Università degli Studi di Udine
Dipartimento Politecnico di Ingegneria ed Architettura
Dottorato di Ricerca in Ingegneria Industriale e dell'Informazione

Ph.D. Thesis

Geometric and Topological aspects of Mimetic Numerical Schemes

Candidate
Silvano Pitassi

Supervisor
Prof. Ruben Specogna

Reviewers
Prof. Ana Alonso Rodríguez
Prof. Daniele Di Pietro

Cycle XXXIV — A.Y. 2020-2021

Institute Contacts
Dipartimento Politecnico di Ingegneria ed Architettura
Università degli Studi di Udine
Via delle Scienze, 206
33100 Udine — Italia
https://dpia.uniud.it/

Author's Contacts
silvano.pitassi@gmail.com

To my family.

Acknowledgements

I would like to thank my scientific advisors Ruben Specogna and Francesco Trevisan. Thanks for their invaluable support, for the many hours spent through explanations and discussions of new and old ideas, to all scientific insights as well as the careful correction of all my drafts. I really liked our collaboration. Special thanks to Prof. Riccardo Ghiloni for the short stints I spent at the University of Trento in the past three years: his level of scientific rigour and knowledge inspired me to be as good as I can be in what I do. Thanks to Michele Libralato for the many interesting conversation and discussion that kept me improving at personal as well as scientific level during these three years Finally, my biggest thanks goes to my whole family. Thanks for supporting me through all the academic degrees and for their sacrifice that I have been appreciating more and more when growing up.

Abstract

Mimetic or compatible numerical schemes are designed to preserve the fundamental properties of physical and mathematical models, such as conservation laws, at the discrete level. To this end, methods of algebraic topology and differential geometry play a fundamental role to design its basic building blocks, like *reconstruction operators* or *discrete Hodge operators* and their algebraic realization given by *mass matrices*.

In this thesis we provide new geometric viewpoints of low-order compatible numerical schemes. In particular, two key principles will guide our constructions. First, a tight relation between reconstruction operators and geometric elements of the *barycentric dual grid*. Second, a decomposition of mass matrices as the sum of a *consistent* and a *stabilization part*. We will use these principles to extended and improve the basic building blocks at the core of mimetic numerical schemes as well as their range of applicability.

We introduce the novel geometric concept of P_0-consistency which generalizes the standard consistency requirement of the mimetic methods. Fundamentally, it shows that geometric elements of a secondary grid, precisely, a *barycentric dual grid*, are not only useful but they are implicitly present in low-order mimetic numerical schemes, even if not made explicit. This fact has two consequences. First, it provides the equivalence between mimetic numerical schemes and discrete geometric approaches. Second, it is the key principle to extend the classical mimetic methods to grids having curved faces. Indeed, all standard mimetic methods only deal with polyhedral grids, thus having planar faces.

Then, we introduce a new construction of *sparse inverse mass matrices* for arbitrary tetrahedral grids and possibly inhomogeneous and anisotropic materials, debunking the conventional wisdom that the barycentric dual grid prohibits a sparse representation for inverse mass matrices. In particular, we provide a unified framework for the construction of both edge and face mass matrices and their sparse inverses as the sum of consistent and a stabilization part.

Next, we address the problem of computing *discrete vector potentials*. Currently, the most efficient methods to compute them are based on the so-called *tree-cotree decomposition*. However, tree-cotree techniques suffer from well-known termination problems that we show be related to topological obstructions of the three-dimensional space. We propose a new algorithm based on *discrete Morse theory* that is able to deal with such topological obstructions.

Finally, we extend the range of applicability of mimetic numerical schemes by introducing a novel mimetic volume integral method to solve eddy current problems. Integral methods for solving eddy current problems use Biot-Savart law to produce non-local constitutive relations that lead to fully populated generalized mass matrices. Yet, these

formulations are very appealing because only the mesh of conductors is needed. We show how our novel mimetic method solves the three main problems of volume integral methods. First, the computation of the inductance matrix elements is slow and also delicate because of the singularity in the integral equation. We exploit constant basis functions that allow a much faster inductance matrix construction with respect to the standard one based on the Rao–Wilton–Glisson (RWG) or Raviart–Thomas (RT) basis functions. Second, our basis functions work for polyhedral elements while producing the same results as RWG and RT basis functions for tetrahedral grids. Third, the new basis functions allow to factorize the inductance matrix and to introduce a novel family of groundbreaking low-rank inductance matrix compression techniques that show several orders of magnitude improvement in memory occupation and computational effort than state-of-the-art alternatives, allowing to solve problems that otherwise cannot be faced.

Contents

I Differential methods **1**

1 Introduction **3**

2 Mimetic numerical schemes **7**
2.1 Geometry of primal and dual grid 7
2.2 Degrees of Freedom 11
2.3 Discrete differential operators 12
2.4 Reconstruction operators 13
2.5 Mimetic inner products: Mass matrices 13
2.5.1 Global mass matrix 15
2.6 Derived discrete operators 16
2.7 Chains and cochains 16
2.8 Mimetic discretization of stationary currrent conduction problem 18
2.8.1 Survey of standard formulations 19
2.8.2 Square resistor benchmark 19

3 Equivalence between mimetic numerical schemes and discrete geometric approaches **21**
3.1 Definition of P_0-consistent reconstruction operators 22
3.1.1 P_0-consistent face reconstruction operators 23
3.1.2 P_0-consistent edge reconstruction operators 24
3.1.3 Linear system formulation of P_0-consistent reconstruction operators 25
3.2 Dual grid P_0-consistent reconstruction operators 27
3.2.1 Reconstruction formulas and their proofs 27
3.2.2 Mass matrices 29
3.2.3 Local mass matrices 29
3.2.4 Optimizing dual grid reconstruction operators 31
3.3 Numerical results 34
3.3.1 Multi-material patch tests 34
3.3.2 Square resistor benchmark 35
3.3.3 Balance laws and dual grid P_0-consistent reconstruction operators 37
3.4 Conclusions 38

4 Curved mimetic method **41**
4.1 Curved MFD method . . . 43
4.2 Numerical results . . . 46
4.2.1 Patch test . . . 46
4.3 Conclusions . . . 46

5 Explicit geometric construction of sparse inverse mass matrices **49**
5.1 Sparse inverse mass matrices . . . 50
5.1.1 Dual cell reconstruction . . . 51
5.1.2 Local inverse mass matrices . . . 55
5.2 Handling material parameter discontinuities inside dual cells . . . 57
5.2.1 Weighted average . . . 57
5.2.2 Piecewise constant vector field . . . 58
5.2.3 Hybrid approach to handle material discontinuities inside dual cells 59
5.3 Numerical results . . . 60
5.3.1 Formulations with one unknown per element . . . 60
5.3.2 Classical patch tests . . . 61
5.3.3 Multi-material patch tests . . . 61
5.3.4 Square resistor benchmark . . . 63
5.4 Conclusions . . . 63

6 Computing discrete vector potentials **65**
6.1 Notation . . . 66
6.2 Discrete Morse Theory . . . 67
6.2.1 Informal introduction to discrete Morse theory . . . 67
6.2.2 Acyclic matchings . . . 68
6.2.3 Basis transformations associated with an acyclic matching . . . 70
6.3 Acyclic matchings and Gaussian elimination . . . 72
6.3.1 Acyclic matchings and Gaussian elimination . . . 72
6.3.2 On the problem of constructing complete acyclic matchings: the case of tree-cotree techniques . . . 73
6.3.3 The case where we cannot find a complete acyclic matching . . . 78
6.4 Algorithm description . . . 79
6.4.1 Greedy approach to construct acyclic matchings . . . 79
6.4.2 Recursive algorithm . . . 80
6.5 Numerical results . . . 82
6.5.1 Triangulations coming from real case boundary value problems . 83
6.5.2 Bing's House . . . 83
6.5.3 Knot-theoretic obstructions . . . 84
6.6 Conclusions . . . 86

II Integral methods **87**

7 A foreword on integral methods to solve eddy current problems **89**
7.1 The eddy current problem . . . 90
7.2 EFIE: discretization on tetrahedral grids . . . 91

7.2.1 Solution based on the electric vector potential and additional DoFs: the CARIDDI code . . . 93
7.2.2 Solution based on mesh current analysis (MCA): the unstructured Partial Elements Equivalent Circuit (PEEC) for eddy currents . 94

8 Mimetic Volume Integral method 95
8.1 Generalization to hexahedral and polyhedral meshes . . . 96
8.2 A novel Mimetic Volume Integral (MVI) method . . . 97
8.2.1 Consistent and positive-definite resistance and magnetic mass matrices . . . 97
8.2.2 Enforcing discrete solenoidal current . . . 98
8.3 Novel interpretation as electrical circuits . . . 100
8.4 Bridging all volumetric integral methods for solving eddy currents . . . 103
8.4.1 Equivalence of MVI with MCA and CARRIDI on tetreahedral grids 103

9 From VU basis functions to inductance matrix factorization 107
9.1 Speeding up assembly with VU basis function . . . 107
9.1.1 Singularity extraction with VU basis functions . . . 108
9.1.2 Factorization of the inductance matrix: MAGICA . . . 109
9.2 A new family of compression techniques . . . 112
9.2.1 LIME: a Lossless Integral Matrix comprEssion . . . 113
9.2.2 Approximated compressions of integral matrices . . . 113
9.3 Numerical results . . . 117
9.3.1 FAIME approach accuracy and convergence . . . 117
9.3.2 Assessing the asymptotically linear behaviour of FAIME . . . 122
9.3.3 A prismatic mesh of a plate with seven holes and a printed circuit board coil . . . 125
9.4 Conclusion . . . 126

10 Conclusion 129

List of Tables

6.1 Running times of Algorithm 3 for the modified TEAM benchmark example for triangulations of decreasing size. 83
6.2 Running times of Algorithm 3 for various triangulations of the thick Bing's House. 84
6.3 Running times of Algorithm 3 for the Furch's knotted balls. 86

9.1 Mesh data and calculation time in the Windows Server for the solid sphere benchmark 119
9.2 Mesh data and calculation time in the Windows Server for the TEAM Workshop Problem 7 119
9.3 Performance comparison in the Windows Server between state-of-the-art approaches and this paper code. Symmetry of the mass matrices is exploited in all the approaches. 125
9.4 Mesh data and calculation time in the Windows Server for the conducting plate with seven holes 126

List of Figures

2.1 A polyhedral element $c \in C$ of K. (a) Primal and (b) dual geometric elements of c. 8

2.2 A tetrahedron $c \in C$ of K. (a) Dual cell $\tilde{c}_{n|c}$. (b) Dual face $\tilde{f}_{e|c}$. (c) Dual edge $\tilde{e}_{f|c}$. 8

2.3 (a) The geometry of the square resistor benchmark ($h = 1\,\mathrm{m}$, $d = 4\,\mathrm{m}$, and $l = 2\,\mathrm{m}$). The two electrodes are depicted in red and blue. (b) Thanks to the symmetry, the computational domain has been reduced to one in eight of the resistor. 20

3.1 Polyhedral grid made by 2 tetrahedra and 3 square pyramids. 26

3.2 A cubic cell c. A set of DoFs $\mathbf{\Phi}$ are attached to faces. This is an example of DoFs that are not image of any constant vector field under the projection map $P_c^{\mathcal{F}}$. 31

3.3 A tetrahedron to illustrate the geometric quantities involved in the proof of Lemma 3 33

3.4 The *series* multi-material patch test (a). We set the voltage between the two electrodes, represented in gray in the picture, to 1 V. (b) Electric field $\boldsymbol{E}$ produced by the SP formulation. (c) Current density field $\boldsymbol{J}$ produced by the SP formulation. (d) Electric field $\boldsymbol{E}$ produced by the VP formulation. (e) Current density field $\boldsymbol{J}$ produced by the VP formulation. 35

3.5 The *parallel* multi-material patch test (a). We set the voltage between the two electrodes, represented in gray in the picture, to 1 V. (b) Electric field $\boldsymbol{E}$ produced by the SP formulation. (c) Current density field $\boldsymbol{J}$ produced by the SP formulation. (d) Electric field $\boldsymbol{E}$ produced by the VP formulation. (e) Current density field $\boldsymbol{J}$ produced by the VP formulation. 35

3.6 The four different types of grids used to discretize the geometry of the square resistor in Fig. 2.3. (a) structured tetrahedral grid. (b) unstructured tetrahedral grid. (c) structured hexahedral grid. (d) polyhedral grid with *subgridding*. 36

3.7 Results for the square resistor benchmark. 37

4.1 Structure of matrix $\mathbb{P}^{\mathcal{F}}$. Face vector of each internal face appears on two different rows with opposite sign, where the two rows correspond to the unique two elements sharing the internal face; instead, each boundary face appears only in one block corresponding to the unique element containing it. For instance, the internal face f_i is shared between elements c, c' and its face vector $\boldsymbol{f}_i$ appears with opposite sign on different rows corresponding to elements c, c'; the boundary face f_k is contained in the unique element c'' and its face vector $\boldsymbol{f}_k$ appears only in the rows corresponding to element c''. 44

4.2 A curved grid partitioning the cube resistor where all internal faces of each element are curved. (a) Uniform electric field reconstructed inside each curved element; it coincides with the analytical value. (b) In red, dual grid structure associated with P_0-consistent face reconstruction operators solution of (4.5). 47

4.3 A curved grid partitioning the cubic resistor where all internal faces of each element are curved. (a) Uniform electric field reconstructed inside each curved element; it coincides with the analytical value. (b) In red, dual grid structure associated with P_0-consistent face reconstruction operators solution of (4.5). 48

5.1 (a) Geometric construction of $s_{n,e_{|c}}$. (b) Geometric construction of $l_{n,f_{|c}}$. 52

5.2 (a) The definition of a patch test as the solution of an electric conduction problem inside a planar resistor. (b) All mentioned formulations are able to retrieve the analytical solution up to machine precision or iterative solver tolerance. 62

5.3 (a) The *series* multi-material patch test. (b) Electrostatic field $\boldsymbol{E}$ obtained with the series. (c) Current density field $\boldsymbol{J}$ obtained with the series. (d) The *parallel* multi-material patch test. (e) Electric field $\boldsymbol{E}$ obtained with the parallel. (f) Current density field $\boldsymbol{J}$ obtained with the parallel. 62

5.4 Results for the square resistor benchmark. 63

6.1 (a) Elementary collapse of free pair (σ, τ). (b) Internal collapse of pair (σ, τ); the resulting cell complex is not more simplicial. 68

6.2 (a) The simplicial complex K. (b) Elementary collapse of the free pair (e_1, f_2). (c) Internal collapse of the pair (e_5, f_2); note that the geometric realization of the resulting cell complex is not more simplicial. 71

6.3 General block structure of matrix $\mathbb{C}$ produced by Algorithm 3. 82

6.4 Two views of the considered 3-dimensional thickening of a Bing's house with two rooms. 84

6.5 A knotted spanning arc in a 3-ball Ω at the core of Furch's construction. 85

8.1 The hexahedron v used in the counterexample. The coordinates of the nodes are $\boldsymbol{p}_{f_1} = (0,0,0)^T$, $\boldsymbol{p}_{f_2} = (2,0,0)^T$, $\boldsymbol{p}_{f_3} = (0,1,0)^T$, $\boldsymbol{p}_4 = (1,1,0)^T$, $\boldsymbol{p}_5 = (0,0,1)^T$, $\boldsymbol{p}_6 = (2,0,1)^T$, $\boldsymbol{p}_7 = (0,1,1)^T$, $\boldsymbol{p}_8 = (1,1,1)^T$. 96

8.2 Examples of cohomology generators $H^2(K, \partial K)$ and $H^1(\partial K)$ for a solid torus. a) The support of a representative $\boldsymbol{g} \in H^2(K, \partial K)$ generator. b) The dual of $\boldsymbol{g}$ is a cycle made of dual edges that are dual to the faces of g. c) The support of two representatives $\boldsymbol{g}_1, \boldsymbol{g}_2$ of the $H^1(\partial K)$ generators of ∂K. d) The support of the homology generator $D(g_1)$ is constituted by dual edges that are dual to the primal edges of $\boldsymbol{g}_1$. 99

9.1 Double integral calculation over a tetrahedron by varying integration order n. For the double numerical integration, at each point of the graph, a pair of n integration orders is intended to be applied. 110
9.2 Double integral calculation over a pair of tetrahedra whose mutual distance is successively increased. As a reference value to compute the error for this test, the "$Sing.\,Extr.$, $\mathcal{O}(x^{16})$" case is used since in the previous plot this approach was shown to be accurate. 111
9.3 Flow chart of the iterative system solution with LIME. 114
9.4 Ohmic losses trend for simplicial and hexahedral meshes successively refined. Top: P_{diss} convergence versus the computation time. Bottom: P_{diss} convergence versus mesh elements n_v. For this chart $P_{diss}^{REF} = 111.49$ mW. 118
9.5 Real($\boldsymbol{J}$) colour map at $f = 50$ Hz in the TEAM 7 conducting plate. . . 120
9.6 Real part of the vertical induction field component along A1-B1 sample line. 120
9.7 Real part of the vertical induction field component along A2-B2 sample line. 121
9.8 Frequency variation and GMRES iterates values on TEAM 7 problem configuration. 122
9.9 Left-bottom corner of TEAM 7 mesh made of mixed hexahedra and triangular prisms. In the inset, for this mesh, Real($\boldsymbol{J}$) colour map at $f = 50.0$ Hz is shown as a comparison to the one previously obtained in Fig. Fig. 9.5.123
9.10 Mesh elements and total computational time comparison 124
9.11 Main picture: colour map of the real part of the current density field in the conducting plate with seven holes. Inset: a detail of the triangular faces of the prisms forming the mesh are depicted in one fourth of the original mesh. 127
9.12 Main picture: colour map of the real part of the current density field in a PCB coil. Inset: zoom of current density distribution. 128

I

Differential methods

1

Introduction

Mimetic or compatible numerical schemes are designed to preserve the fundamental properties of physical and mathematical models, such as conservation laws, at the discrete level. To this end, methods of algebraic topology and differential geometry play a fundamental role. Specifically, all the discrete structures necessary to develop a compatible discrete framework are put together by three basic ingredients.

To begin with, a choice of discrete representation of scalar and vector fields on a given polyhedral grid is necessary, and is given by arrays of *degrees of freedom* (DoFs) [1]. These DoFs are defined through the *projection map* (or de Rham map [2]), i.e. the integration of the fields on specific geometric elements, where their geometric localization results from the physical nature of the fields [1], [3]. An array of DoFs has a counterpart notion in algebraic topology [4], called cochain [1].

Next, *discrete differential operators* acting on degrees of freedom are required. These are obtained by mimicking the fundamental theorem of calculus, namely the Stokes theorem, and they result in topological operators, as they are defined through *incidence matrices* of the grid [1], [5], [3]. This has a counterpart in concepts of algebraic topology as well, being the incidence matrices a matrix representation of the coboundary operator [4]. So, the coboundary operator is the discrete analogue of the gradient, curl and divergence operators [1].

Finally, *reconstruction operators* remap degrees of freedom back to continuous vector-valued functions, thus providing a left inverse of the projection map. The composition of the projection operator with the corresponding reconstruction map provides an interpolation operator. A reconstruction operator is said to be of *low-order* when the action of this interpolation operator leaves fixed element-wise constant vector fields.

In this thesis we focus on low-order methods for various reasons. First of all, most real three-dimensional problems are large and their exact solutions are hardly ever smooth. In these cases, low-order methods are often preferred in order to reduce the number of unknowns. Moreover, high order schemes suffer from the lack of high order representation of curved geometry in off-the-shelf mesh generators. Finally, physical and geometric parameters are generally known with a tolerance in the percent range, which means that extreme accuracy can be hardly justified for real-case industrial applications.

A central role played by reconstruction operators is in the construction of the *discrete Hodge operator* [2], [6], which provides a discrete inner product on the space of discrete vector fields and it is required to satisfy two properties [7], [8]. The first one is a *polynomial consistency condition*, which is an exactness property on a well-defined family of polynomials. In particular, for the low-order case we want to enforce P_0-polynomial consistency, where symbol P_0 refers to exactness for constant fields (i.e. polynomials of degree 0). The second one is a *stability condition*, which ensures the well-posedness of the numerical scheme. From the practical point of view, the *patch test* can be used to test if both requirements are satisfied [9]. Both consistency and stability conditions are stated at a global level, since they are properties of the global mass matrix representing the algebraic realization of the discrete Hodge operator. However, as we will see, the global mass matrix is constructed from local mass matrices, and the latter are constructed from local reconstruction operators. Thus, the design of a global mass matrix that satisfy the consistency and stability conditions must start at the local level with a suitable choice of local reconstruction operators.

A well established design strategy decomposes a local mass matrix as the sum of a *consistent* and a *stabilization* part [10], [8]. Such a decomposition of the local mass matrix is suggestive since each term play a specific role, namely, the consistent term enforces the consistency property while the stabilization term ensures positive-definiteness. Since the building blocks of local mass matrices are reconstruction operators, a similar decomposition can be equivalently performed on reconstruction operators [11].

We can identify two main approaches to design low-order compatible numerical schemes. The first approach relies on a primal-dual grid structure and the related staggered positioning of the discrete variables. In this category fall most physics-compatible discretizations introduced so far, like the Discrete Geometric Approach (DGA) [12], [13], [14], cell method (CM) [15], [5], [16], generalized finite differences [17] as well as in Compatible Discrete Operators (CDO) [18], [19], [11]. Once a dual grid structure is assumed from the beginning as part of the numerical scheme, the consistency condition is satisfied by resorting to the so called *Bossavit's consistency criterion* [7], [17], and, as already recognized in [11], [19], reconstruction operators have a clear geometric interpretation being defined by geometric elements of the dual grid.

In other physics-compatible methods like the Mimetic Finite Difference method (MFD) [10], [8], Finite Volumes (FV) [20], reconstruction operators are defined by mimicking the continuous Stokes theorem, but without introducing at all a dual grid. This strategy of designing reconstruction operators is also employed in the high order generalization of the MFD schemes [21], and in some recent high order compatible methods like Discrete De-Rham Method (DDR) [22] and Virtual Element Methods (VEM) [23], [24], which include the low-order version as a limit case.

In this thesis we provide new geometric and topological viewpoints of low-order compatible numerical schemes. Two key principles will be used extensively throughout this thesis. The first one is the relation between reconstruction operators and the barycentric dual grid and second one is the decomposition of local mass matrices into a consistent and a stabilization part. We will use these principles to improve and extend the basic building blocks of low-order mimetic schemes.

Thesis overview and contributions

The thesis new geometric and topological viewpoints of low-order compatible numerical schemes and is articulated in two parts.

Differential methods

Part I deals with differential formulations using mimetic numerical schemes and the fundamental concepts that will be recurrent throughout the thesis. First of all, the basic building blocks of mimetic numerical schemes are introduced in Chapter 2. In particular, we introduce reconstruction operators and we illustrate how to construct local mass matrices and global mass matrices, in this order, starting from them. The concept of reconstruction operators is further investigated in Chapter 3, where we show that standard reconstruction operators defined using Stokes theorem lead to reconstruction operators defined by geometric elements of the barycentric dual grid. This result shows that the reconstruction operators used in the two main families of low-order compatible numerical schemes are equivalent. Then, the novel class of P_0-consistent reconstruction operators is presented, explaining how it extends the classical consistency requirement of mimetic numerical schemes. In Chapter 4 we exploit P_0-consistent reconstruction operators to design a new mimetic numerical scheme on grids having curved faces. Specifically, we provide a formulation whose convergent scheme is a symmetric discrete problem that uses one DoF for each curved face. In Chapter 5, we provide a new construction of sparse inverse mass matrices on arbitrary tetrahedral grids. This result answers to an open problem on whether the barycentric dual grid allows for such a sparse realization. The key point on which the construction is based are indeed the two principles highlighted at the end of the previous section. To conclude Part I, in Chapter 6 we present a new algorithm to compute discrete vector potentials.

Integral methods

Part II of this thesis is devoted to integral methods. Integral formulations applied to electromagnetic problems have become popular since many years ago especially for the study of full wave models used to study electromagnetic scattering like, among others methods, the partial elements equivalent circuit (PEEC) proposed in 1974 by A. E. Ruehli [25]. The exploitation of such a formulation to study eddy currents successively started around the '90s with the works of R. Rubinacci, R. Martone and G. Albanese [26, 27] by whom the contributions of this thesis are inspired and, simultaneously, with the research activity of L. Kettunen and L. R. Turner [28]. Finally, more recently, the topic has been rediscovered for instance by G. Meunier [29], among others.

Integral formulations applied to eddy current problems are still an open issue in the community of computational electromagnetics due to two main weaknesses. First, the integral constitutive relation used to link the magnetic vector potential to the current density leads to a dense matrix whose size is proportional to the square of the number of DoFs of the problem. This dense matrix leads to storage troubles when it has to be assembled and to memory overflows when computing the solution. As a consequence, researchers have tried to improve the behaviour of the solvers and reduce the size of the matrix by compression techniques. The problem of matrix storage has been successfully

faced initially by resorting to fast multipole methods as in [30, 31] and then by means of effective compression techniques like the adaptive cross approximation (ACA) [32, 33] and the use of hierarchical matrix algebra [34]. These techniques allow to solve problems up to 20,000 DoFs without any compression whereas in past limitations of the calculators memory imposed to deal with a number of DoFs of one order of magnitude smaller. Nevertheless, if from one hand the issue of the system solution has been almost completely faced, on the other hand the proposal of new techniques that can produce further steps forward in the direction of limiting the problem size is an aspect on which some work has still to be done.

This part of the thesis provides the foundations of novel compatible integral methods for solving eddy current problems that mitigate the issues described above of integral methods. It is remarked that this method is deliberately focused on low-order methods for the same reasons detailed in the previous section. In addition, low-order methods enable a clear circuit interpretation of the integral equations which is the main reason for the success of computer codes like FASTHENRY [30].

The state-of-the-art for tetrahedral grids is surveyed in Chapter 7.

In Chapter 8 we present the novel Mimetic Volume Integral (MVI) method. It is based on reconstruction operators defined with geometric elements the barycentric dual grid that can be interpreted as element-wise constant face basis functions They can be thought as the generalization of RWG and RT basis function for hexahedral or even general polyhedral elements because they produce the same stiffness matrix as the RT and RWG in case of tetrahedral meshes. We compare the novel mimetic integral method with respect to previous methods proposed in literature and we show that they produce the same solution in terms of current density in the case of tetrahedral meshes. We also point out how our method should be preferred since it allows a clear interpretation in terms of electric circuits whereas this is not the case with standard approaches where we show that mass matrices produced with RT or RWG face basis functions for hexahedra are *not* consistent. Hence, such mass matrices cannot be generalized to arbitrary polyhedral elements.

Finally, in 9, we present two groundbreaking advantages induced by the use of the novel MVI method. First, it enables a faster computation of the dense integral matrices and, in addition, the use of a faster singularity extraction technique to compute their diagonal terms accurately. Second, it enables an original factorization of these dense integral matrices. The benefits of this new factorization when coupled either with low-rank compression techniques or with black box implementations of the Fast Multipole Method [36] are exposed and compared critically to the state-of-the-art.

2

Mimetic numerical schemes

In this chapter we introduce the main ingredients which lie at the foundations of any mimetic numerical schemes. Specialized definitions are given for the low-order case.

2.1 Geometry of primal and dual grid

The domain of interest of this thesis is a closed and bounded polyhedral domain Ω of $\mathbb{R}^3$ with Lipschitz boundary. We assume that Ω has trivial topology, i.e., it is homeomorphic to a closed 3-dimensional ball or, equivalently, Ω is simply connected and its boundary $\partial\Omega$ is connected ($\partial\Omega$ is homeomorphic to a 2-sphere indeed); see [37] (Section 6) and [38] (Section 3).

We consider a *polyhedral cell complex* (or *polyhedral mesh*) subdivision K of Ω. Elements of K are called *cells*. A k-cell σ is a k-dimensional subset in $\mathbb{R}^3$ homeomorphic to a closed k-dimensional ball. A 0-cell is a point of $\mathbb{R}^3$. Crucially, k-cells as defined above can be *curved*. However, the standard MFD requires k-cells to be *planar* (or *flat*). A k-cell in K is planar if it is contained in a k-dimensional hyperplane of $\mathbb{R}^3$. We equip each k-cell in K with an inner orientation.

The grid K has the structure of a *(regular) cell complex*, namely, the following three conditions hold [39]:

1. For each k-cell σ in K its *boundary* $\partial\sigma$ is a union of $(k-1)$-cells in K for $k \in \{1, 2, 3\}$ [39].

2. Given distinct k-cells σ, τ, their intersection $\sigma \cap \tau$ is either empty or is a union of lower dimensional cells in K.

3. Given a l-cell σ and k-cell τ with $l \leq k$, $\sigma \neq \tau$ and $\sigma \cap \tau \neq \emptyset$, we have $\sigma \cap \tau \subset \partial\tau$.

We focus on the 3-dimensional case, thus we have 3-cells (elements), 2-cells (faces), 1-cells (edges) and 0-cells (nodes or vertices). We denote the cell complex K as $K = (N, E, F, C)$, where N is the set of nodes, E is the set of edges, F is the set of faces and C is the set of elements of K. We denote by c a generic element, by f a face, by e an edge and by v a node.

We denote by $|\cdot|$ either the cardinality of a set or the measure of a cell in K. For instance, $|F|$ is the number of faces in K, $|f|$ is the area of face $f \in F$ and $|c|$ is the volume of element $c \in C$.

A polyhedral cell complex K is *simplicial* if all its cells are simplicies and the boundary of each cell has the natural simplicial decomposition [39]. If K is simplicial, K is also called a triangulation of Ω.

Let X be any set among N, E, F, or C. If σ is a cell of K, we denote by $X(\sigma)$ the subset defined by

$$X(\sigma) \coloneqq \{\tau \in X \mid \sigma \subset \tau\}, \tag{2.1}$$

if (8.9) is not void, or otherwise,

$$X(\sigma) \coloneqq \{\tau \in X \mid \tau \subset \sigma\}. \tag{2.2}$$

As an instance instance, $C(e) = \{c \in C \mid e \subset c\}$ is the set of cells of C containing the edge e and $E(c) = \{e \in E \mid e \subset c\}$ is the set of edges of $c \in C$.

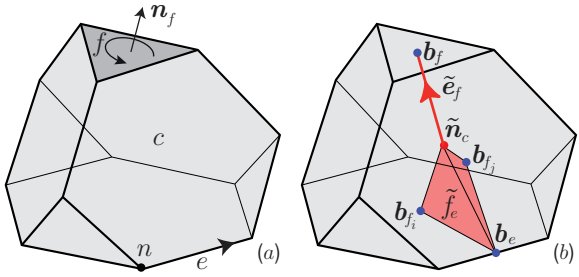

Figure 2.1: A polyhedral element $c \in C$ of K. (a) Primal and (b) dual geometric elements of c.

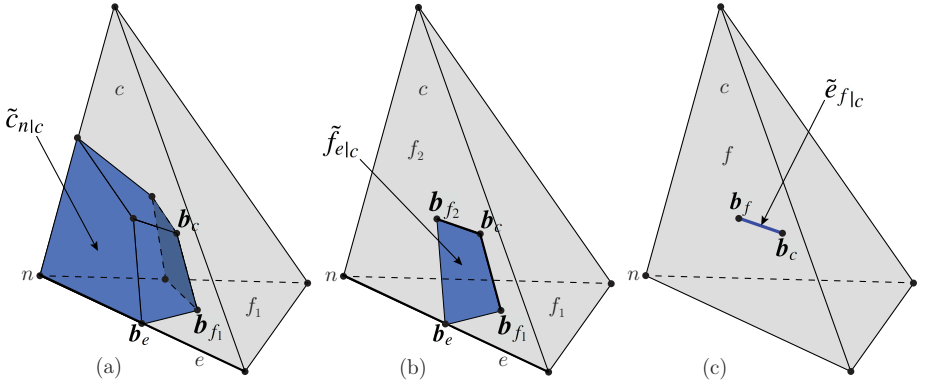

Figure 2.2: A tetrahedron $c \in C$ of K. (a) Dual cell $\tilde{c}_{n|c}$. (b) Dual face $\tilde{f}_{e|c}$. (c) Dual edge $\tilde{e}_{f|c}$.

Interlocked with the primal grid K, a *barycentric dual grid* [4], [1], [3] $\tilde{K} = (\tilde{N}, \tilde{E}, \tilde{F}, \tilde{C})$ is introduced, where the sets $\tilde{N}, \tilde{E}, \tilde{F}$ and $\tilde{C}$ contain *dual nodes*, *dual edges*, *dual faces* and *dual cells*, respectively. Each geometric entity of the dual grid is in one-to-one correspondence (so-called *duality pairing*) with a geometric element of the primal grid and it is constructed by means of the *barycentric subdivision* [4] of the primal grid, see Fig. 2.2. With symbol "$\sim$" we denote geometric elements of the dual grid, thus, we denote a single dual node as $\tilde{n}$, a dual edge by $\tilde{e}$, a dual face by $\tilde{f}$ and a dual cell by $\tilde{c}$, respectively. With a subscript we indicate the corresponding (unique) geometric element of the primal grid, thus, the dual of a primal node n is a dual cell $\tilde{c}_n$, the dual of a primal edge e is a dual face $\tilde{f}_e$, the dual of a primal face f is a dual edge $\tilde{e}_f$ and the dual of a primal cell c is the dual node denoted as $\tilde{n}_c$.

To describe the geometric elements of the pair of grids, we introduce a Cartesian system of coordinates with specified origin and we denote by $\boldsymbol{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ the coordinates of its generic point. Let us first describe the construction of the *restriction of dual geometric elements to an element* c, as the one pictured in Fig. 2.2. The geometric construction of the dual nodes, dual edges, dual faces and dual cells is based on the barycentric subdivision as follows. To begin with, let us define the barycenters of cells in K. The barycenters of an edge e, face f and of an element c are defined, respectively, as follows

$$\boldsymbol{b}_e = \frac{1}{|e|}\int_e \boldsymbol{x}\, de\,, \qquad \boldsymbol{b}_f = \frac{1}{|f|}\int_f \boldsymbol{x}\, df\,, \qquad \boldsymbol{b}_c = \frac{1}{|c|}\int_c \boldsymbol{x}\, dc\,, \tag{2.3}$$

where $|e|, |f|$ and $|c|$ denote the length of edge e, the area of face f and the volume of element c, respectively.

In what follows, we will employ a generalized construction of the barycentric dual grid in which the dual node $\tilde{\boldsymbol{n}}_c$ is an arbitrary point in $\mathbb{R}^3$ and does not necessarily coincide with the barycenter $\boldsymbol{b}_c$ of the cell c. We refer to such a construction again as barycentric dual grid, although the dual node does not coincide with the barycenter $\boldsymbol{b}_c$ of the element c. A dual edge $\tilde{e}_{f|c}$ is a segment which joins $\tilde{\boldsymbol{n}}_c$ with the barycenter $\boldsymbol{b}_f$ of a primal face f. A dual face $\tilde{f}_{e|c}$ is a quadrilateral plane surface whose vertices are $\tilde{\boldsymbol{n}}$, the barycenters of the pair of primal faces having edge e in common and the barycenter $\boldsymbol{b}_e$ of e. A dual cell $\tilde{c}_{n|c}$ is a region whose boundary is made of a disjoint union of dual faces, corresponding to edges of c having the node n in common, and portions of faces of c having the node n in common. Dual edge $\tilde{e}_f$, dual face $\tilde{f}_e$ and dual cell $\tilde{c}_n$ are endowed with outer orientation [1], [3], in such a way that each of the pairs $(e, \tilde{f}_e), (f, \tilde{e}_f), (n, \tilde{c}_n)$ are oriented according to the right-hand rule.

To each of the following geometric elements $e, f, \tilde{f}_{|c}, \tilde{e}_{|c}$ of the primal or of the dual grid, we associate their corresponding vectors $\boldsymbol{e}, \boldsymbol{f}, \boldsymbol{f}_{|c}, \tilde{\boldsymbol{e}}_{|c}$ respectively. Each of these vectors, will be represented with a column vector of its cartesian components. The *edge vector* $\boldsymbol{e}$ is

$$\boldsymbol{e} := \int_e \boldsymbol{t}_e\, de, \tag{2.4}$$

where $\boldsymbol{t}_e$ is the unit vector tangent to edge e at any given point. Next, the *face vector* $\boldsymbol{f}$ is

$$\boldsymbol{f} := \int_f \boldsymbol{n}_f\, df \tag{2.5}$$

where $\boldsymbol{n}_f$ is the unit vector perpendicular to face f at any given point. In a similar way, vector $\tilde{\boldsymbol{e}}_{f|c}$ is the edge vector associated with the dual edge $\tilde{e}_{f|c}$; for instance,

$$\tilde{\boldsymbol{e}}_{f|c} \coloneqq \boldsymbol{b}_f - \tilde{\boldsymbol{n}}_c, \tag{2.6}$$

see Fig. 2.2b. Vector $\tilde{\boldsymbol{f}}_{e|c}$ is the face vector associated with the dual face $\tilde{f}_{e|c}$. Face vector $\tilde{\boldsymbol{f}}_{e|c}$ is equal to

$$\tilde{\boldsymbol{f}}_{e|c} \coloneqq \frac{1}{2}\big(\tilde{\boldsymbol{e}}_{f_i|c} \times (\boldsymbol{b}_e - \tilde{\boldsymbol{n}}_c) - \tilde{\boldsymbol{e}}_{f_j|c} \times (\boldsymbol{b}_e - \tilde{\boldsymbol{n}}_c)\big), \tag{2.7}$$

where f_i, f_j are the unique two faces such that $e = f_i \cap f_j$, with $i = i(e)$ and $j = j(e)$, and in such a way that the expression induces the correct orientation on $\tilde{\boldsymbol{f}}_{e|c}$, see Fig. 2.2b.

Now, starting from the definition of dual geometric elements restricted to a single element c, we introduce dual edges, dual faces and dual cells of the barycentric dual grid. We define a dual edge $\tilde{e}_f$ as the piecewise segment made of the union of the two segments $\tilde{e}_{f|c_1}$ and $\tilde{e}_{f|c_2}$, with $c_1, c_2 \in C(f)$. The corresponding vector $\tilde{\boldsymbol{e}}_f$ is defined as follows

$$\tilde{\boldsymbol{e}}_f \coloneqq \sum_{c \in C(f)} \tilde{\boldsymbol{e}}_{f|c}. \tag{2.8}$$

A dual face $\tilde{f}_e$ is defined as the polyhedral surface made of the union of all $\tilde{f}_{e|c}$, with $c \in C(e)$. The corresponding dual face vector $\tilde{\boldsymbol{f}}_e$ is defined as follows

$$\tilde{\boldsymbol{f}}_e \coloneqq \sum_{c \in C(e)} \tilde{\boldsymbol{f}}_{e|c}. \tag{2.9}$$

A dual cell $\tilde{c}_n$ is defined as the region made of the union of all $\tilde{c}_{n|c}$, with $c \in C(n)$, that is $\tilde{c}_n \coloneqq \bigcup_{c \in C(n)} \tilde{c}_{n|c}$. Note that the boundary of each dual cell decomposes into dual edges and dual faces.

Let X be any set among E, F and let $\tilde{X}$ be any set among $\tilde{E}, \tilde{F}$. Given a non-empty region $\Omega' \subset \Omega$ of $\mathbb{R}^3$, we define

$$X_{\Omega'} \coloneqq \{x \cap \Omega', x \in X\}, \qquad \tilde{X}_{\Omega'} \coloneqq \{\tilde{x} \cap \Omega', \tilde{x} \in X\}, \tag{2.10}$$

as the subsets of geometric elements of X and $\tilde{X}$ restricted to Ω', respectively. In the sequel we consider the cases where Ω' is one among a cell $c \in C$, a dual cell $\tilde{c} \in \tilde{C}$, or the restriction of the dual cell $\tilde{c}$ to the cell c, $\tilde{c}_{|c}$. As an instance, $E_{\tilde{c}} \coloneqq \{e \cap \tilde{c}, e \in E\}$ contains the restriction of primal edges to the dual cell $\tilde{c}$. Observe that E_c, F_c coincide with the subsets of all edges and faces of c. Similarly, $\tilde{E}_{\tilde{c}}, \tilde{F}_{\tilde{c}}$ coincide with the subsets of all dual edges and dual faces of $\tilde{c}$.

Finally, we introduce the corresponding matrices $\mathbb{X}_{\Omega'}, \tilde{\mathbb{X}}_{\Omega'}$ whose rows collect geometric vectors associated with elements of the sets in (2.10). As an instance, rows of $\mathbb{E}_{\tilde{c}}$ contain edge vectors associated with edges in $E_{\tilde{c}}$.

2.2 Degrees of Freedom

We define a *discrete field* as a collection of degrees of freedom (DoFs). DoFs are arrays of real numbers whose entries are obtained by evaluating the physical scalar or vector fields on the geometric elements of the mesh by means of integration [41], [1], [7]. The operation of translating a sufficiently smooth scalar or vector-valued functions into DoFs are performed by *projection maps*, called also *de Rham maps* [2], [7].

Let u be a sufficiently regular scalar function so that we can take its pointwise values. We denote with $S^{\mathcal{N}}$ the space of such functions. For instance, $S^{\mathcal{N}}$ can be taken to be $H^1(\Omega) \cap C^0(\Omega)$. Node projection operator $P^{\mathcal{N}}$ maps u onto its node DoFs $u^{\mathcal{N}} = P^{\mathcal{N}}(u)$, where the entry of $u^{\mathcal{N}}$ corresponding to node n is

$$u_n^{\mathcal{N}} = u(n). \tag{2.11}$$

We denote the vector space of node DoFs by $\mathcal{N} = \mathbb{R}^{|N|}$.

Let $\boldsymbol{u}$ be a sufficiently regular vector-valued function so that the integrals of its tangential component are well defined along the grid edges. We denote with $S^{\mathcal{E}}$ the space of all such functions. For instance, $S^{\mathcal{E}}$ can be taken to be $H^s(\Omega)$, with $s > 1$. Edge projection operator $P^{\mathcal{E}}$ maps $\boldsymbol{u}$ onto its edge DoFs $\boldsymbol{u}^{\mathcal{E}} = P^{\mathcal{E}}(\boldsymbol{u})$, where the entry of $\boldsymbol{u}^{\mathcal{E}}$ corresponding to edge e is

$$\boldsymbol{u}_e^{\mathcal{E}} = \int_e \boldsymbol{u} \cdot \boldsymbol{t}_e \, de. \tag{2.12}$$

We denote the vector space of edge DoFs by $\mathcal{E} = \mathbb{R}^{|E|}$.

Now, let $\boldsymbol{u}$ be a sufficiently regular vector-valued function so that the integrals of its normal component are well defined on the grid faces. We denote with $S^{\mathcal{F}}$ the space of all such functions. For instance, $S^{\mathcal{F}}$ can be taken to be $H^s(\Omega)$, with $s > \frac{1}{2}$. Face projection operator $P^{\mathcal{F}}$ maps $\boldsymbol{u}$ onto its face DoFs $\boldsymbol{u}^{\mathcal{F}} = P^{\mathcal{F}}(\boldsymbol{u})$, where the entry of $\boldsymbol{u}^{\mathcal{F}}$ corresponding to face f is

$$\boldsymbol{u}_f^{\mathcal{F}} = \int_f \boldsymbol{u} \cdot \boldsymbol{n}_f \, df. \tag{2.13}$$

We denote the vector space of face DoFs by $\mathcal{F} = \mathbb{R}^{|F|}$.

Let u be a sufficiently regular scalar function so that we can take its integrals on compact subsets of Ω are well-defined. We denote with $S^{\mathcal{C}}$ the space of such functions. For instance, $S^{\mathcal{C}}$ can be taken to be $L^1(\Omega)$. Cell projection operator $P^{\mathcal{C}}$ maps u onto its node DoFs $u^{\mathcal{N}} = P^{\mathcal{C}}(u)$, where the entry of $n^{\mathcal{C}}$ corresponding to cell c is

$$u_c^{\mathcal{C}} = \int_c u \, dc. \tag{2.14}$$

We denote the vector space of node DoFs by $c = \mathbb{R}^{|C|}$.

Let $\mathcal{X}$ be any set among $\mathcal{N}, \mathcal{E}, \mathcal{F}$ or $\mathcal{C}$. For a given element c, we denote with $\mathcal{X}_c$ the vector subspace of all DoFs of $\mathcal{X}$ that are attached to cells of c. Then, the *local projection operator* on c is defined as $P_c^{\mathcal{X}} : S_c^{\mathcal{X}} \to \mathcal{X}_c$, where $S_c^{\mathcal{X}}$ is finite dimensional vector subspace of $S^{\mathcal{X}}$ chosen in such a way that the map $P_c^{\mathcal{X}}$ is bijective on $\mathcal{X}_c$. Thus, we can set up a correspondence between a vector field $\boldsymbol{u} \in S_c^{\mathcal{X}}$ and its corresponding

DoFs. We point out that in the following we do not need to construct explicitly the space $S_c^{\mathcal{X}}$ since only its generic properties will be used.

Let $\mathbb{P}_c^{\mathcal{X}}$ be the matrix associated with the restriction of the map $P_c^{\mathcal{X}} : S_c^{\mathcal{X}} \to \mathcal{X}_c$ to the vector subspace of *constant* scalar or vector fields. Specifically, rows of $\mathbb{P}_c^{\mathcal{E}}, \mathbb{P}_c^{\mathcal{F}}$ collect the coefficients with respect to the standard basis of $\mathbb{R}^3$ of the quantities $\int_e \boldsymbol{t}_e \, de$ and $\int_f \boldsymbol{n}_f \, df$, respectively. Given a constant vector field $\boldsymbol{u}$ defined on c, by Stokes theorem, the latter integral quantities can also be computed on any homologous geometric element. In particular, matrices $\mathbb{P}_c^{\mathcal{E}}$ and $\mathbb{P}_c^{\mathcal{F}}$ can be written as

$$\mathbb{P}_c^{\mathcal{F}} = \mathbb{F}_c = \begin{pmatrix} \vdots \\ \boldsymbol{f}^T \\ \vdots \end{pmatrix}, \qquad \mathbb{P}_c^{\mathcal{E}} = \mathbb{E}_c = \begin{pmatrix} \vdots \\ \boldsymbol{e}^T \\ \vdots \end{pmatrix}. \tag{2.15}$$

In the sequel we will also need the dual counterpart of matrices $\mathbb{P}_c^{\mathcal{E}}, \mathbb{P}_c^{\mathcal{F}}$. We denote by $\mathbb{P}_{\tilde{\boldsymbol{n}}_c}^{\widetilde{\mathcal{E}}}$ the matrix whose rows collect edge vectors associated with the restriction of dual edges to c constructed using the dual node $\tilde{\boldsymbol{n}}_c$. For the special case where $\tilde{\boldsymbol{n}}_c$ coincide with the origin, we use the symbol $\mathbb{P}_c^{\widetilde{\mathcal{E}}}$. In a similar way, we denote with $\mathbb{P}_{\tilde{\boldsymbol{n}}_c}^{\widetilde{\mathcal{F}}}$ the matrix whose rows collect dual face vectors associated with the restriction of dual faces to c, constructed using the dual node $\tilde{\boldsymbol{n}}_c$. For the special case where $\tilde{\boldsymbol{n}}_c$ coincide with the origin, we use the symbol $\mathbb{P}_c^{\widetilde{\mathcal{F}}}$.

2.3 Discrete differential operators

One kind of equation that we find in physical theories are the *balance equations* [1], [3]. A differential operator usually arises in the continuous form of balance equations and thus, in numerical schemes we need its discrete counterpart. Stokes theorem provides discrete counterparts of the differential operators gradient, curl and divergence [41], [1], [5]. Discrete differential operators can be defined for both the primal and the dual grid and they are defined in matrix form by incidence matrices of the pair of grids [1]. An entry of these matrices corresponding to a given pair of oriented geometric elements of the primal or dual grid is equal to 0, if the two geometric elements are mutually incident, otherwise, 1 if their mutual orientations are compatible and -1 if not [3]. From their definition discrete differential operators encode the topology of problem: being metric free, any stretching or deformation of the mesh does not change their form.

The *discrete gradient operator* maps node DoFs in $\mathcal{N}$ to edge DoFs in $\mathcal{E}$, and is defined through the action of a matrix

$$\mathcal{GRAD}(\boldsymbol{u}^{\mathcal{N}}) \coloneqq \mathbb{G}\, \boldsymbol{u}^{\mathcal{N}}, \tag{2.16}$$

where $\mathbb{G}$ is the $|E| \times |N|$-matrix of incidence numbers between nodes and edges.

The *discrete curl operator* maps edge DoFs in $\mathcal{E}$ to face DoFs in $\mathcal{F}$, and is defined through the action of a matrix

$$\mathcal{CURL}(\boldsymbol{u}^{\mathcal{E}}) \coloneqq \mathbb{C}\, \boldsymbol{u}^{\mathcal{E}}, \tag{2.17}$$

where $\mathbb{C}$ is the $|F| \times |E|$-matrix of incidence numbers between edges and faces.

The *discrete divergence operator* maps face DoFs in $\mathcal{F}$ to cell DoFs in $\mathcal{C}$, and is defined through the action of a matrix

$$\mathcal{DIV}(\boldsymbol{u}^{\mathcal{F}}) \coloneqq \mathbb{D}\,\boldsymbol{u}^{\mathcal{F}}, \tag{2.18}$$

where $\mathbb{D}$ is the $|C| \times |F|$-matrix of incidence numbers between faces and cells.

2.4 Reconstruction operators

Reconstruction operators are designed to remap DoFs into the corresponding continuous vector field. In contrast to the projection operator, where the de Rham map is the obvious candidate, the choice of the reconstruction operator is flexible because there are many possible ways in which global physical quantities can be combined to a local field representation.

Let $R_c^{\mathcal{X}} : \mathcal{X}_c \to S_c^{\mathcal{X}}$ be the reconstruction operator from the vector space $\mathcal{X}_c$ of DoFs restricted to element c. In [8], reconstruction operators are required to satisfy a number of formal properties that involve the projection operator and the continuous and discrete differential operators. However, for the low-order case, from this set of properties it is sufficient to require that reconstruction operators satisfy only the so called *accuracy property* [8]: $R_c^{\mathcal{X}}$ is a left inverse of $P_c^{\mathcal{X}}$, i.e.

$$(R_c^{\mathcal{X}} \circ P_c^{\mathcal{X}})(\boldsymbol{u}) = \boldsymbol{u},\ \forall \boldsymbol{u} \in S_c^{\mathcal{X}}. \tag{2.19}$$

As we will see, since we are interested in the low-order case, we never need an explicit expression for the reconstruction operator, but only its averaged quantities over a cell come into play. To this end in mind, let us introduce the *average reconstruction operator* $\overline{R_c^{\mathcal{X}}} : \mathcal{X}_c \to S_c^{\mathcal{X}}$ as a map whose values are constant vector fields in c, defined as follows

$$\overline{R_c^{\mathcal{X}}}(\boldsymbol{u}_c^{\mathcal{X}}) \coloneqq \frac{1}{|c|}\int_c R_c^{\mathcal{X}}(\boldsymbol{u}_c^{\mathcal{X}})\,dc,\ \forall \boldsymbol{u}_c^{\mathcal{X}} \in \mathcal{X}_c. \tag{2.20}$$

Let $\mathbb{R}_c^{\mathcal{X}}$ be the matrix associated with the linear map $\overline{R_c^{\mathcal{X}}}$. The accuracy property requires that $\mathbb{R}_c^{\mathcal{X}}$ is a left inverse of $\mathbb{P}_c^{\mathcal{X}}$. Indeed, we have $\overline{R_c^{\mathcal{X}}}(\mathrm{im}(\mathbb{P}_c^{\mathcal{X}})) = R_c^{\mathcal{X}}(\mathrm{im}(\mathbb{P}_c^{\mathcal{X}}))$ and $\mathrm{im}(\mathbb{P}_c^{\mathcal{X}}) \subset \mathcal{X}_c$.

An additional property that reconstruction operators have to satisfy is the so called P_0-*consistency* and it will be formally defined in Section 3.1. Informally, these are the reconstruction operators that can be used to construct a global mass matrix, according to the recipe detailed in Section 2.5 and Section 2.5.1, that satisfies the P_0-consistency property.

2.5 Mimetic inner products: Mass matrices

Let us consider a symmetric, positive-definite 3×3 matrix $\mathbb{K}_c$ representing an homogeneous material property of an element c. We want to introduce an inner product on vector spaces $\mathcal{N}, \mathcal{E}, \mathcal{F}$ and $\mathcal{C}$ that is a low-order approximation of classical L^2 product

between two continuous vector fields $\boldsymbol{u}, \boldsymbol{v}$

$$\langle \boldsymbol{u}_c^{\mathcal{X}}, \boldsymbol{v}_c^{\mathcal{X}} \rangle_{\mathcal{X},c} = \int_c \boldsymbol{u} \cdot (\mathbb{K}_c \boldsymbol{v}) dV + \mathcal{O}(h|c|), \; \forall \boldsymbol{u}, \boldsymbol{v} \in S_c^{\mathcal{X}}, \tag{2.21}$$

where h denotes a characteristic size of the mesh and $\boldsymbol{u}_c^{\mathcal{X}}, \boldsymbol{v}_c^{\mathcal{X}}$ are the DoFs restricted to a cell c of $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively. The bilinear form can be expressed in matrix form through a symmetric and positive-definite matrix $\mathbb{M}_c^{\mathcal{X}}$ which acts on the local DoFs

$$\langle \boldsymbol{u}_c^{\mathcal{X}}, \boldsymbol{v}_c^{\mathcal{X}} \rangle_{\mathcal{X},c} = (\boldsymbol{u}_c^{\mathcal{X}})^T \mathbb{M}_c^{\mathcal{X}} \boldsymbol{v}_c^{\mathcal{X}}. \tag{2.22}$$

In view of (2.21), the local matrix $\mathbb{M}_c^{\mathcal{X}}$ must contain information about the material property $\mathbb{K}_c$ on c.

We want to design $\mathbb{M}_c^{\mathcal{X}}$ in such a way that it satisfies the following three conditions [7], [17], [42], [8]:

1. Symmetric.
2. Positive-definite.
3. *Consistency condition* [8]: whenever $\boldsymbol{u}$ is a constant vector field and $\boldsymbol{v} \in S_c^{\mathcal{X}}$, the following equality holds for every element c

$$(\boldsymbol{u}_c^{\mathcal{X}})^T \mathbb{M}_c^{\mathcal{X}} \boldsymbol{v}_c^{\mathcal{X}} = \int_c \boldsymbol{u} \cdot (\mathbb{K}_c \boldsymbol{v}) \, dc, \; \forall \boldsymbol{v} \in S_c^{\mathcal{X}}. \tag{2.23}$$

Suppose that we want to satisfy the three requirements above. Under the above assumptions, we now introduce an equivalent algebraic form of the consistency condition (2.23).

Let $\boldsymbol{u}$ be a constant vector field, and let $\boldsymbol{v}$ be an element of $S_c^{\mathcal{X}}$. We can rewrite in an equivalent way the consistency condition (2.23) as

$$(P_c^{\mathcal{X}}(\boldsymbol{u}))^T \mathbb{M}_c^{\mathcal{X}} \boldsymbol{v}_c^{\mathcal{X}} = \int_c \boldsymbol{u} \cdot (\mathbb{K}_c \, R_c^{\mathcal{X}}(\boldsymbol{v}_c^{\mathcal{X}})) \, dc, \; \forall \boldsymbol{v}_c^{\mathcal{X}} \in \mathcal{X}_c. \tag{2.24}$$

Since $\boldsymbol{u}$ is constant inside c, we introduce the matrix form of the projection operator $\mathbb{P}_c^{\mathcal{X}}$. In addition, since both $\boldsymbol{u}$ and $\mathbb{K}$ are constant inside the cell c, by using the definition of the matrix $\mathbb{R}_c^{\mathcal{X}}$ associated with the average reconstruction operator $\overline{R_c^{\mathcal{X}}}$ on the cell c, it follows that

$$\boldsymbol{u}^T (\mathbb{P}_c^{\mathcal{X}})^T \mathbb{M}_c^{\mathcal{X}} \boldsymbol{v}_c^{\mathcal{X}} = |c| \, \boldsymbol{u} \cdot (\mathbb{K}_c \, \mathbb{R}_c^{\mathcal{X}} \boldsymbol{v}_c^{\mathcal{X}}), \; \forall \boldsymbol{v}_c^{\mathcal{X}} \in \mathcal{X}_c. \tag{2.25}$$

The above equality must hold for every vector $\boldsymbol{u}$ and DoF $\boldsymbol{v}_c^{\mathcal{X}}$, thus we obtain the following condition

$$(\mathbb{P}_c^{\mathcal{X}})^T \mathbb{M}_c^{\mathcal{X}} = |c| \, \mathbb{K}_c \, \mathbb{R}_c^{\mathcal{X}}. \tag{2.26}$$

By transposing both members we can finally write

$$\mathbb{M}_c^{\mathcal{X}} \mathbb{P}_c^{\mathcal{X}} = |c| \, (\mathbb{R}_c^{\mathcal{X}})^T \, \mathbb{K}_c. \tag{2.27}$$

Condition expressed in (2.27) is called *algebraic consistency condition* in mimetic literature [8]. The following theorem show that a solution of (2.27) can always be written

as the sum of two terms, which is the canonical solution proposed in [8].

