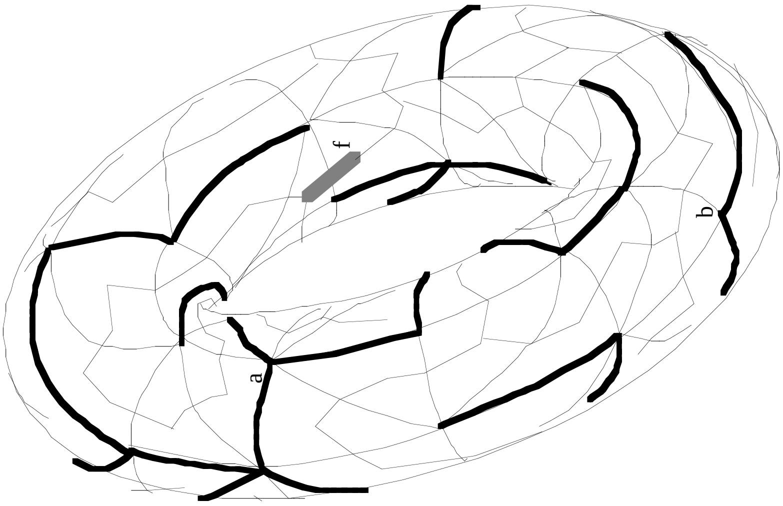
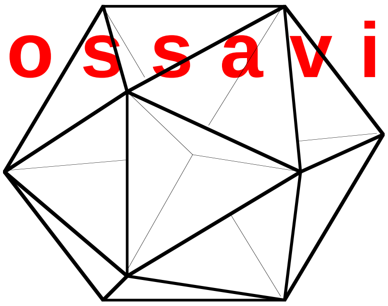
Computational Electromagnetism

Variational Formulations, Complementarity, Edge Elements



Alain Bossavit



COMPUTATIONAL ELECTROMAGNETISM

Variational Formulations,

Complementarity, Edge Elements

ACADEMIC PRESS SERIES IN ELECTROMAGNETISM

Electromagnetism is a classical area of physics and engineering which still plays a very important role in the development of new technology. Electromagnetism often serves as a link between electrical engineers, material scientists, and applied physicists. This series presents volumes on these aspects of theoretical electromagnetism that are becoming increasingly imporant in modern and rapidly developing technology. Its objective is to meet the needs of researchers, students, and practicing engineers.

This is a volume in

ELECTROMAGNETISM

ISAAK MAYERGOYZ, SERIES EDITOR

UNIVERSITY OF MARYLAND, COLLEGE PARK, MARYLAND

COMPUTATIONAL ELECTROMAGNETISM

Variational Formulations, Complementarity, Edge Elements

Alain Bossavit

Électricité de France

ACADEMIC PRESS

San Diego London Boston
New York Sydney Tokyo Toronton

11.7 × 19.6

Use this page to check dimensions and fonts. Configure your printing system to make this page's frame about 11.7 cm large and 19.6 cm high. This font is Palatino 10.

The first edition of this book was plagued by font vagaries, to be blamed on my naive belief in the technical competence of the Publisher's staff in handling .ps files. Please use the checklist below to verify that your copy won't suffer the same way. If all goes well, you should see the following items as described:

- grad, $\mathbb{grad}$
  plain vs. "outline" style, used around pp. 295ff.

- dot-product symbols “ $\cdot$ ” and “ $\circ$ ”
  They should look distinctively different

- $\overline{A}$, $\overline{A}$
  Capital A with overbar

- φ, $\boldsymbol{\varphi}$, Φ, $\boldsymbol{\Phi}$
  Phi, set in plain, **bold**, SMALLCAP, and **BOLD SMALLCAP**

- $\mathcal{A}$
  Curly A.

- $\tilde{\text{H}}$
  Smallcap h, with tilde on top.

- $\mathbb{Z}$, $\mathbb{Z}$
  "Blackboard" capital Z, in sizes 10 and 8 (as used, e.g., p. 151).

Warnings and advice welcome, `bossavit@lgep.supelec.fr`

Errata for the first edition (and later, likely, for *this* one), to be found at
`http://www.lgep.supelec.fr/mse/perso/ab/bossavit.html`

AB, 20 11 03

Foreword

This book is the second volume in the Academic Press Electromagnetism Series, written by Professor Alain Bossavit, one of the most active researchers in the area of electromagnetic field calculations. Professor Bossavit is well known and highly regarded in the electromagnetic community for his seminal contributions to the field of computational electromagnetics. In particular, he has pioneered and strongly advocated the use of edge elements in field calculations. These elements, which are now widely accepted by engineers, have become indispensable tools in numerical analysis of electromagnetic fields. His work on the use of symmetry in numerical calculations, computational implementation of complementarity, and evaluation of electromagnetic forces have also been extremely important for the development of the field.

This book reflects the unique expertise and extensive experience of the author. It is written with a strong emphasis on comprehensive and critical analysis of the foundations of numerical techniques used in field calculations. As a result, the book provides many valuable insights into the nature of these techniques. It contains information hardly available in other sources and no doubt will enrich the reader with new ideas and a better conceptual understanding of computational electromagnetics. The material presented in the book can be expected to contribute to the development of new and more sophisticated software for electromagnetic field analysis.

The book is distinctly unique in its original style of exposition, its emphasis, and its conceptual depth. For this reason, it will be a valuable reference for both experts and beginners in the field. Researchers as well as practitioners will find this book challenging, stimulating, and rewarding.

Isaak Mayergoyz, *Series Editor*

Contents

CHAPTER 7 Infinite Domains 193

CHAPTER 8 Eddy-Current Problems 219

Preface

Computational electromagnetism begins where electromagnetic theory stops, and stops where engineering takes over. Its purpose is not to establish Maxwell equations, and other essential physical theories, but to use them in *mathematical modelling,* with concrete problems in view.

Modelling is this activity by which questions about a physical situation are translated into a mathematical problem—an *equation,* if this word is understood is a general enough sense—that will be solved, in order to answer these questions. "Solving", nowadays, means using a computer, in most cases. The equations one aims at, therefore, can very well be huge systems of linear equations (solved as part of some iterative process, in nonlinear situations, or in the simulation of transient phenomena). Complex shapes, non-uniform physical characteristics, changing configurations, can and should be taken into account. Adequate methods—not necessarily the exclusivity of electromagnetics—have been developed for this: finite elements, boundary integral methods, method of moments . . . Strengthening their foundations, clarifying their presentation, enhancing their efficiency, is the concern of computational electromagnetism.

The contribution of this book to such a large subject will necessarily be limited. Three main topics are treated:

- *Variational formulations,* understood from a functional viewpoint,
- *Edge elements,* and related finite elements,
- *Complementarity.*

These are not, by any means, the definitive pillars on which the whole theory should be erected. Rather three posts, or stakes, on which I believe some platform can be built, with a good view on the whole subject, provided the foundations are steady. A relatively thick Appendix, entitled

- *Mathematical background,*

has been included with this in mind, which should either provide the prerequisites, or give directions to locate and study them.

Such emphasis on foundations does not imply disregard for such issues as implementation, algorithmic efficiency, and relevance of numerical results, which are all important in modelling. There is a rich and fastly growing literature on all this, and lots of opportunities for beginners to get firsthand information at conferences and specialized gatherings (such as, for example, the TEAM Workshop, frequently mentioned in this book). Analyses of the *conceptual bases* of modern methods, on the other hand, are much rarer, and there is a dearth of courses from which rapidly to learn the basic notions gathered in the present book. Yet, these notions are needed by all those who *do* conceive, implement, test, and run electromagnetic software.

I speak here not as a mathematics teacher but as a programmer who had to work on numerical simulations of electrical heating processes, and strongly felt the urge to understand what he was doing, at a time where the ideals of "structured programming" were gaining universal favor, while the current understanding of the finite element method seemed unable to provide the required guidelines.

That was twenty years ago. The finite-element treatment of the Laplace equation ($\Delta\varphi = 0$, or rather, $\operatorname{div}(\mu \operatorname{grad} \varphi) = 0$, in the context of magnetostatics) was well understood, and its application to major problems in electromagnetism, such as computing the pattern of magnetic lines in the cross-section of a rotating machine, was vigorously promoted by energetic leaders, the late P.P. Silvester among them. Tremendous successes were obtained in 2D simulations, and the first commercial codes reached the market. But the passage to three dimensions proved very difficult—an intriguing situation, since it didn't seem to be encountered in neighboring fields such as heat transfer or fluid dynamics.

With hindsight, we now understand why it was so. The Laplace equation is not *the* paradigmatic equations in electromagnetism. There are two of them. The other one is $\operatorname{rot}(\mu^{-1} \operatorname{rot} \mathbf{a}) = \mathbf{j}$, which governs the vector potential in magnetostatics. In dimension 2, where $\mathbf{a} = \{0, 0, \mathrm{a}\}$, with a unique nonzero scalar component a, this vector equation reduces to $-\operatorname{div}(\mu^{-1} \operatorname{grad} \mathrm{a}) = \mathrm{j}$. Hence the easy transposition, and the illusion that one standard model would be enough. But in dimension 3, *the* div–grad *and the* curl–curl *equations are deeply different, and require different treatment.*

This was not obvious in the 1970s. Since $\operatorname{rot}\operatorname{rot} \mathbf{a} = \operatorname{grad}\operatorname{div} \mathbf{a} - \Delta\mathbf{a}$, it could seem possible to transpose the 2D methods to 3D situations by first imposing the "gauge condition" $\operatorname{div} \mathbf{a} = 0$—hence $-\Delta\mathbf{a} = \mathbf{j}$—and then working on the three scalar components of $\mathbf{a}$ separately.[1] Natural as it was, the idea led to a blind alley, and it took years to realize this. We were fortunate, J.C. Vérité and myself, to enter the field at precisely this period of unrest, with a different idea about how to deal with the curl–curl term in the eddy-current equation $\partial_t(\mu\mathbf{h}) + \mathrm{rot}(\sigma^{-1}\mathrm{rot}\,\mathbf{h}) = 0$, and an early implementation (the 'Trifou' code, some salient features of which will be described in this book).

Our idea was inspired by the belief that *network methods* had the best prospects for 3D generalization. In such methods (see [TL], for instance), it was customary to take as basic unknown the electromotive force (emf) along the branch of a circuit and to apply Kirchhoff-like laws to set up the equations. For reasons which will be explained in Chapter 8, the magnetomotive force (mmf) along branches seemed preferable to us, and it was clear that such mmf's, or edge circulations of the magnetic field, *h a d* to be the degrees of freedom in an eddy-current computation. As we had devised variational formulations[2] for the eddy-current equation, the problem was to be able to *interpolate from edge mmf's:* Knowing the circulation of $\mathbf{h}$ along edges of a tetrahedral finite element mesh, by which interpolating formula to express the magnetic field inside each tetrahedron? The help of J.C. Nedelec, who was consultant at EdF at the time (1979), was decisive in providing such a formula: $\mathbf{h}(\mathrm{x}) = \boldsymbol{\alpha} \times \mathrm{x} + \boldsymbol{\beta}$, where x is the position, and $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, two ordinary vectors, tetrahedron-dependent. 'Trifou' was coded with this vector-valued shape function, applied to a test problem [BV], and the idea of "solving directly for $\mathbf{h}$ with edge elements" gradually gained acceptance over the years.[3]

However, the analytic form of these shape functions was a puzzle (why precisely $\boldsymbol{\alpha} \times \mathrm{x} + \boldsymbol{\beta}$?) and remained so for several years. A key piece was provided by R. Kotiuga when he suggested a connection with a little-known compartment of classical differential geometry, Whitney forms [Ko]. This was the beginning of a long work of reformulation and

[1] Plus a fourth unknown field, the electric potential ψ, in transient situations. The pair $\{\mathbf{a}, \psi\}$ fully describes the electromagnetic field, as one well knows, so such a choice of unknown variables was entirely natural. What was wrong was not the $\{\mathbf{a}, \psi\}$ approach per se, but the *separate* treatment of Cartesian components by *scalar*-valued finite elements, which resulted, in all but exceptional cases of uniform coefficients, in hopeless entanglement of components' derivatives. Much ingenuity had to be lavished on the problem before $\{\mathbf{a}, \psi\}$-based methods eventually become operational. Some reliable modern codes do use them.

[2] One of them, using the current density as basic unknown, was implemented [Vé]. It was equivalent (but we were not aware of it at the time) to a low-frequency version of Harrington's "method of moments" [Ha].

[3] 'Trifou', now the reference code of Électricité de France for all electromagnetic problems, continues to be developed, by a task–crew under the direction of P. Chaussecourte.

refoundation, the result of which is given in Chapter 5, where Whitney forms are called "Whitney elements". Although the picture thus given is not definitive, we now can boast of understanding the whole structure of Maxwell equations, both in "continuous" and "discrete" form, with enough detail to firmly ground an orderly object-oriented programming approach.

This is what this book tries to convey. Its strong mathematical component stems from the desire to help the reader track all concepts back to their origin, in a self-contained package. The approach is descriptive and concept-oriented, not proof-oriented. (Accordingly, the Subject Index points only to the page where a *definition* is provided, for most entries.)

Motivations thus being given, it's time to detail the contents. After an introductory Chapter 1, which reviews Maxwell equations and derived models, Chapter 2 focuses on one of them, a very simple model problem in magnetostatics, which is treated in Chapter 3 by solving for the scalar potential. This is intended as an introduction to *variational formulations.*

Variational methods have an ancient and well known history. The solution to some field problem often happens to be the one, among a family of a priori "eligible" fields, that minimizes some energy-related quantity, or at least makes this quantity stationary with respect to small variations. By restricting the search of this optimum to a well-chosen *finite* subfamily of eligible fields, one obtains the desired finite system of equations, the solution of which provides a near-optimum. This powerful heuristics, or *Galerkin method*, applicable to most areas of physics, leads in a quite natural way to finite element methods, as we shall see in Chapters 2 and 3, and suggests a method for error analysis which makes most of Chapter 4.

At this stage, we are through with the div–grad operator, and wish to address the curl–curl one, by treating the same problem in vector potential formulation. The same approach is valid, provided the right finite elements are available, and ...

Edge elements are it; at least, if one takes a large enough view, by considering the whole family of "Whitney elements" (as the foregoing Whitney forms are called in this book), a family which includes the standard node-based scalar elements, edge elements (with degrees of freedom associated with edges of a mesh), face elements, etc. And with these, we discover mathematical structures which, if not foreign to other branches of physics, at least seem ready-made for electromagnetism. Reasons for this marvelous adequacy will be suggested: It has to do with the essentially geometric character of Maxwell equations, which revolve around different versions of the Stokes theorem. This is the substance of Chapter 5. Although the "classical" language used in this work (only vectorial analysis, no differential geometry at all) does not permit to fully explain such things, I hope the reader will feel some curiosity for more advanced books, in which the differential geometric viewpoint would prevail from the onset, and for the specialized literature. Some informal forays into homology and the basics of topology should help show the way.

At the end of Chapter 5, we shall have a better view of the classical scalar potential vs vector potential diptych, and be ready to apply to the curl–curl side what we learned by close scrutiny of the div–grad panel. But the symmetry thus revealed will suggest a new idea, which is the essence of *complementarity.* Indeed, both approaches lead to a solution of the model problem, but using *both* of them, in a deliberately redundant way, brings valuable additional benefits: a posteriori error *bounds* (much better than asymptotic *estimates),* bilateral bounds on some important results, usable error estimators for mesh-refinement procedures. Chapter 6 implements this working program, revisiting the time-honored "hypercircle" idea (cf. p. 171) along the way. Emphasis in this chapter, as elsewhere, is on the structures which underlie the methods, and finer details have been relegated to Appendices B and C.

A technical, but important, issue then comes and makes Chapter 7: how to deal with infinite regions with only a finite number of elements. *Integral methods* on an artificial boundary, in association with finite elements in a bounded region, solve this problem.

It's only then that we shall go beyond magnetostatics. The strong imbalance of this book's contents in favor of such a limited subject may surprise the reader. But this limited scope is precisely what makes magnetostatics such a good model with which to explain concepts of much broader application. The last two chapters are intended to illustrate this point by a rapid examination of a few other, more complex, models.

Chapter 8 is devoted to *eddy current* problems. This can serve as a short theoretical introduction to the Trifou code, with in particular a detailed account of how the hybrid method of Chapter 7 (finite elements inside the conductor, boundary elements on the air–conductor interface) is applied.

Chapter 9 then addresses the "microwave oven" problem, that is, Maxwell equations in harmonic regime in a bounded region. The point is to show how easily, once variational techniques and edge elements are mastered, one can pass from the basic equations to solvable linear systems.

It's also another opportunity to see complementarity and the symmetry of Maxwell equations at work.

Appendix A contains the mathematical background. Although this is independent material, which can be used in many different ways (as a template for a tutorial, as a guide for self-study, or as a reference section, with help of the Subject Index), one may find it easier to read it in parallel with Chapters 1 to 5. From Chapter 6 onwards, the technical level will be significantly higher, and familarity with this mathematical background will be assumed.

References to the literature have been placed at the end of each chapter. This has advantages, but also entails some repetitions, and perhaps difficulty in locating some quotations. The Authors' Index is meant to alleviate such difficulties, as far as possible.

