

where $\{\vec{\Phi}_E\}$ are the basis functions of the piecewise linear Nédélec's edge elements defined in Example 3.27. Thus (after specifying an ordering of the edges in $\mathcal{E}$), $\mathcal{A}$ is the stiffness matrix corresponding to the bilinear form $a(\cdot, \cdot)$ discretised by Nédélec's edge elements, with a natural boundary condition on all of Γ. Because of the non-trivial kernel of the $\vec{\mathrm{curl}}$, as illustrated in Proposition 3.24, the bilinear form $a(\cdot, \cdot)$ is degenerate, and therefore not elliptic on $H(\vec{\mathrm{curl}}, \Omega)$. In fact, let $v \in H^1(\Omega)$, then $a(\vec{\nabla} v, \vec{\Phi}') = 0$, for all $\vec{\Phi}' \in H(\vec{\mathrm{curl}}, \Omega)$. Consequently, $\mathcal{A}$ is singular.

The following result shows that the minor A of this matrix obtained by restricting to $E, E' \in (\mathcal{E}_I \cup \mathcal{E}_D) \backslash \mathcal{H}$, where $\mathcal{H} \subset \mathcal{E}$ as defined in Corollary 3.32, i.e.

$$A_{E,E'} = \mathcal{A}_{E,E'} \qquad \text{for all } E, E' \in (\mathcal{E}_I \cup \mathcal{E}_D) \backslash \mathcal{H}, \tag{4.52}$$

determines the matrix $\mathring{A}$ in (4.7). (This corresponds to imposing an essential boundary condition on Γ_N and restricting to the orthogonal complement of the kernel of $\vec{\mathrm{curl}}$.)

Proposition 4.5. *Let Γ_N and $\mathcal{H} \subset \mathcal{E}$ be as defined in Corollary 3.32. Then*

$$\mathring{A} = A.$$

Proof. First observe that for all $E, E' \in \mathcal{E}$, using (3.60), we have by definition that

$$m(\vec{\Psi}_E, \vec{\Psi}_{E'}) = \int_\Omega D^{-1}(\vec{x}) \vec{\Psi}_E \cdot \vec{\Psi}_{E'} \, d\vec{x} = \int_\Omega D^{-1}(\vec{x}) \vec{\mathrm{curl}} \vec{\Phi}_E \cdot \vec{\mathrm{curl}} \vec{\Phi}_{E'} \, d\vec{x} = a(\vec{\Phi}_E, \vec{\Phi}_{E'}).$$

The result then follows directly from (4.49) and from the definition of A. ☐

In the same way we can identify the rows of the load vector $\mathring{g}$ in (4.7) with the indices $E \in (\mathcal{E}_I \cup \mathcal{E}_D) \backslash \mathcal{H}$, and rewrite (4.17) as

$$\mathring{g}_E = F_{\mathcal{D}}(\vec{\mathrm{curl}} \vec{\Phi}_E), \qquad E \in (\mathcal{E}_I \cup \mathcal{E}_D) \backslash \mathcal{H}. \tag{4.53}$$

Remark 4.6. Hiptmair & Hoppe [59] solve the singular system

$$a(\vec{U}, \vec{\Phi}_E) = F_{\mathcal{D}}(\vec{\mathrm{curl}} \vec{\Phi}_E), \qquad \text{for all} \quad E \in \mathcal{E}_I \cup \mathcal{E}_D,$$

with symmetric positive semidefinite stiffness matrix $\mathcal{A}$ by multilevel preconditioned conjugate gradients without explicitly eliminating columns and rows corresponding to edges $E \in \mathcal{H}$. In their multilevel splitting, they eliminate the kernel $\ker(\vec{\mathrm{curl}})$ of $\vec{\mathrm{curl}}$

only approximately by relaxing the orthogonality with respect to $\ker(\vec{\mathrm{curl}})$ and thus avoid the construction of a basis. Here we eliminate $\ker(\vec{\mathrm{curl}})$ *a priori*, which allows us to then apply the conjugate gradient algorithm with a range of possible preconditioners.

Construction of the spanning tree $\mathbf{H} := (\mathcal{N}, \mathcal{H})$

It remains to discuss, how the set $\mathcal{H}$ can be constructed efficiently. The set $\mathcal{H}$ is a subset of edges $E \in \mathcal{E}$ that form a spanning tree $\mathbf{H} := (\mathcal{N}, \mathcal{H})$ in the graph $\mathbf{G} := (\mathcal{N}, \mathcal{E})$ underlying the triangulation $\mathcal{T}$ of Ω. Such a spanning tree can be found in optimal time, i.e. proportional to the number of edges or, equivalently, the number of nodes, using Algorithm B.7 presented in Appendix B. For mixed boundary conditions, i.e. $s_N \neq 0$, we had to pose an extra condition on the spanning tree $\mathbf{H}$ in Corollary 3.32, namely that for each $\ell = 1, \ldots, s_N - 1$, the restriction $\mathbf{H}_N^\ell := (\mathcal{N}_N^\ell, \mathcal{H} \cap \mathcal{E}_N^\ell)$ of $\mathbf{H}$ to the component Γ_N^ℓ of Γ_N is also a tree. This particular spanning tree can still be calculated, again in optimal time, using a slightly modified version of Algorithm B.7.

Since we assumed that Γ_N^ℓ is connected, each of the graphs $\mathbf{G}_N^\ell := (\mathcal{N}_N^\ell, \mathcal{E}_N^\ell)$ is connected, and we can use Algorithm B.7 to find a spanning tree $\mathbf{H}_N^\ell := (\mathcal{N}_N^\ell, \mathcal{H}_N^\ell)$ for each of them, by restricting only to vertices $y \in \mathcal{N}_N^\ell$. The union $\mathcal{H}_N := \bigcup_{\ell=1}^{s_N-1} \mathcal{H}_N^\ell$ is a cycle-free set of edges in the graph $\mathbf{G} := (\mathcal{N}, \mathcal{E})$, and since $\mathbf{G}$ is connected, we can extend $\mathcal{H}_N$ to a cycle-free set $\mathcal{H}$ of edges containing $\#\mathcal{N} - 1$ edges. It follows from Theorem B.3(ii) that $\mathbf{H} := (\mathcal{N}, \mathcal{H})$ is a spanning tree of $\mathbf{G}$. Since $\mathcal{H}_N \subset \mathcal{H}$, the restrictions $\mathbf{H}_N^\ell := (\mathcal{N}_N^\ell, \mathcal{H} \cap \mathcal{E}_N^\ell)$ of $\mathbf{H}$ to each component Γ_N^ℓ are also trees, as required.

Applying this strategy we can modify Algorithm B.7 in the following way: we choose $x_1 \in \mathcal{N}_N$ (in line `7`), and at first consider only neighbouring nodes $y \in \mathcal{N}_N$ (in line `15`), to find a spanning tree $\mathbf{H}_N^\ell$ for each Γ_N^ℓ; then (without resetting the array `mark[.]`) we call the function `recursive`(x_1) again with the same argument $x_1 \in \mathcal{N}_N$ and now consider $y \in \mathcal{N}_I \cup \mathcal{N}_D$ (in line `15`), to find the rest of the spanning tree.

4.3.2 Analysis

We now give a complexity and conditioning analysis of the decoupled velocity system (4.7) in the light of Proposition 4.5. Let $\{\mathcal{T}_h\}$ be a shape regular family of triangulations of Ω. Again, we consider only the case where each component of Γ_N is simply connected in detail, but we will come back to more general situations in Section 4.3.3.

Complexity

We observe first of all that in 3D the decoupled system (4.7) is about 3 times smaller than the original coupled system (4.2). More precisely, the dimension of (4.7) is smaller than that of (4.2) by a factor

$$\mathcal{C} := \frac{\#\mathcal{F}_I + \#\mathcal{F}_D + \#\mathcal{T}_h}{\#\mathcal{F}_I + \#\mathcal{F}_D - \#\mathcal{T}_h}.$$

Now, in 3D, since $4\#\mathcal{T}_h = 2\#\mathcal{F}_I + \#\mathcal{F}_D + \#\mathcal{F}_N$, we have

$$\mathcal{C} = 3\left\{\frac{\#\mathcal{F}_I + \frac{5}{6}\#\mathcal{F}_D + \frac{1}{6}\#\mathcal{F}_N}{\#\mathcal{F}_I + \frac{3}{2}\#\mathcal{F}_D - \frac{1}{2}\#\mathcal{F}_N}\right\}.$$

Because of the assumed shape regularity of the triangulation, $\#\mathcal{F}_I$ is again the dominant part of $\#\mathcal{F}$ as $h \to 0$, and so $\mathcal{C} \to 3$ as $h \to 0$.

Conditioning

Unfortunately, the matrix $\mathring{A}$ in the decoupled system (4.7) does not take such a simple and well understood form in 3D as in 2D. Nevertheless, $\mathring{A}$ is again symmetric positive definite and because of the interpretation of $\mathring{A}$ in terms of the bilinear form $a(\cdot,\cdot)$ defined in (4.50), the system (4.7) still behaves like a second order elliptic system. The most problematic part in proving this, is obviously the ellipticity which is enforced algebraically, using the spanning tree $\mathbf{H} = (\mathcal{N}, \mathcal{H})$.

Recall that $\{\mathcal{T}_h\}$ is a shape regular family of triangulations of Ω, and let $h(T)$ denote the diameter of an element $T \in \mathcal{T}_h$. The maximum and minimum diameter of any of the elements $T \in \mathcal{T}_h$ are denoted by h and h_{min} again. Furthermore, for each h, let $\mathbf{H}_h := (\mathcal{N}_h, \mathcal{H}_h)$ be a spanning tree for the graph associated with the triangulation $\mathcal{T}_h$, where $\mathcal{N}_h$ denotes the set of nodes in $\mathcal{T}_h$. To theoretically prove that (4.7) behaves like a second order elliptic system, we would need to establish a Poincaré-type inequality

$$\|\vec{u}\|^2_{(L_2(\Omega))^3} \leq \alpha \|\vec{\operatorname{curl}}\,\vec{u}\|^2_{(L_2(\Omega))^3}, \qquad \text{for all } \vec{u} \in \mathring{\mathcal{U}}, \tag{4.54}$$

with α independent of h, where

$$\mathring{\mathcal{U}} := \operatorname{span}\left\{\vec{\Phi}_E : E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}_h\right\}. \tag{4.55}$$

At present, the inequality (4.54) still remains unproved for simplicial triangulations. However, in Cai et al. [21, Lemma 4.1], it is shown for a particular family $\{\mathbf{H}_h\}$ of spanning trees, which can be chosen a priori, that (4.54) holds true for hexahedral Nédélec elements on uniform, rectangular meshes, and there is strong numerical evidence that for a reasonable choice of $\{\mathbf{H}_h\}$, (4.54) also holds true for unstructured simplicial triangulations. We will discuss this issue in more detail below.

Assuming (4.54) for the moment, we show in the following theorem that the condition number of $\mathring{A}$ behaves like $O(h_{min}^{-2})$ when $h \to 0$.

Theorem 4.7. *Assume that $\mathbf{H}_h := (\mathcal{N}_h, \mathcal{H}_h)$ is a family of spanning trees associated with a shape regular family of triangulations $\{\mathcal{T}_h\}$ of Ω such that (4.54) is satisfied with α independent of h. Then there exists a constant $C(\alpha)$ independent of h such that*

$$\kappa(\mathring{A}) \leq C(\alpha)\, h_{min}^{-2} \tag{4.56}$$

Proof. Let

$$\vec{u} := \sum_{E \in (\mathcal{E}_I \cup \mathcal{E}_D) \setminus \mathcal{H}_h} u_E \vec{\Phi}_E \quad \in \quad \mathring{\mathcal{U}},$$

and let $\mathbf{u} := [u_E]_{E \in (\mathcal{E}_I \cup \mathcal{E}_D) \setminus \mathcal{H}_h}$ denote the vector of coefficients of $\vec{u}$. The proof of (4.56) will easily follow from (4.54) and the following results (see Theorems A.1 and A.2 in Appendix A):

$$ch(T) \sum_{E \subset \overline{T}} u_E^2 \;\leq\; \int_T |\vec{u}|^2 d\vec{x} \;\leq\; Ch(T) \sum_{E \subset \overline{T}} u_E^2, \tag{4.57}$$

$$\text{and} \qquad \int_T |\vec{\operatorname{curl}}\, \vec{u}|^2 d\vec{x} \;\leq\; Ch(T)^{-2} \int_T |\vec{u}|^2 d\vec{x}, \tag{4.58}$$

for all $T \in \mathcal{T}_h$, and for c and C independent of h.

Now to prove (4.56), note first that

$$a(\vec{u}, \vec{u}) = \sum_{E, E' \in (\mathcal{E}_I \cup \mathcal{E}_D) \setminus \mathcal{H}_h} u_E a(\vec{\Phi}_E, \vec{\Phi}_{E'}) u_{E'} = \mathbf{u}^T \mathring{A} \mathbf{u}.$$

It follows from (2.2) and (4.58) that

$$\mathbf{u}^T \mathring{A} \mathbf{u} = a(\vec{u}, \vec{u}) \leq \theta^{-1} \sum_{T \in \mathcal{T}_h} \int_T |\vec{\operatorname{curl}}\, \vec{u}|^2 d\vec{x} \leq C\theta^{-1} \sum_{T \in \mathcal{T}_h} h(T)^{-2} \int_T |\vec{u}|^2 d\vec{x},$$

which combined with (4.57) leads to the estimate

$$\mathbf{u}^T \mathring{A} \mathbf{u} \leq C \sum_{T \in \mathcal{T}_h} h(T)^{-1} \sum_{E \subset \overline{T}} u_E^2 \leq C h_{min}^{-1} |\mathbf{u}|^2, \qquad \text{for all } \mathbf{u} \in \mathbb{R}^{\mathring{n}}. \tag{4.59}$$

On the other hand, it follows from (2.2) and (4.54) that

$$\mathbf{u}^T \mathring{A} \mathbf{u} = a(\vec{u}, \vec{u}) \geq \Theta^{-1} \|\vec{\operatorname{curl}}\, \vec{u}\|^2_{(L_2(\Omega))^3} \geq (\alpha\Theta)^{-1} \|\vec{u}\|^2_{(L_2(\Omega))^3},$$

which combined with (4.57) leads to the estimate

$$\mathbf{u}^T \mathring{A} \mathbf{u} \geq c(\alpha\Theta)^{-1} h_{min} |\mathbf{u}|^2, \qquad \text{for all } \mathbf{u} \in \mathbb{R}^{\mathring{n}}. \tag{4.60}$$

Together, (4.59) and (4.60) prove that there are constants C and $c(\alpha)$ independent of h such that

$$\lambda_{max}(\mathring{A}) \leq C h_{min}^{-1}, \qquad \text{and} \qquad \lambda_{min}(\mathring{A}) \geq c(\alpha)\, h_{min}, \tag{4.61}$$

which gives the desired result $\kappa(\mathring{A}) = \frac{\lambda_{max}(\mathring{A})}{\lambda_{min}(\mathring{A})} \leq \frac{C}{c(\alpha)}\, h_{min}^{-2}$. $\square$

Discussion of the Poincaré inequality (4.54)

Let us now discuss (4.54) for the special case of a family $\{\mathcal{T}_h : h := \sqrt{3}/N \text{ and } N \in \mathbb{N}\}$ of uniform simplicial triangulations of the unit cube $(0,1)^3$. The triangulation $\mathcal{T}_h$ is constructed from a uniform rectangular mesh of $N * N * N$ cubes, which are each subdivided themselves into 6 tetrahedra (see Figure 4.1), so that the mesh diameter $h = \sqrt{3}/N$.

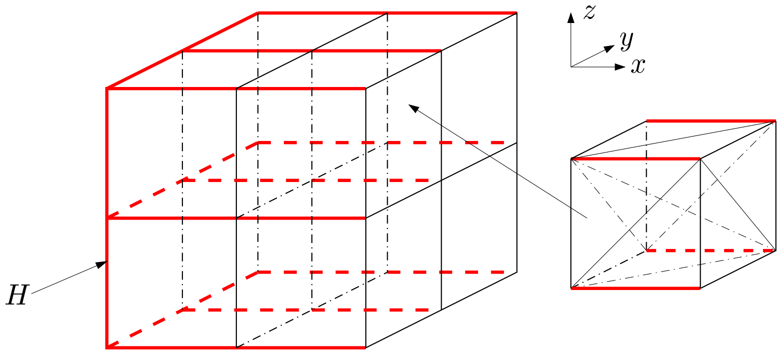


Figure 4.1: Uniform simplicial triangulation of the unit cube $(0,1)^3$ (for $N = 2$), with a "good" spanning tree $\mathbf{H}_h^+$ shown in red.

We do not claim that inequality (4.54) holds true with α independent of h, for an arbitrary family of spanning trees $\{\mathbf{H}_h := (\mathcal{N}_h, \mathcal{H}_h)\}$ associated with $\{\mathcal{T}_h\}$. However, we conjecture that (4.54) will hold true with α independent of h, if $\{\mathbf{H}_h\}$ is chosen as depicted in Figure 4.1. We will denote this family by $\{\mathbf{H}_h^+\}$, where "+" stands for "good". If $\{\mathbf{H}_h\}$ is chosen as depicted in Figure 4.2, on the other hand, then $\alpha = O(h^{-1})$ and (4.54) does not hold true (as we will see in the following Lemma). We will denote this family by $\{\mathbf{H}_h^-\}$, where "−" stands for "bad".

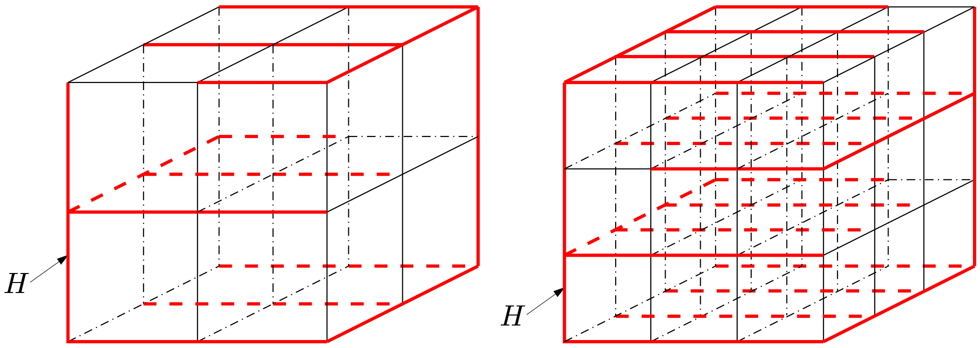


Figure 4.2: A "bad" spanning tree $\mathbf{H}_h^-$ (for $N = 2$ and 3).

Lemma 4.8. *Let $\{\mathcal{T}_h\}$ be the family of uniform simplicial triangulations of the unit cube $(0,1)^3$ as defined above, and let $\{\mathbf{H}_h^- := (\mathcal{N}_h, \mathcal{H}_h^-)\}$ be the "bad" family of spanning trees, as constructed in Figure 4.2. Then there exists an element $\vec{u} \in \mathring{\mathcal{U}}$ such that*

$$ch^{-1}\|\vec{\mathrm{curl}}\,\vec{u}\|^2_{(L_2(\Omega))^3} \leq \|\vec{u}\|^2_{(L_2(\Omega))^3} \leq Ch^{-1}\|\vec{\mathrm{curl}}\,\vec{u}\|^2_{(L_2(\Omega))^3}. \tag{4.62}$$

and (4.54) does not hold true with α independent of h.

Proof. Let H denote the edge between the nodes $(0,0,0)^T$ and $(0,0,\frac{1}{N})^T$ (as marked in Figure 4.2). Then $H \in \mathcal{H}_h^-$, as indicated in Figure 4.2. Furthermore, let $\vec{\Phi}_H$ be the basis function of $\mathcal{ND}_1(\Omega,\mathcal{T}_h)$ associated with edge H. Without loss of generality we assume that $H \in \mathcal{E}_D$.

Since $\vec{\Psi}_H := \vec{\operatorname{curl}}\,\vec{\Phi}_H \in \mathring{\mathcal{V}}$ (cf. Proposition 3.28) and since $\{\vec{\Psi}_{E'} : E' \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}_h^-\}$ is a basis of $\mathring{\mathcal{V}}$ (cf. Corollary 3.32), there has to be a set $\{\beta_{E'} : E' \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}_h^-\}$ such that

$$\vec{\Psi}_H := \sum_{E' \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}_h^-} \beta_{E'} \vec{\Psi}_{E'}. \tag{4.63}$$

As in the proof to Theorem 3.29, for each $E \in \mathcal{E}\backslash\mathcal{H}_h^-$, let $\boldsymbol{\mu}^E$ denote the vector associated with the unique cycle μ^E generated by combining the edge E with the tree $\mathbf{H}_h^-$, with the property that $\mu^E_{E'} := \delta_{E,E'}$, for all $E' \in \mathcal{E}\backslash\mathcal{H}_h^-$ (cf. Theorem B.6). Then using Lemma 3.31, and in particular the bilinear form $d(\cdot,\cdot)$ defined therein, we get

$$\beta_E = \sum_{E' \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}_h^-} \beta_{E'} \mu^E_{E'} = \sum_{E' \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}_h^-} \beta_{E'} d(\mu^E, \vec{\Psi}_{E'}) = d(\mu^E, \vec{\Psi}_H) = \mu^E_H,$$

and therefore

$$\vec{\Psi}_H = \sum_{E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}_h^-} \mu^E_H \vec{\Psi}_E. \tag{4.64}$$

Note that the coefficient $\mu^E_H = \pm 1$, if the edge $H \in \mathcal{H}_h^-$ is in the cycle μ^E associated with edge $E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}_h^-$, and it is 0 otherwise.

Now, let

$$\vec{u} := \sum_{E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}_h^-} \mu^E_H \vec{\Phi}_E. \tag{4.65}$$

Then $\vec{u} \in \mathring{\mathcal{U}}$, with $\mathring{\mathcal{U}}$ defined in (4.55), and as in the proof to Theorem 4.7 it follows from (4.57) that

$$ch \sum_{E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}_h^-} \left(\mu^E_H\right)^2 \le \|\vec{u}\|^2_{(L_2(\Omega))^3} \le Ch \sum_{E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}_h^-} \left(\mu^E_H\right)^2. \tag{4.66}$$

On the other hand, since $\vec{\operatorname{curl}}\,\vec{u} = \vec{\Psi}_H = \vec{\operatorname{curl}}\,\vec{\Phi}_H$, we have

$$ch^{-1} \le \|\vec{\operatorname{curl}}\,\vec{u}\|^2_{(L_2(\Omega))^3} \le Ch^{-1}. \tag{4.67}$$

This follows easily by summation over all $T \in \mathcal{T}_h$ from the following result (see Lemma A.4 in Appendix A):

$$ch(T)^{-1} \;\le\; \int_T |\vec{\Psi}_E|^2 \, d\vec{x} \;\le\; Ch(T)^{-1}, \tag{4.68}$$

for all $E \subset \overline{T}$ and for all $T \in \mathcal{T}_h$.

To complete the proof, let us calculate $\sum_{E\in(\mathcal{E}_I\cup\mathcal{E}_D)\backslash\mathcal{H}_h^-}\left(\mu_H^E\right)^2$. Edge H is used in the cycles μ^E associated with all the vertical edges $E \in \mathcal{E}\backslash\mathcal{H}_h^-$, and therefore $|\mu_H^E| = 1$ for all vertical edges $E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}_h^-$. Since the number of edges per cube is constant, and since the total number of cubes is $N^3 = \sqrt{27}\,h^{-3}$, this implies that

$$ch^{-3} \leq \sum_{E\in(\mathcal{E}_I\cup\mathcal{E}_D)\backslash\mathcal{H}_h^-}\left(\mu_H^E\right)^2 \leq Ch^{-3}.$$

Now, combining this with (4.66) and (4.67), we see that $\vec{u}$ satisfies (4.62) and the inequality (4.54) does not hold true with α independent of h. □

In the case of the "good" family $\{\mathbf{H}_h^+ := (\mathcal{N}_h, \mathcal{H}_h^+\}$ of spanning trees, one of which is depicted in Figure 4.1, on the other hand, edge $H \in \mathcal{H}_h^+$ is not used in any of the cycles μ^E associated with the edges $E \in \mathcal{E}\backslash\mathcal{H}_h^+$ that lie in $(0,1)^2 \times (\frac{1}{N}, 1)$, and therefore $\mu_H^E = 0$, for all edges $E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}_h^+$ in $(0,1)^2 \times (\frac{1}{N}, 1)$. Since the number of edges per cube is constant, and since the number of cubes in $(0,1)^2 \times (0, \frac{1}{N})$ is $N^2 = 3h^{-2}$, this implies

$$\sum_{E\in(\mathcal{E}_I\cup\mathcal{E}_D)\backslash\mathcal{H}_h^+}\left(\mu_H^E\right)^2 \leq Ch^{-2}. \tag{4.69}$$

Combining this with (4.66) and (4.67), we see (following the same steps as in the proof to Lemma 4.8) that for this choice of spanning tree

$$\vec{u} := \sum_{E\in(\mathcal{E}_I\cup\mathcal{E}_D)\backslash\mathcal{H}_h^+}\mu_H^E\vec{\Phi}_E \tag{4.70}$$

satisfies the inequality (4.54) with α independent of h. Similar arguments show that the function $\vec{u}$ defined in (4.70) satisfies (4.54) with α independent of h for all $H \in \mathcal{H}_h^+\backslash\mathcal{E}_N$, in the case of the "good" family $\{\mathbf{H}_h^+\}$ of spanning trees in Figure 4.1. Although this is obviously no proof of (4.54), it is a strong argument in favour, since the functions $\vec{u}$ constructed in (4.70) for each $H \in \mathcal{H}_h^+\backslash\mathcal{E}_N$, are particularly bad functions (with respect to (4.54)). Their support is very large, while their 3D curl vanishes almost everywhere. The numerical results in Section 4.5.2 will underline this point.

Nevertheless, this discussion has definitely shown that for a particular triangulation $\mathcal{T}$ the constant α in (4.54) (and therefore the condition number of $\mathring{A}$) depends on the choice of spanning tree $\mathbf{H} := (\mathcal{N}, \mathcal{H})$, and in particular on

$$\zeta(\mathbf{H}) := \max_{H\in\mathcal{H}\backslash\mathcal{E}_N}\left\{\sum_{E\in(\mathcal{E}_I\cup\mathcal{E}_D)\backslash\mathcal{H}}\left(\mu_H^E\right)^2\right\}. \tag{4.71}$$

We will investigate this relationship numerically in Section 4.5.2.

Despite the fact that the conditioning of $\mathring{A}$ depends on the choice of tree, $\mathring{A}$ is always symmetric positive definite. Thus we can apply preconditioned conjugate gradients to the decoupled velocity system (4.7), and we know that (in the unpreconditioned case) the number of iterations will grow no faster than with the square root of the condition number of $\mathring{A}$ (i.e. $O(h^{-1})$ for the family $\{\mathcal{T}_h\}$ of uniform triangulations defined above, if we assume (4.54) once again). To solve system (4.2) on the other hand, we would have to fall back on the methods presented in Section 2.3.3, such as MINRES (see Figure 2.2). Here (in the unpreconditioned case), the number of iterations (as in 2D) can only be expected to grow no faster than the condition number of the matrix $\mathcal{M}$ (i.e. $O(h^{-1})$ for a quasi-uniform family, see Corollary 2.28).

An optimal preconditioner for $\mathring{A}$, which ensures in theory that the number of iterations for conjugate gradients does not grow as $h \to 0$, is so far only available for uniform, rectangular hexahedral meshes using trees like the "good" tree $\mathbf{H}_h^+$, as presented recently by Cai et al. [21]. The construction of such a preconditioner for unstructured simplicial triangulations, on the other hand, is still an open question in 3D and outside the scope of this thesis.

Nevertheless, we saw in this section that from several points of view the reduction to the decoupled symmetric positive definite velocity system (4.7) makes practical sense also in 3D, and the numerical results in Section 4.5.2 will underline this point.

4.3.3 Extensions

The general situation, where Γ_N^ℓ is not simply connected for some $\ell = 1, \ldots, s_N$, involves the introduction of additional non-local basis functions (see Figure 3.8), and as in 2D these basis functions will lead to a bordered coefficient matrix $\mathring{A}$, that can be obtained from $\mathcal{A}$ using a small number of elementary operations on rows and columns corresponding to certain boundary edges $E \in \mathcal{E}_N$. To simplify the presentation let us suppose, as in Theorem 3.34, that

$$\Gamma_N \text{ is connected, and } \Gamma_D = \Gamma_D^1 \cup \Gamma_D^2, \text{ with } \Gamma_D^1, \Gamma_D^2 \text{ connected and } \overline{\Gamma}_D^1 \cap \overline{\Gamma}_D^2 = \emptyset.$$

Then the basis $\{\vec{\Psi}_i\}$ is given by $\{\vec{\Psi}_E : E \in (\mathcal{E}_I \cup \mathcal{E}_D) \backslash \mathcal{H}\} \cup \{\sum_{E \in \mathcal{E}^{1,2}} \vec{\Psi}_E\}$, as specified in Theorem 3.34. The additional row in Z corresponding to the non-local basis function $\sum_{E \in \mathcal{E}^{1,2}} \vec{\Psi}_E$ will be put to the end, and thus identified with the index $\mathring{n}$. Using this identification we can rewrite the last of the formulae in (4.14) by

$$\sum_{E \in \mathcal{E}^{1,2}} \vec{\Psi}_E = \sum_{F \in \mathcal{F}_I \cup \mathcal{F}_D} Z_{\mathring{n},F} \vec{v}_F, \tag{4.72}$$

where

$$Z_{\mathring{n},F} := \begin{cases} \pm \frac{1}{|F|} & \text{for all } F \in \mathcal{F}_I \cup \mathcal{F}_D \text{ such that } E \subset \overline{F} \text{ for some } E \in \mathcal{E}^{1,2}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.73}$$

Now, let $\mathcal{A}$ be as defined in (4.51) and let A denote the minor of $\mathcal{A}$ defined in (4.52).