Theorem 1. *Let $\mathbb{K}_c$ be a symmetric and positive-definite matrix, and let $\mathbb{R}_c^{\mathcal{X}}$ be a left inverse of $\mathbb{P}_c^{\mathcal{X}}$. Let $\alpha = (\alpha_1, \ldots, \alpha_{n-3}) \in (\mathbb{R}^+)^{n-3}$ be any $(n-3)$-upla of positive real numbers and let $\mathbb{D}_\alpha$ be the diagonal matrix whose diagonal entries are $\alpha_1, \ldots, \alpha_{n-3}$. Let $w = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{n-3})$ be any orthonormal basis of $\mathrm{im}(\mathbb{P}_c^{\mathcal{X}})^\perp$ and let $\mathbb{W}_c$ be the matrix whose columns are the vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{n-3}$. We define the consistent matrix $\mathbb{C}_c$ and stabilization matrix $\mathbb{S}_c$ by setting*

$$\mathbb{C}_c := |c|(\mathbb{R}_c^{\mathcal{X}})^T \mathbb{K}_c \, \mathbb{R}_c^{\mathcal{X}} \tag{2.28}$$

and

$$\mathbb{S}_c := \mathbb{W}_c \mathbb{D}_\alpha \mathbb{W}_c^T. \tag{2.29}$$

.

Then the matrix $\mathbb{M}_c^{\mathcal{X}}$ defined as

$$\mathbb{M}_c^{\mathcal{X}} = \mathbb{C}_c + \mathbb{S}_c \tag{2.30}$$

. is symmetric, consistent and positive definite. We call $\mathbb{C}_c$ the consistent term *of $\mathbb{M}_c^{\mathcal{X}}$, and $\mathbb{S}_c$ the* stabilization term *of $\mathbb{M}_c^{\mathcal{X}}$.*

Proof. It is clear that matrix $\mathbb{M}_c^{\mathcal{X}}$ is symmetric, positive-semidefinite and by direct substitution it satisfies (2.27). Let $\boldsymbol{z} \in \mathbb{R}^{|\mathcal{X}|}$ be such that $\boldsymbol{z}^T \mathbb{M}_c^{\mathcal{X}} \boldsymbol{z} = 0$. In order to prove that $\mathbb{M}$ is positive-definite, we have to show that $\boldsymbol{z} = \boldsymbol{0}$. The condition $\boldsymbol{z}^T \mathbb{M}_c^{\mathcal{X}} \boldsymbol{z} = 0$ is equivalent to require that $(\mathbb{R}_c^{\mathcal{X}} \boldsymbol{z})^T \mathbb{K}_c \, \mathbb{R}_c^{\mathcal{X}} \boldsymbol{z} = 0$ and $(\mathbb{W}_c^T \boldsymbol{z})^T \mathbb{D}_\alpha (\mathbb{W}_c^T \boldsymbol{z}) = 0$. Since $\mathbb{K}_c$ is symmetric and definite-positive, and each α_i is positive, the latter condition is in turn equivalent to $\mathbb{R}_c^{\mathcal{X}} \boldsymbol{z} = \boldsymbol{0}$ and $\boldsymbol{z} \in \mathrm{im}(\mathbb{P}_c^{\mathcal{X}})$. As a consequence, $\boldsymbol{z} = \mathbb{P}_c^{\mathcal{X}} \boldsymbol{y}$ for some $\boldsymbol{y} \in \mathbb{R}^3$ and $\boldsymbol{0} = \mathbb{R}_c^{\mathcal{X}} \boldsymbol{z} = \mathbb{R}_c^{\mathcal{X}} \mathbb{P}_c^{\mathcal{X}} \boldsymbol{y} = \boldsymbol{y}$. Thus $\boldsymbol{z} = \mathbb{P}_c^{\mathcal{X}} \boldsymbol{y} = \boldsymbol{0}$, as desired. □

2.5.1 Global mass matrix

For each $c \in C$, let $\mathbb{O}_c^{\mathcal{X}}$ be the matrix which assigns DoFs of c when it is applied to an element of $\mathcal{X}$. $\mathbb{O}_c^{\mathcal{X}}$ is a matrix of size $|\mathcal{X}_c| \times |\mathcal{X}|$ and every row has exactly one entry equal to 1 corresponding to a element which is in c and zero everywhere else. For each $c \in C$, let $\mathbb{M}_c^{\mathcal{X}}$ be a local mass matrix associated with c, where $\mathbb{M}_c^{\mathcal{X}}$ is defined as in Theorem 1. Using this definition, we obtain

$$(\boldsymbol{u}^{\mathcal{X}})^T \mathbb{M}^{\mathcal{X}} \boldsymbol{v}^{\mathcal{X}} = \sum_{c \in C} \left(\mathbb{O}_c^{\mathcal{X}} \boldsymbol{u}^{\mathcal{X}}\right)^T \mathbb{M}_c^{\mathcal{X}} \left(\mathbb{O}_c^{\mathcal{X}} \boldsymbol{v}^{\mathcal{X}}\right) = (\boldsymbol{u}^{\mathcal{X}})^T \Big(\sum_{c \in C} (\mathbb{O}_c^{\mathcal{X}})^T \mathbb{M}_c^{\mathcal{X}} \mathbb{O}_c^{\mathcal{X}} \Big) \boldsymbol{v}^{\mathcal{X}}, \tag{2.31}$$

from which follows the expression for the global mass matrix

$$\mathbb{M}^{\mathcal{X}} = \sum_{c \in C} (\mathbb{O}_c^{\mathcal{X}})^T \mathbb{M}_c^{\mathcal{X}} \mathbb{O}_c^{\mathcal{X}}. \tag{2.32}$$

2.6 Derived discrete operators

In this subsection we recall the definition of the derived discrete operators $\widetilde{\mathcal{GRAD}}$, $\widetilde{\mathcal{CURL}}$ and $\widetilde{\mathcal{DIV}}$ which are obtained through a duality relation from the discrete operators $\mathcal{GRAD}$, $\mathcal{CURL}$ and $\mathcal{DIV}$, respectively. Let us suppose that the discrete spaces $\mathcal{N}$, $\mathcal{E}$, $\mathcal{F}$ and $\mathcal{C}$ are equipped with inner products as detailed in Section 2.5 and Section 2.5.1.

Let us now define the derived operators. The adjoints of the discrete differential operators $\mathcal{GRAD}$, $\mathcal{CURL}$ and $\mathcal{DIV}$ will be denoted by $\mathcal{GRAD}^*$, $\mathcal{CURL}^*$ and $\mathcal{DIV}^*$. It is convenient to rename the adjoints operators as follows $\widetilde{\mathcal{GRAD}} \cong -\mathcal{DIV}^*$, $\widetilde{\mathcal{CURL}} \cong \mathcal{CURL}^*$ and $\widetilde{\mathcal{DIV}} \cong -\mathcal{GRAD}^*$. By using the definition of $\mathcal{DIV}^*$ and the identification $\widetilde{\mathcal{GRAD}} \cong -\mathcal{DIV}^*$ we obtain

$$\langle \boldsymbol{u}^{\mathcal{F}}, \mathcal{GRAD}\, p^{\mathcal{C}} \rangle_{\mathcal{F}} = -\langle \boldsymbol{u}^{\mathcal{F}}, \mathcal{DIV}^*\, p^{\mathcal{C}} \rangle_{\mathcal{F}} = \langle \mathcal{DIV}\, \boldsymbol{u}^{\mathcal{F}}, p^{\mathcal{C}} \rangle_{\mathcal{C}}, \; \forall \boldsymbol{u}^{\mathcal{F}} \in \mathcal{F}, \; p^{\mathcal{C}} \in \mathcal{C}. \tag{2.33}$$

As $\boldsymbol{u}^{\mathcal{F}}$ and $p^{\mathcal{C}}$ are arbitrary it follows that

$$\widetilde{\mathcal{GRAD}} \cong -\mathcal{DIV}^* = -\mathbb{M}^{\mathcal{F}^{-1}} \mathcal{DIV}^T \mathbb{M}^{\mathcal{C}}. \tag{2.34}$$

By using the definition of $\mathcal{CURL}^*$ and the identification $\widetilde{\mathcal{CURL}} \cong -\mathcal{CURL}^*$ we obtain

$$\langle \boldsymbol{u}^{\mathcal{E}}, \widetilde{\mathcal{CURL}}\, \boldsymbol{w}^{\mathcal{F}} \rangle_{\mathcal{E}} = \langle \boldsymbol{u}^{\mathcal{E}}, \mathcal{CURL}^*\, \boldsymbol{w}^{\mathcal{F}} \rangle_{\mathcal{E}} = \langle \mathcal{CURL}\, \boldsymbol{u}^{\mathcal{E}}, \boldsymbol{w}^{\mathcal{F}} \rangle_{\mathcal{F}}, \; \forall \boldsymbol{u}^{\mathcal{E}} \in \mathcal{E}, \; \boldsymbol{w}^{\mathcal{F}} \in \mathcal{F} \tag{2.35}$$

form which we obtain that

$$\widetilde{\mathcal{CURL}} = \mathbb{M}^{\mathcal{E}^{-1}} \mathcal{CURL}^T \mathbb{M}^{\mathcal{F}}. \tag{2.36}$$

Similarly, the duality relation between the discrete gradient operator $\mathcal{GRAD}$ and its adjoint $\widetilde{\mathcal{DIV}} \cong \mathcal{GRAD}^*$ implies that

$$\langle q^{\mathcal{N}}, \widetilde{\mathcal{DIV}}\, \boldsymbol{w}^{\mathcal{E}} \rangle_{\mathcal{N}} = -\langle q^{\mathcal{N}}, \mathcal{GRAD}^*\, \boldsymbol{w}^{\mathcal{E}} \rangle_{\mathcal{N}} = -\langle \mathcal{GRAD}\, q^{\mathcal{N}}, \boldsymbol{w}^{\mathcal{E}} \rangle_{\mathcal{E}}, \; \forall q^{\mathcal{N}} \in \mathcal{N}, \; \boldsymbol{w}^{\mathcal{E}} \in \mathcal{E} \tag{2.37}$$

from which we obtain that

$$\widetilde{\mathcal{DIV}} = -\mathbb{M}^{\mathcal{N}^{-1}} \mathcal{GRAD}^T \mathbb{M}^{\mathcal{E}}. \tag{2.38}$$

2.7 Chains and cochains

We define a new object, called a real k-chain. A k-chain a of K is a formal linear combination of k-cells $a = \sum_{i=1}^{r} a_i \sigma_i$, where σ_i are k-cells in K and a_i are real coefficients. The number r denotes the cardinality of the collection of k-cells in $\mathcal{K}$ and is any number among $|N|, |E|, |F|$ or $|C|$. The set of k-chains, equipped with the natural addition and scalar multiplication, provides a real vector space. We denote it by $C_k(K)$. Note that each k-cell is also a k-chain. If σ is a k-cell, by $-\sigma$ we denote the cell σ but with opposite

orientation. The set of all k-cells form a basis for $C_k(K)$, which we call *canonical basis* for $C_k(K)$. We identify the boundary of each k-cell with the linear combination of the $(k-1)$-cells in $\partial\sigma$ defined by setting

$$\partial\sigma \coloneqq \sum_{i=1}^{r} w_i \, \rho_i, \tag{2.39}$$

where w_i is different from zero if and only if $\rho_i \subset \partial\sigma$ and in this case, w_i is equal to $+1$ if ρ_i has the orientation induced by that of σ by using the right-hand rule and -1 otherwise [43].

The real vector space of k-chains and the real vector space of $(k-1)$-chains are connected by a linear map called *boundary operator* $\partial_k : C_k(K) \to C_{k-1}(K)$. We define the boundary operator by linearity on the space of chains by setting

$$\partial_k a \coloneqq \sum_{i=1}^{r} a_i \, \partial\sigma_i, \tag{2.40}$$

where $a = \sum_{i=1}^{r} a_i \sigma_i$ as above. Note that (2.40) is well-defined since K is a cell complex. Since K is a cell complex, it can verified that $\partial_{k-1} \circ \partial_k = 0$ for $k \in \{1, 2, 3\}$; see, for example [39].

Let us now consider the concept of a *k-cochain*. A k-cochain b acts on a k-chain to produce a real number and therefore k-cochains are elements of the dual space of $C_k(K)$. We define the vector space of k-cochains $C^k(K)$ to be the dual space of linear functionals $b : C_k(K) \to \mathbb{R}$. We denote the value of a k-chain a under a k-cochain b as $\langle b\,, a\rangle \coloneqq b(a)$. Let us consider the canonical basis $\{\sigma_i \in C_k(K) \mid i = 1, \ldots, r\}$ of the vector space of k-chains $C_k(K)$. From basic linear algebra, there exist unique linear functionals $\{\sigma^i \in C^k(K) \mid i = 1, \ldots, r\}$ such that

$$\langle \sigma^i \,, \sigma_j \rangle = \delta_{ij}, \tag{2.41}$$

where δ_{ij} is the Kronecker delta. The set $\{\sigma^i \in C^k(K) \mid i = 1, \ldots, r\}$ defined by (2.41) form a basis for the vector space of k-cochains $C^k(K)$, which is called *canonical dual basis*. We have established a one-to-one correspondence between chains and cochains. This chain-cochain natural duality yields the real linear isomorphism $\phi_k : C_k(K) \to C^k(K)$ sending each σ_i to σ^i. We will write a generic k-cochain $b \in C^k(K)$ as a sum $b = \sum_{i=1}^{r} b_i \sigma^i$ with real coefficients b_i.

For k-cochains, in intimate analogy with chains, we can define a *coboundary operator* $\delta^k : C^k(K) \to C^{k+1}(K)$ as the *dual of the boundary operator*, i.e., it is defined by requiring that, for every $b \in C^k(K)$ and $a \in C_{k+1}(K)$, the following identity holds

$$\langle \delta^k b \,, a\rangle = \langle b \,, \partial_{k+1} a\rangle. \tag{2.42}$$

In mimetic methods the coboundary operator δ^k acts as a discrete counterpart of the continuous differential operators [44, 43]. Specifically, δ^0 acts as the discrete gradient, δ^1 as the discrete curl and δ^2 as the discrete divergence.

A straightforward calculation using (2.42) shows that $\delta^k \circ \delta^{k-1} = 0$ for $k \in \{1, 2\}$. These relations mimic the structure of continuous differential operators [44, 43]. In

particular, discrete differential operators form a chain complex

$$C^* = \cdots \xrightarrow{\delta^{k-1}} C^k(K) \xrightarrow{\delta^k} C^{k+1}(K) \xrightarrow{\delta^{k+1}} \cdots, \tag{2.43}$$

where $C^k(K) = 0$ if $k < 0$ or $k > 3$. Since the domain Ω is topologically trivial the sequence is *exact* for $k \neq 0$, i.e. it satisfies $\mathrm{im}(\delta^{k-1}) = \ker(\delta^k)$ for $k \neq 0$.

In the case of k-chains, there is a natural choice of a basis given by the canonical basis. Using the isomorphism $\phi_k : C_k(K) \to C^k(K)$, we have also fixed a canonical dual basis for $C^k(K)$. Since the coboundary operator δ^k is a linear map between $C^k(K)$ and $C^{k+1}(K)$, it can be represented, using the fixed bases of $C^k(K)$ and $C^{k+1}(K)$, as a matrix. Thus, to represent the coboundary operator as a matrix, we must always explicitly state which bases are chosen and, in fact, we will soon see the benefits of changing the bases.

Let us now consider an arbitrary basis for the vector space of k-chains $C_k(K)$. We denote it by $\mathcal{B}_k = \{\dots, \xi_i, \dots\}$. Using the isomorphism $\phi_k : C_k(K) \to C^k(K)$, $\phi_k(\mathcal{B}_k) = \{\dots, \xi^i = \phi_k(\xi_i), \dots\}$ is a basis for $C^k(K)$. In what follows, we take this process of choosing a basis for $C^k(K)$ for granted. When this is done, we say that we have chosen a basis $\mathcal{B} = \bigcup_k \mathcal{B}_k$ for the entire chain complex C^*, i.e. a basis for each $C^k(K)$. We write $(C^*, \mathcal{B})$ to denote a chain complex with a basis. We define the *canonical basis* $\widehat{\mathcal{B}}$ for C^* to be the basis $\widehat{\mathcal{B}} = \bigcup_k \widehat{\mathcal{B}}_k$ where each $\widehat{\mathcal{B}}_k$ is the canonical basis for $C_k(K)$.

Having chosen bases in $C_k(K)$ and hence in $C^k(K)$, we denote by $\mathbb{D}_k$ the matrix associated with δ^k for $k \in \{0, 1, 2\}$. Using the definitions of discrete differential operators in Section 2.3 we have that $\mathbb{G} = \mathbb{D}_0$, $\mathbb{C} = \mathbb{D}_1$ and $\mathbb{D} = \mathbb{D}_2$. We can also define $\mathbb{D}_3$ as the null operator from $C_3(K)$ to 0.

We represent k-chains and k-cochains by vectors of size r that contain the real numbers with respect to the ordered bases. A k-chain $a = \sum_{i=1}^r a_i \xi_i$ in a basis $\{\dots, \xi_i, \dots\}$ is represented by the column vector $\mathbb{R}^r \ni \boldsymbol{a} = (a_1 \cdots a_r)^T$ and a k-cochain $b = \sum_{i=1}^r b_i \xi^i$ in the basis $\{\dots, \xi^i = \phi_k(\xi_i), \dots\}$ is represented by the column vector $\mathbb{R}^r \ni \boldsymbol{b} = (b_1, \cdots b_r)^T$.

2.8 Mimetic discretization of stationary currrent conduction problem

The stationary current conduction is a Poisson problem in region Ω of the 3-D Euclidean space

$$\boldsymbol{\nabla} \times \boldsymbol{E} = \boldsymbol{0}, \tag{2.44a}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{J} = 0, \tag{2.44b}$$

$$\boldsymbol{J} = \sigma \boldsymbol{E}, \tag{2.44c}$$

where σ is the material parameter electric conductivity, $\boldsymbol{E}$ and $\boldsymbol{J}$ are the *conservative* electrostatic field and the current density vectors, respectively. The electric conductivity σ is assumed to be a positive scalar value which is piecewise uniform in each material region. The region boundary $\partial\Omega$ is partitioned into a set of N^{i} surfaces of perfect insulators $\partial\Omega^{\mathrm{i}}_k$, and a set of $N^{\mathrm{c}} + 1$ disjoint equipotential surfaces of perfect conductors

$\partial\Omega_k^{\mathrm{c}}$ (usually called electrodes):

$$\partial\Omega = \sum_{k=1}^{N^{\mathrm{i}}} \partial\Omega_k^{\mathrm{i}} + \sum_{k=0}^{N^{\mathrm{c}}} \partial\Omega_k^{\mathrm{c}}. \tag{2.45}$$

Electrode $\partial\Omega_0^{\mathrm{c}}$ is considered as reference for all the voltages of the remaining electrodes, that are supposed to be assigned. $\boldsymbol{J}\cdot\boldsymbol{n}=0$ is set as boundary conditions (b.c.) on each $\partial\Omega_k^{\mathrm{i}}$, where $\boldsymbol{n}$ is the outwards oriented normal unit vector of $\partial\Omega$.

Since $\boldsymbol{\nabla}\times\boldsymbol{E}=\boldsymbol{0}$ and the electrostatic field is a conservative field, we introduce the scalar potential U such that $\boldsymbol{E}=-\boldsymbol{\nabla}U$. Similarly, since $\boldsymbol{\nabla}\cdot\boldsymbol{J}=\boldsymbol{0}$, we introduce a vector potential $\boldsymbol{T}$ such that $\boldsymbol{J}=\boldsymbol{\nabla}\times\boldsymbol{T}$.

2.8.1 Survey of standard formulations

There are several formulations to numerically solve Poisson problems like (2.44a)-(2.44b)-(2.44c). The most common one is the FEM formulation based on the scalar potential SP expanded with the classical nodal basis functions. With this formulation the unknowns $U^{\mathcal{N}}$, are the DoFs of the scalar potentials U sampled on the grid nodes. The voltages $\boldsymbol{E}^{\mathcal{E}}=-\mathbb{G}U^{\mathcal{N}}$, are associated with the grid edges, where $\mathbb{G}$ is the edge-node incidence matrix. Finally, the DoFs of $\boldsymbol{J}$ are attached to dual faces as [45] shows. Faraday's law (2.44a) is enforced implicitly by the scalar potential, whereas the solenoidality of the current (2.44b) is enforced on the boundary of each dual cell by the linear system, see for example [46].

Other possibilities emerge as we exchange the association of physical variables between the primal and dual grids. In *complementary* formulations the DoFs of $\boldsymbol{J}$ are attached to faces of the primal grid, whereas the scalar potential DoFs are attached to dual nodes. Consequently, the voltages $\boldsymbol{E}^{\tilde{\mathcal{E}}}=-\mathbb{D}^T U^{\tilde{\mathcal{N}}}$, are associated with dual edges. There are two methods to obtain a complementary formulation. The first *complementary* formulation VP fulfill (2.44b) by using the electric vector potential $\mathbf{T}$ [46] such that $\boldsymbol{J}^{\mathcal{F}}=\mathbb{C}\boldsymbol{T}^{\mathcal{E}}$ (We remark that in general the so-called *relative cohomology theory* is required to present this formulation, see [46]). We note that in the VP formulation the role of physical laws are exchanged w.r.t. the SP formulation since Faraday's law (2.44a) is enforced with a linear system in the VP formulation.

There is a second method to produce complementary formulations that we call *complementary-dual*: they are complementary formulation but still use the scalar potential $U^{\tilde{\mathcal{N}}}$, which is sampled on grid dual nodes, one-to-one with grid cells. An effective complementary-dual formulation, which is algebraically equivalent to the VP formulation, is the mixed-hybrid MH formulation [47], [48].

2.8.2 Square resistor benchmark

As a real-case benchmark problem, we want to compute the conductance G of the square resistor in Fig. 2.3. A voltage of $u=1\,\mathrm{V}$ is enforced between the two electrodes, which are placed in the two lateral surfaces of the solid square torus. The conductor placed inside the solid torus has an electrical conductivity $\sigma=1\,\mathrm{S/m}$. This problem is interesting because, like most industrial problems, exhibits a singular analytical solution.

Yet, the analytical solution is available $G = 10.23409256$ S and will be used as a reference value. The conductance G is extracted by computing the total dissipated power $P = G\,u^2 = \int_\Omega \frac{|\boldsymbol{J}|^2}{\sigma}$ or alternatively by using the Ohm's law $G = i/u$, where i is the current that flows between the two electrodes.

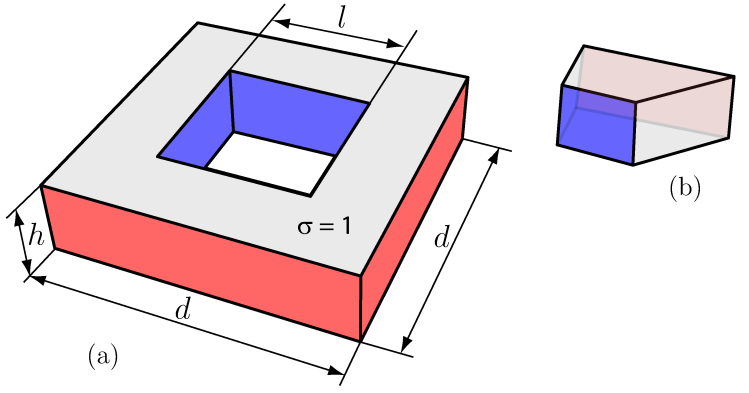

Figure 2.3: (a) The geometry of the square resistor benchmark ($h = 1\,\mathrm{m}$, $d = 4\,\mathrm{m}$, and $l = 2\,\mathrm{m}$). The two electrodes are depicted in red and blue. (b) Thanks to the symmetry, the computational domain has been reduced to one in eight of the resistor.

3

Equivalence between mimetic numerical schemes and discrete geometric approaches

In this chapter we provide new geometric viewpoints of low-order mimetic numerical schemes. In particular, we uncover two main geometric aspects that are hidden in the MFD method.

As a first contribution, in Section 3.2 we show that reconstruction operators defined via the Stokes theorem introduced in [49] lead to reconstruction operators defined by geometric elements of the barycentric dual grid. This result shows that the reconstruction operators used in the two main families of low-order compatible numerical schemes are equivalent. As a consequence, consistent parts of local mass matrices are the same and the only difference is the particular choice of the stabilization part. This result extends the approach proposed in the FV literature [20] by introducing a novel reconstruction formula for edge DoFs. We point out that, although in the approach presented in [20] the reconstruction formula for face DoFs is derived, its identification with dual edges is not highlighted. The derivation of the novel edge formula completes the picture of the relation with the dual barycentric grid. Since the barycentric dual grid is already present in the scheme, although not made explicit, we present a novel geometric numerical scheme which is a hybrid approach between the MFD method and the discrete geometric formulations. Here, a barycentric dual grid is assumed as part of the numerical scheme and the stabilization part proposed in the mimetic literature is employed. The resulting scheme benefits from all the advantages of the two methodologies: the geometric structure allows us to write balance laws explicitly and simplifies the scheme software implementation [3], while the stabilization part offers the possibility of changing properties of the local mass matrices through the selection of some user-dependent parameters. Finally, the equivalence of the consistent terms also implies that mathematical results obtained for the MFD method, such as convergence properties or error estimates, can be readily extended to this unified framework, and in general to the

discrete geometric formulations.

A second contribution of this chapter is to introduce a novel property of the reconstruction operators, to which we refer as P_0-consistency and we present in Section 3.1. As the name suggests, a set of P_0-consistent reconstruction operators are such that the global mass matrix constructed using them satisfy the consistency property. The main difference with respect to the set of requirements that characterize reconstruction operators introduced in [49] is that the new definition of P_0-consistency is a global property since it involves reconstruction operators defined on more than one element. The formal definition of this concept is based on a commuting diagram property of the discrete de Rham complex and its dual, which lie at the foundations of the mathematical structure of physical theories, as represented graphically by Tonti diagrams [3]. This approach extends what is usually done in literature. In [11], a dual grid is introduced as part of the numerical scheme and then its relation with reconstruction operators is identified. Instead, in our approach the barycentric dual grid is derived as a canonical choice of designing P_0-consistent reconstruction operators on general polyhedral elements. Moreover, as an additional result of our analysis, it is the only possible construction if we restrict the design of reconstruction operators at the single element level, namely without taking into account reconstruction operators of other elements. The fundamental role played by Stokes theorem in mimetic numerical schemes appears, therefore, as the geometric manifestation of using the barycentric dual grid to design local reconstruction operators. A geometric characterization of P_0-consistent reconstruction operators can be given as the affine solution space of a linear system of equations, and from a physical point view, we show that it is directly connected with conservation laws of physical theories. An analysis of this linear system constitutes the starting point to geometrically optimize discrete Hodge operators. As an instance, it reveals that, besides the barycentric dual grid, there are other possible dual grids, which are not constructed according to the barycentric subdivision. In addition, for every cell we can choose an arbitrary point in space to geometrically define the entries of local reconstruction operators. We provide an example of how this freedom of choice in the design of reconstruction operators can be used to optimize them, as measured by a L^2 norm.

To conclude, in Section 5.3 we show examples of how the above concepts can be used to design new numerical schemes.

3.1 Definition of P_0-consistent reconstruction operators

In this section we define the novel concept of P_0-consistent reconstruction operators and provide their geometric counterpart. The definition relies on a commuting diagram property of the discrete dual de Rham complex which involves the derived discrete operators and characterizes in a rigorous way the class of all reconstruction operators that can be used to construct a global mass matrix in a such a way that a global patch test is passed. Thus, the P_0-consistency of reconstruction operators refers to the property that we able to reproduce exactly the continuous solution at the discrete level. Instead, the consistency condition (2.23) only expresses an equivalence between discrete and continuous energy, and thus, it is a necessary but not sufficient condition. The

geometry enters into the scheme since for low-order numerical schemes only averages of the reconstruction operator over a given region are measured. In such a case, the matrix representation of the average reconstruction operator has entries which can be interpreted as familiar geometric elements of the space, like segments and polygons. In the following, we will only deal with averages of reconstruction operators, thus, for convenience, when we consider reconstruction operators we will always refer to their average on a suitable region as represented by their associated matrix. The geometric conditions that characterize the class of P_0-consistent reconstruction operators can be written as a linear system of equations. A closer inspection of this linear system reveals that in general it does not have a unique solution. Thus, there is no unique way of designing P_0-consistent reconstruction operators.

3.1.1 P_0-consistent face reconstruction operators

We introduce the definitions through examples of physical theories.

Let us consider the example of magnetic phenomena as described by the following equations

$$\begin{aligned} \boldsymbol{\nabla} \times \boldsymbol{H} &= \boldsymbol{J}, \\ \boldsymbol{\nabla} \cdot \boldsymbol{B} &= 0, \\ \boldsymbol{H} &= \boldsymbol{\nu} \boldsymbol{B}, \\ \boldsymbol{H} \times \boldsymbol{n} &= \boldsymbol{J}_s, \end{aligned} \tag{3.1}$$

where the source of the problem is a known current density vector field $\boldsymbol{J}$. In particular, let us consider boundary conditions in such a way that vector fields $\boldsymbol{B}, \boldsymbol{H}$ are constant in Ω. In this situation, Ampère's law must be $\boldsymbol{\nabla} \times \boldsymbol{H} = \boldsymbol{J} = \boldsymbol{0}$.

To introduce the corresponding mimetic form of the above equations, let $\boldsymbol{B}^{\mathcal{F}}$ be the DoFs of the vector field $\boldsymbol{B}$ attached to faces of the grid. We represent the divergence and curl operators that appear in (3.1) by the discrete operator $\mathcal{DIV}$ and the derived operator $\widetilde{\mathcal{CURL}}$. Using these operators, the mimetic discretization of (3.1) reads as follows

$$\mathcal{DIV}\, \boldsymbol{B}^{\mathcal{F}} = \boldsymbol{0}, \tag{3.2}$$

$$\widetilde{\mathcal{CURL}}\, \boldsymbol{B}^{\mathcal{F}} = \boldsymbol{0}. \tag{3.3}$$

Due to the Dirichlet boundary conditions, (3.3) should be considered only for the interior edges. By using (2.36), (3.3) can be equivalently written as follows

$$\mathbb{M}^{\mathcal{E}^{-1}} \mathcal{CURL}^T \mathbb{M}^{\mathcal{F}} \boldsymbol{B}^{\mathcal{F}} = \boldsymbol{0}. \tag{3.4}$$

We left multiply (3.4) by $\mathbb{M}^{\mathcal{E}}$ obtaining

$$\mathcal{CURL}^T \Big(\sum_{c \in C} (\mathbb{O}_c^{\mathcal{F}})^T \, |c| (\mathbb{R}_c^{\mathcal{F}})^T \boldsymbol{\nu}\, \mathbb{R}_c^{\mathcal{F}} \mathbb{O}_c^{\mathcal{F}} \Big) \boldsymbol{B}^{\mathcal{F}} = \boldsymbol{0}, \tag{3.5}$$

where we have used (2.32) for the definition of the global mass matrix $\mathbb{M}^{\mathcal{F}}$. Next, by

using $\boldsymbol{H} = \boldsymbol{\nu B}$, it follows that

$$\mathcal{CURL}^T \Big(\sum_{c \in C} (\mathbb{O}_c^{\mathcal{F}})^T \, |c| (\mathbb{R}_c^{\mathcal{F}})^T \Big) \, \boldsymbol{H} = \boldsymbol{0}. \tag{3.6}$$

Finally, since $\boldsymbol{H} \in \mathbb{R}^3$ is arbitrary, we obtain the following *geometric condition*

$$\mathrm{row}_e \mathbb{C}^T \Big(\sum_{c \in C} (\mathbb{O}_c^{\mathcal{F}})^T \, |c| (\mathbb{R}_c^{\mathcal{F}})^T \Big) = 0, \; \forall e \in E, e \not\subset \partial\Omega \tag{3.7}$$

where we have used the matrix form of the $\mathcal{CURL}$ operator.

A dimensional analysis informs us that each entry of matrix $|c|(\mathbb{R}_c^{\mathcal{F}})^T$ has units of linear meters. This is a direct consequence of the property of $\mathbb{R}_c^{\mathcal{F}}$ of being a left inverse of $\mathbb{P}_c^{\mathcal{F}}$. Given the physical dimensions, rows of $|c|(\mathbb{R}_c^{\mathcal{F}})^T$ can be regarded as geometric *edges* and (3.7) requires that they can be put one after the other to form a geometric closed path. Thus, they encode the geometric structure of a set of edges of a grid.

Definition 3.1.1 (P_0-consistent face reconstruction operators)**.** A collection of reconstruction operators $\{\mathbb{R}_c^{\mathcal{F}}\}_{c \in C}$ is said to be P_0-consistent if and only if the following *geometric condition* holds

$$\mathrm{row}_e \mathbb{C}^T \Big(\sum_{c \in C} (\mathbb{O}_c^{\mathcal{F}})^T \, |c| (\mathbb{R}_c^{\mathcal{F}})^T \Big) = 0, \; \forall e \in E, e \not\subset \partial\Omega. \tag{3.8}$$

3.1.2 P_0-consistent edge reconstruction operators

Let us consider the example of electric phenomena as described by the following equations

$$\begin{aligned} \boldsymbol{\nabla} \times \boldsymbol{E} &= \boldsymbol{0}, \\ \boldsymbol{\nabla} \cdot \boldsymbol{D} &= \rho, \\ \boldsymbol{D} &= \boldsymbol{\epsilon E}, \\ \boldsymbol{D} \cdot \boldsymbol{n} &= D_s, \end{aligned} \tag{3.9}$$

where the source of the problem is a known charge density scalar field ρ. In particular, let us consider boundary conditions in such a way that vector fields $\boldsymbol{E}, \boldsymbol{D}$ are constant in Ω. In this situation, Gauss' law must be $\boldsymbol{\nabla} \cdot \boldsymbol{D} = 0$.

To introduce the corresponding mimetic form of the above equations, let $\boldsymbol{E}^{\mathcal{E}}$ be the DoFs of the vector field $\boldsymbol{E}$ attached to edges of the grid. We represent the curl and divergence operators that appear in (3.9) by the discrete operator $\mathcal{CURL}$ and the derived operator $\widetilde{\mathcal{DIV}}$. Using these operators, the mimetic discretization of (3.9) reads as follows

$$\mathcal{CURL}\, \boldsymbol{E}^{\mathcal{E}} = \boldsymbol{0}, \tag{3.10}$$

$$\widetilde{\mathcal{DIV}}\, \boldsymbol{E}^{\mathcal{E}} = \boldsymbol{0}. \tag{3.11}$$

Due to the Dirichlet boundary conditions, (3.11) should be considered only for the

interior edges. By using (2.38), (3.11) can be equivalently written as follows

$$-\mathbb{M}^{\mathcal{N}^{-1}}\mathcal{G}\mathcal{R}\mathcal{A}\mathcal{D}^T\mathbb{M}^{\mathcal{E}}\boldsymbol{E}^{\mathcal{F}} = \mathbf{0}. \tag{3.12}$$

We left multiply (3.12) by $\mathbb{M}^{\mathcal{N}}$ obtaining

$$\mathcal{D}\mathcal{I}\mathcal{V}^T\Big(\sum_{c\in C}(\mathbb{O}_c^{\mathcal{E}})^T\,|c|(\mathbb{R}_c^{\mathcal{E}})^T\boldsymbol{\nu}\,\mathbb{R}_c^{\mathcal{E}}\mathbb{O}_c^{\mathcal{E}}\Big)\,\boldsymbol{E}^{\mathcal{E}} = \mathbf{0}, \tag{3.13}$$

where we have used (2.32) for the definition of the global mass matrix $\mathbb{M}^{\mathcal{E}}$. Next, by using $\boldsymbol{D} = \boldsymbol{\epsilon}\boldsymbol{E}$, it follows that

$$\mathcal{D}\mathcal{I}\mathcal{V}^T\Big(\sum_{c\in C}(\mathbb{O}_c^{\mathcal{E}})^T\,|c|(\mathbb{R}_c^{\mathcal{E}})^T\Big)\,\boldsymbol{D} = \mathbf{0}. \tag{3.14}$$

Finally, since $\boldsymbol{D}\in\mathbb{R}^3$ is arbitrary, we obtain the following *geometric condition*

$$\mathrm{row}_n\mathbb{G}^T\Big(\sum_{c\in C}(\mathbb{O}_c^{\mathcal{E}})^T\,|c|(\mathbb{R}_c^{\mathcal{E}})^T\Big) = 0,\ \forall n\in N, n\notin\partial\Omega \tag{3.15}$$

where we have used the matrix form of the $\mathcal{D}\mathcal{I}\mathcal{V}$ operator.

A dimensional analysis informs us that each entry of matrix $|c|(\mathbb{R}_c^{\mathcal{E}})^T$ has units of square meters. This is a direct consequence of the property of $\mathbb{R}_c^{\mathcal{E}}$ of being a left inverse of $\mathbb{P}_c^{\mathcal{E}}$. Given the physical dimensions, rows of $|c|(\mathbb{R}_c^{\mathcal{E}})^T$ can be regarded as geometric *faces* and (3.15) requires that they can be put side to side to form a geometric closed surface. Thus, they encode the geometric structure of a set of faces of a grid.

Definition 3.1.2 (P_0-consistent reconstruction operators for edge DoFs)**.** A collection of reconstruction operators $\{\mathbb{R}_c^{\mathcal{E}}\}_{c\in C}$ is said to be P_0-consistent if and only if the following *geometric condition* holds

$$\mathrm{row}_n\mathbb{G}^T\Big(\sum_{c\in C}(\mathbb{O}_c^{\mathcal{E}})^T\,|c|(\mathbb{R}_c^{\mathcal{E}})^T\Big) = 0,\ \forall n\in N, n\notin\partial\Omega. \tag{3.16}$$

3.1.3 Linear system formulation of P_0-consistent reconstruction operators

The two requirements that reconstruction operators have to satisfy can be encoded into a linear system of equations. The unknown variables are the entries of the matrices $|c|(\mathbb{R}_c^{\mathcal{E}})^T$ and $|c|(\mathbb{R}_c^{\mathcal{F}})^T$ with constraint equations expressed by (3.8), (3.16) and the requirement that each matrix $\mathbb{R}_c^{\mathcal{E}}$ and $\mathbb{R}_c^{\mathcal{F}}$ is a left inverse of the local projection operator, which is the accuracy property for constant vector fields. More specifically, we can write

$$\begin{cases} \mathbb{R}_c^{\mathcal{F}}\mathbb{P}_c^{\mathcal{F}} = \mathbb{I}_3,\ \forall c\in C \\ \mathrm{row}_e\mathbb{C}^T\Big(\sum_{c\in C}(\mathbb{O}_c^{\mathcal{F}})^T\,|c|(\mathbb{R}_c^{\mathcal{F}})^T\Big) = 0,\ \forall e\in E, e\not\subset\partial\Omega \end{cases} \tag{3.17}$$

and

$$\begin{cases} \mathbb{R}_c^{\mathcal{E}} \mathbb{P}_c^{\mathcal{E}} = \mathbb{I}_3, \ \forall c \in C \\ \text{row}_n \mathbb{G}^T \left(\sum_{c \in C} (\mathbb{O}_c^{\mathcal{E}})^T \, |c| (\mathbb{R}_c^{\mathcal{E}})^T \right) = 0, \ \forall n \in N, n \notin \partial\Omega \end{cases} \tag{3.18}$$

where $\mathbb{I}_3$ denotes the identity matrix of dimension 3.

A canonical solution of the two systems of equations is given by the barycentric dual grid. While the geometric conditions in (3.8) and (3.16) are trivially satisfied, proofs that geometric elements of the dual grid satisfy also the accuracy property are given in Section 3.2.

Other solutions are known for particular element shapes. For instance, for cubical grids [1] and Delaunay tetrahedral grids [16], Voronoi dual grids are well defined and satisfy the above properties. In general, the above linear systems do not have a unique solution, see Example 1.

The design of other P_0-consistent reconstruction operators is a global property since it links reconstruction operators defined on more than one element. A crucial observation is the following: if we design reconstruction operator at the single element level, without taking into account adjacent elements, the dual barycentric grid is the unique canonical choice to design reconstruction operators for arbitrary polyhedral elements.

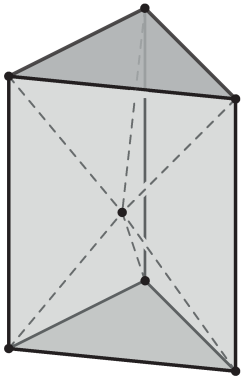

Figure 3.1: Polyhedral grid made by 2 tetrahedra and 3 square pyramids.

Example 1 (Geometric conditions for a polyhedral grid)**.** *Let us consider the polyhedral grid K pictured in Fig. 3.1. The grid is made by 5 cells and 14 faces. As an instance, we analyze the dimension of the affine solution space of the linear system (3.18). We satisfy the constraint expressed by the geometric conditions by identifying with the same unknown the unique two unknowns vector variables corresponding to the same face which lies at the intersection of two cells. Thus, we have* 14 *unknowns vectors variables each associated with a face of K. The requirement that the matrix $\mathbb{R}_c^{\mathcal{F}}$ is a left inverse of $\mathbb{F}_c$ for every cell c of K is encoded into a constraint matrix of size 15×14. A direct computation reveals that the rank of this constraint matrix is 12. Since there exists at least a solution, which is given by the coordinates of the barycenters of faces of K, see Section 3.2, the system is consistent, and we have an infinite number of solutions since*

the constraint matrix is rank deficient.

As a consequence of the conditions (3.8), (3.16), the geometric entries of the reconstruction operators on every element are defined up to an arbitrary point in $\mathbb{R}^3$. The latter, can be chosen freely and can be used to optimize reconstruction operators. We explore this possibility in Section 3.2.4.

3.2 Dual grid P_0-consistent reconstruction operators

In this section we prove that reconstruction operators defined in [49] are equivalent to reconstruction operators defined by geometric elements of the barycentric dual grid. The main novelty with respect to the proofs given in [49] relies in the explicit identification of geometrical elements of dual faces in the expression of edge reconstruction operators. Since, by construction, geometric elements of the barycentric dual grid satisfy the geometric conditions (3.8) and (3.16), the analysis reveals the role of the barycentric dual grid in low-order compatible numerical schemes: even if the dual grid is not assumed as a starting point of the method, for instance in mimetic literature [8], it appears as a canonical choice of constructing P_0-consistent reconstruction operators. According to the results derived in Section 3.1, the expressions of the reconstruction formulas informs us that they are defined up to an arbitrary point in $\mathbb{R}^3$. We show how this additional degree of freedom can be used to optimize specific objective functions.

3.2.1 Reconstruction formulas and their proofs

Let us consider a pair of vector fields $\boldsymbol{u}, \boldsymbol{w} : \mathbb{R}^3 \to \mathbb{R}^3$ along with a scalar field $w : \mathbb{R}^3 \to \mathbb{R}$ defined on a given cell c. The following well-known integration by parts formulas hold

$$\int_c \boldsymbol{u} \cdot \boldsymbol{\nabla} w \, dV = -\int_c \boldsymbol{\nabla} \cdot \boldsymbol{u} \, w \, dV + \int_{\partial c} \boldsymbol{u} \cdot \boldsymbol{n} \, w \, dS, \tag{3.19}$$

$$\int_c \boldsymbol{u} \cdot \boldsymbol{\nabla} \times \boldsymbol{w} \, dV = \int_c \boldsymbol{\nabla} \times \boldsymbol{u} \cdot \boldsymbol{w} \, dV + \int_{\partial c} \boldsymbol{u} \times \boldsymbol{n} \cdot \boldsymbol{w} \, dS. \tag{3.20}$$

Theorem 2 (Reconstruction formulas)**.** *The following tensor identities hold*

$$\sum_{f \in F(c)} \tilde{\boldsymbol{e}}_{f_{|c}} \otimes \boldsymbol{f} = |c| \, \mathbb{I}_3, \tag{3.21}$$

$$\sum_{e \in E(c)} \tilde{\boldsymbol{f}}_{e_{|c}} \otimes \boldsymbol{e} = |c| \, \mathbb{I}_3. \tag{3.22}$$

Proof. Let $\boldsymbol{u}, \boldsymbol{w} \in \mathbb{R}^3$. Thanks to (3.19), we have

$$\begin{aligned} |c|\, \boldsymbol{u} \cdot \boldsymbol{w} &= \int_c \boldsymbol{u} \cdot \boldsymbol{w}\, dV = \int_c \boldsymbol{u} \cdot \boldsymbol{\nabla}(\boldsymbol{w} \cdot (\boldsymbol{x} - \tilde{\boldsymbol{n}}_c))\, dV \\ &= \int_{\partial c} (\boldsymbol{u} \cdot \boldsymbol{n})\, (\boldsymbol{w} \cdot (\boldsymbol{x} - \tilde{\boldsymbol{n}}_c))\, dS \\ &= \sum_{f \in F(c)} (\boldsymbol{u} \cdot \boldsymbol{f})\, ((\boldsymbol{b}_f - \tilde{\boldsymbol{n}}_c) \cdot \boldsymbol{w}). \end{aligned} \tag{3.23}$$

By using the definition of $\tilde{\boldsymbol{e}}_{f_{|c}}$, it follows that

$$|c|\, \boldsymbol{u} \cdot \boldsymbol{w} = \sum_{f \in F(c)} (\boldsymbol{u} \cdot \boldsymbol{f})\, (\tilde{\boldsymbol{e}}_{f_{|c}} \cdot \boldsymbol{w}). \tag{3.24}$$

Since $\boldsymbol{w}$ is arbitrary, (3.21) is proved.

Thanks to (3.20), we have

$$\begin{aligned} 2|c|\, \boldsymbol{u} \cdot \boldsymbol{w} = 2 \int_c \boldsymbol{u} \cdot \boldsymbol{w}\, dV &= \int_c \boldsymbol{u} \cdot \boldsymbol{\nabla} \times \big(\boldsymbol{w} \times (\boldsymbol{x} - \tilde{\boldsymbol{n}}_c)\big)\, dV \\ &= \int_{\partial c} (\boldsymbol{u} \times \boldsymbol{n}) \cdot \big(\boldsymbol{w} \times (\boldsymbol{x} - \tilde{\boldsymbol{n}}_c)\big)\, dS \\ &= \sum_{f \in F(c)} \int_f (\boldsymbol{u} \times \boldsymbol{n}_f) \cdot \big(\boldsymbol{w} \times (\boldsymbol{x} - \tilde{\boldsymbol{n}}_c)\big)\, dS \\ &= \sum_{f \in F(c)} |f|\, (\boldsymbol{u} \times \boldsymbol{n}_f) \cdot (\boldsymbol{w} \times \tilde{\boldsymbol{e}}_{f_{|c}}). \end{aligned} \tag{3.25}$$

Now, by applying the same argument used in (3.23) to the vector $\boldsymbol{u} \times \boldsymbol{n}_f$ restricted to a face f, we obtain

$$|f|\, \boldsymbol{u} \times \boldsymbol{n}_f = \sum_{e \in E(f)} (\boldsymbol{u} \cdot \boldsymbol{e})\, (\boldsymbol{b}_e - \boldsymbol{p}), \tag{3.26}$$

where $\boldsymbol{p} \in \mathbb{R}^3$ is an arbitrary node.

Apply (3.26) to every face f in the last term of (3.25), choosing the same node $\tilde{\boldsymbol{n}}_c$ as the arbitrary point involved in the formula. We obtain

$$\begin{aligned} 2|c|\, \boldsymbol{u} \cdot \boldsymbol{w} &= \sum_{f \in F(c)} \Big(\sum_{e \in E(f)} (\boldsymbol{u} \cdot \boldsymbol{e})\, (\boldsymbol{b}_e - \tilde{\boldsymbol{n}}_c) \Big) \cdot (\boldsymbol{w} \times \tilde{\boldsymbol{e}}_{f_{|c}}) \\ &= \sum_{f \in F(c)} \big(\tilde{\boldsymbol{e}}_{f_{|c}} \times \Big(\sum_{e \in E(f)} (\boldsymbol{u} \cdot \boldsymbol{e})\, (\boldsymbol{b}_e - \tilde{\boldsymbol{n}}_c) \Big)\big) \cdot \boldsymbol{w} \\ &= \sum_{e \in E(c)} (\boldsymbol{u} \cdot \boldsymbol{e})\, ((\tilde{\boldsymbol{e}}_{f_{i|c}} \times (\boldsymbol{b}_e - \tilde{\boldsymbol{n}}_c) - \tilde{\boldsymbol{e}}_{f_{j|c}} \times (\boldsymbol{b}_e - \tilde{\boldsymbol{n}}_c)) \cdot \boldsymbol{w}), \end{aligned} \tag{3.27}$$

where f_i, f_j are the unique faces of c such that $e = f_i \cap f_j$, for suitable indices $i = i(e)$ and $j = j(e)$, and oriented so that they induce opposite orientations on edge e. Now,

dividing by two both members of the last term in (3.27) and using the definition of $\tilde{\boldsymbol{f}}_{e|c}$, it follows that

$$|c|\,\boldsymbol{u}\cdot\boldsymbol{w} = \sum_{e\in E(c)} (\boldsymbol{u}\cdot\boldsymbol{e})\,(\tilde{\boldsymbol{f}}_{e|c}\cdot\boldsymbol{w}). \tag{3.28}$$

Since $\boldsymbol{w}$ is arbitrary, (3.22) is proved. □

Lemma 1 (Dual grid reconstruction operators)**.** *Let us consider a cell c. Choose an arbitrary point $\tilde{\boldsymbol{n}}_c \in \mathbb{R}^3$. Let $\mathbb{R}^{\mathcal{F}}_{\tilde{\boldsymbol{n}}_c} := \frac{1}{|c|}(\mathbb{P}^{\tilde{\mathcal{E}}}_{\tilde{\boldsymbol{n}}_c})^T$ and $\mathbb{R}^{\mathcal{E}}_{\tilde{\boldsymbol{n}}_c} := \frac{1}{|c|}(\mathbb{P}^{\tilde{\mathcal{F}}}_{\tilde{\boldsymbol{n}}_c})^T$. The following formulas hold*

$$\mathbb{R}^{\mathcal{F}}_{\tilde{\boldsymbol{n}}_c}\mathbb{P}^{\mathcal{F}}_c = \tilde{\mathbb{E}}_c^T\mathbb{F}_c = \mathbb{I}_3, \tag{3.29}$$

$$\mathbb{R}^{\mathcal{E}}_{\tilde{\boldsymbol{n}}_c}\mathbb{P}^{\mathcal{E}}_c = \tilde{\mathbb{F}}_c^T\mathbb{E}_c = \mathbb{I}_3. \tag{3.30}$$

Proof. By using (3.21), (3.22) and the definition of the matrices $\mathbb{P}^{\tilde{\mathcal{F}}}_{\tilde{\boldsymbol{n}}_c}, \mathbb{P}^{\tilde{\mathcal{E}}}_{\tilde{\boldsymbol{n}}_c}$, we have the following identities $(\mathbb{P}^{\tilde{\mathcal{E}}}_{\tilde{\boldsymbol{n}}_c})^T\mathbb{P}^{\mathcal{F}}_c = |c|\,\mathbb{I}_3$, $(\mathbb{P}^{\tilde{\mathcal{F}}}_{\tilde{\boldsymbol{n}}_c})^T\mathbb{P}^{\mathcal{E}}_c = |c|\,\mathbb{I}_3$. Dividing by $|c|$ both members of the latter equations we obtain the claimed result. □

3.2.2 Mass matrices

In compatible numerical methods, it is always necessary to map DoFs attached to geometric elements of the primal grid into DoFs attached to geometric elements of the dual grid, or viceversa [1], [45], [3]. This process of converting one type of DoFs to another is performed by the so-called *discrete Hodge star operator* [2].

One possible method for explicitly converting one type of DoFs into another is to reconstruct the polynomial vector field in each cell starting from DoFs attached to a set of geometric elements, and then project this vector field onto the corresponding geometric elements related by duality. We will follow this principle to design both types of local mass matrices, which give a discrete realization of the corresponding discrete Hodge operators.

In this section we show how to construct consistent mass matrices which map DoFs attached to primal geometric elements to DoFs attached to dual geometric elements.

3.2.3 Local mass matrices

Let us focus on a element c where two pairs of constant vector fields are defined, namely $\boldsymbol{u}, \boldsymbol{v}$ and $\boldsymbol{w}, \boldsymbol{z}$. The two pairs are related by two constitutive relations

$$\boldsymbol{v} = \mathbb{K}_{c,1}\boldsymbol{u}, \tag{3.31}$$

$$\boldsymbol{z} = \mathbb{K}_{c,2}\boldsymbol{w}, \tag{3.32}$$

where $\mathbb{K}_{c,1}, \mathbb{K}_{c,2}$ are two symmetric positive definite matrices of order 3, assumed to be uniform in c.

Now, let us introduce the restriction of DoFs of the vector fields to c. In particular, we attach DoFs to geometric elements of the primal and dual grid as follows $\boldsymbol{u}^{\mathcal{F}}_c = \mathbb{F}_c\boldsymbol{u}$, $\boldsymbol{v}^{\tilde{\mathcal{E}}}_c = \tilde{\mathbb{E}}_c\boldsymbol{v}$, $\boldsymbol{w}^{\mathcal{E}}_c = \mathbb{E}_c\boldsymbol{w}$, $\boldsymbol{z}^{\tilde{\mathcal{F}}}_c = \tilde{\mathbb{F}}_c\boldsymbol{z}$.

The local mass matrix $\mathbb{M}_c^{\mathcal{F}}$ maps DoFs of $\boldsymbol{u}$ attached to faces to DoFs of $\boldsymbol{v}$ attached to dual edges of the barycentric dual grid. We say that $\mathbb{M}_c^{\mathcal{F}}$ is a *consistent mass matrix* if

$$\boldsymbol{v}_c^{\tilde{\mathcal{E}}} = \mathbb{M}_c^{\mathcal{F}} \boldsymbol{u}_c^{\mathcal{F}} \tag{3.33}$$

holds exactly for any pair of constant vector fields $\boldsymbol{u}, \boldsymbol{v}$ satisfying (3.31).

Similarly, a local mass matrix $\mathbb{M}_c^{\mathcal{E}}$ maps DoFs of $\boldsymbol{w}$ attached to edges to DoFs of $\boldsymbol{z}$ attached to dual faces of the barycentric dual grid. We say that $\mathbb{M}_c^{\mathcal{E}}$ is a *consistent mass matrix* if

$$\boldsymbol{z}_c^{\tilde{\mathcal{F}}} = \mathbb{M}_c^{\mathcal{E}} \boldsymbol{w}_c^{\mathcal{E}} \tag{3.34}$$

holds exactly for any pair of constant vector fields $\boldsymbol{w}, \boldsymbol{z}$ satisfying (3.32).

An efficient recipe to construct consistent and symmetric matrices $\mathbb{M}_c^{\mathcal{F}}, \mathbb{M}_c^{\mathcal{E}}$ combines the geometric identities in Theorem 2 with the uniformity of the vector fields.