It's a pleasure to acknowledge the influence of many friends in this work: P. Asselin, J.P. Bastos, F. Bouillault, A. di Carlo, J. Carpenter, P. Chaussecourte, R. Kotiuga, A. Nicolet, I. Nishiguchi, L. Pichon, A. Razek, E. Tonti, W. Trowbridge, A. Trykozko, I. Tsukerman, J.C. Vérité, and many other colleagues. I thank Isaac Mayergoyz, Series Editor, for his suggestion to write this book and his help in shaping the contents, and the Academic Press staff for all aspects of its realization. Thanks are due also to R. Temam, D. Ioan, A. Jami for various opportunities to teach this material and test students' reactions. Special thanks to L. Kettunen and I. Munteanu, who read the manuscript at various stages of its inception, and contributed several valuable additions.

The one person, however, whose support most contributed to this book, is my wife. Ginette, my apologies and love.

REFERENCES

[BV] A. Bossavit, J.C. Vérité: "A Mixed FEM-BIEM Method to Solve Eddy-Current Problems", **IEEE Trans., MAG-18**, 2 (1982), pp. 431–435.

[Ha] R.F. Harrington: **Field Computation by Moment Methods,** Macmillan (New York), 1968.

[Ko] P.R. Kotiuga: **Hodge Decompositions and Computational Electromagnetics** (Thesis), Department of Electrical Engineering, McGill University (Montréal), 1984.

[TL] L.R. Turner, R.J. Lari: "Developments in Eddy Current Computation with EDDYNET", **IEEE Trans. Magn., MAG-19** (1983), pp. 2577–2580.

[Vé] J.C. Vérité: “Calcul tridimensionnel des courants de Foucault dans un solide non ferromagnétique”, in **Conference Digests, Compumag 1978** (ENSGP, Grenoble, 1978), paper 7.3.

COMPUTATIONAL ELECTROMAGNETISM

Variational Formulations,

Complementarity, Edge Elements

CHAPTER 1

Introduction: Maxwell Equations

1.1 FIELD EQUATIONS

Computational electromagnetism is concerned with the numerical study of *Maxwell equations,*

(1) $- \partial_t \mathrm{d} + \mathrm{rot}\, \mathrm{h} = \mathrm{j},$ (2) $\partial_t \mathrm{b} + \mathrm{rot}\, \mathrm{e} = 0,$

(3) $\mathrm{d} = \varepsilon_0\, \mathrm{e} + \mathrm{p},$ (4) $\mathrm{b} = \mu_0\, (\mathrm{h} + \mathrm{m}),$

completed by *constitutive laws,* in order to account for the presence of matter and for the field-matter interaction. This introductory chapter will explain the symbols, discuss constitutive laws, and indicate how a variety of mathematical models derive from this basic one.

The vector fields e, h, d, b are called *electric field,*[1] *magnetic field, magnetic induction,* and *electric induction,* respectively. These four[2] vector fields, taken together, should be construed as the mathematical representation of a physical phenomenon, that we shall call the *electromagnetic field.* The distinction thus made between the physical reality one wants to model, on the one hand, and the mathematical structure thanks to which this modelling is done, on the other hand, is essential. We define a *model* as such a mathematical structure,[3] able to account,

[1] *Italics,* besides their standard use for emphasis, signal notions which are implicitly defined by the context.

[2] Two should be enough, after (3) and (4). Reasons for this redundancy will come.

[3] The structure, in this case, is made of the equations *and* of the framework in which they make mathematical sense: Euclidean three-dimensional space, and time-dependent entities, like scalar or vector fields, living there. There are other possible frameworks: the algebra of differential forms ([Mi], Chapter 4), Clifford algebra [Hs, Ja, Sa], etc. As Fig. 1.1 may suggest, Maxwell's theory, as a *physical* theory, should not be confused with any of its mathematical descriptions (which are historically transient; see [Cr, Sp]).

within some reasonably definite limits, for a class of concrete physical situations. To get a quick start, no attempt is made here either to *justify* the present model, on physical grounds, or to *evaluate* it, in comparison with others. (In time, we'll have to pay for this haste.)

The *current density* j, *polarization* p, and *magnetization*[4] m are the *source*-terms in the equations. Each contributes its own part, as we shall see, to the description of electric charges, at rest or in motion, whose presence is the physical cause of the field. Given j, p, and m, as well as initial values (at time $t = 0$, for instance) for e and h, Eqs. (1–4) determine e, h, d, b for $t \geq 0$. (This is no trivial statement, but we shall accept it without proof.) Maxwell's model (1–4) thus accounts for situations where j, p, and m are known in advance and independent of the field. This is not always so, obviously, and (1–4) is only the head of a *series* of models, derived from it by adding features and making specific simplifications, some of which will be described at the end of this chapter.

$$\boxed{\begin{array}{c} 4\pi\mathcal{C} = \mathrm{V}\cdot\nabla\mathcal{H}, \quad \mathcal{C} = \mathrm{C}\mathcal{E} + \dot{\mathcal{D}}, \quad \mathcal{B} = \mathrm{V}\cdot\nabla\mathcal{U}, \quad \mathcal{E} = \mathrm{V}\cdot\dot{\mathcal{R}}\mathcal{B} - \dot{\mathcal{U}} - \nabla\Psi, \\ \mathcal{B} = \mu\mathcal{H}, \qquad\qquad \mathcal{D} = (4\pi)^{-1}\,\mathrm{K}\mathcal{E}, \end{array}}$$

$$\boxed{dF = 0, \quad G = *F, \quad dG = J}$$

FIGURE 1.1. Maxwell equations: as they appear in [Ma], Art. 619 (top box), and in modern differential geometric notation (bottom box). Maxwell's formalism, still influenced by quaternionism (∇ is the operator $i\,d/dx + j\,d/dy + k\,d/dz$, and the V means "vector part" of a quaternionic product), is not so remote from today's standard vector notation, once the symbols have been identified (Ψ and $\mathcal{U}$ are scalar and vector potential, $\dot{\mathcal{R}}$ is material velocity). In this book's notation, and apart from factors 4π, the upper line would read $\mathrm{rot}\ h = C = \sigma e + \partial_t d$, $b = \mathrm{rot}\ a$, $e_{mat} = v \times b - \partial_t a - \nabla\psi$, where e_{mat} is the electric field in the comoving frame of reference.

You may be intrigued, if not put off, by the notation. The choice of symbols goes against recommendations of the committees in charge of such matters, which promote the use of **E**, **H**, **D**, **B**, capital and boldface. Using e, h, d, b instead is the result of a compromise between the desire to keep the (spoken) *names* of the symbols as close as possible to accepted practice and the notational habits of mathematics, capitals for functional spaces and lower case for their elements, according to a hierarchy which reflects

[4] Magnetization could more symmetrically be defined as m such that $b = \mu_0 h + m$. The present convention conforms to the dominant usage.

the *functional point of view* adopted in this book. (Explanations on this fundamental point will recur.) Boldface, still employed in the Preface for 3D vectors, according to the standard convention due to Heaviside [Sp], will from now on be reserved for another use (see p. 71). I should also perhaps call attention to the use of the ∂ symbol: If b is a time-dependent vector field, $\partial_t b$ is the field obtained by differentiating b with respect to time. Having thus $\partial_t b$ instead of $\partial b/\partial t$ is more than a mere ink-saving device: It's a way to establish the status of ∂_t as an *operator,* on the same footing as grad, div, and rot (this will denote the curl operator) which all, similarly, yield a field (scalar- or vector-valued, as the case may be) when acting on a field—the functional viewpoint, again. Other idiosyncrasies include the use of constructs such as $\exp(i\omega t)$ for $e^{i\omega t}$ and, as seen here, of i for the square root of -1, instead of j.[5]

This being said, let's return to our description. Equation (1) is *Ampère's theorem*. Equation (2) is *Faraday's law*. The term $\partial_t d$, whose introduction by Maxwell[6] was the crowning achievement of electromagnetic theory, is called *displacement current.* One defines *electric charge* (expressed in *coulombs* per cubic meter) by

$$q = \text{div } d, \tag{5}$$

a scalar field. According to (1), one has thus

$$\partial_t q + \text{div } j = 0, \tag{6}$$

with j expressed in *ampères* per square meter. Notice that if j is given, from the origin of times to the present, one gets the charge by integration with respect to time: assuming j and q were both null before time 0, then $q(t, x) = -\int_0^t (\text{div } j)(s, x)\, ds$.

If the local differential relation (6) is integrated by applying the Ostrogradskii (Gauss) theorem to a regular[7] spatial domain D bounded by some surface S (Fig. 1.2), one finds that

[5] The shift from i to j was motivated by the desire, at a time when the power of complex numbers in alternating currents theory began to be realized, to denote intensities with the i symbol. Since almost everybody calls "jay" the current density vector, notwithstanding, it makes little sense to perpetuate the dual use of j as the square root of -1. (This remark is respectfully brought to the attention of the above-mentioned Committees.)

[6] Around 1860, and 1873 saw the first edition of his treatise. The classic version we read nowadays [Ma] is the third edition.

[7] "Domain" has a technical meaning: an open set in one piece. Cf. A.2.3. "Regular" means that D is enclosed by one or several surfaces, themselves smooth at all points, with the possible exception of a finite number of corners and edges (Fig. 1.2).

$$\frac{d}{dt}\int_D q+\int_S n\cdot j=0, \tag{7}$$

where n denotes the field of normal vectors, of length 1 and outwardly directed with respect to D, on surface S.[8] The first term in this equality is the increase, per unit of time, of the charge contained in D, whereas the second term is the outgoing flux of charge. They balance, after (7), so (6) is the local expression of *charge conservation*.

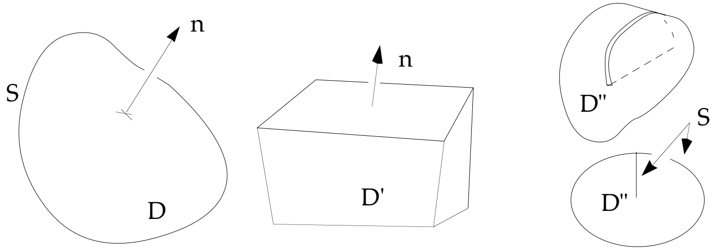
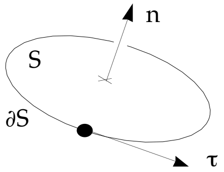


FIGURE 1.2. Notion of *regular domain* (D, on the left), and notations (cf. Note 7). In spite of singularities, D' can still pass as regular (edges and corners form a negligible set), but domain D" on the right (shown in cut view) doesn't qualify, because D" is "on both sides" of a part of its boundary. This geometrical idealization is still useful in the case of small air gaps, cracks, etc., but some care must then be exercised in regard to formulas like (7).

Quite similarly, Eqs. (1) and (2) can be integrated by using the Stokes theorem, hence global (integral) expressions which express flux and current conservation. For instance, the integral form of Faraday's law is

$$\frac{d}{dt}\int_S n\cdot b+\int_{\partial S}\tau\cdot e=0,$$

where S is a surface, ∂S its boundary, and τ a field of unitary tangent vectors on ∂S (inset),



[8] See Appendix A, Subsection A.4.2, for the notions of *flux* $\int n\cdot j$ and *circulation* $\int \tau\ .\ e$, and justification of this notation. When necessary, I denote by dx the volume element, or the area element, according to whether the integral is over a volume or a surface, but each time this does not foster confusion, I'll omit this symbol: $\int f$ rather than $\int f(x)\,dx$. If you do that, don't stop half-way: never $\int f(x)$ alone, without the dx. The symbol x in dx is meant to match with the x in f(x), as demonstrated by the fact that you may substitute at both places some other letter, say y, without changing the meaning. Avoid also constructs such as $\int_S f\ dS$ (although one could make a case for them): it is understood that the integration is with respect to the measure of areas that exists on S, and thus dS is superfluous. The construct $\int_S f$ makes perfect sense by itself (cf. A.4.2): It's the effect on f and S, taken as a pair, of the integration operator.

oriented with respect to n as prescribed by Ampère's rule. Historically, such integral formulations came first and are arguably more germane to physics. Indeed, we shall have to spend some time on correcting some drawbacks of the local differential formulation (or rather, of a too literal interpretation of this formulation).

Treatises on electromagnetism often add two equations to (1–4), namely (5) and $\text{div } b = 0$. But the latter stems from Faraday's law (2), if one assumes a null b (or even just a null $\text{div } b$) before initial time, and (5) is here a definition. So there would be little justification in according to these relations the same status as (1) and (2).

A (rightful) concern for formal symmetry might suggest writing (2) as $\partial_t b + \text{rot } e = - k$, where k would be a given field, the *magnetic current,* and defining *magnetic charge,* expressed in *webers* per cubic meter, as $q_m = \text{div } b$ (electric charge q would then be denoted by q_e), hence the equation $\partial_t q_m + \text{div } k = 0$, which would express magnetic charge conservation. But since k and q_m are null in all known physical situations,[9] this generalization seems pointless.

Now, let us address Eqs. (3) and (4). As the next Section will make clear, the (mathematical) fields e and b suffice to describe the effect of the (physical) electromagnetic field on the rest of the world, in particular on charged particles, whose motion is described by j, p, m, and which in turn constitute the source of the field. The electromagnetic field is thus kinematically[10] characterized by the pair $\{e, b\}$, and fields d and h are auxiliaries in its dynamic description. Moreover, there is some leeway in the very definition of d and h, because the bookkeeping on charge motion can be shared between j, p, and m in different ways.

Exercise 1.1.[11] Rewrite (1–4) by eliminating d and h. Discuss the interchangeability of j, p, and m.

Equations (3) and (4) thus seem to define redundant entities, and indeed, many classical presentations of electromagnetism make do with two vector

[9] Magnetic monopoles, the density of which would be the above q_m, "should exist" [GT], according to theoreticians, but have not been observed yet. (Reports of such observations have been made [Ca, P&], but were not confirmed.) It is comforting to know that the discovery of such particles would not jeopardize Maxwell's theory.

[10] *Kinematics* is about description: which mathematical entities depict the system's state at any instant. *Dynamics* is about evolution laws: how the state will change under external influences.

[11] Texts of exercises are either at the end of each chapter or, when short enough, given on the spot, in which case **Exercise** is in boldface. Look for the "Hints" and "Solutions" sections at the end of each chapter.

fields[12] instead of four. The main advantage of their presence, which explains why this formalism is popular in the computational electromagnetics community, is the possibility this offers to express material properties in a simple way, via "constitutive laws" which relate j, p, and m to the electromagnetic field they generate.

The vacuum, in particular, and more generally, matter that does not react to the field, is characterized by $p = 0$ and $m = 0$, and thus by the coefficients ε_0 and μ_0. In the MKSA system, $\mu_0 = 4\pi\, 10^{-7}$ H/m and $\varepsilon_0 = 1/(\mu_0 c^2)$ F/m, where c is the speed of light (H for *henry* and F for *farad*). These values reflect two things: A fundamental one, which is the very existence of this constant c, and a more contingent one, which is the body of conventions by which historically established units for electric and magnetic fields and forces have been harmonized, once the unity of electromagnetic phenomena was established.

Let us now review these constitutive laws, which we will see are a condensed account ot the laws of charge–matter interaction in specific cases.

1.2 CONSTITUTIVE LAWS

In all concrete problems, one deals with composite *systems,* analyzable into subsystems, or *compartments:* electromagnetical, mechanical, thermal, chemical, etc. Where to put the boundaries between such subsystems is a modelling decision, open to some arbitrariness: elastic forces, for instance, can sensibly be described as electromagnetic forces, at a small enough scale. Each compartment is subject to its own equations (partial differential equations, most often), whose right-hand sides are obtained by solving equations relative to *other* compartments. For instance, Eqs. (1–4) govern the electromagnetic compartment, and we'll soon see how j, p, and m are provided by others. If one had to deal with all compartments at once, and thus with *coupled* systems of partial differential equations of considerable complexity, numerical simulation would be very difficult. *Constitutive laws,* in general, are the device that helps bypass this necessity: They are an approximate but simple summary of a very complex interaction between the compartment of main interest and secondary ones, detailed modelling of which can then be avoided.

[12] As a rule, e and b, but there are dissidents. In Chu's formulation [FC, PH], for instance, e and h are the basic entities.

1.2.1 Dynamics of free charges: the Vlasov–Maxwell model

A concrete example will illustrate this point. Let's discuss the problem of a population of charged particles moving in an electromagnetic field which they significantly contribute to produce. Coupled problems of this kind occur in astrophysics, in plasma physics, in the study of electronic tubes, and so forth. To analyze such a physical system, we may consider it as made of two compartments (Fig. 1.3): the electromagnetic one (EM), and the "charge motion" compartment (CM), which both require a kinematical description, and influence each other's dynamics, in a circular way.

Let's enter this circle at, for instance, CM. A common way to describe its kinematics is to treat charge carriers as a fluid, characterized by its charge density "in configuration space", a function $\tilde{q}(t, x, v)$ of time, position, and (vector-valued) velocity. The actual charge density and current density are then obtained by summing up with respect to v:

$$(8) \qquad q(t, x) = \int \tilde{q}(t, x, v)\, dv, \qquad j(t, x) = \int \tilde{q}(t, x, v)\, v\, dv,$$

where dv is the volume element in the three-dimensional space of velocities. CM thus influences EM by providing a source $\{q, j\}$ for it. (Later we'll see that q is redundant, (6) being satisfied.)