Then by also defining

$$c_E := \sum_{E' \in \mathcal{E}^{1,2}} \mathcal{A}_{E,E'}, \qquad E \in (\mathcal{E}_I \cup \mathcal{E}_D) \backslash \mathcal{H}, \qquad \text{and} \quad r := \sum_{E,E' \in \mathcal{E}^{1,2}} \mathcal{A}_{E,E'},$$

we can prove the following result.

Corollary 4.9. *Let Γ_N and Γ_D be as defined in Theorem 3.34. Then*

$$\mathring{A} = \begin{pmatrix} A & \mathbf{c} \\ \mathbf{c}^T & r \end{pmatrix}.$$

Proof. Using (4.16) again, this follows directly from the proof to Proposition 4.5 and the definition of A, $\mathbf{c}$ and r. □

Thus, the coefficient matrix $\mathring{A}$ is a bordered matrix with major block consisting of the matrix discussed in Section 4.3.2, and can be solved by block elimination (see Section 4.2.2).

4.4 Non-zero divergence - Static condensation

The final theoretical issue that remains to be discussed, is what we do when the underlying continuous problem of (2.29) is not in divergence form (2.18), but in the form (2.15), and $\mathbf{f} \neq 0$ in (4.1). The method we are going to describe was developed by Ewing & Wang in [38, 39], and most of the results in this section are taken from their papers. We include them for completeness and extend them to the case of mixed boundary conditions. Further references are Mathew [71, 72] and Hiptmair & Hoppe [59]

At the beginning of this chapter we saw that (4.1) can be reduced to a problem of the form (4.2) with $\mathbf{f} = 0$, if we know a particular solution $\mathbf{u}^*$ to the constraint equation

$$B^T \mathbf{u}^* = \mathbf{f}. \tag{4.74}$$

Recall that if such a particular solution $\mathbf{u}^*$ is known, by setting $\mathbf{u} := \mathbf{u}^* + \mathbf{u}^0$, system (4.1) is equivalent to solving

$$\begin{pmatrix} M & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}^0 \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{g} - M\mathbf{u}^* \\ \mathbf{0} \end{pmatrix}, \tag{4.75}$$

which is then of same form as (4.2), and can be solved by the decoupling method described in the previous sections.

Thus it remains to solve (4.74), or equivalently the finite element system

$$b(\vec{U}^*, W) = -\int_\Omega f\, W\, d\vec{x}, \qquad \text{for all } W \in \mathcal{W}. \tag{4.76}$$

In Ewing & Wang [38, 39] this is achieved through a variation of domain decomposition (which may also be thought of as a kind of *static condensation*). For ease of presentation, we restrict to the lowest-order case, i.e. $k = 0$.

4.4.1 Two-level approach

Let $\mathcal{T}_0$ be a coarse triangulation of Ω, chosen so that on the boundary it is aligned with the interfaces between Γ_D and Γ_N, i.e. for each element ω of $\mathcal{T}_0$, $\overline{\omega} \cap \Gamma$ is either entirely in Γ_D or in Γ_N or it is empty. The faces of the elements $\omega \in \mathcal{T}_0$ are assumed to align with the faces of the elements $T \in \mathcal{T}$, so that $\mathcal{T}$ can be regarded as a refinement of $\mathcal{T}_0$. Then, the coarse space $\mathcal{W}_0 = \mathcal{P}_0(\Omega, \mathcal{T}_0)$ of discontinuous piecewise constant finite element functions (see the definition in (2.48)) is obviously contained within the fine space $\mathcal{W} = \mathcal{P}_0(\Omega, \mathcal{T})$. Recall that $\{w_T : T \in \mathcal{T}\}$ is the canonical basis of $\mathcal{W}$ defined in (2.76), and let $\{w_\omega : \omega \in \mathcal{T}_0\}$ denote the basis of $\mathcal{W}_0$, which is defined in the same way by

$$w_\omega|_{\omega'} = \delta_{\omega,\omega'}, \qquad \text{for all } \omega' \in \mathcal{T}_0.$$

Furthermore, let $R_0 : \mathcal{W} \to \mathcal{W}_0$ be a restriction from the fine space $\mathcal{W}$ onto the coarse space $\mathcal{W}_0$, such that

$$p_0 := R_0 p = \sum_{\omega \in \mathcal{T}_0} p_{0,\omega} w_\omega, \quad \text{with} \quad p_{0,\omega} := \tfrac{1}{|\omega|} \sum_{T \subset \omega} |T| \, p_T, \tag{4.77}$$

for each $p \in \mathcal{W}$, where $\mathbf{p} = [p_T]_{T \in \mathcal{T}}$ is the coefficient vector of p. Its transpose $R_0^T : \mathcal{W}_0 \to \mathcal{W}$ yields a suitable prolongation.

Furthermore, for each element ω of $\mathcal{T}_0$, let $\mathcal{W}_\omega$ denote the space of piecewise constant functions on ω with respect to the fine mesh $\mathcal{T}$, more precisely $\mathcal{W}_\omega = \mathcal{P}_0(\omega, \mathcal{T} \cap \omega)$. Let $R_\omega : \mathcal{W} \to \mathcal{W}_\omega$ be the restriction operator onto the element ω defined by

$$R_\omega p := \sum_{T \subset \omega} p_T w_T, \tag{4.78}$$

for each $p \in \mathcal{W}$. The prolongation is chosen as the transpose $R_\omega^T : \mathcal{W}_\omega \to \mathcal{W}$.

To find a solution $\vec{U}^*$ of (4.76) we start by defining $f^h \in \mathcal{W}$ such that

$$\int_\Omega (f^h - f) W \, d\vec{x} = 0, \qquad \text{for all } W \in \mathcal{W}. \tag{4.79}$$

i.e. f^h is the orthogonal projection of f onto $\mathcal{W}$ with respect to the L_2 inner product on Ω. Next note that we can decompose $f^h \in \mathcal{W}$ in the following way.

Lemma 4.10.

$$f^h = f_0^h + \sum_{\omega \in \mathcal{T}_0} R_\omega^T f_\omega^h, \qquad \textit{where} \quad f_0^h := R_0 f^h \quad \textit{and} \quad f_\omega^h := R_\omega(f^h - f_0^h).$$

Proof. It follows directly from the fact that $\sum_{\omega \in \mathcal{T}_0} R_\omega^T R_\omega$ is the identity map on $\mathcal{W}$. $\square$

Furthermore, let

$$\mathcal{V}_0 := \{\vec{v} \in \mathcal{RT}_0(\Omega, \mathcal{T}_0) : \vec{v} \cdot \vec{\nu}|_{\Gamma_N} = 0\}$$

and for each $\omega \in \mathcal{T}_0$ let

$$\mathcal{V}_\omega := \{\vec{v} \in \mathcal{RT}_0(\omega, \mathcal{T}_h \cap \omega) : \vec{v} \cdot \vec{\nu}_\omega|_{\partial\omega} = 0\},$$

where $\vec{\nu}_\omega(\vec{x})$ denotes the unit outward normal from ω at $\vec{x} \in \partial\omega$.

Note that we impose pure Neumann conditions on the entire boundary of ω in the local subspaces $\mathcal{V}_\omega$. Since we assumed at the beginning of this thesis that $\Gamma_D \neq \emptyset$ we have not yet discussed the discretisation by Raviart-Thomas-Nédélec elements of mixed problems, which are subject to pure Neumann boundary conditions. However, it is shown in Roberts & Thomas [78, Section 14] that in this case

$$\operatorname{div} \mathcal{V}_\omega = \overline{\mathcal{W}}_\omega := \{W \in \mathcal{W}_\omega : \int_\omega W \, d\vec{x} = 0\}, \tag{4.80}$$

i.e. the spaces $\mathcal{V}_\omega$ and $\overline{\mathcal{W}}_\omega$ satisfy the discrete inf-sup condition (see Lemma 2.8).

Lemma 4.11.

$$f^h_\omega \in \overline{\mathcal{W}}_\omega, \qquad \textit{for all} \quad \omega \in \mathcal{T}_0.$$

Proof. Let $\omega \in \mathcal{T}_0$ and let $\left[f^h_T\right]_{T\in\mathcal{T}}$ be the vector of coefficients of f^h with respect to the basis $\{w_T\}$ of $\mathcal{W}$. We only have to verify that $\int_\omega f^h_\omega d\vec{x} = 0$, but by definition we have

$$\int_\omega f^h_\omega d\vec{x} = \int_\omega f^h d\vec{x} - \int_\omega R_0 f^h d\vec{x} = \sum_{T\subset\omega} f^h_T |T| - f^h_{0,\omega}|\omega| = 0,$$

where in the last two steps we have used (4.77). $\square$

As we shall show in Theorem 4.12, a particular solution $\vec{U}^*$ of (4.76) can now be found by solving for each $\omega \in \mathcal{T}_0$, a local subproblem

$$\left.\begin{array}{llll} m(\vec{U}^*_\omega, \vec{V}) + b(\vec{V}, P^*_\omega) & = & 0\,, & \text{for all } \vec{V} \in \mathcal{V}_\omega, \\ b(\vec{U}^*_\omega, W) & = & -\int_\omega f^h_\omega W \, d\vec{x}, & \text{for all } W \in \overline{\mathcal{W}}_\omega \end{array}\right\} \tag{4.81}$$

for $(\vec{U}^*_\omega, P^*_\omega) \in \mathcal{V}_\omega \times \overline{\mathcal{W}}_\omega$, and by additionally solving the coarse grid problem

$$\left.\begin{array}{llll} m(\vec{U}^*_0, \vec{V}) + b(\vec{V}, P^*_0) & = & 0\,, & \text{for all } \vec{V} \in \mathcal{V}_0, \\ b(\vec{U}^*_0, W) & = & -\int_\Omega f^h_0 W \, d\vec{x}, & \text{for all } W \in \mathcal{W}_0 \end{array}\right\} \tag{4.82}$$

for $(\vec{U}^*_0, P^*_0) \in \mathcal{V}_0 \times \mathcal{W}_0$.

Theorem 4.12. *Let*

$$\vec{U}^* := \vec{U}_0^* + \vec{U}_1^*, \quad \text{where} \quad \vec{U}_1^*\Big|_\omega := \vec{U}_\omega^*, \quad \text{for all} \quad \omega \in \mathcal{T}_0.$$

Then $\vec{U}^*$ satisfies (4.76).

Proof. Since $\operatorname{div} \mathcal{V}_\omega = \overline{\mathcal{W}}_\omega$, the existence of a unique solution of (4.81) follows directly from Theorem 2.9, for all $\omega \in \mathcal{T}_0$. Similarly, since the coarse grid $\mathcal{T}_0$ is assumed to be aligned with the interfaces between Γ_D and Γ_N, the existence of a unique solution of (4.82) also follows from Theorem 2.9 and the fact that $\operatorname{div} \mathcal{V}_0 = \mathcal{W}_0$ (see section 2.2.2).

Secondly, we need to confirm that $\vec{U}^* \in \mathcal{V}$. Since we assumed that the faces of $\mathcal{T}_0$ align with the faces of $\mathcal{T}$ and that each boundary face of $\mathcal{T}_0$ is either entirely in Γ_D or in Γ_N, using Proposition 2.7 we have $\mathcal{V}_0 \subset \mathcal{V}$, and so $\vec{U}_0^* \in \mathcal{V}$. Moreover, using the fact that $\vec{U}_\omega^* \in \mathcal{RT}_0(\omega, \mathcal{T}_h \cap \omega)$ and $\vec{U}_\omega^* \cdot \vec{\nu}\,|_{\partial\omega} = 0$, for all $\omega \in \mathcal{T}_0$, we can apply Proposition 2.7 again and see that $\vec{U}_1^* \in \mathcal{V}$ as well.

Now, using the definition (2.10) of $b(\cdot,\cdot)$ together with the second equation in (4.81) and in (4.82), we have that $\operatorname{div} \vec{U}_0^* = -f_0^h$ in $\mathcal{W}_0$ and $\operatorname{div} \vec{U}_\omega^* = -f_\omega^h$ in $\overline{\mathcal{W}}_\omega \subset \mathcal{W}_\omega$, for all $\omega \in \mathcal{T}_0$. It follows from (4.79) and Lemma 4.10 that

$$b(\vec{U}^*, W) = -\int_\Omega f\, W\, d\vec{x}, \qquad \text{for all } W \in \mathcal{W}.$$

□

Remark 4.13. Since we are only interested in a particular solution to (4.76), the bilinear form $m(\cdot,\cdot)$ in (4.81) and (4.82) can be replaced by any more convenient bilinear form

$$\tilde{m}(\vec{u}, \vec{v}) = \int_\Omega \tilde{D}^{-1}(\vec{x})\vec{u} \cdot \vec{v}\, d\vec{x}, \tag{4.83}$$

e.g. choosing $\tilde{D}(\vec{x}) \equiv I$. This might significantly simplify the solution of (4.81) and (4.82), especially in the presence of large discontinuities in $D(\vec{x})$.

We are not going to give any detail on how this method would be implemented and how the local Neumann problems could be solved, but we refer to Mathew [71], where these issues are discussed in detail. However, to analyse the cost of this method, suppose for example that $\mathcal{T} = \mathcal{T}_h$ and $\mathcal{T}_0 = \mathcal{T}_H$, where $H > h$, and that $\{\mathcal{T}_h\}$, $\{\mathcal{T}_H\}$ are quasi-uniform families of triangulations as $h, H \to 0$. As usual, let h and H denote the largest diameter of the elements $T \in \mathcal{T}$ and $\omega \in \mathcal{T}_0$, respectively. If we choose the coarse grid $\mathcal{T}_0$ such that

$$H = O(h^{1/2}),$$

then the size of the problems (4.81) and (4.82) will grow at the same rate, with $O(h^{-d/2})$ when $h \to 0$. In comparison, the system (4.1) grows with $O(h^{-d})$ when $h \to 0$. So even if a direct method, like the frontal method (see Johnson [63, Section 6.5]) is used to solve each of the linear equation systems resulting from (4.81) and (4.82), the amount of work for each system (including the prolongation from $\mathcal{V}_0$ and $\mathcal{V}_\omega$ onto $\mathcal{V}$) will grow no worse than with $O(h^{-d})$ when $h \to 0$ (see Hagger [51, Section 7.6]).

Since the problem (4.82) and the problems (4.81), for all $\omega \in \mathcal{T}_0$, are fully decoupled, we can solve them in parallel, and thus (provided we have enough processors available) a particular solution $\vec{U}^*$ of (4.76) can be found in asymptotically optimal time. Because of the comments in Remark 4.13, the dependency of the method on discontinuities in $D(\vec{x})$ does not concern us either. We can simply replace the bilinear form $m(\cdot,\cdot)$ in (4.81) and (4.82) by $\tilde{m}(\cdot,\cdot)$, with $\tilde{D}(\vec{x}) = I$.

4.4.2 Extension – multi-level approach

The two-level procedure can be extended to a multilevel procedure which is asymptotically optimal even on a single processor (see Ewing & Wang [39] and Hiptmair & Hoppe [59]), provided we have a hierarchy of simplicial triangulations $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_L := \mathcal{T}$, created by a regular refinement of an initial "coarse" triangulation $\mathcal{T}_0$. This corresponds to breaking up every element of $\mathcal{T}_{i-1}$ into four subtriangles in 2D or eight subtetrahedra in 3D, to obtain $\mathcal{T}_i$.

If this hierarchy is given and if for each level $i = 0, \ldots, L$, the restriction R_i from $\mathcal{W}$ to $\mathcal{W}_i := \mathcal{P}_0(\Omega, \mathcal{T}_i)$ is defined in the same way as R_0 in (4.77) before, then we can extend Lemma 4.10 in the following way.

Lemma 4.14.

$$f^h = f_0^h + \sum_{i=1}^{L} f_i^h, \qquad \textit{where} \quad f_0^h := R_0 f^h \quad \textit{and} \quad f_i^h := R_i f^h - R_{i-1} f^h.$$

Proof. It follows directly from the fact that the spaces $\mathcal{W}_i$, $i = 0, \ldots, L$, are nested (i.e. $\mathcal{W}_i \subset \mathcal{W}_{i-1}$) and that R_L is the identity map on $\mathcal{W} = \mathcal{W}_L$. $\square$

Furthermore, for each level $i = 1, \ldots, L$ and for each element $T \in \mathcal{T}_{i-1}$, we define the local finite element spaces

$$\begin{aligned} \mathcal{V}_{i,T} &:= \{\vec{V} \in \mathcal{RT}_0(T, \mathcal{T}_i \cap T) : \vec{V} \cdot \vec{\nu}_T|_{\partial T} = 0\}, \\ \overline{\mathcal{W}}_{i,T} &:= \{W \in \mathcal{P}_0(T, \mathcal{T}_i \cap T) : \int_\omega W \, d\vec{x} = 0\}. \end{aligned}$$

Lemma 4.15.

$$f_i^h|_T \in \overline{\mathcal{W}}_{i,T}, \qquad \textit{for each level } i = 1, \ldots, L \textit{ and for each element } T \in \mathcal{T}_{i-1}.$$

Proof. Let $i = 1, \ldots, L$ and let $T \in \mathcal{T}_{i-1}$. We only have to verify that $\int_T f_i^h d\vec{x} = 0$, and as in the proof to Lemma 4.11, it follows directly from the definition of the operators

R_i and R_{i-1} that

$$\int_T f_i^h d\vec{x} = \int_T R_i f^h d\vec{x} - \int_T R_{i-1} f^h d\vec{x} = 0.$$

□

The following local subproblems are therefore uniquely solvable again. For all $i = 1, \ldots, L$ and $T \in \mathcal{T}_{i-1}$, find $(\vec{U}^*_{i,T}, P^*_{i,T}) \in \mathcal{V}_{i,T} \times \overline{\mathcal{W}}_{i,T}$ such that

$$\left.\begin{array}{rclll} m(\vec{U}^*_{i,T}, \vec{V}) & + & b(\vec{V}, P^*_{i,T}) & = & 0\ , & \text{for all } \vec{V} \in \mathcal{V}_{i,T}, \\ b(\vec{U}^*_{i,T}, W) & & & = & -\displaystyle\int_T f_i^h W \, d\vec{x}, & \text{for all } W \in \overline{\mathcal{W}}_{i,T}. \end{array}\right\} \quad (4.84)$$

A particular solution $\vec{U}^*$ of (4.76) can now be constructed from the coarse grid solution $\vec{U}^*_0$ of (4.82) and from the solutions $\vec{U}^*_{i,T}$ of the local subproblems (4.84), as the following theorem shows.

Theorem 4.16. *Let*

$$\vec{U}^* := \sum_{i=0}^{L} \vec{U}^*_i, \quad \textit{where} \quad \vec{U}^*_i\Big|_T := \vec{U}^*_{i,T}, \quad \textit{for each } i = 1, \ldots, L, \textit{ and } T \in \mathcal{T}_{i-1}.$$

Then $\vec{U}^*$ satisfies (4.76).

Proof. Analogous to the proof to Theorem 4.12 (see also [39, Lemma 3.2]). □

The solution of each of the local subproblems (4.84) takes a fixed, small amount of elementary operations (see Figure 4.3 for $d = 2$ and $k = 0$), and the number of such

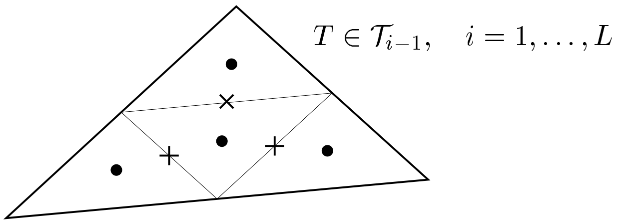


Figure 4.3: Degrees of freedom in $\mathcal{V}_{i,T} \times \mathcal{W}_{i,T}$ (for $k = 0$ and $d = 2$).

problems is proportional to the number of unknowns on the finest mesh $\mathcal{T} = \mathcal{T}_L$. If the coarse triangulation $\mathcal{T}_0$ is fixed, and the triangulation $\mathcal{T}$ is refined by adding new levels $\mathcal{T}_{L+1}, \mathcal{T}_{L+2}, \ldots$, as assumed at the beginning of this section, then $\vec{U}^*$ can be found in an asymptotically optimal number of operations proportional to the number of unknowns on the finest mesh, when $h \to 0$ (even on a single processor).

4.5 Numerical Results

In this final section we will examine numerically the sharpness of the theoretical results for our decoupled iterative method for problem (4.1). We also go slightly beyond what has been proved theoretically.

We will consider the special case when the underlying physical domain is given by $\Omega = (0,1)^d$, for $d = 2, 3$, when $D(\vec{x})$ in (2.1) is a scalar multiple of the identity, and when the right hand sides of the differential equation and of the Neumann boundary condition in the underlying continuous problem (2.1) are zero, i.e. $f \equiv 0$ and $g_N \equiv 0$. More general domains Ω and full tensors $D(\vec{x})$ will be included in the applications in Chapter 5. The different examples in this section are induced by different partitionings of the boundary Γ into Γ_D and Γ_N, as well as by different choices of coefficients $D(\vec{x})$ and of Dirichlet data g_D.

We discretise these problems using the mixed finite element discretisation (2.29) with lowest order Raviart-Thomas-Nédélec elements, i.e. $k = 0$, on simplicial triangulations $\mathcal{T}$ of Ω, and we solve the resulting systems of linear equations (4.2) by applying the decoupled iterative method which we have presented in Sections 4.1–4.3.

Since in all the following examples Γ_D is connected, the basis functions of $\mathring{\mathcal{V}}$ all have local supports (see Chapter 3, and therefore the number of borders in the decoupled velocity system (4.7) is zero. This means we can directly apply preconditioned conjugate gradients to (4.7) without first having to apply the block elimination described in Section 4.2.2.

4.5.1 The two-dimensional case

Let us first look at the two-dimensional case again, i.e. $d = 2$:

Definition 4.17 (Example 1).

$$\Gamma_N = [0,1] \times \{0\} \qquad \text{and} \qquad \Gamma_D = \Gamma \backslash \Gamma_N,$$

$$D(\vec{x}) \equiv I, \quad \text{for all} \quad \vec{x} \in \Omega,$$

and

$$g_D(\vec{x}) := 1 - x_1 \quad \text{for all} \quad \vec{x} := (x_1, x_2)^T \in \Gamma_D.$$

Thus, in this case, problem (2.1) corresponds to a Poisson problem on the unit square with mixed boundary conditions.

We discretise this using the mixed finite element discretisation (2.29) with lowest order Raviart-Thomas elements on a sequence of uniform meshes $\mathcal{T}_h$ obtained by firstly dividing Ω into N^2 equal squares $\left(\frac{i-1}{N}, \frac{i}{N}\right) \times \left(\frac{j-1}{N}, \frac{j}{N}\right)$, and then further subdividing each square into two triangles using a diagonal drawn from top left to bottom right (so that the mesh diameter $h = \sqrt{2}N^{-1}$).

To solve the resulting saddle point system (4.2) we use the decoupled iterative method described above: the construction of the matrix $\mathring{A}$ in the decoupled velocity system (4.7) is carried out in an elementwise fashion as presented in (4.27); the resulting symmetric positive definite system (4.7) is solved with preconditioned conjugate gradients (PCG) and a variety of preconditioners; the matrix A^c in the decoupled pressure system (4.11) is obtained from the original matrix B in (4.2) by deleting some rows and reordering the rows and columns (as mentioned in Section 4.1.3); and the resulting triangular system (4.11) is solved by simple back substitutions. We will test our method with four different preconditioners for $\mathring{A}$: diagonal scaling ($\mathrm{Diag}(\mathring{A})^{-1}$), incomplete LU decomposition with zero fill-in (ILU(0)), and the additive Schwarz preconditioner presented in Section 4.2.4 in the implementation provided by the `DOUG` package [52], both in its two-level form $\mathcal{P}^{-1}_{AS(2)}$ (defined in (4.43)), with an adaptively chosen coarse grid, and in its one-level form

$$\mathcal{P}^{-1}_{AS(1)} := \sum_{i=1}^{S} R_i^T A_i^{-1} R_i \qquad (4.85)$$

with no coarse grid solve. The convergence criterion in the PCG method is the relative reduction of the preconditioned residual by a factor of 10^{-9}.

Robustness with respect to h

Tables 4.1 and 4.2 show the performance of our method for this example when the mesh is refined. First in Table 4.1 we see that when we increase the number of degrees

			MFlops	
N	h	$n_{\mathcal{V}} + n_{\mathcal{W}}$	Decoupling Process	Recovery of Pressure
16	0.088	1296	0.1	0.02
32	0.044	5152	0.4	0.1
64	0.022	20544	1.65	0.25
128	0.011	82048	6.6	1.2
256	0.0055	327936	27	3.6

Table 4.1: Performance of the decoupled method for Example 1 (floating point operations for the decoupling process and for the recovery of the pressure).

of freedom in (4.2) (or equivalently when we reduce the mesh diameter h), the work required to set up the decoupled system (4.7) and to recover the vector of pressures $\mathbf{p}$ from (4.11) is asymptotically optimal, i.e. the number of *floating point operations* (Flops) that are necessary for these processes are growing linearly with the number of degrees of freedom in (4.2).

In Table 4.2, on the other hand, we investigate how the PCG method for the decoupled velocity system (4.7) is affected by a reduction of the mesh diameter h.

As predicted in Section 4.2.3, the dimension $\mathring{n}$ of the reduced system (4.7) is about

N	$\mathring{n}$	$\kappa(\mathring{A})$	Iterations No Prec.	$\mathrm{Diag}(\mathring{A})^{-1}$	ILU(0)	$\mathcal{P}_{AS(1)}^{-1}$	$\mathcal{P}_{AS(2)}^{-1}$
16	272	$9.3 \cdot 10^2$	71	19	20	1	1
32	1056	$3.5 \cdot 10^3$	140	46	35	1	1
64	4160	$1.4 \cdot 10^4$	271	97	66	47	18
128	16512	$5.4 \cdot 10^4$	525	193	126	98	17
256	65792	–	1033	380	247	207	18

Table 4.2: Performance of the decoupled method for Example 1 (iteration count for the solution of the velocity system by the PCG method).

5 times smaller than that of the full system (4.2) (compare Column 2 of Table 4.2 to Column 3 of Table 4.1) and the condition number of the coefficient matrix $\mathring{A}$ in (4.7) grows like $O(\mathring{n}) = O(h^{-2})$ (Column 3). Consequently, the number of iterations for the unpreconditioned conjugate gradient method grows (with the square root of the condition number) like $O(\mathring{n}^{1/2}) = O(h^{-1})$ (Column 4). The iteration counts in Columns 5–8 finally correspond to the different choices of preconditioner in the PCG method, as specified above[1]. While we can see a definite improvement compared to the unpreconditioned case, almost all of them are still affected by the mesh refinement, in that the number of iterations grows like $O(\mathring{n}^{1/2}) = O(h^{-1})$ (Columns 5–7). The only exception is the additive Schwarz preconditioner with coarse grid solve (Column 8). Here, the iterations stay constant when the mesh is refined, which is even better than in our theory where we predicted that they would grow no faster than $O(\mathring{n}^{1/6}) = O(h^{-1/3})$ (see Section 4.2.4).

Altogether we can conclude that for Example 1 our decoupled method (with the additive Schwarz preconditioner for $\mathring{A}$) is extremely robust when $h \to 0$. For the tested range of values for h it performed even optimally.

Comparison with MINRES

To get an idea of how the results for Example 1 compare to other solvers for the saddle point system (4.2), we solve (4.2) directly, using the MINRES algorithm given in Figure 2.2 in Section 2.3.3. To precondition this MINRES algorithm we take the ILU(0) factorisation of an asymptotically optimal, symmetric positive definite, block diagonal preconditioner

$$\mathcal{P}_{RW}^{-1} := \begin{pmatrix} I & 0 \\ 0 & B^T B \end{pmatrix}^{-1}, \tag{4.86}$$

[1] As mentioned in Section 4.2.4, in the additive Schwarz preconditioner ($\mathcal{P}_{AS(1)}^{-1}$ and $\mathcal{P}_{AS(2)}^{-1}$) the subdomains are chosen to contain approximately 1000 degrees of freedom. Since the subdomain solves are carried out by a direct solver, this accounts for the convergence in 1 iteration for $N = 16$ and 32.

presented and analysed in Rusten & Winther [82, Section 5.1] (see also Section 2.3.3). However, we will also include a modified version of $\mathcal{P}_{RW}^{-1}$ given by

$$\mathcal{P}_{MRW}^{-1} := \begin{pmatrix} \mathrm{Diag}(M) & 0 \\ 0 & B^T \mathrm{Diag}(M)^{-1} B \end{pmatrix}^{-1}. \tag{4.87}$$

As for PCG, the convergence criterion for MINRES is the relative reduction of the preconditioned residual by a factor of 10^{-9}.

The results are given in Tables 4.3 and 4.4 for the unpreconditioned and for the preconditioned version of MINRES, respectively. As predicted in Section 2.3.2, the condition number of the coefficient matrix $\mathcal{M}$ in (4.2) grows like $O(h^{-1})$ (Column 2 of Table 4.3), in contrast to the $O(h^{-2})$ growth of $\kappa(\mathring{A})$ in the reduced system (Column 3 of Table 4.3). However, since the number of iterations for MINRES grows linearly

	Condition Number		Iterations		MFlops	
N	$\kappa(\mathcal{M})$	$\kappa(\mathring{A})$	MINRES	Decoupled	MINRES	Decoupled
16	$1.2 \cdot 10^2$	$9.3 \cdot 10^2$	452	71	18.4	0.6
32	$2.2 \cdot 10^2$	$3.5 \cdot 10^3$	887	140	144	4.1
64	$4.7 \cdot 10^2$	$1.4 \cdot 10^4$	1794	271	1160	29.8
128	–	$5.4 \cdot 10^4$	3538	525	9160	223
256	–	–	7062	1033	73100	1740

Table 4.3: Comparison of the decoupled method (using unpreconditioned CG for the velocity system) with unpreconditioned MINRES for Example 1.

in the condition number, the method is affected by mesh refinement in the same way as the decoupled method, in that the number of iterations grows like $O(h^{-1})$ in the unpreconditioned (Column 4 of Table 4.3) as well as in the preconditioned case (Column 2 of Table 4.4).