By applying Theorem 2 and the definitions of DoFs $\boldsymbol{v}_c^{\tilde{\mathcal{E}}}, \boldsymbol{u}_c^{\mathcal{F}}$ of $\boldsymbol{v}$ and $\boldsymbol{u}$, we have that

$$\boldsymbol{v}_c^{\tilde{\mathcal{E}}} = \tilde{\mathbb{E}}_c \boldsymbol{v} = \tilde{\mathbb{E}}_c \mathbb{K}_{c,1} \boldsymbol{u} = \tilde{\mathbb{E}}_c \mathbb{K}_{c,1} \frac{1}{|c|} (\tilde{\mathbb{E}}_c^T \mathbb{F}_c) \boldsymbol{u} = \frac{(\tilde{\mathbb{E}}_c \mathbb{K}_{c,1} \tilde{\mathbb{E}}_c^T)}{|c|} \boldsymbol{u}_c^{\mathcal{F}}, \tag{3.35}$$

and hence, it follows that a symmetric and consistent matrix $\mathbb{M}_c^{\mathcal{F}}$ is given by

$$\mathbb{M}_c^{\mathcal{F}} = \frac{\tilde{\mathbb{E}}_c \mathbb{K}_{c,1} \tilde{\mathbb{E}}_c^T}{|c|}. \tag{3.36}$$

Similarly, by applying Theorem 2 and the definitions of DoFs $\boldsymbol{z}_c^{\tilde{\mathcal{F}}}, \boldsymbol{w}_c^{\mathcal{E}}$ of $\boldsymbol{z}$ and $\boldsymbol{w}$, we have that

$$\boldsymbol{z}_c^{\tilde{\mathcal{F}}} = \tilde{\mathbb{F}}_c \boldsymbol{z} = \tilde{\mathbb{F}}_c \mathbb{K}_{c,2} \boldsymbol{w} = \tilde{\mathbb{F}}_c \mathbb{K}_{c,2} \frac{1}{|c|} (\tilde{\mathbb{F}}_c^T \mathbb{E}_c) \boldsymbol{w} = \frac{(\tilde{\mathbb{F}}_c \mathbb{K}_{c,2} \tilde{\mathbb{F}}_c^T)}{|c|} \boldsymbol{w}_c^{\mathcal{E}}, \tag{3.37}$$

and hence, it follows that a symmetric and consistent matrix $\mathbb{M}_c^{\mathcal{E}}$ is given by

$$\mathbb{M}_c^{\mathcal{E}} = \frac{\tilde{\mathbb{F}}_c \mathbb{K}_{c,2} \tilde{\mathbb{F}}_c^T}{|c|}. \tag{3.38}$$

The matrices $\mathbb{M}_c^{\mathcal{F}}$ and $\mathbb{M}_c^{\mathcal{E}}$, defined in (3.36) and (3.38), are symmetric and consistent but are not positive definite. To achieve this, the idea, developed in Lemma 2, is to add to the consistent positive definite matrices $\mathbb{M}_c^{\mathcal{F}}, \mathbb{M}_c^{\mathcal{E}}$ a *stability matrix*, which is symmetric and positive semidefinite. The stability matrix coincides with one proposed in the mimetic literature [8].

Lemma 2. *Let m be the cardinality of $F(c)$ or $E(c)$. Let $\mathbb{K}_{|c}$ be a symmetric and positive definite matrix of order 3. Let $\alpha = (\alpha_1, \ldots, \alpha_{m-3}) \in (\mathbb{R}^+)^{m-3}$ be any $(m-3)$-upla of positive real numbers and let $\mathbb{D}_\alpha$ be the diagonal matrix whose diagonal entries are $\alpha_1, \ldots, \alpha_{m-3}$. Denote by $\mathbb{W}_c^{\mathcal{F}}$ and $\mathbb{W}_c^{\mathcal{E}}$ the matrices whose columns form an orthonormal*

basis for $\mathrm{im}(\mathbb{F}_c)^\perp$ *and* $\mathrm{im}(\mathbb{E}_c)^\perp$*, respectively. Then, the following matrices*

$$\mathbb{M}_c^{\mathcal{F}} := \frac{1}{|c|}\tilde{\mathbb{E}}_c\mathbb{K}_c\tilde{\mathbb{E}}_c^T + \mathbb{W}_c^{\mathcal{F}}\mathbb{D}_\alpha(\mathbb{W}_c^{\mathcal{F}})^T, \tag{3.39}$$

$$\mathbb{M}_c^{\mathcal{E}} := \frac{1}{|c|}\tilde{\mathbb{F}}_c\mathbb{K}_c\tilde{\mathbb{F}}_c^T + \mathbb{W}_c^{\mathcal{E}}\mathbb{D}_\alpha(\mathbb{W}_c^{\mathcal{E}})^T, \tag{3.40}$$

are symmetric, consistent and positive definite.

In the general case of a grid made of more than one cell, the corresponding global mass matrices $\mathbb{M}^{\mathcal{F}}, \mathbb{M}^{\mathcal{E}}$ are obtained by assembling, cell by cell, the contributions from the local matrices $\mathbb{M}_c^{\mathcal{F}}$ and $\mathbb{M}_c^{\mathcal{E}}$, respectively.

3.2.4 Optimizing dual grid reconstruction operators

The expressions of the dual grid reconstruction operators show that they are uniquely defined up to an arbitrary point $\tilde{\boldsymbol{n}}_c \in \mathbb{R}^3$, which can be chosen independently for every cell. Now, we show how this fact can be used to optimize the dual grid reconstruction operators. In particular, we ask if for a given element c there exists a point such that the reconstructed vector field is optimal with respect to a least square difference.

Let $\mathcal{X} = \mathcal{E}$ or $\mathcal{X} = \mathcal{F}$. We observe that the map $\mathbb{P}_c^{\mathcal{X}}$ has rank 3, thus it is not surjective as a map from $\mathbb{R}^3$ to $\mathbb{R}^{|\mathcal{X}_c|}$.

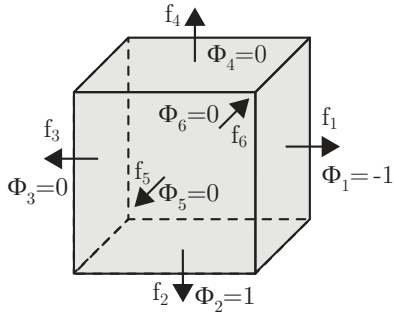

Figure 3.2: A cubic cell c. A set of DoFs $\boldsymbol{\Phi}$ are attached to faces. This is an example of DoFs that are not image of any constant vector field under the projection map $P_c^{\mathcal{F}}$.

Let $\boldsymbol{u}_c^{\mathcal{X}} \notin \mathrm{im}\,\mathbb{P}_c^{\mathcal{X}}$. In this case, among all possible constant vector fields, we select the minimum norm solution of the following least squares problem

$$\boldsymbol{u}^* = \underset{\boldsymbol{u}\in\mathbb{R}^3}{\arg\min} \left\|\mathbb{P}_c^{\mathcal{X}}\boldsymbol{u} - \boldsymbol{u}_c^{\mathcal{X}}\right\|_2. \tag{3.41}$$

The Moore–Penrose matrix inverse of $\mathbb{P}_c^{\mathcal{X}}$, denoted as $(\mathbb{P}^{\mathcal{X}})_c^+$, allows us to write the minimum norm solution $\boldsymbol{u}^*$ of (3.41) into the following form

$$\boldsymbol{u}^* = (\mathbb{P}_c^{\mathcal{X}})^+\boldsymbol{u}_c^{\mathcal{X}}; \tag{3.42}$$

moreover, since $\mathbb{X}_c$ has full rank, $(\mathbb{P}_c^{\mathcal{X}})^+$ can be explicitly written as $(\mathbb{P}_c^{\mathcal{X}})^+ = ((\mathbb{P}_c^{\mathcal{X}})^T\mathbb{P}_c^{\mathcal{X}})^{-1}(\mathbb{P}_c^{\mathcal{X}})^T$, see [50].

Now, we ask the following question: does there exist a point $\tilde{\boldsymbol{n}}_c \in \mathbb{R}^3$ such that $\mathbb{R}_{\tilde{\boldsymbol{n}}_c}^{\mathcal{X}} = (\mathbb{P}_c^{\mathcal{X}})^+$? By Lemma 1, we know that $\mathbb{R}_{\tilde{\boldsymbol{n}}_c}^{\mathcal{X}}$ is a left inverse of $\mathbb{P}_c^{\mathcal{X}}$. Thus, in order that $\mathbb{R}_{\tilde{\boldsymbol{n}}_c}^{\mathcal{X}} = (\mathbb{P}_c^{\mathcal{X}})^+$, it is sufficient that the following property holds

$$\mathbb{P}_c^{\mathcal{X}}\mathbb{R}_{\tilde{\boldsymbol{n}}_c}^{\mathcal{X}} = (\mathbb{P}_c^{\mathcal{X}}\mathbb{R}_{\tilde{\boldsymbol{n}}_c}^{\mathcal{X}})^T, \tag{3.43}$$

see [50]. For tetrahedral and cubic cells we have the following characterization.

Lemma 3. *Let c be a tetrahedral or cubic element. If $\tilde{\boldsymbol{n}}_c$ is the barycenter of c, then $\tilde{\boldsymbol{n}}_c$ satisfies (3.43).*

Proof. First, suppose that c is a tetrahedron, see the tetrahedron pictured in Fig. 3.3. Consider the case $\mathcal{X} = \mathcal{F}$. We have to show that $\mathbb{P}_c^{\mathcal{F}}\mathbb{R}_{\tilde{\boldsymbol{n}}_c}^{\mathcal{F}}$ is symmetric. It is sufficient to consider the pair of faces $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$. By definition, we have

$$\boldsymbol{f}_1 \cdot \tilde{\boldsymbol{e}}_2 = \Big(\frac{\boldsymbol{e}_1 \times \boldsymbol{e}_2}{2}\Big) \cdot \Big(\frac{\boldsymbol{e}_2 + \boldsymbol{e}_3}{3} - \frac{\boldsymbol{e}_1 + \boldsymbol{e}_2 + \boldsymbol{e}_3}{4}\Big) = \frac{1}{24}(\boldsymbol{e}_1 \times \boldsymbol{e}_2) \cdot \boldsymbol{e}_3, \tag{3.44}$$

$$\boldsymbol{f}_2 \cdot \tilde{\boldsymbol{e}}_1 = \Big(\frac{\boldsymbol{e}_2 \times \boldsymbol{e}_3}{2}\Big) \cdot \Big(\frac{\boldsymbol{e}_1 + \boldsymbol{e}_2}{3} - \frac{\boldsymbol{e}_1 + \boldsymbol{e}_2 + \boldsymbol{e}_3}{4}\Big) = \frac{1}{24}(\boldsymbol{e}_2 \times \boldsymbol{e}_3) \cdot \boldsymbol{e}_1. \tag{3.45}$$

By comparing the above two expression we obtain $\boldsymbol{f}_1 \cdot \tilde{\boldsymbol{e}}_2 = \boldsymbol{f}_2 \cdot \tilde{\boldsymbol{e}}_1$.

Consider now the case $\mathcal{X} = \mathcal{E}$. Let us prove that $\mathbb{P}_c^{\mathcal{E}}\mathbb{R}_{\tilde{\boldsymbol{n}}_c}^{\mathcal{E}}$ is symmetric. It is sufficient to consider the pair of edges $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$. By definition, we have

$$\tilde{\boldsymbol{f}}_1 \cdot \boldsymbol{e}_2 = \Big(\frac{1}{2}\Big(\frac{\boldsymbol{e}_1 + \boldsymbol{e}_2}{3} - \frac{\boldsymbol{e}_1 + \boldsymbol{e}_3}{3}\Big) \times \Big(\frac{\boldsymbol{e}_1}{2} - \frac{\boldsymbol{e}_1 + \boldsymbol{e}_2 + \boldsymbol{e}_3}{4}\Big)\Big) \cdot \boldsymbol{e}_2 = \frac{1}{24}(\boldsymbol{e}_1 \times \boldsymbol{e}_3) \cdot \boldsymbol{e}_2, \tag{3.46}$$

$$\tilde{\boldsymbol{f}}_2 \cdot \boldsymbol{e}_1 = \Big(\frac{1}{2}\Big(\frac{\boldsymbol{e}_2 + \boldsymbol{e}_3}{3} - \frac{\boldsymbol{e}_1 + \boldsymbol{e}_2}{3}\Big) \times \Big(\frac{\boldsymbol{e}_2}{2} - \frac{\boldsymbol{e}_1 + \boldsymbol{e}_2 + \boldsymbol{e}_3}{4}\Big)\Big) \cdot \boldsymbol{e}_1 = \frac{1}{24}(\boldsymbol{e}_3 \times \boldsymbol{e}_2) \cdot \boldsymbol{e}_1. \tag{3.47}$$

By comparing the above two expressions we obtain $\tilde{\boldsymbol{f}}_1 \cdot \boldsymbol{e}_2 = \tilde{\boldsymbol{f}}_2 \cdot \boldsymbol{e}_1$.

Let c be a cubic element. Consider the case $\mathcal{X} = \mathcal{F}$. We have to show that $\mathbb{P}_c^{\mathcal{F}}\mathbb{R}_{\tilde{\boldsymbol{n}}_c}^{\mathcal{F}}$ is symmetric. Observe that the (i,j)-entry of this product is of the form $\boldsymbol{f}_i \cdot \tilde{\boldsymbol{e}}_{f_j}$ for some pair of faces $\boldsymbol{f}_i, \boldsymbol{f}_j$. Let $i \neq j$. Define $\tilde{\boldsymbol{e}}_i \coloneqq \tilde{\boldsymbol{e}}_{f_i}$ and $\tilde{\boldsymbol{e}}_j \coloneqq \tilde{\boldsymbol{e}}_{f_j}$. Two cases are possible, either $\boldsymbol{f}_i = -\boldsymbol{f}_j$ or $\boldsymbol{f}_i \cdot \boldsymbol{f}_j = 0$. In the first case, we have $\tilde{\boldsymbol{e}}_i = -\tilde{\boldsymbol{e}}_j$ from which follows that $\boldsymbol{f}_i \cdot \tilde{\boldsymbol{e}}_j = \boldsymbol{f}_j \cdot \tilde{\boldsymbol{e}}_i$. In the second case, we have $\boldsymbol{f}_i \cdot \tilde{\boldsymbol{e}}_j = 0$ and $\boldsymbol{f}_j \cdot \tilde{\boldsymbol{e}}_i = 0$, from which follows that $\boldsymbol{f}_i \cdot \tilde{\boldsymbol{e}}_j = \boldsymbol{f}_j \cdot \tilde{\boldsymbol{e}}_i$.

Consider now the case $\mathcal{X} = \mathcal{E}$. Let us prove that $\mathbb{P}_c^{\mathcal{E}}\mathbb{R}_{\tilde{\boldsymbol{n}}_c}^{\mathcal{E}}$ is symmetric. Observe that the (i,j)-entry of this product is of the form $\boldsymbol{e}_i \cdot \tilde{\boldsymbol{f}}_{e_j}$ for some pair of edges $\boldsymbol{e}_i, \boldsymbol{e}_j$. Let $i \neq j$. Define $\tilde{\boldsymbol{f}}_i \coloneqq \tilde{\boldsymbol{f}}_{e_i}$ and $\tilde{\boldsymbol{f}}_j \coloneqq \tilde{\boldsymbol{f}}_{e_j}$. Two cases are possible, either $\boldsymbol{e}_i = \pm\boldsymbol{e}_j$ or $\boldsymbol{e}_i \cdot \boldsymbol{e}_j = 0$. In the first case, we have $\tilde{\boldsymbol{f}}_i = \pm\tilde{\boldsymbol{f}}_j$ from which follows that $\boldsymbol{e}_i \cdot \tilde{\boldsymbol{f}}_j = \boldsymbol{e}_j \cdot \tilde{\boldsymbol{f}}_i$.

In the second case, we have $\boldsymbol{e}_i \cdot \tilde{\boldsymbol{f}}_j = 0$ and $\boldsymbol{e}_j \cdot \tilde{\boldsymbol{f}}_i = 0$ from which follows that $\boldsymbol{e}_i \cdot \tilde{\boldsymbol{f}}_j = \boldsymbol{e}_j \cdot \tilde{\boldsymbol{f}}_i$. □

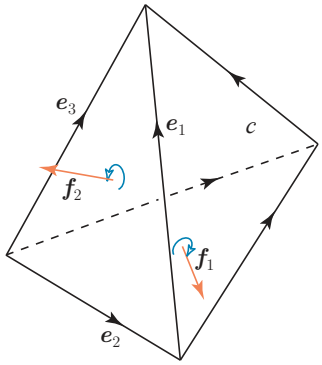

Figure 3.3: A tetrahedron to illustrate the geometric quantities involved in the proof of Lemma 3

Let us denote by D the set of all cells that satisfy (3.43) for some point $\tilde{\boldsymbol{n}}_c \in \mathbb{R}^3$.

Example 2. *Let us consider the case of the face reconstruction operator* $\mathbb{R}^{\mathcal{F}}_{\tilde{\boldsymbol{n}}_c}$. *By definition, we have* $\mathbb{R}^{\mathcal{F}}_{\tilde{\boldsymbol{n}}_c} = \frac{1}{|c|}(\mathbb{P}^{\tilde{\mathcal{E}}}_c - \mathbf{1}\tilde{\boldsymbol{n}}_c^T)^T$, *where* $\mathbf{1}$ *denotes the vector in* $\mathbb{R}^{|F(c)|}$ *whose entries are all equal to* 1. *Let us consider a pyramid with a square base divided in two triangles. The boundary decomposes into six faces. A direct computation reveals that this cell does not belong to* D. *To show this, it is sufficient to directly compute the Moore–Penrose matrix inverse* $(\mathbb{P}^{\mathcal{F}}_c)^+$ *and subtract from* $\mathbb{R}^{\mathcal{F}}_{\tilde{\boldsymbol{n}}_c}$ *the matrix* $|c|(\mathbb{P}^{\mathcal{F}}_c)^+$. *If there exists a point* $\tilde{\boldsymbol{n}}_c$ *satisfying (3.43), then the columns of this matrix difference must be all equal to* $\tilde{\boldsymbol{n}}_c$.

To handle the situations where a cell $c \notin D$, we propose to select a dual node $\tilde{\boldsymbol{n}}_c$ which minimizes the distance with respect to the Moore–Penrose inverse matrix measured by a L^2 metric.

Definition 3.2.1 (Optimal dual node)**.** Let us introduce the following map

$$\begin{aligned} \xi_c \colon \mathbb{R}^3 &\to \mathbb{R} \\ \tilde{\boldsymbol{n}}_c &\mapsto \left\| \mathbb{R}^{\mathcal{X}}_{\tilde{\boldsymbol{n}}_c} - (\mathbb{P}^{\mathcal{X}}_c)^+ \right\|_2^2. \end{aligned} \tag{3.48}$$

We define the *optimal dual node* of a cell c, $\tilde{\boldsymbol{n}}_{\mathrm{opt},c}$, as the unique point which minimizes function ξ_c.

From the practical point of view we can solve the optimization problem in (3.48) for every cell and find two optimal reference points, one for the matrix $\mathbb{R}^{\mathcal{F}}_{\tilde{\boldsymbol{n}}_c}$ and the other one for $\mathbb{R}^{\mathcal{E}}_{\tilde{\boldsymbol{n}}_c}$. To speed up the computation we choose a unique point as the optimal dual node of the matrix $\mathbb{R}^{\mathcal{F}}_{\tilde{\boldsymbol{n}}_c}$. In this case, we give an explicit expression of the optimal dual node.

Lemma 4 (Optimal dual node expression)**.** *The optimal dual node can be written as*

$$\tilde{\boldsymbol{n}}_{opt,c} = \frac{1}{|F(c)|} \sum_{i=1}^{|F(c)|} \mathrm{col}_i \mathbb{A}, \tag{3.49}$$

where $\mathbb{A} \coloneqq (\mathbb{P}_c^{\widetilde{\mathcal{E}}})^T - |c|(\mathbb{P}_c^{\mathcal{F}})^+$ *and the sum is over the set of columns of* $\mathbb{A}$*.*

Proof. By definition $\mathbb{R}_{\tilde{\boldsymbol{n}}_c}^{\mathcal{F}} = \frac{1}{|c|}(\mathbb{P}_c^{\widetilde{\mathcal{E}}} - \mathbf{1}\tilde{\boldsymbol{n}}_c^T)^T$. The objective function can be written as $\frac{1}{|c|^2}\left\|\mathbb{A} - \tilde{\boldsymbol{n}}_c \mathbf{1}^T\right\|_2^2 = \frac{1}{|c|^2}\sum_i \left\|\mathbb{A}_i - \tilde{\boldsymbol{n}}_c\right\|_2^2$, where the sum is over the set of columns of $\mathbb{A}$. By equating the gradient to zero we obtain the minimum point at $\frac{1}{|F(c)|}\sum_{i=1}^{|F(c)|} \mathrm{col}_i \mathbb{A}$. □

3.3 Numerical results

As a reference problem we consider the the stationary current conduction in Section 2.8. We first verified the equivalence of the reconstruction operators introduced in [8] with the reconstructions operators defined by geometric elements of the dual barycentric grid in Section 3.2 by solving two multi-material patch tests on a grid made by general polyhedra. Then, the square resistor benchmark problem in Section 2.8.2 is solved using the geometric optimization of reconstruction operators described in Section 3.2.4. Finally, we show an example of the role of the dual barycentric dual grid for the correct computation of some global variables, in this case the current flowing through the electrodes.

3.3.1 Multi-material patch tests

The multi-material patch tests are Poisson problems designed in such a way that their solution is piecewise-uniform. By interpreting the Poisson problems as direct current conduction problems, a simple way to produce multi-material patch test is to consider a resistor with two conductors with different material properties (called electric resistivity ρ) placed in *series* or in *parallel* as described in detail later.

To test the edge mass matrices we use the classical scalar potential formulation SP. Instead, in order to test the face mass matrices we use the vector potential VP [51] or mixed-hybrid MH formulations [10], [48]. We remark that the VP and MH formulations produce the same results given that they are algebraically equivalent [48]. The polyhedral grid used in the example is formed by 131 nodes, 306 edges, 237 faces and 61 cells. The grid is obtained through two levels of sub-gridding of a few cells, so that the obtained elements are general polyhedra.

In the first multi-material patch test, two different materials with different material properties are places in *series*. From the result represented in Fig. 3.4, we conclude that the tangential component of the electric field $\boldsymbol{E}$ is conserved across the material interface, whereas the tangential component of the current density field $\boldsymbol{J}$ jumps. This result holds for both formulations.

In the second multi-material patch test, two different materials with different material properties are places in *parallel*. From the result represented in Fig. 3.5, we conclude that the normal component of the current density $\boldsymbol{J}$ is conserved across the material

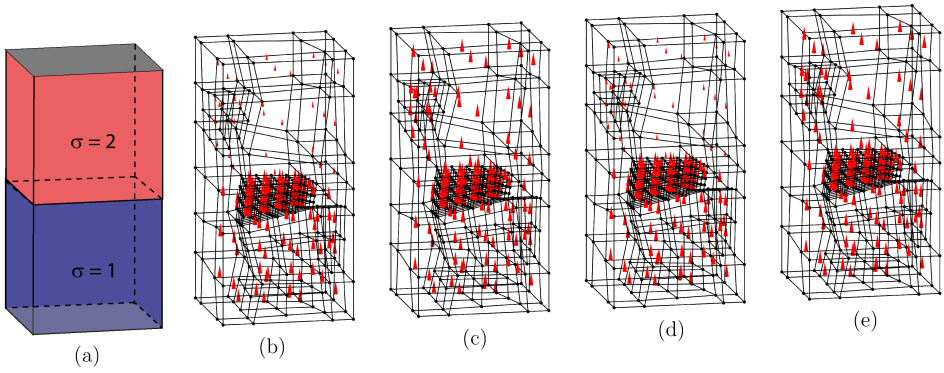

Figure 3.4: The *series* multi-material patch test (a). We set the voltage between the two electrodes, represented in gray in the picture, to 1 V. (b) Electric field $\boldsymbol{E}$ produced by the SP formulation. (c) Current density field $\boldsymbol{J}$ produced by the SP formulation. (d) Electric field $\boldsymbol{E}$ produced by the VP formulation. (e) Current density field $\boldsymbol{J}$ produced by the VP formulation.

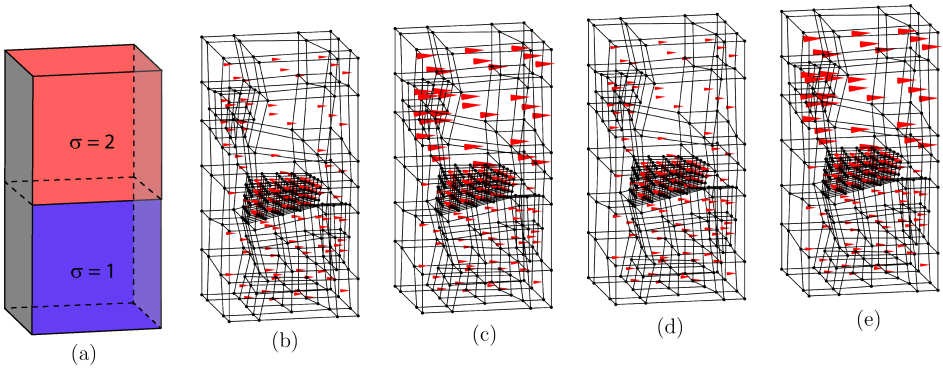

Figure 3.5: The *parallel* multi-material patch test (a). We set the voltage between the two electrodes, represented in gray in the picture, to 1 V. (b) Electric field $\boldsymbol{E}$ produced by the SP formulation. (c) Current density field $\boldsymbol{J}$ produced by the SP formulation. (d) Electric field $\boldsymbol{E}$ produced by the VP formulation. (e) Current density field $\boldsymbol{J}$ produced by the VP formulation.

interface, whereas the normal component of the electric field $\boldsymbol{E}$ jumps. This result holds for both formulations.

3.3.2 Square resistor benchmark

All simulations are performed using the optimal dual node for the geometric dual grid as described in Section 3.2.4. We do not observe a significant improvement in the convergence of the scheme although the reconstruction operator is locally optimized for every single cell. Here, the optimization is substantial for elements which are not

symmetric. This can be qualitatively seen by the results of Section 3.2.4, where it is shown that where the cell is cubic the optimal dual node coincides with the barycenter of the cell. We point out that the optimization can be performed as a preprocessing step for the problem under consideration which defines the optimal geometric structure of reconstruction operators.

The conductance of the square resistor has been evaluated by the SP and $VP = MH$ formulations on refined grids. The results are collected in Fig. 5.4. First, it is interesting to note that the results relative to the scalar potential SP and to the vector potential VP or mixed-hybrid MH formulations provide, respectively, the upper and lower bounds for the conductance [52], [51]. This property is called *complementarity* in the computational electromagnetics community [48]. Second, in Fig. 3.6, four different type of grids have been used. Polyhedral grids with subgridding, i.e. a nonconforming-like local mesh refinement technique, appears to be particularly appealing.

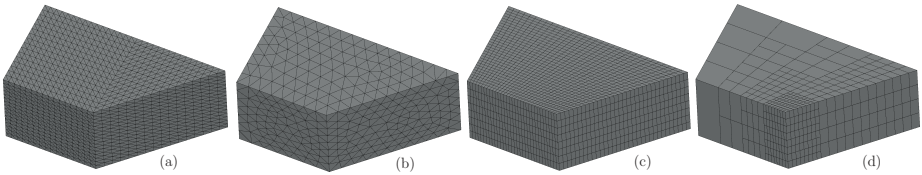

Figure 3.6: The four different types of grids used to discretize the geometry of the square resistor in Fig. 2.3. (a) structured tetrahedral grid. (b) unstructured tetrahedral grid. (c) structured hexahedral grid. (d) polyhedral grid with *subgridding*.

A faster convergence can be reached only by using automatic mesh adaptivity. A comparison of the results obtained with tetrahedral and polyhedral unstructured meshes constructed with automatic mesh adaptivity is prevented by the lack of an automatic polyhedral mesh generator. Such a mesh generator is under development.

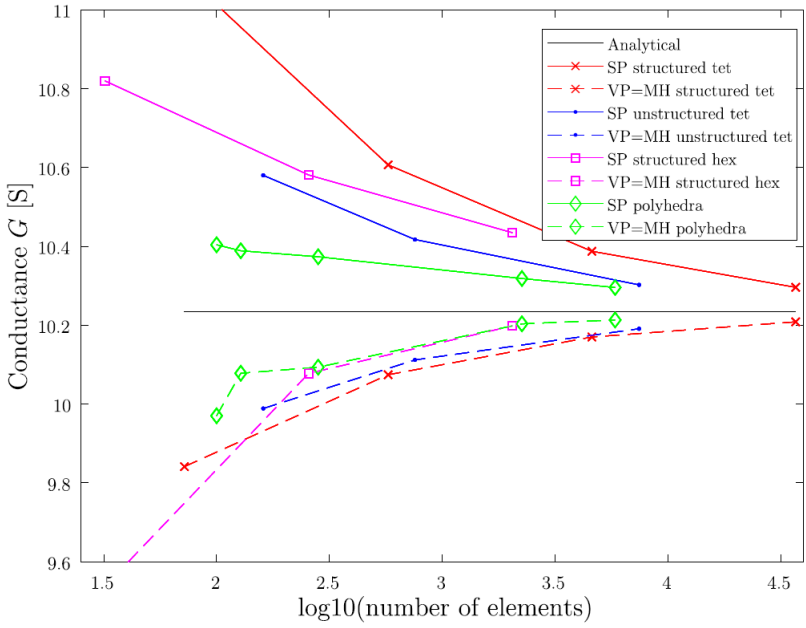

Figure 3.7: Results for the square resistor benchmark.

3.3.3 Balance laws and dual grid P_0-consistent reconstruction operators

Let us consider the problem described in Section 2.8.2. Now we show that the P_0-consistency property of reconstruction operators guarantees the continuous balance laws at the discrete level. To begin with, we observe that

$$\mathcal{CURL}\, \boldsymbol{E}^{\mathcal{E}} = \mathbf{0}, \tag{3.50}$$

so that an essential physical property of the continuous problem is preserved. What about the flux-conservation property of the current density $\boldsymbol{J}$? The question naturally arises since in the mimetic framework DoFs of the current density $\boldsymbol{J}$ are not explicitly defined. As a first approach to assess the conservation property of $\boldsymbol{J}$, let $\boldsymbol{E}_c$ be the element-wise electric field constructed starting from edge DoFs. Then, we denote by $\boldsymbol{J}_r$ the *reconstructed* current density defined on each element multiplying $\boldsymbol{E}_r$ by the corresponding conductivity, assumed to be constant over each element. Although $\boldsymbol{J}_r$ is divergence-free inside cells, there are flux "leaks" between faces of the primal grid. To validate this assertion, let I_1 and I_2 be the inward and outward current of $\boldsymbol{J}_r$ through the two pair of conductor surfaces $\partial\Omega_1^c$ and $\partial\Omega_2^c$, respectively. For the problem under consideration, the following values are found $I_1 = 1.3830$ and $I_2 = 1.0916$. Thus, the computed fluxes are "non-physical" since an essential physical property is violated, namely the fact that I_1 have to be equal to I_2.

The results of Section 3.1 shed light on the reasons why one should not integrate the reconstructed current density $\boldsymbol{J}_r$ on faces on the primal grid. Indeed, by requiring

$\widetilde{\mathcal{DIV}}\, \boldsymbol{E}^{\mathcal{E}} = \mathbf{0}$, we obtain

$$\mathrm{row}_n \mathbb{G}^T \Big(\sum_{c \in C} (\mathbb{O}_c^{\mathcal{E}})^T \, |c| (\mathbb{R}_c^{\mathcal{E}})^T \, \boldsymbol{J}_{r,c} \Big) = 0, \; \forall n \in N, n \notin \partial\Omega \tag{3.51}$$

So discrete solenoidality requires that the flux of $\boldsymbol{J}_r$ through the geometric closed surface associated with P_0-consistent edge reconstructions operators of every inner node vanishes. Therefore, we may aggregate geometric closed surfaces associated with inner nodes and the flux of $\boldsymbol{J}_r$ will vanish across such an aggregate. Based on these results, let us compute the following quantity

$$I_1 = \sum_{n \in \partial\Omega_1^c} \mathrm{row}_n \mathbb{G}^T \Big(\sum_{c \in C} (\mathbb{O}_c^{\mathcal{E}})^T \, |c| (\mathbb{R}_c^{\mathcal{E}})^T \, \boldsymbol{J}_{r,c} \Big) \tag{3.52}$$

as the discrete approximation of the current of $\boldsymbol{J}$ through $\partial\Omega_1^c$, where a similar expression holds for I_2. Then, we see that

$$\sum_{n \in N} \mathrm{row}_n \mathbb{G}^T \Big(\sum_{c \in C} (\mathbb{O}_c^{\mathcal{E}})^T \, |c| (\mathbb{R}_c^{\mathcal{E}})^T \, \boldsymbol{J}_{r,c} \Big) = 0 = I_1 + I_2, \tag{3.53}$$

where we have used the given boundary values of $\boldsymbol{J}$. Thus, we have $I_1 = -I_2 = 1.3225$, proving the physical soundness of the computed quantities (up to a minus sign, which accounts for the orientation of the surfaces).

As shown in Section 3.2, a canonical way of constructing P_0-consistent reconstruction operators is given by barycentric dual grid. In this case, (3.52) can be interpreted as the sum of currents through dual faces. We point out that the current in (3.52) is equal through every homologous surface associated with a different choice of the P_0-consistent edge reconstruction operators.

3.4 Conclusions

In this chapter, we have studied hidden geometric aspects of low-order compatible numerical schemes for arbitrary polyhedral grids. First, we have shown that standard mimetic numerical schemes have noteworthy geometric properties. These geometric properties were demonstrated by reformulating standard mimetic reconstruction operators using geometric elements of the barycentric dual grid, thus proving the equivalence between mimetic and geometric approaches. Second, we have introduced the class of P_0-consistency reconstruction operators which extends the standard consistency requirement of the mimetic framework. A characterization of these operators is given as an affine solution space of a linear system of equations. Given the geometric description of the scheme, this analysis reveals the existence of many dual grids that define reconstruction operators and constitutes the starting point to optimize them. Additionally, it shows that the barycentric dual grid is a canonical way of designing P_0-consistency reconstruction operators at single element level. The fundamental role played by Stokes theorem in mimetic numerical schemes appears, therefore, as the geometric manifestation of using the barycentric dual grid to design local reconstruction operators. Numerical examples show the importance of the geometric interpretation for the correct

evaluation of balance laws of physical theories, thus motivating the practical impact of including them as part of the numerical scheme.

4

Curved mimetic method

In this chapter we seek an extension of the MFD method to *curved grids*, i.e. grids having elements with arbitrary *curved faces*. One of crucial advantages of the MFD method is the possibility of discretizing the computational domain with unstructured *polyhedral grids* having elements with planar faces. For these kind of elements, it yields to a discrete convergent scheme which produces a *symmetric* matrix and uses one *degree of freedom* (DoF) for each face. Yet, many applications require grids whose elements have *curved* (i.e. non-planar) faces; for instance, the modeling of geometries like cylinders or spheres. We emphasize that even if the computational domain is a polyhedron, curved faces may appear in the interior of the domain just because a hexahedral mesh has been obtained with an unstructured meshing algorithm. However, the standard MFD method does not converge on such grids with curved faces [53].

To deal with such convergence problems, different techniques have been proposed in literature. We can group them in two basic approaches.

A first approach is to replace curved faces by triangles (or more generally planar polygons) to obtain a polyhedral grid so that we can apply the standard MFD method. Using this approach we obtain a symmetric discrete problem but the price to pay is the additional number of DoFs, which will be proportional to the number of polygons partitioning each curved face.

As a second approach, we can interpolate with standard least squares methods face DoFs inside each cell to produce a constant vector field. Then, the mass matrix is obtained by projecting the constant vector field onto edges of a *dual* (or *secondary*) grid $\tilde{K}$. A dual grid $\tilde{K}$ is constructed by duality starting from a given discretization grid K, hence, called a *primal* grid. Here, duality is expressed as a bijective correspondence between geometric elements of the pair of grids $(K, \tilde{K})$ such that to each d-dimensional geometric element in K corresponds a (unique) dual $(3-d)$-dimensional geometric element in $\tilde{K}$. Fundamentally, geometric elements of the barycentric dual grid are not only useful but they are implicitly present in the standard MFD method as shown in Chapter 3. Moreover, in Chapter 3 we presented a novel geometric MFD method that produces consistent mimetic inner products and their associated *mass matrices* starting from such a generalized dual grid structure. However, by using this dual grid approach

on curved meshes, the resulting discrete problem is non-symmetric which significantly reduces the number of available efficient solution methods. Basically, the discretization methods in [55], [56] reduce to the above approach, although a dual grid is not explicitly introduced.

The aim of this chapter is to introduce the *curved MFD method*, an extension of the MFD method which yields to a discrete problem which is symmetric and uses only one DoF for each curved face for any curved grid K. We thus answer to an open question raised in [57] on "whether the use of additional degrees of freedom is the only way to preserve symmetry in the discrete problem". To the best of our knowledge, [57] is the only low-order numerical method that can handle curved grids. In [57], a symmetric discrete problem which uses three DoFs for each curved face (more precisely, *strongly* curved faces[1]) is presented. There are other effective approaches for the accurate treatment of curved domains but restricted to the Finite Element framework [58], to 2-dimensional grids [59] or to high-order methods [60]. We note that the lowest order version of the approach in [60] employs more than one DoF for each curved face and thus is not equivalent to the curved MFD method proposed in this chapter.

The principle at the core of our construction is to abandon the definition of *consistency* used in the standard MFD method. Consistency condition (2.23) is a local property of *mimetic inner products* and *reconstruction operators* that encodes an exactness property for constant vector fields [44]. Focusing on local consistency leads to one of the two approaches outlined above, which, however, do not achieve a discrete problem which is both symmetric and uses one DoF for each curved face. Instead, we employ the novel concept of P_0-consistency introduced in Section 3.1. Contrary to the consistency condition (2.23) of the standard MFD method, P_0-consistency is a global property since it involves reconstruction operators defined on more than one element.

As detailed in Section 3.1, the idea behind the definition of P_0-consistency boils down to the fact that a canonical way of designing consistent reconstruction operators is given by geometric elements of the *barycentric dual grid* [54]. Remarkably, standard reconstruction operators in [61] defined using Stokes' theorem are equivalent to reconstruction operators defined by geometric elements of the barycentric dual grid as in Section 3.2. Reconstruction operators defined by the barycentric dual grid satisfy two key properties. First, they are a left inverse (i.e. *unisolvence* or *accuracy* property for constant vector fields) for local projection operators. Second, their entries form geometric closed paths or surfaces supported on different elements. Now, the idea it to abstract the above two properties and reverse our line of reasoning: any grid whose elements satisfy the above two properties leads to consistent reconstruction operators, precisely, P_0-consistent reconstruction operators.

In Section 3.1 a geometric characterization of P_0-consistent reconstruction operators is given as the affine solution space of a linear system of equations. In Chapter 4 we give a purely combinatorial characterization of grids for which there exists a solution of this linear system, or equivalently, there exist P_0-consistent reconstruction operators. This combinatorial condition depends solely on the number of geometric element of the curved grid and requires that the number of curved faces is greater than three times

[1] [57] distinguishes between *moderately* and *strongly* curved faces; the distinction between the two types of curved faces is based, as the names suggest, on a measure of curvature of the face and is used as a definition for theoretical analysis of convergence.

the number of elements in K. If we assume that the curved grid K is made of elements having k curved faces each, the above combinatorial condition is satisfied for $k \geq 6$. Once P_0-consistent reconstruction operators are available, local mass matrices are constructed as in the standard MFD method. Hence, all the solid foundation of the standard MFD can be applied also to our novel curved MFD method.

In Section 4.2 we tested the curved MFD method on curved grids and we show that consistency is achieved.

4.1 Curved MFD method

The novel curved MFD method is as follows. The definition of DoFs and discrete differential operators is the same of the standard MFD method. Instead, a new class reconstruction operators is employed. Accordingly, discrete inner products and derived differential operators are defined starting from reconstruction operators as described in Section 2.5 and Section 2.5.1.

We choose reconstruction operators among the class of P_0-consistent face reconstruction operators defined in (3.8) and (3.18).

P_0-consistent reconstruction operators on every element are defined up to an arbitrary point in $\mathbb{R}^3$. Let us introduce the vector $\boldsymbol{D}_c = (\dots, \mathbb{D}_{c,f}, \dots)^T$ of size $|F(c)| \times 1$ collecting coefficients of row vector $\text{row}_c\mathbb{D}$ corresponding to faces $f \in F(c)$. Given arbitrary nodes $\{\boldsymbol{b}_c\}_{c\in C}$ in $\mathbb{R}^3$ we have that

$$\mathbb{R}^{\mathcal{F}}_{\boldsymbol{b}_c} \coloneqq \mathbb{R}^{\mathcal{F}}_c - \boldsymbol{b}_c \boldsymbol{D}_c^T \; \forall c \in C \tag{4.1}$$

is a well-defined family $\{\mathbb{R}^{\mathcal{F}}_{\boldsymbol{b}_c}\}_{c\in C}$ of P_0-consistent reconstruction operators. This follows from the fact $\boldsymbol{D}_c^T \mathbb{P}^{\mathcal{F}}_c = \boldsymbol{0}$, since by definition of $\mathbb{P}^{\mathcal{F}}_c$, its rows are exactly face vectors decomposing the boundary of c.

The two requirements (3.18) can be encoded into a linear system of equations by introducing the matrices $\mathbb{P}^{\mathcal{F}}$, $\mathbb{R}^{\mathcal{F}}$ and $\mathbb{I}$ defined as follows.

- Matrix $\mathbb{P}^{\mathcal{F}}$ of size $3|C| \times |F|$. $\mathbb{P}^{\mathcal{F}}$ is a block matrix defined as

$$\mathbb{P}^{\mathcal{F}} = \begin{pmatrix} \vdots \\ (\mathbb{P}^{\mathcal{F}}_c)^T \mathbb{O}^{\mathcal{F}}_c \\ \vdots \end{pmatrix}. \tag{4.2}$$

  See Fig. 4.1.

- Matrix $\mathbb{R}^{\mathcal{F}}$ of size $|F| \times 3$. $\mathbb{R}^{\mathcal{F}}$ is defined as

$$\mathbb{R}^{\mathcal{F}} = \begin{pmatrix} \vdots \\ \boldsymbol{b}_f^T \\ \vdots \end{pmatrix}, \tag{4.3}$$

  where each row $\boldsymbol{b}_f$ corresponds to a face $f \in F$. If f is the common face between

$$
\begin{array}{c}
\\
c'' \\
c \\
\\
c'
\end{array}
\begin{pmatrix}
 & 0 & & -\boldsymbol{f}_j & & 0 \\
 & \vdots & & 0 & & \boldsymbol{f}_k \\
 & \boldsymbol{f}_i & & \vdots & & \vdots \\
\cdots & \vdots & \cdots & & \cdots & \\
 & -\boldsymbol{f}_i & & \boldsymbol{f}_j & & \\
 & \vdots & & \vdots & & \\
 & 0 & & 0 & & 0
\end{pmatrix}
$$

Figure 4.1: Structure of matrix $\mathbb{P}^{\mathcal{F}}$. Face vector of each internal face appears on two different rows with opposite sign, where the two rows correspond to the unique two elements sharing the internal face; instead, each boundary face appears only in one block corresponding to the unique element containing it. For instance, the internal face f_i is shared between elements c, c' and its face vector $\boldsymbol{f}_i$ appears with opposite sign on different rows corresponding to elements c, c'; the boundary face f_k is contained in the unique element c'' and its face vector $\boldsymbol{f}_k$ appears only in the rows corresponding to element c''.

two elements c, c', then we identify the columns of $\mathbb{R}_c^{\mathcal{F}}$ and $\mathbb{R}_{c'}^{\mathcal{F}}$ corresponding to face f with the same vector $\boldsymbol{b}_f$.

- Matrix $\mathbb{I}$ of size $3|C| \times 3$. $\mathbb{I}$ is a block matrix defined as

$$
\mathbb{I} = \begin{pmatrix} \vdots \\ |c|\mathbb{I} \\ \vdots \end{pmatrix}. \tag{4.4}
$$

We write the following linear system of equations

$$
\mathbb{P}^{\mathcal{F}}\mathbb{R}^{\mathcal{F}} = \mathbb{I}_3, \tag{4.5}
$$

where $\mathbb{I}_3$ is the identity matrix of order 3. Note that information on volumes of elements is encoded in matrix $\mathbb{I}$.

Let now study the existence of a solution of linear system (4.5). The matrices $\mathbb{P}^{\mathcal{F}}, \mathbb{R}^{\mathcal{F}}$ and $\mathbb{I}$ depend on the geometry of the input mesh K. Hence, conditions that guarantee existence of a solution restrict the class of input meshes. In order to guarantee the existence of a solution of linear system (4.5) a necessary and sufficient condition is that $\operatorname{rank} \mathbb{P}^{\mathcal{F}} = 3|C|$. Hence, a necessary condition is that $|F| \geq 3|C|$.

The next result is the crucial result of the chapter. It shows that for a special class of grids we can always construct P_0-consistent reconstruction operators. This result is remarkable since it extends the standard MFD method to this class of curved grids.

It can be advisable to use Fig. 4.1 as an illustration of the structure of matrix $\mathbb{P}^{\mathcal{F}}$ for the following proof.

Theorem 3 (Existence of P_0-consistent reconstruction operators)**.** *Let $K = (V, E, F, C)$ be a curved grid such that*

$$|F| \geq 3|C|. \tag{4.6}$$

Then, there exists a solution of linear system (4.5).

Proof. Assume that, for the sake of contradiction, $\operatorname{rank} \mathbb{P}^{\mathcal{F}} < 3|C|$. Hence, by definition of the block matrix $\mathbb{P}^{\mathcal{F}}$ in (4.2), there exists a volume c whose block $(\mathbb{P}^{\mathcal{F}}_c)^T \mathbb{O}^{\mathcal{F}}_c$ is linear combinations of other blocks $(\mathbb{P}^{\mathcal{F}}_{c'})^T \mathbb{O}^{\mathcal{F}}_{c'}$ of $\mathbb{P}^{\mathcal{F}}$ with $c' \in C \setminus \{c\}$ as follows

$$(\mathbb{P}^{\mathcal{F}}_c)^T \mathbb{O}^{\mathcal{F}}_c + \sum_{c' \in C \setminus \{c\}} p_{c'} (\mathbb{P}^{\mathcal{F}}_{c'})^T \mathbb{O}^{\mathcal{F}}_{c'} = \mathbf{0}, \tag{4.7}$$

for some real coefficients $p_{c'}$. Let S be the set of volumes in C whose coefficient p_c is different from zero in (4.7), namely, $S = \{c \in C \mid p_c \neq 0\}$. S is not empty since $c \in S$. Moreover, $S = C$. Indeed, if that is not the case, then $C \setminus S \neq \emptyset$. Since K is connected, there exists $c^* \in C \setminus S$ such that $c^* \cap c'$ for some $c' \in S$. It follows that $p_{c^*} \neq 0$ and thus $c^* \in C$ as well. Now, since $S = C$, there exists a boundary face f' contained in some volume $c' \neq c$ such that $p_{c'} \neq 0$. Since f' is a boundary face, c' is the unique volume in C containing it. Hence, the column of $(\mathbb{P}^{\mathcal{F}}_c)^T \mathbb{O}^{\mathcal{F}}_c$ corresponding to the boundary face f' should be different from the zero vector since the expression (4.7) includes the block $p_{c'} (\mathbb{P}^{\mathcal{F}}_{c'})^T \mathbb{O}^{\mathcal{F}}_{c'}$. This gives the desired contradiction since $c \neq c'$. □

We now discuss the implications of the condition (4.6) on the geometry of K. Let us assume that each element c has the same number k of boundary faces. Consider the set T of all pairs (c, f) where f is a boundary face of c. Since every element has k boundary faces, $|T| = k|C|$. But since each internal face is shared by exactly two element and each boundary face is incident to exactly one element we have $|T| = 2|F_i| + |F_b|$, where sets F_i and F_b partition F into internal and boundary faces, respectively. By combining these equations and plugging into (4.6) we get the condition

$$(k - 6)|F_i| \geq (3 - k)|F_b|. \tag{4.8}$$

Summarizing the above reasoning, the following corollary follows immediately. It gives a special characterization of the class of polyhedral meshes K which admit P_0-consistent reconstruction operators.

Corollary 1. *Let $K = (V, E, F, C)$ be a curved grid such that the boundary of every element $c \in C$ decomposes into k faces. Then, there exists a solution of (4.5) if and only if $k \geq 6$.*

Corollary 1 implies that every curved grid such that the boundary of every element decomposes into $k \geq 6$ faces we can write a discrete problem which is symmetric and uses only one DoF for each curved face.

To deal with the cases $k = 4$ or $k = 5$, we propose the following construction. We select a face f of each element c to form a sequence of pairs $T = (\ldots, (f, c), \ldots)$. Then, if $k = 4$, we partition each face in T into three polygons and we group them to form three linearly independent face vectors. Similarly, if $k = 5$, we partition each face in T into two polygons and we group them to form two linearly independent face vectors. At

the end of the procedure, in both cases we get still a curved grid but where each element satisfies the condition $k \geq 6$ so that we can apply Corollary 1. Note that a sequence T as above can be constructed efficiently in linear time and space using standard spanning tree constructions; see Section 6.3.2 for a detailed description of such a construction. The price to pay is the additional number of DoFs which shifts from $|F|$ to $|F| + 2|C|$ for $k = 4$ and $|F| + |C|$ for $k = 5$.

4.2 Numerical results

In this section we present numerical experiments to test the consistency of the curved MFD method. We consider the stationary conduction problem described in Section 2.8.

Similarly to *mixed FE* [62], we write a system with both $\boldsymbol{I}^{\mathcal{F}}$ and $U^{\tilde{\mathcal{N}}}$ as unknowns

$$\mathbb{M}^{\mathcal{F}} \boldsymbol{I}^{\mathcal{F}} - \mathbb{D}^T U^{\tilde{\mathcal{N}}} = \boldsymbol{E}_s^{\tilde{\mathcal{E}}}, \tag{4.9}$$

$$\mathbb{D} \boldsymbol{I}^{\mathcal{F}} = \boldsymbol{0}. \tag{4.10}$$

Even though this saddle-point problem may be casted as a symmetric and positive-definite system by the penalty method (see [62] p. 80), a smarter solution, inspired by mixed-hybrid FEs, involves a domain decomposition with as many sub-domains as mesh elements.

4.2.1 Patch test

We test the consistency of the curved mimetic method by solving Poisson problem whose solution is uniform in Ω. Since we consider the stationary current conduction problem in Section 2.8, a simple way to produce a uniform vector fields is to consider a resistors with known symmetries.

Fig. 4.2 and Fig. 4.3 report the results where it has been verified that the curved MFD method reproduces the exact analytical solution. We note that the standard MFD method and the approach proposed in [57] produce symmetric discrete problems that use more than one DoF for each curved face. Instead, our formulation employs just one DoF for each curved face.

Let us now consider a standard barycentric dual grid to define reconstruction operators. In this case we use one DoF for each curved face. However, the approach does not produce good results. To show this, we compute the dissipated power of the two resistors and we compare it with their analytical values by computing relative errors. For the first test case, we get a 0.02 % relative error whereas for the second one we get a 1 % relative error.

4.3 Conclusions

A new MFD method able to deal with curved faces has been introduced and produces a discrete problem which is symmetric and uses one DoFs for each curved face. This has been achieved by employing the novel concept of P_0-consistent reconstruction operators. As a result, the concept of dual mesh has been generalized. The consistency of the new

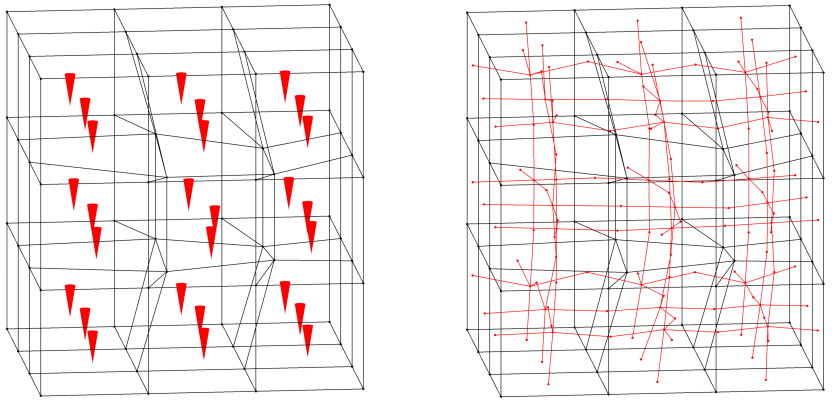

Figure 4.2: A curved grid partitioning the cube resistor where all internal faces of each element are curved. (a) Uniform electric field reconstructed inside each curved element; it coincides with the analytical value. (b) In red, dual grid structure associated with P_0-consistent face reconstruction operators solution of (4.5).

scheme has been tested numerically, demonstrating that the exact solution of the patch test is recovered for domains having curved boundary.

A range of work is slated for future investigation, focusing on further improvements to the properties of P_0-consistent reconstruction operators, as well as the development of fast combinatorial methods for the solution of the linear system to compute the generalized dual mesh. It is also expected that the methods described here can be generalized to high-order methods. However, it is expected that geometric interpretation of the high-order method to be more challenging.

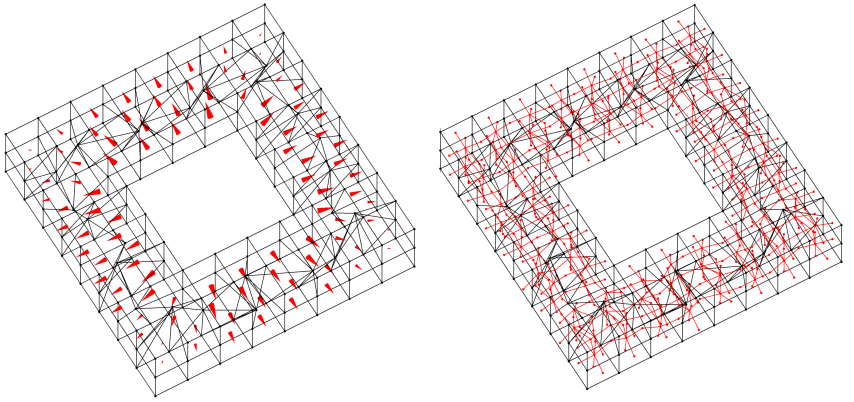

Figure 4.3: A curved grid partitioning the cubic resistor where all internal faces of each element are curved. (a) Uniform electric field reconstructed inside each curved element; it coincides with the analytical value. (b) In red, dual grid structure associated with P_0-consistent face reconstruction operators solution of (4.5).

5

Explicit geometric construction of sparse inverse mass matrices

In this chapter we present a new geometric construction of *sparse inverse mass matrices* for arbitrary *tetrahedral grids* and possibly inhomogeneous and anisotropic materials, debunking the conventional wisdom that the barycentric dual grid prohibits a sparse representation for inverse mass matrices.

What is usually achieved is the construction of sparse mass matrices as in Section 3.2.2 which map DoFs attached to primal faces to DoFs attached to dual edges or DoFs attached to primal edges to DoFs attached to dual faces [2], [7], [13], [14]. This is exactly what the FEM mass matrices computed with Raviart–Thomas and Nédélec basis functions perform, [45], [7], respectively.

However, we can nonetheless define *inverse mass matrices* that map DoFs attached to dual geometric elements to DoFs attached to primal geometric elements. Explicitly computing the algebraic inverse of the mass matrix is not considered a viable solution given that such a matrix would be dense, so of questionable usefulness in practice. Therefore, we are in particular interested in a *sparse realization* of inverse mass matrices.

Inverse mass matrices play a vital role in many applications. Most notably, inverse Hodge operators enable consistent and explicit schemes to solve time-domain wave propagation problems [63], [64], [65], [66]. Other applications enabled by the inverse mass matrices comprise the explicit construction of the *codifferential operator*, the *Laplace–de Rham operator* [67] and compute the discrete Hodge decomposition of discrete fields [67], [68].

The construction of inverse Hodge operators for tetrahedral grids is not entirely new. In CM [1], [69], Discrete Exterior Calculus (DEC) [70] and in the cell-centered Finite Volume literature [71], a *Voronoï dual grid* based on circumcenters is used, and the resulting mass matrices are diagonal in such a way that their inverses can be easily computed. It is important to note that, when the material is anisotripic, the mass matrices are in general not diagonal, so a sparse inverse Hodge operator cannot be easily constructed. Another solution using a Voronoï dual grid is presented in [72]. However, most commonly used mesh generators like NETGEN and GMSH produce tetrahedral

grids that are not Delaunay and in this case the Voronoï dual grid cannot be defined [73].

The aim of this chapter is to extend the construction of sparse inverse Hodge operators on *barycentric dual grids* which can be defined for arbitrary tetrahedral grids. The barycentric dual grids are explicitly used in DGA [14], implicitly in FEM [45] and in the MFD [8] as described in Chapter 3.

Yet, devising a recipe to construct sparse inverse Hodge operators on a barycentric dual grid appears to be a formidable task [74], [65], [67], [72] given that the conventional wisdom is that the barycentric dual grid "prohibits a sparse representation for their inverse operators" [75] and, consequently, only approximate constructions have been proposed [65].

By using a barycentric dual grid, in [64], [76], [66], inverse mass matrices that map from dual faces to primal edges are constructed by assembling local contributions inside dual cells and then computing the algebraic inverse of the resulting local matrices. Yet, we note that the time needed to compute all local inverses is not negligible, because the rank of the local matrices is twenty or more.

In this chapter, we provide a unified framework for the construction of both edge and face mass matrices and their sparse inverses. Such a unifying principle relies on novel geometric reconstruction formulas, from which, according to the well-established design strategy described in Chapter 2,Chapter 3, local mass matrices are constructed as the sum of a consistent and a stabilization part. A major difference with the approaches proposed so far [64], [76], [66] is that the consistent part is defined geometrically and explicitly, that is, without the necessity of computing the inverses of local matrices. This provides a sensible speedup and an easier implementation.