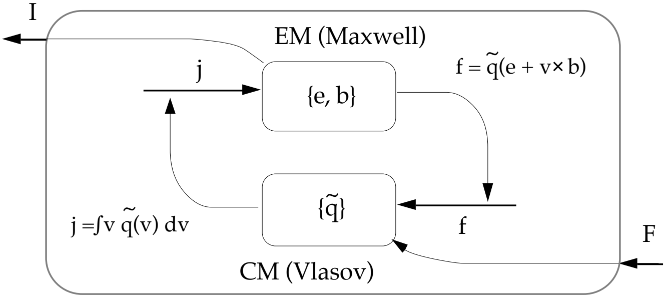


FIGURE 1.3. A typical example of coupling between compartments: the "Vlasov–Maxwell" model. "Vlasov" rules the behavior of a fluid of non-interacting free charges (a "collisionless plasma", for instance). "Maxwell" governs the electromagnetic field. Each compartment influences the other's behavior. External influences (such as, in the present case, the forces F, of non-electromagnetic origin) will in general intervene, and such "input" can then be seen as cause for the coupled system's evolution. Symmetrically, a global parameter (here some macroscopic intensity I) can be designated as "output", and the whole system (here, with its I—F dynamics) can become one compartment in some higher-level description.

The influence of EM on CM is via "Lorentz force". Recall that the force exerted by the field on a point charge Q passing at point x at time t with the (vector-valued) speed v is Q times the vector $e(t, x) + v \times b(t, x)$. (The part independent of celerity, that is $Q\, e(t, x)$, is "Coulomb force".) Here we deal with a continuum of charge carriers, so let us introduce, and denote by $\tilde{f}$, the *density* of force in configuration space: $\tilde{f}(t, x, v)\, dx\, dv$ is thus the force exerted on the packet of charges which are in volume dx around x, and whose speeds are contained in the volume dv of velocity space around v, all that at time t. So we have, in condensed notation,[13]

$$(9) \qquad \tilde{f} = \tilde{q}\,(e + v \times b).$$

These forces do work: We note for further use that the power density π thus communicated from EM to CM is what is obtained by integrating with respect to v:

$$(10) \qquad \pi(t, x) = \int \tilde{q}\,(t, x, v)\,(e(t, x) + v \times b(t, x)) \cdot v\, dv$$
$$= \int \tilde{q}\,(t, x, v)\, e(t, x) \cdot v\ \ dv = j(t, x) \cdot e(t, x),$$

after (8).

The circle around Fig. 1.3 will be closed once we know about the dynamics of CM. Let's suppose (which is a *huge,* but often acceptable over-simplification) that particles do not exchange momentum by collision or other non-electromagnetic interaction. One may as well suppose, in that case, that there is a single species of charge carriers (only electrons, for instance), since otherwise their effects will just add, in all respects, and can be computed separately. Let us call Q_c (specific charge of carrier c) the charge-to-mass ratio[14] of these particles. Charge conservation (or equivalently, mass conservation) then implies the *Vlasov equation,*

$$(11) \qquad \partial_t \tilde{q} + v \cdot \nabla_x \tilde{q} + Q_c\,(e + v \times b) \cdot \nabla_v \tilde{q} = 0,$$

where ∇_x and ∇_v denote partial gradients with respect to position and speed. Exercises 1.2 to 1.5, at the end of this chapter, suggest a road to this result. Exercise 1.6 will then invite you to prove that (6) holds for q and j as given by (8), when (11) is satisfied.

[13] The notation makes sense if v in (9) is understood as a vector *field,* the value of which is v at point $\{x, v\}$ of configuration space. Then $v \times b$ is a field of the same kind.

[14] For electrons, therefore, $Q_e = -\,1.602 \times 10^{-19}/9.109 \times 10^{-31}$ C/kg.

So this is a typical example of a coupled system: Given j, and assuming p = 0 and m = 0, since all charges are accounted for by j, system (1–4) determines e and b, hence the forces by (9), to which one adds other known causative forces, symbolized by F in Fig. 1.3, hence the movement of charge carriers (more generally, of charged matter), hence j again, which must be the same we started from. From a mathematical viewpoint, this is a "fixed point condition", which translates into an equation in terms of j, which will in general have a unique solution. One may then get the field by solving (1–4) with the j just found as source.[15] This can serve as a model for other multi-compartment situations: In general, the coupled problem may be proven, by a similar reasoning, to be well-posed,[16] though overwhelmingly difficult to solve.

With this, we may now elaborate on the notion of constitutive laws as summaries of interactions, or more bluntly, proxies which can take the place of secondary compartments in a modelling. For instance, in Fig. 1.3, it would be nice to have an explicit dependence of j on e and b, allowing us to bypass consideration of the "charge motion" compartment. A *constitutive law* is such a direct, more or less complex, dependence (of course, with limited and conditional validity).

1.2.2 Dynamics of conduction charges: Ohm's law

And indeed there are cases in which the "dynamics" part of the problem is especially simple to solve, at least approximately and accurately enough *as far as the main compartment is concerned*. Two such cases are especially important: conductors and generators.

Conductors (metals, etc.) are those bodies where exists a population of electric charges which are not bound to atoms, but still tightly interact with matter. Stirred by the field, the carriers accelerate for a while, but soon are stopped by collision, and the energy and momentum they acquired via Lorentz force are then transferred to the supporting matter, hence heating and also, possibly, movement of the conductor. Carriers move

[15] This is how the coupled problem is showed to be "well-posed" (see next note), not the way it is solved numerically. The favored technique for that is the "particle-in-cell" method used in "particle pusher" codes [BL, HE], which simulates the electron cloud with a finite family of particles and alternates between determining the motion of charge in a constant known field for one time-step, and updating the field values.

[16] *Well-posed* has a technical meaning: It refers to a problem of which one can prove it has a solution and a unique one, with, moreover, continuity of this solution with respect to the data. (The notion is due to Hadamard [Hd].)

anyway by thermal agitation, and at speeds of much higher magnitude than the additional velocity gained from electromagnetic action. But whereas thermal speeds cancel on the average at macroscopic scales, such is not the case of the motions due to Lorentz force. Their nonzero average is the so-called *drift velocity* (see, e.g., [Fe], II, Sections 32–36, or [We]). This slow[17] but collective motion, which can easily be detected and measured [Kl], results in a macroscopic current density.

This picture of the phenomenon is relatively complex, and one can simplify it as follows: Imagine the carrier population as a fluid, moving at the speed at which Lorentz force is balanced by all "friction-like" forces which tend to slow it down. Friction forces are in general, and are here found to be, *proportional* to the drift velocity. Since the current density due to a particular kind of carriers (ions, electrons, "holes" . . .) is proportional to their speed, one may conclude to a proportionality between e and the current density:

$$(12)\qquad j = \sigma\, e,$$

where σ, the *conductivity*,[18] depends on the material. This is *Ohm's law*.[19] One has $\sigma \geq 0$, and $\sigma = 0$ in *insulators* (dry air, vacuum, etc.).

The law itself is subject to experimental verification and holds with excellent accuracy in many cases, but the explanation behind it was, let's be candid, a *myth*: an explanation of rational appearance, relevant and reasonably consistent, but which openly glosses over fine points of physics, and whose main merit is to get rapidly to the point. (Indeed, the consistency of the foregoing explanation, in spite of its relevance to Ohm's law, can be challenged: cf. Exer. 1.11.)

As for *generators*, they are by definition these regions of space where the current density (then denoted by j^g, g for "given") can be considered as imposed, independently of the local electromagnetic field, and where, therefore, Ohm's law (12) doesn't apply. It is then convenient to set $\sigma = 0$

[17] In Cu, about 0.6 mm/s for 10 A/mm^2. The direction of the drift with respect to the field tells about the sign of the carriers, which are most often electrons, but can also be "holes" [Kl].

[18] Conductivity is measured in siemens per meter. (The siemens, or mho, Ω^{-1}, is the unit of conductance, and the dimension of σ is $(\Omega\ \mathrm{m})^{-1}$.) Fe : 5 to 10×10^6, Al : 36×10^6, Cu : 58 $\times 10^6$. Living tissues: ~ 0.1.

[19] This, in the case of *nonmoving* conductors. The v in (9) is sum of the speed of the free charge with respect to the conductor and of the latter's own speed, V. In case $V \neq 0$, one will have $j = \sigma\,(e + V \times b)$ instead of (12). Problems involving moving conductors will not be addressed in this book (with the advantage of always working within a unique reference frame).

in such regions[20] and to write a generalized Ohm's law, valid for generators, conductors, and insulators alike (and thus, most often, uniformly valid in all space):

(13) $$\mathrm{j} = \sigma\, \mathrm{e} + \mathrm{j}^{\mathrm{g}}.$$

It all goes then as if the charge dynamics problem had been solved in advance, the result being given by (13). One often calls *passive* conductors those ruled by (12), generators being then dubbed *active*.

One may then append (13) to (1–4), with $\mathrm{p} = 0$ and $\mathrm{m} = 0$. The system of equations thus obtained (or "Maxwell's model with linear conductors") embodies the theory of *nonmoving* (cf. Note 19) active and passive conductors which are neither polarizable nor magnetizable (cf. next section). It deals with a two-compartment system, EM and CM again, but the theory we have accepted for the latter is so simple, being all contained in (13), that one may easily overlook the coupled nature of the whole system. (One should not.)

1.2.3 Dynamics of bound charges: dielectric polarization

Now, another case of two-compartment system for which the same approach leads to a specific constitutive law. It deals with polarizable materials, in which charges are too strongly bound to separate from their original sites, but loose enough to be pulled a little off their equilibrium position by Coulomb forces, when the material is subject to a macroscopic electric field. This *polarization* phenomenon is important for some materials, dubbed *dielectric*. The simple reasoning (or myth . . .) that follows shows how to account for it, by a simple relation between e and the p of (3).

Despite its electrical neutrality at a macroscopic scale, matter contains positive and negative charges (+ and – for brevity) which we may imagine as being attached by pairs at certain material sites. Suppose the density of + charges is equal to q_+, a function of position x. In the absence of any macroscopic electric field, the density of – charges must equal $-\mathrm{q}_+$, by electric neutrality. Now, a field e being applied, let's represent by a vector field u the separation of charged pairs that results, as follows: A + charge [resp. a – charge] that was at point x is now at $\mathrm{x} + \mathrm{u}(\mathrm{x})/2$ [resp. at $\mathrm{x} - \mathrm{u}(\mathrm{x})/2$]. To easily compute the new charge density q_p due to this

[20] This amounts to neglecting the internal resistance of the generator. In some modellings, having a nonzero σ there can be useful. Note this wouldn't change the form of (13).

change in localization, let us treat it as a mathematical *distribution,*[21] that is, as the mapping $\psi \to \int q_p\, \psi$, where ψ denotes a so-called "test function". Expanding ψ to first order and integrating by parts, we have

$$\int q_p\, \psi = \int q_+(x)\ [\psi(x + u(x)/2) - \psi(x - u(x)/2)]\ dx$$

$$\sim \int q_+\ u \cdot \text{grad}\, \psi \equiv \int - \text{div}(q_+\ u)\ \psi,$$

hence $q_p = -\ \text{div}\ p$, where $p = q_+\ u$. This field p, soon to be identified with the one in (3), is the polarization of the dielectric.

Exercise 1.7. Try to do the same computation "the other way around", by starting from $\int q_p\, \psi = \int q_+(x - u(x)/2)\ \psi(x)$, etc. Why does it go wrong this way?

The macroscopic manifestation of this local charge splitting is thus the appearance of a distribution of charges in what was initially an electrically neutral medium. Moreover, if the polarization changes with time, the motion of charges $+$ and $-$ in opposite directions amounts to a current density $j_p = \partial_t(qu) = \partial_t\ p$. (Note that $\partial_t\ q_p + \text{div}\ j_p = 0$, as it should be.)

We might treat this current density on the same footing as j, and replace the polarized matter by vacuum plus polarization current. Then $d = \varepsilon_0 e$, and Eqs. (1) and (3) would combine to give

$$(14) \qquad -\ \partial_t(\varepsilon_0\, e) + \text{rot}\ h = j + \partial_t\, p.$$

Instead, we use our option (cf. Exer. 1.1) to charge $\partial_t\, p$ on the account of Eq. (3), by setting $d = \varepsilon_0\, e + p$, hence $-\ \partial_t d + \text{rot}\ h = -\ \partial_t(\varepsilon_0\, e + p) + \text{rot}\ h = j + \partial_t\, p - \partial_t\, p \equiv j$, leaving e unchanged. This separates macroscopic currents j,

[21] In the theory of distributions [Sc], functions are not defined by their values at points of their domain of definition, but via their effect on other functions, called *test functions.* So, typically, a function f over some domain D is known if one is given all integrals $\int_D f\, \psi$, for all smooth ψ supported in D. It is thus allowable to identify f with the linear mapping $\psi \to \int_D f\, \psi$. (The arrowed notation for maps is discussed in A.1.9.) This has the advantage of making functions appear as special cases of such linear *TEST_FUNCTION → REAL_NUMBER* mappings, hence a useful generalization of the notion of function: One calls such maps *distributions*, provided they satisfy some reasonable continuity requirements. For instance, the map $\psi \to \psi(a)$, where a is some point inside D, is a distribution ("Dirac's mass" at point a, denoted δ_a). The generalization is genuine, since there is no function f_a such that $\psi(a) = \int_D f_a \psi$ for all ψ. It is useful, because some theories, such as Fourier transformation, work much better in this framework. The Fourier tranform of the constant 1, for instance, is not defined as a function, but makes perfect sense as a distribution: It's $(2\pi)^{d/2}$ times a Dirac mass at the origin, i.e. $(2\pi)^{d/2}\, \delta_0$, in spatial dimension d.

which continue to appear on the right-hand side of the expression of Faraday's law, and microscopic (polarization) currents $j_P = \partial_t p$, now hidden from view in the constitutive law. Notice that $\mathrm{div}\ d = q$, where q is the macroscopic charge, and $\mathrm{div}(\varepsilon_0 e) = q + q_p$.

All this shuffling, however, leaves the polarization current to be determined. The "(coupled) problem of bound charges" would consist in simultaneously computing p and the electromagnetic field, while taking into account specific laws about the way charges are anchored to material sites. Just as above about conduction, one makes do with a simple—and empirically well confirmed—solution to this problem, which consists in pretending (by invoking a "myth", again) that p and e are proportional: $p = \chi e$, as would be the case if charges were elastically bound, with a restoring force proportional to e, and without any inertia.[22] Now, let us set $\varepsilon = \varepsilon_0 + \chi$. Then, Eqs. (1) and (3) become

(1′) $$-\partial_t d + \mathrm{rot}\, h = j,$$ (3') $$d = \varepsilon e.$$

The coefficient ε in (3') (called *permittivity*, or *dielectric constant* of the medium[23]) thus appears as the simple summary of a complex, but microscopic-scale interaction, which one doesn't wish to know about at the macroscopic scale of interest.

Another, simpler solution of the coupled problem obtains when one may consider the field p, then called *permanent polarization*, as independent of e. The corresponding behavior law, $d = \varepsilon_0 e + p$ with fixed p, is well obeyed by a class of media called *electrets*. Of course one may superpose the two behaviors (one part of the polarization being permanent, the other one proportional to e), whence the law $d = \varepsilon e + p$ instead of (3), with a fixed p.

1.2.4 Magnetization

It is tempting to follow up with a similar presentation of magnetization, where a proportionality between m and h would be made plausible by a simple myth about the interaction of magnetic moments (due to the electrons' spins, mainly) with the magnetic field. This would be a little artificial, however, because too remote from the real physics of magnetism

[22] The latter hypothesis will be reconsidered in the case of high frequencies. Note that χ can be a tensor, to account for anisotropy.

[23] Terminology wavers here. Many authors call "permittivity" the *ratio* between ε and ε_0, and speak of "dielectric constant" when it comes to ε, or even to its real part in the case when ε is complex (see below). Note that ε may be a tensor.

(cf., e.g., [OZ]), and the point is already made anyway: Constitutive laws substitute for a detailed analysis of the interaction, when such analysis is either impossible or unproductive. So let us just review typical constitutive laws about magnetization.

Apart from *amagnetic* materials ($m = 0$), a simple case is that of *paramagnetic* or *diamagnetic* materials, characterized by the linear law $m = \chi h$ (whence $b = \mu h$, with $\mu = (1 + \chi)\mu_0$), where the *magnetic susceptibility* χ is of positive or negative sign, respectively. It can be a tensor, in the case of anisotropic materials. For most bodies, χ is too small to matter in numerical simulations, the accuracy of which rarely exceeds 1 % ($\chi \sim 10^{-4}$ for Al or Cu).