	MINRES ($\mathcal{P}_{RW}^{-1}$)		MINRES ($\mathcal{P}_{MRW}^{-1}$)		Decoupled	
N	Iterations	MFlops	Iterations	MFlops	Iterations	MFlops
16	204	13.0	80	5.2	20	0.32
32	422	108	148	38.1	35	1.85
64	886	910	285	293	66	12.2
128	1882	7730	550	2260	126	86.4
256	3883	63850	1055	17360	247	646

Table 4.4: Comparison of the decoupled method (using ILU(0)-preconditioned CG) with preconditioned MINRES (Rusten & Winther [82]) for Example 1.

To get an idea of how well our decoupling strategy works, the results of MINRES are compared with the results of the decoupled method using a comparable method to solve the velocity system (4.7). In Table 4.3 we compare the number of iterations and the number of floating point operations for the unpreconditioned version of MINRES with the decoupled method using unpreconditioned CG to solve the velocity system, and in Table 4.4 we compare the results for the preconditioned version of MINRES (using the ILU(0) factorisation of $\mathcal{P}_{RW}^{-1}$ and $\mathcal{P}_{MRW}^{-1}$) with the decoupled method using ILU(0)-preconditioned CG for the velocity system. In both cases we can see a clear advantage of the decoupled method. In the unpreconditioned case, the work is reduced by a factor of 42 on the finest mesh. In the preconditioned case on the finest mesh, the work is reduced by a factor of 99 for the original preconditioner $\mathcal{P}_{RW}^{-1}$ proposed by Rusten & Winther and by a factor of 27 for the modified preconditioner $\mathcal{P}_{MRW}^{-1}$.

Now, taking the excellent performance of the additive Schwarz preconditioner (as discussed above) into account, we can conclude that for Example 1 our decoupled method (with $\mathcal{P}_{AS(2)}^{-1}$) will outperform MINRES, even if more sophisticated preconditioning techniques are applied to the full saddle point system (4.2).

Resilience of the method to discontinuities in $D(\vec{x})$

In Section 4.2.4, we also stated that the performance of the method will not deteriorate in the presence of strong discontinuities in the diffusion coefficient $D(\vec{x})$, if the additive Schwarz preconditioner is applied to the velocity system. We will show this resilience for the following example:

Definition 4.18 (Example 2).

$$\Gamma_N = [0,1]\times\{0\} \;\cup\; \{0,1\}\times\left[0,\tfrac{2}{3}\right] \qquad \text{and} \qquad \Gamma_D = \Gamma\backslash\Gamma_N,$$

$$D(\vec{x}) \equiv \begin{cases} \varepsilon\, I & \text{for all} \quad \vec{x} \in \left(\frac{1}{3},\frac{2}{3}\right)\times\left(\frac{1}{3},1\right), \\ I & \text{otherwise,} \end{cases} \tag{4.88}$$

with $\varepsilon \in (0,\infty)$ (see also Figure 4.4), and

$$g_D(\vec{x}) := 1 - x_1 \quad \text{for all} \quad \vec{x} := (x_1,x_2)^T \in \Gamma_D.$$

We discretise this problem using lowest order Raviart-Thomas elements on a tensor product mesh $\mathcal{T}$, which is graded in each direction in order to refine near the discontinuity, as depicted in Figure 4.4. The number of elements in this mesh is 18432, which corresponds to 46048 degrees of freedom in (4.2) and to 9184 degrees of freedom in the decoupled velocity system (4.7). As for Example 1, we solve the resulting saddle point system (4.2) using the decoupled iterative method described above, with diagonal scaling or additive Schwarz preconditioner for $\mathring{A}$. Again, the convergence criterion is the relative reduction of the preconditioned residual by a factor of 10^{-9}.

We investigate the robustness of the method with respect to discontinuities in $D(\vec{x})$ by varying ε in (4.88) from 10^{-6} to 10^{+6}. The results of this experiment are presented in Table 4.5. The solution for $\varepsilon = 10^{-6}, 10^{0}$ and 10^{+6} is depicted in Figure 4.5.

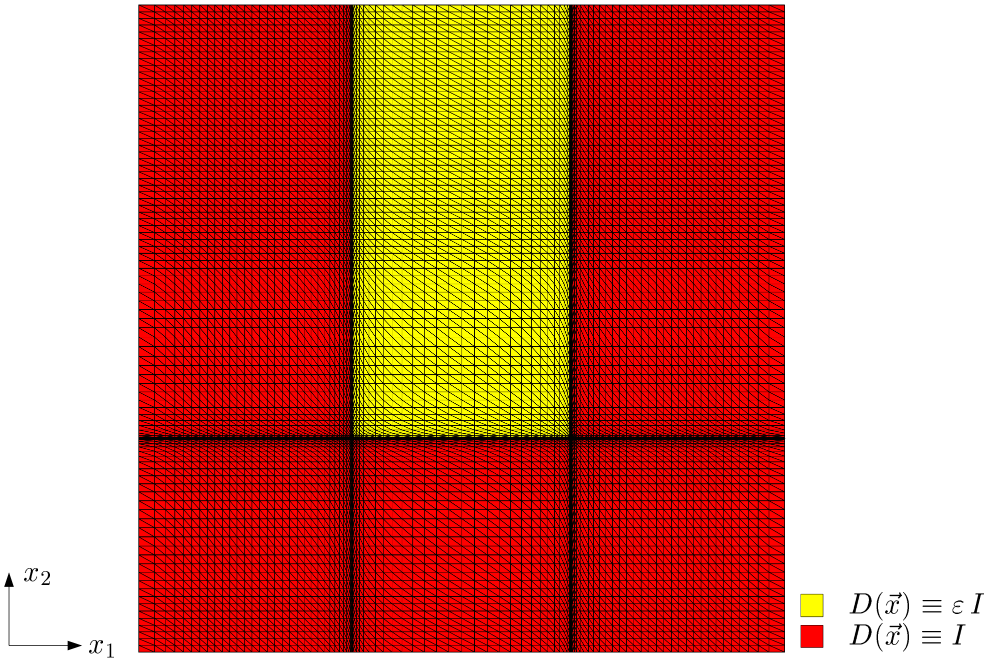


Figure 4.4: Regions of constant coefficient $D(\vec{x})$ and graded mesh $\mathcal{T}$ for Example 2.

The results reflect exactly the theory developed in Graham & Hagger [45, 46]. Note first that the Neumann boundary condition $\vec{u} \cdot \vec{\nu} = 0$ on Γ_N is an essential condition in the mixed formulation. Therefore Γ_N corresponds to the Dirichlet boundary for the bilinear form $a(\cdot,\cdot)$ underlying the coefficient matrix $\mathring{A}$ (recall (4.29) for the definition of $a(\cdot,\cdot)$). Furthermore as $\varepsilon \to 0$ in (4.29), the (diffusion) coefficient $\mathcal{D}(\vec{x})^{-1}$ in $a(\cdot,\cdot)$ tends to ∞, for $\vec{x} \in \left(\frac{1}{3}, \frac{2}{3}\right) \times \left(\frac{1}{3}, 1\right)$ (conversely as $\varepsilon \to \infty$, $\mathcal{D}(\vec{x})^{-1} \to 0$ on $\left(\frac{1}{3}, \frac{2}{3}\right) \times \left(\frac{1}{3}, 1\right)$).

Now, by the theory in [45, 46], for this particular choice of Dirichlet boundary and (diffusion) coefficient and for any of the three preconditioners in Table 4.5, the preconditioned matrix $\mathcal{P}^{-1}\mathring{A}$ should have one "bad" eigenvalue which approaches 0 as $\varepsilon \to 0$, independent of the mesh $\mathcal{T}$. The remaining spectrum should be bounded away from 0 independently of ε. And indeed, as can be seen in Table 4.6, the smallest eigenvalue of $(\mathrm{Diag}(\mathring{A})^{-1}\mathring{A})$ tends to 0 linearly with ε, while the remaining eigenvalues stay bounded independent of ε. As explained in [45, 46], this leads to a logarithmic growth with ε^{-1} of the number of iterations of our decoupled method for diagonal scaling and for the one-level additive Schwarz preconditioner (Rows 3–9 and Columns 4–5 of Table 4.5), which has to be compared to the linear growth with ε^{-1} of the condition numbers of $\mathring{A}$ and of $(\mathrm{Diag}(\mathring{A})^{-1}\mathring{A})$. Moreover, the results for the two-level additive Schwarz preconditioner (Column 6), which show only a very mild dependency on the discontinuity and remain bounded when $\varepsilon \to 0$, are a further indication that a sufficiently well-designed coarse mesh may completely remove the dependence on the coefficient, even when it does not resolve the discontinuity (see the remark at the end of [45]). Recall from Section 4.2.4 that in `DOUG` the coarse mesh is produced automatically,

	Condition Number		Iterations		
ε	$\kappa(\mathring{A})$	$\kappa\left(\mathrm{Diag}(\mathring{A})^{-1}\mathring{A}\right)$	$\mathrm{Diag}(\mathring{A})^{-1}$	$\mathcal{P}^{-1}_{AS(1)}$	$\mathcal{P}^{-1}_{AS(2)}$
10^{-6}	$1.1 \cdot 10^{11}$	$9.2 \cdot 10^{9}$	1808	122	36
10^{-5}	$1.1 \cdot 10^{10}$	$9.2 \cdot 10^{8}$	1745	118	36
10^{-4}	$1.1 \cdot 10^{9}$	$9.2 \cdot 10^{7}$	1688	112	36
10^{-3}	$1.1 \cdot 10^{8}$	$9.2 \cdot 10^{6}$	1614	107	38
10^{-2}	$1.0 \cdot 10^{8}$	$9.6 \cdot 10^{5}$	1539	103	38
10^{-1}	$1.0 \cdot 10^{8}$	$1.9 \cdot 10^{5}$	1480	96	37
10^{0}	$1.2 \cdot 10^{8}$	$1.9 \cdot 10^{5}$	1403	99	38
10^{+1}	$1.6 \cdot 10^{8}$	$2.3 \cdot 10^{5}$	1343	93	37
10^{+2}	$7.3 \cdot 10^{8}$	$2.4 \cdot 10^{5}$	1420	80	32
10^{+3}	$6.9 \cdot 10^{9}$	$2.5 \cdot 10^{5}$	1419	76	32
10^{+4}	$6.9 \cdot 10^{10}$	$2.5 \cdot 10^{5}$	1417	72	31
10^{+5}	$6.9 \cdot 10^{11}$	$2.5 \cdot 10^{5}$	1414	68	31
10^{+6}	$6.9 \cdot 10^{12}$	$2.5 \cdot 10^{5}$	1413	63	31

Table 4.5: Performance of the decoupled method for Example 2 in the presence of discontinuous coefficients (iteration count for the velocity system).

ε	$\lambda_{min}(\mathrm{Diag}(\mathring{A})^{-1}\mathring{A})$	$\leq$	$\lambda_2(\mathrm{Diag}(\mathring{A})^{-1}\mathring{A})$	$\leq \ldots \leq$	$\lambda_{max}(\mathrm{Diag}(\mathring{A})^{-1}\mathring{A})$
10^{0}	$1.1 \cdot 10^{-5}$		$1.3 \cdot 10^{-5}$	...	2.1
10^{-1}	$1.1 \cdot 10^{-5}$		$1.3 \cdot 10^{-5}$	...	2.1
10^{-2}	$2.2 \cdot 10^{-6}$		$1.3 \cdot 10^{-5}$	...	2.1
10^{-3}	$2.3 \cdot 10^{-7}$		$1.3 \cdot 10^{-5}$	...	2.1
10^{-4}	$2.3 \cdot 10^{-8}$		$1.3 \cdot 10^{-5}$	...	2.1
10^{-5}	$2.3 \cdot 10^{-9}$		$1.3 \cdot 10^{-5}$	...	2.1
10^{-6}	$2.3 \cdot 10^{-10}$		$1.3 \cdot 10^{-5}$	...	2.1

Table 4.6: Spectrum of the diagonally scaled matrix $(\mathrm{Diag}(\mathring{A})^{-1}\mathring{A})$ when $\varepsilon \to 0$ in Example 2.

and does not necessarily resolve the discontinuity.

If $\varepsilon \to \infty$, on the other hand, then the spectrum of $(\mathrm{Diag}(\mathring{A})^{-1}\mathring{A})$ is bounded independent of ε, again in correspondence with the theory in [45, 46], and therefore the number of iterations for our decoupled method does not increase for any of the three preconditioners (Rows 9–15 in Table 4.5).

We can conclude that the decoupled method with additive Schwarz preconditioner is very robust in the presence of large discontinuities in $D(\vec{x})$, and we will see in Section 5.2.4 that this robustness can be maintained even in the extreme case, where $D(\vec{x})$ is a random field.

Large aspect ratios - violation of the shape regularity condition

The grading of the mesh in Example 2 results in very elongated and thin elements near the discontinuity. In fact, the largest aspect ratio (the ratio between the lengths of the longest and the shortest edge of an element) of any of the elements of the mesh depicted in Figure 4.4 is 64.9. This is a very common feature in real applications, such as groundwater flow problems, and we will encounter even worse situations in the examples considered in Chapter 5.

Such meshes represent an extreme test for any iterative solver. We will investigate the behaviour of our decoupled method in this situation, by solving Example 2 with $\varepsilon = 10^{-6}$, on a sequence of graded tensor product meshes $\mathcal{T}_h$ of the form depicted in Figure 4.4, which are obtained by varying the number N of subdivisions in each coordinate direction (or equivalently by varying the mesh diameter h). For every element $T \in \mathcal{T}_h$, let $h(T)$ be the length of the longest edge of T and let $h_{min}(T)$ be the length of the shortest edge of T. As we decrease h, the elements will become more and more elongated near the discontinuity, and the maximum aspect ratio

$$\rho_{max} := \max_{T \in \mathcal{T}_h} \frac{h(T)}{h_{min}(T)} \tag{4.89}$$

will grow, thus violating the shape regularity condition (2.35).

The results of this experiment are presented in Table 4.7. The work which is required

	Mesh Parameters			Dimensions			Iterations	
N	h	h_{min}	ρ_{max}	$n_{\mathcal{V}} + n_{\mathcal{W}}$	$\mathring{n}$	$\kappa(\mathring{A})$	$\mathcal{P}_{AS(1)}^{-1}$	$\mathcal{P}_{AS(2)}^{-1}$
24	0.095	0.011	9.6	2872	568	$1.2 \cdot 10^{9}$	1	1
48	0.050	0.0022	24.3	11504	2288	$1.1 \cdot 10^{10}$	26	17
96	0.029	0.00051	64.9	46048	9184	$1.1 \cdot 10^{11}$	122	36
192	0.016	0.00012	166.1	184256	36800	$8.9 \cdot 10^{11}$	329	59

Table 4.7: Performance of the decoupled method for Example 2 for increasing aspect ratios (iteration count for the solution of the velocity system).

to set up the decoupled velocity system (4.7) and to recover the vector of pressures $\mathbf{p}$ from (4.11), is independent of the aspect ratio and grows only linearly with the number of unknowns (as in Example 1). Thus Table 4.7 deals only with the solution of (4.7) by preconditioned conjugate gradients (PCG).

The results show that in each level of refinement in our experiment, the number of unknowns grows by a factor of about 4 (Columns 5 and 6), while the aspect ratio grows by a factor of about 2.5 (Column 4). This implies a relationship roughly like $\rho_{max} = O(\mathring{n}^{2/3})$, and therefore violates the shape regularity condition (2.35), which is crucial in proving the theoretical results on our method. Moreover, the condition number of $\mathring{A}$ grows by a factor of about 8 (Column 7), which corresponds to $\kappa(\mathring{A}) = O(\mathring{n}^{3/2})$. In comparison, if the mesh was uniform (as in Example 1), the condition number would grow only linearly in the number of unknowns.

Considering these facts, our method is performing surprisingly well. If we apply the one-level additive Schwarz preconditioner $\mathcal{P}_{AS(1)}^{-1}$ to the matrix $\mathring{A}$, then the number of iterations for the PCG method grows by a factor of about 2.7 (Column 8). This suggests an asymptotic growth of about $O(\mathring{n}^{3/4}) = O(\kappa(\mathring{A})^{1/2})$, as expected. If we apply the two-level preconditioner $\mathcal{P}_{AS(2)}^{-1}$, on the other hand, then the number of iterations grows by a factor of about 1.6 (Column 7), which suggests an asymptotic growth of about $O(\mathring{n}^{1/3}) = O(\kappa(\mathring{A})^{2/9})$. This is only slightly worse than in the uniform mesh case, where we would expect an asymptotic growth of $O(\kappa(\mathring{A})^{1/6})$.

However, the amount of data is very restricted and the asymptotics are hard to estimate. It might be the case that for finer grids an asymptotic growth of $O(\kappa(\mathring{A})^{1/6})$ is possible even in the presence of large aspect ratios. Nevertheless, it is safe to conclude that our method performs extremely well in the presence of large aspect ratios, and the results in Chapter 5 (where we encounter aspect ratios of up to 3547) confirm this point.

Parallel performance

For tests on the parallel performance of our decoupled method with the additive Schwarz preconditioner, we refer to Sections 5.1.3 and 5.2.4 in the next chapter, where we will extensively investigate this point for up to 14 processors.

4.5.2 The three-dimensional case

Now, let us look at the three-dimensional case, i.e. $\Omega = (0,1)^3$:

Definition 4.19 (Example 3 and Example 4).
Let

$$D(\vec{x}) \equiv I, \quad \text{for all} \quad \vec{x} \in \Omega,$$

and

$$g_D(\vec{x}) := 1 - x_1, \quad \text{for all} \quad \vec{x} := (x_1, x_2, x_3)^T \in \Gamma_D,$$

so that problem (2.1) corresponds (as in Example 1 for 2D) to a Poisson problem on the unit cube with mixed boundary conditions. The difference between Example 3 and Example 4 lies in the partitioning of the boundary Γ. In Example 3 we let

$$\Gamma_N = [0,1]^2 \times \{0\} \ \cup \ [0,1] \times \{0,1\} \times [0,1] \qquad \text{and} \qquad \Gamma_D = \Gamma \backslash \Gamma_N.$$

In Example 4, on the other hand, we let

$$\Gamma_N = [0,1]^2 \times \{0\} \ \cup \ [0,1] \times \{0,1\} \times [0,1] \ \cup \ \{0,1\} \times [0,1]^2 \qquad \text{and} \qquad \Gamma_D = \Gamma \backslash \Gamma_N.$$

The Neumann boundary Γ_N is illustrated in Figure 4.6 for each case.

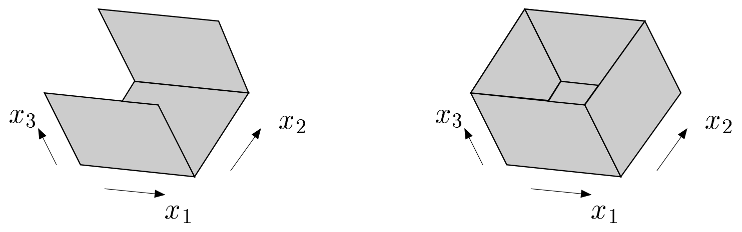


Figure 4.6: The Neumann boundary Γ_N for Examples 3 and 4, respectively

As in 2D, we discretise these problems using the mixed finite element discretisation (2.29) with lowest-order Raviart-Thomas-Nédélec elements on a sequence of uniform meshes $\mathcal{T}_h$ obtained by firstly dividing Ω into N^3 equal cubes $\left(\frac{i-1}{N}, \frac{i}{N}\right) \times \left(\frac{j-1}{N}, \frac{j}{N}\right) \times \left(\frac{k-1}{N}, \frac{k}{N}\right)$, and then further subdividing each cube into 6 tetrahedra (so that the mesh diameter $h = \sqrt{3}N^{-1}$).

To solve the resulting saddle point system (4.2) we use the decoupled iterative method described above: the construction of the matrix $\mathring{A}$ in the decoupled velocity system (4.7) is carried out in an elementwise fashion as presented in (4.48); for each $\mathcal{T}_h$ the spanning tree is chosen to be the "good" spanning tree $\mathbf{H}_h^+$ depicted in Figure 4.1; the resulting symmetric positive definite system (4.7) is solved with preconditioned conjugate gradients (PCG); the matrix A^c in the decoupled pressure system (4.11) is obtained (as in 2D) from the original matrix B in (4.2) by deleting some rows and reordering the rows and columns; and the resulting triangular system (4.11) is solved by simple back substitutions again. In these 3D examples, we will test our method in two case: (i) using no preconditioning for $\mathring{A}$ or (ii) using an incomplete LU decomposition with zero fill-in (ILU(0)). The convergence criterion in the PCG method is the relative reduction of the preconditioned residual by a factor of 10^{-5}.

Asymptotic behaviour – the h-dependency

Tables 4.8 and 4.9 show the performance of our method for Examples 3 and 4 when the mesh is refined. First in Table 4.8 we see that when we increase the number of degrees

	Example 3 (MFlops)			Example 4 (MFlops)		
N	$n_{\mathcal{V}} + n_{\mathcal{W}}$	Decoupling	Recovery	$n_{\mathcal{V}} + n_{\mathcal{W}}$	Decoupling	Recovery
2	144	0.032	0.0021	128	0.024	0.0016
4	1152	0.34	0.020	1088	0.29	0.017
8	9216	3.0	0.17	8960	2.8	0.16
16	73728	25.5	1.4	72704	24.7	1.36

Table 4.8: Performance of the decoupled method for Examples 3 and 4 (floating point operations for the decoupling process and for the recovery of the pressure).

of freedom in (4.2) (or equivalently when we reduce the mesh diameter $h := \sqrt{3}N^{-1}$), the work required to set up the decoupled system (4.7) and to recover the vector of pressures $\mathbf{p}$ from (4.11) is asymptotically optimal, i.e. the number of *floating point operations* (Flops) that are necessary for these processes are growing linearly with the number of degrees of freedom $n_{\mathcal{V}} + n_{\mathcal{W}}$ in (4.2).

In Table 4.9, on the other hand, we investigate how the PCG method for the decoupled velocity system (4.7) is affected by a reduction of the mesh diameter h. As

	Example 3 (Iterations)				Example 4 (Iterations)			
N	$\mathring{n}$	$\kappa(\mathring{A})$	No Prec.	ILU(0)	$\mathring{n}$	$\kappa(\mathring{A})$	No Prec.	ILU(0)
2	48	$1.5 \cdot 10^2$	42	14	32	$8.7 \cdot 10^1$	25	9
4	384	$6.1 \cdot 10^2$	124	26	320	$4.5 \cdot 10^2$	80	18
8	3072	$2.4 \cdot 10^3$	265	45	2816	$2.1 \cdot 10^3$	196	35
16	24576	$9.2 \cdot 10^3$	523	97	23552	$8.7 \cdot 10^3$	430	75

Table 4.9: Performance of the decoupled method for Example 3 and 4 (iteration count for the solution of the velocity system by the PCG method).

predicted in Section 4.3.2, the dimension $\mathring{n}$ of the reduced system (4.7) in 3D is about 3 times smaller than that of the full system (4.2) (compare Columns 2 and 5 of Table 4.8 with Columns 2 and 6 of Table 4.9, respectively).

More interestingly, the condition number of the coefficient matrix $\mathring{A}$ in (4.7) grows, as predicted, like $O(\mathring{n}^{2/3}) = O(N^2) = O(h^{-2})$ (Columns 3 and 7). This underlines our conjecture made in Section 4.3.2 that for the "good" family $\{\mathbf{H}_h^+\}$ of spanning trees in Figure 4.1 the Poincaré inequality (4.54) holds true with α independent of h, and that therefore (4.7) behaves like a second order elliptic problem. Consequently, the number of iterations for the unpreconditioned conjugate gradient method grows (with the square root of the condition number) like $O(\mathring{n}^{1/3}) = O(N) = O(h^{-1})$ (Columns 4 and 8). The iteration counts in Columns 5 and 9 finally correspond to the ILU(0)-preconditioned CG method. While we can see a definite improvement compared to the unpreconditioned case, the number of iterations still grows like $O(h^{-1})$. Although the effect of the ILU preconditioner deteriorates as the grid size decreases, it is extremely cheap to invert and reduces the number of iterations considerably. It remains a cost effective way of preconditioning this system. In future it would be useful to carry out a more detailed investigation of various preconditioners for this system (like the additive Schwarz preconditioner which we successfully used in 2D).

Comparison with MINRES

As in 2D, we compare the performance of our decoupled method to the MINRES algorithm (see Figure 2.2), applied to the full coupled saddle point system (4.2). We will only explicitly present our results for Example 4.

The asymptotically optimal, symmetric positive definite, block diagonal preconditioner $\mathcal{P}_{RW}^{-1}$ defined in (4.86) was only presented and analysed for 2D in Rusten & Winther [82], but the analysis can be extended in a straightforward way to the three-dimensional case. As before, we will use its ILU(0) factorisation and the ILU(0) factorisation of the modified preconditioner $\mathcal{P}_{MRW}^{-1}$ defined in (4.87) to precondition the MINRES algorithm. As for PCG, in 3D the convergence criterion for MINRES is the relative reduction of the preconditioned residual by a factor of 10^{-5}.

The results are given in Tables 4.10 and 4.11 for the unpreconditioned and preconditioned version of MINRES, respectively. As predicted in Section 2.3.2, the condition

	Condition Number		Iterations		MFlops	
N	$\kappa(\mathcal{M})$	$\kappa(\mathring{A})$	MINRES	Decoupled	MINRES	Decoupled
2	$2.2 \cdot 10^1$	$8.7 \cdot 10^1$	99	25	0.41	0.047
4	$4.7 \cdot 10^1$	$4.5 \cdot 10^2$	237	80	8.7	1.12
8	$9.9 \cdot 10^1$	$2.1 \cdot 10^3$	483	196	150	21.9
16	–	$8.7 \cdot 10^3$	955	430	2430	386

Table 4.10: Comparison of the decoupled method (using unpreconditioned CG for the velocity system) with unpreconditioned MINRES for Example 4.

number of the coefficient matrix $\mathcal{M}$ in (4.2) grows like $O(N) = O(h^{-1})$ (Column 2 of Table 4.10), in contrast to the $O(N^2) = O(h^{-2})$ growth of $\kappa(\mathring{A})$ in the reduced system (Column 3 of Table 4.10). However, since the number of iterations for MINRES grows linearly in the condition number, the method is affected by mesh refinement in the same way as the decoupled method, in that the number of iterations grows like $O(N) = O(h^{-1})$ in the unpreconditioned (Column 4 of Table 4.10), as well as in the preconditioned case (Columns 2 and 4 of Table 4.11).

	MINRES ($\mathcal{P}_{RW}^{-1}$)		MINRES ($\mathcal{P}_{MRW}^{-1}$)		Decoupled	
N	Iterations	MFlops	Iterations	MFlops	Iterations	MFlops
2	42	0.28	28	0.19	9	0.038
4	105	6.6	50	3.2	18	0.62
8	237	127	93	50	35	8.7
16	557	2460	163	723	75	132

Table 4.11: Comparison of the decoupled method (using ILU(0)-preconditioned CG) with preconditioned MINRES (Rusten & Winther [82]) for Example 4.

Again, to get an idea of how well our decoupling strategy works, the results of MINRES are compared with the results of the decoupled method using a comparable method to solve the velocity system (4.7). In Table 4.10 we compare the number of iterations and the number of floating point operations for the unpreconditioned version of MINRES with the decoupled method using unpreconditioned CG to solve the velocity system, and in Table 4.11 we compare the results for the preconditioned version of MINRES (using the ILU(0) factorisation of $\mathcal{P}_{RW}^{-1}$ and $\mathcal{P}_{MRW}^{-1}$) with the decoupled method using ILU(0)-preconditioned CG for the velocity system. In both cases we can see a clear advantage of the decoupled method. Although the benefit is not as dramatic as in the 2D examples, the work is still reduced by a more than significant factor. On the finest mesh, this factor is still as large as 6.3 in the unpreconditioned case, and as large as 5.5, if the modified preconditioner $\mathcal{P}_{MRW}^{-1}$ is used in the MINRES algorithm. Note that this improvement factor is much higher, if we use the "standard" preconditioner $\mathcal{P}_{RW}^{-1}$ for (4.2).

The significance of the choice of spanning tree

In the experiments above, to find the basis $\{\vec{\Psi}_E : E \in (\mathcal{E}_I \cup \mathcal{E}_D)\backslash\mathcal{H}\}$ for the divergence-free Raviart-Thomas-Nédélec elements (necessary in the decoupling process of system (4.2) in 3D), we chose the "good" spanning tree $\mathbf{H}_h^+ := (\mathcal{N}_h, \mathcal{H}_h^+)$ (as depicted in Figure 4.1) on each mesh $\mathcal{T}_h$. As predicted in Section 4.3.2, in this case the condition number of the coefficient matrix $\mathring{A} = [\mathring{A}_{E,E'}]_{E,E'\in(\mathcal{E}_I\cup\mathcal{E}_D)\backslash\mathcal{H}_h^+}$ in (4.7) grew like $O(\mathring{n}^{2/3}) = O(h^{-2})$ (see for example Columns 3 and 7 of Table 4.9). This underlines our conjecture made in Section 4.3.2 that for the "good" family $\{\mathbf{H}_h^+\}$ of spanning trees in Figure 4.1, the Poincaré inequality (4.54) holds true with α independent of h.