In general, the major contribution of our construction is not only to have avoided the computation of the local inverses, but rather the fact that it results in symmetric expressions for both types of local mass matrices, thus highlighting the duality relationship between the pair of grids. A key requirement of our setting would be that the material parameters are constant on dual cells associated with the nodes of the primal simplicial grid. Instead, to deal with the general case of material parameters that are constant on the grid cells, but arbitrary discontinuous across the interfaces between the cells, two different approaches are proposed. The first one is a weighted averaging technique classically used in Finite Volumes Methods [71]. The second one is based on an extension of the approach proposed [64],[66].

In Section 5.3 we use these new sparse inverse mass matrices to discretize the reference problems in Section 2.8, providing the comparison between the results obtained by various formulations on a benchmark problem with analytical solution.

5.1 Sparse inverse mass matrices

To derive inverse mass matrices we will mimic the reasoning of Section 3.2.2. In particular, we will derive a local mass matrix starting from reconstruction formulas and subsequently construct a global mass matrix by applying a standard FE assembling process. The construction of inverse mass matrices is carried out only for tetrahedral grids through the introduction of specialized geometric identities. This approach differs from the one described in Section 3.2.2. Indeed, the same arguments used in the

proof of Theorem 2 cannot be applied to the barycentric dual grid since dual cells have non-planar geometric elements.

5.1.1 Dual cell reconstruction

In the following, we give novel proofs of reconstruction formula that provide a constant vector field defined on $\tilde{c}$ starting from DoFs attached to dual edges and dual faces.

Let c be a tetrahedron and let n be one of its nodes. Let $i \in \{1,2,3\}$. For $e_i \in E(n) \cap E(c)$, consider the unique face $f_i \in F(n) \cap F(c)$ to which e_i does not belong. In this way, for every $i \in \{1,2,3\}$, there exists, and is unique, $r_i \in \{-1,1\}$ such that $r_i \boldsymbol{f}_i \cdot \boldsymbol{e}_i > 0$.

Lemma 5 (Tensor identity)**.** *The following tensor identity holds*

$$\sum_{i=1}^{3} r_i(\boldsymbol{f}_i \otimes \boldsymbol{e}_i) = 3|c|\,\mathbb{I}_3. \tag{5.1}$$

Proof. Let $j \in \{1,2,3\}$ and let $\boldsymbol{m}_j$ be the unit vector of $\boldsymbol{f}_j$. Let us prove that

$$\sum_{i=1}^{3} r_i(\boldsymbol{f}_i \otimes \boldsymbol{e}_i)\,\boldsymbol{m}_j = \sum_{i=1}^{3} r_i(\boldsymbol{e}_i \cdot \boldsymbol{m}_j)\,\boldsymbol{f}_i = 3|c|\,\boldsymbol{m}_j = 3|c|\,\mathbb{I}_3\boldsymbol{m}_j, \tag{5.2}$$

from which the claimed identity follows, since the set of vectors $\boldsymbol{m}_j$ are linearly independent. Observe that if $i \neq j$, then $\boldsymbol{e}_i \cdot \boldsymbol{m}_j = 0$. It follows that

$$\sum_{i=1}^{3} r_i(\boldsymbol{e}_i \cdot \boldsymbol{m}_j)\,\boldsymbol{f}_i = r_j(\boldsymbol{e}_j \cdot \boldsymbol{m}_j)\,\boldsymbol{f}_j = r_j(\boldsymbol{e}_j \cdot \boldsymbol{m}_j)|f_j|\,\boldsymbol{m}_j = 3|c|\,\boldsymbol{m}_j, \tag{5.3}$$

where we have used the expression of the volume of a tetrahedron [40]. □

Now, we define the restriction of the following oriented geometric boundary terms $s_{n,e_{|c}}, l_{n,f_{|c}}$ to a tetrahedron c, which are in bijection with edges $e \in E(n) \cap E(c)$ and faces $f \in F(n) \cap F(c)$, respectively.

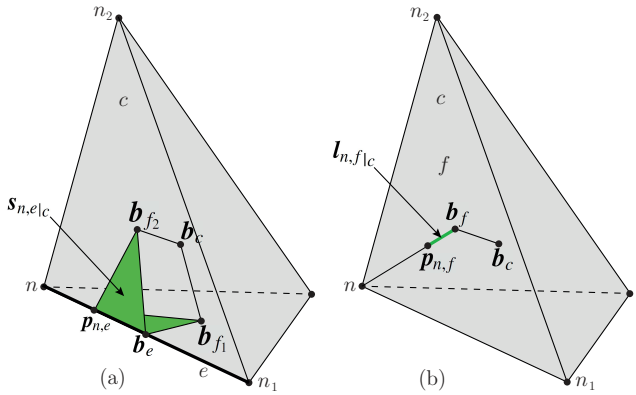

Figure 5.1: (a) Geometric construction of $s_{n,e_{|c}}$. (b) Geometric construction of $l_{n,f_{|c}}$.

We first define $s_{n,e_{|c}}$. It decomposes into the union of two triangles, $t_{n,e_{|c}}^{(1)}$ and $t_{n,e_{|c}}^{(2)}$, associated with the two faces $f_1, f_2 \in F(e)$, see Fig. 5.1(a). The first has the vertices $\{\boldsymbol{b}_e, \boldsymbol{b}_{f_1}, \boldsymbol{p}_{n,e}\}$, the second the vertices $\{\boldsymbol{b}_e, \boldsymbol{b}_{f_2}, \boldsymbol{p}_{n,e}\}$, with node $\boldsymbol{p}_{n,e}$ defined as follows

$$\boldsymbol{p}_{n,e} \coloneqq \frac{3}{4}\boldsymbol{n} + \frac{1}{4}\boldsymbol{n}_1,\ n_1 \in N(e) \setminus \{n\}. \tag{5.4}$$

By construction, both triangles $t_{n,e_{|c}}^{(1)}, t_{n,e_{|c}}^{(2)}$ are adjacent to the restriction of the dual face $\tilde{f}_{e|c}$. Thus, we orient them in such a way that each pair of surfaces $t_{n,e_{|c}}^{(1)}, \tilde{f}_{e|c}$ and $t_{n,e_{|c}}^{(2)}, \tilde{f}_{e|c}$ induce opposite orientations on their common edge [3]. We define $s_{n,e_{|c}}$ to be the surface made of the union of the two triangles $t_{n,e_{|c}}^{(1)}, t_{n,e_{|c}}^{(2)}$. We denote with $\boldsymbol{s}_{n,e_{|c}}$ the face vector associated with $s_{n,e_{|c}}$. It is defined as the sum of the face vectors $\boldsymbol{t}_{n,e_{|c}}^{(1)}, \boldsymbol{t}_{n,e_{|c}}^{(2)}$, that is $\boldsymbol{s}_{n,e_{|c}} \coloneqq \boldsymbol{t}_{n,e_{|c}}^{(1)} + \boldsymbol{t}_{n,e_{|c}}^{(2)}$. We define $\boldsymbol{s}_{n,e}$ to be $\boldsymbol{s}_{n,e} \coloneqq \sum_{c \in C(e)} \boldsymbol{s}_{n,e_{|c}}$.

Next, we define $l_{n,f_{|c}}$ to be the segment joining $\boldsymbol{b}_f$ with node $\boldsymbol{p}_{n,f}$ defined as follows

$$\boldsymbol{p}_{n,f} \coloneqq \frac{1}{2}\boldsymbol{n} + \frac{1}{4}\boldsymbol{n}_1 + \frac{1}{4}\boldsymbol{n}_2,\ n_1, n_2 \in N(f) \setminus \{n\}, \tag{5.5}$$

see Fig. 5.1(b). By construction, $l_{n,f_{|c}}$ is adjacent to the restriction of the dual edge $\tilde{e}_{f_{|c}}$. Thus, we orient $l_{n,f_{|c}}$ in such a way the two segments $l_{n,f_{|c}}, \tilde{e}_{f_{|c}}$ induce opposite orientations on their common node [3]. We denote with $\boldsymbol{l}_{n,f_{|c}}$ the edge vector associated with $l_{n,f_{|c}}$. We define $\boldsymbol{l}_{n,f}$ to be $\boldsymbol{l}_{n,f} \coloneqq \sum_{c \in C(f)} \boldsymbol{l}_{n,f_{|c}}$.

By mimicking the notation introduced at the end of Section 2.1, we denote by $\mathbb{B}_{|\tilde{c}}$ and $\mathbb{L}_{|\tilde{c}}$ the matrices whose rows collect vectors $\boldsymbol{s}_{n,e}$ and $\boldsymbol{l}_{n,f}$ with $e \in E(n)$ and $f \in F(n)$, respectively. Similarly, we denote by $\mathbb{B}_{\tilde{c}_{|c}}$ and $\mathbb{L}_{\tilde{c}_{|c}}$ the matrices whose rows collect vectors $\boldsymbol{s}_{n,e_{|c}}$ and $\boldsymbol{l}_{n,f_{|c}}$ with $e \in E(n) \cap E(c)$ and $f \in F(n) \cap F(c)$, respectively.

The geometric boundary terms satisfy the following identities

$$\frac{r_i \boldsymbol{f}_i}{6} = \tilde{\boldsymbol{f}}_{e_{i|c}} + \boldsymbol{s}_{n,e_{i|c}}, \tag{5.6}$$

$$\frac{r_i \boldsymbol{e}_i}{4} = \tilde{\boldsymbol{e}}_{f_{i|c}} + \boldsymbol{l}_{n,f_{i|c}}, \tag{5.7}$$

which can be proven after trivial algebraic manipulations that we omit. We point out that if we require $s_{n,e_{|c}}, l_{n,f_{|c}}$ to satisfy (5.6), (5.7), the nodes $\boldsymbol{p}_{n,e}, \boldsymbol{p}_{n,f}$ are uniquely identified by (5.4), (5.5).

(5.6), (5.7) are the main geometric identities that together with Lemma 5 allow us to prove the following result.

Lemma 6 (Reconstruction formulas with boundary terms)**.** *We focus attention on a tetrahedron c and on a dual cell $\tilde{c}_n$, dual to a node n of c. The following tensor identities hold*

$$\sum_{e \in E(n) \cap E(c)} \boldsymbol{e}_{|\tilde{c}} \otimes (\tilde{\boldsymbol{f}}_{e|c} + \boldsymbol{s}_{n,e_{|c}}) = |\tilde{c}_{|c}|\, \mathbb{I}_3, \tag{5.8}$$

$$\sum_{f \in F(n) \cap F(c)} \boldsymbol{f}_{|\tilde{c}} \otimes (\tilde{\boldsymbol{e}}_{f_{|c}} + \boldsymbol{l}_{n,f_{|c}}) = |\tilde{c}_{|c}|\, \mathbb{I}_3. \tag{5.9}$$

Proof. Let us apply Lemma 5 to c. We rewrite (5.1) as follows

$$\frac{1}{12} \sum_{i=1}^{3} r_i (\boldsymbol{e}_i \otimes \boldsymbol{f}_i) = |\tilde{c}_{|c}|\, \mathbb{I}_3, \tag{5.10}$$

where we have used the fact that $4|\tilde{c}_{|c}| = |c|$, as a result of barycentric subdivision [40]. Now, using (5.6) it follows that

$$\sum_{i=1}^{3} \frac{\boldsymbol{e}_i}{2} \otimes (\tilde{\boldsymbol{f}}_{e_{i|c}} + \boldsymbol{s}_{n,e_{i|c}}) = \sum_{i=1}^{3} \frac{\boldsymbol{e}_i}{2} \otimes \frac{r_i \boldsymbol{f}_i}{6} = \frac{1}{12} \sum_{i=1}^{3} \boldsymbol{e}_i \otimes (r_i \boldsymbol{f}_i) = |\tilde{c}_{|c}|\, \mathbb{I}_3. \tag{5.11}$$

This proves (5.8), since the length of each edge in $E_{\tilde{c}}$ is half of the length a primal edge.

Similarly, using (5.7) it follows that

$$\sum_{i=1}^{3} \frac{\boldsymbol{f}_i}{3} \otimes (\tilde{\boldsymbol{e}}_{f_{i|c}} + \boldsymbol{l}_{n,f_{i|c}}) = \sum_{i=1}^{3} \frac{\boldsymbol{f}_i}{3} \otimes \frac{r_i \boldsymbol{e}_i}{4} = \frac{1}{12} \sum_{i=1}^{3} \boldsymbol{f}_i \otimes (r_i \boldsymbol{e}_i) = |\tilde{c}_{|c}|\, \mathbb{I}_3. \tag{5.12}$$

This proves (5.9), since the area of each face in $F_{\tilde{c}}$ is a third of the area of a primal face. □

Now we are ready to prove the main result of this section, which is the dual counterpart of Theorem 2. This result is remarkable since dual cells have non-planar geometric elements.

Theorem 4 (Dual reconstruction formulas)**.** *Let $\tilde{c}_n$ be a dual cell. The following two assertions hold.*

(1) If $\tilde{c}$ does not intersect with $\partial\Omega$, then

$$\sum_{e\in E(n)} \boldsymbol{e}_{|\tilde{c}} \otimes \tilde{\boldsymbol{f}}_e = |\tilde{c}|\,\mathbb{I}_3,, \tag{5.13}$$

$$\sum_{f\in F(n)} \boldsymbol{f}_{|\tilde{c}} \otimes \tilde{\boldsymbol{e}}_f = |\tilde{c}|\,\mathbb{I}_3. \tag{5.14}$$

(2) If $\tilde{c}$ does intersect with $\partial\Omega$, then

$$\sum_{e\in E(n)} \boldsymbol{e}_{|\tilde{c}} \otimes (\tilde{\boldsymbol{f}}_e + \boldsymbol{s}_{n,e}) = |\tilde{c}|\,\mathbb{I}_3, \tag{5.15}$$

$$\sum_{f\in F(n)} \boldsymbol{f}_{|\tilde{c}} \otimes (\tilde{\boldsymbol{e}}_f + \boldsymbol{l}_{n,f}) = |\tilde{c}|\,\mathbb{I}_3. \tag{5.16}$$

Proof. Proof of (5.13) and (5.15). Let us apply (5.8) to every tetrahedron $c \in C(n)$. By summing over the set $C(n)$ we have

$$\sum_{c\in C(n)} \left(\sum_{e\in E(n)\cap E(c)} \boldsymbol{e}_{|\tilde{c}} \otimes (\tilde{\boldsymbol{f}}_{e|c} + \boldsymbol{s}_{n,e_{|c}}) \right) = \sum_{c\in C(n)} |\tilde{c}_{|c}|\,\mathbb{I}_3 = |\tilde{c}|\,\mathbb{I}_3. \tag{5.17}$$

We can rewrite (5.17) as a sum over the set $E(n)$ as follows

$$\sum_{c\in C(n)} \left(\sum_{e\in E(n)\cap E(c)} \boldsymbol{e}_{|\tilde{c}} \otimes (\tilde{\boldsymbol{f}}_{e|c} + \boldsymbol{s}_{n,e_{|c}}) \right) = \sum_{e\in E(n)} \boldsymbol{e}_{|\tilde{c}} \otimes \left(\sum_{c\in C(e)} \tilde{\boldsymbol{f}}_{e|c} + \boldsymbol{s}_{n,e_{|c}} \right). \tag{5.18}$$

By using the definition of $\tilde{\boldsymbol{f}}_e$ and $\boldsymbol{s}_{n,e}$, and combining the two expressions (5.17), (5.18) we obtain

$$\sum_{e\in E(n)} \boldsymbol{e}_{|\tilde{c}} \otimes (\tilde{\boldsymbol{f}}_e + \boldsymbol{s}_{n,e}) = |\tilde{c}|\,\mathbb{I}_3. \tag{5.19}$$

This proves (5.15). To prove (5.13), note that if $\tilde{c}$ does not intersect with $\partial\Omega$, then

$$\boldsymbol{s}_{n,e} = \sum_{e\in C(e)} \boldsymbol{s}_{n,e_{|c}} = \boldsymbol{0} \tag{5.20}$$

because each term $\boldsymbol{s}_{n,e_{|c}}$ decomposes into the union of two triangles, and each of them appears in the sum exactly two times and with opposite orientation.

Proof of (5.14) and (5.16). Let us apply (5.9) to every tetrahedron $c \in C(n)$. By summing over the set $C(n)$ we have

$$\sum_{c\in C(n)} \left(\sum_{f\in F(n)\cap F(c)} \boldsymbol{f}_{|\tilde{c}} \otimes (\tilde{\boldsymbol{e}}_{f_{|c}} + \boldsymbol{l}_{n,f_{|c}}) \right) = \sum_{c\in C(n)} |\tilde{c}_{|c}|\,\mathbb{I}_3 = |\tilde{c}|\,\mathbb{I}_3. \tag{5.21}$$

We can rewrite (5.21) as a sum over the set $F(n)$ as follows

$$\sum_{c\in C(n)}\left(\sum_{f\in F(n)\cap F(c)} \boldsymbol{f}_{|\tilde{c}}\otimes(\tilde{\boldsymbol{e}}_{f_{|c}}+\boldsymbol{l}_{n,f_{|c}})\right)=\sum_{f\in F(n)}\boldsymbol{f}_{|\tilde{c}}\otimes\left(\sum_{c\in C(f)}\tilde{\boldsymbol{e}}_{f_{|c}}+\boldsymbol{l}_{n,f_{|c}}\right). \tag{5.22}$$

By using the definition of $\tilde{\boldsymbol{e}}_f$ and $\boldsymbol{l}_{n,f}$, and combining the two expressions (5.21), (5.22) we obtain

$$\sum_{f\in F(n)}\boldsymbol{f}_{|\tilde{c}}\otimes(\tilde{\boldsymbol{e}}_f+\boldsymbol{l}_{n,f})=|\tilde{c}|\,\mathbb{I}_3. \tag{5.23}$$

This proves (5.16). To prove (5.14), note that if $\tilde{c}$ does not intersect with $\partial\Omega$, then

$$\boldsymbol{l}_{n,f}=\sum_{c\in C(f)}\boldsymbol{l}_{n,f_{|c}}=0 \tag{5.24}$$

because each term $\boldsymbol{l}_{n,f_{|c}}$ appears in the sum exactly two times and with opposite orientation. □

The results of Theorem 4 cannot be extended to arbitrary polyhedral grids. To see this, it is sufficient to consider a generic 3D polyhedral grid and check that the relations in Theorem 4 are not satisfied. Simple random examples provide readily a counterexample.

5.1.2 Local inverse mass matrices

Let us focus on a single dual cell $\tilde{c}_n$ where two pairs of constant vector fields are defined, namely $\boldsymbol{u},\boldsymbol{v}$ and $\boldsymbol{w},\boldsymbol{z}$. For the sake of clarity, we suppose that $\tilde{c}$ does not intersect with $\partial\Omega$. Otherwise, it is sufficient to repeat the same reasoning using $\tilde{\mathbb{F}}_{\tilde{c}}+\mathbb{B}_{\tilde{c}}$ and $\tilde{\mathbb{E}}_{\tilde{c}}+\mathbb{L}_{\tilde{c}}$ in place of $\tilde{\mathbb{F}}_{\tilde{c}}$ and $\tilde{\mathbb{E}}_{\tilde{c}}$, respectively. The two pairs are related by two constitutive relations

$$\boldsymbol{v}=\tilde{\mathbb{K}}_{\tilde{c},1}\boldsymbol{u}, \tag{5.25}$$

$$\boldsymbol{z}=\tilde{\mathbb{K}}_{\tilde{c},2}\boldsymbol{w}, \tag{5.26}$$

where $\tilde{\mathbb{K}}_{\tilde{c},1},\tilde{\mathbb{K}}_{\tilde{c},2}$ are two symmetric positive definite matrices of order 3, assumed to be uniform in $\tilde{c}$.

Now, let us introduce the restriction of DoFs of the vector fields to $\tilde{c}$. In particular, we attach DoFs to geometric elements of the primal and dual grid as follows $\boldsymbol{u}_{\tilde{c}}^{\tilde{\mathcal{F}}}=\tilde{\mathbb{F}}_{\tilde{c}}\boldsymbol{u}$, $\boldsymbol{v}_{\tilde{c}}^{\mathcal{E}}=\mathbb{E}_{\tilde{c}}\boldsymbol{v}$, $\boldsymbol{w}_{\tilde{c}}^{\tilde{\mathcal{E}}}=\tilde{\mathbb{E}}_{\tilde{c}}\boldsymbol{w}$, $\boldsymbol{z}_{\tilde{c}}^{\mathcal{F}}=\mathbb{F}_{\tilde{c}}\boldsymbol{z}$.

The local mass matrix $\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{F}}}$ maps DoFs of $\boldsymbol{u}$ attached to dual faces to DoFs of $\boldsymbol{v}$ attached to edges of the primal grid. We say that $\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{F}}}$ is a *consistent mass matrix* if

$$\boldsymbol{v}_{\tilde{c}}^{\mathcal{E}}=\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{F}}}\boldsymbol{u}_{\tilde{c}}^{\tilde{\mathcal{F}}} \tag{5.27}$$

holds exactly for any pair of constant vector fields $\boldsymbol{u},\boldsymbol{v}$ satisfying (5.25).

Similarly, a local mass matrix $\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{E}}}$ maps DoFs of $\boldsymbol{w}$ attached to dual edges to DoFs of $\boldsymbol{z}$ attached to faces of the primal grid. We say that $\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{E}}}$ is a *consistent mass matrix* if

$$\boldsymbol{z}_{\tilde{c}}^{\mathcal{F}} = \mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{E}}} \boldsymbol{w}_{\tilde{c}}^{\tilde{\mathcal{E}}} \tag{5.28}$$

holds exactly for any pair of constant vector fields $\boldsymbol{w}, \boldsymbol{z}$ satisfying (5.26).

An efficient recipe to construct a consistent and symmetric matrices $\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{F}}}, \mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{E}}}$ combines the geometric identities in Theorem 4 with the uniformity of the vector fields.

By applying Theorem 4, and the definitions of DoFs $\boldsymbol{v}_{\tilde{c}}^{\mathcal{E}}, \boldsymbol{u}_{\tilde{c}}^{\tilde{\mathcal{F}}}$ of $\boldsymbol{v}$ and $\boldsymbol{u}$, we have that

$$\boldsymbol{v}_{\tilde{c}}^{\mathcal{E}} = \mathbb{E}_{\tilde{c}} \boldsymbol{v} = \mathbb{E}_{\tilde{c}} \tilde{\mathbb{K}}_{\tilde{c},1} \boldsymbol{u} = \mathbb{E}_{\tilde{c}} \tilde{\mathbb{K}}_{\tilde{c},1} \frac{1}{|\tilde{c}|} (\mathbb{E}_{\tilde{c}}^T \tilde{\mathbb{F}}_{\tilde{c}}) \boldsymbol{u} = \frac{(\mathbb{E}_{\tilde{c}} \tilde{\mathbb{K}}_{\tilde{c},1} \mathbb{E}_{\tilde{c}}^T)}{|\tilde{c}|} \boldsymbol{u}_{\tilde{c}}^{\tilde{\mathcal{F}}}, \tag{5.29}$$

and hence, it follows that a symmetric and consistent matrix $\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{F}}}$ is given by

$$\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{F}}} = \frac{\mathbb{E}_{\tilde{c}} \tilde{\mathbb{K}}_{\tilde{c},1} \mathbb{E}_{\tilde{c}}^T}{|\tilde{c}|}. \tag{5.30}$$

Similarly, by applying Theorem 4, and the definitions of DoFs $\boldsymbol{z}_{\tilde{c}}^{\mathcal{F}}, \boldsymbol{w}_{\tilde{c}}^{\tilde{\mathcal{E}}}$ of $\boldsymbol{z}$ and $\boldsymbol{w}$, we have that

$$\boldsymbol{z}_{\tilde{c}}^{\mathcal{F}} = \mathbb{F}_{\tilde{c}} \boldsymbol{z} = \mathbb{F}_{\tilde{c}} \tilde{\mathbb{K}}_{\tilde{c},2} \boldsymbol{w} = \mathbb{F}_{\tilde{c}} \tilde{\mathbb{K}}_{\tilde{c},2} \frac{1}{|\tilde{c}|} (\mathbb{F}_{\tilde{c}}^T \tilde{\mathbb{E}}_{\tilde{c}}) \boldsymbol{w} = \frac{(\mathbb{F}_{\tilde{c}} \tilde{\mathbb{K}}_{\tilde{c},2} \mathbb{F}_{\tilde{c}}^T)}{|\tilde{c}|} \boldsymbol{w}_{\tilde{c}}^{\tilde{\mathcal{E}}}, \tag{5.31}$$

and hence, it follows that a symmetric and consistent matrix $\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{E}}}$ is given by

$$\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{E}}} = \frac{\mathbb{F}_{\tilde{c}} \tilde{\mathbb{K}}_{\tilde{c},2} \mathbb{F}_{\tilde{c}}^T}{|\tilde{c}|}. \tag{5.32}$$

The matrices $\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{F}}}$ and $\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{E}}}$ defined in (5.30), (5.32) are symmetric and consistent but are not positive definite. To achieve this, we add a stability matrix just as we did for the primal mass matrix.

Lemma 7. *Let m be the cardinality of $F(n)$ or $E(n)$. Let $\tilde{\mathbb{K}}_{\tilde{c}}$ be a symmetric and positive definite matrix of order 3. Let $\alpha = (\alpha_1, \ldots, \alpha_{m-3}) \in (\mathbb{R}^+)^{m-3}$ be any $(m-3)$-upla of positive real numbers and let $\mathbb{D}_\alpha$ be the diagonal matrix whose diagonal entries are $\alpha_1, \ldots, \alpha_{m-3}$. Denote by $\mathbb{W}_{\tilde{c}}^{\tilde{\mathcal{F}}}$ and $\mathbb{W}_{\tilde{c}}^{\tilde{\mathcal{E}}}$ the matrices whose columns form an orthonormal basis for $\mathrm{im}(\tilde{\mathbb{F}}_{\tilde{c}} + \mathbb{B}_{|\tilde{c}})^\perp$ and $\mathrm{im}(\tilde{\mathbb{E}}_{\tilde{c}} + \mathbb{L}_{|\tilde{c}})^\perp$, respectively. Then, the following matrices*

$$\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{F}}} := \frac{1}{|\tilde{c}|} \mathbb{E}_{\tilde{c}} \tilde{\mathbb{K}}_{|\tilde{c}} \mathbb{E}_{\tilde{c}}^T + \mathbb{W}_{|\tilde{c}}^{\tilde{\mathcal{F}}} \mathbb{D}_\alpha \mathbb{W}_{|\tilde{c}}^{\tilde{\mathcal{F}}^T}, \tag{5.33}$$

$$\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{E}}} := \frac{1}{|\tilde{c}|} \mathbb{F}_{\tilde{c}} \tilde{\mathbb{K}}_{|\tilde{c}} \mathbb{F}_{\tilde{c}}^T + \mathbb{W}_{|\tilde{c}}^{\tilde{\mathcal{E}}} \mathbb{D}_\alpha \mathbb{W}_{|\tilde{c}}^{\tilde{\mathcal{E}}^T}, \tag{5.34}$$

are symmetric, consistent and positive definite.

In the general case of a dual grid made of more than one dual cell, the corresponding global mass matrices $\mathbb{M}^{\tilde{\mathcal{F}}}, \mathbb{M}^{\tilde{\mathcal{E}}}$ are obtained by assembling, dual cell by dual cell, the

contributions from the local matrices $\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{F}}}$ and $\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{E}}}$, respectively. However, we need also to take into account the case where $\tilde{c}$ does intersect with $\partial\Omega$. In this case, (5.15), (5.16) tell us that in order to obtain a valid reconstruction we have to attach DoFs of vector fields also to geometric elements defined by $\boldsymbol{s}_{n,e}$ and $\boldsymbol{l}_{n,f}$. In the numerical examples we show how the proposed sparse inverse mass matrices are used in formulations to solve boundary value problems. In this case, the additional DoFs on the boundary are either known because of boundary conditions or not used in the laws enforced by the formulation. We emphasize that the additional DoFs on the boundary appears as a canonical and necessary construction to fulfil the geometrical formulas (5.6) and (5.7).

5.2 Handling material parameter discontinuities inside dual cells

In this section we are interested in the discretization of problems with discontinuous material parameters, which may appear for example when inhomogeneous Poisson or wave propagation problems are considered. In order to apply the setting detailed in Section 5.1.2, a key requirement is that the material parameters are uniform inside dual cells. In this case, discontinuities of the material parameters do not create any complication if the grid is chosen in such a way that the interfaces between different materials are aligned with the boundaries of the dual cells. Hence, in this case, the assignment of the material parameters must be based on the dual cells. This is a common practice in the Finite Volume literature [71]. We observe that such an assignment is correctly defined since dual cells provide a partition of the computational domain.

However, in most cases, it is desired that the material parameters are constant on cells of the primal grid and that they can be arbitrary discontinuous across interfaces between cells. To each $c \in C$ is assigned a possibly different material parameter $\mathbb{K}_{|c}$. As a result, the material parameters of each dual cell are no longer uniform. In order to handle this kind of situations we can use two different strategies which differ on how the constant vector field is reconstructed inside dual cells. The first strategy is to use a *weighted average* of the material property in the scheme. The reconstructed vector field is uniform and the construction technique proposed in Section 5.1.2 can be applied, by using a suitable choice of the material parameter. The second strategy consists of using a *piecewise constant vector field* representation inside dual cells, which accounts for discontinuities of the vector field components due to material parameters discontinuities.

5.2.1 Weighted average

Let us consider a dual cell $\tilde{c}_n \in \tilde{C}$. To motivate the definition of the new material parameter, let us consider a uniform vector field $\boldsymbol{u}$ defined in $\tilde{c}$. Then, we define the vector field $\boldsymbol{v}$ whose restriction to each cell $c \in C(n)$ satisfy $\boldsymbol{v}_{|c} := \mathbb{K}_c \boldsymbol{u}_{|c}$. Thus, $\boldsymbol{v}$ is a piecewise constant vector field that is spatially constant in each $c \in C(n)$. The weighted

average of $\boldsymbol{v}$ over $\tilde{c}$ is

$$\bar{\boldsymbol{v}} = \frac{1}{|\tilde{c}|} \sum_{c \in C(n)} |\tilde{c}_{|c}| \, \mathbb{K}_c \boldsymbol{u} = \left(\frac{1}{|\tilde{c}|} \sum_{c \in C(n)} |\tilde{c}_{|c}| \, \mathbb{K}_c \right) \boldsymbol{u}, \tag{5.35}$$

where we have used the fact that $\boldsymbol{u}$ is uniform in $\tilde{c}$. To each dual cell $\tilde{c}$ we assign a new material parameter defined by a weighted average as follows

$$\tilde{\mathbb{K}}_{\tilde{c}} := \left(\frac{1}{|\tilde{c}|} \sum_{c \in C(n)} |\tilde{c}_{|c}| \, \mathbb{K}_c \right)^{-1}. \tag{5.36}$$

We use (5.36) as the material parameter appearing in the expressions of the local mass matrices in Lemma 7. As a consequence, if the material parameter is constant over the whole computational domain, the same material parameter is assigned to every dual cell.

5.2.2 Piecewise constant vector field

Let us consider a dual cell $\tilde{c}_n \in \tilde{C}$. Since material parameters differ on each $c \in C$, we use a piecewise constant vector field representation that is spatially constant on each $c \in C(n)$. Thus, let us consider two pairs of piecewise constant vector fields defined on $\tilde{c}$, namely $\boldsymbol{u}, \boldsymbol{v}$ and $\boldsymbol{w}, \boldsymbol{z}$. The restriction of the vector fields $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{z}$ to a cell $c \in C(n)$ satisfy the following constitutive relations

$$\boldsymbol{v}_{|c} = \mathbb{K}_{c,1} \boldsymbol{u}_{|c}, \tag{5.37}$$

$$\boldsymbol{z}_{|c} = \mathbb{K}_{c,2} \boldsymbol{w}_{|c}, \tag{5.38}$$

where $\mathbb{K}_{c,1}, \mathbb{K}_{c,2}$ are two symmetric positive definite matrices of order 3. In addition, let us suppose that $\boldsymbol{u}, \boldsymbol{z}$ and $\boldsymbol{v}, \boldsymbol{w}$ preserve the *normal* and *tangential* components over the boundaries of $\tilde{c}_{|c}$, respectively. Thus, the restriction of DoFs associated with the vector fields defined in $\tilde{c}$ are well defined. In particular, we attach DoFs to geometric elements of the primal and dual grid as follows $\boldsymbol{u}_{\tilde{c}}^{\mathcal{F}} = \mathbb{F}_{\tilde{c}} \boldsymbol{u}$, $\boldsymbol{v}_{\tilde{c}}^{\tilde{\mathcal{E}}} = \tilde{\mathbb{E}}_{\tilde{c}} \boldsymbol{v}$, $\boldsymbol{w}_{\tilde{c}}^{\mathcal{E}} = \mathbb{E}_{\tilde{c}} \boldsymbol{w}$, $\boldsymbol{z}_{\tilde{c}}^{\tilde{\mathcal{F}}} = \tilde{\mathbb{F}}_{\tilde{c}} \boldsymbol{z}$.

The idea underlying the construction is to mimic the reasoning proposed in Section 3.2.3 and Section 5.1.2 to construct global mass matrices starting from local mass matrices defined on cells and dual cells, respectively. To be more precise, "local" contributions from matrices $\mathbb{M}_{\tilde{c}_{|c}}^{\mathcal{F}}, \mathbb{M}_{\tilde{c}_{|c}}^{\mathcal{E}}$ restricted to each region $\tilde{c}_{|c}$ are assembled to construct "global" mass matrices $\mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{F}}}, \mathbb{M}_{\tilde{c}}^{\tilde{\mathcal{E}}}$ on the dual cell $\tilde{c}$. The analogy is well defined since the regions $\tilde{c}_{|c}$ provide a partition of the cell $\tilde{c}$ and the material is uniform on each of them. This latter assumption, allows us to apply the construction of a consistent term using the recipe detailed in Section 5.1.2 to each region $\tilde{c}_{|c}$.

Thus, by applying Lemma 6, a *consistent* mass matrix $\mathbb{M}_{\tilde{c}_{|c}}^{\mathcal{E}}$ restricted to $\tilde{c}_{|c}$ is given by

$$\mathbb{M}_{\tilde{c}_{|c}}^{\mathcal{E}} = \frac{(\tilde{\mathbb{F}}_{|\tilde{c}_{|c}} + \mathbb{B}_{\tilde{c}_{|c}}) \mathbb{K}_{|c} (\tilde{\mathbb{F}}_{|\tilde{c}_{|c}} + \mathbb{B}_{\tilde{c}_{|c}})^T}{|\tilde{c}_{|c}|}. \tag{5.39}$$

A local mass matrix $\mathbb{M}^{\mathcal{E}}_{\tilde{c}}$ is obtained by assembling the local contributions of each $\mathbb{M}^{\mathcal{E}}_{\tilde{c}_{|c}}$. $\mathbb{M}^{\mathcal{E}}_{\tilde{c}}$ constructed in this way is symmetric and positive definite since each matrix (5.39) has rank 3. Moreover, $\mathbb{M}^{\mathcal{E}}_{\tilde{c}}$ satisfy the following *piecewise consistency property*, that is

$$\boldsymbol{z}^{\tilde{\mathcal{F}}}_{\tilde{c}} = \mathbb{M}^{\mathcal{E}}_{\tilde{c}} \boldsymbol{w}^{\mathcal{E}}_{\tilde{c}} \tag{5.40}$$

holds exactly for any pair of piecewise constant vector fields $\boldsymbol{w}, \boldsymbol{z}$ satisfying (5.38). This follows from the fact that the normal component of $\boldsymbol{z}$ is preserved over the boundaries of $\tilde{c}_{|c}$ and hence the geometric boundary terms cancel out, as shown in the proof of Theorem 4. We observe that the piecewise consistency property implies the consistency property (5.27). A local mass matrix $\mathbb{M}^{\tilde{\mathcal{F}}}_{\tilde{c}}$ is obtained by computing the algebraic inverse of $\mathbb{M}^{\mathcal{E}}_{\tilde{c}}$,

$$\mathbb{M}^{\tilde{\mathcal{F}}}_{\tilde{c}} = \mathbb{M}^{\mathcal{E}^{-1}}_{\tilde{c}}. \tag{5.41}$$

In a similar way, by applying Lemma 6, a *consistent* mass matrix $\mathbb{M}^{\mathcal{F}}_{\tilde{c}_{|c}}$ restricted to $\tilde{c}_{|c}$ is given by

$$\mathbb{M}^{\mathcal{F}}_{\tilde{c}_{|c}} = \frac{(\tilde{\mathbb{E}}_{|\tilde{c}_{|c}} + \mathbb{L}_{\tilde{c}_{|c}}) \mathbb{K}_{|c} (\tilde{\mathbb{E}}_{|\tilde{c}_{|c}} + \mathbb{L}_{\tilde{c}_{|c}})^T}{|\tilde{c}_{|c}|}. \tag{5.42}$$

A local mass matrix $\mathbb{M}^{\mathcal{F}}_{\tilde{c}}$ is obtained by assembling the local contributions of each $\mathbb{M}^{\mathcal{F}}_{\tilde{c}_{|c}}$. $\mathbb{M}^{\mathcal{F}}_{\tilde{c}}$ constructed in this way is symmetric and positive definite since each matrix (5.42) has rank 3. Moreover, $\mathbb{M}^{\mathcal{F}}_{\tilde{c}}$ satisfy the following *piecewise consistency property*, that is

$$\boldsymbol{v}^{\tilde{\mathcal{E}}}_{\tilde{c}} = \mathbb{M}^{\mathcal{F}}_{\tilde{c}} \boldsymbol{u}^{\mathcal{F}}_{\tilde{c}} \tag{5.43}$$

holds exactly for any pair of piecewise constant vector fields $\boldsymbol{v}, \boldsymbol{u}$ satisfying (5.37). This follows from the fact that the tangential component of $\boldsymbol{v}$ is preserved over the boundaries of $\tilde{c}_{|c}$ and hence the geometric boundary terms cancel out, as shown in the proof of Theorem 4. We observe that the piecewise consistency property implies the consistency property (5.28). Then, a local mass matrix $\mathbb{M}^{\tilde{\mathcal{E}}}_{\tilde{c}}$ is obtained by computing the algebraic inverse of $\mathbb{M}^{\mathcal{E}}_{\tilde{c}}$,

$$\mathbb{M}^{\tilde{\mathcal{E}}}_{\tilde{c}} = \mathbb{M}^{\mathcal{F}^{-1}}_{\tilde{c}}. \tag{5.44}$$

We observe that even if the material parameter is uniform inside $\tilde{c}$, using this approach it is in any case necessary to compute the inverses of matrices $\mathbb{M}^{\mathcal{E}}_{\tilde{c}}, \mathbb{M}^{\mathcal{F}}_{\tilde{c}}$. Instead, in Section 5.1.2, explicit expressions of $\mathbb{M}^{\tilde{\mathcal{F}}}_{\tilde{c}}, \mathbb{M}^{\tilde{\mathcal{E}}}_{\tilde{c}}$ are derived.

5.2.3 Hybrid approach to handle material discontinuities inside dual cells

The proposed approach to handle discontinuities of the material parameter inside dual cells is the following. First, if material parameters are uniform inside a dual cell we can

apply the construction detailed in Section 5.1.2. Otherwise, we apply one among the approaches described in Section 5.2.1, Section 5.2.2. We point out that these constructions are applied only to cells which lie at the intersection between regions enclosing different materials. Thus, the required computational effort to compute local inverses in the approach described in Section 5.2.2 is negligible in practice.

An important remark is that if we want to satisfy exactly a *multi-material patch test* (see Section 5.3), it is necessary in our hybrid approach to resort to the construction proposed in Section 5.2.2. This is because, being the reconstructed field piecewise constant, it can represent exactly the discontinuities of the vector field components that appear at interfaces between different materials. Instead, in Section 5.2.1, the reconstructed field is constant.

5.3 Numerical results

Beside all the other applications of sparse inverse mass matrices, to validate the construction proposed in this paper we concentrate on the solution of Poisson boundary value problems formulated with one unknown per element.

We first verified that the proposed technique is able to pass a patch test on grids made by general tetrahedra. Then, a stationary current conduction problem representing a square resistor as in Section 2.8.2 is solved.

5.3.1 Formulations with one unknown per element

There are other complementary-dual formulations, much less explored in the literature, that feature one unknown per element [71], [77], [73]. As pointed out in [77], there has been a long-standing interest to reduce the system to one potential value per element, to reduce unknowns and obtain a positive-definite system.

In [77] two approaches are proposed. The first one requires the solution of local problems on patches of elements to produce local flux expressions. The second one is based on the use of the algebraic inverse of the mass matrix and has been deemed as impossible in [77] because the "the inverse of a mass matrix is not sparse". We follow the latter approach enabled by the efficient construction of sparse inverse of mass matrices proposed for the first time in this paper. To show the details, let us consider the dual scalar potential formulations *DSP* obtained by writing the problem in the geometric framework [45] as

$$\mathbb{C}^T \boldsymbol{E}^{\tilde{\mathcal{E}}} = \mathbf{0} \tag{5.45}$$

$$\mathbb{D}\boldsymbol{J}^{\mathcal{F}} = \mathbf{0} \tag{5.46}$$

$$\boldsymbol{J}^{\mathcal{F}} = \mathbb{M}^{\tilde{\mathcal{E}}} \boldsymbol{E}^{\tilde{\mathcal{E}}}, \tag{5.47}$$

where $\mathbb{D}$ is the cell-face incidence matrix, $\mathbb{C}$ is the face-edge incidence matrix and $\mathbb{M}^{\tilde{\mathcal{E}}}$ is the inverse mass matrix of $\mathbb{M}^{\mathcal{E}}$, which maps DoFs of $\boldsymbol{E}$ attached to dual edges to DoFs of $\boldsymbol{J}$ attached to primal faces. Thus, we can interpret $\mathbb{M}^{\tilde{\mathcal{E}}}$ as a *dual conductance matrix*. To implicitly satisfy (5.45), the scalar potential $U^{\tilde{\mathcal{N}}}$ in the dual nodes is introduced

through

$$\boldsymbol{E}^{\tilde{\mathcal{E}}} = -\mathbb{D}^T U^{\tilde{\mathcal{N}}} + \boldsymbol{E}_s^{\tilde{\mathcal{E}}}, \tag{5.48}$$

where $\boldsymbol{E}_s^{\tilde{\mathcal{E}}}$ is introduced to take into account Dirichlet b.c., so that $\mathbb{C}^T \boldsymbol{E}_s^{\tilde{\mathcal{E}}} = \mathbf{0}$ [74], [48]. Its construction is straightforward and detailed in [48]. By substituting (5.48) and (5.47) into (5.46), one gets

$$(\mathbb{D}\mathbb{M}^{\tilde{\mathcal{E}}}\mathbb{D}^T) U^{\tilde{\mathcal{N}}} = \mathbb{D}\mathbb{M}^{\tilde{\mathcal{E}}} \boldsymbol{E}_s^{\tilde{\mathcal{E}}}, \tag{5.49}$$

having the scalar potential on dual nodes as unknowns.

A symmetric and positive-definite $\mathbb{M}^{\tilde{\mathcal{E}}}$ may be computed as $\mathbb{M}^{\tilde{\mathcal{E}}} = \mathbb{M}^{\mathcal{F}^{-1}}$, where the Raviart–Thomas mass matrix $\mathbb{M}^{\mathcal{F}}$ can be interpreted as a *resistance mass matrix*. Yet, the obtained matrix $\mathbb{M}^{\tilde{\mathcal{E}}}$ using this recipe is fully populated, hence not usable in practice [77]. Another solution is enabled by the geometric construction of $\mathbb{M}^{\tilde{\mathcal{E}}}$ by using (5.34), which produces a symmetric and positive definite matrix for any tetrahedral grid given as input. We point out that the additional voltages DoFs on the boundary are zero on the electrodes, because electrodes are equipotential by hypothesis. The other additional voltages DoFs on the boundary are not needed because the boundary conditions are applied on the currents of the corresponding boundary faces.

We remark that another formulation arises that we may call dual vector potential DVP, which is the dual of the VP formulation. I.e. one obtains a system in which the unknowns are the DoFs of the vector potential attached on dual edges. This formulation is not presented in detail since it has been verified that it is far to be competitive with the others in terms of computational efficiency.

5.3.2 Classical patch tests

The patch tests are Poisson problems designed in such a way that their analytical solution is uniform. By interpreting the Poisson problems as direct current conduction problems, a simple way to produce a patch test is to consider a planar resistor, as described in Fig. 5.2(a). It has been verified that the SP, $VP = MH$, and DSP formulations produce the analytical solution as represented in Fig. 5.2(b).

5.3.3 Multi-material patch tests

In the first multi-material patch test, two different materials with different material properties are placed in *series*. From the result represented in Fig. 5.3(b-c), we conclude that the tangential component of the electric field $\boldsymbol{E}$ is conserved across the material interface, whereas the tangential component of the current density field $\boldsymbol{J}$ jumps. This result holds for all formulations.

In the second multi-material patch test, two different materials with different material properties are placed in *parallel*. From the result represented in Fig. 5.3(e-f), we conclude that the normal component of the current density $\boldsymbol{J}$ is conserved across the material interface, whereas the normal component of the electric field $\boldsymbol{E}$ jumps. This result holds for all formulations.

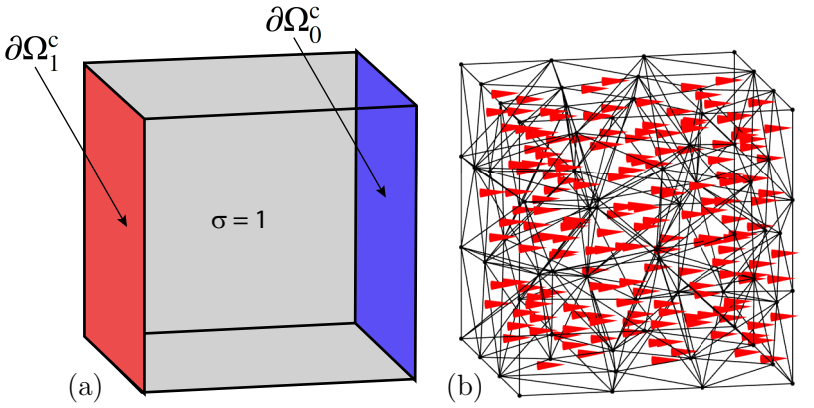

Figure 5.2: (a) The definition of a patch test as the solution of an electric conduction problem inside a planar resistor. (b) All mentioned formulations are able to retrieve the analytical solution up to machine precision or iterative solver tolerance.

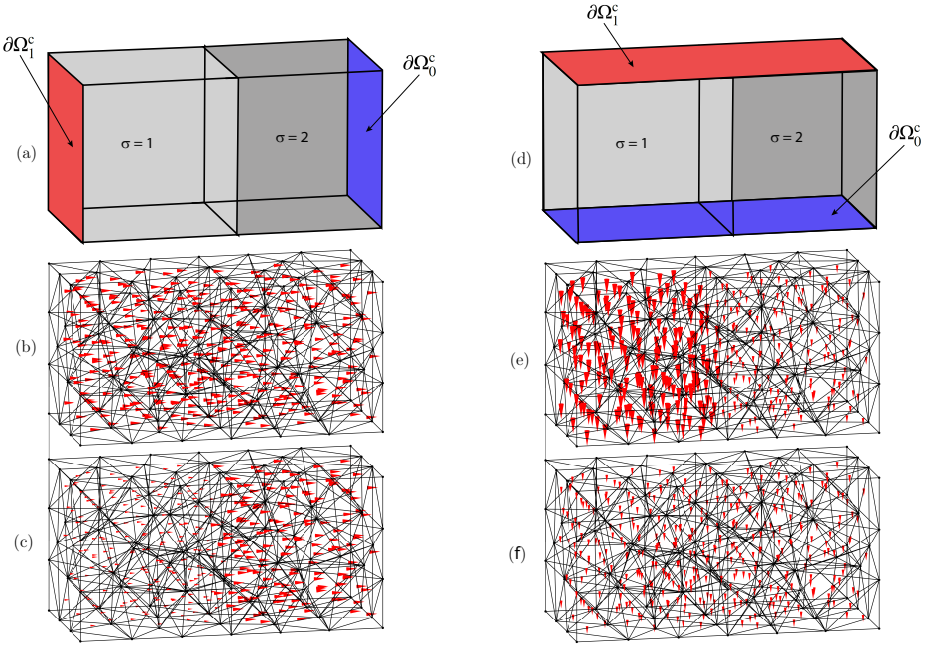

Figure 5.3: (a) The *series* multi-material patch test. (b) Electrostatic field $\boldsymbol{E}$ obtained with the series. (c) Current density field $\boldsymbol{J}$ obtained with the series. (d) The *parallel* multi-material patch test. (e) Electric field $\boldsymbol{E}$ obtained with the parallel. (f) Current density field $\boldsymbol{J}$ obtained with the parallel.

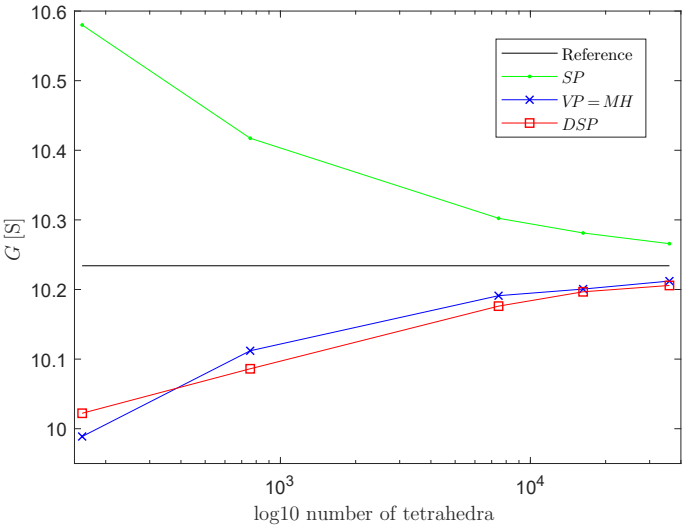

Figure 5.4: Results for the square resistor benchmark.

5.3.4 Square resistor benchmark

We recall that the VP and MH formulations produce the same results given that they are algebraically equivalent [48]. The conductance of the square resistor has been evaluated by the SP, $VP = MH$, and DSP formulations on refined grids. The results are collected in Fig. 5.4. First, it is interesting to note that the results relative to the scalar potential SP and to the vector potential VP or mixed-hybrid MH formulations provide, respectively, the upper and lower bounds for the conductance [52], [46]. This property is called *complementarity* in the computational electromagnetics community [7], [48].

5.4 Conclusions

In this chapter, we have proposed a new construction of sparse inverse edge and face mass matrices for arbitrary tetrahedral grids. The proposed framework unifies the construction of both mass matrices and their inverses and is based on novel geometric reconstruction formulas, from which, local mass matrices are defined as the sum of a consistent and a stabilization term. The consistent term is defined geometrically and explicitly, thus providing a sensible speedup and an easier implementation, whereas the stabilization term is constructed according to a well-established design strategy. The resulting expressions for both types of local mass matrices are highly symmetric, thus highlighting the duality relationship between the pair of grids. Global mass matrices are then constructed using a standard assembling process. In this way we obtain a sparse matrix representation also for the inverse mass matrix. This avoids explicitly computing the algebraic inverse of the mass matrix which is found to be typically dense. A key

aspect of the proposed method is that being based on geometric elements of the barycentric dual grid, it can be applied to arbitrary tetrahedral grids. Two different strategies are proposed to deal with the case of problems involving discontinuous material parameters for the case where the interfaces are aligned with the tetrahedral grid, and an hybrid approach between the two is employed in practical applications. We tested the newly derived inverse mass matrices by solving a Poisson problem and we verified that patch test is passed, even when material discontinuities are present. A real case problem involving a square resistor benchmark is analyzed. Given that reconstruction formulas in each dual cell provide a first-order approximation of a vector field, we expect that first-order error estimates can be derived for the numerical scheme.

6

Computing discrete vector potentials

In this chapter we present a novel framework to solve the discrete version of the following potential problem: determine a vector field with specified curl, i.e. a vector potential of a given vector field.

Let us consider a bounded domain Ω of $\mathbb{R}^3$ as in Section 2.1, where we assume that Ω is topologically trivial, i.e., it is homeomorphic to a closed 3-ball (or, equivalently, to a cube). Let $\boldsymbol{J}$ be a vector field defined in Ω. A vector field $\boldsymbol{H}$ is a vector potential of $\boldsymbol{J}$, namely,

$$\boldsymbol{\nabla} \times \boldsymbol{H} = \boldsymbol{J}, \tag{6.1}$$

if and only if the divergence of $\boldsymbol{J}$ is zero and its flux is vanishing across all the (but one) connected components of $\partial\Omega$. In our case, $\partial\Omega$ is connected, so the flux condition on $\boldsymbol{J}$ can be omitted since it is automatically verified.

By considering the mimetic discretization of (6.1), the discrete vector potential problem is to find a discrete vector potential $\boldsymbol{H}^{\mathcal{E}} \in \mathcal{E}$ such that

$$\mathbb{C}\boldsymbol{H}^{\mathcal{E}} = \boldsymbol{J}^{\mathcal{F}}. \tag{6.2}$$

In intimate analogy with the continuous case, a necessary and sufficient condition for the existence of a discrete vector potential $\boldsymbol{H}^{\mathcal{E}}$ is that array $\boldsymbol{J}^{\mathcal{F}}$ represents a *discrete solenoidal* vector field, i.e. it verifies $\mathbb{D}\boldsymbol{J}^{\mathcal{F}} = \boldsymbol{0}$.

Our main motivation to solve problem (6.2) stems from the fact that this algorithmic primitive is an enabling technology for solving many problems arising in computational physics, from electromagnetism to elasticity and fluid mechanics [78]. First, it can be used to solve the vector laplacian in nearly linear time [79]. The idea is that, instead formulating the vector laplacian by using a vector potential, the scalar potential [80] or the mixed-hybrid [81] formulations could be used instead, which produce linear systems that can be solved in nearly linear time by using algebraic multigrid methods. Second, inverse discrete curl is at the root of efficient algorithms to compute a cohomology basis

and source fields for solving magnetostatics and eddy current problems by mimetic or finite element methods [82, 83, 84, 85]. We think $\boldsymbol{H}^{\mathcal{E}}$ as a discrete magnetic field and $\boldsymbol{J}^{\mathcal{F}}$ as a discrete current; then (6.2) expresses the so-called discrete Ampère's law. In the electromagnetic literature, discrete fields $\boldsymbol{H}^{\mathcal{E}}$ satisfying (6.2) are often called *source fields*. A different application in computational electromagnetics is to find a magnetic vector potential from a magnetic induction field [86].

There are two analogous discrete potential problems: the problem of determining a scalar potential with assigned gradient and a vector field with assigned divergence. However, as we will see in our discourse, and as been already pointed out in literature [87], these two problems are less challenging than problem (6.2), since they are easily solved in linear worst-case complexity using standard spanning tree constructions.

Currently, The most efficient methods to solve linear system (6.2) are based on the so-called *tree-cotree decomposition*, whose basic idea was introduced in the works [80, 82]. However, tree-cotree techniques suffer from well-known termination problems [88, 87]. In Section 6.3 we show that termination problems of tree-cotree decomposition techniques are equivalent to the problem of deciding whether a given mesh has a topological property called *collapsibility*. It is known that there exist 3-balls which are not collapsible [89]. Moreover, the problem of deciding whether an arbitrary (embedded or not) simplicial complex is collapsible is NP-complete [90]. This limits deeply the theoretical usage of tree-cotree techniques. In Section 6.4 we propose a new algorithm based on *discrete Morse theory* that is able to deal with non-collapsible complexes while showing the same linear computational complexity of tree-cotree techniques.