Ferromagnetic metals (Fe, Co, Ni) and their alloys are the exception, with susceptibilities up to 10^5, but also with a nonlinear (and hysteretic[24]) behavior beyond some threshold. In practice, one often accepts the linear law $b = \mu h$ as valid as far as the modulus of b does not exceed 1 tesla.[25]

For *permanent magnets* [La, Li] a convenient law is $m = \chi h + h_m$, where h_m is a vector field independent of h and of time, supported by the magnet (that is, zero-valued outside it), with χ roughly independent of h, too, and on the order of 1 to 4, in general [La]. This law's validity, however, is limited to the normal working conditions of magnets, that is, for h and b of opposite signs, and not too large. The characteristic $b = \mu h + \mu_0 h_m$ is then called the "first order reversal curve".

1.2.5 Summing up: Linear materials

Hysteresis, and nonlinearity in general, are beyond our scope, and we shall restrict to the "Maxwell model of memoryless linear materials with Ohm's law":

[24] *Hysteresis* occurs when the value of b at time t depends not only on $h(t)$, but on past values. *Linearity* does not preclude hysteresis, for it just means that if two field histories are physically possible, their superposition is possible too. This does not forbid behavior laws "with memory", but only allows "convolution laws" of the form $b(t) = \int^t \mathcal{M}(t - s)\, h(s)\, ds$. As we shall see in Section 1.4, this amounts to $\mathrm{B} = \mu \mathrm{H}$, in Fourier space, with a complex and frequency-dependent μ.

[25] The unit for b is the tesla (T), or weber (Wb) per square meter. (One tesla is 10 000 gauss, the cgs unit still in use, alas.) The field h is measured in ampères per meter (A/m). An ordinary magnet creates an induction on the order of .1 to 1 T. The Earth field is about 0.4×10^{-4} tesla.

(15) $- \partial_t d + \mathrm{rot}\, h = j \equiv j^g + \sigma e,$ (16) $\partial_t b + \mathrm{rot}\, e = 0,$

(17) $d = \varepsilon e,$ (18) $b = \mu h\,,$

plus occasionally some constant term on the right of (17) or (18), in order to model electrets or permanent magnets. In most modellings, these equations correctly describe what we shall from now on call "the electromagnetic compartment" (and still denote by EM, although it has been slightly enlarged). But let's not forget the complexity of field-matter interactions that are thus hidden beyond a neat façade, and the relative arbitrariness with which compartment boundaries have been moved in order to incorporate microscopic interactions in (15–18), leaving only macroscopic interactions with other compartments to describe. We now turn to this.

1.3 MACROSCOPIC INTERACTIONS

Most engineering applications have to do with power conversion. In this respect, what we have established in (10) has general validity:

Proposition 1.1. *The power density yielded by the electromagnetic compartment of a system to other compartments is given by* $\pi(t, x) = j(t, x) \cdot e(t, x)$, that is, as an equality between scalar fields,

(19) $\pi = j \cdot e,$

at all times. (Be aware that j is the total current, $j = \sigma e + j^g$.)

In the case of a passive and immobile conductor, $j \cdot e = \sigma\, |e|^2$, so this is Joule loss, and therefore, thermal power. In the case of generators, $-\pi(t, x)$ is the density of power needed to push charges up the electric field (and thus given to the EM compartment). In the case of moving conductors, $j \cdot e$ is in part Joule heating, and for the other part mechanical work. In all cases, the total yielded power is thus[26] $\Pi = \int_{E_3} \pi(x)\, dx$, that is, $\Pi = \int_{E_3} j \cdot e$. This is the bottom-line figure in the inter-compartment trade balance.

[26] See Appendix A, Subsections A.2.4 and A.2.5, for E_3. This symbol stands for "oriented three-dimensional Euclidean affine space": ordinary space, equipped with a notion of orientation (i.e., a way to distinguish direct and skew reference frames, cf. A.2.5), and with the dot-product here denoted by " · ", which gives sense to the notions of distance, area, volume, etc.

1.3.1 Energy balance

Compartmentalization, however, is not limited to physically distinct subsystems, and may concern distinct regions of space too. In this respect, energetical exchanges through spatial boundaries are important. Let thus a closed surface S separate a domain D from the rest of space. Take the scalar product of both sides of (1) and (2) by $-\mathrm{e}$ and h, respectively, add, and integrate over D:

$$\int_D (\mathrm{h} \cdot \partial_t \mathrm{b} + \mathrm{e} \cdot \partial_t \mathrm{d}) + \int_D (\mathrm{h} \cdot \mathrm{rot}\, \mathrm{e} - \mathrm{e} \cdot \mathrm{rot}\, \mathrm{h}) = -\int_D \mathrm{j} \cdot \mathrm{e}.$$

The result is then transformed by the following integration by parts formula, to which we shall return in the next chapter:

$$(20) \qquad \int_D \mathrm{h} \cdot \mathrm{rot}\, \mathrm{e} = \int_D \mathrm{e} \cdot \mathrm{rot}\, \mathrm{h} - \int_D (\mathrm{n} \times \mathrm{h}) \cdot \mathrm{e},$$

and by setting

$$\mathrm{W}_D(t) = \frac{1}{2} \int_D (\mu \; |\mathrm{h}(t)|^2 + \varepsilon \; |\mathrm{e}(t)|^2),$$

hence

$$(21) \qquad \partial_t \mathrm{W}_D + \int_S \mathrm{n} \cdot (\mathrm{e} \times \mathrm{h}) = -\int_D \mathrm{j} \cdot \mathrm{e}.$$

A special case of this[27] is when D is all space:

$$(22) \qquad \partial_t \mathrm{W} = -\Pi,$$

where $\mathrm{W}(t) = \mathrm{W}_{E_3}(t)$, a quantity that may thus legitimately be called *electromagnetic energy:* indeed, (22) points to it as being the energy stored in the electromagnetic compartment[28] of the system.

So if we turn to (21), its interpretation in similar terms is immediate: The "subcompartment EM-D" cedes the power $\int_D \mathrm{j} \cdot \mathrm{e}$ to D-based subsidiaries of all non-EM compartments, and exports $\int_S \mathrm{n} \cdot (\mathrm{e} \times \mathrm{h})$ to other regions of the EM compartment, which themselves, of course, may trade with non-EM entities in their own domain. The vector field $\mathrm{e} \times \mathrm{h}$, which records these trans-boundary exchanges, is *Poynting's vector.*[29]

[27] One has $\int_{E_3} \mathrm{h} \cdot \mathrm{rot}\, \mathrm{e} = \int_{E_3} \mathrm{e} \cdot \mathrm{rot}\, \mathrm{h}$ (no "surface term at infinity") provided e and h both belong to the functional space $\mathbb{L}^2_{\mathrm{rot}}(E_3)$, to be studied in more detail in Chapter 5, where this assertion will be proved. Its physical content is just that h and e *decrease fast enough* at infinity; hence the absence of the boundary term when S recedes to infinity, and this we can accept without qualms for the moment.

[28] In the extended sense in which we now understand "electromagnetic" compartment. If $\varepsilon \neq \varepsilon_0$, for instance, part of this energy is in dipole vibration.

Note that, by applying Ostrogradskii's formula to (21),

$$(23) \qquad \partial_t w + j \cdot e + \operatorname{div}(e \times h) = 0,$$

where w is the scalar field $x \to \frac{1}{2}(\mu \,|h(x)|^2 + \varepsilon \,|e(x)|^2)$. Just as (6) expressed "local" charge conservation, (23) is the local expression of energy conservation, the integrated or "global" form of which is (21). It is tempting to call w the (electromagnetic) *energy density*, and we shall do that. See, however, Remark 1.1 below.

As an illustration, let us mention thermal exchanges (induction heating, direct heating, microwave heating, welding . . .). The "thermal compartment" (TM) of a system is governed by the heat equation, in all its guises, the best known of which, valid when most thermal exchanges are by conduction and diffusion, is

$$(24) \qquad \partial_t(c\,\theta) - \operatorname{div}(\kappa \operatorname{grad} \theta) = \pi,$$

where θ (a scalar field) stands for the temperature, c for the volumic heat, κ for the thermal conductivity, and π for the injected power density. When this power is Joule loss, one has $\pi(x) = \sigma(x)\ |e(x)|^2$. Since σ, as well as coefficients ε and μ for that matter, may depend on temperature, studying electrothermal interactions amounts to studying the coupled system (15–18)(24).

Most often, there is a natural division into subcompartments. In induction heating, for instance, if the workpiece (the passive conductor) occupies domain D, TM will be restricted to D, with of course adequate boundary conditions for (24) on its boundary S. A partition of EM into D and $E_3 - D$ is then the obvious thing to do, especially at low frequencies, where the equations in the non-conducting region take a simple form, as we shall see in Chapter 8. The flux of $e \times h$ through S is then the heating power, and thus of particular significance.

Remark 1.1. Just as W_{E_3} is the energy of EM, we may *define* W_D as "the energy of EM-D". But to say that this energy *is inside* D, which amounts to saying that $w(x)\,dx$ is the energy "effectively present" in volume dx, goes much further, since it asserts that energy is *localized*, as a substance can be, and this is controversial. Some authors, comparing this with localizing the beauty of a painting at specific parts of it, protest they "do

[29] An instance of this appalling habit many physicists have to call "vector" what is actually a vector *field*. Such sloppiness about the *type* (cf. A.1.2) of the entities one deals with, harmless as it may be in the present case, should not be condoned. Vector fields are objects of type $A_3 \to V_3$, in the notation of A.2.2 and A.2.4, vectors being elements of V_3.

not believe that 'Where?' is a fair or sensible question to ask concerning energy. Energy is a function of configuration, just as (. . .) beauty (. . .)". (Cf. [MW], pp. 266–267.) The problem is inherent in field theory, and not special to electromagnetism [KB]. ◊

Remark 1.2. The Poynting vector field also is a bone of contention. There are totally *static* situations in which the energy flux $e \times h$ is not zero (**Exercise 1.8:** find one). It all goes then as if energy was perpetually flowing in circles. The idea may seem unattractive, and alternatives have been proposed, based on the fact that the flux of a curl through a closed surface is always zero, so one may add to $e \times h$ the curl of any vector field u one fancies, without changing *any* power flux, whatever the domain of interest. (This is clear on the local expression (23), since $\text{div}(\text{rot } u) = 0$.) Slepian [Sl] thus could list no fewer than eight plausible expressions for the energy-flow vector, including Poynting's. The debate rebounds regularly [Ly, Lo, He]. There is an old argument ([Bi], discussed in [Ro]) to the effect that if $\text{rot } u$ is to be a function of e and b only, then $\text{rot } u$ is a constant, so Poynting's vector is the natural choice (the "gauge invariant" one) among these alternatives. But this leaves some unconvinced [BS]. ◊

1.3.2 Momentum balance

Even more controversial[30] is the question of momentum: One century ago, Abraham and Minkowski disagreed about the correct expression of the linear momentum of the electromagnetic field [Cs]. The question is still debated, and what follows will not resolve it. But having discussed energy, we cannot elude momentum, since they are two observer-dependent aspects of one and the same objective entity (the four-dimensional energy–momentum, or "momenergy" [TW]). Moreover, the expression of forces exerted by EM on conductors and polarized or magnetized matter derives from momentum conservation, and forces are an often-desired output in computations, even those restricted to immobile bodies, to which we limit consideration here.

First let us introduce a notation (local to this section): if v is a vector field and φ a scalar field, $\nabla_v \varphi$ may conveniently denote the scalar field $v \cdot \text{grad } \varphi$, so that $\nabla_v \varphi(x)$ is "the derivative of φ in the direction of v", at point x. Now if u is another vector field, one can form $\nabla_v u^i$ for its three Cartesian coordinates u^i, hence the three scalar components of a vector, which will be denoted $\nabla_v u$. Next move, please, is yours:

[30] See R.H. Romer: "Question #26: Electromagnetic field momentum", **Am. J. Phys., 63,** 9 (1995), pp. 777–779, and the answers provided in **Am. J. Phys., 64**, 1 (1996), pp. 15–16.

Exercise 1.9. Show that $\int_D \nabla_v u = -\int_D u \ \mathrm{div}\, v + \int_S n \,.\, v \ u$ (with D, S, and n as usual, cf. Fig. 1.2).

Exercise 1.10. Show that, in case $v = \alpha\, u$, where α is a scalar field,

$$(25) \qquad \int_D v \times \mathrm{rot}\, u = -\int_D \nabla_v u + \tfrac{1}{2} \int_S u \cdot v \ n - \tfrac{1}{2} \int_D |u|^2 \nabla\alpha.$$

We can then do the following calculation. Starting from (15) and (16), take the cross product (from the left) of both sides by b and d, respectively, add, and integrate. This gives

$$\partial_t \int_D d \times b + \int_D b \times \mathrm{rot}\, h + \int_D d \times \mathrm{rot}\, e = -\int_D j \times b.$$

Now apply (25) and Exer. 1.9 to d and e, with $\alpha = \varepsilon$:

$$\int_D d \times \mathrm{rot}\, e = -\int_S n \cdot d \ e + \int_D e \ \mathrm{div}\, d + \tfrac{1}{2} \int_S d \cdot e \ n - \tfrac{1}{2} \int_D |e|^2 \nabla\varepsilon$$

(recall that div d = q), then to b and h, quite similarly, and gather the results to get

$$(26) \qquad \partial_t \int_D d \times b + \int_S (\tfrac{1}{2}\, b \cdot h \ n - n \cdot b \ h) + \int_S (\tfrac{1}{2}\, d \cdot e \ n - n \cdot d \ e) = \\ -\int_D [j \times b + q\, e - \tfrac{1}{2} |h|^2 \nabla\mu - \tfrac{1}{2} |e|^2 \nabla\varepsilon].$$

This (to be compared with (21), which had the same structure) is the momentum balance: $\int_D d \times b$ is the momentum "of" (same caveats as above) subcompartment EM-D, its flux is governed by the so-called "Maxwell stress tensor", here[31] $\mathrm{M} = \tfrac{1}{2}\, b \cdot h - b \otimes h + \tfrac{1}{2}\, d \cdot e - d \otimes e$, and the right-hand side of (26) is, up to sign, the resultant of body forces.[32] Note the unexpected gradients, which should be interpreted in the sense of distributions when ε or μ are discontinuous.

The local version of (26), quite similar to (23), is

$$(27) \qquad \partial_t [d \times b] + j \times b + q\, e - \tfrac{1}{2} |h|^2 \nabla\mu - \tfrac{1}{2} |e|^2 \nabla\varepsilon + \mathrm{div}\, \mathrm{M} = 0.$$

But the global version (26) is more popular: A standard way to obtain the total force on some object (in a time-independent situation) is to compute the flux of M through some boundary S enclosing it.

[31] This tensor product symbol $\otimes$ will not be used again. Owing to our sign conventions, M is actually *minus* the Maxwell tensor of tradition.

[32] No proof has been offered here as to the validity of these interpretations. But if one accepts the expression of body-force density (which is standard, cf. e.g., [Rb]), the rest follows. See [Bo] for a *direct* derivation of the body force expression.

When studying the dynamics of moving conductors, one should take into account the momentum of the moving bodies *and* the momentum of the field in the expression of momentum conservation.[33] In an interaction between two solids, for instance, momentum lost by one of them may temporarily be stored in the field, before being restituted to the other body. Thus, action and reaction may seem not to balance, in apparent violation of Newton's third law [Ke]. See for instance [Co], [Ho], and the abundant literature on Feynman's "disk paradox", a situation in which a disk, initially at rest in a static field, can acquire angular momentum without any mechanical action, just because of a change in the electromagnetic environment [Lm].

Remark 1.3. So there are *static* configurations in which $\int \mathrm{d} \times \mathrm{b} \neq 0$: Surprising as this may appear, a static electromagnetic field *can* possess linear momentum. (Cf. R.H. Romer, **Am. J. Phys., 62,** 6 (1994), p. 489. See also [PP].) ◊

Remark 1.4. The cross product is an orientation-dependent operation: its very definition requires a rule for orienting ambient space. Yet we see it appear in expressions such as $\mathrm{e} \times \mathrm{h}$ or $\mathrm{d} \times \mathrm{b}$, which account for energy or momentum flux, physical quantities which obviously do *not* depend on orientation conventions. How come? It must be that some of the vector fields e, d, b, h themselves depend on orientation. No surprise in that: The *mathematical* entities by which the physical field is represented may depend on the structures of Euclidean space, whereas the objective phenomena do not. The question is further discussed in Section A.3 of Appendix A. ◊

1.4 DERIVED MODELS

Concrete problems in electromagnetism rarely require the solution of Maxwell equations in full generality, because of various simplifications due to the smallness of some terms. The displacement currents term $\varepsilon\, \partial_t \mathrm{e}$, for instance, is often negligible; hence an important submodel, *eddy-currents* theory, which we shall later study in its own right:

$$(28) \qquad \partial_t \mathrm{b} + \mathrm{rot}\, \mathrm{e} = 0, \quad \mathrm{rot}\, \mathrm{h} = \mathrm{j}, \quad \mathrm{j} = \sigma \mathrm{e} + \mathrm{j}^{\mathrm{g}},$$

[33] Many papers in which this commonsense rule is neglected get published, notwithstanding, in refereed Journals. It has been asserted, for example, that the operation of a railgun cannot be explained in terms of classical electrodynamics. See a refutation of this crankish claim in [AJ].

with in particular, in passive conductors (where one may eliminate e from (28) after division by σ), $\partial_t(\mu h) + \mathrm{rot}\,(\sigma^{-1}\,\mathrm{rot}\,h) = 0$.