However, for an arbitrary family $\{\mathbf{H}_h = (\mathcal{N}_h, \mathcal{H}_h)\}$ of spanning trees this is not the case, as we see in Table 4.12. In this table for Example 4, we present the smallest

	"Good" Tree		Arbitrary Spanning Trees					
N	$\lambda_{min}(\mathring{A})$	$\zeta(\mathbf{H}_h^+)$	$\lambda_{min}(\mathring{A})$	$\zeta(\mathbf{H}_h^1)$	$\lambda_{min}(\mathring{A})$	$\zeta(\mathbf{H}_h^2)$	$\lambda_{min}(\mathring{A})$	$\zeta(\mathbf{H}_h^3)$
2	$2.7 \cdot 10^{-1}$	21	$4.5 \cdot 10^{-1}$	13	$3.3 \cdot 10^{-1}$	13	$2.2 \cdot 10^{-1}$	21
4	$1.2 \cdot 10^{-1}$	45	$7.8 \cdot 10^{-2}$	61	$7.2 \cdot 10^{-2}$	71	$4.4 \cdot 10^{-2}$	85
8	$5.5 \cdot 10^{-2}$	93	$1.1 \cdot 10^{-2}$	377	$9.6 \cdot 10^{-3}$	391	$6.4 \cdot 10^{-3}$	623
16	$2.6 \cdot 10^{-2}$	–	$1.3 \cdot 10^{-3}$	–	$1.2 \cdot 10^{-3}$	–	$8.2 \cdot 10^{-4}$	–

Table 4.12: The relationship between $\lambda_{min}(\mathring{A})$ and the quantity $\zeta(\mathbf{H})$ defined in (4.71), for the "good" family $\{\mathbf{H}_h^+\}$ of spanning trees, as well as for the arbitrary families $\{\mathbf{H}_h^1\}$, $\{\mathbf{H}_h^2\}$, $\{\mathbf{H}_h^3\}$ of spanning trees in Example 4.

eigenvalues of $\mathring{A} = [\mathring{A}_{E,E'}]_{E,E'\in(\mathcal{E}_I\cup\mathcal{E}_D)\backslash\mathcal{H}_h}$ for different choices of the spanning tree $\mathbf{H}_h$ on each triangulation $\mathcal{T}_h$. The first one is the "good" tree $\mathbf{H}_h^+$, depicted in Figure 4.1. In fact, this tree is also obtained by applying Algorithm B.7 with the numbering of the nodes of the mesh taken to be that provided by the mesh generator. The three other spanning trees $\mathbf{H}_h^1$, $\mathbf{H}_h^2$, and $\mathbf{H}_h^3$ were obtained using Algorithm B.7 with a different (random) numbering of the nodes each time. Indeed for $\{\mathbf{H}_h^+\}$, the smallest eigenvalue of $\mathring{A}$ tends to zero linearly with h (Column 2) in correspondence with the bound (4.61) established in the proof to Theorem 4.7. However, looking at Columns 4, 6, and 8, on the other hand, we can see that for an arbitrary family of spanning trees $\{\mathbf{H}_h\}$ (with $\mathbf{H}_h \in \{\mathbf{H}_h^1, \mathbf{H}_h^2, \mathbf{H}_h^3\}$ chosen randomly on each $\mathcal{T}_h$), the smallest eigenvalue of $\mathring{A}$ tends to zero at least with $O(h^2)$, which violates (4.61), and suggests that in this case the Poincaré inequality (4.54) does not hold true with α independent of h. Moreover, since the largest eigenvalue of $\mathring{A}$ does not depend on the choice of spanning tree (cf. the proof to Theorem 4.7), this also means that $\kappa(\mathring{A}) \geq O(h^{-3})$. Consequently, the number of iterations necessary to solve (4.7) with preconditioned conjugate gradients and ILU(0) preconditioner in these cases, grows faster than with $O(h^{-3/2})$ (e.g. 458 iterations for $\mathbf{H}_h = \mathbf{H}_h^3$, when $N = 16$, compared to 123 iterations, when $N = 8$).

For each tree $\mathbf{H}_h$ in Table 4.12, we also present the quantity $\zeta(\mathbf{H}_h)$ (Columns 3, 5, 7, and 9), which we defined in (4.71), and which turned out to play an essential rôle in the conditioning analysis of $\mathring{A}$ in Section 4.3.2. The first thing we note is the large variation in $\zeta(\mathbf{H}_h)$ for different choices of $\mathbf{H}_h$ on a fixed mesh $\mathcal{T}_h$ (see for example Row 5, for $N = 8$). However, we also note that an increase in $\zeta(\mathbf{H}_h)$ is directly coupled to a decrease in $\lambda_{min}(\mathring{A})$, and so $\zeta(\mathbf{H}_h)$ gives us an indication of whether or not a mesh is "good" or "bad" for our decoupling method. Obviously it would not be feasible to search for the tree $\mathbf{H}_h^{opt}$ which minimises $\zeta(\mathbf{H}_h)$ over all possible trees $\mathbf{H}_h$ associated with $\mathcal{T}_h$, since the total number of trees associated grows exponentially with the number of nodes of the mesh. Nevertheless, our experiments show that (at least for the examples considered here) the family $\{\mathbf{H}_h^+\}$ of trees that we obtain with Algorithm B.7 using the "natural" (geometrically based) numbering of the nodes provided by the mesh generator, is sufficient for the Poincaré inequality (4.54) to hold with α independent of h, and nothing more is necessary.

Non-uniform triangulations

To be confident about applying our decoupled method to more difficult geometries, it is necessary to have an idea about its performance on unstructured simplicial triangulations. For that reason we repeat the tests for Examples 3 and 4 above, on a sequence of non-uniform triangulations $\mathcal{T}_h$ obtained by firstly dividing Ω into N^3 non-uniform hexahedra, and then further subdividing each hexahedron into 24 tetrahedra (see Figure 4.7). This leads to a quasi-uniform family $\{\mathcal{T}_h\}$ of triangulations with $h := 1.4 \cdot N^{-1}$ and $h_{min} := 0.3 \cdot N^{-1}$. Again, the spanning tree $\mathbf{H}_h$ for each triangulation $\mathcal{T}_h$, is calculated with Algorithm B.7 using the numbering of the nodes provided by the mesh generator.

As in the uniform mesh case, the work required to set up the decoupled system (4.7) and to recover the vector of pressures $\mathbf{p}$ from (4.11) is asymptotically optimal, and grows only linearly with the number of degrees of freedom in (4.2). Therefore, we will only present the results for the solution of the velocity system (4.7) in Example 4 explicitly, and compare the results with preconditioned MINRES using the modified preconditioner $\mathcal{P}_{MRW}^{-1}$ (see Table 4.13).

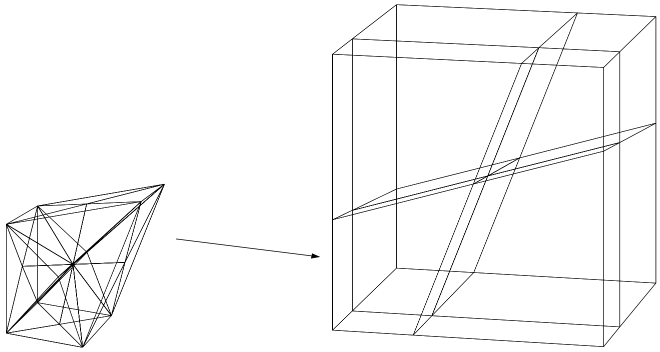


Figure 4.7: Non-uniform triangulation of the unit cube (for $N = 2$).

	MINRES ($\mathcal{P}_{MRW}^{-1}$)				Decoupled Method			
N	$n_{\mathcal{V}} + n_{\mathcal{W}}$	$\kappa(\mathcal{M})$	Its.	MFlops	$\mathring{n}$	$\kappa(\mathring{A})$	Its.	MFlops
2	544	$1.8 \cdot 10^2$	62	2.0	160	$3.6 \cdot 10^2$	24	0.35
4	4480	$5.2 \cdot 10^2$	119	32.2	1408	$2.8 \cdot 10^3$	80	8.0
8	36352	–	240	540	11776	$2.1 \cdot 10^4$	187	145

Table 4.13: Comparison of the decoupled method (using ILU(0)-preconditioned CG) with preconditioned MINRES (Rusten & Winther [82]) for Example 4 on a sequence of non-uniform meshes.

The reduced system (4.7) is about 3 times smaller than the full coupled system (4.2) again (compare Columns 2 and 6), but the condition numbers $\kappa(\mathcal{M})$ (Column 3) and $\kappa(\mathring{A})$ (Column 7) grow faster for this range of N than asymptotically. On a sequence $\{\mathcal{T}_h\}$ of quasi-uniform triangulations, we would expect them to grow with $O(h^{-1})$ and $O(h^{-2})$, respectively. This is probably a preasymptotic effect: The mesh diameter h is too large and the amount of data is far too small, to say anything about the asymptotic behaviour. However, in the case of $\kappa(\mathring{A})$ it could also mean that the family $\{\mathbf{H}_h\}$ of spanning trees, which we found with Algorithm B.7, does not satisfy the Poincaré inequality (4.54) with α independent of h. Unfortunately we are not able to answer this question, but taking for example $N = 4$, we find $\zeta(\mathbf{H}_h) = 111$, which is significantly smaller than $\zeta(\tilde{\mathbf{H}}_h)$ in the case of an arbitrary spanning tree $\tilde{\mathbf{H}}_h$. (In our tests for $N = 4$ the value $\zeta(\tilde{\mathbf{H}}_h)$, for arbitrary spanning trees $\tilde{\mathbf{H}}_h$, ranged from 165 to 375, and the condition number $\kappa(\mathring{A})$ ranged from $4.6 \cdot 10^3$ to $2.4 \cdot 10^4$.) This suggests that our choice of spanning tree to obtain the results in Table 4.13 was “good”.

As a consequence of the faster growth of the condition numbers, the number of iterations necessary in our decoupled method, to solve (4.7) with ILU(0)-preconditioned CG also grows faster than expected (for the tested range of values of h). Nevertheless, we can still see a clear advantage of our decoupled method over preconditioned MINRES. On the finest mesh, the work is still reduced by a factor of 3.7.

Behaviour with respect to discontinuities in $D(\vec{x})$

Finally, in the last example in this section, we will see that even in 3D the decoupled method is very robust with respect to discontinuities in the coefficient $D(\vec{x})$.

Definition 4.20 (Example 5).
Let $\Omega = (0,1)^3$, and let

$$\Gamma_N = [0,1]^2 \times \{0\} \;\cup\; [0,1]\times\{0,1\}\times[0,1] \;\cup\; \{0\}\times[0,1]^2 \qquad \text{and} \qquad \Gamma_D = \Gamma\backslash\Gamma_N,$$

$$D(\vec{x}) \equiv \begin{cases} \varepsilon\, I & \text{for all} \quad \vec{x} \in \left\{(x_1,x_2,x_3)^T \in \Omega : \frac{1+x_2}{10} < x_3 < \frac{8-2x_1-3x_2}{10}\right\}, \\ I & \text{otherwise,} \end{cases} \tag{4.90}$$

with $\varepsilon \in (0,\infty)$ (see also Figure 4.8), and

$$g_D(\vec{x}) := 1 - \frac{x_1 + x_2}{2} \quad \text{for all} \quad \vec{x} := (x_1,x_2,x_3)^T \in \Gamma_D.$$

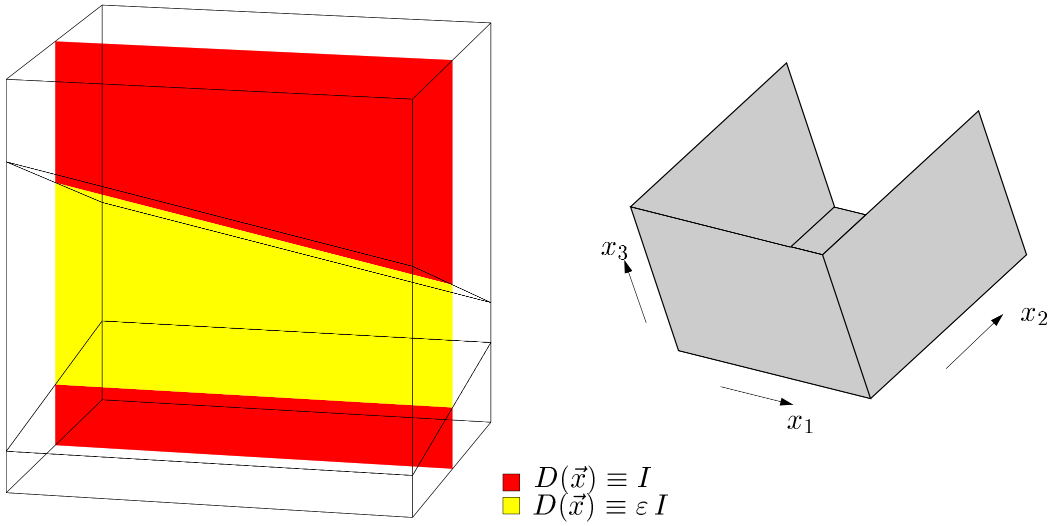


Figure 4.8: Regions of constant coefficient $D(\vec{x})$ (left) and Neumann boundary Γ_N (right) for Example 5.

As in all the previous examples, we discretise this problem using lowest order Raviart-Thomas-Nédélec elements on a simplicial mesh $\mathcal{T}$ with 4374 elements, which corresponds to 12960 degrees of freedom in (4.2) and to 4212 degrees of freedom in the decoupled velocity system (4.7). The tree $\mathbf{H}$ is chosen to be the "good" tree $\mathbf{H}^+$ in Figure 4.1 again. We solve the resulting saddle point system (4.2) using the decoupled iterative method, using the same components as specified for Examples 3 and 4. In addition to the incomplete LU factorisation, we will also use diagonal scaling to precondition the reduced matrix $\mathring{A}$. The convergence criterion is again the relative reduction of the preconditioned residual by a factor of 10^{-5}.

As in 2D for Example 2, we investigate the robustness of the method with respect to discontinuities in $D(\vec{x})$ by varying ε in (4.90) from 10^{-6} to 10^{+6}. The results of this experiment are presented in Table 4.14. The largest jump in the coefficient $D(\vec{x})$ is

	Condition Number		Iterations	
ε	$\kappa(\mathring{A})$	$\kappa\Big(\mathrm{Diag}(\mathring{A})^{-1}\mathring{A}\Big)$	$\mathrm{Diag}(\mathring{A})^{-1}$	ILU(0)
10^{-6}	$2.2 \cdot 10^9$	$6.1 \cdot 10^8$	1717	368
10^{-5}	$2.2 \cdot 10^8$	$6.1 \cdot 10^7$	1496	317
10^{-4}	$2.2 \cdot 10^7$	$6.1 \cdot 10^6$	1236	262
10^{-3}	$2.2 \cdot 10^6$	$6.1 \cdot 10^5$	988	207
10^{-2}	$2.2 \cdot 10^5$	$6.1 \cdot 10^4$	741	150
10^{-1}	$2.2 \cdot 10^4$	$6.8 \cdot 10^3$	445	92
10^{0}	$4.8 \cdot 10^3$	$2.3 \cdot 10^3$	275	62
10^{+1}	$3.3 \cdot 10^4$	$1.2 \cdot 10^4$	424	94
10^{+2}	$3.1 \cdot 10^5$	$1.1 \cdot 10^5$	689	161
10^{+3}	$3.1 \cdot 10^6$	$1.1 \cdot 10^6$	891	204
10^{+4}	$3.1 \cdot 10^7$	$1.1 \cdot 10^7$	1082	258
10^{+5}	$3.1 \cdot 10^8$	$1.1 \cdot 10^8$	1297	314
10^{+6}	$3.1 \cdot 10^9$	$1.1 \cdot 10^9$	1517	366

Table 4.14: Performance of the decoupled method for Example 5 in the presence of discontinuous coefficients (iteration count for the velocity system).

given by

$$\mathcal{J}_{max} := \frac{\max_{\vec{x}\in\Omega} |D(\vec{x})|}{\min_{\vec{x}\in\Omega} |D(\vec{x})|} = \begin{cases} \varepsilon & \text{if } \varepsilon \geq 1, \\ \varepsilon^{-1} & \text{if } \varepsilon < 1. \end{cases} \tag{4.91}$$

Therefore, as $\varepsilon \to 0$ and as $\varepsilon \to \infty$ in (4.90), $\mathcal{J}_{max}$ tends to ∞. Now interestingly, as $\mathcal{J}_{max} \to \infty$, the number of iterations of preconditioned conjugate gradients, which are necessary to solve the decoupled velocity system (4.7), grows only logarithmically with $\mathcal{J}_{max}$ (Columns 4 and 5), even though the condition numbers of $\mathring{A}$ and of $(\mathrm{Diag}(\mathring{A})^{-1}\mathring{A})$ grow linearly with $\mathcal{J}_{max}$ (Columns 2 and 3).

This is similar to the behaviour in 2D (see Example 2 in Section 4.5.1), where $\mathring{A}$ was the matrix analysed in Graham & Hagger [45, 46]. In 3D, $\mathring{A}$ is an entirely different matrix and it would be a useful focus for future work to investigate whether the resilience of preconditioned conjugate gradients in this case can be explained in a similar way.

4.6 Summary

In this chapter we formulated, analysed, and tested an iterative method for solving indefinite saddle point systems of the form (4.1), arising from mixed finite element approximations of second-order elliptic boundary value problems. The central idea in our method was to decouple the velocity unknown $\mathbf{u}$ in (4.1) from the pressure unknown $\mathbf{p}$ by using the basis for the subspace $\mathring{\mathcal{V}}$ of divergence-free Raviart-Thomas-Nédélec elements constructed in Chapter 3.

We introduced the decoupling process as an abstract algebraic procedure for (4.1) (in the case where $\mathbf{f} = 0$) and then analysed the resulting decoupled systems for $\mathbf{u}$ and $\mathbf{p}$ in the particular case where (4.1) represents the approximation of a second-order elliptic boundary value problem using Raviart-Thomas-Nédélec elements. To decouple the velocity unknown $\mathbf{u}$ in (4.1) from the pressure unknown $\mathbf{p}$, we made essential use of the basis for $\mathring{\mathcal{V}}$. The resulting system (4.7) for $\mathbf{u}$ was then symmetric positive definite. Additionally, it was about 5 times smaller than the original system (4.1) in 2D (see Section 4.2.3) and about 3 times smaller in 3D (see Section 4.3.2). Moreover, using the particular basis for the complementary space $\mathcal{V}^c$ of $\mathring{\mathcal{V}}$ found in Section 3.3 of the previous chapter, the resulting system (4.11) for the pressure unknown $\mathbf{p}$ turned out to be triangular (cf. Proposition 4.1), and was solved by simple back substitutions.

The bulk of the computational work lay in the solution of the symmetric positive definite velocity system (4.7). It was shown (in Proposition 4.2 for 2D and in Proposition 4.5 and Corollary 4.5 for 3D) that for general mixed boundary conditions, the computation reduces to the solution of a bordered system (4.35) with the width n_C of the border (i.e. the number of columns in C) depending on the connectivity of the domain Ω and on the partitioning $\Gamma_D \cup \Gamma_N$ of the boundary. For most applications n_C will be small, and in all applications it does not increase as the mesh is refined (e.g. if Γ_D is connected, $n_C = 0!$). Thus we considered solving (4.35) by block elimination, leading to $n_C + 1$ linear systems with a sparse, symmetric positive definite coefficient matrix A, which we then solved by preconditioned conjugate gradients (PCG).

The components in (4.35) can be obtained by elementary row and column operations on a stiffness matrix $\mathcal{A}$ corresponding to a discretisation of the well-known bilinear form $a(u,v) := \int_\Omega \mathcal{D}^{-1}(\vec{x})\vec{\nabla} u \cdot \vec{\nabla} v \, d\vec{x}$ by C^0-elements in 2D (cf. Proposition 4.2), and to a discretisation of the less well-known bilinear form $a(\vec{u},\vec{v}) := \int_\Omega D^{-1}(\vec{x}) \overrightarrow{\mathrm{curl}}\, \vec{u} \cdot \overrightarrow{\mathrm{curl}}\, \vec{v} \, d\vec{x}$ by Nédélec's edge elements in 3D (cf. Proposition 4.5, Corollary 4.9). Therefore for the remainder of the chapter we had to distinguish between 2D and 3D again.

In 2D, because of the crucial link to the bilinear form $a(\cdot,\cdot)$, we were able to propose a fully parallel, overlapping two-level additive Schwarz preconditioner for A (the major block in (4.35)) that guarantees in theory that the number of iterations of the PCG method will grow no faster than $O(n^{1/6})$, where n is the dimension of A (cf. Section 4.2.4). Moreover in the presence of discontinuous coefficients $\mathcal{D}^{-1}(\vec{x})$, the number of PCG iterations will only grow logarithmically in the largest jump of $\mathcal{D}^{-1}(\vec{x})$ in Ω. In a series of experiments in Section 4.5.1 we confirmed numerically the sharpness of these estimates and demonstrated the superior performance of the decoupled method for (4.1) in comparison with Rusten and Winther's preconditioned MINRES algorithm [82]. In view of the applications in Chapter 5, we also tested the behaviour of our decoupled method for meshes with large aspect ratios and noticed only a mild dependency.

In 3D, on the other hand, to obtain the matrix A from $\mathcal{A}$, we had to eliminate the rows and columns corresponding to edges E in a spanning tree $\mathbf{H}$ of the graph $\mathbf{G}$ underlying the triangulation (in order to extract a basis for $\mathring{\mathcal{V}}$ as discussed in Section 3.2.2 of the previous chapter). Obviously the choice of $\mathbf{H}$ is not unique, and we saw in Section 4.3.2 that the condition number of A in fact depends on this choice. The numerical tests in Section 4.5.2 confirmed this dependency. In particular, we showed in Lemma 4.8 that there exists a family of trees $\{\mathbf{H}_h\}$ on a particular sequence of uniform triangulations $\{\mathcal{T}_h\}$, for which the condition number of A increases rather more rapidly than the "usual" rate of h^{-2} for second order problems, as the mesh diameter $h \to 0$ (compare also Theorem 4.7). Nevertheless, in Section 4.3.2 we also established a heuristic criterion on how to choose spanning trees which will not lead to such an increase. This involves the quantity $\zeta(\mathbf{H})$ defined in (4.71). In a series of experiments in Section 4.5.2 we put this criterion to the test and found that (in the tested examples at least) the "natural" tree $\mathbf{H}_h^+$ (induced by the ordering of the nodes provided by the mesh generator) satisfied our criterion. As a consequence, for $\mathbf{H}_h^+$ the number of iterations of (unpreconditioned) conjugate gradients grew linearly with respect to the parameter h^{-1} as in 2D. Moreover, using an ILU preconditioner for A, we were able to outperform preconditioned MINRES in 3D as well, and achieve a resilience to discontinuous coefficients similar to the 2D case.

Finally we discussed, at least in theory, the case $\mathbf{f} \neq \mathbf{0}$ in (4.1). It was shown in Ewing & Wang [38, 39] that for $\Gamma_D = \emptyset$, the case $\mathbf{f} \neq \mathbf{0}$ can be reduced to the case $\mathbf{f} = \mathbf{0}$, if a particular solution $\mathbf{u}^*$ to the constraint equation $B^T\mathbf{u}^* = \mathbf{f}$ is known. They find this particular solution through a variation of domain decomposition and multilevel methods in asymptotically optimal time, proportional to the number of unknowns. In Section 4.4 we extend their results to the case of mixed boundary conditions (cf. Theorems 4.12 and 4.16).

Literature related to our decoupled iterative method can already be found in [21, 24, 38, 39, 59, 71, 72]. In this chapter we presented for the first time a unified theory for the method, which was only possible with the results of Chapter 3 of this thesis in hand. In particular, we extended the method to mixed boundary conditions and to multiply connected domains in 2D, and in the case of lowest order elements, to (unstructured) simplicial triangulations on simply connected three-dimensional domains (with mixed boundary conditions). Additionally, we also provided for the first time a simple algorithm to recover the pressure in optimal time.

While our theoretical and numerical results clearly showed the extraordinary potential of this decoupled iterative method for our saddle point systems in 2D, the theory in 3D is far from being complete. The most pressing issue which needs to be addressed in 3D is the construction of a robust and efficient preconditioner for the matrix A, but it would also be of interest to prove the Poincaré inequality (4.54) (for a particular choice of spanning trees), or to explain theoretically the resilience of the PCG method to discontinuous coefficients in 3D.

Chapter 5

Applications in Groundwater Flow

The application we particularly had in mind, when we constructed the method described in Chapter 4, is the modelling of single phase flow in saturated porous media. The flow of fluids in the rocks comprising the earth's crust is important in a number of technological and industrial fields, most notably the hydrocarbon and water resources industries. In the former, one is motivated to understand the underground flow of oil (and gas) in order to recover as much of this resource as possible. In the latter, a proper understanding of the flow of groundwater and of the transport of chemicals in it is essential not only for good resource management and quality control but also for applications in pollution modelling. One option for the long-term disposal of radioactive waste is storage in an underground repository. In order to scientifically assess the safety of this option it is necessary to model the transport of radionuclides in flowing groundwater. Thus this topic is of general environmental importance.

From now on we restrict our attention to groundwater flow. Let Ω be a bounded two or three-dimensional domain. Then the classical equations governing groundwater flow in the steady-state case are *Darcy's Law*,

$$\vec{q}(\vec{x}) = -\frac{k(\vec{x})}{\mu}\,\vec{\nabla} p_R(\vec{x}), \qquad \text{for all } \vec{x} \in \Omega, \tag{5.1}$$

and the *incompressibility constraint*,

$$\operatorname{div}(\,\vec{q}(\vec{x})\,) = 0, \qquad \text{for all } \vec{x} \in \Omega, \tag{5.2}$$

subject to appropriate boundary conditions. Here $k(\vec{x})$ is the *permeability tensor* for the porous medium; μ is the *viscosity* (assumed constant over the entire domain), $\vec{q}(\vec{x})$ is the *specific discharge* (Darcy velocity), and $p_R(\vec{x})$ is the *residual pressure* of the fluid. The actual pressure is $p_R - \rho g z$, where z is the *fluid height*, ρ is the *density* of the fluid and g is the acceleration due to gravity.

The boundary Γ of Ω is assumed to be partitioned into $\Gamma_D \cup \Gamma_N$, so that it is possible to prescribe the residual pressure on part of the boundary, i.e.

$$p_R(\vec{x}) = g_D(\vec{x}), \qquad \text{for all } \vec{x} \in \Gamma_D, \tag{5.3}$$

and to prohibit flux otherwise, i.e.

$$\vec{q}(\vec{x}) \cdot \vec{\nu}(\vec{x}) = 0, \qquad \text{for all } \vec{x} \in \Gamma_N, \tag{5.4}$$

where $\vec{\nu}(\vec{x})$ denotes the outward unit normal from Ω at $\vec{x} \in \Gamma$.

Assuming sufficient regularity of the functions appearing in (5.1)–(5.4) (as required at the beginning of Chapter 2), we can set $D(\vec{x}) := k(\vec{x})/\mu$ and define

$$m(\vec{q}, \vec{v}) := \int_\Omega \mu k^{-1} \vec{q} \cdot \vec{v} \, d\vec{x}, \tag{5.5}$$

$$b(\vec{v}, w) := \int_\Omega \operatorname{div} \vec{v} \, w \, d\vec{x}, \tag{5.6}$$

$$F_{\mathcal{D}}(\vec{v}) := \langle \vec{v} \cdot \vec{\nu}, g_D \rangle_\Gamma. \tag{5.7}$$

The weak form of (5.1)–(5.4) is to determine $(\vec{q}, p_R) \in H_{0,N}(\operatorname{div}, \Omega) \times L_2(\Omega)$ such that

$$\left.\begin{aligned} m(\vec{q}, \vec{v}) \;+\; b(\vec{v}, p_R) \;&=\; F_{\mathcal{D}}(\vec{v}) \;, &&\text{for all } \vec{v} \in H_{0,N}(\operatorname{div}, \Omega), \\ b(\vec{q}, w) \;&=\; 0 \;, &&\text{for all } w \in L_2(\Omega). \end{aligned}\right\} \tag{5.8}$$

Therefore, in this case, the mixed formulation (2.18) is the natural formulation of (2.1), and a discretisation of (5.8) by Raviart-Thomas-Nédélec elements (as described in Sections 2.2 and 2.3) leads to a linear equation system of the form (4.2), i.e.

$$\begin{pmatrix} M & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{q} \\ \mathbf{p}_R \end{pmatrix} = \begin{pmatrix} \mathbf{g} \\ \mathbf{0} \end{pmatrix} \qquad \text{in} \quad \mathbb{R}^{n_\mathcal{V}} \times \mathbb{R}^{n_\mathcal{W}} \;, \tag{5.9}$$

which is solvable by the decoupled iterative method described in Chapter 4. We will only consider two-dimensional examples here, and we can therefore apply the fast parallel iterative domain decomposition method described in Section 4.2.4 for the core task of solving the resulting decoupled symmetric positive definite system

$$\mathring{A}\mathring{\mathbf{q}} = \mathring{\mathbf{g}} \qquad \text{in} \quad \mathbb{R}^{\mathring{n}} \tag{5.10}$$

(cf. (4.7)). Therefore for the rest of this chapter, <u>let $\Omega \subset \mathbb{R}^2$ and $d = 2$</u>.

The most testing feature of realistic groundwater flow problems for any iterative solution method, is the variation in the permeability tensor k. Not only does k usually vary immensely over the entire domain, but in addition, these variations are not gradual. Commonly, the models provided by geologists are layered media, comprising strata of different rock type, which can have largely differing permeabilities. Furthermore, these strata are usually intersected by fault lines with very large permeabilities, and so in general k will be highly discontinuous. In addition, the creation of an accurate model of a typical underground rock structure often necessitates highly unstructured finite element meshes with a range of element sizes and possibly large aspect ratios. Therefore, in Section 5.1, we are going to test our decoupled method on two of our industrial collaborator AEA Technology's actual case studies from sites in the UK. The first one is a very basic model comprising only three different rock strata. The second one, on the other hand, is a detailed model taken from a recent study of a repository site. We would like to thank United Kingdom Nirex Limited for kindly providing us with the data and the finite element grid for this second model.

These "layered media" models assume that in each layer the permeability k is constant, neglecting the possibility of localised variations in the permeability on a shorter length scale. In some models there is a need for this spatial variation (or *heterogeneity*) to be taken into account. Heterogeneity gives rise to variability in the flow velocity, which in turn would affect the transport of dissolved chemicals or pollutants. This heterogeneity has two main aspects.