6.1 Notation

In what follows, the concept of a partition of a given index set will play an important role. Let I be a finite index set. A partition of I is a family of disjoint subsets $\{I_1, \ldots, I_p\}$ of I such that $\bigcup_{l=1}^{p} I_l = I$. The *subvector of* $\boldsymbol{v} = (v_i)_{i \in I} \in \mathbb{R}^I$ *induced by* I_l is

$$\boldsymbol{v}_{|I_l} \coloneqq (v_i)_{i \in I_l}, \tag{6.3}$$

for $l \in \{1, \ldots, p\}$. A representation of the vector $\boldsymbol{v}$ as a block vector is given by

$$\boldsymbol{v} = (\boldsymbol{v}_{|I_l})_{l \in \{1,\ldots,p\}}. \tag{6.4}$$

Let us consider a product of index sets I and J. We will need a corresponding notion of (6.3) for a matrix whose entries are indexed by elements in $I \times J$. Let us consider partitions of I and J as $\{I_1, \ldots, I_p\}$ and $\{J_1, \ldots, J_q\}$, respectively. We have a corresponding partition of $I \times J$ as a family of disjoint subsets $\{O_1, \ldots, O_n\}$ such that $O_i = I_l \times J_m$ for some $l \in \{1, \ldots, p\}$, $m \in \{1, \ldots, q\}$ and $I \times J = \bigcup_{i=1}^{n} O_i$. The *submatrix of* $\mathbb{A} = (\mathbb{A}_{i,j})_{i \in I, j \in J} \in \mathbb{R}^{I \times J}$ *induced by* $I_l \times J_m$ is

$$\mathbb{A}_{|I_l \times J_m} \coloneqq (\mathbb{A}_{i,j})_{i \in I_l, j \in J_m} \in \mathbb{R}^{I_l \times J_m}. \tag{6.5}$$

A representation of the matrix $\mathbb{A}$ as a block matrix is given by

$$\mathbb{A} = (\mathbb{A}_{|I_l \times J_m})_{l \in \{1,\ldots,p\}, m \in \{1,\ldots,q\}}. \tag{6.6}$$

6.2 Discrete Morse Theory

The underlying principle of our construction follows an ad hoc reformulation of Forman's Discrete Morse theory [91] given by Kozlov [92], where the basic tool is a combinatorial object called acyclic matching. Several special cases of our construction have already appeared in literature. We present a formulation of discrete Morse theory adapted to our purposes, along with smaller, more illustrative instances, which will provide insights on the structure of our algorithm.

6.2.1 Informal introduction to discrete Morse theory

The first concept is that of *elementary collapse*. One may view discrete Morse theory as a generalization of the theory of simplicial collapses. The concept of collapse, originated in Whitehead's work [93], provides a combinatorial operation that is analogous to the continuous operation called deformation retraction, i.e., the operation of continuously shrinking a topological space to a subset. More specifically, let (σ, τ) be a pair of cells such that $\sigma \subset \tau$ and $\dim \sigma = \dim \tau - 1$. For this pair, to induce an elementary collapse, we require τ to be a cell of maximal dimension in K and the only one cell of K containing σ; we refer to this as saying that the pair (σ, τ) is *free* in K. Equivalently, we also say that σ is free in K; see Fig. 6.1(a). We say that K *collapses to* L if one could get from K to L in a finite sequence of elementary collapses. If K is equivalent to a single vertex, then we say that K is *collapsible*; in this case there exists a sequence of elementary collapses leaving a single node.

Dropping the uniqueness condition on τ, we obtain what we refer to as an *internal collapse*, see Fig. 6.1(b).

Geometrically, in both cases, we obtain a collapse of the pair (σ, τ) by contracting the whole cell τ onto $\partial\tau \setminus \sigma$.

In intimate analogy with elementary collapses, we may combine many internal collapses to form a sequence of internal collapses, again without affecting the homotopy type.

We thus have a family of pairs $\{(\sigma_1, \tau_1), \ldots, (\sigma_n, \tau_n)\}$ to be collapsed, in this order. One may view the set of all such pairs as a *matching* on K. Accordingly, we refer to cells contained in some pair as *matched* and other cells as *unmatched* or *critical* (with respect to the matching).

Let $K^{(i)}$ be the resulting cell complex after the first i collapses. For the pairs to form a sequence of elementary collapses, we require that each new pair (σ_i, τ_i) is free in $K^{(i-1)}$. For generic collapses we apply the same requirement, except that we restrict our attention to the family of matched cells. Specifically, we do not require (σ_i, τ_i) to be free in $K^{(i-1)}$, but τ_i must be the only matched cell of $K^{(i-1)}$ containing σ_i. Equivalently, for each i, we should have that σ_i is not contained in $\tau_{i+1}, \ldots, \tau_n$ for $i \in \{1, \ldots, n-1\}$. We refer to a matching on K admitting an ordering with this property as *acyclic*. We formalize all this concepts in Section 6.2.2.

The main theorem of discrete Morse theory states that an acyclic matching induces a homotopy equivalence between K and the so-called Morse complex, a cell complex formed by critical cells only [92] (Theorem 11.13 (b)).

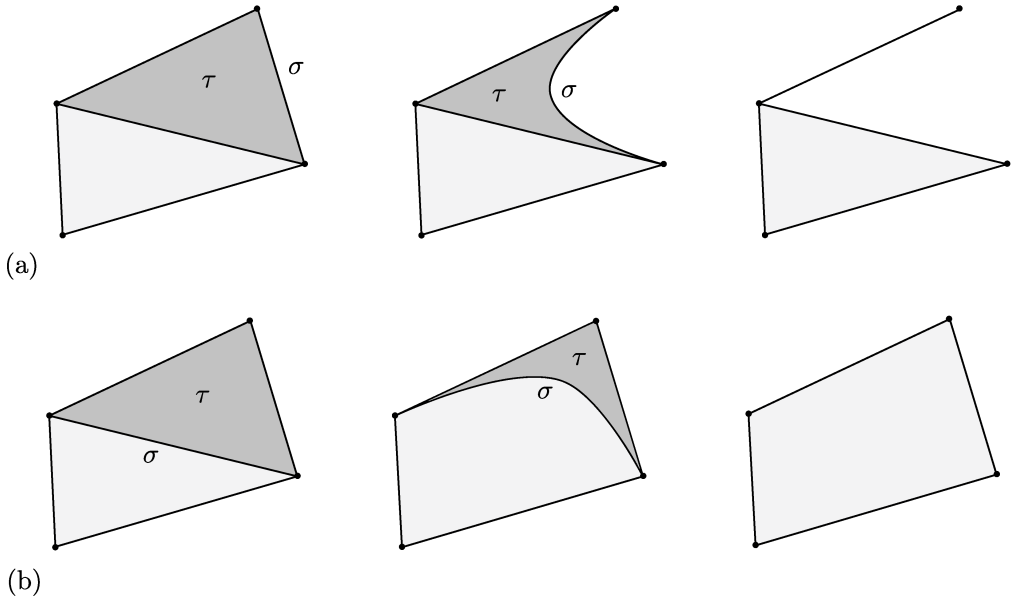

Figure 6.1: (a) Elementary collapse of free pair (σ, τ). (b) Internal collapse of pair (σ, τ); the resulting cell complex is not more simplicial.

6.2.2 Acyclic matchings

We start our exposition by examining acyclic matchings from a purely combinatorial point of view without any reference to topology. Indeed, our interest is in using discrete Morse theory to develop a fast algorithm for the solution of linear system (6.2) to be applied to cell complexes arising from experimental or numerical meshes of real case problems. We give a specific version of combinatorial discrete Morse theory by Kozlov [92] that is adapted to our purposes.

For any $\sigma, \tau \in \widehat{\mathcal{B}}_k$, define $\langle \sigma \, , \tau \rangle$ to be 1 if $\sigma = \tau$ and 0 otherwise. Extend, by linearity, $\langle \cdot \, , \cdot \rangle$ to a scalar product on $C_k(K)$. Note that we can identify the scalar product $\langle \cdot \, , \cdot \rangle$ with the duality product between chains and cochains in Section 6.1 via the isomorphism $\phi_k : C_k(K) \to C^k(K)$ in (2.41), i.e. $\langle \sigma \, , \tau \rangle = \langle \phi_k(\sigma) \, , \tau \rangle = \phi_k(\sigma)(\tau)$.

Let us consider the chain complex $(C^*, \mathcal{B})$ with basis $\mathcal{B}$. We define a relation $\prec$ on $\mathcal{B}$ as follows. Given distinct basis elements $\sigma \in \mathcal{B}_k$ and $\tau \in \mathcal{B}_{k+1}$,

$$\sigma \prec \tau \iff \langle \sigma \, , \partial_{k+1} \tau \rangle \neq 0. \tag{6.7}$$

If $\sigma \prec \tau$, then we say that σ and τ are *incident*.

We introduce the *boundary set* and *coboundary set* of $\sigma \in \mathcal{B}_k$ as

$$\mathrm{bd}_{\mathcal{B}}(\sigma) \coloneqq \{ \rho \in \mathcal{B}_{k-1} \mid \rho \prec \sigma \}, \tag{6.8}$$

and

$$\mathrm{cobd}_{\mathcal{B}}(\sigma) \coloneqq \{\, \rho \in \mathcal{B}_{k+1} \mid \sigma \prec \rho \}, \tag{6.9}$$

respectively.

Let $\sigma \in \mathcal{B}_k$. If the cardinality of $\mathrm{cobd}_{\mathcal{B}}(\sigma)$ is one, then we say that σ is *free*. In this case, there exists a unique basis element τ such that $\sigma \prec \tau$, and we also say that the pair (σ, τ) is free. If the cardinality of $\mathrm{cobd}_{\mathcal{B}}(\sigma)$ is greater than one, then we say that σ in *internal*. In this case, if $\tau \in \mathrm{cobd}_{\mathcal{B}}(\sigma)$, then we also say that the pair (σ, τ) is internal.

Definition 6.2.1 (Matching, acyclic matching)**.** A *matching* $\mathcal{M}$ on $\mathcal{B}$ is a family of pairs $\{(\sigma, \tau)\}$ with $\sigma, \tau \in \mathcal{B}$ such that:

1. $(\sigma, \tau) \in \mathcal{M}$ implies $\sigma \prec \tau$.
2. each $\sigma \in \mathcal{B}$ is the first component of at most one pair (σ, τ) in $\mathcal{M}$.

A matching $\mathcal{M}$ is called *acyclic* if there does not exist a cycle

$$\tau_1 \succ \sigma_1 \prec \tau_2 \succ \cdots \prec \tau_h \succ \sigma_h \prec \tau_1, \tag{6.10}$$

with $h \geq 2$, $(\sigma_i, \tau_i) \in \mathcal{M}$ for all $i \in \{1, \ldots, h\}$ and all $\tau_i \in \mathcal{B}$ being distinct.

A matching $\mathcal{M}_k$ of k-chains on $\mathcal{B}$ is a matching such that if $(\sigma, \tau) \in \mathcal{M}_k$ then $\sigma \in \mathcal{B}_k$.

The following result is a reformulation of Theorem 11.2 in [92] by Kozlov. It describes the crucial combinatorial property that characterizes acyclic matchings. Its proof can be obtained by a suitable adaption of the mentioned Theorem 11.2, see pages 181-182 of [92].

Theorem 5. *A matching $\mathcal{M}$ on $\mathcal{B}$ is acyclic if and only if there exists a total order of pairs of $\mathcal{M}$ as $\{(\sigma_1, \tau_1), \ldots, (\sigma_n, \tau_n)\}$ such that, for every $i \in \{1, \ldots, n-1\}$, σ_i is not incident to any $\tau_{i+1}, \ldots, \tau_n$.*

In what follows, we will write an acyclic matching $\mathcal{M}$ as a sequence $\{(\sigma_1, \tau_1), \ldots, (\sigma_n, \tau_n)\}$, where it is understood that the total order is chosen according to Theorem 5.

Let $\mathcal{M}$ be a matching on $\mathcal{B}$. We say that a basis element in $\mathcal{B}$ is *matched* (with respect to $\mathcal{M}$) if it is contained in some pair in $\mathcal{M}$ (both as first or second component of the pair), otherwise, we say that it is *unmatched* or *critical* (with respect to $\mathcal{M}$).

We denote by $\mathcal{U}_k$ the set of all $\tau \in \mathcal{B}_k$ such that τ is matched with some $\sigma \in \mathcal{B}_{k-1}$. Similarly, we denote by $\mathcal{D}_k$ the set of all $\sigma \in \mathcal{B}_k$ such that σ is matched with some $\tau \in \mathcal{B}_{k+1}$. Given $\mathcal{D}_k$ and $\mathcal{U}_k$, there is a corresponding set of *critical* k-chains

$$\mathcal{A}_k \coloneqq \mathcal{B}_k \setminus (\mathcal{D}_k \cup \mathcal{U}_k). \tag{6.11}$$

Finally, we set $\mathcal{U} \coloneqq \bigcup_k \mathcal{U}_k$, $\mathcal{D} \coloneqq \bigcup_k \mathcal{D}_k$ and $\mathcal{A} \coloneqq \bigcup_k \mathcal{A}_k$. It easy to see that the sets $\mathcal{U}, \mathcal{D}, \mathcal{A}$ provide a partition of $\mathcal{B}$

$$\mathcal{B} = \mathcal{U} \cup \mathcal{D} \cup \mathcal{A}. \tag{6.12}$$

Our definition of a matching is related to the presentation of the combinatorial Morse theory of Forman [91] and in particular the more recent formulation given by

Kozlov [92]. In earlier presentations, elements in $\mathcal{U}$ and $\mathcal{D}$ are not explicitly introduced since what is important is only the bijective pairing between their elements. Instead, in our setting they will play a fundamental role since we use discrete Morse theory from a purely combinatorial point of view and elements in $\mathcal{U}$ and $\mathcal{D}$ will be used to select suitable submatrices. The set of critical elements is present also in classical discrete Morse theory. The set of critical elements will play a fundamental role in Algorithm 3, where critical elements become the new input for subsequent iterations. In Algorithm 3, it is also essential to be able to express new basis elements as a function of the previous ones. Hence, we need to keep track of basis structure at each iteration.

An important difference with classical discrete Morse theory is that in Forman [91] a new chain complex, the so-called Morse complex, is constructed. Our version of discrete Morse theory operates directly on basis elements by performing elementary operations on it. This is needed to describe how incidence matrices transform in the new bases due to our combinatorial operations.

6.2.3 Basis transformations associated with an acyclic matching

We begin by considering simple examples to develop geometric intuition behind the general definitions. When simplifying a cell complex, the effect of a collapse is that of changing the structure of the basis $\mathcal{B}$, by performing elementary operations on it. There are three elementary operations to obtain a new basis from a previous one. If $\mathcal{B}_k = \{\dots, \xi_i, \xi_j, \dots\}$ is a basis of $C_k(K)$, then a new basis may be obtained by

1. Exchanging elements ξ_i and ξ_j.
2. Multiplying ξ_i by -1.
3. Replacing ξ_j by $\xi_j + q\xi_i$ with $q \in \mathbb{R}$.

Let us consider a prototype example of a simplicial complex K in Fig. 6.2. The set of all 1-chains is generated by the canonical basis $\widehat{\mathcal{B}}_1 = \{e_1, e_2, e_3, e_4, e_5\}$ and the set of all 2-chains by the canonical basis $\widehat{\mathcal{B}}_2 = \{f_1, f_2\}$. In Fig. 6.2 edges e_1, e_2, e_3, e_4 are free since each of them is incident to exactly one face in K. Instead, edge e_5 is internal since is the common edge of f_1 and f_2. Depending on whether a collapse is elementary or internal we have corresponding elementary operations on $\mathcal{B}$.

Example 3 (Elementary collapse)**.** *Let us consider the elementary collapse of the free pair (e_1, f_2). The obtained cell complex in Fig. 6.2(b) is generated by the set of critical basis elements $\mathcal{A}$. In fact, we have $\mathcal{A}_1 = \{e_2, e_3, e_4, e_5\}$ and $\mathcal{A}_2 = \{f_1\}$. We get a new basis $\mathcal{B}'$ of K as $\mathcal{B}'_1 = \mathcal{D}_1 \cup \mathcal{A}_1$ and $\mathcal{B}'_2 = \mathcal{U}_2 \cup \mathcal{A}_2$, where $\mathcal{D}_1 = \{e_1\}$ and $\mathcal{U}_2 = \{f_2\}$. We see that, in the case of an elementary collapse, we get a new basis $\mathcal{B}'$ of K by performing elementary operations of type 1.*

Example 4 (Interior collapse)**.** *Let us now consider the collapse of the pair (e_5, f_2). Contrary to the previous case, edge e_5 is not free, so we cannot consider an elementary collapse of the pair (e_5, f_2). However, we can collapse (e_5, f_2) as an internal collapse. The obtained cell complex in Fig. 6.2(c) is not generated by critical basis elements in $\mathcal{A}$ as in the previous case. In fact, we have $\mathcal{A}_1 = \{e_1, e_2, e_3, e_4\}$ and $\mathcal{A}_2 = \{f_1\}$. Instead, it is generated by a new set of basis elements obtained from $\mathcal{A}$ by adding a linear combination*

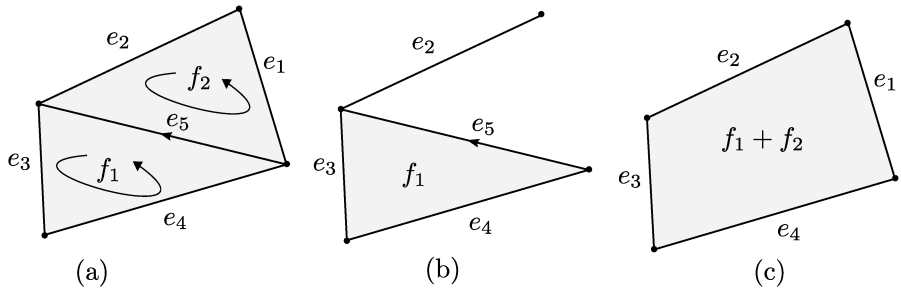

Figure 6.2: (a) The simplicial complex K. (b) Elementary collapse of the free pair (e_1, f_2). (c) Internal collapse of the pair (e_5, f_2); note that the geometric realization of the resulting cell complex is not more simplicial.

of other basis elements. We consider the linear transformation $\mathcal{A}_2 \ni f_1 \mapsto f_1 + f_2$. *The new set of critical basis elements is* $\mathcal{A}_1 = \{e_1, e_2, e_3, e_4\}$ *and* $\mathcal{A}_2 = \{f_1 + f_2\}$. *We get a different basis* $\mathcal{B}'$ *of the cell complex* K *as* $\mathcal{B}'_1 = \mathcal{D}_1 \cup \mathcal{A}_1$ *and* $\mathcal{B}'_2 = \mathcal{U}_2 \cup \mathcal{A}_2$, *where* $\mathcal{D}_1 = \{e_5\}$ *and* $\mathcal{U}_2 = \{f_2\}$. *We see that, in the case of an internal collapse, we get a new basis* $\mathcal{B}'$ *of* K *by performing elementary operations of type 1, 2 and 3.*

Let $\mathcal{M}_k = \{(\sigma_1, \tau_1), \ldots, (\sigma_n, \tau_n)\}$ be an acyclic matching of k-chains on $\mathcal{B}$. We now give a recursive definition of the *change of basis associated with* $\mathcal{M}_k$ and we denote it by $\mathcal{B} \cdot \mathcal{M}_k$. To start with, given the matched pair (σ, τ), we define the basis $\mathcal{B} \cdot (\sigma, \tau)$ as follows. $\mathcal{B} \cdot (\sigma, \tau)$ is obtained from $\mathcal{B}$ by performing two actions. First, we consider the partition of $\mathcal{B} = \mathcal{U} \cup \mathcal{D} \cup \mathcal{A}$, in particular, we have

$$\mathcal{B}_k = \mathcal{D}_k \cup \mathcal{A}_k, \tag{6.13}$$

$$\mathcal{B}_{k+1} = \mathcal{U}_{k+1} \cup \mathcal{A}_{k+1}. \tag{6.14}$$

Second, the pair (σ, τ) acts on the set of critical elements $\mathcal{A}$ as follows

$$\mathcal{A}_k \ni \sigma' \mapsto \sigma', \tag{6.15}$$

$$\mathcal{A}_{k+1} \ni \tau' \mapsto \tau' - \frac{\langle \sigma \, , \partial_{k+1} \tau' \rangle}{\langle \sigma \, , \partial_{k+1} \tau \rangle} \tau. \tag{6.16}$$

We see that $\mathcal{B} \cdot (\sigma, \tau)$ is again a basis of C^*. Indeed, it is obtained from $\mathcal{B}$ by adding linear combinations of other basis elements. We define $\mathcal{B} \cdot \mathcal{M}_k$ recursively by the rule

$$\mathcal{B} \cdot \{(\sigma_1, \tau_1), \ldots, (\sigma_n, \tau_n)\} \coloneqq \big(\mathcal{B} \cdot \{(\sigma_1, \tau_1), \ldots, (\sigma_{n-1}, \tau_{n-1})\}\big) \cdot (\sigma_n, \tau_n), \tag{6.17}$$

with $n \geq 2$. We see that (6.17) is well-defined. Indeed, if $\mathcal{B} \cdot \{(\sigma_1, \tau_1), \ldots, (\sigma_{n-1}, \tau_{n-1})\}$ is basis, then $\mathcal{B} \cdot \{(\sigma_1, \tau_1), \ldots, (\sigma_n, \tau_n)\}$ is obtained by adding linear combinations of

basis elements, hence it is a basis. Moreover, $\mathcal{M}_k$ is an acyclic matching on $\mathcal{B} \cdot \{(\sigma_1, \tau_1), \ldots, (\sigma_{n-1}, \tau_{n-1})\}$, since transformations (6.15), (6.16) leaves matched elements in $\mathcal{B}$ invariant.

Note that, when (σ, τ) is a free pair, the transformation (6.16) leaves elements in $\mathcal{A}_{k+1}$ invariant. Hence, no algebraic operations are required and the new basis is simply a permutation of the previous one. Instead, when (σ, τ) is internal, we have to apply at least one transformation (6.16).

6.3 Acyclic matchings and Gaussian elimination

We shall now head towards an algorithm to reduce the matrix $\mathbb{C}$ to a row echelon form by means of elementary operations on the basis $\mathcal{B}$. The whole procedure boils down to a standard train of thought used in basic linear algebra. When the coboundary operator δ^1 is given as a finite matrix $\mathbb{C}$, the bases and orders are already determined. However, we can get any other bases by applying elementary row and column operations on the matrix $\mathbb{C}$. Our algorithm will produce a change of bases in such a way that the matrix $\mathbb{C}$ in the new bases has an invertible upper triangular submatrix (i.e. a matrix with non-zeros only in its upper triangle and main diagonal) and thus it can be transformed in row echelon form. Hence, we can fast solve the system by processing the unknown variables in reverse order, a standard process known as back substitution.

6.3.1 Acyclic matchings and Gaussian elimination

The crucial observation is the following general novel result.

Lemma 8. *Denote by $\mathbb{D}_k$ one among the matrices $\mathbb{G}$, $\mathbb{C}$ or $\mathbb{D}$ for k equal to 0,1 or 2, respectively. Let $\mathcal{M}_k$ be an acyclic matching of k-chains on $\mathcal{B}$. Then, $\mathbb{D}_{k|\mathcal{U}_{k+1} \times \mathcal{D}_k}$, the submatrix of $\mathbb{D}_k$ induced by $\mathcal{U}_{k+1} \times \mathcal{D}_k$, is upper triangular and invertible.*

Proof. Since the matching $\mathcal{M}_k = \{(\sigma_1, \tau_1), \ldots, (\sigma_n, \tau_n)\}$ is acyclic, for every $i \in \{1, \ldots, n-1\}$, the basis element σ_i is not incident to any basis element $\tau_{i+1}, \ldots, \tau_n$. Thus, all the non-zero entries in each column are above the diagonal as the rows and columns are arranged in the total order induced by the matching. □

By Lemma 8, for every acyclic matching there is a corresponding upper triangular submatrix $\mathbb{D}_{k|\mathcal{U}_{k+1} \times \mathcal{D}_k}$ of matched rows and columns of $\mathbb{D}_k$. However, Lemma 8 is decisive only when the number of matched pairs in $\mathcal{M}_k$ is equal to the rank of $\mathbb{D}_k$. In fact, in this case, $\mathbb{D}_{k|\mathcal{U}_{k+1} \times \mathcal{D}_k}$ is an invertible submatrix of $\mathbb{D}_k$ of order equal to the rank $\mathbb{D}_k$, hence, we can write a solution of (6.2) after setting some free variables to zero. For this reason, we introduce the new concept of complete acyclic matching.

Definition 6.3.1 (Complete acyclic matching)**.** We say that a matching $\mathcal{M}_k$ of k-chains on $\mathcal{B}$ is *complete* if the number of matched pairs in $\mathcal{M}_k$ is equal to the rank of $\mathbb{D}_k$.

Now, we illustrate how the upper triangular submatrix $\mathbb{D}_{k|\mathcal{U}_{k+1} \times \mathcal{D}_k}$ induced by a complete acyclic matching allows us to operationally obtain a solution of (6.2) using back substitution.

Let us consider a complete acyclic matching $\mathcal{M}_1$ of 1-chains. By applying Lemma 8, the action of $\mathcal{M}_1$ on (6.2) is equivalent to Gaussian elimination. It is thus sufficient to invert submatrix $\mathbb{C}_{|\mathcal{U}_2 \times \mathcal{D}_1}$ after setting free variables in $\mathcal{A}_1$ to zero. We write a discrete potential $\boldsymbol{H}^{\mathcal{E}}$ solution of (6.2) as

$$\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1} = \mathbb{C}^{-1}_{|\mathcal{U}_2 \times \mathcal{D}_1} \boldsymbol{J}^{\mathcal{F}}_{|\mathcal{U}_2}, \tag{6.18}$$

$$\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{A}_1} = \mathbf{0}. \tag{6.19}$$

Since $\mathbb{C}_{|\mathcal{U}_2 \times \mathcal{D}_1}$ is upper triangular, we can evaluate $\mathbb{C}^{-1}_{|\mathcal{U}_2 \times \mathcal{D}_1} \boldsymbol{J}^{\mathcal{F}}_{|\mathcal{U}_2}$ by back substitution in linear time [94].

The aim of the back substitution is to determine the coefficient values of $\boldsymbol{H}^{\mathcal{E}}$. Let $\mathcal{M}_1 = \{(e_1, f_1), \ldots, (e_n, f_n)\}$. The process of back substitution is so-called because one determines the coefficient values backwards, by first computing $\boldsymbol{H}^{\mathcal{E}}_{e_n}$, then substituting back into the previous equation to solve for $\boldsymbol{H}^{\mathcal{E}}_{e_{n-1}}$ and repeating through $\boldsymbol{H}^{\mathcal{E}}_{e_{n-1}}$. A naive combination of these coefficients in one step for each $\boldsymbol{H}^{\mathcal{E}}_{e_i}$ is very time consuming since large intermediate expressions are generated. This is avoided by combining coefficients pairwise as in the following standard back substitution algorithm.

Algorithm 1 Back substitution process

1: **procedure** BACKSUBSTITUTION($\mathcal{M}_1, \mathcal{B}$)
2: **for** i ranging from n down to 1 **do**
3: $\boldsymbol{H}^{\mathcal{E}}_{e_i} = (\mathbb{C}_{f_i,e_i})^{-1}(\boldsymbol{J}^{\mathcal{F}}_{f_i} - \sum_{e \in \mathrm{bd}_{\mathcal{B}}(f_i)} \mathbb{C}_{f_i,e} \boldsymbol{H}^{\mathcal{E}}_{e})$
4: **return** $\boldsymbol{H}^{\mathcal{E}}$;

In the above discussion, although we focused on matrix $\mathbb{C}$, we have actually detailed the proof of the following theorem.

Theorem 6. *Given a complete acyclic matching $\mathcal{M}_k$ of k-chains on $\mathcal{B}$, for $k \in \{0, 1, 2\}$ we can find in linear time a discrete potential $\boldsymbol{v} \in C^k(K)$ of $\boldsymbol{w} \in C^{k+1}(K)$ with $\mathbb{D}_{k+1}\boldsymbol{w} = \mathbf{0}$, namely a solution of $\mathbb{D}_k \boldsymbol{v} = \boldsymbol{w}$.*

6.3.2 On the problem of constructing complete acyclic matchings: the case of tree-cotree techniques

We now turn to the theoretical issue of constructing complete acyclic matchings $\mathcal{M}_k$ of k-chains on the canonical basis $\widehat{\mathcal{B}}$.

Case $k = 0$. We need to construct a complete acyclic matching $\mathcal{M}_0$ of 0-chains on $\widehat{\mathcal{B}}$. By definition, the number of matched elements in $\mathcal{M}_0$ has to be equal to the rank of $\mathbb{G}$, which is $|N| - 1$. Let us consider a *spanning tree* T on K. It can be constructed in worst-case linear time using standard graph algorithms, for instance *breadth-first search* (BFS) algorithm [95]. There is a standard reasoning to define an acyclic matching corresponding to a spanning tree T [96]. It is obtained by mimicking a spanning tree traversal process. The construction goes as follows. Since T is a tree, there exists at least a *leaf* in T, namely a vertex $v \in N$ with only one incident edge in T. Pick a leaf v and pair it with the unique edge e containing it. Then, add the pair (v, e) to $\mathcal{M}_0$ and

remove v, e from T. Since v is a leaf, the obtained graph is again a tree. By iteratively repeating this process, we define a complete acyclic matching $\mathcal{M}_0$. The matching $\mathcal{M}_0$ is well-defined: if a vertex v is matched during the process, it is removed from T and thus cannot appear in any other pair; it is acyclic since every tree is. The matching $\mathcal{M}_0$ is complete, since K is connected, T is spanning, namely all vertices of K are in T and they are eventually added to $\mathcal{M}_0$ in the above process, except the last one.

Case $k = 2$. Before considering the case $k = 1$, we show the similarity of the construction for $k = 2$ with the case $k = 0$. We need to construct a complete acyclic matching $\mathcal{M}_2$ of 2-chains on basis $\widehat{\mathcal{B}}$. By definition, the number of matched elements in $\mathcal{M}_2$ has to equal to the rank of $\mathbb{D}$, which is $|C|$. Note that the cell complex K defines a manifold with boundary. It implies that if c, c' are distinct elements and f is a face such that $f \subset \partial c \cap \partial c'$, then c, c' are the only elements that contain f. Thanks to the manifold condition, it is well-defined the so-called *complete dual graph* of K. The dual graph $\tilde{G} = (\tilde{V} \cup \tilde{V}_\infty, \tilde{E} \cup \tilde{E}_\infty)$ of K is a graph with set of vertices $\tilde{V}$ given by elements of K and $\{c, c'\} \in \tilde{E}$ if $\dim(c \cap c') = 2$. There is an additional vertex $\tilde{v}_\infty \in \tilde{V}_\infty$ and there are additional edges $\tilde{E}_\infty$ in $\tilde{G}$. Each edge in $\tilde{E}_\infty$ corresponds to an element c whose boundary ∂c contains a boundary face f; thus, if f is a boundary face and c is the unique element incident to it, then $\{\tilde{v}_\infty, c\} \in \tilde{E}_\infty$. Let us consider a spanning tree $\tilde{T}$ on $\tilde{G}$. By repeating the same construction detailed for $k = 0$ on $\tilde{T}$, we obtain a finite sequence S of pairs of the form $(c, \tilde{e})$, where c is a vertex in $\tilde{V} \cup \tilde{V}_\infty$ (i.e. an element of K or $\tilde{v}_\infty$) and $\tilde{e} = \{c, c'\}$ is an edge in $\tilde{E} \cup \tilde{E}_\infty$. For each pair $(c, \tilde{e})$, we choose a face f of K such that $f \subset \partial c \cap \partial c'$ if $\tilde{v}_\infty \notin \tilde{e} = \{c, c'\}$, $f \subset \partial c$ if $c' = \tilde{v}_\infty$ and $f \subset \partial c'$ if $c = \tilde{v}_\infty$. Replace in S each pair $(c, \tilde{e})$ with either (f, c) if $c \neq \tilde{v}_\infty$ or (f, c') if $c = \tilde{v}_\infty$. In this way, we obtain a complete acyclic matching $\mathcal{M}_2$ of 2-chains of K. Thus, all vertices of the dual graph are eventually added to $\mathcal{M}_2$, except the vertex $\tilde{v}_\infty$, and all elements in K are matched.

By combining Theorem 6 with the above results, we state the following theorem which solves the discrete potential problem for $k \in \{0, 2\}$.

Theorem 7. *For $k \in \{0, 2\}$, we can find in linear time a discrete potential $\boldsymbol{v} \in C^k(K)$ of $\boldsymbol{w} \in C^{k+1}(K)$ with $\mathbb{D}_{k+1}\boldsymbol{w} = \boldsymbol{0}$, namely a solution of $\mathbb{D}_k\boldsymbol{v} = \boldsymbol{w}$.*

Case $k = 1$. With no surprise, it turns out to be the most challenging case. To begin with, we show the following result.

Theorem 8. *Let K be a 3-dimensional topologically trivial simplicial complex embedded in $\mathbb{R}^3$ with Lipschitz boundary as in Section 6.1. Then, K is collapsible if and only if there exists a complete acyclic matching $\mathcal{M}_1$ of 1-chains on the canonical basis $\widehat{\mathcal{B}}$.*

Proof. Let us consider a sequence of elementary collapses $\mathcal{M}$ of K leading from K to a vertex. We define an acyclic matching $\mathcal{M}_1$ of 1-chains (on $\widehat{\mathcal{B}}$) by selecting, according to the total order of $\mathcal{M}$, all pairs made by matched edges and faces in $\mathcal{M}$. Note that $\mathcal{M}_1$ is necessarily acyclic since $\mathcal{M}$ is. Next, since K collapses to a vertex, all faces in K are matched in $\mathcal{M}$. In particular, all elements in K are matched in $\mathcal{M}$ and there are $|C|$ corresponding matched faces. Thus, there are $|F| - |C|$ matched faces in $\mathcal{M}_1$. It follows that $\mathcal{M}_1$ is complete since the rank of $\mathbb{C}$ is exactly $|F| - |C|$.

To prove the converse result, let us consider a complete acyclic matching of 1-chains $\mathcal{M}_1$.

We need the following construction. First let us consider the complete dual graph $\tilde{G} = (\tilde{V} \cup \tilde{V}_\infty, \tilde{E} \cup \tilde{E}_\infty)$ of K as defined above for the case $k = 2$. Consider the subgraph $\tilde{G}_{\mathcal{M}_1} = (\tilde{V} \cup \tilde{V}_\infty, \tilde{E}_{\mathcal{M}_1})$ of $\tilde{G}$ which has the same vertices of $\tilde{G}$ and $\tilde{E}_{\mathcal{M}_1}$ includes an edge $\{c, c'\} \in \tilde{E}$ of $\tilde{G}$ if the unique face f of K such that $f \subset c \cap c'$ is critical with respect to $\mathcal{M}_1$. We now show that $\tilde{G}_{\mathcal{M}_1}$ has $|C|$ edges (i.e, the number of vertices of $\tilde{G}$ minus one) and is connected. First, since $\mathcal{M}_1$ is a complete acyclic matching the number of critical faces, which is equal to the number of edges in $\tilde{E}_{\mathcal{M}_1}$, is $|F| - (|F| - |C|) = |C|$. Second, assume that, for the sake of contradiction, $\tilde{G}_{\mathcal{M}_1}$ is disconnected. Let $\tilde{V}^*$ be the set of nodes in a fixed connected component of $\tilde{G}_{\mathcal{M}_1}$. Let $\tilde{E}_{\tilde{V}^*} \subset \tilde{E} \cup \tilde{E}_\infty$ be the subset of *cut edges*, namely edges of $\tilde{G}$ with one vertex in $\tilde{V}^*$ and one vertex in its complement $(\tilde{V} \cup \tilde{V}_\infty) \setminus \tilde{V}^*$. Note that since $\tilde{G}$ is connected, $\tilde{E}_{\tilde{V}^*}$ is not empty. Moreover, by definition of $\tilde{G}_{\mathcal{M}_1}$, to each $\tilde{e} \in \tilde{E}_{V^*}$ corresponds a matched face f of K with respect to $\mathcal{M}_1$ forming the pair $(e, \tilde{e}) = (e, f) \in \mathcal{M}_1$ for a unique edge e of K.

Now we construct a cycle of the form $\tilde{e}_1 \succ e_1 \prec \tilde{e}_2 \succ \cdots \prec \tilde{e}_h \succ e_h \prec \tilde{e}_1$ with $h \geq 2$, $(e_i, \tilde{e}_i) \in \mathcal{M}_1$ for all $i \in \{1, \ldots, h\}$ and all $\tilde{e}_i \in \tilde{E}_{V^*}$ being distinct. Start with a cut edge $\tilde{e}_1$ and consider the unique edge e_1 of K forming the matched pair $(e_1, \tilde{e}_1) \in \mathcal{M}_1$. Then, e_1 is contained in at least one other cut edge $\tilde{e}_2 \in \tilde{E}_{V^*}$, otherwise $\tilde{e}_1$ cannot be a cut edge. By iteratively repeating this process, we obtain a sequence S of pairs $(e_1, \tilde{e}_1), \ldots, (e_i, \tilde{e}_i)$ such that $\tilde{e}_1 \succ e_1 \prec \tilde{e}_2 \succ \cdots \prec \tilde{e}_{i-1} \succ e_i \prec \tilde{e}_i$. Since we have finite graphs, eventually we will run out of those cut edges which do not appear as the second component of a pair in S. Thus, at some step i of the above process, we must get a cut edge that is the second component of a pair in S and we get a cycle of the form (6.10), which is a contradiction since $\mathcal{M}_1$ is acyclic. We have proved that $\tilde{G}_{\mathcal{M}_1}$ is a connected subgraph of $\tilde{G}$ with $|C|$ edges and thus it is a spanning tree of $\tilde{G}$. We define a complete acyclic matching $\mathcal{M}_2$ of 2-chains by using the spanning tree construction detailed above for $k = 2$ on $\tilde{G}_{\mathcal{M}_1}$.

Now, $\mathcal{M}_2$ and $\mathcal{M}_1$ define a sequence of elementary collapses leading from K to a single vertex as follows. Since every element of K is matched with respect to $\mathcal{M}_2$, the set of critical cells of K with respect to $\mathcal{M}_2$ form a 2-dimensional subcomplex $K^{(1)}$ of K. By applying Theorem 11.13 (a) in [92], there exists a sequence of elementary collapses leading from K to $K^{(1)}$. Next, observe that, by construction, the set of matched faces in K with respect to $\mathcal{M}_2$ is exactly the set of critical faces of K with respect to $\mathcal{M}_1$. Thus, the set of critical cells of $K^{(1)}$ with respect to $\mathcal{M}_1$ form a 1-dimensional subcomplex $K^{(2)}$ of $K^{(1)}$. By applying Theorem 11.13 (a) in [92] there exists a sequence of elementary collapses leading from $K^{(1)}$ to $K^{(2)}$. $K^{(2)}$ is a topologically trivial 1-dimensional subcomplex, thus it is a spanning tree on K. Thus K collapses to a single vertex. □

Now we focus on the *Spanning Tree Technique* (STT) algorithm introduced in [88] and implicitly used in many works [82, 83, 87, 80]. The algorithm, given as input a spanning tree T on K, computes a discrete vector potential $\boldsymbol{H}^{\mathcal{E}}$ as follows. To start with, STT sets the value $\boldsymbol{H}^{\mathcal{E}}_e = 0$ if $e \in T$, otherwise the value $\boldsymbol{H}^{\mathcal{E}}_e$ at this stage is unknown. Next, all faces f of K are loaded into a list L. The main loop of STT works until there are no more faces in L. In each iteration, we randomly search for a face f with two boundary edges $e_1, e_2 \subset \partial f$ such that the values $\boldsymbol{H}^{\mathcal{E}}_{e_1}, \boldsymbol{H}^{\mathcal{E}}_{e_2}$ are known. Then, the value of $\boldsymbol{H}^{\mathcal{E}}_e$ of the remaining boundary edge $e \subset \partial f$ is determined by

$$\boldsymbol{H}^{\mathcal{E}}_e = (\mathbb{C}_{f,e})^{-1}(\boldsymbol{J}^{\mathcal{F}}_f - \mathbb{C}_{f,e_1}\boldsymbol{H}^{\mathcal{E}}_{e_1} - \mathbb{C}_{f,e_2}\boldsymbol{H}^{\mathcal{E}}_{e_2}) \tag{6.20}$$

and face f is removed from the list L. In the case when L is non-empty and there is no available face f satisfying the above property, then STT does not terminate since it stalls in a infinite loop. As shown in [88], STT termination depends on the choice of the input spanning tree T.

We now show that the STT algorithm boils down to a procedure to construct complete acyclic matchings of 1-chains on the canonical basis $\widehat{\mathcal{B}}$. This result completes the picture of the equivalence between tree-cotree techniques and the problem of finding complete acyclic matching of k-chains. The idea behind this observation has it root on the fact that tree-cotree techniques describe the same actions of acyclic matchings although using a different language.

To formally state the next theorem we introduce the following notations. Given a spanning tree T on K, we say that STT *terminates* (with input T) if it does not stall in an infinite loop. We say that STT *uses the pair* (e, f) if STT determines the value of edge $e \subset \partial f$ via (6.20) during its main loop execution. Finally, we say that STT *uses edge* e if STT uses the pair (e, f) for some face $f \in L$.

Theorem 9. *Let K be a simplicial complex as in Theorem 8. There exists a spanning tree on K for which STT terminates if and only if there exists a complete acyclic matching of 1-chains on the canonical basis $\widehat{\mathcal{B}}$.*

Proof. Let T be a spanning tree such that STT terminates. We now define an acyclic matching $\mathcal{M}_T$ of 1-chains on $\widehat{\mathcal{B}}$ such that every *cotree edge* of T, namely an edge $e \in E$ which is not an edge of T, is matched in $\mathcal{M}_T$. $\mathcal{M}_T$ is constructed during STT execution as follows. We initialize $\mathcal{M}_1 = \emptyset$. Next, if STT uses a pair (e, f) during its execution then we add (e, f) to $\mathcal{M}_1$. In this way we get a total order of pairs in $\mathcal{M}_1$ where a pair $(e_i, f_i) \in \mathcal{M}_1$ comes before than a pair $(e_j, f_j) \in \mathcal{M}_1$ in this total order with $i < j$ if STT uses (e_i, f_i) before than (e_j, f_j). The set $\mathcal{M}_T$ of all such pairs is a complete acyclic matching. First, we see that $\mathcal{M}_T$ is a matching since if STT uses the pair (e, f), then the value of $\boldsymbol{H}_e^{\mathcal{E}}$ set by STT can never be reassigned. Second, the matching $\mathcal{M}_T$ is acyclic since if STT uses the pair (e_i, f_i), then e_i is not contained in the boundary of any matched face f_j in $\mathcal{M}_1$ for $1 \leq j < i$. Third, since STT terminates, every cotree edge e of T is matched in $\mathcal{M}_1$. Using Euler's formula, the fact that K is topologically trivial and that T is a spanning tree on K we get $|E| - (|N| - 1) = |F| - |C|$. Thus $\mathcal{M}_1$ is complete.

Conversely, let $\mathcal{M}_1$ be a complete acyclic matching of 1-chains on $\widehat{\mathcal{B}}$. We need the following construction, which mimics the one described in Theorem 8. Consider the subgraph $G_{\mathcal{M}_1} = (V, E_{\mathcal{M}_1})$ of K with set of vertices given by vertices of K and $E_{\mathcal{M}_1}$ includes an edge $e \in E$ if e is critical with respect to $\mathcal{M}_1$. We now show that $G_{\mathcal{M}_1}$ has $|N| - 1$ edges and is connected. First, since $\mathcal{M}_1$ is a complete acyclic matching the number of critical edges, which is equal to the number of edges in $E_{\mathcal{M}_1}$, is $|E| - (|F| - |C|) = |N| - 1$ where we have used Euler's formula $|N| - |E| + |F| - |C| = 1$ and the fact that K is topologically trivial. Second, assume that, for the sake of contradiction, $G_{\mathcal{M}_1}$ is disconnected. Let V^* be the set of nodes in a fixed connected component of $G_{\mathcal{M}_1}$. Let $E_{V^*} \subset E$ be the subset of *cut edges*, namely edges of G with one vertex in V^* and one vertex in its complement $V \setminus V^*$. Note that since K is connected, E_{V^*} is not empty. Moreover, by definition of $G_{\mathcal{M}_1}$, to each $e \in E_{V^*}$ corresponds a matched edge e of K with respect to $\mathcal{M}_1$ forming the pair $(e, f) \in \mathcal{M}_1$ for a unique face f of K.

Now we construct a cycle of the form $f_1 \succ e_1 \prec f_2 \succ \cdots \prec f_h \succ e_h \prec f_1$ with $h \geq 2$, $(e_i, f_i) \in \mathcal{M}_1$ for all $i \in \{1, \ldots, h\}$ and all $e_i \in E_{V^*}$ being distinct. Start with a cut edge e_1 and consider the unique face f_1 of K forming the matched pair $(e_1, f_1) \in \mathcal{M}_1$. Then, f_1 contains at least one other cut edge $e_2 \in E_{V^*}$, otherwise e_1 cannot be a cut edge. By iteratively repeating this process, we obtain a sequence S of pairs $(e_1, f_1), \ldots, (e_i, f_i)$ such that $f_1 \succ e_1 \prec f_2 \succ \cdots \prec f_{i-1} \succ e_i \prec f_i$. Since we have finite graphs, eventually we will run out of those cut edges which do not appear as the first component of a pair in S. Thus, at some step i of the above process, we must get a cut edge that is the first component of a pair in S and we get a cycle of the form (6.10), which is a contradiction since $\mathcal{M}_1$ is acyclic. We have proved that $G_{\mathcal{M}_1}$ is a connected subgraph of K with $|N| - 1$ edges and thus it is a spanning tree on K.

Now we prove that STT with input $G_{\mathcal{M}_1}$ terminates. Note that, by definition of $G_{\mathcal{M}_1}$, if $(e, f) \in \mathcal{M}_1$ then e is a *cotree edge* of $G_{\mathcal{M}_1}$, namely e is not an edge of $G_{\mathcal{M}_1}$. Denote by C the set of all cotree edges of $G_{\mathcal{M}_1}$. Moreover, since $\mathcal{M}_1$ is complete, reasoning as above using Euler's formula, we get that every cotree edge of $G_{\mathcal{M}_1}$ is matched with respect to $\mathcal{M}_1$. Since $\mathcal{M}_1 = \{(e_1, f_1), \ldots, (e_n, f_n)\}$ is acyclic, we can order pairs in $\mathcal{M}_1$ in such a way that, for every $i \in \{1, \ldots, n-1\}$, f_i is not incident to any $e_{i+1}, \ldots, e_n$. Let E_{STT} be the set of cotree edges used by STT during its execution. Note that E_{STT} is not empty since STT can use at least the pair (e_1, f_1), thanks to the total order chosen on $\mathcal{M}_1$. Suppose that STT does not terminate. This means that, during its execution, it does not use any cotree edge and $C \setminus E_{\mathrm{STT}} \neq \emptyset$. We will show that this is impossible. Let i be the minimum integer such that the cotree edge e_i belongs to $C \setminus E_{\mathrm{STT}}$. Since pairs in $\mathcal{M}_1$ are ordered as described above, $i > 1$ and f_i can be only incident to cotree edges e_j with $j \leq i$. Moreover, by definition of i, there is no j with $j < i$ such that the cotree edge e_j belongs to $C \setminus E_{\mathrm{STT}}$. Hence, STT should have at least used the pair (e_i, f_i), i.e. e_i belongs to E_{STT} as well. This gives the desired contradiction and completes the proof. □

The proof of Theorem 9 shows that, if the STT algorithm terminates for a given spanning tree T, starting from T we can construct a complete acylclic matching $\mathcal{M}_T$ of 1-chains on $\widehat{\mathcal{B}}$. But also the other way around, namely, if we have a complete acyclic matching $\mathcal{M}_1$ of 1-chains on $\widehat{\mathcal{B}}$, starting from $\mathcal{M}_1$ we can construct a spanning tree $G_{\mathcal{M}_1}$ for which STT terminates.

By combining Theorem 8 and Theorem 9 we now state the following result which gives a topological characterization of termination problems of tree-cotree techniques.

Theorem 10. *Let K be a simplicial complex as in Theorem 8. Then, there exists a spanning tree on K for which STT terminates if and only if K is collapsible.*

There are known examples of triangulations of 3-balls which are not collapsible [89]. Hence, using Theorem 10, there are triangulations K of 3-balls such that, for every possible spanning tree of K given as input, STT does not terminate.

We remark that it is NP-complete to decide whether a given 3-dimensional simplicial complex (embedded or not) is collapsible [90]. However, to the authors knowledge, the related question for the case of 3-dimensional simplicial complexes embedded in $\mathbb{R}^3$ and with Lipschitz boundary is still an open problem. Yet, numerical evidence shows that this problem is very difficult in general although good heuristics exists [97].

These results are the cause of the well-known termination problems of tree-cotree techniques [88]; see also the discussion in Section 6.5.

6.3.3 The case where we cannot find a complete acyclic matching

As shown by Theorem 8 and the discussion at the end of Section 6.3.2, for cell complex K given as input, we cannot find in general a complete acyclic matching $\mathcal{M}_1$ of 1-chains on the canonical basis $\widehat{\mathcal{B}}$. Consequently, Theorem 6 cannot be applied in general for $\mathcal{B} = \widehat{\mathcal{B}}$.

Let us now assume that acyclic matching $\mathcal{M}_1$ is not complete. We shall now show that, after solving another linear system, we can still get a discrete vector potential solution $\boldsymbol{H}^{\mathcal{E}}$ solution of (6.2) by exploiting back substitution.

As shown in Section 6.2.3, an acyclic matching acts on a basis $\mathcal{B}$ by performing elementary operations on it. Let us consider the new basis $\mathcal{B}' = \mathcal{B} \cdot \mathcal{M}_1 = \mathcal{U} \cup \mathcal{D} \cup \mathcal{A}$. There is a corresponding block partition of linear system (6.2) as

$$\begin{pmatrix} \mathbb{C}_{|\mathcal{U}_2 \times \mathcal{D}_1} & \mathbb{C}_{|\mathcal{U}_2 \times \mathcal{A}_1} \\ \mathbb{C}_{|\mathcal{A}_2 \times \mathcal{D}_1} & \mathbb{C}_{|\mathcal{A}_2 \times \mathcal{A}_1} \end{pmatrix} \begin{pmatrix} \boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1} \\ \boldsymbol{H}^{\mathcal{E}}_{|\mathcal{A}_1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{J}^{\mathcal{F}}_{|\mathcal{U}_2} \\ \boldsymbol{J}^{\mathcal{F}}_{|\mathcal{A}_2} \end{pmatrix}. \tag{6.21}$$

The crucial fact turns out to be that $\mathbb{C}_{|\mathcal{A}_2 \times \mathcal{D}_1}$ is a zero matrix. Thus, we can determine a discrete vector potential $\boldsymbol{H}^{\mathcal{E}} = (\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1}, \boldsymbol{H}^{\mathcal{E}}_{|\mathcal{A}_1})^T$ solution of (6.2) by solving, in order,

$$\mathbb{C}_{|\mathcal{A}_2 \times \mathcal{A}_1} \boldsymbol{H}^{\mathcal{E}}_{|\mathcal{A}_1} = \boldsymbol{J}^{\mathcal{F}}{}_{|\mathcal{A}_2}, \tag{6.22}$$

$$\mathbb{C}_{|\mathcal{U}_2 \times \mathcal{D}_1} \boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1} = \boldsymbol{J}^{\mathcal{F}}_{|\mathcal{U}_2} - \mathbb{C}_{|\mathcal{U}_2 \times \mathcal{A}_1} \boldsymbol{H}^{\mathcal{E}}_{|\mathcal{A}_1}, \tag{6.23}$$

where (6.23) is solved by exploiting back substitution as in Section 6.3.1.

Let us prove that $\mathbb{C}_{|\mathcal{A}_2 \times \mathcal{D}_1}$ is a zero matrix. Let $\mathcal{M}_1 = \{(\sigma_1, \tau_1) \ldots, (\sigma_n, \tau_n)\}$. Let us consider bases $\{\ldots, \xi^i = \phi_k(\xi_i), \ldots\}$ and $\{, \ldots, \pi_i, \ldots, \}$ of $C^1(K)$ and $C_2(K)$, respectively. We see that the (i,j)-entry of $\mathbb{C}$ is $\mathbb{C}_{i,j} = \langle \delta^1(\xi^j), \pi_i \rangle = \langle \xi^j, \partial_2 \pi_i \rangle = \langle \xi_j, \partial_2 \pi_i \rangle$, where we have used (2.42) and the isomorphism $\phi_k : C_k(K) \to C^k(K)$ in (2.41). Thus, to prove that $\mathbb{C}_{|\mathcal{A}_2 \times \mathcal{D}_1}$ is a zero matrix we have to show that every $\sigma_i \in \mathcal{D}_1$ with $i \in \{1, \ldots, n\}$ is not incident to any $\rho \in \mathcal{A}_2$.

We proceed by induction on n.

Suppose that $n = 1$. Necessarily, $\tau_1 \in \mathcal{U}_2$ is the only basis element in $\mathcal{B} \cdot \mathcal{M}_1 = \mathcal{U} \cup \mathcal{D} \cup \mathcal{A}$ incident on σ_1. In fact, if $\tau \in \mathcal{B}$ is incident on σ_1, then its image under the transformation (6.16) is not incident on σ_1. It follows that σ_1 is not incident to any $\rho \in \mathcal{A}_2$.

We now assume the statement true for $n-1$ and we prove it for n. Let us consider the basis $\mathcal{B} \cdot \mathcal{M}_1 = \mathcal{U} \cup \mathcal{D} \cup \mathcal{A}$. Proceeding as above, we see that σ_n is not incident to any $\rho \in \mathcal{A}_2$. To conclude, it is sufficient to show that also each σ_i with $i \in \{1, \ldots, n-1\}$ is not incident to any $\rho \in \mathcal{A}_2$. Basis elements in $\mathcal{B} \cdot \mathcal{M}_1$ are obtained from that of $\mathcal{B} \cdot \{(\sigma_1, \tau_1), \ldots, (\sigma_{n-1}, \tau_{n-1})\}$ by applying transformation (6.16) with $(\sigma, \tau) = (\sigma_n, \tau_n)$. Using the induction hypothesis we see that the first term in (6.16) (i.e., $\tau' \in C_2(K)$) is not incident on any σ_i with $i \in \{1, \ldots, n-1\}$. Moreover, each σ_i with $i \in \{1, \ldots, n-1\}$

is not incident to τ_n as pairs in $\mathcal{M}_1$ are ordered as in Theorem 5. This completes the proof.

We now state the following theorem which combines the results of this section with those of Section 6.3.1. It can be thought as a specific algebraic version of Forman's discrete Morse complex construction [91]. We think that our presentation and terminology shed light on the linear algebra behind the more abstract discrete Morse theory constructions.

Theorem 11. *Denote by $\mathbb{D}_k$ one among the matrices $\mathbb{G}$, $\mathbb{C}$ or $\mathbb{D}$ for k equal to 0,1 or 2, respectively. Let $\mathcal{M}_k$ be an acyclic matching of k-chains on $\mathcal{B}$. Then, there is a corresponding block partition of $\mathbb{D}_k$ as*

$$\mathbb{D}_k = \begin{pmatrix} \mathbb{D}_{k|\mathcal{U}_{k+1}\times\mathcal{D}_k} & \mathbb{D}_{k|\mathcal{U}_{k+1}\times\mathcal{A}_k} \\ 0 & \mathbb{D}_{k|\mathcal{A}_{k+1}\times\mathcal{A}_k} \end{pmatrix}, \tag{6.24}$$

where $\mathbb{D}_{k|\mathcal{U}_{k+1}\times\mathcal{D}_k}$ is upper triangular and invertible.

6.4 Algorithm description

The goal of this section is a recursive algorithm that reduces matrix $\mathbb{C}$ into a row echelon form by means of elementary operations on the basis $\mathcal{B} = \bigcup_k \mathcal{B}_k$ and we present in Section 6.4.2. We first describe in Section 6.4.1 our novel greedy procedure to construct acyclic matchings.

6.4.1 Greedy approach to construct acyclic matchings

To minimize computational effort of change of basis in (6.22), we have to carefully choose how to construct the acyclic matching $\mathcal{M}_1$. It is visible from Section 6.2.3 that constructing acyclic matchings by internal collapses can get more complicated than by elemetary collapses since changes of basis of type 2 and 3 are involved. In this case, we have to express the new basis elements as a linear combination of the previous ones and hence they cannot be removed from the data structure after each collapse.

The above discussion motivates the concept of *degree* of a basis element $\sigma \in \mathcal{B}$. We define $\deg(\sigma)$ as the cardinality of the coboundary of σ, i.e. the set $\mathrm{cobd}_{\mathcal{B}}(\sigma)$ defined in (6.9). Note that if $\deg(\sigma)$ is 1 then σ is free. In our approach, also the case $\deg(\sigma) = 2$ will play a fundamental role. We define $\sigma \in \mathcal{B}$ to be *flat* if $\deg(\sigma) = 2$.

To reduce the amount of computation during Gaussian elimination, it is wise to first search for basis elements with lowest degree. Indeed, if $\deg(\sigma)$ is 1, then σ is free and the new basis is a selection of the previous one and no algebraic operations are needed. In this case, it should be possible to avoid all the matrix algebra computation and efficiently organize the basis so that matrices are in triangular form as in Lemma 8.

Keeping in mind the above heuristics, we search for basis elements having smallest degree. Specifically, we do not strictly choose basis elements with minimum degree but instead we proceed in a sequential manner. First, we search for all free basis elements until exhaustion. Next, we search for all flat basis elements until exhaustion. We have pursued this method because been motivated by its practical implementation and

performance on test problems, rather than by following the best theoretical greedy approach; see the discussion in Section 6.5.