Another frequent simplification is the passage to complex numbers representations. If the source current j^g is sinusoidal in time,[34] that is, of the form $j^g(t, x) = \mathrm{Re}[\mathrm{J}^g(x)\ \exp(i\omega\,t)]$, where J^g is a *complex-valued* vector field, and *if* all constitutive laws are linear, one may[35] look for the electromagnetic field in similar form, $h(x) = \mathrm{Re}[\mathrm{H}(x)\ \exp(i\omega t)]$, etc., the unknowns now being the complex fields H, E, etc., independent of time. Maxwell's model with Ohm's law (15–18) then assumes the following form:

$$(29) \qquad -\,i\omega\,\mathrm{D} + \mathrm{rot}\,\mathrm{H} = \mathrm{J}^g + \sigma\,\mathrm{E}, \quad i\omega\,\mathrm{B} + \mathrm{rot}\,\mathrm{E} = 0, \quad \mathrm{D} = \varepsilon\,\mathrm{E}, \quad \mathrm{B} = \mu\,\mathrm{H}.$$

It is convenient there to *redefine* ε by assigning to this symbol the complex value $\varepsilon + \sigma/(i\omega)$, which allows the incorporation of the term $\sigma\,\mathrm{E}$ into $i\omega\,\mathrm{D}$, whence the model

$$(29') \qquad -\,i\omega\,\mathrm{D} + \mathrm{rot}\,\mathrm{H} = \mathrm{J}^g, \quad i\omega\,\mathrm{B} + \mathrm{rot}\,\mathrm{E} = 0, \quad \mathrm{D} = \varepsilon\,\mathrm{E}, \quad \mathrm{B} = \mu\,\mathrm{H},$$

which is, with appropriate boundary conditions, the *microwave oven* problem. In (29'), ε is now complex, and one often writes it as $\varepsilon = \varepsilon' - i\varepsilon''$, where the real coefficients ε' and ε'', of same physical dimension as ε_0, are nonnegative. (They often depend on temperature, and are measured and tabulated for a large array of products, foodstuffs in particular. Cf. eg., [FS, St, Jo]. Figure 1.4 gives an idea of this dependence.)

Nothing forbids accepting complex μ's as well, and not only for the sake of symmetry. This really occurs with ferrites[36] [La, Li], and also in some modellings, a bit simplistic[37] perhaps, of hysteresis.

[34] One often says "harmonic", but be wary of this use, not always free of ambiguity.

[35] This procedure is valid, a priori, each time one is certain about the *uniqueness* of the solution of the problem "in the time domain", for if one finds a solution, by whatever method, it's bound to be the right one. But it's the *linearity* of constitutive laws (cf. Note 24) that makes the procedure effective. Moreover, linearity allows one to extend the method to non-periodic cases, thanks to Laplace transform (then one has p, complex-valued, in lieu of $i\omega$). The passage to complex numbers is *in principle* of no use in nonlinear cases (for instance, when iron or steel is present), and the notion of "equivalent (complex) permeability", often invoked in applications to induction heating, is not theoretically grounded. (Its possible empirical value is another question, to be considered in each particular instance.)

[36] One refers to *linear* behavior there, and this complex permeability is not of the same nature as the one of the previous note.

[37] Because of their essentially *linear* nature. Law $\mathrm{B} = (\mu' - i\mu'')\mathrm{H}$ amounts to $\mu''\,\partial_t h = \omega(\mu'\,h - b)$ in the time domain.

An even more drastic simplification obtains when one may consider the phenomena as independent of time (steady direct current at the terminals, or current with slow enough variations). Let us review these models, dubbed *stationary*, derived from Maxwell's model by assuming that all fields are independent of time.

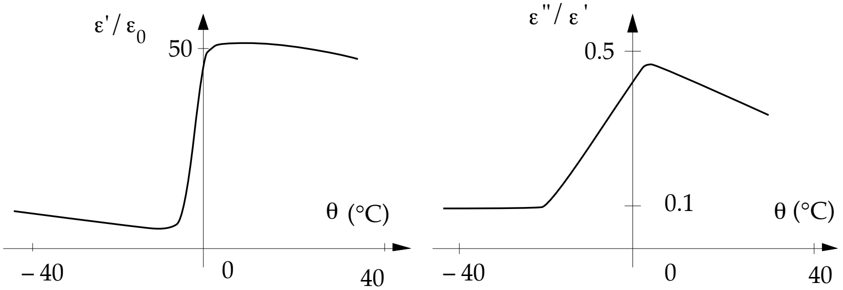


FIGURE 1.4. Typical curves for ε' and ε'' as functions of temperature, for a stuff with high water content. The ratio $\varepsilon''/\varepsilon'$, shown on the right, is often denoted by $\tan\delta$.

In this case, one has in particular $\partial_t b = 0$, and thus $\operatorname{rot} e = 0$. So, after (5) and (17),

(30) $\operatorname{rot} e = 0, \quad d = \varepsilon e, \quad \operatorname{div} d = q,$

and this is enough to determine e and d in all space, if the electric charge q is known: Setting $e = -\operatorname{grad}\psi$, where ψ is the *electric potential*, one has indeed $-\operatorname{div}(\varepsilon\operatorname{grad}\psi) = q$, a Poisson problem which is, as one knows, well posed. In the case where $\varepsilon = \varepsilon_0$ all over, the solution is given by

$$\psi(x) = \frac{1}{4\pi\varepsilon_0}\int_{E_3}\frac{q(y)}{|x-y|}\,dy,$$

as one will check (cf. Exers. 4.9 and 7.5) by differentiating under the summation sign in order to compute $\Delta\psi$. Model (30) is the core of linear *electrostatics*.

In a similar way, one has $\operatorname{rot} h = j$, after (1), whence, taking into account $\operatorname{div} b = 0$ and (18), the model of linear *magnetostatics*:

(31) $\operatorname{rot} h = j, \quad b = \mu h, \quad \operatorname{div} b = 0,$

and this determines b and h in all space when j is given. If $\mu = \mu_0$ all over, the solution is obtained in closed form by introducing the vector field

$$a(x) = \frac{\mu_0}{4\pi} \int_{E_3} \frac{j(y)}{|x - y|}\, dy,$$

called *magnetic vector potential,* and by setting $b = \text{rot } a$. (By differentiating inside the integral, one will find *Biot and Savart's formula,* which directly gives h in integral form:

$$h(x) = \frac{1}{4\pi} \int_{E_3} \frac{j(y) \times (x - y)}{|x - y|^3}\, dy.) \tag{32}$$

When, as in the case of ferromagnetic materials, constitutive laws more involved than $b = \mu\, h$ occur, problem (31) appears as an intermediate in calculations (one step in an iterative process, for instance), with then in general a position-dependent μ. An important variant is the magnetostatics problem for a given distribution of currents and magnets, the latter being modelled by $b = \mu h + \mu_0 h_m$ with known μ and (vector-valued) h_m.[38] Setting $h_m = 0$ in the air, one gets

$$\text{rot } h = j, \quad b = \mu h + \mu_0 h_m, \quad \text{div } b = 0.$$

An analogous situation may present itself in electrostatics: $d = \varepsilon e + p$, with p given, as we saw earlier.

Still under the hypothesis of stationarity, one has $\partial_t q = 0$, and thus $\text{div } j = 0$, after (6), hence

$$\text{rot } e = 0, \quad j = \sigma e, \quad \text{div } j = 0, \tag{33}$$

in passive conductors. This is the *conduction* or *electrokinetics model*. In contrast to the previous ones, it does not usually concern the whole space, and thus requires boundary conditions, at the air-conductor interfaces, in order to be properly posed.

The formal similarity between these static models is obvious, and we need examine only one in detail to master the others. We'll focus on

[38] A legitimate question, at this stage, would be, "How does one know h_m, for a given permanent magnet?". Giving a rigorous answer would require the knowledge of the conditions under which the material has been magnetized, as well as the details of its hysteretic response, and a feasible simulation method of this process. In practice, most often, a uniform magnetization field parallel to one of the edges of the magnets is a fair representation. However, as more and more complex magnetization patterns are created nowadays, the problem may arise to find h_m from measurements of b by a computation (solving an inverse problem).

magnetostatics in this book, with only a few indications about the other models in Chapters 8 and 9. This disproportion is to some extent mitigated by the paradigmatic character of the magnetostatics model. As pointed out in the Preface, the difficulties encountered in computational electromagnetism in the 1970s, when one tried to extend then well-established finite element or boundary integral 2D methods to three-dimensional situations, appear in retrospect to be due not to the increased dimensionality per se, but to the essential difference between the "curl–curl" operator and the "div–grad" operator to which it reduces in two dimensions, and fortunately, all essential points about the curl–curl operator can be understood in the simple, limited, and well-defined framework of linear magnetostatics.

EXERCISES

The text for Exer. 1.1 is on p. 5.

Exercise 1.2. Let X be an affine space and V the associated vector space, $f : \mathbb{R} \times X \times V \rightarrow \mathbb{R}$ a *repartition function,* interpreted as the time-dependent density of some fluid in configuration space X × V. Let $\gamma(t, x)$ be the acceleration imparted at time t to particles passing at x at this instant, by some given external force field. Show that

$$(34) \qquad \partial_t f + v \cdot \nabla_x f + \gamma \cdot \nabla_v f = 0$$

expresses *mass conservation* of this fluid. What if γ, instead of being a data, depended on velocity?

Exercise 1.3. What is the divergence of the field $x \rightarrow a \times x$, of type $E_3 \rightarrow V_3$, where a is a fixed vector? Its curl? Same questions for $x \rightarrow x$. (Cf. Subsection A.1.2 for the notion of type, and the notational convention, already evoked in Note 13, and Note 29.)

Exercise 1.4. In the context of Exer. 1.2, what is the divergence of the field $v \rightarrow e + v \times b$?

Exercise 1.5. Establish Vlasov's equation (11).

Exercise 1.6. Prove, using (11), that charge and current as given by (8) do satisfy the charge conservation relation (6).

See p. 12 for Exer. 1.7, pp. 18 and 19 for Exers. 1.8 to 1.10.

Exercise 1.11. Show that, in a region of a conductor where σ is not constant (due to variations in temperature, or in the composition of an alloy, etc.), $q = \text{div } d$ may not be zero, and that this can happen in stationary situations (continuous current). Thus, there can exist a permanent charge imbalance at some places in the conductor. But Lorentz force acts on this charge packet. *Why doesn't it move?*

HINTS

1.1. Don't worry about differentiability issues: Assume all fields are smooth.

1.2. Imitate the classical computation about the convective derivative in fluid dynamics (which is very close to our treatment of charge conservation, p. 4).

1.3. For $x \to a \times x$, divergence: 0, curl: 2a. For $x \to x$, curl-free, the divergence is the constant scalar field $x \to 3$.

1.4. Mind the trap. Contrary to e and b, this field does *not* live in 3D Euclidean space! The *type* of the map will tell you unambiguously what "divergence" means.

1.5. Apply Exer. 1.2, acceleration being $Q_e(e + v \times b)$. By Exer. 1.4, there is no extra term.

1.6. Ostrogradskii on $\{t, x\} \times V$. Ensure suitable boundary conditions by assuming, for instance, an upper bound for the velocity of charges.

1.7. A careless attempt, like[39] $q_+(x - u(x)/2)\ \psi(x) \overset{*}{=} -\frac{1}{2}\ \nabla q_+ \cdot u$, would seem to lead to $-\int q_+ \text{ div}(\psi u)$, and hence to a different result than above if $\text{div } u \neq 0$. This is the key: Why does this divergence matter?

1.8. A bar magnet between the plates of a condenser.

1.9. This is an extension of the integration by parts formula (2.9)[40] of the next chapter, $\int_D v \cdot \text{grad } u^i = -\int_D u^i \text{ div } v + \int_S n \,.\, v\ u^i$, $i = 1, 2, 3$.

[39] The star in $\overset{*}{=}$ serves as a warning that the assertion should not be believed blindly.

[40] As a rule, we'll refer to "Eq. (n)" inside a chapter, and to "Eq. (X.n)" for the equation labelled (n) in Chapter X.

1.10. The simplest way is probably to work in Cartesian coordinates, starting from

$$(\int_D v \times \mathrm{rot}\, u)^i = \sum_j \int_D v^j\, (\partial_i u^j - \partial_j u^i), \quad i = 1, 2, 3.$$

Then the last term is $-\int_D \nabla_v u$, and $\int_D v^j\, \partial_i u^j = \int_D \alpha\, u^j\, \partial_i u^j = \frac{1}{2} \int_D \alpha\, \partial_i\, |u^j|^2 = \frac{1}{2} \int_D \partial_i(\alpha\, |u^j|^2) - \frac{1}{2} \int_D |u^j|^2\, \partial_i \alpha$.

1.11. It doesn't move, but *they* do: Charge carriers may very well pass through the region of charge imbalance, being accelerated by the electric field and slowed down by the invoked "friction" along the way, and leave the apparent net charge constant. But how does the charge dynamics account for this behavior? Imagine two kinds of carriers, positive and negative but identical in all other respects, and argue against the logical consistency of the myth we used to justify Ohm's law. (This is more than a mere exercise, rather a theme for reflection. See the **Int. Compumag Society Newsletter, 3,** 3 (1996), p. 14.)

SOLUTIONS

1.1. Eliminate h and d: Then $\partial_t b + \mathrm{rot}\, e = 0$, unchanged, and

$$-\varepsilon_0\, \partial_t e + \mathrm{rot}(\mu_0^{-1}\, b) = j + \partial_t p + \mathrm{rot}\, m,$$

so j can "absorb" p and m at leisure. Alternatively, p can assume the totality of charge fluxes (integrate $j + \mathrm{rot}\, m$ in t). But one can't put all of them in rot m, since $j + \partial_t p$ may not be divergence-free. One calls rot m the density of *Amperian currents*.

1.2. Consider a domain D in configuration space (Fig. 1.5). The decrease of the mass it contains, which is $-\int_D \partial_t f$, equals outgoing mass. The latter is the flux through the boundary S of the vector field $\{v, \gamma\}\, f$, which is the speed, not of a particle in physical space, but of the representative point $\{x, v\}$ in configuration space. By Ostrogradskii, $\partial_t f + \mathrm{div}(\{v, \gamma\}\, f) = 0$. Since γ does not depend on v, $\mathrm{div}(\{v, \gamma\}) = 0$. So $\mathrm{div}(\{v, \gamma\}\, f) = \{v, \gamma\} \cdot \nabla f \equiv v \cdot \nabla_x f + \gamma \cdot \nabla_v f$. (Be wary of the wavering meaning of the dot, which stands for the dot-product in V × V left to the ≡ sign, but for the one in V right to it.) If γ depends on v, an additional term $f\, \mathrm{div}_v \gamma$ will appear on the left-hand side of (34). (Here, $\mathrm{div}_v \gamma$ is the divergence of γ considered as a field on V, the x–coordinates being mere parameters.)

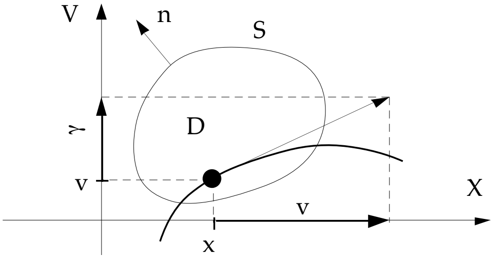


FIGURE 1.5. Notations for Exer. 1.2. The open curve is the trajectory of {x, v} in configuration space.

1.4. Let X be E_3, V the associated vector space (denoted V_3 in A.2.2). With v → e + v × b, we are dealing with a V-valued function, the domain[41] of which is all or part of the vector space V, considered with its affine structure, and position x and time t (which are what e and b may depend on) are parameters. (This is an illustration of the notion of section, cf. A.1.1: section by {x, t} of the function {t, x, v} → e(t, x) + v × b(t, x).) Now, e does not depend on v, and since this is also the case for b, one has div(v → v × b) = 0, after the result of Exercise 1.3.

1.6. Last term in (11) is $\mathrm{div}(v \to Q_c\,(e + v \times b)\,\tilde{q})$, the integral of which over V (with t and x as parameters) is zero if $\tilde{q}$ vanishes fast enough. And by (8), $\partial_t q + \mathrm{div}\, j = \int_V [\partial_t \tilde{q} + \mathrm{div}(v\, \tilde{q})] = \int_V (\partial_t \tilde{q} + \ v \cdot \nabla_x \tilde{q})$, thus 0 after (11).

1.7. The *density* q_p does not transform like a *function* in the change of reference frame defined by x → x + u(x)/2, because the volume element also changes, unless div u = 0, which characterizes volume-preserving deformations. A correct computation must therefore explicitly take into account the Jacobian of the mapping x → x + u(x)/2. Hence a more involved computation in the case when div u ≠ 0, for of course the same final result.