1. *Uncertainty*: Because the rock properties are varying in a complicated way and it is not possible to measure the permeability at each point in space, there is inevitably a degree of uncertainty concerning the values of the permeability. However, the permeability is in principle required at every point in order to model the flow of groundwater. Simple-minded interpolation of measured values yields rather inaccurate permeability fields that do not reflect the heterogeneity which is known to be present.

2. *Dispersion*: The heterogeneity means that the velocity field varies on a range of length scales and so, particle paths that are initially close together can become progressively separated. This phenomenon – called *hydrodynamic dispersion* – is the primary mechanism for the spreading of a plume of pollutant as it is transported by the groundwater flow (see Dagan [33]).

A widely used method of treating heterogeneity, capable of dealing with both these aspects, is stochastic modelling, and we will discuss this approach for a model problem in Section 5.2 below. The basic idea is to model the permeability field k as a stochastic spatial process, assuming that a single realisation of this stochastic process is a reasonable representation of the permeability field and that any of the realisations are equally probable given the information available from geological measurements. This approach leads to a system (5.1)–(5.4) of stochastic PDEs, where $\vec{q}$ and p_R are now random variables. Given certain statistical properties of k, it is of interest to study statistical properties of those random variables. Thus, it is essential that we can quickly and efficiently solve the system (5.1)–(5.4) of stochastic PDEs and, most importantly, compute the velocity $\vec{q}$ for each realisation of k.

Although we do not carry out here any statistical analysis involving multiple simulations, a prime motivation of the tests in Section 5.2 is to establish whether our decoupled method is sufficiently accurate and efficient to make such a statistical analysis possible. After discretisation a typical simulation of (5.1)–(5.4) will involve the solution of very large highly ill-conditioned indefinite linear systems of the form (5.9) and the fast parallel iterative method proposed in Chapter 4 constitutes an essential tool which can be used in later statistical analyses.

The results in Section 5.2 were obtained in collaboration with K. A. Cliffe, I. G. Graham, and L. Stals and have already been published in the joint paper [29] (see also Cliffe et al. [28] concerning some of the theoretical results). Other iterative methods for related problems are reported, for example, in Ashby et al. [11] and Wagner et al. [91], although there the emphasis is on finite volume/multigrid techniques.

The applications in this chapter represent an extreme test for our decoupled method, and we are pleased to be able to report good performance of our solver under these circumstances. This performance is even of greater significance, if we take the method's almost optimal parallel efficiency into account.

5.1 Layered media

In this section the permeability tensor $k(\vec{x})$ is modelled as a piecewise constant function, with constant values on relatively large regions compared to the size of the finite element mesh. To be precise, let Ω be partitioned into a family of open subdomains Λ_j (corresponding to the different rock types), with polygonal boundary $\partial\Lambda_j$, such that

$$\overline{\Omega} = \overline{\Lambda}_1 \cup \ldots \cup \overline{\Lambda}_L, \quad \text{and} \quad \Lambda_j \cap \Lambda_{j'} = \emptyset, \quad \text{for all } 1 \leq j \neq j' \leq J.$$

Then, for each $j = 1, \ldots, J$, we assume that

$$k(\vec{x}) := k^{(j)} = \begin{pmatrix} k_{1,1}^{(j)} & k_{1,2}^{(j)} \\ k_{1,2}^{(j)} & k_{2,2}^{(j)} \end{pmatrix}, \qquad \text{for all} \quad \vec{x} \in \Lambda_j,$$

where $k_{1,1}^{(j)} k_{2,2}^{(j)} - \left(k_{1,2}^{(j)}\right)^2 \geq \alpha^{(j)} > 0$. The values of k at the interfaces are arbitrary, and so $k(\vec{x})$ satisfies (2.2) almost everywhere. Consequently, k is invertible almost everywhere and the components of k and k^{-1} are in $L_\infty(\Omega)$, as required.

The prescribed pressure $g_D(\vec{x})$ on the part Γ_D of the boundary, on the other hand, is modelled as a piecewise linear function, and is therefore in $H^{1/2}(\Gamma_D)$, as required. The viscosity μ of the groundwater is taken to be 10^{-3} $\left[\frac{\mathrm{N\,s}}{\mathrm{m}^2}\right]$.

The data and the meshes for the following two examples were generated using the computer package NAMMU (Numerical Assessment Method for Migration Underground) [53], which has been developed and is marketed by our industrial collaborators at AEA

Technology.[1] It can be used to numerically model two and three-dimensional groundwater flow through complicated layered media with varying geological properties, but it also allows the coupling with and the simulation of many other geological processes, like heat transport, radionuclide transport, or transport of salinity.

5.1.1 The Harwell site

The first example is taken from the `NAMMU` User Guide [53, Section 6]. It simulates steady state groundwater flow beneath the site of AEA Technology in Harwell, Oxfordshire, UK. A two-dimensional vertical cross section (SSW – NNE) consisting of three rock types is chosen to model the local geological features. Figure 5.1 shows this simplified model. The cross section is about 13000 [m] long (x_1-direction) and about 222 [m] deep at the left hand side. The various rock strata in the cross section may be broadly classified as comprising layers of high permeability (*chalk* and *corallian*) separated by a layer of low permeability (*clay*). The permeability tensors for each rock type are assumed to be diagonal, in this simplified model, and they are given by

$$\begin{aligned} k^{(chalk)} &= \begin{pmatrix} 3.3 \cdot 10^{-13} & 0 \\ 0 & 3.3 \cdot 10^{-13} \end{pmatrix} [\mathrm{m}^2], \\ k^{(clay)} &= \begin{pmatrix} 1.3 \cdot 10^{-17} & 0 \\ 0 & 8.2 \cdot 10^{-19} \end{pmatrix} [\mathrm{m}^2], \\ k^{(coral)} &= \begin{pmatrix} 5.2 \cdot 10^{-13} & 0 \\ 0 & 5.2 \cdot 10^{-13} \end{pmatrix} [\mathrm{m}^2]. \end{aligned}$$

Therefore the largest jump in k in any component is $6.3 \cdot 10^5$ (on the interface between Clay and Corallian). Note also that the horizontal permeability $k_{1,1}^{(clay)}$ of clay is much larger than the vertical permeability $k_{2,2}^{(clay)}$. Hence we have an anisotropy in the problem which is even increased by the big difference of the length scales in x_1 and x_2-direction.

The left hand boundary of the cross section corresponds to the observed groundwater divide where we assume zero horizontal flow, i.e. $\vec{q} \cdot \vec{n} = 0$. Also, the right hand boundary corresponds to the Thames valley, and coincides with a point of low groundwater head with zero horizontal flow. The lower boundary is formed by a layer of almost impervious Oxford Clay, and will also be specified as impermeable, i.e. $\vec{q} \cdot \vec{n} = 0$. Therefore these three parts of the boundary form Γ_N. The upper boundary of the section is taken to be the phreatic surface (water table), which does not coincide with the physical surface. It is denoted by Γ_D and we prescribe a (residual) pressure depending on the height x_2 of the fluid above sea level, i.e.

$$g_D(\vec{x}) = 9731.5 \cdot x_2 \ \left[\frac{\mathrm{N}}{\mathrm{m}^2}\right].$$

To discretise (5.1)–(5.4) in this case, we use a very simple non-uniform triangulation

[1] AEA Technology plc., Harwell, Oxfordshire OX11 0Qj, UK

$\mathcal{T}$ of Ω with 12200 elements (see Figure 5.1) which was created with NAMMU. The minimum and maximum diameter of the elements $T \in \mathcal{T}$ are $h_{min} = 0.105$ and $h_{max} = 205.7$ and the maximum aspect ratio of any of the elements $T \in \mathcal{T}$ is $\rho_{max} = 142.3$. A discretisation with lowest-order Raviart-Thomas elements yields a system of the form (5.9) with $n_{\mathcal{V}} + n_{\mathcal{W}} = 30475$ degrees of freedom.

We solve this system using the decoupled iterative method described in Chapter 4, and apply a preconditioned conjugate gradient (PCG) method for the core task of solving the decoupled symmetric positive definite velocity system (5.10). The convergence criterion in the PCG method is the relative reduction of the preconditioned residual by a factor of 10^{-9}, and to precondition (5.10) we employ the additive Schwarz preconditioner presented in Section 4.2.4 in the implementation provided by the DOUG package [52], both in its two-level form $\mathcal{P}^{-1}_{AS(2)}$ (defined in (4.43)) with an adaptively chosen coarse grid, and in its one-level form $\mathcal{P}^{-1}_{AS(1)}$ (defined in (4.85)) with no coarse grid solve. Once the velocity vector $\mathbf{q}$, or equivalently $\mathring{\mathbf{q}}$ (see Section 4.1.1), is known, we recover the pressure $\mathbf{p}_R$ in (5.9) by solving the lower triangular system (4.11), by simple back substitutions. The computed velocity field and pressure contours are plotted in Figure 5.2.

The dimension of the velocity system (5.10) in this case is $\mathring{n} = 6075$, and our simulations show that 41 iterations of PCG are necessary to solve it, if we employ the two-level preconditioner $\mathcal{P}^{-1}_{AS(2)}$ with the adaptive coarse mesh provided by DOUG. Even without a coarse mesh, using $\mathcal{P}^{-1}_{AS(1)}$ instead, our method converges in 68 iterations.

For both choices of preconditioner this is a more than reasonable performance of our method, taking into account the large jumps in the permeability field, the large aspect ratios in the mesh, and the strong anisotropy in the problem.

5.1.2 The Sellafield site

The second example is taken from a recent case study of a repository site in Sellafield, Cumbria, UK, carried out by United Kingdom Nirex Limited[2]. We would like to thank Nirex for kindly providing us with the data sets `cbase10a.dat`, `cbase10b.dat`, which were developed for the Nirex 97 groundwater flow calculations [62].

In this study, a two-dimensional vertical cross section through the Sellafield site that runs approximately perpendicular to the coast is chosen to model the geological features. It is about 22300 [m] long and about 4030 [m] deep at the right hand side, and follows a flow line taken from an old 3D model which passes through the proposed repository location. The geological configuration of the cross section is shown in Figure 5.3, and is illustrated further by shading the rock strata and fault lines using different colours (see Jackson & Watson [62, Section 4.1.1] for a detailed description).

Although the transport of salinity is a very important feature in modelling groundwater flow in coastal regions, we will neglect any variations in the density ρ of the

[2]United Kingdom Nirex Limited, Harwell, Oxon OX11 0RH, UK.

groundwater and assume that ρ is constant throughout the section. Under this assumption it makes sense to simulate steady-state groundwater flow through the cross section by solving the system (5.1)–(5.4).

The permeabilities in the principal directions of the different rock types and fault lines, which are present in the cross section, range from $2.8 \cdot 10^{-18}$ to $2.3 \cdot 10^{-12}$ [m^2], and the largest jump in any direction is $8.4 \cdot 10^5$. As in the previous example, for some of the rock types the permeability has anisotropic characteristics, i.e. the permeability in the first principal direction is larger than the permeability in the second. However, to obtain the permeability tensor k in the (x_1, x_2)-coordinate system, the principal directions are rotated to be parallel to the faults and strata, and so in contrast to the example considered in Section 5.1.1, the permeability tensor k is full here, i.e. $k_{1,2} \not\equiv 0$. This means that in this problem the anisotropy is not simply confined to differences in the x_1 and x_2-direction.

The boundary conditions are chosen in a similar way as in Section 5.1.1. On the right hand boundary, on the left hand boundary, and on the lower boundary the model assumes no flow, i.e. $\vec{q}\cdot\vec{n} = 0$, and so these three parts of the boundary form Γ_N again. The upper boundary, on the other hand, forms the part Γ_D. Offshore, it coincides with the bottom of the sea, and we prescribe a bathymetric pressure depending on the depth x_2 of the sea, i.e.

$$g_D(\vec{x}) = -258.8 \cdot x_2 \;\; \left[\tfrac{\mathrm{N}}{\mathrm{m}^2}\right].$$

Inland, it is taken to be the phreatic surface as in the previous example, and we prescribe a topographical pressure depending on the height x_2 of the fluid above sea level, i.e.

$$g_D(\vec{x}) = 9797.0 \cdot x_2 \;\; \left[\tfrac{\mathrm{N}}{\mathrm{m}^2}\right].$$

To discretise (5.1)–(5.4) in this case, we use a very complicated, highly unstructured, non-uniform triangulation $\mathcal{T}$ of Ω with 46598 elements (see Figure 5.3) which Nirex developed for their groundwater flow calculations within `NAMMU`. The minimum and maximum diameter of the elements $T \in \mathcal{T}$ are $h_{min} = 0.004$ and $h_{max} = 688.7$ and the maximum aspect ratio of any of the elements $T \in \mathcal{T}$ is $\rho_{max} = 3546.6$, which is extremely large and a true test for our solver. A discretisation with lowest-order Raviart-Thomas elements yields a system of the form (5.9) with $n_{\mathcal{V}} + n_{\mathcal{W}} = 116473$ degrees of freedom.

As in the previous example, we solve system (5.9) using the decoupled iterative method described in Chapter 4 and apply a preconditioned conjugate gradient (PCG) method for the core task of solving the decoupled symmetric positive definite system (5.10). The convergence criterion in the PCG method is the relative reduction of the preconditioned residual by a factor of 10^{-9}, and once again, to precondition (5.10) we employ the one and two-level additive Schwarz preconditioners $\mathcal{P}^{-1}_{AS(1)}$ and $\mathcal{P}^{-1}_{AS(2)}$ (defined in (4.43) and (4.85) respectively). The pressure $\mathbf{p}_R$ in (5.9) is recovered in a post-processing step by solving the lower triangular system (4.11) by simple back substitutions. The computed velocity field and pressure contours are plotted in Figure 5.4. See also Figure 5.5 for a magnification of the velocity field in the coastal region in the middle third of the cross section.

The dimension of the reduced velocity system (5.10) in this case, is $\mathring{n} = 23277$, and our simulations show that 370 iterations of PCG are necessary to solve this system if we employ the one-level overlapping additive Schwarz preconditioner $\mathcal{P}_{AS(1)}^{-1}$ (with minimal overlap and without any coarse grid solve). If we employ the two-level variant $\mathcal{P}_{AS(2)}^{-1}$ instead, which additionally uses an adaptive coarse mesh (automatically created in the `DOUG` implementation), the number of iterations is reduced to 101, a good result!

Under the extreme circumstances in this case study, in particular in view of the huge aspect ratios in the mesh and in view of the strong variability of the permeability field throughout the cross section, our method is performing exceptionally well. The significance of this good performance is complemented by its optimal parallel efficiency which we will now demonstrate in Section 5.1.3.

5.1.3 Parallel Efficiency

Table 5.1 illustrates the parallel efficiency of our method for the Sellafield problem described in Section 5.1.2. The iterations and times recorded are those for the solution of the decoupled velocity system (5.10) using the parallel iterative method with additive Schwarz preconditioner described in Section 4.2.4. They were obtained on the 20 node SGI Origin2000 at the Faculty of Science of the University of Bath, UK (Peak Performance: 390 MFlops/sec per processor).

	With coarse grid			Without coarse grid		
Slaves	Iterations	Time (sec)	Efficiency	Iterations	Time (sec)	Efficiency
1	101	23.1	100 %	370	60.6	100 %
2	131	13.3	87 %	449	36.2	89 %
3	79	6.1	126 %	324	16.4	123 %
4	129	6.8	85 %	421	15.5	98 %
5	97	4.3	107 %	486	14.4	84 %
6	112	4.1	94 %	414	11.0	92 %
7	106	3.50	94 %	361	8.10	107 %
8	100	2.96	98 %	420	8.15	93 %
9	81	2.37	108 %	272	4.65	145 %
10	119	3.16	73 %	408	6.60	92 %
11	116	2.88	73 %	439	6.40	86 %
12	115	2.81	69 %	335	5.24	96 %
13	95	2.34	76 %	274	4.27	109 %

Table 5.1: Parallel Efficiency for Sellafield problem on an Origin2000.

To understand these results, in particular the efficiency columns in Table 5.1, recall that the `DOUG` solver described in Section 4.2.4 is organised on a master/slave model.

In the construction of the preconditioner the coarse mesh is assembled and solved on the master processor while the slaves handle the solves on the subdomains. Similarly, in the execution of dot and matrix-vector products, the slaves do the local calculations while the master is responsible for collating global information (see also Hagger [51]). In Table 5.1 we give the parallel efficiency results as a function of the number of slave processors. The Efficiency column is computed in each case as $t(1)/(st(s))$, where $t(s)$ is the time required by the solver when s slaves are used. In effect, the bulk of the computation is done on the slaves and so the figures in Table 5.1 give an accurate impression of the parallel efficiency of the algorithm.

An important thing to note are the variations in the number of iterations for different numbers of slave processors in Table 5.1 (Columns 2 and 5). This is due to the fact that the number S of subdomains which are used in the additive Schwarz preconditioner in `DOUG`, varies with the number of slave processors ("load-balancing"). Since the partitioning of the domain Ω in `DOUG` is carried out automatically using only the mesh topology, this might lead to very different partitionings of Ω in each case, in particular on a highly unstructured mesh with large aspect ratios (as in this case).

Nevertheless, the results in Columns 3, 4, 6, and 7 of Table 5.1 still reflect the success of the parallelisation.

5.2 Heterogeneous media

In contrast to the classical deterministic models considered in Section 5.1 for layered media, k will, in the heterogeneous case considered here, be modelled using a Gaussian random field. The numerical treatment of the resulting system of stochastic PDEs then involves the solution of (5.1)–(5.4) for many different *realisations* of k and subsequent computation of statistical properties of the resulting velocity and/or pressure fields.

The Gaussian random fields which determine k are characterised by a pair of parameters (σ^2, λ), where σ^2 is the *variance* and λ is the *length scale* over which the field is correlated. It is known that any realisation of the Gaussian random field k is Hölder continuous but not in general differentiable and so the resulting velocity and pressure fields have only low regularity throughout the domain. Since this irregularity is global it cannot be compensated by local mesh refinement, and the only known way to achieve acceptable accuracy for these problems is to use a mesh which is (uniformly) as fine as possible throughout the domain. In typical 2D simulations the required number of degrees of freedom n for acceptable accuracy typically lies in the range 10^6 to 10^8. A key aim of this section is to provide usable methods for problems of this sort. The use of parallel computing power plays an essential rôle in achieving this aim.

Because the variable of prime interest in this computation is the (Darcy) velocity $\vec{q}$, the discretisation schemes of most interest are those which preserve conservation of mass in an appropriate way, with the prime candidates being mixed finite element or finite volume techniques. Because of the lack of regularity in this problem, high order elements are inefficient and so we will discretise (5.1)–(5.4) using lowest order Raviart-Thomas elements, as presented in Chapter 2.

In this section we apply the fast parallel iterative method described in Chapter 4 to the resulting linear equation system. The results in Section 5.2.4 show that, using our solver with a fixed number of processors, the time taken for a solve scales almost linearly (i.e. optimally) in n and is remarkably robust to variations in σ^2 and λ. Moreover, almost 100% parallel efficiency is observed when the algorithm is tested on a machine with up to 10 processors, with a modest decrease in efficiency for higher numbers of processors.

The layout of Section 5.2 is as follows. In Section 5.2.1 we describe the stochastic model that is going to be used for the permeability k and some of its statistical properties. In Section 5.2.2 we describe the model problem and in Section 5.2.3 we select a stopping criterion which ensures reasonably uniform accuracy across the parameter range. Finally, in Section 5.2.4 we give a sequence of experiments which show the performance of the method.

We would like to note again at this point that the results in this section were achieved in collaboration with Andrew Cliffe, Ivan Graham, and Linda Stals (see Cliffe et al. [28, 29]). In particular, all the computations in this section have been carried out by Linda Stals.

5.2.1 Stochastic modelling of heterogeneous media

A widely used method of treating heterogeneity in porous media is stochastic modelling. The basic idea is to model the permeability field k in (5.1) as a stochastic spatial process, assuming that a single realisation of this stochastic process is a reasonable representation of the permeability field and that any of the realisations are equally probable given the information available from geological measurements. This approach leads to a system (5.1)–(5.4) of stochastic PDEs, where $\vec{q}$ and p_R are now random variables. Given certain statistical properties of k (which we now describe), it is of interest to study statistical properties of those random variables. More detail on the following statistical background can be found in Adler [1] or Cressie [31].

We recall the notion of a *Gaussian random variable* Z which is specified by its mean (or expectation) $m = E\{Z\}$ and variance σ^2. More generally, a vector $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$ of n Gaussian random variables is completely specified by the vector $\mathbf{m} \in \mathbb{R}^n$ (containing the mean values of each of these variables) together with its $n \times n$ *covariance matrix* $\Sigma = E\{(\mathbf{Z} - \mathbf{m})(\mathbf{Z} - \mathbf{m})^T\}$.

Generalising this to the infinite dimensional setting, a *random field* on an open domain $\Omega \subset \mathbb{R}^2$ (also called a *spatial process*) is a set of random variables $Z(\vec{x})$, each of which is associated with a point $\vec{x} \in \Omega$. This random field is called *Gaussian* if for each arbitrary n, each set of n random variables located at n arbitrarily chosen spatial points is Gaussian. Such fields can be completely specified by their (spatially varying) mean and covariance functions, denoted respectively by $m(\vec{x})$ and

$$\Sigma(\vec{x},\vec{y}) := E\{(Z(\vec{x}) - m(\vec{x}))\,(Z(\vec{y}) - m(\vec{y}))\}, \quad \vec{x},\vec{y} \in \Omega. \tag{5.11}$$

In this paper we will be concerned only with *statistically homogeneous isotropic* Gaussian random fields whose mean and covariance have the particular forms:

$$m(\vec{x}) := m, \quad \Sigma(\vec{x},\vec{y}) := \sigma^2 \exp(-|\vec{x}-\vec{y}|/\lambda), \tag{5.12}$$

for positive constants m, σ and λ. Note that evaluating (5.11) at $\vec{y} = \vec{x}$ and combining with (5.12) shows that for each $\vec{x} \in \Omega$, the random variable $Z(\vec{x})$ is normally distributed with mean m and variance σ^2 (independent of $\vec{x}$). Moreover, expanding (5.11), rearranging and using (5.12) shows

$$E\{Z(\vec{x})Z(\vec{y})\} = \Sigma(\vec{x},\vec{y}) + m^2 = \sigma^2 \exp(-|\vec{x}-\vec{y}|/\lambda) + m^2 \ , \tag{5.13}$$

from which we can deduce

$$\begin{aligned} E\{(Z(\vec{x}) - Z(\vec{y}))^2\} &= E\{Z(\vec{x})^2\} + E\{Z(\vec{y})^2\} - 2E\{Z(\vec{x})Z(\vec{y})\} \\ &= 2\sigma^2\,[1 - \exp(-|\vec{x}-\vec{y}|/\lambda)] \\ &= \gamma(|\vec{x}-\vec{y}|) \ , \end{aligned} \tag{5.14}$$

where

$$\gamma(r) := 2\sigma^2(1 - \exp(-r/\lambda)) \ . \tag{5.15}$$

The function γ is often called a *variogram* - see, e.g. Cressie [31].

Now let us assume that the permeability *tensor* $k(\vec{x})$ is a scalar multiple of the identity, i.e.

$$k(\vec{x}) := k_{iso}(\vec{x})\, I, \qquad \text{for all} \quad \vec{x} \in \Omega, \tag{5.16}$$

and by an abuse of notation let us denote the scalar function k_{iso} by k again. We shall solve (5.1)–(5.4) in the case when $\log(k)$ is a realisation of a statistically homogeneous isotropic Gaussian random field. There is some evidence from field data that this gives a reasonable representation of reality in certain cases (see Gelhar [42], Hoeksema & Kitanidis [61]). There are many good methods for generating realisations of Gaussian random fields, including those based on FFT [49, 79], direct simulation [34], and the Turning Bands method [68, 69, 89]. Here we use the Turning Bands approach that represents the field as a superposition of one-dimensional fields, which are generated along lines radiating from the origin using a spectral technique.

Of particular importance to the accuracy of any discretisation is the question of regularity of this realisation. This question is thoroughly investigated in the statistical literature and the following theorem can be deduced from Adler [1].

Theorem 5.1. *Let $0 < \alpha < 1/2$ and let X denote any realisation of the Gaussian random field Z introduced above. Then with probability 1,*

$$|X(\vec{x}) - X(\vec{y})| \leq C|\vec{x} - \vec{y}|^{\alpha} \ , \quad \vec{x}, \vec{y} \in \Omega$$

for some positive constant C.

Proof. Without loss of generality we can assume that the polygonal domain Ω is a subset of $[0,1] \times [0,1]$. (If this is not the case, choose $a \in \mathbb{R}$ and $\vec{b} \in \mathbb{R}^2$ such that the affine map $\tilde{\vec{x}} \mapsto a\tilde{\vec{x}} + \vec{b}$ provides a bijection between a polygon $\tilde{\Omega} \subset [0,1] \times [0,1]$ and Ω. Then $\tilde{Z}(\tilde{\vec{x}}) := Z(a\tilde{\vec{x}} + \vec{b})$ defines a statistically homogeneous Gaussian random field on $\tilde{\Omega}$, with variogram $\tilde{\gamma}(r) = 2\sigma^2(1 - \exp(-ar/\lambda))$ and Hölder continuity of realisations of $\tilde{Z}$ will imply the analogous result for Z.)

So, assuming that $\Omega \subset [0,1] \times [0,1]$, we first note that the random field Z is just the restriction to Ω of the Gaussian random field on $[0,1] \times [0,1]$ with the same variogram. The regularity of any realisation of Z can then be deduced from the asymptotics of the variogram $\gamma(r)$ as $r \to 0$. Because $\sqrt{\gamma(r)} = O(r^{1/2})$ it follows that $Z - m$ is an index-$\frac{1}{2}$ (2,1) Gaussian field in the sense of Adler [1, Definition 8.3.1] and the result follows from Theorem 8.3.2 of the same reference. $\square$

Once system (5.1)–(5.4) has been solved for multiple realisations of k and the statistical properties of the velocity field have been found, the dispersion present in the system can (at least when the molecular diffusion is small) be studied by looking at the statistics of particle paths moving in the velocity field. In fact if X'_j denotes the jth coordinate of the particle displacement from its mean position then the spreading can be characterised by the second order moment of the particle paths:

$$X_{jl} := E\{X'_j X'_l\} \ , \quad j, l = 1, 2 \ . \tag{5.17}$$

X'_j satisfies the differential equation

$$\frac{dX'_j}{dt} = q_j \quad , \quad j = 1, 2 \tag{5.18}$$

where q_j is the jth component of the velocity field. This model highlights another advantage of the low order mixed finite element method: Since the computed velocity field is constant on each element, the differential equation for X'_j is trivially integrated in an element by element fashion, thus allowing the efficient computation of the many particle paths which would be required in statistical analyses.

Before continuing we remark that there are many stochastic models for groundwater flow (see, e.g. Kolterman & Gorelick [66] for a review). We have chosen the given model here because it is a relatively simple model which applies to fully saturated flows but still has many of the features of some of the more complicated models. We remark also that more complicated models require more data to support them, and data is very often difficult to come by, especially in the case of deep geological waste disposal, our chief motivation for studying this problem.

5.2.2 Numerical solution of a model problem

Let us consider the following model problem: Once more the domain Ω is taken to be the unit square $(0,1)^2$, the viscosity μ is set to 1, and the permeability k is chosen so that $\log(k)$ is a realisation of a statistically homogeneous isotropic Gaussian random field on Ω with zero mean, variance σ^2 and length scale λ, as described in Section 5.2.1.

To achieve acceptable accuracy with our numerical solution in this case, we need a mesh that is (uniformly) as fine as possible throughout the domain, justifying the use of a uniform triangulation $\mathcal{T}$ of Ω. We obtain $\mathcal{T}$ by firstly dividing Ω into N^2 equal squares $(\frac{i-1}{N}, \frac{i}{N}) \times (\frac{j-1}{N}, \frac{j}{N})$, and then further subdividing each square into two triangles. This is done by colouring the squares in a red/black checkerboard pattern, and then using a diagonal drawn from bottom left to top right for red squares and from top right to bottom left for black squares. As in the deterministic case, a discretisation of (5.1)–(5.4) with lowest-order Raviart-Thomas elements on $\mathcal{T}$ yields a sparse, indefinite, and highly ill-conditioned saddle point system of the form (5.9).

As we mentioned in Section 5.2.1, to generate realisations of k is expensive and it is important to sample k at as few points as possible. Therefore, in the practical implementation of (5.9) we shall replace $m(\cdot,\cdot)$ in (5.5) by

$$\widetilde{m}(\vec{q},\vec{v}) := \int_\Omega \mu \widetilde{k}^{\,-1}\, \vec{q}\cdot\vec{v}\, d\vec{x}, \tag{5.19}$$

where $\widetilde{k}$ denotes the piecewise constant interpolation of k at the centroids of the triangles in the mesh $\mathcal{T}$. It is shown in the Appendix to Cliffe et al. [28], by the use of the First Strang Lemma (see, e.g. Ciarlet [27]), that this approach maintains the accuracy of the discretisation.

The direct simulation approach which we have taken here, only makes statistical sense when the length scale λ is of the order of the mesh diameter, equivalently

$$\lambda = C_\ell / N \tag{5.20}$$

for some constant $C_\ell \geq 1$, as $N \to \infty$. However, since N must already be large enough to ensure acceptable accuracy (i.e. $N \sim 10^3$ or 10^4), fairly fine length scales are treatable by this choice, and it is widely used in hydrogeological modelling. Smaller length scales could be treated by an appropriate upscaling of k in each element, but this is expensive and it is not clear how to do it accurately. From now on, k is replaced by its piecewise constant interpolant $\widetilde{k}$, which is computed using the Turning Bands Algorithm [68, 69, 89].