We construct acyclic matchings using a standard elementary collapse greedy procedure, where we search for collapsing sequences of free basis elements in a monotone-like fashion [97]. That is, we proceed in sequential order with respect to the dimension of the basis elements by first collapsing 2-chains and then 1-chains.

Algorithm 2 Random strategy to construct acyclic matchings

```
1: procedure ConstructAcyclicMatchings(B)
2:     for k ranging from 2 down to 1 do
3:         while there exists a free pair (σ, τ) in B_k do
4:             elementary collapse (σ, τ)
5:             insert (σ, τ) to M_k
6:         while there exists a flat pair (σ, τ) in B_k do
7:             internal collapse of (σ, τ)
8:             insert (σ, τ) to M_k
9:     return M_2 ∪ M_1
```

It is clear that Algorithm 2 always terminates. Moreover, the obtained matching is acyclic since at each iteration collapses are performed. Note that Algorithm 2 requires no backtracking since new free or flat pairs can only appear after each new collapse. Thus, the worst-case complexity is linear using a suitable algorithm implementation that employs a list data structure.

As proved in Section 6.3.2, we can always find a complete acyclic matching of 2-chains by using a standard spanning tree construction. The proof of this result implies that the order in which we collapse 2-chains in Algorithm 2 is not important. Therefore, there always exists a complete acyclic matching of 2-chains and we can get one by performing collapses in a random fashion.

6.4.2 Recursive algorithm

We now present a recursive construction of acyclic matchings.

Normally one aims at finding a complete acyclic matching, namely a matching that reaches the needed number of matched pairs so that we can apply Theorem 6. However, as proved in Theorem 8, determining a complete acyclic matching $\mathcal{M}_1$ of 1-chains is a hard algorithmic problem and we do not tackle it. In Algorithm 2 we employed a greedy strategy that randomly selects an acyclic matching but it can easily happen that the obtained acyclic matching $\mathcal{M}_1$ is not complete.

Let us consider the case where $\mathcal{M}_1$ is not complete. After applying basis transformations associated with $\mathcal{M}_1$ to the current basis $\mathcal{B}$, we get a new basis $\mathcal{B}' = \mathcal{B} \cdot \mathcal{M}_1 = \mathcal{U} \cup \mathcal{D} \cup \mathcal{A}$. In particular, the set of critical basis elements $\mathcal{A}$ in $\mathcal{B}'$ forms a linearly independent subset of $\mathcal{B}'$. As pointed out in Section 6.3.3, the focus now shifts to the set of critical basis elements $\mathcal{A}$.

The novel idea is that we may call the whole routine recursively, where the output set of critical basis elements obtained from the previous iteration becomes the input basis

for the next iteration. Fundamentally, we think linear system (6.22) as an instance of the original linear system (6.2), although considering a subset of the previous basis. This operation is well-defined since we have the partition $\mathcal{B}' = \mathcal{U} \cup \mathcal{D} \cup \mathcal{A}$ and the set $\mathcal{A}$ of critical basis elements form a linearly independent subset of $\mathcal{B}$.

If at a certain recursion stage we get a complete acyclic matching $\mathcal{M}_1$ of 1-chains, then we can find a discrete vector potential $\boldsymbol{H}^{\mathcal{E}}$ solution of (6.2) by recursively applying the reasoning described in Section 6.3.3. Otherwise, we recursively apply the same routine on the obtained set of critical basis elements $\mathcal{A}$.

Algorithm 3 Construction of a discrete vector potential $\boldsymbol{H}^{\mathcal{E}}$ solution of (6.2)

procedure ConstructDiscreteVectorPotential($\mathcal{B}$) ▷ at the beginning $\mathcal{B}$ is the canonical basis $\widehat{\mathcal{B}}$
  construct an acyclic matching $\mathcal{M}_2 \cup \mathcal{M}_1$ on $\mathcal{B}$ using Algorithm 2
  if $\mathcal{M}_1$ is complete **then**
    set $\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{A}_1} = \boldsymbol{0}$
    determine $\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1}$ by back substitution using Algorithm 1
    return $(\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1}, \boldsymbol{H}^{\mathcal{E}}_{|\mathcal{A}_1})^T$
  else
    compute new basis $\mathcal{B}'$ from $\mathcal{B}$ by change of basis associated with $\mathcal{M}_2 \cup \mathcal{M}_1$
    if $\mathcal{M}_1$ is empty **then**
      determine $\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{A}_1}$ in (6.22) using a sparse linear system solver
      return $\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{A}_1}$
    else
      recursively call this routine with input the set of critical elements $\mathcal{A}$ of $\mathcal{B}' = \mathcal{U} \cup \mathcal{D} \cup \mathcal{A}$
      set $\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{A}_1}$ to the value returned by the recursive process
      determine $\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1}$ by back substitution using Algorithm 1
      return $(\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1}, \boldsymbol{H}^{\mathcal{E}}_{|\mathcal{A}_1})^T$

For a given input basis $\mathcal{B}$, Algorithm 3 produces a sequence of bases $\mathcal{B} = \mathcal{B}^{(0)}, \mathcal{B}^{(1)}, \ldots, \mathcal{B}^{(n)}$, where index i keeps track of the recursion depth. Each $\mathcal{B}^{(i+1)}$ is constructed from $\mathcal{B}^{(i)}$ by recursion as follows. We have the partition of $\mathcal{B}'^{(i)} = \mathcal{B}^{(i)} \cdot (\mathcal{M}_2^{(i)} \cup \mathcal{M}_1^{(i)}) = \mathcal{U}^{(i)} \cup \mathcal{D}^{(i)} \cup \mathcal{A}^{(i)}$, associated with acyclic matching $\mathcal{M}_2^{(i)} \cup \mathcal{M}_1^{(i)}$ for $i \in \{0, \ldots, n\}$. Note that from the discussion at the end of Section 6.4.1 it follows that $\mathcal{M}_2^{(i)}$ is not empty only for $i = 0$. We set $\mathcal{B}^{(i+1)} = \mathcal{A}^{(i)}$.

It is clear that the cardinality of each basis $\mathcal{B}^{(i)}$ decreases as long as there are some collapses to be made by Algorithm 2. If no free or flat basis elements are available in Algorithm 2, then the recursion is stopped and a sparse linear system solver is used. So the algorithm always terminates.

At the end of the recursion, we can determine a discrete vector potential $\boldsymbol{H}^{\mathcal{E}}$ solution of (6.2) by recursively applying the reasoning described in Section 6.3.3. We write the general form a discrete vector potential $\boldsymbol{H}^{\mathcal{E}}$ solution of (6.2) as

$$\boldsymbol{H}^{\mathcal{E}} = (\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1^{(0)}}, \ldots, \boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1^{(n)}}, \boldsymbol{H}^{\mathcal{E}}_{|\mathcal{A}_1^{(n)}})^T, \tag{6.25}$$

where each $\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1^{(i)}}$ is the subvector of $\boldsymbol{H}^{\mathcal{E}}$ induced by $\mathcal{D}_1^{(i)}$, see Fig. 6.3. We start by possibly determining $\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{A}_1^{(n)}}$ using a linear system solver. Next, we determine each $\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1^{(0)}}, \ldots, \boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1^{(n)}}$ by back substitution starting from $\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1^{(n)}}$ down to $\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1^{(0)}}$. This proves the correctness of Algorithm 3.

Theorem 12. *Given as input the canonical basis $\widehat{\mathcal{B}}$, Algorithm 3 returns a discrete vector potential $\boldsymbol{H}^{\mathcal{E}}$ solution of (6.2).*

$$\begin{pmatrix} \mathbb{C}|_{\mathcal{U}_2^{(0)}\times\mathcal{D}_1^{(0)}} & & & & \\ 0 & \mathbb{C}|_{\mathcal{U}_2^{(1)}\times\mathcal{D}_1^{(1)}} & & & \\ \vdots & \ddots & \ddots & & \\ & & & \mathbb{C}|_{\mathcal{U}_2^{(n)}\times\mathcal{D}_1^{(n)}} & \\ 0 & & \cdots & 0 & \mathbb{C}|_{\mathcal{A}_2^{(n)}\times\mathcal{A}_1^{(n)}} \end{pmatrix}$$

Figure 6.3: General block structure of matrix $\mathbb{C}$ produced by Algorithm 3.

Concerning the computational complexity of Algorithm 3, if at a certain recursion stage no free or flat basis elements are available in Algorithm 2, a linear system solver is employed. Thus, the worst-case complexity is cubical with respect to the size of the input mesh. Yet, the average complexity of Algorithm 3 in all tested problems has been linear. This is because, in practice, only one recursively call of Algorithm 3 is needed. Indeed, the first recursive call of Algorithm 3 always finds a complete acyclic matching $\mathcal{M}_1$ with respect to new basis $\mathcal{B}^{(1)}$. In other words, there is no need in practice to solve a linear system of the form (6.22) with matrix $\mathbb{C}_{|\mathcal{A}_2^{(1)}\times\mathcal{A}_1^{(1)}}$ but instead we can recursively apply back substitution to determine $\boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1^{(1)}}, \boldsymbol{H}^{\mathcal{E}}_{|\mathcal{D}_1^{(0)}}$, in this order, after setting free variables in $\mathcal{A}_1^{(1)}$ to zero.

6.5 Numerical results

In this section we illustrate the performance of Algorithm 3. We consider different sets of test problems. In the first set, we focus on simple triangulations that appear in practical boundary value problems. Next, we present more complicated benchmark triangulations.

For all triangulations that appear in practical boundary value problems, our greedy procedure in Algorithm 2 always finds a complete acyclic matching $\mathcal{M}_1^{(0)}$. Only for the more complicated benchmark problems it is necessary to exploit the novel recursive procedure in Algorithm 3. However, to compute a discrete vector potential it is enough, in all tested problems, only one recursively call of Algorithm 3.

The algorithm has been implemented in C++. All the numerical computations have been performed in a Intel Core i7-3720QM, with a processor at 2.60 GHz in a laptop with 16 GB of RAM.

6.5.1 Triangulations coming from real case boundary value problems

We consider a triangulation coming from a computational electromagnetics application. As an example, the computation of the source magnetic field for the TEAM problem 7 has been addressed [98].

Table 6.1 contains information on the number of cells of triangulations of different sizes together with the time (in milliseconds) required to compute the discrete vector potential $\boldsymbol{H}^{\mathcal{E}}$ using Algorithm 3. It is worth noticing that in a triangulation with about 2 million tetrahedra our procedure computes a discrete vector potential under a second. We run our Algorithm 2 with different triangulations of the same metal plate and on each example Algorithm 2 finds a complete acyclic matching $\mathcal{M}_1^{(0)}$. Thus, there is no recursive call of Algorithm 3. For the considered examples in Table 6.1 we can clearly see the linear behaviour of the computational time with respect to the size of the triangulations.

Table 6.1: Running times of Algorithm 3 for the modified TEAM benchmark example for triangulations of decreasing size.

Name	Tets. ($\|C\|$)	Faces ($\|F\|$)	Edges ($\|E\|$)	Vertices ($\|N\|$)	Time [ms]
Mesh 1	1,851,493	3,871,379	2,419,350	399,465	992
Mesh 2	1411688	2847256	1683787	248220	756
Mesh 3	529,664	1,065,104	626,566	91,127	284
Mesh 4	186264	378588	226584	34261	101

6.5.2 Bing's House

A Bing's House is now considered [99]. The simplicial complex, homeomorphic to a 3-dimensional ball, can be obtained by replacing every surface in the Bing's House by a thick wall made of 3-cells. At the end of this procedure we obtain the polyhedron in Fig. 6.4. Although we can informally identify two "chambers", it can be demonstrated that the Bing's House is homeomorphic to the three-dimensional ball.

As in the previous set of tests, Table 6.2 contains information about the number of cells of the considered triangulations together with the computational time required to compute a discrete vector potential. We have found that in almost all runs of Algorithm 2 we get a complete acyclic matching $\mathcal{M}_1^{(0)}$. Only in a few cases we need to resort to a recursive call of Algorithm 3. However, to compute a discrete vector potential it is enough, in all these cases, only one recursive call of Algorithm 3, given that in the first recursion we always find a complete acyclic matching $\mathcal{M}_1^{(1)}$.

To measure the complexity of the first recursive call, we consider the cardinality of the basis $\mathcal{B}_2^{(1)}$, namely the output basis of 2-chains becoming the input for the first recursive call of Algorithm 3. We have found that the cardinality of $\mathcal{B}_2^{(1)}$ is always less than 10 on thousands of algorithm runs with different choices of the acyclic matching $\mathcal{M}_2^{(0)}$ of 2-chains. Accordingly, as reported in Table 6.2, we observe no influence of the recursive call in Algorithm 3 on the linear behaviour of the running times with respect

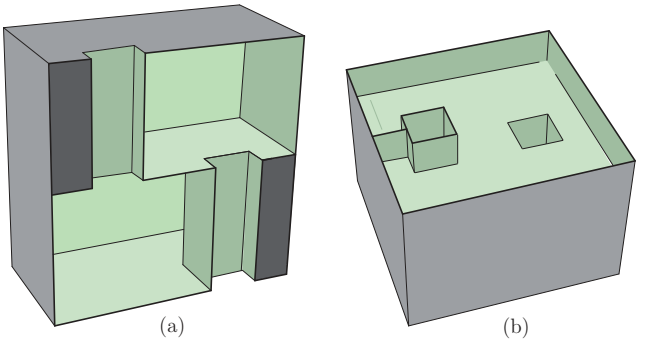

Figure 6.4: Two views of the considered 3-dimensional thickening of a Bing's house with two rooms.

to the size of the triangulations.

Table 6.2: Running times of Algorithm 3 for various triangulations of the thick Bing's House.

Name	Tets. ($\|C\|$)	Faces ($\|F\|$)	Edges ($\|E\|$)	Vertices ($\|N\|$)	Time [ms]
Bing 1	800,020	1,600,537	937,631	137,115	429
Bing 2	87,221	175,317	102,212	14,117	47

6.5.3 Knot-theoretic obstructions

We consider 3-balls of $\mathbb{R}^3$ which admit non-collapsible triangulations.

Obstructions coming from short knots have been considered first in the works [100, 101].

In [100], Bing proved, using knot theory, that some triangulations of the 3-ball are not collapsible. Bing's construction works as follows. One starts with a triangulated 3-ball Ω and introduces a "knotted spanning arc" in its 1-skeleton. A knotted spanning arc is an arc as in Fig. 6.5. We dig a knot-shaped tubular hole inside Ω starting from the top and we stop digging one step before the tunnel go through the bottom of Ω. In this way we obtain a 3-ball Ω' containing a knot having all its edges on the boundary of Ω', except for a single interior edge.

If the knot is sufficiently complicated (like a double, or a triple trefoil), Bing's ball cannot be collapsible [100, 89]. In contrast, if the knot is simple enough (like a single trefoil), then Bing's ball may be collapsible. The construction also appears in the 1924 work [102] of Furch and for the present discussion we refer to it as *Furch's knotted ball*; see [103] (Section 3.1) for an historical account.

Firstly, we consider the simplest case of Furch's knotted ball with only one trefoil knot. Table 6.3 summarizes the geometrical information of this triangulation named Furch 1. In each run of Algorithm 3, we do not find a complete acyclic matching $\mathcal{M}_1^{(0)}$ of 1-chains. Thus, Algorithm 3 is recursively called. However, only one recursive call

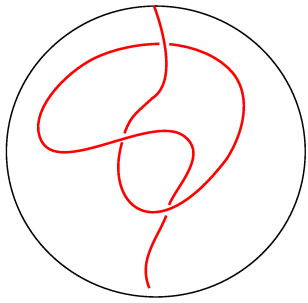

Figure 6.5: A knotted spanning arc in a 3-ball Ω at the core of Furch's construction.

is needed in all the considered runs of Algorithm 3, given that in the first recursion we always find a complete acyclic matching $\mathcal{M}_1^{(1)}$.

To measure the complexity of the first recursive call, we consider, as done in Section 6.5.2, the cardinality of the basis $\mathcal{B}_2^{(1)}$. We have found that the cardinality of $\mathcal{B}_2^{(1)}$ is always less than 30 on thousands of algorithm runs with different choices of the acyclic matching $\mathcal{M}_2^{(0)}$ of 2-chains.

As the last benchmark problem, we consider a more complicated obstruction. We dig one hundred trefoil knots in a parallel-like fashion starting from the top of Ω. Table 6.3 reports again the results for this triangulation named Furch 100. We found that the cardinality of $\mathcal{B}_2^{(1)}$ is always less than 360 on thousands of algorithm runs with different choices of the acyclic matching $\mathcal{M}_2^{(0)}$ of 2-chains. Also in this case, only one recursive call is needed in all considered runs of Algorithm 3, given that in the first recursion we always find a complete acyclic matching $\mathcal{M}_1^{(1)}$.

We observe no influence of the recursive call in Algorithm 3 on the linear behaviour of the running times with respect to the size of the triangulations. This is because of the small cardinality of $\mathcal{B}_2^{(1)}$.

This example show the effectiveness and generality of our method, at least from the practical point of view. In fact, applying the approach in [87] on the Furch 100 triangulation required 5.1 seconds, where we have computed more than one hundred double integral evaluations. Similarly, applying the approach in [104] on the Furch 100 triangulation required 4.5 seconds, where we have constructed and then solved a sparse linear system of more than one hundred equations having as unknowns the symbolic variables employed in the approach. Therefore, Furch 100 is an explicit example of a triangulation on which the approaches [87, 104] perform poorly compared to Algorithm 3, where the same solution is obtained in 129 milliseconds. To have a provably good method which is reliable in practice, Algorithm 3 is expressly needed.

An important point is that in all tested problems we do not see a dependence between the cardinality of the basis $\mathcal{B}_2^{(1)}$ and the number of cells of the input triangulations. Thus, this quantity can be used as an indicator of how easy it is to find a complete acyclic matching $\mathcal{M}_1^{(0)}$ on a given input triangulation, namely, it quantifies the "topological complexity" of the triangulation.

Table 6.3: Running times of Algorithm 3 for the Furch's knotted balls.

Name	Tets. $\|C\|$	Faces ($\|F\|$)	Edges ($\|E\|$)	Vertices ($\|N\|$)	Time [ms]
Furch 100	243062	506619	311547	47991	129
Furch 1	31189	63830	38445	5805	17

6.6 Conclusions

The novel algorithm presented in this chapter was proved to be general, straightforward to implement and outperforms competing state-of-art algorithms in the class of admissible meshes while showing linear average complexity with respect to the input mesh size. By applying discrete Morse theory, we have shown that for the important class of simplicial triangulations we achieve linear computational complexity for all considered test problems. These include, besides real case triangulations having simple topological properties, also pathological triangulations. A challenging test case made of one hundred trefoil knots have been considered and yet the proposed algorithm succeeds in computing a discrete vector potential. Thus, we expect that our algorithm works for every practical mesh that one encounters in practical applications.

Worst-case complexity analysis can be misleading in the analysis of our Algorithm 3. Worst-case complexity analysis of our algorithm assumes that at certain iteration no new matched pairs are possible. In this case we need to employ a sparse linear system solver, which leads to a cubical worst-case complexity with respect to mesh size parameters. Yet, the average complexity is linear in all tested problems. Our point of view is that the reason we see linear computational complexity is because our examples restrict to 3-dimensional cell complexes decomposing bounded domains of $\mathbb{R}^3$ with sufficiently regular boundaries. We have observed that our recursive strategy based on algebraic discrete Morse theory is efficient at solving potential topological obstructions that can appear in 3-dimensional cases. However, to prove that linear worst-case times are guaranteed is still an open problem and will be subject to future work.

II

Integral methods

7

A foreword on integral methods to solve eddy current problems

This part of the thesis is devoted to the analysis of eddy current problems obtained using *integral methods*.

Integral methods are very appealing because, unlike the standard differential formulations discussed in the first part of this thesis, the computational domain is formed by conductors only, so that modelling and meshing of insulators are avoided. This fact renders integral methods particularly efficient for problems involving moving conductors. Considering only conductors is made possible by formulating the eddy current problem with the *Electric Field Integral Equation* (EFIE), which uses the Biot–Savart law as a non-local constitutive relation.

However, integral methods also have two serious drawbacks.

First, the discretization of the EFIE leads to a fully populated generalized mass matrix, called *inductance matrix* or *magnetic matrix* in the electromagnetic context. Dealing with full matrices means that the time spent for their construction and the computer memory to store them scale quadratically when the mesh is refined.

Second, computing entries of inductance matrix is computationally costly since it requires the evaluation of double integrals. Moreover, these double integrals become singular for diagonal entries of inductance matrix due to the singularity of the EFIE. A standard solution, even in recent contributions [105], is to use two different numerical integration rules for the two nested integrals. Unfortunately, as it will be shown in Chapter 9, this solution leads to numerical instability and to poor accuracy in the computation the diagonal elements of the inductance matrix. For instance, as shown in Chapter 9, errors up to 200% may occurs in the evaluation of the diagonal elements if they are computed with a double numerical integration.

The interest in integral methods revived when inductance matrix compression techniques were developed. These techniques exploit the fact that the inductance matrices

have low-rank off-diagonal blocks, so that they can be approximated by using hierarchical matrices ($\mathcal{H}$-matrices) and *Adaptive Cross Approximation* (ACA), see for example [106, 107, 32, 108, 105]. The compression techniques mitigate the first drawback of integral methods, providing a typical compression which ranges from 30% [32] up to 95% [105] of the total occupation of the full matrix. Yet, the time saving is very limited given that the construction of the matrix requires nearly half of the time required to compute the full matrix [32, 105]. This fact limits drastically the range of problems addressable with integral methods.

To overcame the limits due to computational performances, the VoxHenry technique [109, 110] has been recently introduced to solve exactly the same problem addressed in this part of the thesis. It uses the Fast Fourier Transform (FFT) to sensibly speed up the simulation, but it has the strong limitation of working with a Cartesian grids. Such voxelized geometries present the well-known "staircase" error in the geometric representation when a slanted or curved boundary is rendered on a Cartesian grid, the same error which makes one to prefer FE method over Finite Differences. Moreover, a voxels grid prevents the use of local mesh refinement.

This part of the thesis introduces the foundations of a novel compatible integral method for solving eddy current problems that mitigate both issues of integral methods while extending the simulation speed of VoxHenry to arbitrary polyhedral grids.

In this chapter, we survey the state-of-the-art of EIFE for tetrahedral grids. In particular, we recall the standard EFIE and its discretization using the lowest-order Raviart–Thomas (RT) and Rao–Wilton–Glisson (RWG) face basis functions. Finally, we review two standard approaches proposed in literature to enforce implicitly current conservation in the EFIE.

7.1 The eddy current problem

We begin by describing the eddy current problem. As a physical reference framework, we consider a *conducting domain* Ω_c of arbitrary topology. Ω_c is placed under the influence of a time-varying magnetic field. Such a source magnetic field, produced by a known current density $\boldsymbol{J}_s(\boldsymbol{x}, t)$—where $\boldsymbol{x} \in \mathbb{R}^3$ is a point of Ω and t a time instant—flowing in a *source domain* Ω_s, produces an unknown induced current $\boldsymbol{J}$ in Ω_c accordingly to the Faraday–Neumann law. The conductor is characterized by its resistivity ρ (or its reciprocal, i.e. the conductivity σ) while, for simplicity, the whole domain is considered a medium whose magnetic permeability μ is constant in time and uniform in space and it is equal to the vacuum permeability μ_0. In fact, the contributions introduced in what follows can be extended to problems involving magnetic materials by using consolidated techniques to deal with them like [25]. Thus, the extension of the formulation to problems containing magnetic materials is left for further developments. In the continuation, the time-and-space variation of the fields might be sometimes omitted for the sake of brevity.

We consider the total magnetic field

$$\boldsymbol{H}_{\mathrm{t}} \coloneqq \boldsymbol{H} + \boldsymbol{H}_s$$

as a sum of the unknown reaction $\boldsymbol{H}$ of the conducting domain—produced by the current

density $\boldsymbol{J}$ in Ω_c—with the field $\boldsymbol{H}_s$ generated by $\boldsymbol{J}_s$, and the same for the total magnetic induction field

$$\boldsymbol{B}_{\mathrm{t}} = \boldsymbol{B} + \boldsymbol{B}_s$$

in which $\boldsymbol{B}$ is the unknown contribution and $\boldsymbol{B}_s$ the known one produced by the current density $\boldsymbol{J}_s$.

In this framework, it is known that, under the hypothesis of magneto quasi-static approximation, the following set of equations that characterizes the sources of the problem holds in Ω

$$\boldsymbol{\nabla} \cdot \boldsymbol{B}_s = 0, \tag{7.1}$$
$$\boldsymbol{\nabla} \cdot \boldsymbol{J}_s = 0, \tag{7.2}$$
$$\boldsymbol{\nabla} \times \boldsymbol{H}_s = \boldsymbol{J}_s, \tag{7.3}$$

in addition the one that has to be enforced to determine the unknown eddy currents

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0, \tag{7.4}$$
$$\boldsymbol{\nabla} \cdot \boldsymbol{J} = 0, \tag{7.5}$$
$$\boldsymbol{\nabla} \times \boldsymbol{H} = \boldsymbol{J}, \tag{7.6}$$
$$\boldsymbol{\nabla} \times \boldsymbol{E} - \frac{\partial}{\partial t}\boldsymbol{B}_{\mathrm{t}} = \boldsymbol{0}. \tag{7.7}$$

Moreover, constitutive laws read

$$\boldsymbol{B}_{\mathrm{s}} = \mu_0 \boldsymbol{B}_{\mathrm{s}},\ \boldsymbol{x} \in \Omega \tag{7.8}$$
$$\boldsymbol{B}_{\mathrm{t}} = \mu_0 \boldsymbol{H}_{\mathrm{t}},\ \boldsymbol{x} \in \Omega \tag{7.9}$$
$$\boldsymbol{E} = \rho \boldsymbol{J},\ \boldsymbol{x} \in \Omega_c \tag{7.10}$$

where $\boldsymbol{E}$ refers to the unknown electric field related to the unknown current $\boldsymbol{J}$ flowing in the conductor Ω_c.

For the well posedness of the eddy current problem, we also consider regularity condition at infinity for $\boldsymbol{H}$, $\boldsymbol{B}$ and $\boldsymbol{E}$ since Ω is unbounded. In addition, we impose the boundary conditions on $\boldsymbol{J}$ in such a way that

$$\boldsymbol{J} \cdot \boldsymbol{n} = 0,\ \forall \boldsymbol{x} \in \partial\Omega_c$$

where $\partial\Omega_c$ is the boundary of Ω_c and $\boldsymbol{n}$ is an outgoing vector normal to $\partial\Omega_c$ in $\boldsymbol{x}$. For the sake of simplicity in the exposition, we assume that no electrodes are present. The extension to the case with electrodes presents no difficulty and will be presented elsewhere.

7.2 EFIE: discretization on tetrahedral grids

Following [111], we introduce the Hilbert vector spaces $L^2(\Omega_c)$ and $\boldsymbol{L}^2(\Omega_c)$ with the usual scalar products. Then, given the subspaces $H^1_{\mathrm{grad}}(\Omega_c) := \{\varphi \in L^2(\Omega_c) : \boldsymbol{\nabla}\varphi \in \boldsymbol{L}^2(\Omega_c)\}$ and $\boldsymbol{H}_{\mathrm{div}}(\Omega_c) = \{\boldsymbol{J} \in \boldsymbol{L}^2(\Omega_c) : \boldsymbol{\nabla} \cdot \boldsymbol{J} \in L^2(\Omega_c)\}$, we define the closed subspace

$\boldsymbol{H}_{\mathrm{div},0}(\Omega_c) \coloneqq \{\boldsymbol{J} \in \boldsymbol{H}_{\mathrm{div}} : \boldsymbol{\nabla} \cdot \boldsymbol{J} = 0 \text{ in } \Omega_c, \boldsymbol{J} \cdot \boldsymbol{n} = 0 \text{ in } \Gamma\}$.

Because of (7.1) and (7.4), it is customary to introduce a magnetic vector potential $\boldsymbol{A}_{\mathrm{t}}$ so that $\boldsymbol{B}_{\mathrm{t}} = \boldsymbol{\nabla} \times \boldsymbol{A}_{\mathrm{t}}$ thus rewriting Faraday's law in (7.7) as

$$\frac{\partial}{\partial t}\boldsymbol{A}_{\mathrm{t}} + \boldsymbol{E} = -\boldsymbol{\nabla}\varphi \tag{7.11}$$

by using the electric scalar potential φ.

In order to solve eddy currents by means of an integral approach, the linear integral relation called Biot–Savart law linking the magnetic vector potential on an arbitrary point of the space $\boldsymbol{x} \in \Omega$ to a given current density field $\boldsymbol{J}_d$ reads as

$$\boldsymbol{A}_d(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int_\Omega \frac{\boldsymbol{J}_d(\boldsymbol{x}')}{\|\boldsymbol{x} - \boldsymbol{x}'\|}\, d\Omega = \mathcal{G}(\boldsymbol{J}_d) \tag{7.12}$$

via the integral linear operator $\mathcal{G}$ [111]. We note that $\mathcal{G}(\boldsymbol{J}_s)$ gives the vector potential $\boldsymbol{A}_s$ relative to the solution, in a given time instant t, of the magnetostatic problem expressed by (7.1), (7.2), (7.3) and (7.8).

By separating the known contribution $\boldsymbol{A}_s$ of the source domain Ω_s from the unknown part $\boldsymbol{A}$ due the eddy currents in Ω_c, namely $\boldsymbol{A}_{\mathrm{t}} = \boldsymbol{A} + \boldsymbol{A}_s$ with $\boldsymbol{A} = \mathcal{G}(\boldsymbol{J})$ and $\boldsymbol{A}_s = \mathcal{G}(\boldsymbol{J}_s)$, we obtain the Electric Field Integral Equation (EFIE) from (7.11) as

$$\frac{\partial}{\partial t}\mathcal{G}(\boldsymbol{J}) + \frac{\partial}{\partial t}\boldsymbol{A}_s + \boldsymbol{E} = -\nabla\varphi. \tag{7.13}$$

To develop the finite element formulation of (7.13), the partial differential equations must be restated in a *weak form*, which reads

$$\frac{d}{dt}\langle \mathcal{G}(\boldsymbol{J}), \boldsymbol{J}' \rangle + \langle \rho \boldsymbol{J}, \boldsymbol{J}' \rangle = -\frac{d}{dt}\langle \boldsymbol{A}_s, \boldsymbol{J}' \rangle, \ \forall \boldsymbol{J}' \in \boldsymbol{H}_{\mathrm{div},0} \tag{7.14}$$

where (7.10) has been used and plugged into (7.13).

Let us consider a partition of Ω_c into a tetrahedral grid K . The partitioning into tetrahedral elements allows the interpolation of the current density vector field $\boldsymbol{J}$ by means of suitable *face basis functions* $\{\boldsymbol{w}_f\}_{f\in F}$ as

$$\boldsymbol{J} = \sum_{f\in F} \boldsymbol{w}_f \boldsymbol{J}_f^{\mathcal{F}}. \tag{7.15}$$

The standard choice for the vector functions $\{\boldsymbol{w}_f\}_{f\in F}$ is represented by *Raviart–Thomas (RT) basis functions*, also known as *Whitney facet elements*. Another choice, which is equivalent to RT [112] and is popular in the computational electromagnetics community, is the use of *Rao–Wilton–Glisson (RWG) shape functions* firstly introduced for 2D simplicial elements [113] and then extended to tetrahedra [114].

Equation (7.14) with the use of (7.15) and the standard Galerkin method, which sets $\boldsymbol{J}' = \boldsymbol{w}_f$, yields a symmetric system of linear equations as

$$\mathbb{R}\boldsymbol{J}^{\mathcal{F}} + \mathbb{M}\frac{d}{dt}\boldsymbol{J}^{\mathcal{F}} = -\frac{d}{dt}\boldsymbol{A}_s^{\tilde{\mathcal{E}}}. \tag{7.16}$$

Moreover, in (7.16), we impose the boundary conditions $\boldsymbol{J} \cdot \boldsymbol{n} = 0$ on $\partial\Omega_c$ by setting to zero all the DoFs $\boldsymbol{J}_f^{\mathcal{F}}$ related to boundary faces of K.

The *resistance (mass) matrix* $\mathbb{R}$ is constructed by means of a standard FE assembly of local matrices $\mathbb{R}_c$

$$\mathbb{R} = \sum_{c \in C} (\mathbb{O}_c^{\mathcal{F}})^T \mathbb{R}_c \mathbb{O}_c^{\mathcal{F}}. \tag{7.17}$$

On each element c of K the local matrix $\mathbb{R}_c$ is defined as

$$(\mathbb{R}_c)_{f,f'} \coloneqq \int_c \boldsymbol{w}_{f|c}(\boldsymbol{x}) \cdot \rho\, \boldsymbol{w}_{f'|c}(\boldsymbol{x})\, dc. \tag{7.18}$$

being $\boldsymbol{w}_{f|c}$ the restriction of $\boldsymbol{w}_f$ to c.

Similarly, the *magnetic (mass) matrix* $\mathbb{M}$ is constructed by means of a standard FE assembly of local matrices $\mathbb{M}_{cc'}$

$$\mathbb{M} = \sum_{c,c' \in C} (\mathbb{O}_c^{\mathcal{F}})^T \mathbb{M}_{cc'} \mathbb{O}_{c'}^{\mathcal{F}}. \tag{7.19}$$

By considering two elements c and c', that can also be coincident, the local matrix $\mathbb{M}_{cc'}$ is defined as

$$(\mathbb{M}_{cc'})_{f,f'} \coloneqq \frac{\mu_0}{4\pi} \int_c \int_{c'} \frac{\boldsymbol{w}_{f|c}(\boldsymbol{x}) \cdot \boldsymbol{w}_{f'|c'}(\boldsymbol{x}')}{\|\boldsymbol{x} - \boldsymbol{x}'\|}\, dc\, dc'. \tag{7.20}$$

We now deal with the issue of enforcing the condition $\boldsymbol{\nabla} \cdot \boldsymbol{J} = 0$ on (7.15). We present two different standard approaches proposed in literature.

7.2.1 Solution based on the electric vector potential and additional DoFs: the CARIDDI code

Historically, a first solution was proposed in 1985 by Albanese and Rubinacci [26]. The idea is to represent the solenoidal current density as the curl of an *electric vector potential* $\boldsymbol{T}$

$$\boldsymbol{J} = \boldsymbol{\nabla} \times \boldsymbol{T}. \tag{7.21}$$

In the FE, the current density is represented inside the tetrahedron c by

$$\boldsymbol{J} = \sum_{e \in E(c)} \boldsymbol{\nabla} \times \boldsymbol{N}_{e|c}\, \boldsymbol{T}_e^{\mathcal{E}}, \tag{7.22}$$

where $\boldsymbol{N}_{e|c}$ are the *Nédélec's edge basis functions* [115] restricted to element c and DoFs of $\boldsymbol{T}$ are attached to edges of K. This gives rise to the system of equations

$$\mathbb{R}^{\mathrm{CAR}} \boldsymbol{T}^{\mathcal{E}} + \mathbb{L}^{\mathrm{CAR}} \frac{d}{dt} \boldsymbol{T}^{\mathcal{E}} = -\frac{d}{dt} \boldsymbol{E}^{\tilde{\mathcal{E}}}. \tag{7.23}$$

Following the discussion in [27], we have that the local matrix $\mathbb{R}_c^{\mathrm{CAR}}$ is defined as

$$(\mathbb{R}_c^{\mathrm{CAR}})_{e,e'} \coloneqq \int_c \boldsymbol{\nabla} \times \boldsymbol{N}_{e|c}(\boldsymbol{x}) \cdot \rho \boldsymbol{\nabla} \times \boldsymbol{N}_{e'|c}(\boldsymbol{x})\, dc, \tag{7.24}$$

and the local matrix $\mathbb{L}_{|c}^{\mathrm{CAR}}$ is defined as

$$(\mathbb{L}_c^{\mathrm{CAR}})_{e,e'} \coloneqq \frac{\mu_0}{4\pi} \int_c \int_{c'} \frac{\boldsymbol{\nabla} \times \boldsymbol{N}_{e|c}(\boldsymbol{x}) \cdot \boldsymbol{\nabla} \times \boldsymbol{N}_{e'|c'}(\boldsymbol{x}')}{\|\boldsymbol{x} - \boldsymbol{x}'\|}\, dc\, dc'. \tag{7.25}$$

To reduce the unknowns and obtain a full rank system, the so-called *tree-cotree gauge* [116] was developed, which sets to zero entries of $\boldsymbol{T}^{\mathcal{E}}$ corresponding to edges of a suitable spanning tree of K. More details of gauging will be provided in Section 8.4.

It is fundamental to note that (7.23) holds only for simply connected conductors. The most up-to-date extension to conductors that are not simply connected is presented in [117] and it is based on adding some "additional DoFs". However, these additional DoFs do not have a precise mathematical definition so the approach lacks clear theoretical foundations. More importantly, an efficient way to find the global basis functions is missing. With the approach proposed in [117], the computation of such global basis functions may easily take hours given that it requires the solution of many global linear systems.

7.2.2 Solution based on mesh current analysis (MCA): the unstructured Partial Elements Equivalent Circuit (PEEC) for eddy currents

A second approach exploits the so-called *electric circuit interpretation* of integral methods, see for example [30, 105]. Then, the problem is solved by applying the *mesh current analysis* (MCA), which is a standard method of network theory [118, 119]. This approach, proposed for the first time as far as we know in [30], computes a cycle basis on the circuit graph with the help of a tree-cotree decomposition. In particular, once one adds a cotree dual edge to the tree of the dual graph, exactly one loop is created. The set of these loops built for all cotree edges forms a cycle basis of the graph [118, 119]. These *global cycles* may be interpreted on the primal complex as a set of faces which produce a basis for solenoidal currents. Greater details are provided in Section 8.4. This approach has been recently rediscovered and called *unstructured* Partial Elements Equivalent Circuit (PEEC) for eddy currents, see for example [120].

8

Mimetic Volume Integral method

In this chapter, we present the *Mimetic Volume Integral* (MVI) method, an extension of integral methods to general polyhedral grids.

For meshes composed by general polyhedra, the only integral method that we are aware to produce a consistent, symmetric and positive definite mass matrices is the approach in [121]. It is based on the piecewise uniform basis functions introduced in [122]. Moreover, for hexahedra, prisms and pyramids one may use standard RT basis functions.

In Section 8.1 we show that mass matrices produced with RT or RWG face basis functions for hexahedra are *not* consistent. Hence, such mass matrices cannot be generalized to arbitrary polyhedral elements.

In Section 8.2 we present a novel construction of resistance and magnetic mass matrice that are positive definite, symmetric and consistent for arbitrary polyhedral elements. The construction of these mass matrices is inspired by a well established design strategy used throughout this thesis which decomposes local mass matrices as the sum of a consistent and a stabilization part. These novel mass matrices can be thought as the generalization of RWG and RT basis function for hexahedral or even general polyhedral elements because, as shown in Section 8.4, they produce the same stiffness matrix as the RT and RWG in case of tetrahedral grids. In Section 8.4 we also show that our novel MVI produce the same solution in terms of current density of the standard approaches for tetrahedral grids reviewed in Chapter 7. Yet, our novel MVI should be preferred with respect to others also for the particular case of tetrahedral grids. Indeed, the splitting of matrices into a consistent and a stabilization part has strong implications from both a theoretical and a computational point of view that will be analysed in Chapter 9.

Finally, it is also mentioned that another point of view of integral methods is their interpretation in terms of electrical networks, see for example [30, 105]. In Section 8.3 we show how a rigorous interpretation of equation (7.16) in terms of electric circuits is available in the MVI framework.

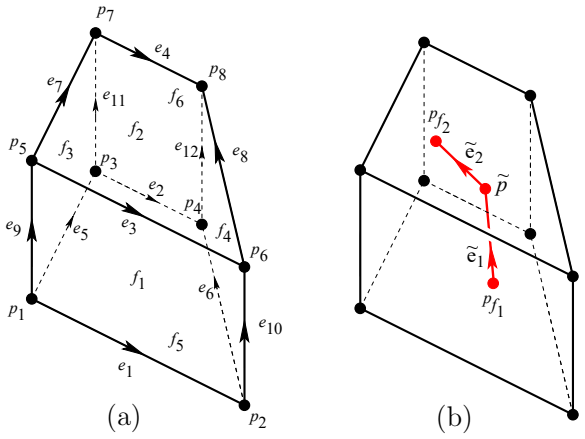

Figure 8.1: The hexahedron v used in the counterexample. The coordinates of the nodes are $\boldsymbol{p}_{f_1} = (0,0,0)^T$, $\boldsymbol{p}_{f_2} = (2,0,0)^T$, $\boldsymbol{p}_{f_3} = (0,1,0)^T$, $\boldsymbol{p}_4 = (1,1,0)^T$, $\boldsymbol{p}_5 = (0,0,1)^T$, $\boldsymbol{p}_6 = (2,0,1)^T$, $\boldsymbol{p}_7 = (0,1,1)^T$, $\boldsymbol{p}_8 = (1,1,1)^T$.

8.1 Generalization to hexahedral and polyhedral meshes

The aim of this section is to show that extending the integral method on hexahedral or general polyhedral meshes is not trivial. A first indication is that the circuit interpretation of CARIDDI and PEEC methods on hexahedral meshes by using standard Raviart–Thomas face basis functions [123] lacks theoretical foundations.

We start by claiming that the Raviart–Thomas mass matrix on a hexahedron is *not consistent*. Let us consider the hexahedron c as in Fig. 8.1. We assume a unitary uniform resistivity. We denote by p_f the intersection between each face f of c and the corresponding dual edge $\tilde{e}_{f|c}$. Let $\tilde{n}_c$ be the dual node dual c. We stress that, for the present discussion, the points p_f and $\tilde{n}_c$ are arbitrary positions and do not necessarily coincide with the barycenter of face f and hexahedron c, respectively. A necessary and sufficient condition for the consistency of $\mathbb{R}_c$ according to the definition reported in [17, 124] is

$$\mathbb{R}_c\mathbb{F}_c = \tilde{\mathbb{E}}_c, \tag{8.1}$$

where $\mathbb{F}_{|c}$ and $\tilde{\mathbb{E}}_c$ are defined as in . In what follows we will prove that condition (8.1) are *not* satisfied for *any* choice of the points $\boldsymbol{p}_f$ for $f \in F(c)$. By direct computation, the right hand side of (8.1) yields

$$\begin{aligned}
&\mathrm{row}_i(\mathbb{R}_c\mathbb{F}_c) = (0,0,3\log(2)/4), && \text{with } i = 1,\ 2\\
&\mathrm{row}_i(\mathbb{R}_c\mathbb{F}_c) = (3/4,0,0), && \text{with } i = 3,\ 4\\
&\mathrm{row}_i(\mathbb{R}_c\mathbb{F}_c) = (-1/4,1/2,0), && \text{with } i = 5,\ 6,
\end{aligned}$$

where with row_i we denote the ith row of a matrix. Let us consider the vectors $\tilde{\boldsymbol{e}}_{f_1|c}$, $\tilde{\boldsymbol{e}}_{f_2|c}$ associated with the dual edges $\tilde{e}_{f_1|c}$, $\tilde{e}_{f_1|c}$ respectively; in order to guarantee that

$\tilde{\boldsymbol{e}}_{f_1|c}$, $\tilde{\boldsymbol{e}}_{f_2|c}$ are parallel to the vectors $\mathrm{row}_i(\mathbb{R}_c\mathbb{F}_c) = (0, 0, 3\log(2)/4)$, with $i = 1, 2$, it is necessary that $\tilde{\boldsymbol{n}}_c$, $\boldsymbol{p}_{f_1}$ and $\boldsymbol{p}_{f_2}$ are on a straight line parallel to the z-axis of the Cartesian coordinate system. Thus, by assuming for $\tilde{\boldsymbol{n}}_c = (x_2, y_2, x_2)^T$ it results in $\boldsymbol{p}_{f_1} = (x_2, y_2, 0)^T$, $\boldsymbol{p}_{f_2} = (x_2, y_2, 1)^T$. Then, given that $\hat{\boldsymbol{z}}$ is the unitary vector $(0, 0, 1)^T$, it is

$$\begin{aligned}&\frac{3}{2}\log(2) = (\mathrm{row}_1(\mathbb{R}_c\mathbb{F}_c) + \mathrm{row}_2(\mathbb{R}_c\mathbb{F}_c)) \cdot \hat{\boldsymbol{z}} = \\ &(\tilde{\boldsymbol{e}}_{f_1|c} + \tilde{\boldsymbol{e}}_{f_2|c} \cdot \hat{\boldsymbol{z}} = ((\tilde{\boldsymbol{n}}_c - \boldsymbol{p}_{f_1}) + (\boldsymbol{p}_{f_2} - \tilde{\boldsymbol{n}}_c)) \cdot \hat{\boldsymbol{z}} = \\ &((0, 0, x_2)^T + (0, 0, 1 - x_2)^T) \cdot \hat{\boldsymbol{z}} = 1\end{aligned}$$

which is clearly false.

Hence, the rows of the matrix

$$\mathbb{R}_c^{\mathrm{RT}}\mathbb{F}_c, \tag{8.2}$$

where $\mathbb{R}_c^{\mathrm{RT}}$ is the Raviart–Thomas resistance mass matrix , do not represent a dual grid structure on which a constant electric field in c can be evaluated. This, in our opinion, renders the interpretation in terms of circuits questionable whenever applied to an hexahedral grids whose constitutive matrices are computed by means of Raviart–Thomas shape functions.

To generalize in a consistent way the integral formulations for eddy currents to hexahedral meshes and even the most general polyhedral meshes we need to use our novel MVI framework. This framework not only allows to obtain a consistent matrix on a mesh constituted by arbitrary polyhedra but it also enables to present the novel and original results contained in the next sections.

8.2 A novel Mimetic Volume Integral (MVI) method

8.2.1 Consistent and positive-definite resistance and magnetic mass matrices

An efficient recipe to construct consistent and symmetric mass matrices exploits the constant basis function introduced in Theorem 2, which are able to represent a constant vector field defined in each element.

The local resistance matrix $\mathbb{R}_c$, built for c whose resistivity tensor is $\mathbb{K}_\rho$, is constructed as

$$(\mathbb{R}_c)_{f,f'} = \int_c \tilde{\boldsymbol{e}}_{f|c} \cdot \mathbb{K}_\rho \tilde{\boldsymbol{e}}_{f'|c}\, dc = |c|\, \tilde{\boldsymbol{e}}_{f|c} \cdot \mathbb{K}_\rho \tilde{\boldsymbol{e}}_{f'|c}. \tag{8.3}$$

Similarly, the local magnetic mass matrix $\mathbb{M}_{cc'}$ between two considered elements c and c' is

$$(\mathbb{M}_c)_{f,f'} = \frac{\mu_0}{4\pi}\int_c\int_{c'} \frac{\tilde{\boldsymbol{e}}_{f|c} \cdot \tilde{\boldsymbol{e}}_{f'|c'}}{\|\boldsymbol{x} - \boldsymbol{x}'\|}\, dc\, dc' = t^{cc'}\, \tilde{\boldsymbol{e}}_{f|c} \cdot \tilde{\boldsymbol{e}}_{f'|c}, \tag{8.4}$$

where $t^{cc'}$ is the positive number defined as follows

$$t^{cc'} := \frac{\mu_0}{4\pi} \int_c \int_{c'} \frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|} dc, dc', \tag{8.5}$$

Global matrices $\mathbb{R}$ and $\mathbb{M}$ are assembled as in (7.17) and (7.19), respectively.

To obtain positive definite mass matrices, we add to the local mass matrix $\mathbb{R}_c$ a stabilization matrix which is symmetric and positive semidefinite and we look for a suitable stabilization matrix to be added to the dense $\mathbb{M}_{cc'}$.

Thus, as done in Chapter 2, the local matrix $\mathbb{R}_{s|c}$ defined as

$$\mathbb{R}_c^s = \mathbb{R}_c + \mathbb{S}_c. \tag{8.6}$$

is symmetric, consistent and positive definite.

Now, let $\mathbb{M}_{cc'}^s$ be the local matrix defined as follows

$$\mathbb{M}_{cc'}^s := \begin{cases} \mathbb{M}_{cc'} + \mathbb{S}_c & \text{if } c = c' \\ \mathbb{M}_{cc'} & \text{if } c \neq c' \end{cases}. \tag{8.7}$$

Theorem 13. *Matrix $\mathbb{M}^s$ is symmetric and positive-definite.*

Proof. Let $\boldsymbol{z} \in \mathbb{R}^{|F|}$ such that $\boldsymbol{z}^T \mathbb{M}^s \boldsymbol{z} = 0$. In order to prove that $\mathbb{M}_s$ is positive definite, we have to show that $\boldsymbol{z} = 0$. We have

$$\begin{aligned} 0 &= \boldsymbol{z}^T \mathbb{M}^s \boldsymbol{z} \\ &= \sum_{c,c' \in C} \boldsymbol{z}^T (\mathbb{O}_c^{\mathcal{F}})^T \mathbb{M}_{cc'}^s \mathbb{O}_{c'}^{\mathcal{F}} \boldsymbol{z} \\ &= \sum_{c \in C} \boldsymbol{z}^T (\mathbb{O}_c^{\mathcal{F}})^T \mathbb{M}_{cc}^s \mathbb{O}_c^{\mathcal{F}} \boldsymbol{z} + \sum_{c,c' \in C, c' \neq c} \boldsymbol{z}^T (\mathbb{O}_c^{\mathcal{F}})^T \mathbb{M}_{cc'}^s \mathbb{O}_{c'}^{\mathcal{F}} \boldsymbol{z}. \end{aligned} \tag{8.8}$$

(8.8) is equivalent to require that

$$\sum_{c \in C} \boldsymbol{z}^T (\mathbb{O}_c^{\mathcal{F}})^T (\mathbb{M}_{cc} + \mathbb{S}_c) \mathbb{O}_c^{\mathcal{F}} \boldsymbol{z} = 0, \tag{8.9}$$

$$\sum_{c,c' \in C, c' \neq c} \boldsymbol{z}^T (\mathbb{O}_c^{\mathcal{F}})^T \mathbb{M}_{cc'} \mathbb{O}_{c'}^{\mathcal{F}} \boldsymbol{z} = 0, \tag{8.10}$$

where we have used (8.7). By repeating the same argument used in the proof of Theorem 1, it follows that each matrix $\mathbb{M}_{cc}^s + \mathbb{S}_c$ is positive definite. Thus, from (8.9) it follows that $\boldsymbol{z} = 0$. As a consequence, (8.10) is also satisfied. □

8.2.2 Enforcing discrete solenoidal current

In the MVI framework we enforce a discrete solenoidal current (formally, a 2-cocycle, as defined in) into the EFIE as

$$\mathbb{D}\boldsymbol{J}^{\mathcal{F}} = \boldsymbol{0}. \tag{8.11}$$

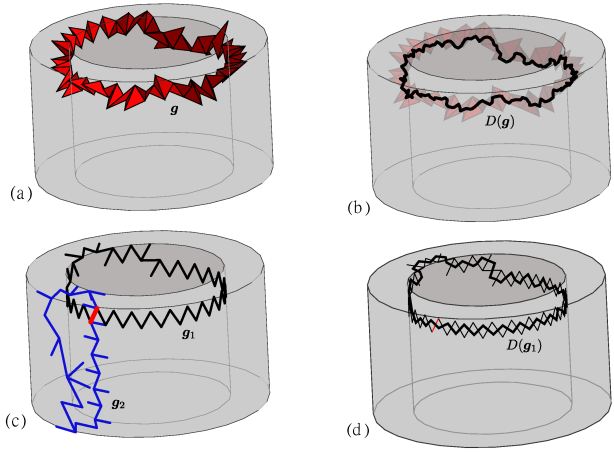

Figure 8.2: Examples of cohomology generators $H^2(K,\partial K)$ and $H^1(\partial K)$ for a solid torus. a) The support of a representative $\boldsymbol{g} \in H^2(K,\partial K)$ generator. b) The dual of $\boldsymbol{g}$ is a cycle made of dual edges that are dual to the faces of g. c) The support of two representatives $\boldsymbol{g}_1, \boldsymbol{g}_2$ of the $H^1(\partial K)$ generators of ∂K. d) The support of the homology generator $D(g_1)$ is constituted by dual edges that are dual to the primal edges of $\boldsymbol{g}_1$.

The so-called *cohomology theory* tells us formally—by its very definition—that all discrete solenoidal currents $\boldsymbol{J}^{\mathcal{F}}$ satisfying (8.11) can be spanned as

$$\boldsymbol{J}^{\mathcal{F}} = \mathbb{C}\boldsymbol{T}^{\mathcal{E}} + \mathbb{W}\boldsymbol{I}^{\mathcal{F}}, \tag{8.12}$$

where $\boldsymbol{T}^{\mathcal{E}}$ are DoFs array of the electric vector potential attached on grid edges (formally, a 1-cochain), $\boldsymbol{I}^{\mathcal{F}}$ the so-called *independent currents* [125], [121] and the columns of $\mathbb{W}$ store the representatives of generators of the *second relative cohomology group* $H^2(K,\partial K)$ [126], see Fig. 8.2a.

The second relative cohomology group $H^2(K,\partial K)$, by its very definition, spans solenoidal fields tangent to ∂K that are not curl of anything. As an example, the matrix $\mathbb{W}$ for a solid toric conductor is formed by a single column whose entries, interpreted as electric current DoFs, form a unit current that flows through the red thin tube around the torus, see Fig. 8.2a. The independent currents $\boldsymbol{I}^{\mathcal{F}}$ are additional unknowns of the eddy current problem. Their number is usually very small since it depends on the topology of the conductor, in particular its number of "handles". We also note that the dual of the faces in the red thin tube form a dual cycle made by dual edges that goes around the torus like in Fig. 8.2b.

Efficient algorithms for the automatic computation of matrix $\mathbb{W}$ have been proposed [125]. As shown in [125], $\mathbb{W}$ may be constructed as $\mathbb{W} = \mathbb{C}\mathbb{H}$, where columns of matrix $\mathbb{H}$ stores suitable representative of the *first cohomology group* $H^1(\partial K)$ [126] generator. Indeed, for efficiency, it is preferable to construct $\mathbb{W}$ by working on ∂K only. This is because there are less geometric elements to process in ∂K than in K and the algorithms are intrinsically simpler since they exhibit linear worst-case complexity. In case of the example, the representatives $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ of the two boundary generators for a torus are

shown in Fig. 8.2c; they are stored into the column of $\mathbb{H}$. In this case, $\boldsymbol{g}$ can be obtained as $\mathbb{C}\boldsymbol{g}_1$. Moreover, the dual $D(\boldsymbol{g}_1)$ of $\boldsymbol{g}_1$, showed in Fig. 8.2d, is a dual cycle homologous to $D(\boldsymbol{g})$, where D denotes the duality map.

Yet, there is also a theoretical downside of using ∂K only: the major difficulty here is that the $H^1(\partial K)$ cohomology group produces twice the number of generators of a $H^2(K, \partial K)$ basis. For example, when dealing with the solid torus depicted in Fig. 8.2, the two boundary generators able to represent the poloidal and toroidal currents that flow in ∂K appear. A first solution proposed in [125] and [121] suggests to use all representatives of the $H^1(\partial K)$ basis to produce the $\mathbb{W}$ matrix by pre-multiplying the representative by the $\mathbb{C}$ matrix. The obtained $\mathbb{W}$ is called a *lazy cohomology basis* and the obtained system turns out to be singular. Yet, most iterative and direct solvers do not have any problem in solving it since it is algebraically consistent. On the contrary, if one wants for some reason to obtain a full rank system, a cheap technique to find the required change of cohomology basis to obtain the matrix $\mathbb{H}$ has been introduced in [125] and described in more details in [105]. Consequently, by using $\mathbb{W} = \mathbb{C}\mathbb{H}$, the current is thus represented by

$$\boldsymbol{J}^{\mathcal{F}} = \mathbb{C}\left(\boldsymbol{T}^{\mathcal{E}} + \mathbb{H}\boldsymbol{I}^{\mathcal{F}}\right). \tag{8.13}$$

By enforcing the discrete Faraday's law locally on the boundary of all dual faces as $\mathbb{C}^T\boldsymbol{E}^{\tilde{\mathcal{E}}} + i\omega\mathbb{C}^T(\boldsymbol{A}^{\tilde{\mathcal{E}}} + \boldsymbol{A}_s^{\tilde{\mathcal{E}}}) = \boldsymbol{0}$ and globally on the non-local dual cycles like $D(\boldsymbol{g}) = D(\mathbb{C}\boldsymbol{g}_1)$ of Fig. 8.2b as $\mathbb{H}^T\left(\mathbb{C}^T\boldsymbol{E}^{\tilde{\mathcal{E}}} + i\omega\mathbb{C}^T(\boldsymbol{A}^{\tilde{\mathcal{E}}} + \boldsymbol{A}_s^{\tilde{\mathcal{E}}})\right) = \boldsymbol{0}$, the complete set of equations reads as

$$\begin{pmatrix} \mathbb{K} & \mathbb{K}\mathbb{H} \\ \mathbb{H}^T\mathbb{K} & \mathbb{H}^T\mathbb{K}\mathbb{H} \end{pmatrix} \begin{pmatrix} \boldsymbol{T}^{\mathcal{E}} \\ \boldsymbol{I}^{\mathcal{F}} \end{pmatrix} = \begin{pmatrix} -i\omega\mathbb{C}^T\boldsymbol{A}_s^{\tilde{\mathcal{E}}} \\ -i\omega\mathbb{H}^T\mathbb{C}^T\boldsymbol{A}_s^{\tilde{\mathcal{E}}} \end{pmatrix}, \tag{8.14}$$

where

$$\mathbb{K} = \mathbb{C}^T\left(\mathbb{R} + i\omega\mathbb{M}\right)\mathbb{C}. \tag{8.15}$$

Exactly like in CARIDDI, we set to zero the entries of the array $\boldsymbol{T}^{\mathcal{E}}$ relative to edges on ∂K because of boundary conditions and, to reduce the unknowns and obtain a full rank system, one may apply the *tree-cotree gauge* [116] by setting to zero the entries of the array $\boldsymbol{T}^{\mathcal{E}}$ on a suitable tree inside K.