1.11. Let $\rho = 1/\sigma$ be the conductivity. Assume steady currents. Then div j = 0 by (1), $e = \rho j$ if Ohm's law is valid, and $q = \mathrm{div}(\varepsilon_0\, e) = \varepsilon_0\, e \cdot \nabla\rho$, nonzero if ρ varies with position. This result clashes with the predictions of the simple-minded model in which there would be two symmetrical, but oppositely charged, kinds of carriers. Charges of opposite signs moving

[41] Be aware that "domain" has a dual meaning, open connected set as in Note 7, or domain of definition of a map, as here. Cf. Appendix A for precise definitions.

in opposite directions yield a net nonzero current, but a zero macroscopic charge. Under the basic assumption of the myth (speed proportional to electric field), the symmetry between the two kinds of charge is total, and hence $q = 0$. This is enough to show there is a problem. See the **Int. Compumag Society Newsletter, 4,** 1 (1997), pp. 13–18, for a discussion, including my own answer (the *inertia* of charge carriers plays a role in suppressing what would be otherwise a logical conundrum) and two other approaches [Cp], [Ni].

REFERENCES

[AJ] J.E. Allen, T.V. Jones: "Relativistic recoil and the railgun", **J. Appl. Phys., 67**, 1 (1990), pp. 18–21.

[BS] U. Backhaus, K. Schäfer: "On the uniqueness of the vector for energy flow density in electromagnetic fields", **Am. J. Phys., 54**, 3 (1986), pp. 279–280.

[BL] C.K. Birdsall, A.B. Langdon: **Plasma Physics via Computer Simulation,** McGraw-Hill (New York), 1985.

[Bi] K. Birkeland: "Ueber die Strahlung electromagnetischer Energie im Raume", **Annalen der Physik, 52** (1894), pp. 357–380.

[Bo] A. Bossavit: "On local computation of the force field in deformable bodies", **Int. J. Applied Electromagnetics in Materials**, **2**, 4 (1992), pp. 333–343.

[Ca] B. Cabrera : "First results from a superconducting detector for moving magnetic monopoles", **Phys. Rev. Lett., 48** (1982), pp. 1378–1380.

[Cp] J. Carpenter: "Why don't they move?", **Int. Compumag Society Newsletter, 4,** 1 (1997), p. 13.

[Cs] B.R. Casserberg: "Electromagnetic momentum introduced simply", **Am. J. Phys., 50**, 5 (1982), pp. 415–416.

[Co] E. Comay: "Exposing 'hidden momentum' ", **Am. J. Phys., 64**, 8 (1996), pp. 1028–1033.

[Cr] M.J. Crowe: **A History of Vector Analysis**, University of Notre Dame Press, 1967 (Dover edition, New York, 1985).

[FC] R.M. Fano, L.J. Chu, R.B. Adler: **Electromagnetic Fields, Energy, and Forces,** J. Wiley (New York), 1960.

[Fe] R.P. Feynman, R.B. Leighton, M. Sands: **The Feynman Lectures on Physics**, Addison-Wesley (Reading, MA), 1964.

[FS] K.P. Foster, H.P. Schwan: "Dielectric properties of tissues and biological materials: a critical review", **Critical Reviews in Biomedical Engineering, 17**, 1 (1989), pp. 25–104.

[GT] A.S. Goldhaber, W.P. Trower: "Resource Letter MM-1: Magnetic monopoles", **Am. J. Phys., 58**, 5 (1990), pp. 429–439.

[Hd] J.S. Hadamard: **Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques**, Hermann (Paris), 1932.

[He] M.A. Heald: “Energy flow in circuits with Faraday emf”, **Am. J. Phys., 56**, 6 (1988), pp. 540–547.

[Hs] D. Hestenes: "Vectors, Spinors, and Complex Numbers in Classical and Quantum Physics", **Am. J. Phys., 39,** 9 (1971), pp. 1013–1027.

[HE] R.W. Hockney, J.W. Eastwood: **Computer Simulation Using Particles,** McGraw-Hill (New York), 1981.

[Ho] G.W.O. Howe: “Some electromagnetic problems”, **Proc. IEE, 97 I,** 106 (1950), pp. 129–135.

[Ja] B. Jancewicz: **Multivectors and Clifford Algebra in Electrodynamics,** World Scientific (Singapore), 1988.

[Jo] A.K. Jonscher: “The 'universal' dielectric response”, **Nature**, 267 (23 June 1977), pp. 673–679.

[Ke] J.M. Keller: “Newton's third law and electrodynamics”, **Am. J. Phys., 10**, (1942), pp. 302–307.

[Kl] W. Klein: “In memoriam J. Jaumann: A direct demonstration of the drift velocity in metals”, **Am. J. Phys., 55**, 1 (1987), pp. 22–23.

[KB] P. Krumm, D. Bedford: “The gravitational Poynting vector and energy transfer”, **Am. J. Phys., 55**, 4 (1987), pp. 362–363.

[La] G. Lacroux: **Les aimants permanents**, Technique et Documentation-Lavoisier (Paris), 1989.

[Ly] C.S. Lai: “Alternative choice for the energy flow vector of the electromagnetic field”, **Am. J. Phys., 49** (1981), pp. 841–843.

[Lm] G.G. Lombardi: “Feynman's disk paradox”, **Am. J. Phys., 51**, 3 (1983), pp. 213–214.

[Lo] P. Lorrain: “Alternative choice for the energy flow vector of the electromagnetic field”, **Am. J. Phys., 50**, 6 (1982), p. 492.

[Li] J.D. Livingston: **Driving Force,** The natural magic of magnets, Harvard U.P. (Cambridge, Mass.), 1996.

[MW] M. Mason, W. Weaver: **The Electromagnetic Field,** University of Chicago (Chicago), 1929 (Dover edition).

[Ma] J.C. Maxwell: **A Treatise on Electricity and Magnetism**, Clarendon Press, 3^d ed., 1891 (Dover edition, New York, 1954).

[Mi] C.W. Misner, K.S. Thorne, J.A. Wheeler: **Gravitation**, Freeman (New York), 1973.

[Ni] A. Nicolet: “And yet they move . . . A classical approach to Ohm’s law”, **Int. Compumag Society Newsletter, 4,** 1 (1997), pp. 14–16.

[OZ] S.L. O'Dell, R.K.P. Zia: “Classical and semiclassical diamagnetism: A critique of treatment in elementary texts”, **Am. J. Phys., 54**, 1 (1986), pp. 32–35.

[PH] P. Penfield, Jr., H.A. Haus: **Electrodynamics of Moving Media**, The MIT Press (Cambridge, MA), 1967.

[P&] P.B. Price, E.K. Shirk, W.Z. Osborne, L.S. Pinsky: “Evidence for Detection of a Moving Magnetic Monopole”, **Phys. Rev. Lett., 35** (1975), pp. 487–490.

[PP] E.M. Pugh, G.E. Pugh: "Physical significance of the Poynting vector in static fields", **Am. J. Phys., 35**, 1 (1967), pp. 153–156.

[Rb] F.N.H. Robinson: **Macroscopic Electromagnetism**, Pergamon Press (Oxford), 1973.

[Ro] R.H. Romer: “Alternatives to the Poynting vector for describing the flow vector of electromagnetic energy”, **Am. J. Phys., 50**, 12 (1982), pp. 1166–1168.

[Sa] N. Salingaros: “Invariants of the electromagnetic field and electromagnetic waves”, **Am. J. Phys., 53,** 4 (1985), pp. 361–364.

[Sc] L. Schwartz: **Théorie des distributions**, Hermann (Paris), 1966.

[Sl] J. Slepian: “Energy and Energy Flow in the Electromagnetic Field”, **J. Appl. Phys., 13** (1942), pp. 512–518.

[Sp] R.J. Stephenson: "Development of vector analysis from quaternions", **Am. J. Phys., 34**, 2 (1966), pp. 194-201.

[St] M.A. Stuchly, S.S. Stuchly: “Dielectric Properties of Biological Substances — Tabulated”, **J. Microwave Power, 15**, 1 (1980), pp. 19–26.

[TW] E.F. Taylor, J.A. Wheeler: **Spacetime Physics**, Freeman (New York), 1992.

[We] V.F. Weisskopf: “On the Theory of the Electric Resistance of Metals”, **Am. J. Phys., 11**, 1 (1943), pp. 1–12.

CHAPTER 2

Magnetostatics: “Scalar Potential” Approach

2.1 INTRODUCTION: A MODEL PROBLEM

Let us now tackle problem (31) from Chapter 1: magnetostatics. We need a model problem for this discussion; we need it to be as simple as possible, and still come from the real world.

The following, known as the “Bath cube” problem [DB], will do. It is concerned with a device, built around 1979 at Bath University, which was essentially a hollow box between the poles of a large electromagnet (Fig. 2.1). In this almost closed experimental volume, various conducting or magnetizable objects could be placed, and probes could be installed to measure the field. The purpose was to confront what computational codes would predict with what these probes recorded. The problem was one in a series of such benchmark problems, regularly discussed in an ad-hoc forum (the TEAM international workshop [T&]). Comparative results for this one (known as “Problem 5”) can be found in [B5].

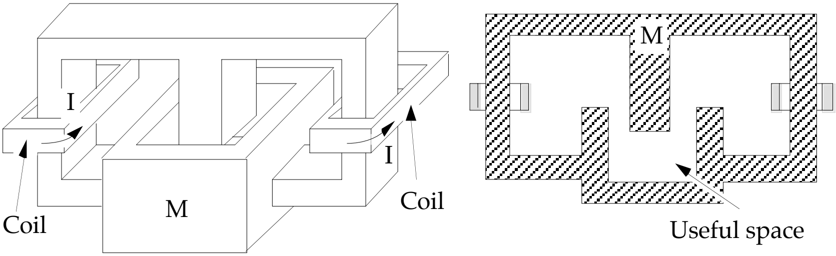


FIGURE 2.1. The “Bath cube” benchmark. Both coils bear the same intensity I. The magnetic circuit M is made of laminated iron, with high permeability ($\mu > 1000\ \mu_0$). Various objects can be placed in the central experimental space.

Problem 5 was actually an eddy-current problem, with alternating current in both coils, and we shall address it in Chapter 8. What we discuss here is the corresponding *static* problem, with DC currents: given the coil-current, find the field inside the box.

It will be some time before we are in a position to actually *solve* this problem, despite its obvious simplicity. For before solving it, we must *set* it properly. We have a physical situation on the one hand, with a description (dimensions, values of physical parameters) and a query (more likely, an endless series of queries) about this situation, coming from some interested party (the Engineer, the Experimenter, the Customer, . . .). To be definite about that here, we shall suppose the main query is, "What is the reluctance of the above device?" The task of *our* party (the would-be Mathematical Modeller, Computer Scientist, and Expert in the Manipulation of Electromagnetic Software Systems) is to formulate a *relevant* mathematical problem, liable to approximate solution (usually with a computer), and this solution should be in such final form that the query is answered, possibly with some error or uncertainty, but within a controlled and predictable margin. (Error *bounds* would be ideal.)

Mathematical modelling is the process by which such a correspondence between a physical situation and a mathematical problem is established.[1] In this chapter, a model for the above situation will be built, based on the so-called "scalar potential variational formulation". We shall spiral from crude *attempts* to set a model to refined ones, via *criticism* of such attempts. Some points about modelling will be made along the way, but most of the effort will be spent on sharpening the mathematical tools.

First attempt, based on a literal reading of Eqs. (1.31). We are given a scalar field μ, equal to μ_0 in the air region, and a time-independent vector field j (actually, the j^g of (1.31), but we may dispense with the superscript g here). From this data, *find vector fields* b *and* h *such that*

(1) $\quad \mathrm{rot}\, h = j,$

(2) $\quad b = \mu\, h,$

(3) $\quad \mathrm{div}\, b = 0,$

in all space.

The first remark, predictable as it was, may still come as a shock: *This formulation doesn't really make sense; the problem is* not *properly posed this way.*

[1] It requires from *both* parties a lot of give and take.

2.2 HONING OUR TOOLS

At least two things disqualify (1–3) as a proper formulation. One is the non-uniqueness of b and h, a mild problem which we'll address later. The other is the implicit and unwarranted assumption of *regularity*, or smoothness, of these fields. For instance, div b = 0 makes perfect sense if the three components b^1, b^2, b^3, in Cartesian coordinates, are differentiable. Then $(\text{div } b)(x) = \partial_1 b^1(x) + \partial_2 b^2(x) + \partial_3 b^3(x)$, a well-defined function of position x, and the statement "div b = 0" means that this function is identically 0. No ambiguity about that. But we can't assume such differentiability.[2] As one knows, and we'll soon reconfirm this knowledge, the components of b are *not* differentiable, not even continuous, at some material interfaces. Still, conservation of the induction flux implies a very definite "transmission condition" on S.

2.2.1 Regularity and discontinuity of fields

Since smoothness, or lack thereof, is the issue, let's be precise, while introducing some shorthands. D being a space domain,[3] the set of all functions continuous at all points of D is denoted $C^0(D)$. A function is *continuously differentiable* in D if all its partial derivatives are in $C^0(D)$, and one denotes by $C^1(D)$ the set of such functions (an infinite-dimensional linear space). Similarly, $C^k(D)$ or $C^\infty(D)$ denote the spaces composed of functions which have continuous partial derivatives of all orders up to k or of all orders without restriction, inside D. In common parlance, one says that a function "is C^k", or "is C^∞" in some region, implying that there is a domain D such that $C^k(D)$, or $C^\infty(D)$, includes the restriction of this function to D as a set element. "Smooth" means by default C^∞, but is often used noncommittally to mean "as regular as required", that is, C^k for k high enough. (I'll say "k-smooth" in the rare cases when definiteness

[2] This is not mere nit-picking, not one of these gratuitous "rigor" or "purity" issues. We have here a tool, differential operators, that fails to perform in some cases. So it's not the right tool, and a better one, custom-made if necessary, should be proposed, one which will work also in borderline cases. Far from coming from a position of arrogance, this admission that a mismatch exists between some mathematical concepts and the physical reality they are supposed to model, and the commitment to correct it, are a manifestation of modesty: When the physicist says "this tool works well almost all the time, and the exceptions are not really a concern, so let's not bother", the mathematician, rather than hectoring, "But you have no *right* to do what you do with it," should hone the tool in order to make it able *also* to handle the exceptions.

[3] Recall the dual use of "domain", here meaning "open connected set" (cf. Appendix A, Subsection A.2.3).

on this point is important.) These notions extend to vector fields by applying them coordinatewise.

In principle, the gradient of a function is only defined at *interior* points of its domain[4] of definition, since the gradient is a record of variation rates in all directions. Depending on the local shape of the boundary, it may still be possible to define a gradient at a boundary point, by taking directional derivatives. How to do that is clear in the case of a smooth boundary (on each line through a boundary point, there is a half-line going *inwards).* But it's more problematic at a corner, at the tip of a cusp, etc. This is why the concept of smoothness *over* a region (not only *inside* it), including the boundary, is delicate. To avoid ambiguities about it, I'll say that a function f is *smooth over* a region R (which may itself be very irregular, devoid of a smooth boundary) if there is a domain D containing R *in* which some extension of f (cf. A.1.2) is smooth. (See Fig. 2.2.)

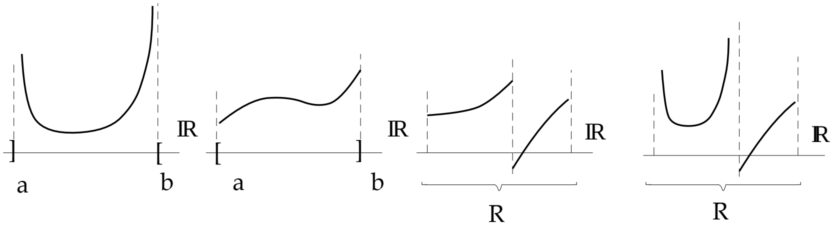


FIGURE 2.2. Notions of smoothness, for a function of a real variable. Left to right, functions which are: smooth *in*]a, b[, smooth *over* [a, b], *piecewise* smooth in region R, *not* piecewise smooth.

Piecewise smooth, then, has a precise meaning: It refers to a function, the domain of which can be partitioned into a mosaic of regions, in finite number, *over* each of which the function is smooth. This does not exclude discontinuities across inner boundaries, but allows only frank discontinuities (of the "first kind"), or as we shall say below, "jumps".

Exercise 2.1. Check that a piecewise smooth function f has a definite integral $\int_D |f|$ on a *bounded* domain. Is this latter assumption necessary?

Now let's return to the case at hand and see where exceptions to smoothness can occur. In free space ($\mu = \mu_0$, and $j = 0$), $\text{rot}\, h = 0$ and $\text{div}\, h = 0$, and the same is true of b. We have this well-known formula which says that, for a C^2-vector field u,

$$\text{(4)} \qquad \text{rot rot } u = \text{grad div } u - \Delta u,$$

[4] The other meaning of the word (Subsection A.1.2).

where Δu is the field, the components of which are $\{\Delta u^1, \Delta u^2, \Delta u^3\}$. So both h and b are *harmonic*, $\Delta h = 0$ and $\Delta b = 0$, in free space. A rather deep result, *Weyl's lemma*, can then be invoked: *harmonic functions are* C^∞. So both b and h are smooth.[5] The same argument holds unchanged in a region with a uniform μ, instead of μ_0.