We assume that there is zero flow across the bottom and top of Ω, i.e.

$$\Gamma_N := [0,1] \times \{0,1\} \tag{5.21}$$

in (5.4), and that the residual pressure p_R is required to have value 1 at the left hand boundary and 0 at the right hand boundary (corresponding to a prescribed pressure gradient across the domain). Thus in (5.3) we have

$$\Gamma_D = \{0,1\} \times (0,1) \tag{5.22}$$

and

$$g_D(\vec{x}) \equiv 1 \;\text{ for } \vec{x} \in \{0\} \times [0,1], \quad g_D(\vec{x}) \equiv 0 \;\text{ for } \vec{x} \in \{1\} \times (0,1)\ . \tag{5.23}$$

We shall give results here only for the computation of the velocity $\mathbf{q}$ in (5.9) by solving the decoupled system (5.10) for $\mathring{\mathbf{q}}$. In the case of the particular boundary conditions (5.21)–(5.23), the computation reduces to the solution of the linear system

$$\begin{bmatrix} A & \mathbf{c} \\ \mathbf{c}^T & r \end{bmatrix} \begin{bmatrix} \mathring{\mathbf{q}}^{(A)} \\ \mathring{q}_{\mathring{n}} \end{bmatrix} = \begin{bmatrix} \mathring{\mathbf{g}}^{(A)} \\ \mathring{g}_{\mathring{n}} \end{bmatrix}, \tag{5.24}$$

where A is a square sparse matrix, and (since Γ_N here contains only two components) $\mathbf{c}$ is a single column vector and r is a scalar. All of these are obtained by elementary row and column operations on a standard piecewise linear finite element matrix (see Section 4.2.1).

The block elimination procedure in Section 4.2.2 in this case requires solutions of two systems of the form

$$A\mathbf{u} = \mathbf{b}. \tag{5.25}$$

In the special case here, where the Dirichlet data g_D is constant on each component of Γ_D, it turns out that $\mathring{\mathbf{g}}^{(A)} = \mathbf{0}$ and we only need to solve (5.25) once. The timings in Section 5.2.4 are for this task.

The sparse, symmetric positive definite, and highly ill-conditioned problem (5.25) is solved by the parallel iterative method described in Section 4.2.4. There are 3 parameters which determine the difficulty of (5.25): the mesh parameter N, the variance σ^2, and the length scale λ. We are interested in the efficiency of this parallel method as well as its robustness with respect to these parameters. For our tests we allow σ^2 to vary independently, and λ to vary as in (5.20), for some constant C_ℓ to be specified below. From a numerical point of view these are particularly difficult problems, since the realisation of k varies from element to element and may take wildly differing values across the domain. As C_ℓ decreases the probability of large jumps in k between neighbouring elements increases. On the other hand, to illustrate the effect of increasing σ^2, in Figure 5.6 we give a grey scale plot of the values of $\log(k)$ for a single realisation in the case $N = 256$ and $\lambda = 10/N$ for two different values of σ^2. Observe that the pattern is independent of σ^2, but that the scale changes as σ^2 increases. In fact the numerical range of $\log(k)$ grows linearly in $\sqrt{\sigma^2}$, and so the condition number of the matrix A will grow like $\exp(2\sqrt{\sigma^2})$ as σ^2 increases. To emphasise the effect that this will have on the conditioning of (5.25) observe, for example, that $\frac{\max|k(\vec{x})|}{\min|k(\vec{x})|} \sim 10^9$ when $\sigma^2 = 8$.

5.2.3 Selection of a stopping criterion

Since we have in mind here the solution of a range of problems of varying difficulty by an iterative method, it is important to design a stopping criterion which ensures reasonably uniform accuracy across all problems. This ensures that subsequent comparison of solution times and iteration counts will be meaningful. In this subsection we describe a heuristically based approach to designing such a stopping criterion.

The preconditioned conjugate gradient (PCG) method for (5.25) with a symmetric positive definite preconditioner $\mathcal{P}^{-1}$ produces a sequence of iterates $\mathbf{u}^i$ and residuals $\mathbf{r}^i$ which satisfy $\mathbf{r}^i = \mathbf{b} - A\mathbf{u}^i = A\mathbf{e}^i$ where $\mathbf{e}^i = \mathbf{u} - \mathbf{u}^i$ is the error at the ith iterate. This algorithm also computes the *preconditioned residual* $\mathbf{z}^i = \mathcal{P}^{-1}\mathbf{r}^i = (\mathcal{P}^{-1}A)\mathbf{e}^i$.

Typical stopping criteria for the PCG method involve requiring that $\mathbf{z}^i$ is small in some norm. More precisely we have the standard estimate for the relative error reduction

$$\frac{\|\mathbf{e}^i\|_2}{\|\mathbf{e}^0\|_2} \leq \kappa \frac{\|\mathbf{z}^i\|_2}{\|\mathbf{z}^0\|_2}, \tag{5.26}$$

where $\kappa := \kappa(\mathcal{P}^{-1}A)$ denotes the spectral condition number of $\mathcal{P}^{-1}A$ as defined in (2.83). From this it follows that the stopping criterion:

$$\frac{\|\mathbf{z}^i\|_2}{\|\mathbf{z}^0\|_2} \leq \varepsilon/\kappa \tag{5.27}$$

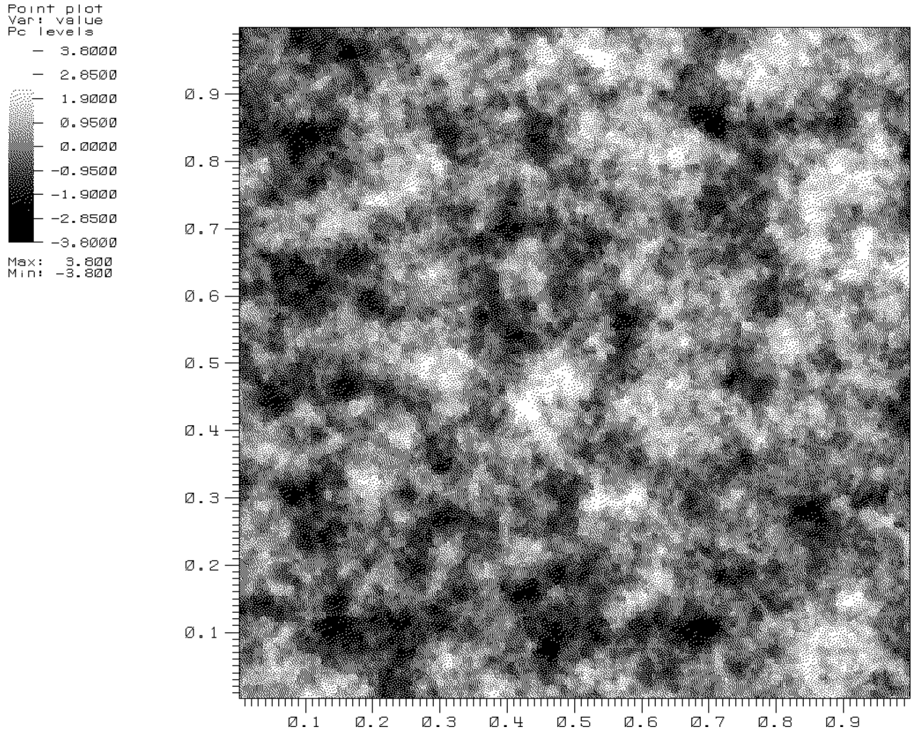
is sufficient to ensure the required relative error reduction $\|\mathbf{e}^i\|_2 \ \|\mathbf{e}^0\|_2 \leq \varepsilon$.

The difficulty with implementing (5.27) is the problem of computing κ. In Kaasschieter [64] (for the unpreconditioned case $\mathcal{P} = I$) it is proposed to estimate κ dynamically using the Lanczos procedure (for which some of the data is already computed during the CG iteration). However, even if such a procedure is adopted, the resulting stopping criterion (5.27) is often over-pessimistic due to the fact that the smallest constant κ such that (5.26) holds for all i is often very much smaller than the true condition number of $\mathcal{P}^{-1}A$.

Here we are interested in a class of problems which depend on parameters σ^2, N, and λ. For a restricted range of problems (which are small enough so that the exact solution can be computed by a direct solver), we compute the *effective condition number*:

$$\tilde{\kappa} := \tilde{\kappa}(\sigma^2, N, \lambda) := \frac{\|\mathbf{e}^i\|_2}{\|\mathbf{e}^0\|_2} \frac{\|\mathbf{z}^0\|_2}{\|\mathbf{z}^i\|_2} , \tag{5.28}$$

for some specified i as the parameters σ^2, N and λ change. Our practical stopping



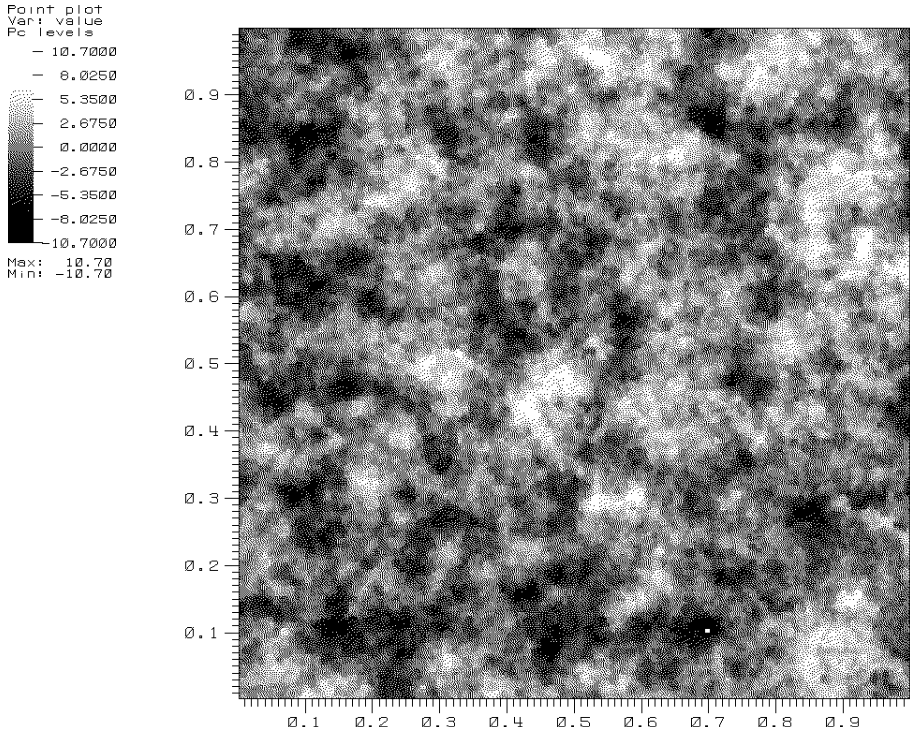


Figure 5.6: Grey scale plot of $\log(k)$ for $\sigma^2 = 1$ (top) and $\sigma^2 = 8$ (bottom).

criterion is then to choose the first i such that

$$\frac{\|\mathbf{z}^i\|_2}{\|\mathbf{z}^0\|_2} \le \varepsilon/\tilde{\kappa}. \tag{5.29}$$

The result of this exercise is that $\tilde{\kappa}$ is found to vary only very mildly with these parameters (see (5.30) below).

To obtain $\tilde{\kappa}(\sigma^2, N, \lambda)$ experimentally, we solved the test problems using the conjugate gradient method with additive Schwarz preconditioner $\mathcal{P}^{-1}_{AS(2)}$, as described in Section 4.2.4, with initial guess $\mathbf{u}^0 = \mathbf{0}$, and we iterated until the relative error $\|\mathbf{e}^i\|_2/\|\mathbf{e}^0\|_2$ was less than $\varepsilon = 10^{-4}$ (with the exact solution $\mathbf{u}$ found using a direct solver). From this solution we computed $\tilde{\kappa}$ above.

First we studied the variation with respect to σ^2, and here we fixed $N = 32$ and $\lambda = 10/N$. In Figure 5.7 (left) we plot computed values of $\tilde{\kappa}$ against σ^2 (solid line). The best least squares straight line fit to these points (dotted line) yields an empirical

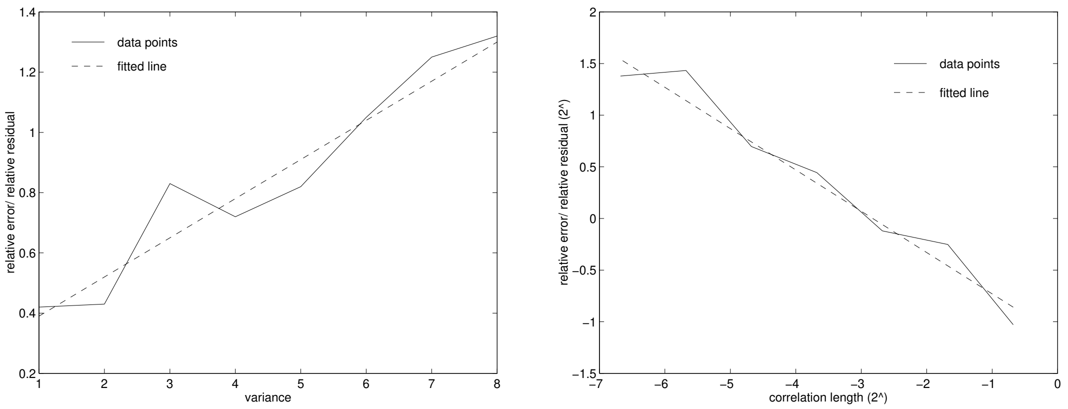


Figure 5.7: Variation of $\tilde{\kappa}$ with σ^2 (left), and with λ for $\sigma^2 = 4$ (right) ($N = 32$).

approximation for the variation of $\tilde{\kappa}$ with σ^2 as: $0.26 + 0.13\sigma^2$. To test the validity of this, we recomputed the above experiments using the stopping criterion (5.29) with $\tilde{\kappa} = 0.26 + 0.13\sigma^2$ and $\varepsilon = 10^{-4}$. The relative error $\|\mathbf{e}^i\|_2/\|\mathbf{e}^0\|_2$ remained in the interval $[2 \times 10^{-5}, 1.4 \times 10^{-4}]$ as σ^2 ranged between 1 and 8, indicating that this is a reasonable approximation of how $\tilde{\kappa}$ varies with σ^2.

To study variation with respect to λ, we set $N = 32$ and $\sigma^2 = 4$ and computed $\tilde{\kappa}$ for $\lambda = 10/16, 10/32, \cdots, 10/1024$. A $\log_2 - \log_2$ plot of these results is given in Figure 5.7 (right) (solid line). The dotted line shows the best computed straight line fit and suggests that $\tilde{\kappa}$ decreases with $O(\lambda^{-0.4})$ as λ increases. From this observation we propose the empirical model $\tilde{\kappa} = (0.26 + 0.13\sigma^2)(0.46\lambda^{-0.4})$. To demonstrate the validity of this we recomputed these experiments using this value of $\tilde{\kappa}$ in stopping criterion (5.29) where $\sigma^2 = 4$, $N = 32$ and $\varepsilon = 10^{-4}$. We found that the resulting relative error lay in the range $[6 \times 10^{-5}, 1.4 \times 10^{-4}]$ indicating a stopping criterion which is robust to variations in λ.

Finally, to model variations with respect to N we computed $\tilde{\kappa}$ in (5.28) for $N = 16, 32, 64, 128$ in the case $\sigma^2 = 4$ and $\lambda = 10/16$. These experiments suggested that there is no noticeable increase in the value of $\tilde{\kappa}$ as N increases. Thus we postulate that

$$\tilde{\kappa}(\sigma^2, N, \lambda) \approx (0.26 + 0.13\sigma^2)(0.46\lambda^{-0.4}) \tag{5.30}$$

as σ^2, λ and N vary. In the experiments in the next subsection we use this formula for $\tilde{\kappa}$ in the stopping criterion (5.29).

5.2.4 Performance of the iterative method

Our first set of results – Table 5.2 – illustrates the performance of the PCG method for (5.25) with additive Schwarz preconditioner $\mathcal{P}^{-1}_{AS(2)}$ and $\mathcal{P}^{-1}_{AS(1)}$, with and without coarse grid solve respectively, for various values of N and σ^2. (See (4.43) and (4.85) for the definition of the preconditioner $\mathcal{P}^{-1}_{AS(2)}$ and $\mathcal{P}^{-1}_{AS(1)}$.) The length scale λ varies as in (5.20) with $C_\ell = 10$, and n denotes the number of unknowns in system (5.25). The stopping criterion was (5.29) with $\varepsilon = 10^{-9}$ and $\tilde{\kappa}$ given by (5.30). The value

			With coarse grid		Without coarse grid	
N	n	σ^2	Iterations	$\|\|\mathbf{z}^i\|\|_2/\|\|\mathbf{z}^0\|\|_2$	Iterations	$\|\|\mathbf{z}^i\|\|_2/\|\|\mathbf{z}^0\|\|_2$
128	16383	1	21	$7.7 \cdot 10^{-10}$	123	$1.8 \cdot 10^{-9}$
		2	22	$1.2 \cdot 10^{-9}$	137	$1.2 \cdot 10^{-9}$
		4	26	$5.7 \cdot 10^{-10}$	174	$8.1 \cdot 10^{-10}$
		6	29	$7.0 \cdot 10^{-10}$	198	$6.8 \cdot 10^{-10}$
		8	33	$4.0 \cdot 10^{-10}$	223	$4.2 \cdot 10^{-10}$
256	65535	1	23	$9.5 \cdot 10^{-10}$	270	$1.3 \cdot 10^{-9}$
		2	26	$7.8 \cdot 10^{-10}$	320	$1.1 \cdot 10^{-9}$
		4	31	$5.5 \cdot 10^{-10}$	454	$6.7 \cdot 10^{-10}$
		6	34	$5.6 \cdot 10^{-10}$	602	$5.6 \cdot 10^{-10}$
		8	41	$3.8 \cdot 10^{-10}$	740	$3.8 \cdot 10^{-10}$
512	262143	1	27	$4.5 \cdot 10^{-10}$	593	$1.1 \cdot 10^{-9}$
		2	29	$7.9 \cdot 10^{-10}$	742	$8.3 \cdot 10^{-10}$
		4	38	$5.0 \cdot 10^{-10}$	1155	$5.5 \cdot 10^{-10}$
		6	46	$4.3 \cdot 10^{-10}$	1677	$4.1 \cdot 10^{-10}$
		8	57	$2.6 \cdot 10^{-10}$	> 2000	–
1024	1048575	1	33	$3.4 \cdot 10^{-10}$	1059	$8.7 \cdot 10^{-9}$
		2	35	$6.4 \cdot 10^{-10}$	1598	$6.5 \cdot 10^{-9}$
		4	45	$4.2 \cdot 10^{-10}$	> 2000	–
		6	57	$3.1 \cdot 10^{-10}$	> 2000	–
		8	70	$1.7 \cdot 10^{-10}$	> 2000	–

Table 5.2: Study of the iteration count ($\lambda = 10/N$).

of $\|\mathbf{z}^i\|_2/\|\mathbf{z}^0\|_2$ given is the value of this quantity when the iteration stops (where $\mathbf{z}^i$ denotes the preconditioned residual, as described above).

The first thing to note is the observed success of the strategy for computing the coarse grid as outlined in Section 4.2.4. Since the coarse grid is constructed just from the geometry of the fine grid, ignoring the fact that the coefficient k is varying from element to element, one may be concerned that the coarse grid may not model the underlying fine scales of the problem (at the fine grid level) well enough to be effective. While there is clearly some dependence on the fine scale of the coefficient (the iteration numbers increase slightly as σ^2 increases) this dependence is mild (see below) and the addition of the coarse grid solve is clearly having a big effect on the robustness of the preconditioner. In the case $N = 512$ and $\sigma^2 = 8$, the addition of the coarse grid solve improved the computation time by a factor of about 30.

In the next two tables we investigate the robustness of the iterative method with respect to the various parameters in the problem in more detail. First, in Table 5.3 we investigate the behaviour of the method as N grows. We know from the discussion

N	n	With coarse grid	Without coarse grid
256	65535	26	320
512	262143	29	742
1024	1048575	35	1598

Table 5.3: Number of iterations as N (and therefore n) increases ($\lambda = 10/N$, $\sigma^2 = 2$).

in Section 4.2.4 that, for a *fixed smooth coefficient function*, as N (and therefore n) increases, we expect the number of PCG iterations to grow at worst with $O(n^{1/2}) = O(N)$, when the coarse solve is not included in the preconditioner, and with $O(n^{1/6}) = O(N^{1/3})$ when the coarse solve is included. The results in Table 5.3 indicate a growth no worse than this, even though in this case the coefficient is extremely rough.

In Table 5.4 we illustrate how the iteration numbers are affected by growth in σ^2 for $N = 256$ and $\lambda = 10/N$. The rate of growth of the number of PCG iterations

$\sqrt{\sigma^2}$	With coarse grid	Without coarse grid
1	23	270
1.4	26	320
2	31	454
2.4	34	602
2.8	41	740

Table 5.4: Number of iterations as σ^2 increases ($N = 256$, $\lambda = 10/N$).

is approximately linear in $\sqrt{\sigma^2}$. This behaviour is observed both with and without a coarse grid solve, although with a considerably larger asymptotic constant in the latter case. This should be compared with the fact that the *condition number* of the stiffness matrix A in (5.25) grows like $\exp(2\sqrt{\sigma^2})$. This observed behaviour (where the growth of the number of iterations is logarithmic in the condition number) is exactly as proved in Graham & Hagger [45, 46] (see Sections 4.2.4) for the special case when the number of regions in which the coefficient has a constant value is small compared to the number of elements on the fine mesh (compare also the results in Section 4.5.1). Here we have computed the harder problem where the coefficient has a different value on each element, but we still observe the same good behaviour as predicted in Graham & Hagger [45, 46]. It remains an open question to give a proof of this observation.

Recall that for a physically realistic model we assume (see (5.20)) that the length scale λ decreases linearly in $1/N$. In the previous Tables 5.2–5.4 we took $C_\ell = 10$ in (5.20). In Table 5.5 we illustrate the cases $C_\ell = 5, 20$. As expected the smaller value of C_ℓ leads to neighbouring values of k being less well-correlated and thus a larger number of PCG iterations are needed to solve this "rougher" problem.

	$C_\ell = 5$		$C_\ell = 20$	
σ^2	Iterations	$\|\|\mathbf{z}^i\|\|_2/\|\|\mathbf{z}^0\|\|_2$	Iterations	$\|\|\mathbf{z}^i\|\|_2/\|\|\mathbf{z}^0\|\|_2$
1	29	$6.2 \cdot 10^{-10}$	27	$5.9 \cdot 10^{-10}$
2	34	$5.1 \cdot 10^{-10}$	27	$5.7 \cdot 10^{-10}$
4	46	$3.9 \cdot 10^{-10}$	30	$6.5 \cdot 10^{-10}$
6	62	$3.1 \cdot 10^{-10}$	35	$4.2 \cdot 10^{-10}$
8	82	$1.9 \cdot 10^{-10}$	41	$3.4 \cdot 10^{-10}$

Table 5.5: Affect of C_ℓ on the iteration count ($N = 512$, with coarse grid).

In groundwater flow calculations in practice it is often necessary to study flows in long thin regions. In Table 5.6 we repeat some of the above calculations for the case when the domain Ω is $[0, L] \times [0, 1]$ and we study the effect of varying the aspect ratio $L \geq 1$ of the domain. In the absence of any additional information concerning

L	No. It.	$\|\|\mathbf{z}^i\|\|_2/\|\|\mathbf{z}^0\|\|_2$
1	31	$5.5 \cdot 10^{-10}$
4	42	$6.0 \cdot 10^{-10}$
16	50	$5.1 \cdot 10^{-10}$
64	65	$7.6 \cdot 10^{-10}$

Table 5.6: Effect of aspect ratio L of Ω on iteration count (with coarse grid).

anisotropy, in general for such problems we would need to take the same mesh diameter in both coordinate directions to ensure adequate accuracy. Thus, for each value of L, we here construct a uniform tensor product mesh with N_L subdivisions along the side $[0,1]$ and $L \times N_L$ subdivisions along the side $[0,L]$. However, in order to compare problems of the same dimension, N_L is chosen to ensure that the total number of degrees of freedom in the system is fixed at $n = (N+1)*(N-1)$, with $N = 256$. For these experiments $\sigma^2 = 4$ and $\lambda = 10/256$. The iteration count, as L increases, is given in Table 5.6. A very modest growth with L is observed.

Our final table – Table 5.7 – illustrates the parallel efficiency of the algorithm. The times recorded are those obtained on the 16 node IBM SP2 at the Daresbury Laboratory, UK (Peak Performance: 480 MFlops/sec per processor). The efficiency column is computed as discussed in Section 5.1.3.

	Without coarse grid		With coarse grid	
Slaves	Time (sec)	Efficiency	Time (sec)	Efficiency
1	156.7	100 %	12.65	100 %
2	79.9	98 %	6.25	101%
4	43.5	90 %	3.15	100%
6	30.2	86 %	2.04	103%
8	24.4	80 %	1.62	98%
10	20.6	76 %	1.34	94%
12	19.6	67 %	1.19	89%
14	17.2	65 %	1.10	82%

Table 5.7: Parallel Efficiency on an SP2 ($\sigma^2 = 2$, $N = 256$, $\lambda = 1/N$).

In Table 5.7 note especially the improved parallel efficiency of the method with the coarse grid compared to that without. This indicates the success of the the parallelisation strategy implemented in `DOUG`: the coarse solve is not only necessary to obtain good theoretical results, it also gives much improved timings and efficiency even though in principle much more communication is needed. The key is the overlapping of communication with computation implemented in `DOUG` [51, 52].

Efficiencies of greater than 100% for small numbers of processors are not unusual, due to cache effects as well as small differences in the actual quality of the solution produced at the end of the PCG iteration (see Hagger [51]). Nevertheless, the numbers of iterations, which are necessary to reach the required accuracy for different numbers of slave processors, vary only moderately here, i.e. between 26 and 28 for the two-level method with coarse grid and between 346 and 364 for the one-level method (compare the results in Section 5.1.3).

Finally, in Figure 5.8 we plot the computed velocity fields corresponding to (5.1)–(5.4) for the model problem defined in Section 5.2.2 with boundary conditions (5.23) in the case $N = 256$, $\lambda = 10/N$ and $\sigma^2 = 1, 4, 8$ respectively. Note the increased dispersion in the flow paths as σ^2 increases.

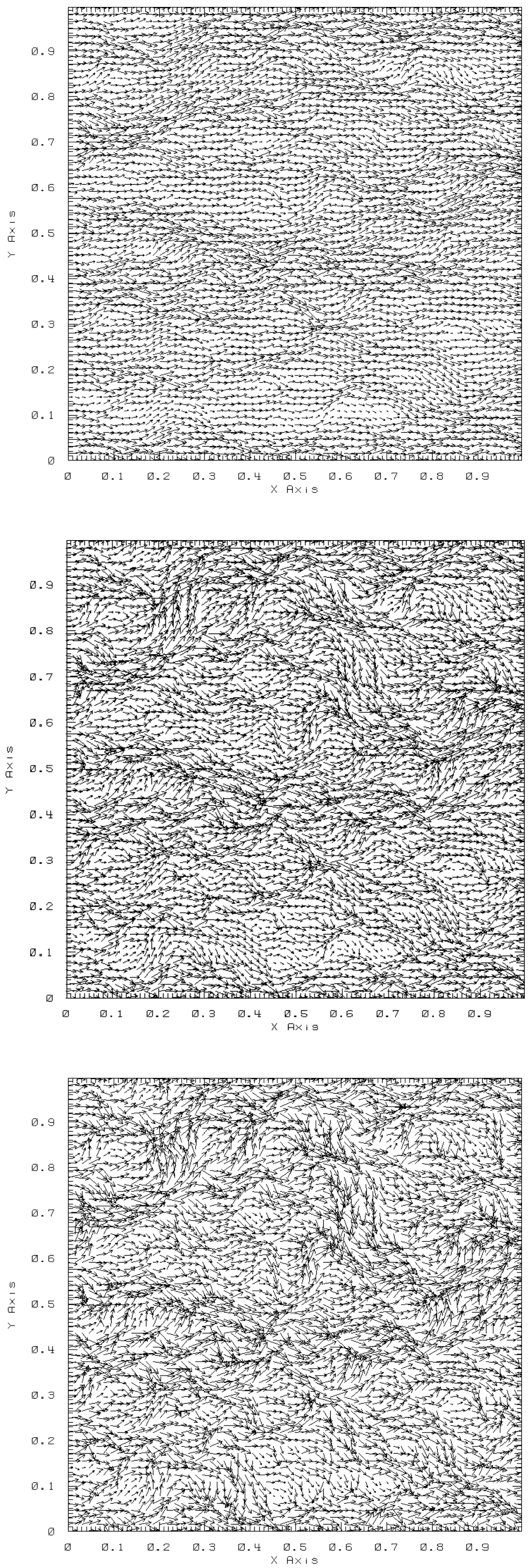


Figure 5.8: Vector plot of the velocity for $\sigma^2 = 1$, 4, and 8 ($N = 256$, $\lambda = 10/N$).

5.3 Summary

In this chapter we applied the decoupled iterative method which we constructed in Chapter 4 to realistic problems arising in two-dimensional groundwater flow computations. A discretisation by Raviart-Thomas elements of the equations (5.1)–(5.4) that model single phase flow in saturated porous media, led to saddle point systems of the form (5.9) which are tractable by our decoupled method. Moreover, since we only considered two-dimensional examples here, we were able to apply the fast parallel domain decomposition method described in Section 4.2.4 of the previous chapter for the core task of solving the symmetric positive definite velocity system (5.10).

First, in Section 5.1 we considered two of our industrial collaborator AEA Technology's actual case studies from sites in the UK (the first one being a very basic model, and the second one being a detailed model taken from a recent study of a waste repository site). The two most testing features of realistic groundwater flow problems for iterative methods, are on the one hand, the large jumps in the permeability field $k(\vec{x})$ between different rock strata, and on the other hand, the highly unstructured meshes, which are necessary to accurately model the various rock strata and fault lines. As a consequence, the resulting system (5.9) was highly ill-conditioned and anisotropic in both case studies, and represented a challenging test for our iterative method. However, much to our satisfaction, we were able to report a more than reasonable performance of our method under the circumstances. The method proved to be extremely robust and highly efficient, in particular taking into account its almost optimal parallel efficiency (see Section 5.1.3).