8.3 Novel interpretation as electrical circuits

The idea of interpreting (7.16) in terms of electric circuits has been already discussed in the literature but without any theoretical explanation on how this is possible. The aim of this section is to provide a rigorous interpretation of equation (7.16) in terms of electric circuits.

Let us start by showing what is the physical interpretation of the edge DoFs $\boldsymbol{E}^{\tilde{\mathcal{E}}}$ in

$$\boldsymbol{E}^{\tilde{\mathcal{E}}} = \mathbb{R}\boldsymbol{J}^{\mathcal{F}}, \tag{8.16}$$

where $\boldsymbol{J}^{\mathcal{F}}$ is the vector of face DoFs and $\mathbb{R}$ is the resistance mass matrix. It turns out, from the study of the measurement units, that each DoF of $\boldsymbol{E}^{\tilde{\mathcal{E}}}$ is a voltage. It is therefore legitimate to ask whether there exists a path on which this voltage is sampled.

Let us now consider the element-wise uniform current density

$$\boldsymbol{J} = \sum_{f \in F} \boldsymbol{w}_f \boldsymbol{J}_f^{\mathcal{F}} \tag{8.17}$$

and define the element-wise uniform electric field

$$\boldsymbol{E} = \rho \boldsymbol{J}. \tag{8.18}$$

Since also $\boldsymbol{E}$ is uniform inside each element c and the following property holds, see [17],

$$\int_c \boldsymbol{w}_{f|c}\, dc = \tilde{\boldsymbol{e}}_{f|c} \tag{8.19}$$

we have

$$\int_c \boldsymbol{E} \cdot \boldsymbol{w}_{f|c}\, dc = \boldsymbol{E} \cdot \int_c \boldsymbol{w}_{f|c}\, dc = \boldsymbol{E} \cdot \tilde{\boldsymbol{e}}_{f|c}. \tag{8.20}$$

Thus, we can finally claim that the entry $\boldsymbol{E}_e^{\mathcal{E}}$ of the vector $\boldsymbol{E}^{\tilde{\mathcal{E}}}$ stores the integral of the electric field $\boldsymbol{E}$ along the *dual edge* $\tilde{e}_f$

$$\boldsymbol{E}_e^{\mathcal{E}} = \int_\Omega \boldsymbol{E} \cdot \boldsymbol{w}_{f|c}\, dc = \boldsymbol{E} \cdot \left(\tilde{\boldsymbol{e}}_{f|c} + \tilde{\boldsymbol{e}}_{f|c'}\right) = \int_{\tilde{e}_f} \boldsymbol{E} \cdot d\boldsymbol{l}, \tag{8.21}$$

where $\{c, c'\} = C(f)$ are the only two tetrahedra sharing the face f. This geometric interpretation of the finite elements is at the root of the DGA method [122], [124].

Afterwards, we move on to the physical interpretation of $\boldsymbol{A}^{\tilde{\mathcal{E}}} = \mathbb{M}\boldsymbol{J}^{\mathcal{F}}$. It turns out, from the study of measurement units, that each DoF of $\boldsymbol{A}^{\tilde{\mathcal{E}}}$ is the line integral of the magnetic vector potential on the dual edges, whose time derivative is related to the voltage on dual edges. Thus, in analogy to what was done before, we ask ourself if there exists a path on which this voltage is sampled. To this end, let us consider the vector field

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi} \int_\Omega \frac{\boldsymbol{J}(\boldsymbol{x}')}{\|\boldsymbol{x} - \boldsymbol{x}'\|}\, d\Omega. \tag{8.22}$$

where $\boldsymbol{J}$ is defined as in (8.17). Yet, contrary to the previous case, the entry of $\boldsymbol{A}^{\tilde{\mathcal{E}}}$ corresponding to dual edge $\tilde{e}_f$, namely $\boldsymbol{A}_f^{\mathcal{E}}$, cannot be directly interpreted as the line integral of the magnetic vector potential on dual edges. Indeed, we have that

$$\boldsymbol{A}_f^{\mathcal{E}} = \int_\Omega \boldsymbol{A} \cdot \boldsymbol{w}_f\, d\Omega \tag{8.23}$$

is in general different from

$$\int_{\tilde{e}_{f_i}} \boldsymbol{A} \cdot d\boldsymbol{l}. \tag{8.24}$$

To get an interpretation of $\boldsymbol{A}^{\tilde{\mathcal{E}}}$ as the line integral of the magnetic vector potential

on dual edges, we introduce the element-wise uniform vector field

$$\bar{\boldsymbol{A}} = \frac{1}{|c|}\int_c \boldsymbol{A}\, dc, \tag{8.25}$$

namely the average vector field of $\boldsymbol{A}$ on each volume c. Since $\bar{\boldsymbol{A}}$ is element-wise uniform we clearly have

$$\int_c \bar{\boldsymbol{A}}\cdot \boldsymbol{w}_{f_{|c}}\, dc = \bar{\boldsymbol{A}}\cdot\int_c \boldsymbol{w}_{f_{|c}}\, dc = \bar{\boldsymbol{A}}\cdot\tilde{\boldsymbol{e}}_{f_{|c}}. \tag{8.26}$$

so that (8.23) can be recast as

$$\bar{\boldsymbol{A}}_f^{\mathcal{E}} = \int_\Omega \bar{\boldsymbol{A}}\cdot\boldsymbol{w}_f\, d\Omega = \bar{\boldsymbol{A}}\cdot\left(\tilde{\boldsymbol{e}}_{f_{|c}} + \tilde{\boldsymbol{e}}_{f_{|c'}}\right) = \int_{\tilde{e}_f}\bar{\boldsymbol{A}}\cdot d\boldsymbol{l}, \tag{8.27}$$

where, as before, c and c' are the only two tetrahedra sharing the face f.

We now prove that the difference between $\bar{\boldsymbol{A}}_f^{\mathcal{E}}$ and $\boldsymbol{A}_f^{\mathcal{E}}$ exhibits a faster convergence order than the discretization error when the mesh is refined. Thus, the approximation made in replacing $\boldsymbol{A}$ with $\bar{\boldsymbol{A}}_f^{\mathcal{E}}$ is negligible in practice and yet it allows to formally introduce a network interpretation of the physical quantities. Let us consider the difference on a single volume c

$$\int_c(\boldsymbol{A}-\bar{\boldsymbol{A}})\cdot\boldsymbol{w}_{f_{|c}}\, dc = \frac{\tilde{\boldsymbol{e}}_{f_{|c}}}{|c|}\cdot\int_c(\boldsymbol{A}-\bar{\boldsymbol{A}})\, dc + \int_c(\boldsymbol{A}-\bar{\boldsymbol{A}})\cdot\left(\boldsymbol{w}_{f_{|c}} - \frac{\tilde{\boldsymbol{e}}_{f_{|c}}}{|c|}\right) dc. \tag{8.28}$$

First, note that $\int_c(\boldsymbol{A}-\bar{\boldsymbol{A}})\, dc$ is zero because of the definition of $\bar{\boldsymbol{A}}$. Second, $(\boldsymbol{A}-\bar{\boldsymbol{A}})$ and $(\boldsymbol{w}_{f_{|c}} - \frac{\tilde{\boldsymbol{e}}_{f_{|c}}}{|c|})$ are at least linear fields with zero average so that $|\boldsymbol{A}_f^{\mathcal{E}} - \bar{\boldsymbol{A}}_f^{\mathcal{E}}|$ exhibits a second convergence order with respect to the mesh size.

Only now that we know that a path on which voltages are defined exists, we can interpret the discretized EFIE (7.16) with an electric circuit that can be solved by standard methods of *network analysis*. The graph of the electrical network is thus formed by dual nodes and dual edges of $\mathcal{K}$. We remark that, because of boundary conditions, the dual edges which are dual to faces in $\partial\mathcal{K}$ do not belong to the dual graph.

Thanks to this interpretation, the left hand side of equation (7.16) can be regarded as the resistive and inductive voltage drop caused by the current flow. This, in the frequency domain, configures $\mathbb{R}+i\omega\mathbb{M}$ as the impedance of Ω_c which writes

$$\mathbb{Z}\boldsymbol{J}^{\mathcal{F}} = -i\omega\boldsymbol{A}_s^{\tilde{\mathcal{E}}} \tag{8.29}$$

being $\omega = 2\pi f$ the angular frequency and i the imaginary unit.

The easiest method to introduce the EFIE for eddy currents is the one based on MCA, known also as unstructured PEEC for eddy currents, because all topological issues are seamlessly taken into account[1] when computing the cycle basis of the dual graph. Yet, as a downside, there is no control on the length of the cycle basis. In

[1] Actually, the computation of the cycle basis on the dual graph is exactly the computation of the generators of the $H_1(\mathcal{C},\mathbb{Z})$ homology group [127, p. 506].

practice each cycle runs through a large portion of the mesh. This fact has terrible consequences, given that state-of-the-art system matrix compression techniques do not work at all on matrices produced with such a cycle basis.

It is therefore natural to ask if there are techniques to find the *shortest cycle basis* or at least one of its close approximations. *Shortest* in our case means a basis with the minimum number of total dual edges contained in all the cycles. This is not directly feasible with graph theoretic algorithms, since the complexity of general algorithms to compute a minimal cycle basis is cubical in the number of graph edges. Therefore, a practical solution has to exploit some additional structure of the problem.

8.4 Bridging all volumetric integral methods for solving eddy currents

In this section we show the relationship between MVI, MCA and CARRIDI. Even if developed independently in the literature, it will be shown that all three methods return the same solution up to machine precision or linear solver algebraic error because they are algebraically equivalent (i.e. they enforce equivalent constraints) for tetrahedral grids. Nevertheless, even if they are identical concerning accuracy of the solution, they are *not* in terms of computational efficiency especially when applied to large problems that requires advanced compression algorithms, see Chapter 9.

8.4.1 Equivalence of MVI with MCA and CARRIDI on tetreahedral grids

We first remark that when constructing $\mathbb{C}^T \left(\mathbb{R}_s + i\omega\mathbb{M}_s\right)\mathbb{C}$ in tetrahedral meshes the stabilization part is not needed. In other words

$$\mathbb{C}^T \left(\mathbb{R}_s + i\omega\mathbb{M}_s\right)\mathbb{C} = \mathbb{C}^T \left(\mathbb{R} + i\omega\mathbb{M}\right)\mathbb{C}. \tag{8.30}$$

This fact has two consequences. First, the construction of the resistance and inductance matrices may be performed by using just the consistent part encompassing the constant vector fields inside the elements, since the construction of the stabilization matrix can be avoided. Second, since the constant vector field reconstructed from solenoidal currents on grid faces is unique, it turns out that

$$\mathbb{C}^T \left(\mathbb{R} + i\omega\mathbb{M}\right)\mathbb{C} = \mathbb{C}^T \left(\mathbb{R}^{\mathrm{RWG}} + i\omega\mathbb{M}^{\mathrm{RWG}}\right)\mathbb{C}. \tag{8.31}$$

Yet another consequence is that the RT and RWG mass matrices can be interpreted as the one obtained with VU basis functions with a different choice for the stabilization part.

All of this implies that the accuracy of the novel method based on VU basis functions produces *exactly* the same matrices and the *same solution* of the one from standard Raviart–Thomas or RWG basis functions. In Chapter 9 we will show why the VU in the MVI framework should be preferred even in case of tetrahedral meshes.

We now show that MCA and CARRIDI just differ in the dual cycles bases used, therefore they can be thought as equivalent from the theoretical point of view and their

solutions are the same.

VINCO is a MCA method with a short cycle basis

The main idea is that the boundary of a dual face is a "short" dual cycle and we can try to build a cycle basis from the *local cycles* produced by taking the boundaries of all dual faces. What is missing is just a recipe to extract a set of local cycles in such a way they form a basis.

How many independent local loops one has to add? To answer this question we note that, thanks to Stokes theorem, local cycles are dependent if their dual faces form a closed surface. Therefore, we have to avoid closed surfaces in the set of dual faces. This can be realized by removing the local cycles relative to dual faces that are dual to a primal edges which belong to a suitable spanning tree of K.

There is another issue. When the topology is not trivial, local loops alone are not able to span all the cycle basis. By definition of homology, what is needed in addition is a $H_1(\tilde{K})$ homology basis. The representatives of this homology basis are called *global dual cycles* and they are obtained as the dual of the cohomology generators. For example, the global dual cycle in Fig. 8.2b is obtained by $D(\boldsymbol{g})$ or by $D(\mathbb{C}\boldsymbol{g}_1)$.

What follows is a formal proof of the claim of this section. We show in particular that the rank of the cycle basis obtained by MVI method is the same as the one obtained by the MCA method. We start from the Euler formula for combinatorial 3-manifolds

$$|N| - |E| + |F| - |C| = \beta_0(K) - \beta_1(K) + \beta_2(K), \tag{8.32}$$

where $\beta_i(K)$ is the ith *Betti number* [126] of the polyhedral grid K. We write also the Euler formula for combinatorial 2-manifolds like the boundary of K:

$$|N_\partial| - |E_\partial| + |F_\partial| = \beta_0(\partial K) - \beta_1(\partial K) + \beta_2(\partial K), \tag{8.33}$$

where $|N_\partial|$, $|E_\partial|$ and $|F_\partial|$ are the number of nodes, edges and faces contained in ∂K, respectively. We remark that in the considered case ∂K is a surface without boundary. A more general setting will be considered in a forthcoming work.

Here we find how many cycles are contained inside the cycle basis of the MCA method. Considering the boundary conditions, the number c_{MCA} of internal loop currents is

$$m^{\mathrm{MCA}} = (|F| - |F_\partial|) - |C| + \beta_0(K), \tag{8.34}$$

since the edges of the tree are number of available dual edges minus dual nodes plus the number of connected components.

From (8.32) and (8.33) we have that

$$|F| - |C| = \beta_0(K) - \beta_1(K) + \beta_2(K) + |E| - |N| \tag{8.35}$$

and

$$|F_\partial| = \beta_0(\partial K) - \beta_1(\partial K) + \beta_2(\partial K) + |E_\partial| - |N_\partial|. \tag{8.36}$$

Let us substitute these two (8.35) and (8.36) inside (8.34) to get

$$m^{\mathrm{MCA}} = [\beta_0(K) - \beta_1(K) + \beta_2(K) + |E| - |N|] + \beta_0(K) \tag{8.37}$$
$$- [\beta_0(\partial K) - \beta_1(\partial K) + \beta_2(\partial K) + |E_\partial| - |N_\partial|]\,. \tag{8.38}$$

Let us rearrange the terms as

$$m^{\mathrm{MCA}} = |E| - |N| + \beta_0(K) + [\beta_0(K) - \beta_0(\partial K)] \tag{8.39}$$
$$- [\beta_1(K) - \beta_1(\partial K)] + [\beta_2(K) - \beta_2(\partial K)] - [|E_\partial| - |N_\partial|]\,. \tag{8.40}$$

Let us use the relationship between Betti numbers to simplify the last formula. Let us call $\beta_0(K) = q$ the number of connected components of K. Let us call $\beta_1(K) = g$, where g is the *genus* of K. Then, it is well known that $\beta_1(\partial K) = 2g$. Let us call $\beta_2(K) = p$, where p is the number of *cavities* (or *voids*) of K. Then, the number of connected components $\beta_0(\partial K)$ of ∂K is $\beta_0(\partial K) = p + q$. Concerning the number $\beta_2(\partial K)$ of cavities of ∂K, they are $p + q$. Finally, $\beta_3(K) = 0$. By substituting these results we have

$$\beta_0(K) + [\beta_0(K) - \beta_0(\partial K)] - [\beta_1(K) - \beta_1(\partial K)] \tag{8.41}$$
$$+ [\beta_2(K) - \beta_2(\partial K)] = q + [q - p - q] - [g - 2g] + [p - q - p] = g - p. \tag{8.42}$$

By substituting this result inside (8.39) we get

$$m^{\mathrm{MCA}} = |E| - |N| + g - p - [|E_\partial| - |N_\partial|]\,. \tag{8.43}$$

Let us now compute the number m^{VI} loop currents produced by the VINCO framework and show that they are the same as m^{MCA}. The primal tree produced by taking into account the gauging constraints contains the following number of edges

$$(|N| - |N_\partial|) - q + |N_\partial| - (p + q) + p + q = |N| - q, \tag{8.44}$$

where $(|N| - |N_\partial|) - q$ is the number of edges of an internal tree (i.e. a tree made by using the mesh edges and nodes in $K \setminus \partial K$), $|N_\partial| - (p + q)$ the boundary tree (i.e. a tree made by using the mesh edges and nodes in ∂K). To get a spanning tree of K, $p + q$ edges have to be added. We remark that this is the same number of edges of any unconstrained tree of K. The constraint is needed just to being able to enforce boundary conditions. Therefore, the unknowns of the MVI formulation are internal cotree edges (on the boundary they are set to zero because of boundary conditions) plus g cohomology generators

$$m^{\mathrm{VI}} = |E| - |E_\partial| - [(|N| - |N_\partial|) - q + p + q] + g = m^{\mathrm{MCA}}, \tag{8.45}$$

since the obtained expression is exactly equation (8.43).

To conclude, we can state that the MVI is able to obtain a quasi-minimal cycle basis. There exist techniques also to minimize the global loops which result from cohomology when the domain is not simply connected, see for example [105], but, in the authors' opinion, usually there is no sensible gain in doing it in our setting.

We also remark that in general it is also interesting to use an *ungauged* formulation[2], lazy cohomology generators or both. We will explore also these solutions in the numerical results section.

CARIDDI is a particular case of VINCO

The equivalence between CARIDDI and VINCO in case of a simply-connected conductor could be soon established by using the results of [42]. By applying them, we get

$$\mathbb{K} = \mathbb{C}^T (\mathbb{R} + i\omega\mathbb{M})\, \mathbb{C} = \mathbb{C}^T \mathbb{R} \mathbb{C} + i\omega \mathbb{C}^T \mathbb{M} \mathbb{C} = \mathbb{R}^{\mathrm{CAR}} + i\omega \mathbb{L}^{\mathrm{CAR}}. \tag{8.46}$$

Thus, all the results on cohomology computation obtained in MVI can be directly applied to CARIDDI.

[2] The ungauged formulation is the formulation as exposed in Sec. 8.2.2 in which the tree-cotree gauge is not applied thus producing an underdetermined system of equations.

9

From VU basis functions to inductance matrix factorization

EFIE formulations for solving eddy current problems, as discussed in this thesis, are mainly afflicted by two shortcomings. First, the construction of the entries of the magnetic mass matrix $\mathbb{M}$ matrix is slow due to the presence of the double integral that, in addition, for its self terms becomes singular; as a second aspect, the inductance matrix is fully populated and thus extremely expensive to be assembled and stored. Indeed, in this chapter, we illustrate how these two issues can be faced for arbitrary polyhedral meshes (wherein hexahedra, tetrahedra, prisms and pyramids are naturally included) thanks to the novel MVI method proposed in Chapter 8.

9.1 Speeding up assembly with VU basis function

Let us consider a pair of elements c, c' of a grid K. An entry of the local magnetic mass matrix $\mathbb{M}_{cc'}$ is generally defined, as in (7.20),

$$(\mathbb{M}_{cc'})_{f,f'} = \frac{\mu_0}{4\pi} \int_c \int_{c'} \frac{\boldsymbol{w}_{f|c}(\boldsymbol{x}) \cdot \boldsymbol{w}_{f|c'}(\boldsymbol{x}')}{\|\boldsymbol{x} - \boldsymbol{x}'\|} \, dc \, dc'. \tag{9.1}$$

If VU basis functions (3.21) are employed in (9.1), then the computational cost for $\mathbb{M}_{cc'}$ assembly is reduced, even when K is a tetrahedral grid. Indeed, if standard RWG or RT basis functions are employed, one has to directly use (9.1). Instead, as reported in Section 8.2.1, when VU basis functions are considered, (9.1) becomes

$$(\mathbb{M}_{cc'})_{f,f'} = t^{cc'} \boldsymbol{w}_{f|c} \cdot \boldsymbol{w}_{f'|c'}, \tag{9.2}$$

where $t^{cc'}$ is defined in (8.5).

Let us consider distinct elements c, c'. Then, under the hypothesis that the floating point operations cost is ruled by the computation of $1/\|\boldsymbol{x} - \boldsymbol{x}'\|$[1] it turns out that the obtained speed up when using (9.2) instead of (9.1) is around $|F(c)||F(c')|$. In fact, with our approach one has to compute just one double numerical integral (8.5) whereas in the standard case (9.1) one has to compute $|F(c)||F(c')|$ numerical integrals.

9.1.1 Singularity extraction with VU basis functions

The evaluation of the double integral (9.1) is singular whenever $c = c'$, i.e. for all the diagonal terms of $\mathbb{M}$ [116].

From a survey of literature on this topic, it is soon clear that the presence of affine basis functions in (9.1) leads to complex recipes to get rid of the singularity because each proposed technique is necessarily influenced by the integration domain shape (triangular, tetrahedral, prismatic, polyhedral) and by the basis functions definition; see, for instance, [116] and [128] for 3D integration domains or [129] and [130] for 2D. Hence, the possibility of dealing with volume uniform basis functions has a great potential in this respect since it allows to drastically simplify and speed up the computation. Yet, in case of a simplicial grid, the germ of this idea is presented in [116], but no volume uniform basis functions are proposed. On the contrary, the idea seems to be entirely new when working with hexahedra, pyramids and polyhedral volumes in general.

We use the well established approach called *singularity extraction* (sometimes also referred to as singularity *subtraction*) that consists of extracting a singular term from the double integral and integrate it in closed form to then treat the obtained expression numerically by mean of quadrature rules [128, 130]. The singularity extraction has been already applied in literature, but not with volume uniform basis functions. In fact, when the new volume uniform basis functions are used, the singularity extraction can be applied to equation (9.2), in which the singular double integral $t^{cc'}$ can be computed separately from the calculation of the involved basis functions values. This yields a speed up of at least $|F(c)||F(c')|$ because only one numerical integral has to be computed (namely $t^{cc'}$) in place of $|F(c)||F(c')|$ numerical integrals in (9.1). The obtained gain is even more, because the integral (8.5) is simpler. In fact, since the $\boldsymbol{w}_{f_{|c}}, \boldsymbol{w}_{f_{|c'}}$ are no more part of the singular integral, several closed-form exact expressions as the ones in [131], can now be applied to analytically calculate the innermost integral thus eliminating the singularity.

Given that it has just been shown that our (9.2) has to be preferred with respect to (9.1), now we investigate the difference between the results obtained with singularity extraction and the common solution that uses a double numerical integration. To this aim, a standard tetrahedron c whose vertices are the nodes $(0,0,0)^T$, $(1,0,0)^T$, $(0,1,0)^T$, $(0,0,1)^T$ is considered on which the double integral

$$DD = \int_c \int_{c'} \frac{1}{\|\boldsymbol{x} - \boldsymbol{x}'\|} \, dc \, dc',$$

is computed in three different ways: with a double numeric integration, with the singularity extraction approach and, finally, with an analytic formula developed for this case

[1] CPU cycles to perform division and square root dominate the dot products between the basis functions

only, by successively applying the Gauss Divergence Theorem [132, 133] and by means of a symbolic calculus software. This last one will be considered as the reference value for the double integral.

In figure 9.1, we report the results of this computation with several integration orders: the top plot shows the actual values obtained, the bottom plot shows the percentage error computed as

$$100\frac{|DD - DD_{\text{ref}}|}{|DD_{\text{ref}}|}.$$

Indeed, as reported in literature, the singularity extraction approach yields accurate values also for low numerical integration orders, whereas the purely numeric integration results to be highly inaccurate.

To complete the picture, in figure 9.2 we propose the same computation wherein one of the two domains of the double integral changes its position with respect to a first fixed tetrahedron. In this case we just compare the values obtained by applying the singularity extraction with the ones obtained with a double numerical integration: the error is computed against the value obtained with the singularity extraction combined with the highest order of numerical integration (namely, $Sing.Extr.$, $\mathcal{O}(x^{16})$) given that the previous plot showed the accuracy of this approach. The plot shows that when the two domains superpose, the double numerical integration yields inaccurate and oscillating results, whereas without a superposition of the two tetrahedra a good accuracy is obtained also with double numeric integration since for disjoint integration domains there is no singularity.

9.1.2 Factorization of the inductance matrix: MAGICA

We now recall the fact that, being $\mathbb{M}$ or $\mathbb{K}$ dense matrices, commonly advanced compression techniques must be taken into consideration whenever eddy currents have to be computed on large conducting domains that otherwise would be not affordable by means of EFIE. This happens both because of the too large memory requirements for the matrix storage and because of the prohibitive computation time for the matrix assembly, see for example [32, 105].

Usually one computes and stores the dense matrix $\mathbb{M}$ whose dimension is $|F| \times |F|$ to then assemble the complex system of equation (8.14) by means of the matrix-matrix product

$$\mathbb{K}_M := \mathbb{C}^T \mathbb{M} \mathbb{C} \tag{9.3}$$

finally obtaining a dense $|E| \times |E|$; also, as an alternative, the matrix-matrix product can be performed locally on each mesh element and then matrix $\mathbb{K}_M$ without computing the full $\mathbb{M}$. A third alternative for tetrahedral meshes is to assemble $\mathbb{K}_M$ directly by using the curl of the Nédélec basis functions. We remark that, in all the three case, the number of elements in the system matrix scales quadratically with the number of the grid edges $|E|$.

The seed of the idea here exposed stems from the fact that it might be much more convenient to compute and store a dense matrix $\mathbb{N}$ of dimension $|C| \times |C|$ containing all the results of the computation of the double integral $t^{cc'}$ for each pair of elements c, c' of the grid and then reconstruct $\mathbb{K}_M$ via a factorized expression involving sparse matrices whose memory occupation scales linearly with the number of unknowns of the

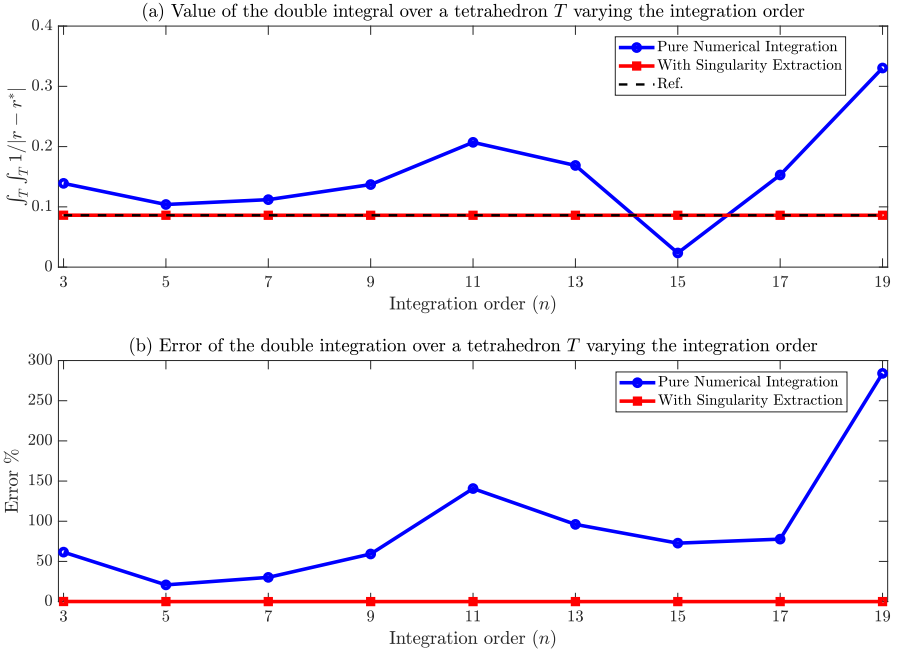

Figure 9.1: Double integral calculation over a tetrahedron by varying integration order n. For the double numerical integration, at each point of the graph, a pair of n integration orders is intended to be applied.

problem. Since in a mesh K the number of volumes $|C|$ is typically lower than the number of edges $|E|$ (or even faces), a memory saving is obtained. In addition, it is here remarked that this reformulation is obtained by algebraic manipulation only, without any approximation or loss of accuracy.

Once the interaction matrix $\mathbb{N}$ is computed and defined as

$$\mathbb{N}_{c,c'} \coloneqq \frac{t^{cc'}}{|c||c'|}, \tag{9.4}$$

it is then worth solving the system iteratively, in a matrix-free fashion, in order to just calculate matrix-vector products and never assemble the whole $\mathbb{K}_M$ as it should be done in case of system solutions by means of direct linear solvers.

If, for the sake of simplicity, we for now consider the case of a tetrahedral grids only, from (9.3), (9.2) and (8.7), an equivalent expression for $\mathbb{K}_M$ based on the here introduced *Matrix fActorization for Geometrical Integral mAtrices* (MAGICA) can be obtained as

$$\mathbb{K}_M = \mathbb{C}^T \mathbb{O}_{F_B} \left(\tilde{\mathbb{E}}_x \mathbb{N} \tilde{\mathbb{E}}_x^T + \tilde{\mathbb{E}}_y \mathbb{N} \tilde{\mathbb{E}}_y^T + \tilde{\mathbb{E}}_z \mathbb{N} \tilde{\mathbb{E}}_z^T \right) \mathbb{O}_{F_B}^T \mathbb{C}. \tag{9.5}$$

Let us now describe the terms in the last expression. We introduce the set F_B of

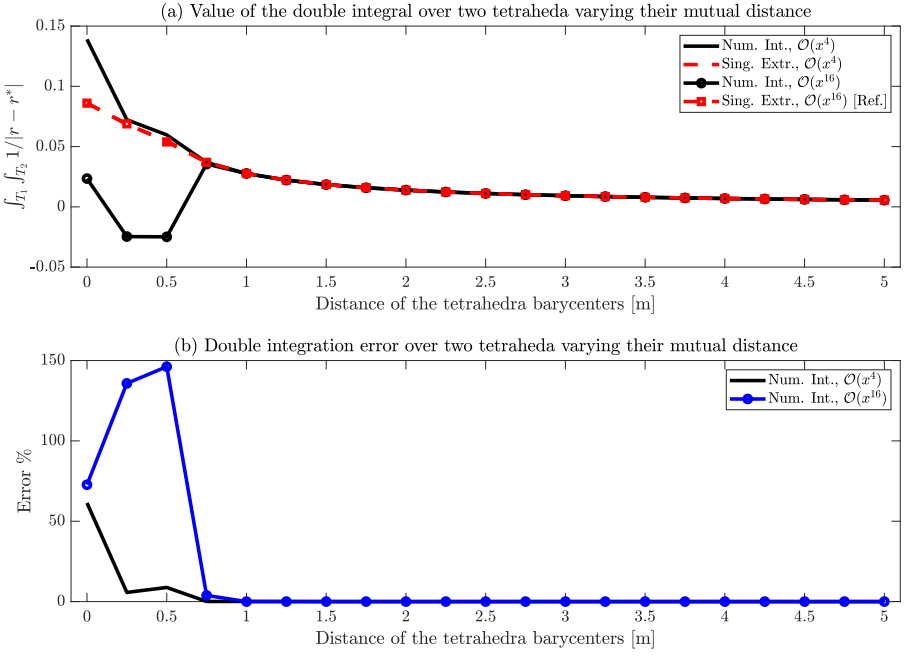

Figure 9.2: Double integral calculation over a pair of tetrahedra whose mutual distance is successively increased. As a reference value to compute the error for this test, the "*Sing. Extr.*, $\mathcal{O}(x^{16})$" case is used since in the previous plot this approach was shown to be accurate.

blossomed faces that is obtained by renumbering independently faces of each element of K, namely $F_B \coloneqq \{(f,c) \in F \times C \mid f \subset \partial c\}$. This means that all the faces shared between two elements, i.e. not belonging to the boundary of K, are repeated [2] Then, we define the sparse matrix $\mathbb{O}_{F_B}$ of dimensions $|F_B| \times |F|$ mapping the faces F of the grid K into blossomed faces in F_B. It has only one non-zero entry per row that is equal to one. Finally, $\tilde{\mathbb{E}}_x$, $\tilde{\mathbb{E}}_y$ and $\tilde{\mathbb{E}}_z$ are *sparse* matrices of dimensions $|F_B| \times |C|$. Each row of $\tilde{\mathbb{E}}_x$, $\tilde{\mathbb{E}}_y$ and $\tilde{\mathbb{E}}_z$, corresponding to an element c, is equal to $(\tilde{\mathbb{E}}_{|c}\boldsymbol{x})^T$, $(\tilde{\mathbb{E}}_{|c}\boldsymbol{y})^T$ and $(\tilde{\mathbb{E}}_{|c}\boldsymbol{z})^T$ respectively, where $\hat{\boldsymbol{x}} = (1,0,0)^T$, $\hat{\boldsymbol{y}} = (0,1,0)^T$, $\hat{\boldsymbol{z}} = (0,0,1)^T$.

[2] As a consequence, in a simplicial mesh $|F_B| = 4|C|$.

As an instance, matrix $\tilde{\mathbb{E}}_x$ has the following block structure

$$\tilde{\mathbb{E}}_x = \begin{pmatrix} \ddots & 0 & \dots & \dots & 0 \\ 0 & (\tilde{\mathbb{E}}_{|c}\boldsymbol{x})^T & & & \vdots \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & & & (\tilde{\mathbb{E}}_{|c'}\boldsymbol{x})^T & 0 \\ 0 & 0 & & 0 & \ddots \end{pmatrix}.$$

When the mesh is constituted by tetrahedra only, the stabilization matrix is not necessary. Thus, the expression in (9.5) is complete and the only hurdle left is the construction of $\tilde{\mathbb{E}}_x$, $\tilde{\mathbb{E}}_y$ and $\tilde{\mathbb{E}}_z$ matrices.

Differently, for general polyhedra, the factorization in (9.5) is not complete and also the stabilization part has to be taken into account in order to be compliant with the definition of $\mathbb{M}^s$ of (8.7). In this case is necessary to add new terms to (9.5) expressing the stabilization matrix $\mathbb{S}_c$ of (8.7). The most efficient way to that goal is the direct assembly of the global stabilization matrix $\mathbb{S}$ as

$$\mathbb{S} = \sum_{c \in C} (\mathbb{O}_c^{\mathcal{F}})^T \mathbb{S}_c \mathbb{O}_c^{\mathcal{F}}. \tag{9.6}$$

Indeed, thanks to the definition in (8.7) in which the local stabilization matrix $\mathbb{S}_c$ has to be summed to $\mathbb{M}_{cc'}$ only in the case $c = c'$, equation (9.6) produces a sparse matrix with exactly the same sparsity of $\mathbb{R}^s$. Hence, it affects neither the assembly time nor the memory consumption for its storage since it scales linearly with the mesh dimension. In addition, it can be efficiently assembled simultaneously to $\mathbb{R}^s$, since the $\mathbb{S}_c$ matrix has to be added both to $\mathbb{R}_c$ and to $\mathbb{M}_{cc'}$, hence it can be constructed only once per considered volume c.

As a matter of fact, in case of polyhedra, equation (9.5) becomes

$$\mathbb{K}_M = \mathbb{C}^T \mathbb{O}_{F_B}^T \left(\tilde{\mathbb{E}}_x^T \mathbb{N} \tilde{\mathbb{E}}_x + \tilde{\mathbb{E}}_y^T \mathbb{N} \tilde{\mathbb{E}}_y + \tilde{\mathbb{E}}_z^T \mathbb{N} \tilde{\mathbb{E}}_z \right) \mathbb{O}_{F_B} \mathbb{C} + \mathbb{C}^T \mathbb{S} \mathbb{C}. \tag{9.7}$$

It is here also noticed that *MAGICA* expression in (9.7) does not reflect the actual implementation. In fact, from a practical point of view, instead of storing $\mathbb{C}$ and $\mathbb{O}_{F_B}$ to then perform the matrix-matrix product with $\tilde{\mathbb{E}}_x$, $\tilde{\mathbb{E}}_y$ and $\tilde{\mathbb{E}}_z$ it is wiser to directly assemble each of the three products $\tilde{\mathbb{E}}_x \mathbb{O}_{F_B} \mathbb{C}$, $\tilde{\mathbb{E}}_y \mathbb{O}_{F_B} \mathbb{C}$ and $\tilde{\mathbb{E}}_z \mathbb{O}_{F_B} \mathbb{C}$ as sparse $|N| \times |E|$ matrices. In addition, we also recall for the sake of precision, that equation (9.7) takes into account neither the boundary conditions nor the gauging: to this purpose, the same techniques described in Section 7.2 can be directly applied without any loss of generality.

9.2 A new family of compression techniques

In the previous section the MAGICA factorization of the dense inductance matrix $\mathbb{K}_M$ of (9.3) was introduced. Indeed, it was shown that thanks to (9.5) for simplicial meshes and to (9.7) for polyhedral ones, a lossless compression of the memory occupation can

be usually achieved. What is missing now is the system solution that can be achieved by means of two different iterative approaches: a lossless technique that directly applies (9.7) to (9.3) and an approximated scheme that can exploit either the Fast Multipole Method [30], ACA or any other fast summation algorithm in order to efficiently compute the off-diagonal terms of $\mathbb{N}$, with a consequent drastic reduction of the memory footprint and of the computation time. Let us carefully delineate them both.

9.2.1 LIME: a Lossless Integral Matrix comprEssion

The first possibility is represented by the application of the factorized expression of (9.7) into (9.3). In order to avoid the construction of a full, for instance, $\tilde{\mathbb{E}}_x^T \mathbb{N} \tilde{\mathbb{E}}_x$ matrix, only on-the-fly matrix-vector products are allowed. Hence, matrix-free algorithms for the iterative solution of symmetric positive definite (SPD) systems, like GMRES [134], can be used. As far as the preconditioner is concerned, the diagonal part of $\mathbb{K} = \mathbb{K}_R + i\omega\mathbb{K}_M$ (i.e diag($\mathbb{K}$)) has been used.

In addition to this, also resorting to the solution of an *ungauged* system results to be very effective when working with Krylov's subspace-based iterative solvers.

MVI solution with LIME The process described above yields the flow chart of Fig. 9.3 in which it is illustrated how to face the solution of an eddy current problem as expressed by EFIE in which MAGICA is applied for the compression and factorization of $\mathbb{K}_M$.

We here remark that in this section and thus in the proposed flow chart too, we did not specify whether the domain is simply connected or not; in fact, this distinction is not necessary since equation (9.7) can be plugged into (8.14) without loss of generality.

9.2.2 Approximated compressions of integral matrices

In order to further increase the size of the problems affordable with the EFIE formulation here exposed, it may be very useful to apply more advanced compression techniques than the lossless factorization based on the full computation of $\mathbb{N}$ matrix exposed in the last section.

Literature mainly illustrates two possible approaches to be employed in case of EFIE: from the one hand there are *algebraic* methods based on hierarchical matrices algebra like the Adaptive Cross Approximation that are the most diffused, to the other hand also *analytical* approaches that relies on an analytic expansion of the $1/\boldsymbol{x}$-kernel-based expressions, as it is done by the Fast Multipole Method, can be extremely interesting since these methods are able to rapidly and efficiently calculate the interactions that have to be computed when assembling the dense magnetic matrix $\mathbb{M}$. Yet, even if both of them may seem to be equally appealing and theoretically effective in reducing both the peak memory usage and the computational time during the system assembly, in truth, when applied to the formulation without any factorization as exposed until section 8.4, both of them exhibit important limitations and downsides.

First, ACA-based approaches are not very efficient when applied to the system matrix $\mathbb{K}$. Indeed, the tricky aspect resides in the fact that rows and columns of $\mathbb{K}$ are referred to the grid edges whereas usually the assembly of $\mathbb{M}$ has to be constructed sequentially

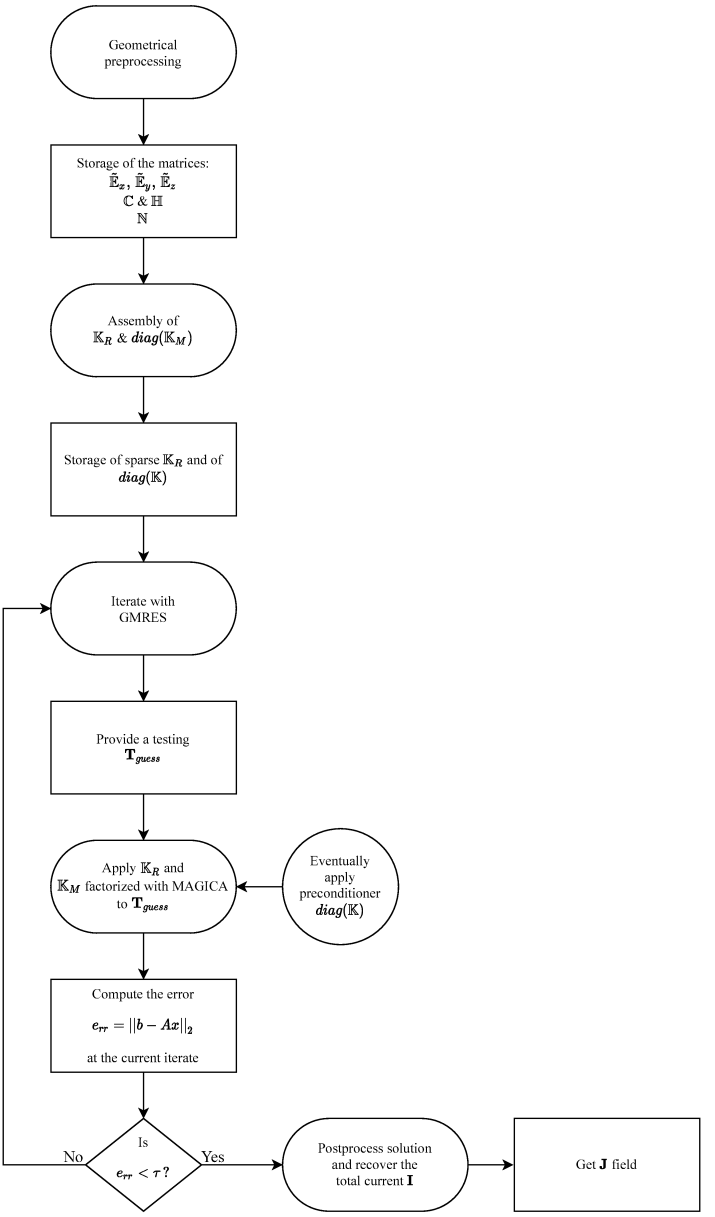

Figure 9.3: Flow chart of the iterative system solution with LIME.

by looping within the grid elements. When an ACA library is interfaced to $\mathbb{K}$, this traduces either into huge assembly inefficiencies. In both cases, the result is that, even if a memory usage reduction could be achieved, the same does not happen for the assembly time which is reduced to half at most, see for example [105] or [32].

Second, as extensively explained in [105], standard ACA-based approaches applied on K show poor performances whenever the problem needs many cohomology generators (i.e., when the conducting domain presents several holes). This is caused by the non-local basis functions expressed in $\mathbb{H}$ matrix that unavoidably have a very large support. As a consequence, most entries of K have to be computed anyway. In [105] again, the problem is minimized by the retrieval of cohomology generators whose length is quasi-minimal, but also using a minimal cohomology basis the time required is huge.

Last but not least, we cannot forget that $\mathbb{K}$ matrix is a complex operator. Since off-the-shelf ACA libraries offer facilities for the linear system solution, most contributions in literature apply ACA to the complex matrix $\mathbb{K}$. Thus, the memory allocation is almost *doubled* with respect to what would have been possible, considering that the real part of $\mathbb{K}$, i.e. $\mathbb{K}_R$, is instead very sparse.

Eventually, we must remark that FMM approaches are not free from criticism too. In this case the main limit is represented by the fact that the FMM works smoothly only in case of point charges. It follows that the standard definition of $\mathbb{M}$ in which the $\boldsymbol{w}_f$ face basis functions are involved is not well suited by itself for such an approach. In fact, the presence of linear $\boldsymbol{w}_f$ in the double integral to be approximated with FMM yields, in the end, the necessity of resorting to complicated strategies in order to make FMM fit into the EFIE framework as it is done for instance in [31].

In conclusion, the proposal of a new form of (8.14) able to totally eliminate all the previously exposed drawbacks both for ACA-based and FMM-based approaches, represents without any doubt a ground-breaking step forward for EFIE formulation for eddy currents. Indeed, we claim that the factorization expressed by (9.5) and (9.7) perfectly reaches this goal. In the continuation we describe in detail how and why.

FAIME: a Fast Approximated Integral Matrix comprEssion

In the authors' opinion, the most interesting and effective technique able both to reduce the memory requirements and to shrink the computation time too, which is the main bottleneck of EFIE formulations, is represented by the *Fast Approximated Integral Matrix comprEssion* approach (FAIME) in which FMM is applied to the computation of $\mathbb{N}$. More precisely, this new approach is based on the splitting of the dense matrix $\mathbb{N}$ into

$$\mathbb{N} = \mathbb{N}^{\mathrm{NEAR}} + \mathbb{N}^{\mathrm{FAR}} \tag{9.8}$$

in which $\mathbb{N}^{\mathrm{NEAR}}$ is a very sparse matrix representing the *near-field interactions* and $\mathbb{N}^{\mathrm{FAR}} = \mathbb{N} - \mathbb{N}^{\mathrm{NEAR}}$ is the dense remaining part representing the *far-field interactions* on which FMM is applied. We remark that this is different with respect to what done previously in the literature, where the matrix $\mathbb{K}$ or $\mathbb{K}_M$ is split; instead, we split matrix $\mathbb{N}$. A possible choice is to consider in $\mathbb{N}_D$ the diagonal term plus some of the off-diagonal terms of $\mathbb{N}$, specifically the ones relative to the neighbouring volumes.

In addition, when two elements of the grid c and c' are not close from each other,

the following approximation holds

$$t^{cc'} = \frac{\mu_0}{4\pi}\int_c\int_{c'}\frac{1}{\|\boldsymbol{x}-\boldsymbol{x}'\|}dc\,dc' \approx \frac{\mu_0}{4\pi}\frac{|c||c'|}{\|\boldsymbol{b}_c-\boldsymbol{b}_{c'}\|}, \tag{9.9}$$

in which $\boldsymbol{b}_c$ and $\boldsymbol{b}_{c'}$ are the barycenters of c and c', respectively.

Once this approximation is introduced, then *any* off-the-shelf FMM library can be applied in order to rapidly and efficiently compute the effects of the off-diagonal terms relative to $\mathbb{N}_{OD}$. Moreover, any alternative fast-summation technique developed for N-body simulation, molecular dynamics or electrosatics, like Mesh particle method (M3P), fast Ewald summation, etc, can be easily employed thanks to the MAGICA factorization.

From the implementation point of view, FMM results to be the most effective tool in drastically reducing the part of $\mathbb{N}$ to be stored that can be thus limited to the sparse $\mathbb{N}_D$. This traduces into the possibility of rearrange $\mathbb{K}$ too by following the same splitting as of (9.8), thus obtaining

$$\mathbb{K} = \mathbb{K}^{\text{NEAR}} + i\omega\mathbb{K}_M^{\text{FAR}}, \tag{9.10}$$

in which

$$\mathbb{K}^{\text{NEAR}} = \mathbb{K}_R + i\omega\mathbb{C}^T\mathbb{O}_{F_B}^T\left(\tilde{\mathbb{E}}_x^T\mathbb{N}^{\text{NEAR}}\tilde{\mathbb{E}}_x + \tilde{\mathbb{E}}_y^T\mathbb{N}^{\text{NEAR}}\tilde{\mathbb{E}}_y + \tilde{\mathbb{E}}_z^T\mathbb{N}^{\text{NEAR}}\tilde{\mathbb{E}}_z\right)\mathbb{O}_{F_B}\mathbb{C} + \mathbb{C}^T\mathbb{S}\mathbb{C} \tag{9.11}$$

and

$$\mathbb{K}^{\text{FAR}} = \mathbb{C}^T\mathbb{O}_{F_B}^T\left(\tilde{\mathbb{E}}_x^T\mathbb{N}_{OD}\tilde{\mathbb{E}}_x + \tilde{\mathbb{E}}_y^T\mathbb{N}_{OD}\tilde{\mathbb{E}}_y + \tilde{\mathbb{E}}_z^T\mathbb{N}_{OD}\tilde{\mathbb{E}}_z\right)\mathbb{O}_{F_B}\mathbb{C}. \tag{9.12}$$

It turns out that, by resorting to the proposed FAIME approach, we can represent the otherwise dense EFIE matrix $\mathbb{K}$ just by only storing a set of sparse *real* matrices, i.e. $\mathbb{K}_R = \mathbb{C}^T\mathbb{R}_s\mathbb{C}$, $\mathbb{K}_S = \mathbb{C}^T\mathbb{S}\mathbb{C}$, $\mathbb{O}_{F_B}$, $\tilde{\mathbb{E}}_x$, $\tilde{\mathbb{E}}_y$, $\tilde{\mathbb{E}}_z$ and $\mathbb{N}^{\text{NEAR}}$, because $\mathbb{K}^{\text{FAR}}$ and thus $\mathbb{N}^{\text{FAR}}$ too, are never actually assembled but it is computed their application to a given DoFs array provided by GMRES (namely, $\boldsymbol{T}_{\text{guess}}$ of Fig. 9.3) on the fly at each iterate by means of the FMM. In addition, it is worth recalling that the computation of $\mathbb{N}^{\text{NEAR}}$ is very efficient too, because it is highly parallelizable without any peculiar effort due to the fact that $\mathbb{N}^{\text{NEAR}}$ rows and columns directly refer to the volumes mesh and not to edges or faces as it happens for $\mathbb{M}$ or $\mathbb{K}_M$, respectively.

We also mention that the use of FMM for the fast computation of $\mathbb{N}^{\text{FAR}} = \mathbb{N} - \mathbb{N}^{\text{NEAR}}$ applied to a given DoFs array results to be a successful recipe because by doing this the FMM library is applied to a minimal set of points when computing point-to-point interactions between the barycenters of the mesh elements thus saving time and reducing memory occupation for the same reasons explained in Sec. 9.1.2.

Algebraic methods

When discussing about algebraic method, also ACA is suitable to be used in order to compress $\mathbb{N}$. Again, we remark that this is different with respect to what appears in literature, where ACA compression can be applied only on matrix $\mathbb{K}$ or $\mathbb{K}_M$, and not on matrix $\mathbb{N}$ as we propose. The strategy is exactly the same as that one described for FMM with the only difference that in this case the computation of the application of $\mathbb{N}$ to $\boldsymbol{T}$ is not performed on the fly as for the FMM but it has to be treated as a preprocessing step before iteratively solving the system. When ACA library is invoked, a compressed

hierarchical expression of the whole $\mathbb{N}$ is computed and stored in the calculator memory. Yet, even if it can be shown that the overall compression ration between FMM and ACA is very similar, the same cannot be said for the peak memory usage, that is much higher when ACA is applied. Thus, using FMM extends the applicability of the solver, where ACA is inapplicable because it would use too much memory. By the way, also for this case, the application of ACA for the calculation of $\mathbb{N}$ instead of $\mathbb{K}$ automatically resolves, in a efficient and accurate way, all the troubles and limitations above exposed both in terms of implementation complications and in terms of theoretical shortcomings caused by cohomology generators and vector basis functions.

9.3 Numerical results

In this section we verify the correctness and accuracy of the implementation of the new MVI method.

Intel®MKL routines are exploited for specific tasks, like sparse matrix-vector products and sparse matrix factorizations that will be shown to be required when preconditioning in the low-frequency regime. As far as the FMM implementation is concerned, we make use of the parallel *FMM3D* library proposed by the Flatiron Institute [36].

Simulations are performed on a Windows server (hereafter, *the Windows Server*) in which an AMD Ryzen™Threadripper PRO 3975WX (32 cores/64 threads @3.49GHz) processor runs endowed with 256 GB of RAM. Yet, it is fundamental to highlight that most of the results can be also obtained on a standard Windows laptop (now on, *the Windows Laptop*) equipped with an Intel®Core™i7 processor (4 cores @2.9GHz) with 16 GB of RAM without incurring in any memory saturation.

In the following, for the first set of results, we resort to the eddy currents computation in the frequency domain in a solid conductive sphere of radius $R_0 = 50.0$ mm immersed in a uniform induction field $B_z = 1.0$ mT vertically directed along the z axis. The choice of this geometry as a test bench is driven by the fact that an analytical solution is provided in [135], thus allowing for thorough comparisons and analysis. Later on, once that the code validity is assessed, eddy currents phenomena are calculated in more general not simply connected geometries, more specifically, on the TEAM Workshop Problem 7 [136], on a conducting plate with seven holes and a coil realized on a printed circuit board.

9.3.1 FAIME approach accuracy and convergence

As a successive step, the attention is now focused on the solution accuracy that can be achieved with the FAIME implementation. In more details, the solution of a solid sphere with differently grained meshes is faced first, both with tetrahedral and hexahedral elements; then, also TEAM 7 benchmark is proposed as a more complicate geometry example. As a third paramount aspect that is analysed in the following, the behaviour of the iterative solver convergence when the frequency varies is reported too.

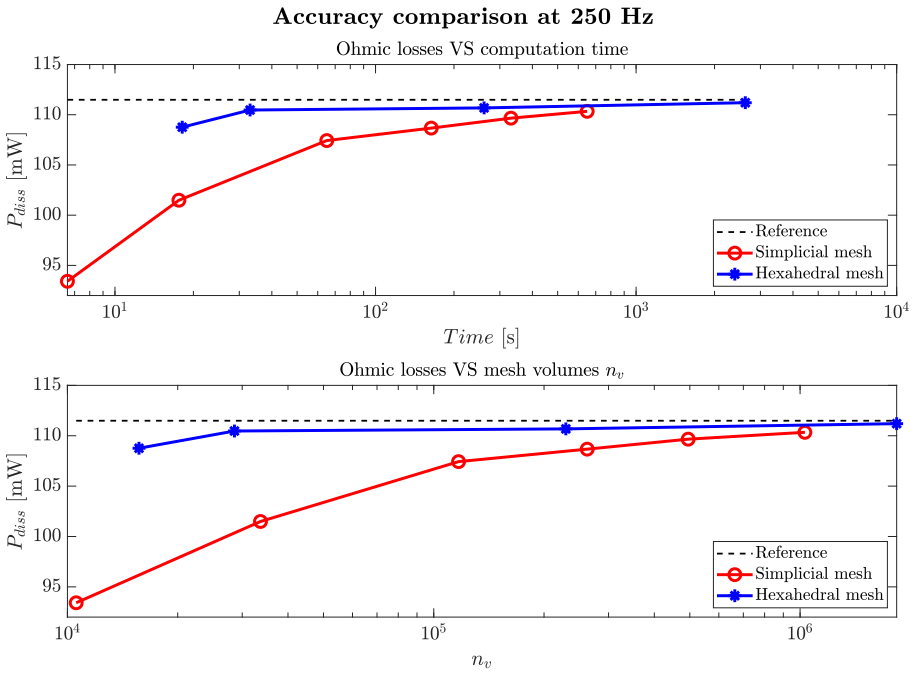

Figure 9.4: Ohmic losses trend for simplicial and hexahedral meshes successively refined. Top: P_{diss} convergence versus the computation time. Bottom: P_{diss} convergence versus mesh elements n_v. For this chart $P_{diss}^{REF} = 111.49$ mW.

Ohmic losses in a solid sphere

In Fig. 9.4 a comparison versus the number of grid elements in the solid conducting sphere of radius $R_0 = 50$ mm is proposed. For this test, the frequency is $f = 250$ Hz, the resistivity $\rho = 1.68 \cdot 10^{-8}\,\Omega \cdot$ m and the external induction field is uniform: $B_z = 1.0$ mT.