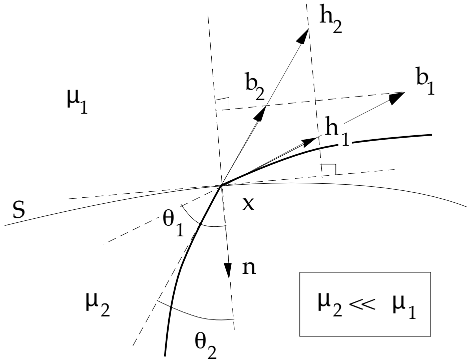


FIGURE 2.3. Flux line deviation at a material interface. The pair $\{h_i, b_i\}$ is the field on side i, where $i = 1$ or 2.

In case two regions with different permeabilities μ_1 and μ_2 are separated by a smooth surface S (Fig. 2.3), b and h will therefore be smooth on both sides, and thus have well-defined flux lines.[6] But the latter will not go straight through S. They deviate there, according to the following "law of tangents":

(5) $$\mu_2 \tan\theta_1 = \mu_1 \tan\theta_2,$$

where θ_1 and θ_2 are the angles the flux half-lines make with the normal n at the traversal point x. So if $\mu_1 \neq \mu_2$, neither b nor h can be continuous at x. Formula (5) is an immediate consequence of the two equalities, illustrated in Fig. 2.3,

(6) $$n \cdot b_1 = n \cdot b_2 \text{ on } S,$$ (7) $$n \times h_1 = n \times h_2 \text{ on } S,$$

called *transmission conditions*, which assert that the normal part of b

[5] A similar, stronger result by Hörmander [Hö] implies that h and b are smooth if μ itself is C^∞. Cf. [Pe]. All this has to do with one (number 19) of the famous Hilbert problems [Br].

[6] A *flux line* of field b through point x_0 is a trajectory $t \to x(t)$ such that $x(0) = x_0$ and $(\partial_t x)(t) = b(x(t))$. If b is smooth and $b(x_0) \neq 0$, there is such a trajectory in some interval $]-\beta, \alpha[$ including 0, by general theorems on ordinary differential equations. See, e.g., [Ar], [CL], [Fr], [LS].

and the tangential part of h are continuous across S, and which we now proceed to prove.

The proof of (6) comes from an integral interpretation of Faraday's law. By the latter, the induction flux through any *closed* surface vanishes. Let's apply this to the surface of the "flat pillbox" of Fig. 2.4, built from the patch Ω (lying in S) by extrusion. This surface is made of two surfaces Ω_1 and Ω_2 roughly parallel to S, joined by a thin lateral band. Applying Ostrogradskii[7] and letting the box thickness d go to 0, one finds that $\int_\Omega (n \cdot b_1 - n \cdot b_2) = 0$, because the contribution of the lateral band vanishes at the limit, whereas $n \cdot b$ on Ω_1 and Ω_2 respectively tend to the values $n \cdot b_1$ and $n \cdot b_2$ of $n \cdot b$ on both sides of Ω. Hence (6), since Ω is arbitrary.

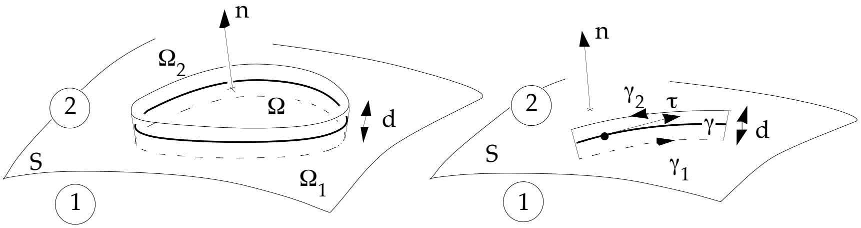


FIGURE 2.4. Setup and notations for the proof of (6) and (7).

As for (7), we rely on the integral interpretation of Ampère's theorem: the circulation of h along a closed curve is equal to the flux of $j + \partial_t d$ through a surface bound by this curve. Here we apply this to the "thin ribbon" of Fig. 2.4, built by extrusion from the curve γ lying in S. Since $j + \partial_t d = 0$ in the present situation, the circulation of h along the boundary of the ribbon is zero. Again, letting the ribbon's width d go to 0, we obtain $\int_\gamma (\tau \cdot h_1 - \tau \cdot h_2) = 0$, which implies, since γ is arbitrary, the equality of the projections (called "tangential parts") of h_1 and h_2 onto the plane tangent to S. This equality is conveniently expressed by (7).

Fields therefore fail to be regular at all material interfaces where μ presents a discontinuity, and div or rot cease to make sense there. Some regularity subsists, however, which is given by the interface conditions (6) and (7). For easier manipulation of these, we shall write them $[n \cdot b]_S = 0$ and $[n \times h]_S = 0$, and say that the *jumps* of the normal part of b and of the tangential part of h vanish at all interfaces. Before discussing the possibilities this offers to correctly reformulate (1–3), let's explain the notation and digress a little about jumps.

[7] Flux, circulation, and relevant theorems are discussed in detail in A.4.2.

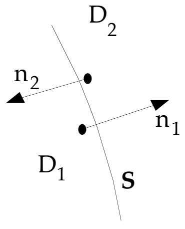
2.2.2 Jumps

This section is a partly independent development about the bracket notation [ ] for jumps, which anticipates further uses of it.

Consider a field (scalar- or vector-valued) which is smooth on both sides of a surface S, but may have a discontinuity across S, and suppose S is provided with a crossing direction. The *jump* across S of this quantity is by definition equal to its value just before reaching the surface, minus its value just after. (The jump is thus counted downwards; rather a "drop", in fact.) Giving a crossing direction through a surface is equivalent to providing it with a continuous field of normals. One says then that the surface has been *externally oriented*.[8]

Not all surfaces can thus be oriented. For one-sided surfaces (as happens with a Möbius band), defining a continuous normal field is not possible, and the crossing direction can only be defined locally, not consistently over the whole surface. For surfaces which enclose a volume, and are therefore two-sided, the convention most often adopted consists in having the normal field point outwards. This way, if a function φ is defined inside a domain D, and equal to zero outside, its jump $[\varphi]_S$ across the surface S of D is equal to the trace φ_S of φ, that is, its restriction to S if φ is smooth enough, or its limit value from inside otherwise. The conventions about the jump and the normal thus go together well.

For interfaces between two media, there may be no reason to favor one external orientation over the other. Nonetheless, some quantities can be defined as jumps in a way which does not depend on the chosen crossing direction.



Consider for instance the flux of some vector field j through an interface S between two regions D_1 and D_2 (inset). Let n_1 and n_2 the normal fields defined according to both possible conventions: n_1 points from D_1 towards D_2, and n_2 points the other way. Suppose we choose n_1 as the crossing direction, and thus set $n = n_1$. Then the jump of the normal component $n . j$ across S is by definition equal to its value on the D_1 side, that is, $n_1 . j$, minus the value of $n . j$ on the D_2 side, which is $- n_2 \cdot j$. The jump is thus the *sum* $n_1 \cdot j + n_2 \cdot j$. This is symmetrical with respect to D_1 and D_2, so we are entitled to speak of "the jump of $n \cdot j$ across S" without specifying a crossing direction. Whichever this direction, the decrease of $n \cdot j$ when crossing will be the same, because the sign of n intervenes twice, in the choice of direction,

[8] Which suggests there is also a different concept of *internal* orientation (cf. Chapter 5).

and in the choice of which side one "jumps from". Hence the definition of the jump of j as $[n \cdot j] = n_1 \cdot j_1 + n_2 \cdot j_2$, where j_1 and j_2 are the values on each side of S.

Such jumps often have interesting physical interpretations. For instance, if j is a current density, the jump is equal to the intensity that is "instilled in S", and is withdrawn by some mechanism. When no such mechanism exists, as for instance at the interface between two conductors, the jump must vanish. But it may happen otherwise. For instance, if S corresponds to a highly conducting inclusion inside a normal conductor, the current $[n \cdot j]$ withdrawn at some place will be conveyed along S and reinjected at other places, where $[n \cdot j]$ will be negative. Note that such considerations would apply to a Möbius band without any problem.

In the case of the electric induction d, the jump $[n \,.\, d]_S$ is the density of electric charge present on surface S; hence the interface condition $[n \,.\, d]_S = 0$, unless there is a physical reason to have electric charge concentrated there. Same thing with b, and magnetic charge. Our proof above that $[n \cdot b] = 0$ across all interfaces made implicit use of the absence of such charge. But there are problems in which the jump of a quantity denoted $n \cdot b$ can be nonzero. This happens, for instance, when fictitious surface magnetic charges are used as auxiliary quantities in integral methods, and then $[n \cdot b] = q$, the fictitious charge density.

A bit different is the case of vector quantities, such as the magnetic field. The jump $[h]_S$ is simply the field, defined on S, obtained by taking the jumps of the three coordinates. The subject of interest, however, is more often the jump of the *tangential* part of h.

If h is smooth, we call *tangential part* and denote by h_S the field of vectors tangent to S obtained by projecting h(x), for all x in S, on the tangent plane T_x at x (Fig. 2.5, left). If h is smooth on both faces of S but discontinuous there, there are two bilateral projections h_{S1} and h_{S2}, and the jump of h_S, according to the general definition, is $[h_S]_S = h_{S1} - h_{S2}$ in the case of Fig. 2.5. The sign of this of course depends on the crossing direction. But the remark

$$(8) \qquad [h_S]_S = - n \times [n \times h]_S = - n \times (n_1 \times h_1 + n_2 \times h_2)$$

points to the orientation-independent surface vector field $[n \times h]_S$. This is equal to the jump $[h_S]_S$ of the tangential part, up to a 90° rotation, counterclockwise, around the normal. The cancellation of the tangential jump is thus conveniently expressed by $[n \times h]_S = 0$.

This field $[n \times h]$ is interesting for another reason. As suggested by Fig. 2.5, right, $[n \times h]$ is always equal to minus the current density j_S (a surface vector field, thus modelling a "current sheet") supported by the interface. For instance, if the crossing direction is from region 1 to region 2, and thus $n = n_1 = - n_2$, then $[n \times h] = n_1 \times h_1 + n_2 \times h_2$, which is $- j_S$ by Ampère's theorem. We find the same result with the other choice. Again, if there is no way to carry along the excess current (such as, for instance, a thin sheet of high conductivity borne by S), then $j_S = - [n \times h] = 0$, which is the standard transmission condition about h we derived earlier.

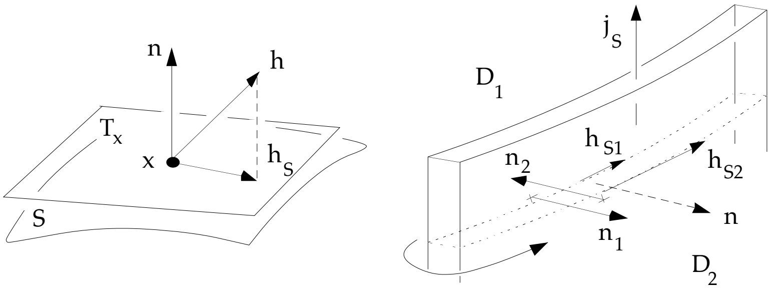


FIGURE 2.5. Left: Definition of h_S. Right: Relation between j_S and the jump of h_S. (Take the circulation of h along the circuit indicated.)

In a quite similar way, $[n \times e]$ is equal, irrespective to the choice of normal, to the time-derivative of the induction flux vector, $\partial_t b_S$, along the surface. This is most often 0—hence the transmission condition $[n \times e] = 0$—but not always so. By way of analogy with the previous example, the case of a thin highly permeable sheet will come to mind. But there are other circumstances, when modelling a thin air gap, for instance, or a crack within a conductor in eddy-current testing simulations, when it may be necessary to take account of the induction flux in a direction tangential to such a surface.

2.2.3 Alternatives to the standard formalism

Back to our critical evaluation of the ill-specified equation $\text{div } b = 0$: What can be done about it? A simple course would be to explicitly acknowledge the exceptions, and say, "We want $\text{div } b = 0$ wherever b is effectively differentiable, and $[n \cdot b] = 0$ across all material interfaces and surfaces where a singularity might occur." Indeed, many textbooks list transmission conditions as equations to be satisfied, and add them to Maxwell equations, on almost the same footing.

It would be quite awkward, however, always to be forced to dot i's and cross t's that way. Besides, and more importantly, the practice of finite elements does not suggest that material interfaces should contribute *additional* equations. So there must be a way to stretch the meaning of statements such as div b = 0 in order to *imply* the transmission conditions. In fact there are two main ways. A radical one: differential forms; and a moderate one: weak formulations, laid on top of the bedrock of classical vector calculus.

The radical way will not be followed here, but must be mentioned, because being aware of its existence helps a lot in understanding the surprising analogies and formal symmetries that abound in the classical approach. When looking for substitutes for the *differential, local* equations div b = 0 and rot h = 0, we invoked *integral, global* relations: flux conservation, Ampère's theorem. All electromagnetic laws (apart from constitutive laws) say things like "This circulation along that line is equal to this flux across that surface, this volume integral equals that charge," and so forth, with line, surface, and volume in a definite and simple relationship, such as "is the boundary of". The laws thus appear as relations between real quantities assigned to geometrical elements (points, lines, surfaces, volumes), and the scalar or vector fields are there as a way to compute these quantities.

Once we begin to see things in this light, some patterns appear. Fields like e and h are definitely associated with *lines:* One takes their *circulations,* which are electromotive forces (emf) and magnetomotive forces (mmf). The same can be said about the vector potential a. And it can't be a coincidence either if when a curl is taken, one of these fields is the operand. Fields like b and d, or j, in contrast, are *surface* oriented, their *fluxes* matter, and it's div, rarely rot, which is seen acting on them. Even the scalar fields of the theory (charge density q, magnetic potential φ, electric potential ψ) have an associated dimensionality: Point values of φ and ψ matter, but only volume integrals of q are relevant, and terms like grad q are never encountered, contrary to grad φ or grad ψ.

This forces us to shift attention from the fields to the linear mappings of type *GEOMETRIC_ELEMENT → REAL_NUMBER* they help realize. For instance, what matters about h, physically, is not its pointwise values, but its circulations along lines (mmf). Thus, the status of h as a *LINE → REAL* linear map is more important than its status as a vector field. The status of b as a *SURFACE → REAL* linear map is what matters (and in this respect, b and h are different kinds of vector fields). The (mathematical) fields thus begin to appear as mere props, auxiliaries in the description of the (physical) field as a connector between geometrical entities.

Which somewhat devalues differential operators, too: grad, rot and div, in this light, appear as auxiliaries in the expression of conservation relations, as expressed by the Ostrogradskii and Stokes theorems. Their failure to make sense locally is thus not to be taken too seriously.

Proper form is given to the foregoing ideas in *differential geometry*. There, one forgets about the scalar or vector fields and one focuses on the mappings they represent (and thus, to some extent, hide). Fields of linear mappings of $GEOMETRIC\_ELEMENT \rightarrow REAL$ type are called *differential forms*, of *degree* 0 to 3 according to the dimension of the geometric objects they act upon, and under regularity assumptions which are milder than for the scalar or vector proxies, one defines a unique operator d, the *exterior differential*, which is realized as grad, rot, or div, depending on the dimension. *All* laws of electromagnetism can be cast in this language (including constitutive laws, which are mappings from p-forms ($p = 0$ to 3) to $(3 - p)$-forms).

The moderate approach we now follow does not go so far, and keeps the fields as basic objects, but stretches the meaning of the differential operators, so that they continue to make sense for some discontinuous fields. The main idea is borrowed from the theory of distributions: Instead of seeing fields as collections of pointwise values, we consider how they act on other fields, by integration. But the full power of the theory of distributions is not required, and we may eschew most of its difficulties.

2.3 WEAK FORMULATIONS

First, some notation. Symbols C^k and C^∞ for smoothness have already been introduced, compact support[9] is usually denoted by a subscripted 0, and blackboard capitals are used in this book to stress the vector vs scalar opposition when referring to spaces of fields. Putting all these conventions together, we shall thus have the following list of infinite-dimensional linear spaces:

- $C^k(E_3)$: The vector space of all k-smooth functions in E_3,
- $\mathbb{C}^k(E_3)$: All k-smooth vector fields in E_3,

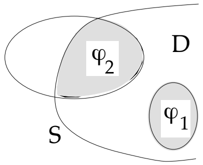
[9] The *support* of a function, real- or vector-valued, is the closure of the set of points where it doesn't vanish. Cf. A.2.3.

- $C_0^k(D)$, $\mathbb{C}_0^k(D)$: Same, with compact support contained in D,

being understood that domain D can be all E_3, and finally,

- $C^k(\overline{D})$, $\mathbb{C}^k(\overline{D})$,

for the vector spaces of restrictions to $\overline{D}$ (the closure of D), of k-smooth functions or vector fields. In all of these, k can be replaced by ∞. (In the inset, the supports of a $\varphi_1 \in C_0^\infty(D)$ and a $\varphi_2 \in C^\infty(\overline{D})$, which is thus the restriction of some function defined beyond D, whose support is sketched.)