Then, in Section 5.2 we reported on a series of experiments, where the permeability k of a heterogeneous porous medium was modelled using a stochastic spatial process. This approach led to a system (5.1)–(5.4) of stochastic PDEs, where in contrast to Section 5.1, the velocity $\vec{q}$ and the pressure p_R were now random variables. It was pointed out that in order to study statistical properties of those random variables, it is essential to have a fast and efficient solver for (5.1)–(5.4) in the case when k is a typical realisation of the stochastic process. The key aim of this section was therefore to establish whether our numerical method for (5.1)–(5.4) would be accurate and fast enough to serve as such a tool.

In order to test our method, we studied a model problem on the unit square. To begin with, in Section 5.2.1 we defined an appropriate stochastic process that can be used to model heterogeneity in porous media, i.e. we chose k in such a way that $\log(k)$ is a realisation of a Gaussian random field, with mean m, variance σ^2, and correlation length scale λ. This means that k is Hölder continuous, but not in general differentiable (cf. Theorem 5.1). Because of the lack of regularity in this problem, to achieve acceptable numerical accuracy of the solution with our numerical solution, we had to use a mesh which was (uniformly) as fine as possible (i.e. up to 10^6 degrees of freedom). After describing the model problem (cf. Section 5.2.2) and after specifying a suitable stopping criterion (cf. Section 5.2.3), we then tested the robustness and efficiency of our solver. The results showed that the time taken for a solve (with a fixed number of processors) scaled almost linearly (i.e. optimally) in the number of degrees of freedom and was remarkably robust to variations in σ^2 and λ. Moreover, as in Section 5.1.3 we observed almost 100% parallel efficiency.

Appendix A

Asymptotic Mesh Dependency

To obtain asymptotic bounds on the condition number of finite element matrices in the presence of mesh refinement, two kinds of relationships turn out to be crucial:

- *Stability estimates*, which relate the L_2-norm of a finite element function to the Euclidean norm of its coefficient vector (with respect to a chosen basis).
- *Inverse estimates*, which provide mesh dependent bounds for the norms of differential operators.

Either result can be proved using affine techniques by switching to a reference element $\widehat{T}$. While this proof is a simple exercise for scalar valued finite element spaces (see, for example Johnson [63, Section 7.7] for C^0-elements), it is much more complicated in the case of vector valued finite element spaces like $\mathcal{RT}_k(\Omega, \mathcal{T})$ or $\mathcal{ND}_{k+1}(\Omega, \mathcal{T})$. A special transformation known as Piola's transformation has to be used to preserve normal or tangential components of the vector fields, and citing Hiptmair [57, Page 14]: "Putting it bluntly, in the bulk of the finite element literature (c.f. [20, 73]) these [transformation] rules are simply conjured up." Nevertheless, Hiptmair offers a "canonical way" to rigorously develop these rules using differential forms (see also Hiptmair [58]). The results in Theorems A.1 and A.2 (for the lowest order case) are taken from Hiptmair [57, Section 2.5].

Let $\Omega \subset \mathbb{R}^3$ and let $\{\mathcal{T}_h\}$ be a shape regular family of triangulations. Furthermore, let T be an arbitrary element of $\mathcal{T}_h$ with diameter $h(T) \leq h$. As usual, c and C will denote generic positive constants independent of h. Then we have the following estimates for the finite element spaces $RT_0(T)$ and $ND_1(T)$.

Theorem A.1 (Stability estimates).

(a) *Let $\vec{u} \in RT_0(T)$. Then*

$$ch(T)^{-1} \sum_{F \subset \overline{T}} \bar{u}_F^2 \;\leq\; \int_T |\vec{u}|^2 d\vec{x} \;\leq\; Ch(T)^{-1} \sum_{F \subset \overline{T}} \bar{u}_F^2, \tag{A.1}$$

where $\bar{u}_F := \int_F \vec{u} \cdot \vec{\nu}_F ds$, for all faces $F \subset \overline{T}$.

(b) *Let $\vec{u} \in ND_1(T)$. Then*

$$ch(T) \sum_{E\subset\overline{T}} u_E^2 \;\le\; \int_T |\vec{u}|^2 d\vec{x} \;\le\; Ch(T) \sum_{E\subset\overline{T}} u_E^2, \tag{A.2}$$

where $u_E := \int_E \vec{u}\cdot\vec{\tau}_E\, ds$, for all edges $E \subset \overline{T}$.

Proof. See Hiptmair [57, Page 33]. □

Additionally, we have the following inverse estimate for $\vec{\text{curl}}$ on $ND_1(T)$.

Theorem A.2 (Inverse estimate). *Let $\vec{u} \in ND_1(T)$. Then*

$$\int_T |\vec{\text{curl}}\,\vec{u}|^2 d\vec{x} \le Ch(T)^{-2} \int_T |\vec{u}|^2 d\vec{x}. \tag{A.3}$$

Proof. See Hiptmair [57, Theorem 2.38]. □

Note that the coefficients $\bar{u}_F$, $F \subset \overline{T}$, in the stability estimate for $RT_0(T)$ in Theorem A.1(a) correspond to a different basis of $RT_0(T)$ than the one we introduced in (2.75). Since for all $\vec{u} \in RT_0(T)$, the normal component $\vec{u}\cdot\vec{\nu}_F$ is constant on each face F of T, a stability estimate that corresponds to the basis (2.75) can be deduced easily from Theorem A.1(a).

Corollary A.3. *Let $\vec{u} \in RT_0(T)$. Then*

$$ch(T)^3 \sum_{F\subset\overline{T}} u_F^2 \;\le\; \int_T |\vec{u}|^2 d\vec{x} \;\le\; Ch(T)^3 \sum_{F\subset\overline{T}} u_F^2, \tag{A.4}$$

where $u_F := \vec{u}\cdot\vec{\nu}_F$, for all faces $F \subset \overline{T}$.

Proof. Let F be a face of T. The shape regularity condition (2.35) guarantees that the length $h_{min}(T)$ of the smallest edge of T satisfies $h_{min}(T) \ge 2\rho(T) \ge 2\kappa h(T)$ and therefore

$$ch(T)^2 \le |F| \le Ch(T)^2. \tag{A.5}$$

The proof of (A.4) follows directly from (A.1) and (A.5), using the fact that

$$\bar{u}_F = \int_F \vec{u}\cdot\vec{\nu}_F ds = |F|\, u_F, \qquad \text{for all} \quad F \subset \overline{T}.$$

□

Finally, in the proof to Lemma 4.8 we will also make use of the following result.

Lemma A.4. *Let E be an edge of T and let $\vec{\Psi}_E := \vec{\text{curl}}\vec{\Phi}_E$ as defined in (3.60). Then*

$$ch(T)^{-1} \;\le\; \int_T |\vec{\Psi}_E|^2\, d\vec{x} \;\le\; Ch(T)^{-1}. \tag{A.6}$$

Proof. Let E be an edge of T and let $\vec{\Psi}_E := \vec{\text{curl}}\vec{\Phi}_E$ as defined in (3.60). It is shown in the proof to Proposition 3.28 that $\vec{\Psi}_E|_T \in RT_0(T)$. Therefore, we can apply Corollary A.3 and we obtain

$$ch(T)^3 \sum_{F\subset\overline{T}} (\vec{\Psi}_E \cdot \vec{\nu}_F)^2 \;\leq\; \int_T |\vec{\Psi}_E|^2 \, d\vec{x} \;\leq\; Ch(T)^3 \sum_{F\subset\overline{T}} (\vec{\Psi}_E \cdot \vec{\nu}_F)^2. \tag{A.7}$$

Now, adopting the local notation on T, which we introduced in the proof to Proposition 3.28, we can calculate the bounds in (A.7) using (3.63), i.e.

$$ch(T)^3(|F^c|^{-2} + |F^d|^{-2}) \;\leq\; \int_T |\vec{\Psi}_E|^2 \, d\vec{x} \;\leq\; Ch(T)^3(|F^c|^{-2} + |F^d|^{-2}), \tag{A.8}$$

where F^c and F^d are the two faces of T that contain edge E (see also the figure on page 59). The proof of (A.6) follows directly from (A.8) and (A.5). $\square$

Appendix B

Some Results from Graph Theory

Definition B.1. (Berge [15])

(a) A *graph* (or more precisely a *1-graph*) $\mathbf{G}$ is defined to be a pair $(\mathcal{X},\mathcal{U})$, where $\mathcal{X}$ is a set $\{x_1, x_2, \ldots, x_n\}$ of elements called *vertices* (or nodes), and $\mathcal{U}$ is a subset $\{u_1, u_2, \ldots, u_m\}$ of $\mathcal{X} \times \mathcal{X}$ of elements called *arcs* (or *orientated edges*). For an arc $u = (x, y) \in \mathcal{X} \times \mathcal{X}$, the vertex x is called its *initial endpoint*, and the vertex y is called its *terminal endpoint*. A vertex $y \in \mathcal{X}$ is called a *neighbour* of $x \in \mathcal{X}$, if either $(x, y) \in \mathcal{U}$ or $(y, x) \in \mathcal{U}$. The set of all neighbours of a vertex x in the graph $\mathbf{G}$ will be denoted by $\Gamma_{\mathbf{G}}(x)$.

(b) A *partial graph* of a graph $\mathbf{G} = (\mathcal{X},\mathcal{U})$ is a graph $\mathbf{H} = (\mathcal{X}, \mathcal{V})$ with $\mathcal{V} \subset \mathcal{U}$.

(c) A *chain* is a sequence $\mu = (u_{i_1}, u_{i_2}, \ldots, u_{i_q})$ of arcs of a graph $\mathbf{G}$ such that each arc in the sequence has one endpoint in common with its predecessor and its other endpoint in common with its successor. A chain that does not encounter the same vertex twice is called *elementary*. A chain that does not use the same arc twice is called *simple*.

(d) For two vertices x and y of a graph $\mathbf{G}$ let us define the equivalence relation $x \equiv y$ by:

$$[x = y, \text{ or } x \neq y \text{ and there exists a chain in } \mathbf{G} \text{ connecting } x \text{ and } y].$$

The equivalence classes of $\equiv$ are called the *connected components* of $\mathbf{G}$. A *connected graph* is a graph that consists only of one connected component.

(e) A *cycle* is a simple chain whose terminal endpoint coincides with its initial endpoint. Let m be the number of arcs in $\mathbf{G}$. With each cycle μ of $\mathbf{G}$ we can associate a vector $\boldsymbol{\mu} \in \mathbb{Z}^m$ with

$$\mu_i = \begin{cases} 0 & \text{if } u_i \text{ is not in } \mu \\ 1 & \text{if } u_i \text{ is in } \mu \text{ and shares its initial endpoint with its predecessor} \\ -1 & \text{if } u_i \text{ is in } \mu \text{ and shares its terminal endpoint with its predecessor} \end{cases}$$

The set of all those vectors $\boldsymbol{\mu} \in \mathbb{Z}^m$ generates a vector space over $\mathbb{Z}$. We denote this vector space by $\mathcal{V}(\mathbf{G})$.

(f) A *forest* is defined to be a graph without cycles. A *tree* is defined to be a connected graph without cycles.

Theorem B.2. *Let* $\mathbf{G}$ *be a graph with* n *vertices,* m *arcs and* p *connected components. The dimension of* $\mathcal{V}(\mathbf{G})$ *is* $m - n + p$.

Proof. See Berge [15, p.16]. $\square$

Theorem B.3. *Let* $\mathbf{H} = (\mathcal{X}, \mathcal{U})$ *be a graph with* $n > 2$ *vertices. The following properties are equivalent and each characterises a tree:*

(i) $\mathbf{H}$ *is connected and has no cycles.*

(ii) $\mathbf{H}$ *has* $n - 1$ *arcs and has no cycles.*

(iii) $\mathbf{H}$ *is connected and contains* $n - 1$ *arcs.*

(iv) $\mathbf{H}$ *has no cycles and adding an arc creates a unique cycle.*

(v) $\mathbf{H}$ *is connected and removing an arc leaves the remaining graph disconnected.*

(vi) Every pair of vertices x, y *of* $\mathbf{H}$ *is connected by a unique chain.*

Proof. See Berge [15, p.24]. $\square$

Theorem B.4. *Let* $\mathbf{G} = (\mathcal{X}, \mathcal{U})$ *be a connected graph. There exists a partial graph* $\mathbf{H} = (\mathcal{X}, \mathcal{V})$ *such that* $\mathbf{H}$ *is a tree.*

Proof. See Berge [15, p.25]. $\square$

The tree $\mathbf{H}$ obtained from $\mathbf{G}$ as above is called a *spanning tree*. An optimal algorithm to find a spanning tree $\mathbf{H}$ of a connected graph $\mathbf{G}$ is presented in Algorithm B.7.

Theorem B.5. *Let* $\mathbf{G}$ *be a graph with* n *vertices and* $m \geq n$ *arcs. The time spent on Algorithm B.7 is proportional to the number of arcs, i.e.* $O(m)$.

Proof. See Aho et al. [2]. $\square$

Theorem B.6. *Let* $\mathbf{G} = (\mathcal{X}, \mathcal{U})$ *be a connected graph with* n *vertices and* m *arcs, let* $\mathbf{H} = (\mathcal{X}, \mathcal{V})$ *be a spanning tree of* $\mathbf{G}$, *and let* $u_i \in \mathcal{U}$ *be an arc of* $\mathbf{G}$ *not in tree* $\mathbf{H}$, *i.e.* $u_i \notin \mathcal{V}$. *Adding* u_i *to* $\mathbf{H}$ *creates a unique cycle* μ^i *and its associated vector* $\boldsymbol{\mu}^i$ *satisfies* $\mu^i_i = 1$. *The set* $\{\boldsymbol{\mu}^i : u_i \in \mathcal{U} \backslash \mathcal{V}\}$ *forms a basis of* $\mathcal{V}(\mathbf{G})$.

Proof. The existence of $\boldsymbol{\mu}^i$, for all $u_i \in \mathcal{U} \backslash \mathcal{V}$ is guaranteed by virtue of Theorem B.3 (iv). The vectors are linearly independent, since $\mu_i^j = \delta_{i,j}$, for all $u_i, u_j \in \mathcal{U} \backslash \mathcal{V}$. Moreover,

$$\dim\{\boldsymbol{\mu}^i : u_i \in \mathcal{U} \backslash \mathcal{V}\} = \#\mathcal{U} - \#\mathcal{V} = m - (n-1) = \dim \mathcal{V}(\mathbf{G})$$

where in the last step we used Theorem B.2 with $p = 1$. ☐

Algorithm B.7.

```
variables
  n -- number of vertices
  mark[1:n] -- array of flags
begin
  𝒱 = ∅
  for each vertex x ∈ 𝒳 do mark[x] := unvisited
  recursive(x₁)
end

procedure recursive( x -- vertex )
  variables
    y -- vertex
  begin
    mark[x] := visited
    for each vertex y ∈ Γ_G(x) do
      if mark[y] = unvisited then
        𝒱 = 𝒱 ∪ {u}   --   where u ∈ 𝒰 with endpoints x and y
        recursive(y)
  end
```

Appendix C

A Topological Result on Simplicial Triangulations

Definition C.1. (The Fundamental Group) (Armstrong [7, Ch.5])

(a) A *topological space* is a set S together with a collection $\mathcal{U}$ of subsets of S satisfying the following conditions:

(1) $\emptyset \in \mathcal{U}$, $S \in \mathcal{U}$.

(2) If $U_1, \ldots, U_n \in \mathcal{U}$, then $\bigcap_{i=1}^{n} U_i \in \mathcal{U}$.

(3) If $\tilde{\mathcal{U}} \subset \mathcal{U}$, then $\bigcup_{U \in \tilde{\mathcal{U}}} U \in \mathcal{U}$.

The elements of $\mathcal{U}$ are called *open sets* in S. $\mathcal{U}$ is called a *topology* on S.

(b) Let X be a topological space. A *path* in X from x_0 to x_1 (with origin x_0 and end x_1) is a continuous map $\alpha : [0,1] \to X$ such that $\alpha(0) = x_0$ and $\alpha(1) = x_1$. Let α be a path in X from x_0 to x_1 and let β be a path in X from x_1 to x_2. The *product* of α and β is the path $\alpha\beta$ from x_0 to x_2 defined by

$$\alpha\beta(t) = \begin{cases} \alpha(2t) & \text{for } t \in [0, 1/2] \\ \beta(2t-1) & \text{for } t \in [1/2, 1]. \end{cases}$$

The *inverse* of α is the path α^{-1} from x_1 to x_0 defined by $\alpha^{-1}(t) = \alpha(1-t)$.

(c) Two paths α and β from x_0 to x_1 are *homotopic* (written $\alpha \simeq \beta$) if there exists a continuous map $F : [0,1] \times [0,1] \to X$ such that

$$\begin{array}{llll} F(0,t) = x_0 & \text{and} & F(1,t) = x_1 & \text{for all } t \in [0,1], \\ F(s,0) = \alpha(s) & \text{and} & F(s,1) = \beta(s) & \text{for all } s \in [0,1]. \end{array}$$

(d) Let X be a topological space and let $x_0 \in X$. The set of $\simeq$ equivalence classes of paths with origin x_0 and end x_0 forms a group under the operations of multiplication and inverse as defined above. This group is denoted $\pi_1(X, x_0)$ and is called the *fundamental group* of the pair (X, x_0). X is called *simply connected* if its fundamental group is trivial.

Definition C.2. (The First Homology Group) (Armstrong [7, Ch.8])

(a) Let V be a vector space over $\mathbb{R}$, and let $\{v_0, v_1, \dots, v_k\} \subset V$ such that the set $\{v_1 - v_0, \dots, v_k - v_0\}$ is linearly independent. The smallest convex set containing $\{v_0, v_1, \dots, v_k\}$, i.e. the convex hull

$$\{v := \sum_{i=0}^{k} \lambda_i v_i : \lambda_i \geq 0 \text{ and } \sum_{i=0}^{k} \lambda_i = 1\},$$

is called a *simplex of dimension k* (or a *k-simplex*). The points $v_0, v_1, \dots, v_k$ are called the *vertices* (or nodes) of the simplex. The simplices formed by the subsets of $\{v_0, v_1, \dots, v_k\}$ are called the *faces* of the simplex.

(b) A *simplicial complex* K is a finite set of simplices in V such that

(1) if $A \in K$, then the faces of A are also in K;

(2) if $A, B \in K$ and $A \cap B \neq \emptyset$, then $A \cap B \in K$.

The *dimension* of K is the maximum dimension of the simplices of K. The point set union of all simplices in K is denoted by $|K|$.

(c) Let K be a simplicial complex. An *orientated edge* in K is an ordered pair (u, v) such that u and v lie in some simplex of K. An *orientated triangle* in K is an ordered triple (u, v, w) such that u, v, w lie in some simplex of K. Note that $(u, v, w) = (v, w, u) = (w, u, v)$. A change of orientation is denoted by a minus sign, thus $(v, u) = -(u, v)$ and $(v, u, w) = -(u, v, w)$. The *boundary* of the orientated edge (u, v) is defined to be

$$\partial(u, v) = v - u$$

The boundary of the orientated triangle (u, v, w) is

$$\partial(u, v, w) = (v, w) + (w, u) + (u, v)$$

Let n be the number of all edges in K. A linear combination of orientated edges

$$\sum_{i=1}^{n} \lambda_i (u_i, v_i) \quad \text{with the property that} \quad \sum_{i=1}^{n} \lambda_i \partial(u_i, v_i) = 0$$

and $\lambda_i \in \mathbb{Z}$ for all $i = 1, \dots, n$, is called a *(one-dimensional) cycle* of K. A cycle

β is called a *bounding cycle*, if we can find a linear combination

$$\sum_{j=1}^{k} \alpha_j (u_j, v_j, w_j)$$

of orientated triangles in K such that

$$\beta = \sum_{j=1}^{k} \alpha_j \partial (u_j, v_j, w_j).$$

(d) The set of all cycles of K forms an abelian group under the addition

$$\sum_{i=1}^{n} \lambda_i (u_i, v_i) + \sum_{i=1}^{n} \mu_i (u_i, v_i) = \sum_{i=1}^{n} (\lambda_i + \mu_i)(u_i, v_i).$$

We denote this group by $Z_1(K)$. The bounding cycles form a subgroup $B_1(K)$ of $Z_1(K)$. The quotient group

$$H_1(K) = Z_1(K) \backslash B_1(K)$$

is called the *first homology group* of K.

We will only need the following fundamental theorem which is a corollary to the Simplicial Approximation Theorem (Armstrong [7, p.128]).

Theorem C.3. *Let K be a simplicial complex, and let v be a vertex of K. If $|K|$ is connected, abelianising $\pi_1(|K|, v)$ gives the first homology group $H_1(K)$.*

Proof. See Armstrong [7, p.182] □

Corollary C.4. *If $|K|$ is simply connected, then each cycle of K is a bounding cycle.*

Proof. From Definition C.1(d) we know that if $|K|$ is simply connected, then $\pi_1(|K|, v)$ is trivial for any vertex v of K. As a consequence of Theorem C.3 this also implies that $H_1(K)$ is trivial (since abelianising the trivial group has to result in the trivial group again). Therefore $B_1(K) = Z_1(K)$. □

Appendix D

List of Notations

Symbol			**Page**
Ω	:=	Bounded and connected open subset of $\mathbb{R}^d$	8
d	:=	Dimension of Ω (2 or 3)	8
$\vec{x}$	:=	(Physical) vector in $\mathbb{R}^d$	8
$\Gamma = \Gamma_D \cup \Gamma_N$	:=	Polygonal (polyhedral) boundary of Ω partitioned into Γ_D (Dirichlet) and Γ_N (Neumann) with $\Gamma_D \neq \emptyset$ and Γ_N closed	8
$\vec{\nu}(\vec{x})$	:=	Outward unit normal from Ω at $\vec{x} \in \Gamma$	8
$L_p(\Omega)$	:=	$\{u\!:\!\Omega \to \mathbb{R} : \|u\|_{L_p(\Omega)} < \infty\}$, for $1 \leq p \leq \infty$	8
$\|u\|^p_{L_p(\Omega)}$	:=	$\int\limits_\Omega \lvert u\rvert^p d\vec{x}$, for $1 \leq p < \infty$	8
$\|u\|_{L_\infty(\Omega)}$	:=	$\inf\limits_{\substack{A \subset \Omega \text{ s.t.} \\ \mu(A)=0}} \sup\limits_{\vec{x} \in \Omega \setminus A} \lvert u(\vec{x})\rvert$, where $\mu(A)$ is the measure of the set A	8
$f(\vec{x})$	:=	Source term of the PDE in $L_2(\Omega)$	8
$H^{1/2}(\Gamma)$	:=	$\{u \in L_2(\Omega) : \lvert u\rvert_{H^{1/2}(\Gamma)} < \infty\}$	8
$\lvert u\rvert^2_{H^{1/2}(\Gamma)}$	:=	$\int\limits_\Gamma \int\limits_\Gamma (u(\vec{x}) - u(\vec{y}))^2 \lvert\vec{x} - \vec{y}\rvert^{-(d+1)} \, d\vec{x} d\vec{y}$	8
$g_D(\vec{x})$	:=	Dirichlet data in $H^{1/2}(\Gamma_D)$	8
$g_N(\vec{x})$	:=	Neumann data in $L_2(\Gamma_N)$	8
$D(\vec{x})$	:=	Diffusion coefficient tensor with $D_{i,j} \in L_\infty(\Omega)$, $D^{-1}_{i,j} \in L_\infty(\Omega)$	8
$H^m(\Omega)$	:=	$\left\{u \in L_2(\Omega) : \|\partial^{\vec{\alpha}} u\|_{L_2(\Omega)} < \infty \text{ for all } 0 \leq \sum\limits_{i=1}^{d} \alpha_i \leq m\right\}$, where $\partial^{\vec{\alpha}} u$ is a derivative in the sense of distributions	9
$\lvert u\rvert_{H^m(\Omega)}$	:=	$\sum\limits_{\lvert\vec{\alpha}\rvert = m} \|\partial^{\vec{\alpha}} u\|_{L_2(\Omega)}$ where $\lvert\vec{\alpha}\rvert := \sum\limits_{i=1}^{d} \alpha_i$	9
$\|u\|_{H^m(\Omega)}$	:=	$\|u\|_{L_2(\Omega)} + \sum\limits_{j=1}^{m} \lvert u\rvert_{H^j(\Omega)}$	9
$H^1_{0,D}(\Omega)$	:=	$\{u \in H^1(\Omega) : u\rvert_{\Gamma_D} = 0\}$	9
$\operatorname{div} \vec{u}(\vec{x})$	:=	$\vec{\nabla} \cdot \vec{u}(\vec{x})$ (:= Divergence operator)	9
$H(\operatorname{div}, \Omega)$	:=	$\{\vec{v} \in (L_2(\Omega))^d : \operatorname{div} \vec{v} \in L_2(\Omega)\}$	9
$(\vec{u}, \vec{v})_{H(\operatorname{div},\Omega)}$	:=	$\int\limits_\Omega (\vec{u} \cdot \vec{v} + \operatorname{div} \vec{u} \operatorname{div} \vec{v}) \, d\vec{x}$	9
$\|\vec{v}\|^2_{H(\operatorname{div},\Omega)}$	:=	$(\vec{v}, \vec{v})_{H(\operatorname{div},\Omega)}$	9

Symbol			**Page**
$H^{-1/2}(\Gamma)$	:=	Dual space to $H^{1/2}(\Gamma)$	10
$\langle\cdot,\cdot\rangle_\Gamma$	:=	Duality between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$	10
$\vec{v}\cdot\vec{\nu}\vert_\Gamma$	:=	Normal trace of $\vec{v}$ on Γ	10
$H_{0,N}(\mathrm{div},\Omega)$	:=	$\{\vec{v}\in H(\mathrm{div},\Omega) : \langle\vec{v}\cdot\vec{\nu},\Phi\rangle_\Gamma = 0 \text{ for all } \Phi\in H^1_{0,D}(\Omega)\}$	10
$m(\vec{u},\vec{v})$	:=	$\int_\Omega D^{-1}(\vec{x})\vec{u}\cdot\vec{v}\,d\vec{x}$	10
$b(\vec{v},w)$	:=	$\int_\Omega \mathrm{div}\,\vec{v}\,w\,d\vec{x}$	10
$G(\vec{v})$	:=	$\langle\vec{v}\cdot\vec{\nu},g_D\rangle_\Gamma$	11
$F(w)$	:=	$-\int_\Omega f w\,d\vec{x}$	11
$\vec{f}_{\mathcal{D}}(\vec{x})$	:=	Source term for the divergence form of the PDE in $(L_2(\Omega))^d$	11
$F_{\mathcal{D}}(\vec{v})$	:=	$-m(\vec{f}_{\mathcal{D}},\vec{v}) + \langle\vec{v}\cdot\vec{\nu},g_D\rangle_\Gamma$	12
$\mathcal{Z}$	:=	$\{\vec{v}\in H_{0,N}(\mathrm{div},\Omega) : b(\vec{v},w)=0 \text{ for all } w\in L_2(\Omega)\}$	12
$\vec{\nu}_T(\vec{x})$	:=	Outward unit normal from T at $\vec{x}\in\partial T$	14
$u\vert_T$	:=	Restriction of a function $u(\vec{x})$ to $\vec{x}\in T$	14
$\mathcal{D}(\Omega)$	:=	Space of infinitely differentiable fcts. with compact support	15
$\mathcal{T},\mathcal{T}_h,\mathcal{T}_H,\ldots$	:=	Simplicial triangulations of Ω	17
$h(T)$	:=	Diameter of an element $T\in\mathcal{T}$	17
h	:=	Maximum diameter of any of the elements $T\in\mathcal{T}_h$	17
$\mathcal{N},\mathcal{N}_I,\mathcal{N}_D,\mathcal{N}_N$	:=	Sets of nodes of the triangulation $\mathcal{T}$	17
$\mathcal{F},\mathcal{F}_I,\mathcal{F}_D,\mathcal{F}_N$	:=	Sets of faces of the triangulation $\mathcal{T}$	17
$\mathcal{E},\mathcal{E}_I,\mathcal{E}_D,\mathcal{E}_N$	:=	Sets of edges of the triangulation $\mathcal{T}$ in 3D	17
R^+	:=	$\{\vec{x}\in\mathbb{R}^2 : x_1>0\}\cup\left\{\begin{pmatrix}0\\1\end{pmatrix}\right\}$, if $d=2$, otherwise $\{\vec{x}\in\mathbb{R}^3 : x_1>0\}\cup\{\vec{x}\in\mathbb{R}^3 : x_1=0,\ x_2>0\}\cup\left\{\begin{pmatrix}0\\0\\1\end{pmatrix}\right\}$	18
$\vec{\nu}_F$	:=	Unit normal vector in R^+ associated with face F	18
$\vec{\tau}_E$	:=	Unit tangent vector in R^+ associated with edge E	18
k	:=	Order of the finite element space	18
$P_k(T)$	:=	Space of multivariate polynomials of degree $\le k$ over T	18
$RT_k(T)$	:=	Local Raviart-Thomas-Nédélec space over T	18
$\mathcal{RT}_k(\Omega,\mathcal{T})$	:=	Raviart-Thomas-Nédélec space of $H(\mathrm{div},\Omega)$-conforming finite element functions of order k over the triangulation $\mathcal{T}$	20
$\mathcal{P}_k(\Omega,\mathcal{T})$	:=	Space of $L_2(\Omega)$-conforming, discontinuous, piecewise polynomial finite element fcts. of order k over the triangulation $\mathcal{T}$	20
$\mathcal{V}$	:=	$\{\vec{v}\in\mathcal{RT}_k(\Omega,\mathcal{T}) : \vec{v}\cdot\vec{\nu}\vert_{\Gamma_N}=0\}$	21
$n_{\mathcal{V}}$	:=	$\dim\mathcal{V}$	21
$\mathcal{W}$	:=	$\mathrm{div}\,\mathcal{V} \;=\; \mathcal{P}_k(\Omega,\mathcal{T})$	21
$n_{\mathcal{W}}$	:=	$\dim\mathcal{W}$	21
$\#A$	:=	Number of elements of a (finite) set A	21
C, c	:=	Generic constants independent of the mesh diameter h	22