Table 9.1 contains the geometric information about the number of elements of the tetrahedral and hexahedral grids. It is worth noticing that our implementation can handle up to two million elements. Also, GMRES iterated 11 times for every simplicial mesh and 14 times for the hexahedral grids. No variation of this parameter was observed during the trials when increasing n_v. Further details about these simulations are reported in Table 9.1.

Generally, from this comparison, it turns out that hexahedral meshes perform better than simplicial ones in terms of accuracy both when time and mesh volumes are considered and, in particular, when $|C|$ is lower than $100,000$ elements. Yet, simplicial grids have the advantage of a faster computation hence, especially when the number of volumes and DoFs grows, their performance becomes comparable to that one obtained with the hexahedral grids. Moreover, from a look at Table 9.1, it should be noticed that, as already stated, hexahedral meshes exhibits a DoFs number which is at least twice $|C|$

Table 9.1: Mesh data and calculation time in the Windows Server for the solid sphere benchmark

Mesh type	n_v	N_{dofs}	Assembly [s]	Solve [s]	Tot. time [s]
Simplicial	10566	9517	4.35	2.0	6.57
	33652	31405	14.0	2.6	17.6
	116863	111677	54.5	8.9	65.0
	262349	253072	144.4	14.6	163.6
	494865	481383	249.5	63.0	331.0
	1030656	1008768	541.8	89.0	649.0
Hexahedral	15680	30773	14.0	3.2	18.12
	28572	56577	28.0	7.1	33.0
	229376	455681	211.0	48.0	261.0
	1835008	3657729	1809.6	758.0	2625.0

Table 9.2: Mesh data and calculation time in the Windows Server for the TEAM Workshop Problem 7

Frequency [Hz]	$\|C\|$	N_{dofs}	Iterates	Assembly [s]	Solve [s]	Tot. time [s]
50.0	575064	1099925	10	578.6	61.6	662.0
200.0	575064	1099925	16	579.0	88.0	690.0

whereas for simplicial ones this ratio is always equal to 1 or lower: another argument that makes simplicial grids preferable when $|C|$ exceeds one million elements.

TEAM7 benchmark: accuracy and frequency sweep

TEAM Workshop Problem 7 is faced. For the sake of concision the geometrical set up of the problem is not here reported given that it is carefully described in [136]. To solve the problem, a multi-block hexahedral mesh was created. The grid is made by $575,064$ elements for which correspond $1,099,925$ DoFs to be solved and the source field computation is tackled by means of exact closed-form formula. In Fig. 9.5 the map of the real part of the current density in the conductor is showcased for the 50 Hz case whereas in Table 9.2 all the meaningful aspects of the simulation performance are listed.

Last but not least, in Fig. 9.6 and Fig. 9.7 the traditional induction field plots along A1-B1 and A2-B2 lines for the two test frequencies are proposed. Also in this case, an excellent correspondence between the simulated data and the TEAM 7 reference field is shown.

Furthermore, the same problem is solved by discretizing the conductor with a mesh constituted by $132,000$ hexahedra and $514,410$ triangular prisms in order to asses the method effectiveness also in presence of a mixed mesh. Results reported in Fig. 9.9 are in perfect agreement with those ones previously obtained with a pure hexahedral grid.

Eventually, Fig. 9.8 is about a fundamental aspect when using iterative solvers like

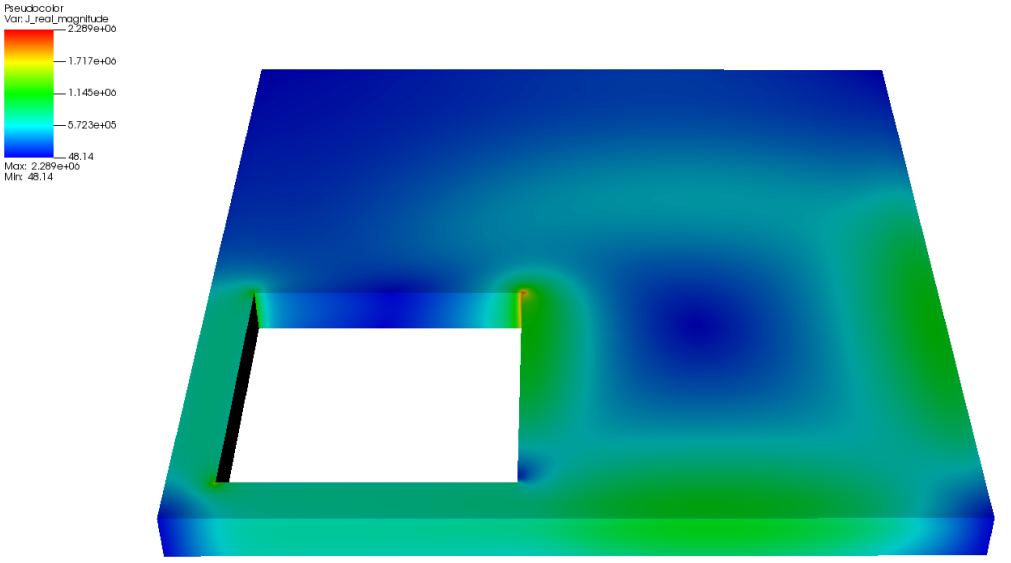

Figure 9.5: Real($\boldsymbol{J}$) colour map at $f = 50$ Hz in the TEAM 7 conducting plate.

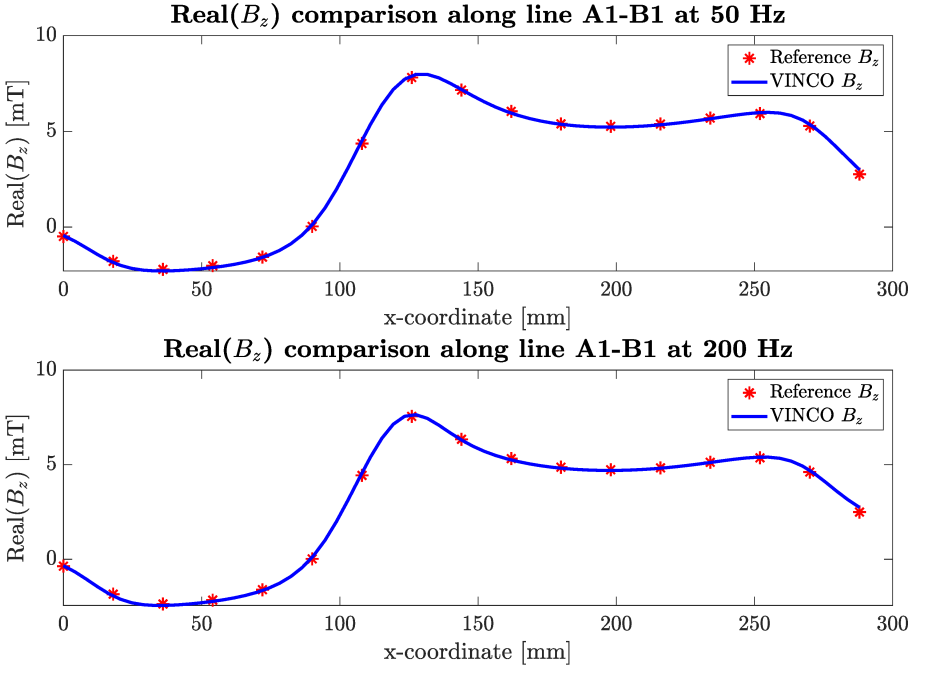

Figure 9.6: Real part of the vertical induction field component along A1-B1 sample line.

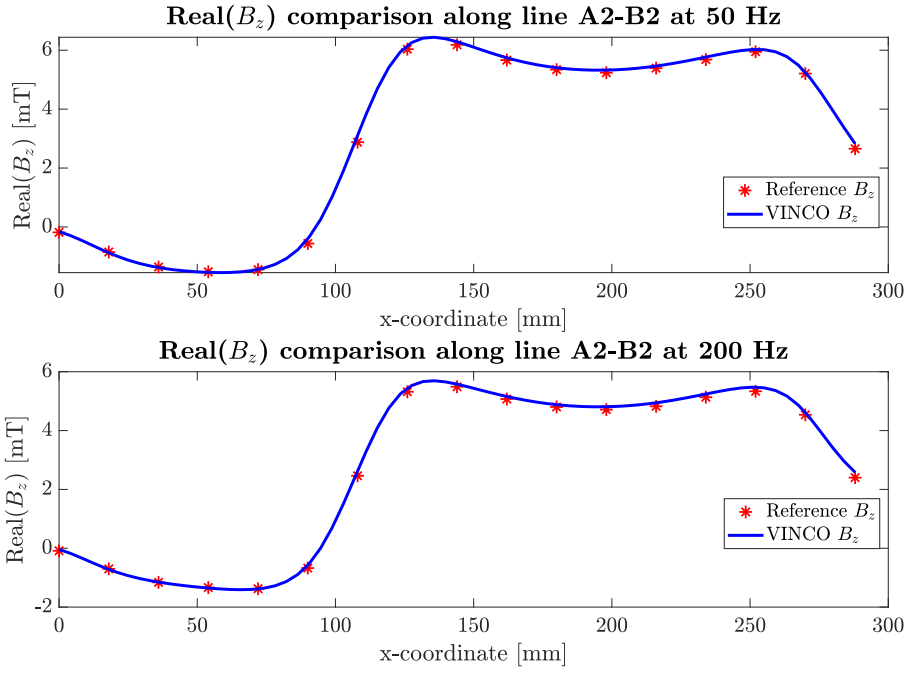

Figure 9.7: Real part of the vertical induction field component along A2-B2 sample line.

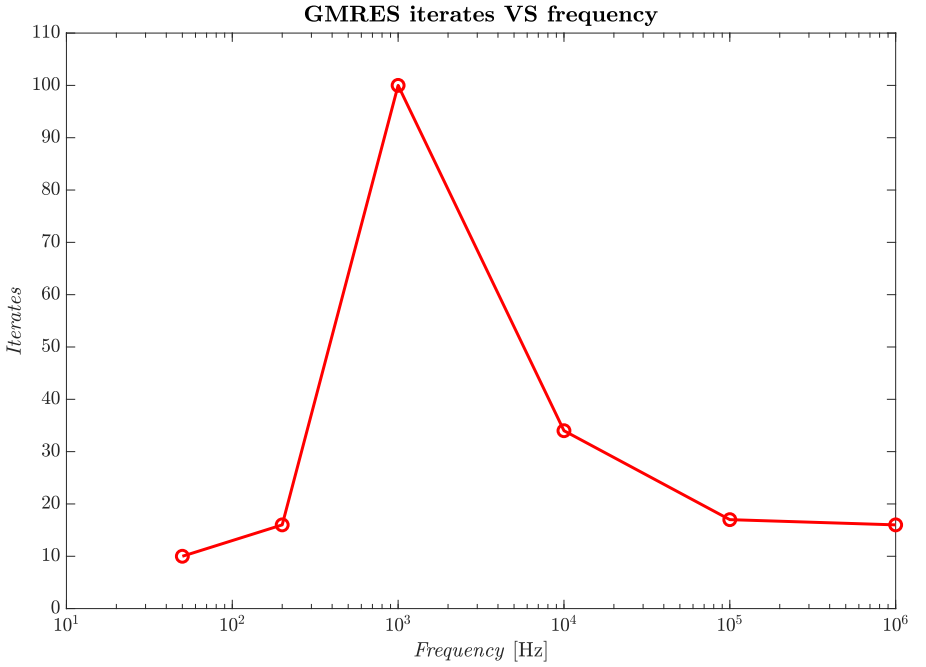

Figure 9.8: Frequency variation and GMRES iterates values on TEAM 7 problem configuration.

GMRES to obtain the solution of EFIE system: variation of the iterates number when changing the frequency. To that end, TEAM 7 problem is used as a common problem setting.

It is now worth mentioning that the chart is the result of two different strategies to precondition the system during its iterative solution: from the one hand, at low frequency i.e. $f < 500.0$ Hz, a sparse preconditioner coincident to the real $\mathbb{K}_R$ matrix is effectively applied to the *gauged* system of equations by means of its pre-factorized version obtained with PARDISO, to deeply accelerate the convergence of GMRES; to the other hand, when f exceeds that threshold another preconditioning technique is pursued that consists of solving an *ungauged* system whose rows are then rescaled thanks to a complex diagonal preconditioner $\mathbb{K}_P = \mathrm{diag}(\mathbb{K}_R + i\omega\mathbb{K}_M)$.

This choice reflects into limiting the iterates to some tens in the high and low frequency range and to obtain just a dull and affordable peak of one hundred iterates for the intermediate frequencies in the range from one to some kHz.

9.3.2 Assessing the asymptotically linear behaviour of FAIME

This section goal is testing the asymptotically linear behaviour that is achieved thanks to the new formulation based on the factorization of the inductance matrix for the frequency range typical of application of power electronics and inductive sensors (from

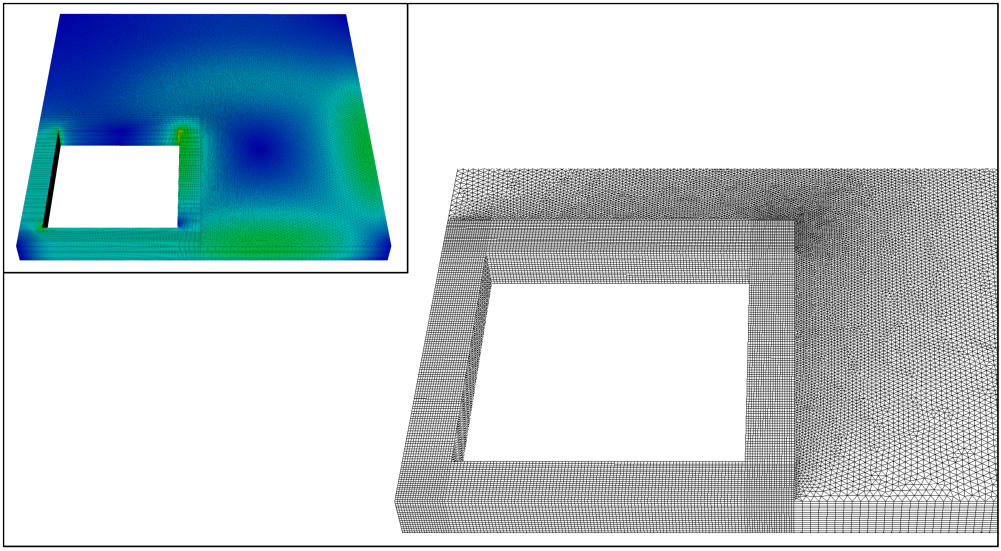

Figure 9.9: Left-bottom corner of TEAM 7 mesh made of mixed hexahedra and triangular prisms. In the inset, for this mesh, Real($\boldsymbol{J}$) colour map at $f = 50.0$ Hz is shown as a comparison to the one previously obtained in Fig. Fig. 9.5.

statics to a few MHz). In addition to this, the performance of EFIE system solution when ACA is directly applied to $\mathbb{K}$ matrix in place of its factorized version presented in this paper is compared to FAIME implementation in which FMM is used to get rid of the bottlenecks caused by being $\mathbb{K}_M$ a dense matrix. Benefits and observations on this last point are exposed.

Computational time scaling for a solid sphere

The asymptotically linear behaviour of FAIME is distinctly depicted in Fig. 9.10. The same result can be obtained if in place of the total computational time the Peak Memory Usage (PMU) is put in the ordinate axis.

This plot justifies the impact of the use of FMM in addition to the factorized expression of $\mathbb{K}_M$ and validates the correctness of the code structure thus eliminating the otherwise quadratic scaling of time and memory occupation typical of all EFIE formulations. In other words, it is possible to state that FAIME exhibits a better asymptotic computational complexity than FEM codes based on direct solvers like PARDISO (iterative solvers usually converge very slowly for eddy current problems formulated with FEM). In the following section, we also show that the same results cannot be obtained when algebraic compression techniques are directly applied to the whole K.

Overall performance comparison with state-of-the-art

Table 9.3 relates the performance of the solution of EFIE when afforded with three different approaches: the standard solution of (8.14) where K is computed and stored

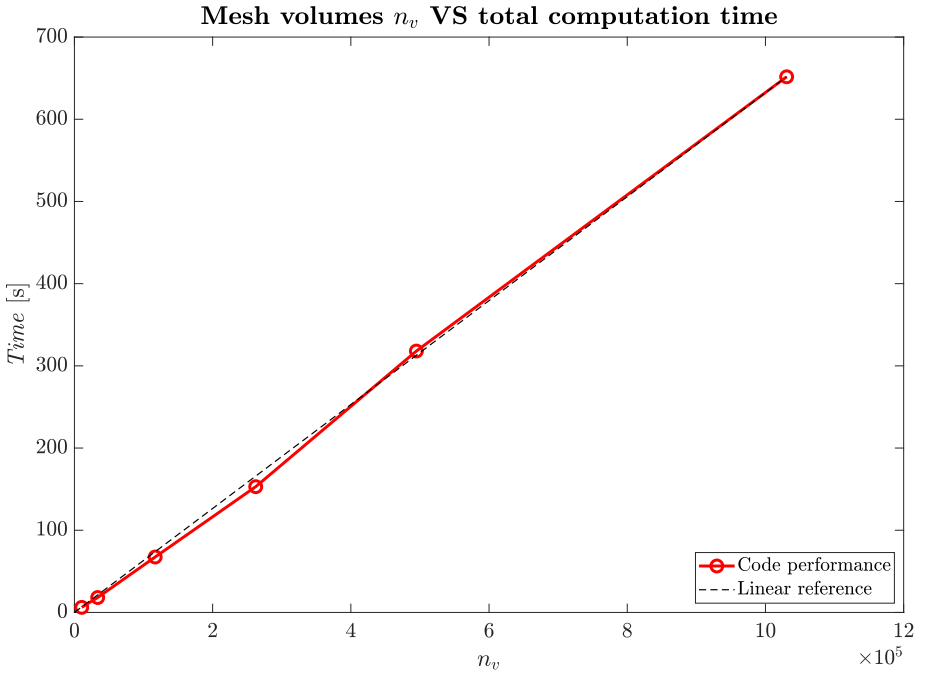

Figure 9.10: Mesh elements and total computational time comparison

Table 9.3: Performance comparison in the Windows Server between state-of-the-art approaches and this paper code. Symmetry of the mass matrices is exploited in all the approaches.

	Standard EFIE	ACA/HLIBpro	FAIME
Matrices allocated memory [GB]	8141	-	0.883
Non-zero entries stored	$1.0 \cdot 10^{12}$	-	$95 \cdot 10^6$
Tot. time	-	> 24 h	649 s
PMU [GB]	-	-	13.5

as a fully populated complex matrix (*Standard EFIE* in the table header), the algebraic compression of the whole K entries by means of HLIBpro library [137] (*HLIBpro* in the table header). These two approaches are regarded as state-of-the-art techniques. The third missing term of the comparison naturally is the novel FAIME approach presented in this paper (*FAIME* in the table).

In the proposed example, the considered configuration is the solid sphere problem at $f = 50$ Hz, $\boldsymbol{B}_z = 1.0$ mT, in which the conductor is meshed with $|C| = 1,030,656$ tetrahedra with corresponding $1,008,768$ DoFs.

From a comparison of the data two main considerations can be done: the first, the trivial one, is that the standard approach is not suitable to deal with such a problem because of memory and time requirements that are anyhow prohibitive for a practical use of the code. The second aspect to be highlighted is that, even if HLIBpro might succeed in reducing and limiting the memory occupation, the same cannot be said for the computational time that still dramatically impacts into the overall performance making also this method not considerable to solve problems with a large number of unknowns. Differently, FAIME approach is effective in both squeezing the memory occupation and in drastically reducing the computation time too thus rendering this novel approach very promising for a wide range of practical problems.

For the sake of precision, it is here remarked that, in virtue of FAIME effectiveness in eliminating EFIE bottlenecks, the same problem has been also solved with FAIME approach in the Windows Laptop (16 GB of RAM), scoring a total computation time of $1,220$ s of which 788.5 s for the matrices assembly. It can be deduced that the gap with respect to the computation time of the Windows Server is mainly due to the parallel implementation of FMM3D library and MKL PARDISO solver (4 vs 32 cores) used to factorize the preconditioner K_R applied to solve the problem.

9.3.3 A prismatic mesh of a plate with seven holes and a printed circuit board coil

To conclude the numerical results section, it is proposed in Fig. 9.11 the solution of the same problem considered by the authors in [138] whose results will be regarded as reference values for the present test. This last benchmark introduces two peculiar aspects not faced yet in the previous cases: the presence of a wider number of holes and thus of cohomology generators and the discretization of the conductor by means of triangular prismatic elements. The computed ohmic losses in the discrete grid are

Table 9.4: Mesh data and calculation time in the Windows Server for the conducting plate with seven holes

Frequency [Hz]	n_v	N_{dofs}	Iterates	Assembly [s]	Solve [s]	Tot. time [s]
50.0	290246	528093	19	265.8	38.8	312.6

$P_{diss} = 0.891$ mW that perfectly matches the values in [138]. Other aspects of the simulation performance are reported in Table 9.4.

Finally, in Fig. 9.12, a colour map of the real part of the current density field in a PCB coil fed by a sinusoidal voltage at 5 MHz is shown. The coil is composed by 8 copper turns and the outer turns have a radius of 20 mm. The copper section is 0.2 mm $\times$ 35μm. The mesh used consists of 514,285 tetrahedra that yields 416,049 unknowns. The total simulation time is 242 s.

9.4 Conclusion

Starting from the classical EFIE in the magneto-quasistatic limit, a novel compatible volume integral method to solve the eddy current problem has been introduced. The method roots on volume uniform basis functions, which provide the rational for taming the two main weaknesses of integral formulations. On one hand, the computation of the elements of the system matrix is improved with respect to speed and implementation simplicity. On the other hand, a novel factorization of the inductance matrix is introduced. This factorization induces a ground-braking speedup of various orders of magnitude with respect to the state-of-the-art solutions (in particular, a popular ACA library) thanks to the use of fast summation techniques, like the Fast Multipole Method, to perform the matrix-vector product. Moreover, in our framework, one can use any off-the-shelf libraries developed for fast charge summation.

It is clear that all the techniques introduced in this paper for tetrahedra and general polyhedra can be readily adapted to triangles and general polygons. To deal with electrodes, magnetic materials and full Maxwell problems are the topics currently under investigation. In particular, the proposed basis functions and factorization can be universally applicable to any generalized mass matrices arising with any integral method or boundary element method.

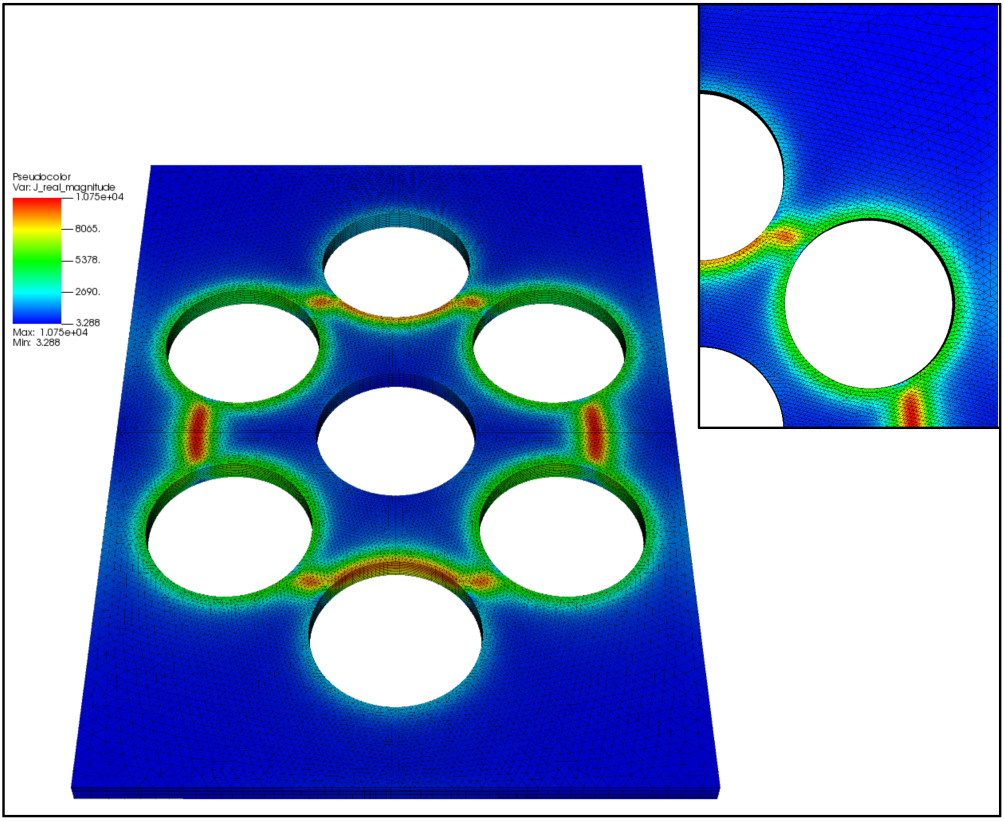

Figure 9.11: Main picture: colour map of the real part of the current density field in the conducting plate with seven holes. Inset: a detail of the triangular faces of the prisms forming the mesh are depicted in one fourth of the original mesh.

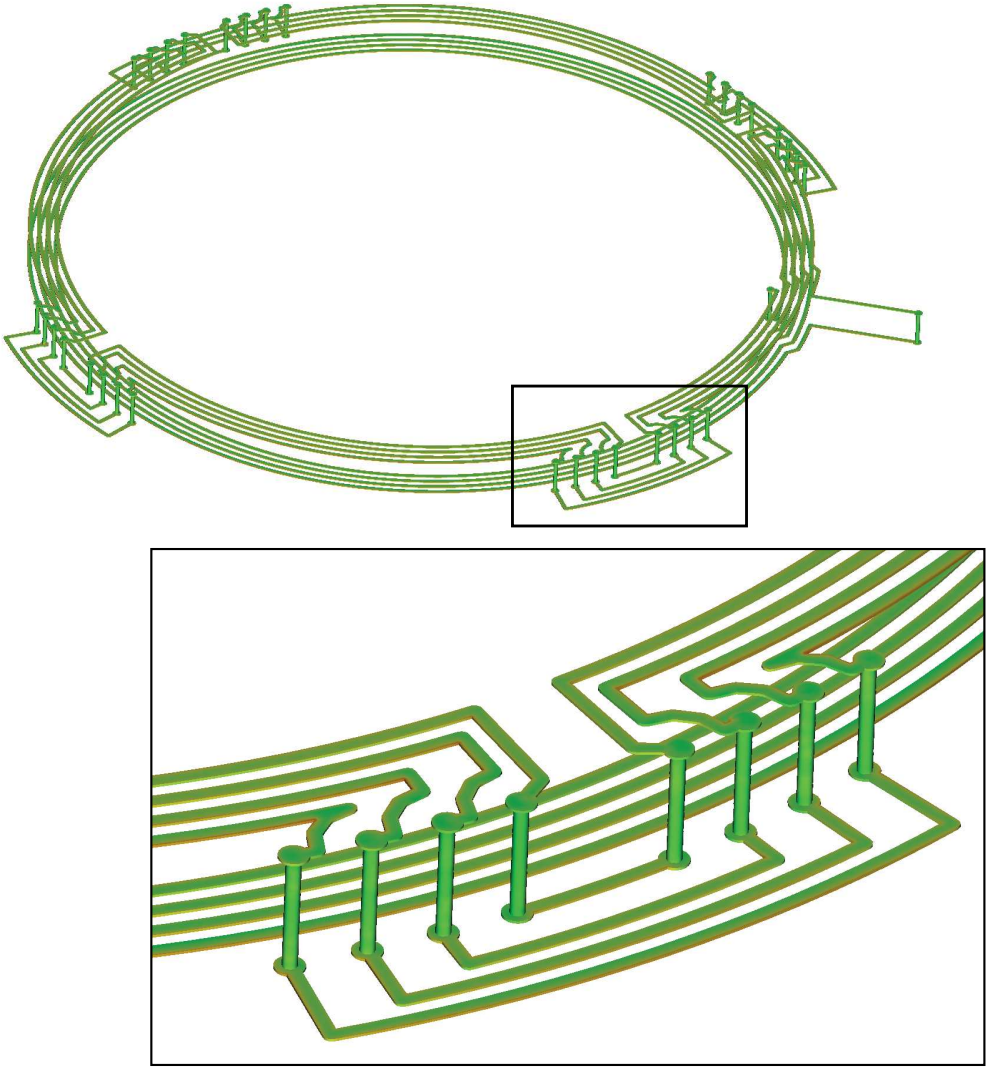

Figure 9.12: Main picture: colour map of the real part of the current density field in a PCB coil. Inset: zoom of current density distribution.

10

Conclusion

The results of the previous chapters allow us to draw the main conclusion of this thesis: the ability of expressing the MFD method using the geometric language is very important for improving its basic building blocks. Thus we emphasize the importance of seeing the geometry hidden behind the standard MFD method.

Many directions for follow up research can be imagined. We mention the high-order generalization of the constructions proposed in Part I, like the sparse inverse mass matrices, the methods for handling curved faces and computing discrete vector potentials. Regarding Part II, it is clear that all the techniques introduced for tetrahedra and general polyhedra can be readily adapted to triangles and general polygons. To deal with electrodes, magnetic materials and full Maxwell problems are the topics currently under investigation. In particular, the proposed basis functions and factorization can be universally applicable to any generalized mass matrices arising with any integral method or boundary element method.

Bibliography

[1] E. Tonti, *On the formal structure of physical theories.* Monograph of the Italian National Research Council, 1975.

[2] T. Tarhasaari, L. Kettunen, and A. Bossavit, "Some realizations of a discrete hodge operator: a reinterpretation of finite element techniques [for em field analysis]," *IEEE Trans. Magn.*, vol. 35, no. 3, pp. 1494–1497, 1999.

[3] E. Tonti, *The mathematical structure of classical and relativistic physics. A general classification diagram.* Birkhäuser, Basel, 01 2013.

[4] J. Munkres, *Elements of algebraic topology.* Perseus Books, Cambridge, MA, 1984.

[5] C. Mattiussi, "An analysis of finite volume, finite element and finite difference methods using some concepts from algebraic topology," *J. Comput. Phys.*, vol. 133, pp. 289–309, 1997.

[6] R. Hiptmair, "Discrete hodge-operators: an algebraic perspective," *PIER*, vol. 32, pp. 247–269, 2001.

[7] A. Bossavit, *Computational Electromagnetism.* Cambridge, MA: Academic Press, 1998.

[8] K. Lipnikov, G. Manzini, and M. Shashkov, "Mimetic finite difference method," *J. Comput. Phys.*, vol. 257, no. PB, pp. 1163–1227, jan 2014.

[9] O. Zienkiewicz and R. Taylor, "The finite element patch test revisited a computer test for convergence, validation and error estimates," *Computer Methods in Applied Mechanics and Engineering*, vol. 149, no. 1, pp. 223 – 254, 1997, containing papers presented at the Symposium on Advances in Computational Mechanics. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0045782597000856

[10] F. Brezzi, K. Lipnikov, and V. Simoncini, "A family of mimetic finite difference methods on polygonal and polyhedral meshes," *Mathematical Models and Methods in Applied Sciences*, vol. 15, no. 10, pp. 1533–1551, Oct. 2005.

[11] J. Bonelle, D. A. Di Pietro, and A. Ern, "Low-order reconstruction operators on polyhedral meshes: Application to compatible discrete operator schemes," *Computer Aided Geometric Design*, vol. 35-36, pp. 27–41, may 2015.

[12] L. Codecasa, R. Specogna, and F. Trevisan, "A new set of basis functions for the discrete geometric approach," *J. Comput. Phys.*, vol. 229, no. 19, pp. 7401–7410, 2010.

[13] ——, "Base functions and discrete constitutive relations for staggered polyhedral grids," *Comput. Meth. Appl. Mech. Eng.*, vol. 198, no. 9, pp. 1117 – 1123, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0045782508004234

[14] ——, "A new set of basis functions for the discrete geometric approach," *J. Comput. Phys.*, vol. 229, no. 19, pp. 7401 – 7410, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0021999110003384

[15] E. Tonti, "A direct discrete formulation of field laws: The cell method," *CMES*, vol. 2, no. 2, pp. 237–258, 2001. [Online]. Available: http://www.techscience.com/CMES/v2n2/24731

[16] ——, "Finite formulation of the electromagnetic field," *PIER*, vol. 32, pp. 1–44, 2001.

[17] A. Bossavit, "'Generalized Finite Differences' in computational electromagnetics," *Progress In Electromagnetics Research*, vol. 32, pp. 45–64, 2001.

[18] A. Palha, P. P. Rebelo, R. Hiemstra, J. Kreeft, and M. Gerritsma, "Physics-compatible discretization techniques on single and dual grids, with application to the poisson equation of volume forms," *J. Comput. Phys.*, vol. 257, pp. 1394 – 1422, 2014, physics-compatible numerical methods. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0021999113005330

[19] J. Bonelle and A. Ern, "Analysis of compatible discrete operator schemes for elliptic problems on polyhedral meshes," *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, vol. 48, no. 2, pp. 553–581, 2014. [Online]. Available: http://www.numdam.org/item/M2AN_2014__48_2_553_0

[20] J. Droniou, R. Eymard, T. Gallouet, and R. Herbin, "A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods," *Mathematical Models and Methods in Applied Sciences*, vol. 20, pp. 265–295, 2010.

[21] K. Lipnikov and G. Manzini, "A high-order mimetic method on unstructured polyhedral meshes for the diffusion equation," *J. Comput. Phys.*, vol. 272, pp. 360–385, 2014.

[22] D. A. D. Pietro and J. Droniou, "An arbitrary-order discrete de rham complex on polyhedral meshes: Exactness, poincaré inequalities, and consistency," *ArXiv*, vol. abs/2101.04940, 2021.

[23] L. Beirão da Veiga, F. Dassi, G. Manzini, and L. Mascotto, "Virtual elements for maxwell's equations," *Computers & Mathematics with Applications*, 2021.

[24] L. Beirão da Veiga, F. Brezzi, F. Dassi, L. Marini, and A. Russo, "Lowest order virtual element approximation of magnetostatic problems," *Computer Methods in Applied Mechanics and Engineering*, vol. 332, pp. 343–362, 2018.

[25] A. Ruehli, G. Antonini, and L. Jiang, *Circuit Oriented Electromagnetic Modeling Using the PEEC Techniques*, ser. Wiley - IEEE. Wiley, 2017.

[26] R. Albanese, G. Miano, G. Rubinacci, and R. Martone, "A T formulation for 3D finite element eddy current computation," *IEEE Transactions on Magnetics*, vol. 21, no. 6, pp. 2299–2302, 1985.

[27] R. Albanese and G. Rubinacci, "Finite Element Methods for the Solution of 3D Eddy Current Problems," *Advances in Imaging and Electron Physics*, 1997.

[28] L. Kettunen and L. R. Turner, "A volume integral formulation for nonlinear magnetostatics and eddy currents using edge elements," *IEEE Transactions on Magnetics*, vol. 28, pp. 1639–1642, 1992.

[29] J. F. Siau, G. Meunier, O. Chadebec, J.-M. Guichon, and R. Perrin-Bit, "Volume integral formulation using face elements for electromagnetic problem considering conductors and dielectrics," *IEEE Transactions on Electromagnetic Compatibility*, vol. 58, pp. 1587–1594, 2016.

[30] M. Kamon, M. J. Tsuk, and J. K. White, "Fasthenry: a multipole-accelerated 3-d inductance extraction program," *IEEE Transactions on Microwave Theory and Techniques*, vol. 42, no. 9, pp. 1750–1758, 1994.

[31] G. Rubinacci, A. Tamburrino, S. Ventre, and F. Villone, "A fast 3-d multipole method for eddy-current computation," *IEEE Transactions on Magnetics*, vol. 40, no. 2, pp. 1290–1293, 2004.

[32] S. Kurz, O. Rain, and S. Rjasanow, "The adaptive cross-approximation technique for the 3d boundary-element method," *IEEE Transactions on Magnetics*, vol. 38, no. 2, pp. 421–424, 2002.

[33] P. Alotto, P. Bettini, and R. Specogna, "Sparsification of bem matrices for large-scale eddy current problems," *IEEE Transactions on Magnetics*, vol. 52, pp. 1–4, 2016.

[34] W. Hackbusch, "A sparse matrix arithmetic based on h-matrices. part i: Introduction to h-matrices," *Computing*, vol. 62, pp. 89–108, 1999.

[35] S. Pitassi, F. Trevisan, and R. Specogna, "Explicit geometric construction of sparse inverse mass matrices for arbitrary tetrahedral grids," *Computer Methods in Applied Mechanics and Engineering*, vol. 377, p. 113699, 2021.

[36] "FMM3D," https://www.simonsfoundation.org/flatiron/software/, Accessed: June 2020.

[37] J. Cantarella, D. DeTurck, and H. Gluck, "Vector calculus and the topology of domains in 3-space," *The American Mathematical Monthly*, vol. 109, pp. 409 – 442, 2002.

[38] R. Benedetti, R. Frigerio, and R. Ghiloni, "The topology of Helmholtz domains," *arXiv: Geometric Topology*, 2010.

[39] S. H. Christiansen, "A construction of spaces of compatible differential forms on cellular complexes," *Mathematical Models and Methods in Applied Sciences*, vol. 18, pp. 739–757, 2008.

[40] M. Berger, M. Cole, and S. Levy, *Geometry I*, ser. Universitext. Springer Berlin Heidelberg, 2009.

[41] F. J. Branin, "The algebraic-topological basis for network analogies and the vector calculus," *Proceedings of the Symposium on Generalized Networks, Polytechnic Institute of Brooklyn*, vol. 16, pp. 453–491, 1966.

[42] M. Marrone, "Properties of constitutive matrices for electrostatic and magnetostatic problems," *IEEE Transactions on Magnetics*, vol. 40, no. 3, pp. 1516–1520, 2004.

[43] E. Tonti, "The mathematical structure of classical and relativistic physics: A general classification diagram," 2013.

[44] K. Lipnikov, G. Manzini, and M. J. Shashkov, "Mimetic finite difference method," *J. Comput. Phys.*, vol. 257, pp. 1163–1227, 2014.

[45] A. Bossavit, "How weak is the "weak solution" in finite element methods?" *IEEE Transactions on Magnetics*, vol. 34, no. 5, pp. 2429–2432, 1998.

[46] R. Specogna, "Complementary geometric formulations for electrostatics," *Int. J. Numer. Meth. Eng.*, vol. 86, no. 8, pp. 1041–1068, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/nme.3089

[47] F. Brezzi, K. Lipnikov, M. Shashkov, and V. Simoncini, "A new discretization methodology for diffusion problems on generalized polyhedral meshes," *Comput. Meth. Appl. Mech. Eng.*, vol. 196, no. 37, pp. 3682 – 3692, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0045782507000965

[48] R. Specogna, "One stroke complementarity for poisson-like problems," *IEEE Trans. Magn.*, vol. 51, no. 3, pp. 1–4, 2015.

[49] F. Brezzi, A. Buffa, and G. Manzini, "Mimetic scalar products of discrete differential forms," *J. Comput. Phys.*, vol. 257, no. PB, pp. 1228–1259, jan 2014.

[50] S. L. Campbell and C. D. Meyer, *Generalized Inverses of Linear Transformations*. Society for Industrial and Applied Mathematics, jan 2009.

[51] R. Specogna, "Complementary geometric formulations for electrostatics," *Int. J. Numer. Meth. Eng.*, vol. 86, no. 8, pp. 1041–1068, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/nme.3089

[52] J. Synge, *The hypercircle in mathematical physics*. Cambridge University Press, Cambridge, 1957.

[53] K. Lipnikov, M. J. Shashkov, and D. Svyatskiy, "The mimetic finite difference discretization of diffusion problem on unstructured polyhedral meshes," *J. Comput. Phys.*, vol. 211, pp. 473–491, 2006.

[54] S. Pitassi, R. Ghiloni, F. Trevisan, and R. Specogna, "The role of the dual grid in low-order compatible numerical schemes on general meshes," *J. Comput. Phys.*, vol. 436, p. 110285, 2021.

[55] I. D. Mishev, "Nonconforming finite volume methods," *Computational Geosciences*, vol. 6, pp. 253–268, 2002.

[56] I. Aavatsmark, "An introduction to multipoint flux approximations for quadrilateral grids," *Computational Geosciences*, vol. 6, pp. 405–432, 2002.

[57] F. Brezzi, K. Lipnikov, and M. J. Shashkov, "Convergence of mimetic finite difference method for diffusion problems on polyhedral meshes with curved faces," *Mathematical Models and Methods in Applied Sciences*, vol. 16, pp. 275–297, 2006.

[58] R. Sevilla, S. Fernández-Méndez, and A. Huerta, "Comparison of high-order curved finite elements," *International Journal for Numerical Methods in Engineering*, vol. 87, pp. 719–734, 2011.

[59] L. B. da Veiga, A. Russo, and G. Vacca, "The virtual element method with curved edges," *ESAIM: Mathematical Modelling and Numerical Analysis*, 2019.

[60] L. Botti and D. A. D. Pietro, "Assessment of hybrid high-order methods on curved meshes and comparison with discontinuous galerkin methods," *J. Comput. Phys.*, vol. 370, pp. 58–84, 2018.

[61] F. Brezzi, A. Buffa, and G. Manzini, "Mimetic scalar products of discrete differential forms," *J. Comput. Phys.*, vol. 257, pp. 1228–1259, 2014.

[62] F. Brezzi and M. Fortin, "Mixed and hybrid finite element methods," in *Springer Series in Computational Mathematics*, 1991.

[63] K. S. Yee, "Numerical Solution of Initial Boundary Value Problems Involving Maxwell's Equations in Isotropic Media," *IEEE Transactions on Antennas and Propagation*, vol. 14, no. 3, pp. 302–307, 1966.

[64] L. Codecasa and M. Politi, "Explicit, consistent, and conditionally stable extension of FD-TD to tetrahedral grids by FIT," *IEEE Transactions on Magnetics*, vol. 44, no. 6, pp. 1258–1261, jun 2008.

[65] B. He and F. L. Teixeira, "Differential forms, galerkin duality, and sparse inverse approximations in finite element solutions of maxwell equations," *IEEE Transactions on Antennas and Propagation*, vol. 55, no. 5, pp. 1359–1368, 2007.

[66] L. Codecasa, B. Kapidani, R. Specogna, and F. Trevisan, "Novel FDTD technique over tetrahedral grids for conductive media," *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 10, pp. 5387–5396, oct 2018.

[67] N. Bell and L. N. Olson, "Algebraic multigrid for k-form laplacians," *Numerical Linear Algebra with Applications*, vol. 15, no. 2âĂŘ3, pp. 165–185, 2008. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/nla.577

[68] A. Lemoine, J. Caltagirone, M. Azaïez, and S. Vincent, "Discrete helmholtz-hodge decomposition on polyhedral meshes using compatible discrete operators," *J. Sci. Comput.*, vol. 65, no. 1, pp. 34–53, 2015. [Online]. Available: https://doi.org/10.1007/s10915-014-9952-8

[69] E. Tonti, "A direct discrete formulation of field laws: The cell method," *CMES - Computer Modeling in Engineering and Sciences*, vol. 2, no. 2, pp. 237–258, 2001.

[70] A. N. Hirani, "Discrete Exterior Calculus Thesis by," Tech. Rep., 2003.

[71] R. Eymard, T. Gallouët, R. Herbin, and andRaphà ele Herbin, "Handbook of Numerical Analysis, 9780444503503. which appeared in Handbook of Numerical Analysis," *P.G. Ciarlet, J.L. Lions eds*, vol. 7, pp. 713–1020, 2000. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02100732v2

[72] A. Gillette and C. Bajaj, "Dual formulations of mixed finite element methods with applications," *Computer-Aided Design*, vol. 43, no. 10, pp. 1213 – 1221, 2011, solid and Physical Modeling 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S001044851100159X

[73] F. Bellina and R. Specogna, "Diagonal material matrices for arbitrary simplicial meshes for solving poisson problems with one unknown per element," *IEEE Transactions on Magnetics*, vol. 56, no. 2, pp. 1–4, 2020.

[74] A. Bossavit, "Mixed-hybrid methods in magnetostatics: complementarity in one stroke," *IEEE Transactions on Magnetics*, vol. 39, no. 3, pp. 1099–1102, 2003.

[75] M. S. Mohamed, A. N. Hirani, and R. Samtaney, "Numerical convergence of discrete exterior calculus on arbitrary surface meshes," *International Journal for Computational Methods in Engineering Science and Mechanics*, vol. 19, no. 3, pp. 194–206, 2018.

[76] L. Codecasa, R. Specogna, and F. Trevisani, "The discrete geometric approach for wave propagation problems," in *2009 International Conference on Electromagnetics in Advanced Applications*, 2009, pp. 59–62.

[77] M. VohralÃĐík and B. I. Wohlmuth, "Mixed finite element methods: implementation with one unknown per element, local flux expressions, positivity, polygonal meshes, and relations to other methods," *Math. Mod. Meth. Appl. S.*, vol. 23, no. 05, pp. 803–838, 2013.

[78] T. Tran-Cong, "On the potential of a solenoidal vector field," *Journal of Mathematical Analysis and Applications*, vol. 151, no. 2, pp. 557–580, 1990. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022247X9090166D

[79] M. B. Cohen, B. T. Fasy, G. L. Miller, A. Nayyeri, R. Peng, and N. J. Walkington, "Solving 1-laplacians in nearly linear time: Collapsing and expanding a topological ball," in *SODA*, 2014.

[80] J. P. Webb and B. Forghani, "A single scalar potential method for 3d magnetostatics using edge elements," *International Magnetics Conference*, pp. JD1–JD1, 1989.

[81] S. Pitassi, R. Ghiloni, F. Trevisan, and R. Specogna, "The role of the dual grid in low-order compatible numerical schemes on general meshes," *Journal of Computational Physics*, vol. 436, p. 110285, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999121001807

[82] Y. Le Menach, S. Clenet, and F. Piriou, "Determination and utilization of the source field in 3d magnetostatic problems," *IEEE Transactions on Magnetics*, vol. 34, no. 5, pp. 2509–2512, 1998.

[83] A. A. Rodríguez, E. Bertolazzi, R. Ghiloni, and A. Valli, "Construction of a finite element basis of the first de Rham cohomology group and numerical solution of 3d magnetostatic problems," *SIAM Journal on Numerical Analysis*, vol. 51, no. 4, pp. 2380–2402, 2013.

[84] P. Dłotko and R. Specogna, "Physics inspired algorithms for (co)homology computations of three-dimensional combinatorial manifolds with boundary," *Computer Physics Communications*, vol. 184, no. 10, pp. 2257–2266, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010465513001665

[85] A. Alonso Rodríguez, E. Bertolazzi, R. Ghiloni, and A. Valli, "Finite element simulation of eddy current problems using magnetic scalar potentials," *Journal of Computational Physics*, vol. 294, pp. 503–523, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999115002211

[86] Z. J. Silberman, T. R. Adams, J. A. Faber, Z. B. Etienne, and I. Ruchlin, "Numerical generation of vector potentials from specified magnetic fields," *Journal of Computational Physics*, vol. 379, pp. 421–437, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999118307940

[87] A. A. Rodríguez and A. Valli, "Finite element potentials," *Applied Numerical Mathematics*, vol. 95, pp. 2–14, 2015.

[88] P. Dlotko and R. Specogna, "Critical analysis of the spanning tree techniques," *SIAM J. Numer. Anal.*, vol. 48, pp. 1601–1624, 2010.

[89] B. Benedetti and F. Lutz, "Knots in collapsible and non-collapsible balls," *Electron. J. Comb.*, vol. 20, p. P31, 2013.

[90] M. Tancer, "Recognition of collapsible complexes is np-complete," *Discrete & Computational Geometry*, vol. 55, pp. 21–38, 2016.

[91] R. Forman, "Morse theory for cell complexes," *Advances in Mathematics*, vol. 134, pp. 90–145, 1998.

[92] D. Kozlov, "Combinatorial algebraic topology," in *Algorithms and computation in mathematics*, 2008.

[93] J. Whitehead, "Simplicial spaces, nuclei and mâĂŘgroups," *Proceedings of The London Mathematical Society*, pp. 243–327, 1939.

[94] G. Strang, "Introduction to linear algebra," 1993.

[95] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, "Introduction to algorithms, third edition," 2009.

[96] T. Lewiner, H. Lopes, and G. Tavares, "Optimal discrete morse functions for 2-manifolds," *Comput. Geom.*, vol. 26, pp. 221–233, 2003.

[97] B. Benedetti and F. H. Lutz, "Random discrete Morse theory and a new library of triangulations," *Experimental Mathematics*, vol. 23, pp. 66 – 94, 2014.

[98] K. Fujiwara and T. Nakata, "Results for benchmark problem 7 (asymmetrical conductor with a hole)," *Compel-the International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, vol. 9, pp. 137–154, 1990.

[99] M. M. Cohen, "A course in simple-homotopy theory," 1973.

[100] R. Bing, "Some aspects of the topology of 3-manifolds related to the poincare conjecture," 1964.

[101] R. Goodrick, "Non-simplicially collapsible triangulations of In," 1968.

[102] R. Furch, "Zur grundlegung der kombinatorischen topologie," *Abh.Math.Semin.Univ.Hambg.*, vol. 3, p. 69âĂŞ88, 1924.

[103] G. M. Ziegler, "Shelling polyhedral 3-balls and 4-polytopes," *Discrete & Computational Geometry*, vol. 19, pp. 159–174, 1998.

[104] P. Dlotko and R. Specogna, "Efficient generalized source field computation for h-oriented magnetostatic formulations," *European Physical Journal-applied Physics*, vol. 53, p. 20801, 2011.

[105] D. Voltolina, R. Torchio, P. Bettini, R. Specogna, and P. Alotto, "Optimized cycle basis in volume integral formulations for large scale eddy-current problems," *Computer Physics Communications*, vol. 265, p. 108004, 2021.

[106] M. Bebendorf and S. Rjasanow, "Adaptive low-rank approximation of collocation matrices," *Computing*, vol. 70, no. 1, pp. 1–24, 2003.

[107] Kezhong Zhao, M. N. Vouvakis, and Jin-Fa Lee, "The adaptive cross approximation algorithm for accelerated method of moments computations of emc problems," *IEEE Transactions on Electromagnetic Compatibility*, vol. 47, no. 4, pp. 763–773, 2005.

[108] R. Torchio, "A volume PEEC formulation based on the cell method for electromagnetic problems from low to high frequency," *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 12, pp. 7452–7465, 2019.

[109] A. C. Yucel, I. P. Georgakis, A. G. Polimeridis, H. Bagci, and J. K. White, "Voxhenry: FFT-accelerated inductance extraction for voxelized geometries," *IEEE Transactions on Microwave Theory and Techniques*, vol. 66, no. 4, pp. 1723–1735, 2018.

[110] R. Torchio, F. Lucchini, J.-L. Schanen, O. Chadebec, and G. Meunier, "FFT-PEEC: A fast tool from CAD to power electronics simulations," *IEEE Transactions on Power Electronics*, pp. 1–1, 2021.

[111] A. Bossavit, "On the numerical analysis of eddy-current problems," *Computer Methods in Applied Mechanics and Engineering*, vol. 27, no. 3, pp. 303 – 318, 1981.

[112] S. M. Rao, D. R. Wilton, and A. W. Glisson, *RWG Functions: Evolution and Progress*, 2016, pp. 1–13.

[113] S. Rao, D. Wilton, and A. Glisson, "Electromagnetic scattering by surfaces of arbitrary shape," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 3, pp. 409–418, 1982.

[114] D. Schaubert, D. Wilton, and A. Glisson, "A tetrahedral modeling method for electromagnetic scattering by arbitrarily shaped inhomogeneous dielectric bodies," *IEEE Transactions on Antennas and Propagation*, vol. 32, no. 1, pp. 77–85, 1984.

[115] J. C. Nedelec, "Mixed finite elements in $\mathbb{R}^3$," *Numerische Mathematik*, vol. 35, no. 3, pp. 315 – 341, 1980.

[116] R. Albanese and G. Rubinacci, "Integral formulation for 3D eddy-current computation using edge elements," *IEE Proceedings A*, vol. 135, no. 7, pp. 457–462, 1988.

[117] G. Rubinacci and A. Tamburrino, "Automatic treatment of multiply connected regions in integral formulations," *IEEE Transactions on Magnetics*, vol. 46, no. 8, pp. 2791–2794, aug 2010.

[118] H. A. Haus and J. R. Melcher, *Basic Circuit Theory.* New York: MacGraw-Hill, 1969.

[119] N. Balabian, *Electrical Network Theory.* Wiley, 1969.

[120] J. Siau, G. Meunier, O. Chadebec, J. Guichon, and R. Perrin-Bit, "Volume integral formulation using face elements for electromagnetic problem considering conductors and dielectrics," *IEEE Transactions on Electromagnetic Compatibility*, vol. 58, no. 5, pp. 1587–1594, 2016.

[121] P. Bettini, M. Passarotto, and R. Specogna, "A volume integral formulation for solving eddy current problems on polyhedral meshes," *IEEE Transactions on Magnetics*, vol. 53, no. 6, 2017.

[122] L. Codecasa, R. Specogna, and F. Trevisan, "A new set of basis functions for the discrete geometric approach," *J. Comput. Phys.*, vol. 229, no. 19, pp. 7401 – 7410, 2010.

[123] P. Dular, J. . Hody, A. Nicolet, A. Genon, and W. Legros, "Mixed finite elements associated with a collection of tetrahedra, hexahedra and prisms," *IEEE Transactions on Magnetics*, vol. 30, no. 5, pp. 2980–2983, 1994.

[124] S. Pitassi, R. Ghiloni, F. Trevisan, and R. Specogna, "The role of the dual grid in low-order compatible numerical schemes on general meshes," *Journal of Computational Physics*, vol. 436, p. 110285, 2021.

[125] P. Dłotko and R. Specogna, "Physics inspired algorithms for (co)homology computations of three-dimensional combinatorial manifolds with boundary," *Computer Physics Communications*, vol. 184, no. 10, pp. 2257–2266, 2013.

[126] J. R. Munkres, *Elements Of Algebraic Topology.* Westview Press, 1993.

[127] P. Bamberg and S. Sternberg, *A Course in Mathematics for Students of Physics.* Cambridge University Press, Cambridge, UK, 1988.

[128] S. Jarvenpaa, M. Taskinen, and P. YlÃď-Oijala, "Singularity extraction technique for integral equation methods with higher order basis functions on plane triangles and tetrahedra," *International Journal for Numerical Methods in Engineering*, vol. 58, no. 8, pp. 1149–1165, 2003.

[129] S. Jarvenpaa, M. Taskinen, and P. Yla-Oijala, "Singularity subtraction technique for high-order polynomial vector basis functions on planar triangles," *IEEE Transactions on Antennas and Propagation*, vol. 54, no. 1, pp. 42–49, 2006.

[130] D. Wilton, S. Rao, A. Glisson, D. Schaubert, O. Al-Bundak, and C. Butler, "Potential integrals for uniform and linear source distributions on polygonal and polyhedral domains," *IEEE Transactions on Antennas and Propagation*, vol. 32, no. 3, pp. 276–281, 1984.

[131] R. A. Werner and D. J. Scheeres, "Exterior gravitation of a polyhedron derived and compared with harmonic and mascon gravitation representations of asteroid 4769 castalia," *Celestial Mechanics and Dynamical Astronomy*, vol. 65, no. 3, 1997.

[132] ——, "Mutual potential of homogeneous polyhedra," vol. 91, no. 3-4, pp. 337–349, Mar. 2005.

[133] S. Bao, D. Wang, Y. Mo, S. Hu, J. Gu, and W. Tang, "Fully analytical evaluation of singular integrals with rwg and rooftop basis functions," *IEEE Journal on Multiscale and Multiphysics Computational Techniques*, vol. 5, pp. 217–226, 2020.

[134] Y. Saad and M. Schultz, "GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems," *SIAM Journal on Scientific and Statistical Computing*, vol. 7, no. 3, pp. 856–869, 1986.

[135] M. Nayfeh, *Electricity and magnetism.* Mineola, New York: Dover Publications, Inc, 2015.

[136] K. Fujiwara and T. Nakata, "RESULTS FOR BENCHMARK PROBLEM 7 (ASYMMETRICAL CONDUCTOR WITH a HOLE)," *COMPEL - The international journal for computation and mathematics in electrical and electronic engineering*, vol. 9, no. 3, pp. 137–154, Mar. 1990.

[137] "HLIBpro," https://www.hlibpro.com/, Accessed: May 2020.

[138] M. Passarotto, D. Klis, O. Rain, and R. Specogna, "Mirror symmetry in integral formulations for eddy currents," *IEEE Transactions on Magnetics*, vol. 57, no. 6, 2021.