Symbol			**Page**
$\{\vec{v}_i\}$	:=	Basis for $\mathcal{V}$	24
$\{w_j\}$	:=	Basis for $\mathcal{W}$	24
$M_{i,i'}$	:=	$m(\vec{v}_i, \vec{v}_{i'})$:= Raviart-Thomas-Nédélec mass matrix	24
$B_{i,j}$	:=	$b(\vec{v}_i, w_j)$:= Discrete divergence operator in matrix form	24
g_i	:=	$G(\vec{v}_i)$:= Right hand side	25
f_j	:=	$F(w_j)$:= Right hand side	25
$\mathbf{x}$	:=	(General) coefficient vector in $\mathbb{R}^n$, $n \in \mathbb{N}$ (cf. definition of $\vec{x}$ above)	25
$\delta_{i,j}$	:=	Kronecker delta	25
$\vec{v}_F(\vec{x})$	:=	Canonical basis function of $\mathcal{V}$ for $k=0$ associated with face F	25
$w_T(\vec{x})$	:=	Canonical basis function of $\mathcal{W}$ for $k=0$ associated with element T	25
$\|F\|$	:=	Length of face F in 2D, area of face F in 3D	26
$\mathcal{M}$	:=	$\begin{pmatrix} M & B \\ B^T & 0 \end{pmatrix}$	26
μ_{min}, μ_{max}	:=	Minimum and maximum eigenvalues of M	27
σ_{min}, σ_{max}	:=	Minimum and maximum singular values of B	27
$\kappa(A)$	:=	Spectral condition number of a matrix A	28
h_{min}	:=	Minimum diameter of any of the elements $T \in \mathcal{T}_h$	28
$\|\mathbf{x}\|$	:=	$\{\mathbf{x}^T\mathbf{x}\}^{1/2}$:= Euclidean vector norm in $\mathbb{R}^n$	28
$\|T\|$	:=	Area of element T in 2D, volume of element T in 3D	28
$\mathring{\mathcal{V}}$	:=	$\{\vec{V} \in \mathcal{V} : b(\vec{V}, W) = 0 \text{ for all } W \in \mathcal{W}\}$	38
$\mathcal{V}^c$	:=	Complementary space of $\mathring{\mathcal{V}}$ in $\mathcal{V}$	38
$\mathring{n}$	:=	$\dim \mathring{\mathcal{V}}$	38
$\vec{\mathrm{curl}}\,\Phi(\vec{x})$	:=	$(\partial\Phi/\partial x_2, -\partial\Phi/\partial x_1)^T$:= 2D curl operator	40
$C^0(\overline{\Omega})$	:=	Space of continuous functions over $\overline{\Omega}$	41
$\Sigma_{k+1}(T)$	:=	Local set of degrees of freedom for $P_{k+1}(T)$ on T	41
$\mathcal{S}_{k+1}(\Omega, \mathcal{T})$	:=	Space of $H^1(\Omega)$-conforming C^0 or Lagrange finite element functions of order $k+1$ over the triangulation $\mathcal{T}$	41
Σ_{k+1}	:=	Global set of degrees of freedom for $\mathcal{S}_{k+1}(\Omega, \mathcal{T})$	42
$\Phi_P(\vec{x})$	:=	Canonical basis function of $\mathcal{S}_{k+1}(\Omega, \mathcal{T})$ associated with $P \in \Sigma_{k+1}$	42
$\vec{\Psi}_P(\vec{x})$	:=	$\vec{\mathrm{curl}}\,\Phi_P(\vec{x})$:= Basis function of $\mathring{\mathcal{V}}$ in 2D	43
$\mathrm{supp}\, u(\vec{x})$	:=	Support of the function $u(\vec{x})$	44
s_N	:=	Number of connected components in Γ_N	45
Γ_N^ℓ	:=	Connected component in Γ_N	45
$\mathcal{N}_N^\ell$	:=	Set of nodes on Γ_N^ℓ	45
$\sum_{P \in \mathcal{N}_N^\ell} \vec{\Psi}_P(\vec{x})$	:=	Non-local basis function of $\mathring{\mathcal{V}}$ for $k=0$ in 2D associated with Γ_N^ℓ	45
$\chi(M)$	:=	Euler characteristic of compact, two-dimensional surface M in $\mathbb{R}^3$	46
S	:=	$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$	48
Σ_{k+1}^ℓ	:=	Degrees of freedom for $\mathcal{S}_{k+1}(\Omega, \mathcal{T})$ on Γ_N^ℓ	48
$\sum_{P \in \Sigma_{k+1}^\ell} \vec{\Psi}_P(\vec{x})$	:=	Non-local basis function of $\mathring{\mathcal{V}}$ associated with Γ_N^ℓ (general case)	48

Symbol			**Page**
s	:=	Number of connected components in Γ	50
Γ^ℓ	:=	Connected component in Γ	50
$\vec{\Psi}_{2,3}(\vec{x})$	:=	Non-local basis function of $\mathring{\mathcal{V}}$ for multiply connected Ω in 2D	52
$\vec{\mathrm{curl}}\,\vec{\Phi}(\vec{x})$	:=	$\left(\frac{\partial\Phi_3}{\partial x_2} - \frac{\partial\Phi_2}{\partial x_3}\,,\, \frac{\partial\Phi_1}{\partial x_3} - \frac{\partial\Phi_3}{\partial x_1}\,,\, \frac{\partial\Phi_2}{\partial x_1} - \frac{\partial\Phi_1}{\partial x_2}\right)^T$:= 3D curl operator	55
$H(\vec{\mathrm{curl}},\Omega)$	:=	$\{\vec{\Phi} \in (L_2(\Omega))^3 : \vec{\mathrm{curl}}\,\vec{\Phi} \in (L_2(\Omega))^3\}$	55
$\langle\cdot,\cdot\rangle_\Gamma$	:=	Duality between $(H^{-1/2}(\Gamma))^3$ and $(H^{1/2}(\Gamma))^3$	55
$\vec{\Phi} \times \vec{\nu}\vert_\Gamma$	:=	Tangential trace of $\vec{\Phi}$ on Γ	55
$H_0(\vec{\mathrm{curl}},\Omega)$	:=	$\{\vec{\Phi} \in H(\vec{\mathrm{curl}},\Omega) : \left\langle \vec{\Phi} \times \vec{\nu}, \vec{\xi} \right\rangle_\Gamma = 0$ for all $\vec{\xi} \in (H^1(\Omega))^3\}$	55
$ND_{k+1}(T)$	:=	Local Nédélec space over T	57
$(H_0^1(\Omega))^3$	:=	$\{\vec{\xi} \in (H^1(\Omega))^3 : \vec{\xi}\vert_\Gamma = \vec{0}\}$	58
$\mathcal{ND}_{k+1}(\Omega,\mathcal{T})$	:=	Nédélec space of $H(\vec{\mathrm{curl}},\Omega)$-conforming finite element functions of order $k+1$ over the triangulation $\mathcal{T}$	58
$\vec{\Phi}_E(\vec{x})$	:=	Canonical basis function of $\mathcal{ND}_1(\Omega,\mathcal{T})$ associated with edge E	58
$\vec{\Psi}_E(\vec{x})$	:=	$\vec{\mathrm{curl}}\,\vec{\Phi}_E(\vec{x})$:= Basis function of $\mathring{\mathcal{V}}$ in 3D	59
$\vert E\vert$	:=	Length of edge E	60
$\mathbf{G}$	:=	$(\mathcal{N},\mathcal{E})$:= Graph formed by the nodes and edges of $\mathcal{T}$ in 3D	62
$\mathbf{H}$	:=	$(\mathcal{N},\mathcal{H})$:= Spanning tree of $\mathbf{G}$	62
$\mathcal{H}$	:=	Subset of edges (which form the spanning tree $\mathbf{H}$)	62
$\mathcal{V}(\mathbf{G})$	:=	Vector space over $\mathbb{Z}$ generated by the cycles of $\mathbf{G}$	62
μ^F	:=	Elementary cycle of $\mathbf{G}$ formed by the edges E of face F	62
$\boldsymbol{\mu}^F$	:=	Vector in $\mathcal{V}(\mathbf{G})$ associated with μ^F	62
K	:=	Simplicial complex underlying $\mathcal{T}$ ($\vert K\vert$:= its point set union)	62
μ^E	:=	Unique cycle of $\mathbf{G}$ generated by taking edge E into the tree $\mathbf{H}$	63
$\boldsymbol{\mu}^E$	:=	Vector in $\mathcal{V}(\mathbf{G})$ associated with μ^E	63
$\mathcal{E}_N^\ell, \mathcal{F}_N^\ell, \mathbf{H}_N^\ell$	:=	Restrictions of $\mathcal{E},\mathcal{F},\mathbf{H}$ to the component Γ_N^ℓ	64
$\sum_{E\in\mathcal{E}^{1,2}} \vec{\Psi}_E(\vec{x})$	:=	Non-local basis fct. of $\mathring{\mathcal{V}}$ between to disjoint components of Γ_D	65
s_D	:=	Number of connected components in Γ_D	67
Γ_D^j	:=	Connected component in Γ_D	67
$n_{\mathcal{ND}}$	:=	$\dim \mathcal{ND}_{k+1}(\Omega,\mathcal{T})$	68
$\mathcal{F}^c$	:=	Subset of faces that constitutes an index set for the basis of $\mathcal{V}^c$	71
$\mathbf{u}^*$	:=	Particular solution to the constraint problem $B^T\mathbf{u}^* = \mathbf{f}$	76
$\ker A$	:=	Kernel of a matrix or of a linear operator A	78
Z	:=	Matrix with rows $\mathbf{z}_1^T,\ldots,\mathbf{z}_{\mathring{n}}^T$, where $\{\mathbf{z}_i\}$ is a basis of $\ker B^T$	78
$\mathring{A}$	:=	ZMZ^T	78
$\mathring{\mathbf{g}}$	:=	$Z\mathbf{g}$	78
Z^c	:=	Matrix with rows $\mathbf{z}_{\mathring{n}+1}^T,\ldots,\mathbf{z}_{n_\mathcal{V}}^T$, where $\{\mathbf{z}_k\}$ is a complementary basis to the basis $\{\mathbf{z}_i\}$ of $\ker B^T$ above	79
A^c	:=	Z^cB	79

Symbol			**Page**
$\mathbf{g}^c$	:=	$Z^c(\mathbf{g} - M\mathbf{u})$	79
$\mathcal{D}(\vec{x})$	:=	$S^T D(\vec{x}) S$	84
$a(\Phi, \Phi')$	:=	$\int_\Omega \mathcal{D}^{-1}(\vec{x}) \vec{\nabla}\Phi \cdot \vec{\nabla}\Phi' \, d\vec{x}$ (Associated bilinear form in 2D)	84
$\mathcal{A}_{P,P'}$	:=	$a(\Phi_P, \Phi_{P'})$ (Associated matrix in 2D)	84
$A_{P,P'}$	:=	$\mathcal{A}_{P,P'}$ for all $P, P' \in \mathcal{N}_I \cup \mathcal{N}_D$	84
$\mathcal{P}^{-1}_{AS(1)}, \mathcal{P}^{-1}_{AS(2)}$	:=	Overlapping additive Schwarz preconditioners (1-level, 2-level)	89
$a(\vec{\Phi}, \vec{\Phi}')$	:=	$\int_\Omega D^{-1}(\vec{x}) \vec{\text{curl}}\vec{\Phi} \cdot \vec{\text{curl}}\vec{\Phi}' \, d\vec{x}$ (Associated bilinear form in 3D)	94
$\mathcal{A}_{E,E'}$	:=	$a(\vec{\Phi}_E, \vec{\Phi}_{E'})$ (Associated matrix in 3D)	94
$A_{E,E'}$	:=	$\mathcal{A}_{E,E'}$ for all $E, E' \in (\mathcal{E}_I \cup \mathcal{E}_D) \backslash \mathcal{H}$	94
$\mathcal{N}_h$	:=	Set of nodes in a triangulation $\mathcal{T}_h$ of diameter h	96
$\mathbf{H}_h, \mathbf{H}_h^+, \mathbf{H}_h^-$	:=	Spanning trees associated with a triangulation $\mathcal{T}_h$ of diameter h	96
$\mathring{\mathcal{U}}$	:=	$\text{span}\{\vec{\Phi}_E : E \in (\mathcal{E}_I \cup \mathcal{E}_D) \backslash \mathcal{H}_h\}$ where $\mathbf{H}_h := (\mathcal{N}_h, \mathcal{H}_h)$	96
α	:=	Parameter in the Poincaré inequality (4.54)	96
$C(\alpha)$, $c(\alpha)$	:=	Constants independent of h, but depending on α	96
$\zeta(\mathbf{H})$	:=	$\max_{H \in \mathcal{H} \backslash \mathcal{E}_N} \left\{ \sum_{E \in (\mathcal{E}_I \cup \mathcal{E}_D) \backslash \mathcal{H}} \left(\mu_H^E\right)^2 \right\}$ where $\mathbf{H} := (\mathcal{N}, \mathcal{H})$	100
$\text{Diag}(A)$	:=	Diagonal matrix with entries $(\text{Diag}(A))_{i,i} := A_{i,i}$	109
MFlops	:=	Mega Flops := 10^6 floating point operations	109
$\mathcal{P}^{-1}_{RW}, \mathcal{P}^{-1}_{MRW}$	:=	Rusten & Winther preconditioner (and modified version)	110
ρ_{max}	:=	Maximum aspect ratio of a triangulation $\mathcal{T}$	116
$k(\vec{x})$	:=	Permeability tensor	128
$\vec{q}(\vec{x})$	:=	Specific discharge (Darcy velocity)	128
$p_R(\vec{x})$	:=	$p + \rho g z$:= (Residual) fluid pressure	128
μ	:=	Dynamic viscosity of the fluid	128
σ^2	:=	Variance of Gaussian random field	141
λ	:=	Correlation length scale of Gaussian random field	141
$\mathbf{z}^i$	:=	Preconditioned residual at ith iteration	147
$\tilde{\kappa}$	:=	Effective condition number	147

References

[1] R. J. Adler. *The Geometry of Random Fields.* John Wiley, Chichester, 1980.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *Data Structures and Algorithms.* Addison-Wesley, Reading, Massachusetts, 1983.

[3] M. Ainsworth and S. Sherwin. Domain decomposition preconditioners for p and hp finite element approximation of stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 175:243–266, 1999.

[4] R. Albanese and G. Rubinacci. Integral formulation for 3D eddy-current computation using edge-elements. *IEE Proceedings A*, 135(7):457–462, 1988.

[5] P. Alotto and I. Perugia. Mixed finite element methods and tree-cotree implicit condensation. *Calcolo*, 36:233–248, 1999.

[6] T. Arbogast and Z. Chen. On the implementation of mixed methods as nonconforming methods for second-order elliptic problems. *Mathematics of Computation*, 64(211):943–972, 1995.

[7] M. A. Armstrong. *Basic Topology.* Springer, New York, 1983.

[8] D. N. Arnold and F. Brezzi. Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates. *RAIRO Modelisation Mathematique et Analyse Numérique*, 19:7–32, 1985.

[9] D. N. Arnold, R. S. Falk, and R. Winther. Preconditioning in H(div) and applications. *Mathematics of Computation*, 66(219):957–984, 1997.

[10] D. N. Arnold, R. S. Falk, and R. Winther. Multigrid preconditioning in H(div) on non-convex polygons. *Computational and Applied Mathematics*, 17(3):303–315, 1998.

[11] S. F. Ashby, R. D. Falgout, S. G. Smith, and T. W. Fogwell. Multigrid preconditioned conjugate gradients for the numerical solution of groundwater flow on the Cray T3D. In *Proceedings of ANS Conference on Mathematics and Computations, Reactor Physics, and Environmental Analysis*, 1, pages 405–413, Portland, OR, 1995.

[12] R. E. Bank and L. R. Scott. On the conditioning of finite element equations with highly refined meshes. *SIAM Journal on Numerical Analysis*, 26(6):1383–1394, 1989.

[13] J. Baranger, J.-F. Maitre, and F. Oudin. Connection between finite volume and mixed finite element methods. *RAIRO Modelisation Mathematique et Analyse Numérique*, 30(4):445–465, 1996.

[14] S. J. Benbow. *Iterative Methods for Augmented Linear Systems*. PhD thesis, University of Bath, 1997.

[15] C. Berge. *Graphs and Hypergraphs*. North Holland, Amsterdam, 1973.

[16] N. L. Biggs, E. K. Lloyd, and R. J. Wilson. *Graph theory 1736–1936*. Oxford University Press, 1976.

[17] A. Bossavit. *Computational Electromagnetism: Variational Formulation, Complementarity, Edge Elements*. Academic Press, San Diego, 1998.

[18] S. C. Brenner. A multigrid algorithm for the lowest-order raviart-thomas mixed triangular finite element method. *SIAM Journal on Numerical Analysis*, 29(3):647–678, 1992.

[19] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer, Berlin, 1994.

[20] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer, New York, 1991.

[21] Z. Cai, R. R. Parashkevov, T. F. Russell, and X. Ye. Domain decomposition for a mixed finite element method in three dimensions. To appear in *SIAM Journal on Numerical Analysis*.

[22] A.-L. Cauchy. Recherches sur les polyèdres – premier mémoire. *Journal de l'École Polytechnique*, 9(16):68–86, 1813.

[23] T. F. Chan, B. F. Smith, and J. Zou. Overlapping Schwarz methods on unstructured meshes using non-matching course grids. *Numerische Mathematik*, 73:149–167, 1996.

[24] G. Chavent, G. Cohen, J. Jaffre, M. Dupuy, and I. Ribera. Simulation of two-dimensional waterflooding by using mixed finite elements. *Society of Petroleum Engineers Journal*, 24:382–390, 1984.

[25] Z. Chen, R. E. Ewing, and R. Lazarov. Domain decomposition algorithms for mixed methods for second-order elliptic problems. *Mathematics of Computation*, 65(214):467–490, 1996.

[26] Z. Chen, R. E. Ewing, R. Lazarov, S. Maliassov, and Y. A. Kuznetsov. Multilevel preconditioners for mixed methods for second order elliptic problems. *Numerical Linear Algebra with Applications*, 3(5):427–453, 1996.

[27] P.G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North Holland, Amsterdam, 1978.

[28] A. Cliffe, I. G. Graham, R. Scheichl, and L. Stals. Parallel computation of flow in heterogeneous media modelled by mixed finite elements. Bath Mathematics Preprint 99/16, University of Bath, 1999.

[29] A. Cliffe, I. G. Graham, R. Scheichl, and L. Stals. Parallel computation of flow in heterogeneous media modelled by mixed finite elements. *Journal of Computational Physics*, 164(2):258–282, 2000.

[30] L. C. Cowsar, J. Mandel, and M. F. Wheeler. Balancing domain decomposition for mixed finite elements. *Mathematics of Computation*, 64(211):989–1015, 1995.

[31] N. A. C. Cressie. *Statistics for Spatial Data.* Wiley, London, 1993.

[32] M. Crouzeix. In *Proceedings of Journées éléments finis*, Université de Rennes, 1976.

[33] G. Dagan. *Flow and Transport in Porous Formations.* Springer, New York, 1989.

[34] C. R. Dietrich and G. N. Newsam. A fast and exact method for multidimensional Gaussian stochastic simulations. *Water Resources Research*, 29(8):2861–2869, 1993.

[35] J. Douglas, Jr. and J. E. Roberts. Global estimates for mixed methods for second order elliptic equations. *Mathematics of Computation*, 44(169):39–52, 1985.

[36] M. Dryja and O. B. Widlund. Some domain decomposition methods for elliptic problems. In L. Hayes and D. Kincaid, editors, *Iterative Methods for Large Linear Systems*, pages 273–291. Academic Press, San Diego, 1989.

[37] F. Dubois. Discrete vector potential representation of a divergence-free vector field in 3D: Numerical analysis of a model problem. *SIAM Journal of Numerical Analysis*, 27(5):1103–1141, 1990.

[38] R. E. Ewing and J. Wang. Analysis of the Schwarz algorithm for mixed finite element methods. *RAIRO Modelisation Mathematique et Analyse Numérique*, 26(6):739–756, 1992.

[39] R. E. Ewing and J. Wang. Analysis of multilevel decomposition iterative methods for mixed finite element methods. *RAIRO Modelisation Mathematique et Analyse Numérique*, 28(4):377–398, 1994.

[40] R. S. Falk and J. E. Osborn. Error estimates for mixed methods. *RAIRO Modelisation Mathematique et Analyse Numérique*, 14:249–277, 1980.

[41] M. Fortin and R. Glowinski. *Augmented Lagrangian Methods.* North-Holland, Amsterdam, 1983.

[42] L. W. Gelhar. A stochastic conceptual analysis of one-dimensional groundwater flow in nonuniform homogeneous media. *Water Resources Research*, 11:725–741, 1975.

[43] V. Girault. Incompressible finite element methods for Navier-Stokes equations with nonstandard boundary conditions in $\mathbb{R}^3$. *Mathematics of Computation*, 51(183):55–74, 1988.

[44] V. Girault and P. A. Raviart. *Finite Element Methods for Navier-Stokes Equations.* Springer, Berlin, 1986.

[45] I. G. Graham and M. J. Hagger. Additive Schwarz, CG and discontinuous coefficients. In P. E. Bjørstad, M. S. Espedal, and D. E. Keyes, editors, *Domain Decomposition Methods in Science and Engineering*, pages 113–120. Domain Decomposition Press, Bergen, 1998.

[46] I. G. Graham and M. J. Hagger. Unstructured additive Schwarz-CG method for elliptic problems with highly discontinuous coefficients. *SIAM Journal of Scientific Computing*, 20(6):2041–2066, 1999.

[47] D. F. Griffiths. The construction of approximately divergence-free finite elements. In *The Mathematics of Finite Elements and its Applications*, volume 3, pages 237–245. Academic Press, New York, 1979.

[48] K. Gustafson and R. Hartman. Divergence-free bases for finite element schemes in hydrodynamics. *SIAM Journal of Numerical Analysis*, 20(4):697–721, 1983.

[49] A. L. Gutjahr, D. McKay, and J. L. Wilson. Fast Fourier transform methods for random field generation. *Eos Trans. AGU*, 68(44):1265, 1987.

[50] W. Hackbusch. *Elliptic Differential Equations. Theory and Numerical Treatment.* Springer, Berlin, 1987.

[51] M. J. Hagger. Automatic domain decomposition on unstructured grids (DOUG). *Advances in Computational Mathematics*, 9:281–310, 1998.

[52] M. J. Hagger and L. Stals. DOUG *User Guide, Version 1.98.* University of Bath, 1998.

[53] L. J. Hartley, C. P. Jackson, and S. P. Watson. NAMMU *(Release 6.4) User Guide.* AEA Technology, Harwell, 1998.

[54] F. Hecht. *Construction d'une base d'un élément fini P_1 non conforme à divergence nulle dans $\mathbb{R}^3$*. PhD thesis, Université de Paris VI, 1980.

[55] F. Hecht. Construction d'une base de fonctions P1 non-conformes à divergence nulle dans $\mathbb{R}^3$. *RAIRO Modelisation Mathematique et Analyse Numérique*, 15(2):119–150, 1981.

[56] F. Hecht. Construction d'une base pour des éléments finis mixtes à divergence faiblement nulle. Unpublished report, 1988.

[57] R. Hiptmair. *Multilevel Preconditioning for Mixed problems in Three Dimensions.* PhD thesis, University of Augsburg, 1996.

[58] R. Hiptmair. Canonical construction of finite elements. *Mathematics of Computation*, 228(68):1325–1346, 1999.

[59] R. Hiptmair and R.H.W. Hoppe. Multilevel methods for mixed finite elements in three dimensions. *Numerische Mathematik*, 82(2):253–279, 1999.

[60] R. Hiptmair, T. Schiekofer, and B. Wohlmuth. Multilevel preconditioned Augmented Lagrangian techniques for 2nd order mixed problems. *Computing*, 57:25–48, 1996.

[61] R. J. Hoeksema and P. K. Kitanidis. Analysis of the spatial structure of properties of selected aquifers. *Water Resources Research*, 21:563–572, 1985.

[62] C. P. Jackson and S. P. Watson. *Hydrogeological Model Development – Effective Parameters and Calibration*, volume 2 of *Nirex 97: An Assessment of the Post-closure Performance of a Deep Waste Repository at Sellafield.* Nirex Science Report S/97/012, UK Nirex Ltd., Harwell, 1997.

[63] C. Johnson. *Numerical Solutions of Partial Differential Equations by the Finite Element Method.* Cambridge University Press, 1987.

[64] E. F. Kaasschieter. A practical termination criterion for the conjugate gradient method. *BIT*, 28:308–322, 1988.

[65] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal of Scientific Computing*, 20(1):359–392, 1998.

[66] C. E. Kolterman and S. M. Gorelick. Heterogeneity in sedimentary deposits: A Review of structure-imitating, process-imitating and descriptive approaches. *Water Resources Research*, 32(9):2617–2658, 1996.

[67] A. N. F. Mack. An element level zero-divergence finite element approach. *International Journal for Numerical Methods in Fluids*, 19:795–813, 1994.

[68] A. Mantoglou. Digital simulation of multivariate two- and three-dimensional stochastic processes with a spectral turning bands method. *Mathematical Geology*, 19(2):129–149, 1987.

[69] A. Mantoglou and J. L. Wilson. The Turning Bands method for simulation of random fields using line generation by a spectral method. *Water Resources Research*, 18(5):1379–1394, 1982.

[70] W. S. Massey. *Algebraic Topology: An Introduction.* Springer, New York, 1967.

[71] T. P. Mathew. Schwarz alternating and iterative refinement methods for mixed formulations of elliptic problems, part 1: algorithms and numerical results. *Numerische Mathematik*, 65:445–468, 1993.

[72] T. P. Mathew. Schwarz alternating and iterative refinement methods for mixed formulations of elliptic problems, part 2: convergence theory. *Numerische Mathematik*, 65:469–492, 1993.

[73] J. C. Nédélec. Mixed finite elements in $\mathbb{R}^3$. *Numerische Mathematik*, 35:315–341, 1980.

[74] J. C. Nédélec. Éléments finis mixtes incompressibles pour l'équation de Stokes dans $\mathbb{R}^3$. *Numerische Mathematik*, 39:97–112, 1982.

[75] 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.

[76] L.F. Pavarino and M. Ramé. Numerical experiments with an overlapping additive Schwarz solver for 3D parallel reservoir simulation. *International Journal of Supercomputer Applications*, 9(1):3–17, 1995.

[77] P. A. Raviart and J. M. Thomas. A mixed finite element method for 2-nd order elliptic problems. In I. Galligani and E. Magenes, editors, *Mathematical Aspects of the Finite Element Method*, Lecture Notes in Mathematics 606, pages 292–315. Springer, New York, 1977.

[78] J. E. Roberts and J. M. Thomas. Mixed and hybrid methods. In P. G. Ciarlet and J. L. Lions, editors, *Handbook of Numerical Analysis*, volume 2. North Holland, Amsterdam, 1991.

[79] M. J. L. Robin, A. L. Gutjahr, E. A. Sudicky, and J. L. Wilson. Cross-correlated random field generation with the direct Fourier transform method. *Water Resources Research*, 29:2385–2397, 1993.

[80] T. Rusten, P. S. Vassilevski, and R. Winther. Interior penalty preconditioners for mixed finite element approximations of elliptic problems. *Mathematics of Computation*, 65(214):447–466, 1996.

[81] T. Rusten, P. S. Vassilevski, and R. Winther. Domain embedding preconditioners for mixed systems. *Numerical Linear Algebra with Applications*, 5:321–345, 1998.

[82] T. Rusten and R. Winther. A preconditioned iterative method for saddlepoint problems. *SIAM Journal on Matrix Analysis and Applications*, 13(3):887–904, 1992.

[83] T. Rusten and R. Winther. Substructure preconditioners for elliptic saddle point problems. *Mathematics of Computation*, 60(201):23–48, 1993.

[84] Y. Saad. *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company, Boston, 1996.

[85] R. Scheichl. A decoupled iterative method for mixed problems using divergence-free finite elements. Bath Mathematics Preprint 00/11 (submitted to *SIAM Journal on Scientific Computing*), University of Bath, 2000.

[86] D. Silvester and A. Wathen. Fast iterative solution of stabilised Stokes systems. part II: Using general block preconditioners. *SIAM Journal on Numerical Analysis*, 31(5):1352–1367, 1994.

[87] F. Thomasset. *Numerical Solution of the Navier-Stokes Equations by Finite Element Methods*. Von Karman Institute Lecture Notes, no. 86. (Computational Fluid Dynamics, March 21-25, 1977). Springer, New York, 1977.

[88] F. Thomasset. *Implementation of Finite Element Methods for Navier-Stokes Equations*. Springer, New York, 1981.

[89] A. F. B. Tompson, R. Ababou, and L. W. Gelhar. Implementation of the three-dimensional Turning Bands random field generator. *Water Resources Research*, 25(10):2227–2243, 1989.

[90] P. S. Vassilevski and R. D. Lazarov. Preconditioning mixed finite element saddle-point elliptic problems. *Numerical Linear Algebra with Applications*, 3(1):1–20, 1996.

[91] C. Wagner, W. Kinzelbach, and G. Wittum. Schur-complement multigrid, a robust method for groundwater flow and transport problems. *Numerische Mathematik*, 75:523–545, 1997.

[92] A. Wathen and D. Silvester. Fast iterative solution of stabilised Stokes systems. part I: Using simple diagonal preconditioners. *SIAM Journal on Numerical Analysis*, 30(3):630–649, 1993.

[93] B.I. Wohlmuth, A. Toselli, and O.B. Widlund. An iterative substructuring method for Raviart-Thomas vector fields in three dimensions. *SIAM Journal on Numerical Analysis*, 37(5):1657–1676, 2000.

[94] X. Ye and G. Anderson. The derivation of minimal support basis functions for the discrete divergence operator. *Journal of Computational and Applied Mathematics*, 61:105–116, 1995.

[95] X. Ye and C. A. Hall. The construction of an optimal weakly divergence-free macroelement. *International Journal for Numerical Methods in Engineering*, 36:2245–2262, 1993.