2.3.1 The "divergence" side

Now, let's establish a technical result, which generalizes integration by parts. Let D be a regular domain (not necessarily bounded), S its boundary, b a smooth vector field, and φ a smooth function, both with compact support in E_3 (but their supports may extend beyond D). Form $u = \varphi\, b$. Ostrogradskii's theorem asserts that $\int_D \operatorname{div} u = \int_S n \cdot u$, with n pointing outwards, as usual. On the other hand, we have this vector analysis formula,

$$\operatorname{div}(\varphi\, b) = \varphi \operatorname{div} b + b \cdot \operatorname{grad} \varphi.$$

Both things together give

$$(9) \qquad \int_D \varphi \operatorname{div} b = -\int_D b \cdot \operatorname{grad} \varphi + \int_S n \cdot b\ \varphi,$$

a fundamental formula.

By (9), we see that a 1-smooth divergence-free field b in D is characterized by

$$(10) \qquad \int_D b \cdot \operatorname{grad} \varphi = 0 \ \ \forall\, \varphi \in C_0^1(D),$$

since with $\varphi = 0$ on the boundary, there is no boundary term in (9). But (10) makes sense for fields b which are only *piecewise* smooth.[10] We now take a bold step:

Definition 2.1. *A piecewise smooth field* b *which satisfies* (10) *will be said to be* divergence-free, *or solenoidal,* in the weak sense.

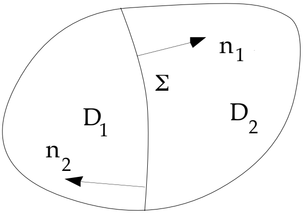
The φ's in (10) are called *test functions*.

[10] All that is required is the integrability of $b \cdot \operatorname{grad} \varphi$ in (10), so 0-smoothness, that is, continuity, of each "piece" of b is enough.

A solenoidal smooth field is of course weakly divergence-free. But from our earlier discussion, we know that the physical b, in magnetostatics, is only piecewise smooth, and satisfies transmission conditions. Hence the interest of the following result:

Proposition 2.1. *For a piecewise-smooth* b, (10) *is equivalent to* $\operatorname{div} b = 0$ *inside regularity regions and* $[n \cdot b] = 0$ *at their interfaces.*

Proof. Recall that "piecewise" means that D can be partitioned into a *finite* number of subdomains D_i in which b is smooth, so a proof with *two* subdomains will be enough. It's a long proof, which will require two steps.



We begin with the shorter one, in which b is supposed to be solenoidal in D_1 and D_2 separately (notation in inset), with $[n \cdot b] = 0$ on the interface Σ. Our aim is to prove (10). Let φ be a test function. Formula (9) holds in D_1 and D_2 separately, and gives

$$\begin{aligned}\int_D b \cdot \operatorname{grad} \varphi &= \int_{D_1} b \cdot \operatorname{grad} \varphi + \int_{D_2} b \cdot \operatorname{grad} \varphi \\ &= -\int_{D_1} \varphi \operatorname{div} b + \int_\Sigma n_1 \cdot b\ \varphi - \int_{D_2} \varphi \operatorname{div} b + \int_\Sigma n_2 \cdot b\ \varphi \\ &= \int_\Sigma [n \cdot b]\ \varphi = 0\end{aligned}$$

since $[n \cdot b] = 0$ has been assumed, hence (10). The absence of surface terms on S is due to the inclusion $\operatorname{supp}(\varphi) \subset D$.

Conversely, suppose (10) holds. Since (10) says "*for all* test functions φ", let's pick one which is supported in D_1, and apply (9). There is no surface term, since $\operatorname{supp}(\varphi_1)$ does not intersect Σ, so $0 = \int_{D_1} \operatorname{div} b\ \varphi$. (Note that div b is a smooth function there.) This holds for *all* $\varphi \in C_0^1(D_1)$.[11] But the only way this can happen (see A.2.3 if this argument doesn't sound obvious) is by having $\operatorname{div} b = 0$ in D_1. Same reasoning in D_2, leading to $\operatorname{div} b = 0$ also in D_2. Now, start from (10) again, with a test function which does not necessarily vanish on Σ, and use the newly acquired knowledge that $\operatorname{div} b = 0$ in D_1 and D_2:

$$\begin{aligned}0 = \int_D b \cdot \operatorname{grad} \varphi = \int_{D_1} b \cdot \operatorname{grad} \varphi + \int_{D_2} b \cdot \operatorname{grad} \varphi = \\ = \int_\Sigma n_1 \cdot b\, \varphi + \int_\Sigma n_2 \cdot b\, \varphi = \int_\Sigma [n \cdot b]\ \varphi\end{aligned}$$

for all $\varphi \in C_0^1(D)$. But such test functions can assume any value on Σ, so

[11] Which is why the presence of the quantifier ∀ in Eq. (10) is mandatory. Without it, the meaning of the statement would change totally.

again, the only way $\int_\Sigma [\mathrm{n} \cdot \mathrm{b}]\, \varphi$ can vanish for all of them is by $[\mathrm{n} \cdot \mathrm{b}]_\Sigma$ being 0. ◊

Interface conditions are thus implicit in the "weak solenoidality" condition (10). We shall therefore acquire the "weak formulation reflex": Each time a statement of the form "div b = 0" appears in the formulation of a problem (this is what one calls the "strong formulation"), replace it by the weak formulation (10). This does no harm, since there is equivalence in case b has a divergence in the ordinary ("strong") sense. It does some good if b is only piecewise smooth, since there is no need to make explicit, or even mention, the transmission conditions $[\mathrm{n} \cdot \mathrm{b}\,] = 0$, which are implied by (10), as Prop. 2.1 has shown.

We now see why the φ's are called "test functions": By carefully selecting them, we were able to "test" the equality div b = 0 inside regularity regions, to "test" the transmission condition over Σ, etc. The function div b and the constant 0 were thus deemed equal not because their *values* would coincide, but on the ground that their *effects* on test functions were the same. (This principle, duly abstracted, was the foundation of the theory of distributions.)

Remark 2.1. The reader aware of the "virtual work principle" in mechanics will have recognized the analogy: There too, fields of forces are tested for equality by dot-multiplying them by fields of virtual displacements and integrating, and two force fields are equal if their virtual works always coincide. ◊

An obvious generalization of (10) is

$$(11) \qquad -\int_D \mathrm{b} \cdot \mathrm{grad}\, \varphi = \int_D \mathrm{f}\, \varphi \quad \forall\, \varphi \in C_0^{\,1}(D),$$

where f is a given function (piecewise smooth). This means "div b = f in the weak sense". (**Exercise 2.2**: Check that.)

Remark 2.2. One may wonder to which extent weak solenoidality depends on the regularity of test functions, and this is a good question, since of course the following statement, for instance,

$$(10') \qquad \int_D \mathrm{b} \cdot \mathrm{grad}\, \varphi = 0 \quad \forall\, \varphi \in C_0^{\,\infty}(D),$$

is logically *weaker* than (10): There are fewer test functions, hence fewer constraints imposed on b, and hence, conceivably, more weakly solenoidal fields in the sense of (10') than in the sense of (10). The notion would be of dubious value if things went that way. But fortunately (10) *and* (10') *are equivalent:* This results from the density property proved in Section A.2.3: Given a $C_0^{\,1}$ test function φ, there exists a sequence $\{\varphi_n\}$ of $C_0^{\,\infty}(D)$ functions such that $\int_D |\operatorname{grad}(\varphi_n - \varphi)|^2 \to 0$. Then (**Exercise 2.3**: Check it) $\int_D b \cdot \operatorname{grad} \varphi = \lim_{n \to \infty} \int_D b \cdot \operatorname{grad} \varphi_n$ if (10') holds, which implies (10). ◊

Remark 2.3. As a corollary of Remark 2.2, any function φ such that $\operatorname{grad} \varphi$ is the limit, in the above sense, of a sequence of gradients of test functions, also qualifies as a test function. We shall remember this in due time. ◊

Remark 2.4. If the superscript k in $C_0^k(D)$ is thus not crucial (all that is required of φ, in terms of regularity, is to have a square-integrable gradient), what about the subscript 0, denoting compact support ? *That* is essential. If test functions could assume nonzero values on the boundary, this would put *more* constraints on b than mere solenoidality. We shall make use of this too, when treating boundary conditions. ◊

2.3.2 The "curl" side

All of this cries out for symmetrization: What we did with the divergence operator should have counterparts with the curl operator. This time we know the way and will go faster.

Let a and h both belong to $\mathbb{C}_0^1(E_3)$, and let D be as above. Form $u = h \times a$. We have this other vector analysis formula,

$$\operatorname{div}(h \times a) = a \cdot \operatorname{rot} h - h \cdot \operatorname{rot} a,$$

and by Ostrogradskii again, we get

$$(12) \qquad \int_D h \cdot \operatorname{rot} a = \int_D a \cdot \operatorname{rot} h - \int_S n \times h \cdot a,$$

the second fundamental integration by parts formula, on a par with (9).

Remark 2.5. Note the formal analogies, and also the differences, between (9) and (12): grad became rot, – div became rot too, × replaced the dot, signs changed in puzzling patterns . . . Obviously these two formulas are, in some half-veiled way, realizations of a *unique* one, which would call for different notation and concepts: those of differential geometry. ◊

By (12), a smooth curl-free field h in D is characterized by

$$(13) \qquad \int_D h \cdot \operatorname{rot} a = 0 \quad \forall\, a \in \mathbb{C}_0^1(D).$$

Again (13) makes sense for non-smooth fields h, if they are square-integrable, and hence (now the obvious thing to do):

Definition 2.2. *A piecewise smooth field* h *which satisfies* (13) *will be said to be* curl-free, *or irrotational,* in the weak sense.

We can prove, in quite the same way as Prop. 2.1, what follows:

Proposition 2.2. *For a piecewise-smooth* h, (13) *is equivalent to* $\operatorname{rot} h = 0$ *inside regularity regions and* $[n \times h] = 0$ *at their interfaces.*

Proof. This should be an exercise. If $\operatorname{rot} h = 0$ in the regularity regions D_1 and D_2, and $[n \times h] = 0$ at the interface Σ, then

$$\begin{aligned}\int_D h \cdot \operatorname{rot} a &= \int_{D_1} h \cdot \operatorname{rot} a + \int_{D_2} h \cdot \operatorname{rot} a \\ &= \int_{D_1} a \cdot \operatorname{rot} h - \int_\Sigma n_1 \times h \cdot a + \int_{D_2} a \cdot \operatorname{rot} h - \int_\Sigma n_2 \times h \cdot a \\ &= -\int_\Sigma [n \times h] \cdot a = 0\end{aligned}$$

for all test fields in $\mathbb{C}_0^1(D)$, which is (13). Conversely, assuming (13), we first obtain $\operatorname{rot} h = 0$ in D_1 and D_2 separately by the same maneuvers as above, then, backtracking,

$$\begin{aligned}0 = \int_D h \cdot \operatorname{rot} a &= \int_{D_1} h \cdot \operatorname{rot} a + \int_{D_2} h \cdot \operatorname{rot} a = \\ &= -\int_\Sigma n_1 \times h \cdot a - \int_\Sigma n_2 \times h \cdot a = -\int_\Sigma [n \times h] \cdot a\end{aligned}$$

for all $a \in \mathbb{C}_0^1(D)$. Surface values of a are not constrained on Σ, so the only way this equality can hold is by having $[n \times h]_\Sigma = 0$. ◊

One may generalize there too:

$$(14) \qquad \int_D h \cdot \operatorname{rot} a = \int_D j \cdot a \quad \forall\, a \in \mathbb{C}_0^1(D),$$

with j given, piecewise smooth. This means (**Exercise 2.4:** Make sure you understand this) "rot h = j in the weak sense". Here also, $\mathbb{C}_0^1(D)$ can be replaced by $\mathbb{C}_0^\infty(D)$.

This was only a first brush with weak formulations, and the full potential of the idea has not been exploited yet. Instead of (10) or (10'), we could have characterized divergence-free vector fields by the weak formulation

$$(15) \qquad \int_D b \cdot \operatorname{grad} \varphi = \int_S n \cdot b\ \varphi \quad \forall\, \varphi \in C^\infty(\bar{D}),$$

for instance, which suggests (cf. Remark 2.4) that not only right-hand sides, as in (11), but also some boundary conditions may be accommodated. The symmetrical formula on the curl side is

$$(16) \qquad \int_D h \cdot \operatorname{rot} a = \int_S n \times h \cdot a \quad \forall\, a \in \mathbb{C}^\infty(\bar{D}).$$

With experience, this flexibility turns out to be the most compelling reason to use weak formulations.

2.3.3 The uniqueness issue

So we got rid of the ambiguities hidden in the "strong" formulations rot h = j and div b = 0. A different kind of problem arises about the *uniqueness* of the solution, assuming there is one. Take $j = 0$ in (1), and $\mu = \mu_0$ in all space. The physical solution is then $h = 0$ and $b = 0$. But this is *not* implied by Eqs. (1–3): Take $h = \text{grad}\, \varphi$, where φ is a harmonic function in all space (for instance, to exhibit only one among an infinity of them, $\varphi(x, y, z) = xy$, in x–y–z Cartesian coordinates). Then $\text{rot}\, h = 0$, and $\text{div}(\mu_0 h) = \mu_0\, \Delta\varphi = 0$. So we have here an example of a nonzero static field that satisfies the equations, although there is no source to create it.

All fields of this kind, however, have in common the property of carrying infinite energy, which is the criterion by which we shall exclude them: We want[12] fields with *finite* energy. From Chapter 1, the expression of the energy of the magnetic field is

$$(17) \qquad W_{\text{mag}} = {}^1/_2 \int_{E_3} \mu\, |h|^2 = {}^1/_2 \int_{E_3} \mu^{-1}\, |b|^2.$$

Since $\mu \geq \mu_0$ all over, the first integral is bounded from below by $\mu_0 \int_{E_3} |h|^2/2$; hence the eligible h's are *square-integrable:* $\|\, h\, \| < \infty$, where $\|\ \|$ denotes the quadratic norm, thus defined:

$$\|\, h\, \| = [\int_{E_3} |h(x)|^2\, dx]^{1/2}.$$

If there is also an upper bound μ_1 to μ, which we assume, the same reasoning with the other integral shows that b should be square-integrable as well.

To be consistent with this requirement of finite energy, we shall also assume that j, besides being piecewise smooth, has compact support: this excludes cases such as, for instance, that of a uniform current density in all space, which would generate a field of infinite energy.

All that is required of μ, then (last item in our critical review of (1–3)), is not to spoil these arrangements. We want the integrals in (17) to make sense for all eligible fields h and b, that is, square-integrable fields, and this is the case if μ is piecewise smooth (a reasonable requirement, as regards a material property) and if there exist two positive

[12] Note this is a *modelling choice*, justified in the present situation, not a dogma.

constants μ_0 and μ_1 such that[13]

$$(18) \qquad \mu_0 \leq \mu(x) \leq \mu_1 \text{ a.e. in } E_3.$$

Remark 2.6. The two values in (17) are equal when $b = \mu h$. But notice we have there two different expressions of the energy, one in terms of h, the other in terms of b. It's customary to call *energy* of a vector field b (any vector field b, not necessarily the physical induction) the integral $\frac{1}{2}\int \mu^{-1} |b|^2$, and *coenergy* of h the integral $\frac{1}{2}\int \mu\ |h|^2$. Note that energy(b) + coenergy(h) $\geq \int b \cdot h$, with equality only when $b = \mu h$. ◊

2.4 MODELLING: THE SCALAR POTENTIAL FORMULATION

At last, we found a problem that, first, is relevant to the situation, and second, makes mathematical sense:

Given μ *and* j, *piecewise smooth, with* μ *as in* (18) *and* j *with compact support, find piecewise smooth fields* b *and* h *such that*

$$(19) \qquad \left| \begin{array}{l} \int_{E_3} b \cdot \operatorname{grad} \varphi = 0 \quad \forall\, \varphi \in C_0^\infty(E_3), \\ b = \mu\, h, \\ \int_{E_3} h \cdot \operatorname{rot} a = \int_{E_3} j \cdot a \quad \forall\, a \in \mathbb{C}_0^\infty(E_3). \end{array} \right.$$

Whether there is a solution and how to get it is another story, but at least we have, for the first time so far, a *model*.

2.4.1 Restriction to a bounded domain

For the moment, let us return to physics, and criticize this model on the grounds of an element of the situation which has been neglected up to now: the large value of μ in the magnetic core M of the apparatus. A look at Fig. 2.6 shows that flux lines will arrive almost orthogonally to the "magnetic wall" ∂M (the boundary of M). On the other hand, if μ is large, h must be small in M, since the magnetic energy is finite.[14] We are thus entitled to neglect M in the eventual calculation, and to set

[13] The abbreviation "a.e." stands for "almost everywhere", meaning "at all points except those of some negligible set". (The latter notion is discussed in Appendix A, Subsection A.4.2.) The a.e. clause is a necessary precaution since μ has no definite value at discontinuity points.

[14] Be wary of this line of reasoning, which is correct in the present case, but can lead to unexpected trouble in some topologically complex situations [Bo].

$$n \times h = 0 \text{ on } \partial M$$

as a boundary condition for a problem that will now be posed in the complementary domain $E_3 - M$.

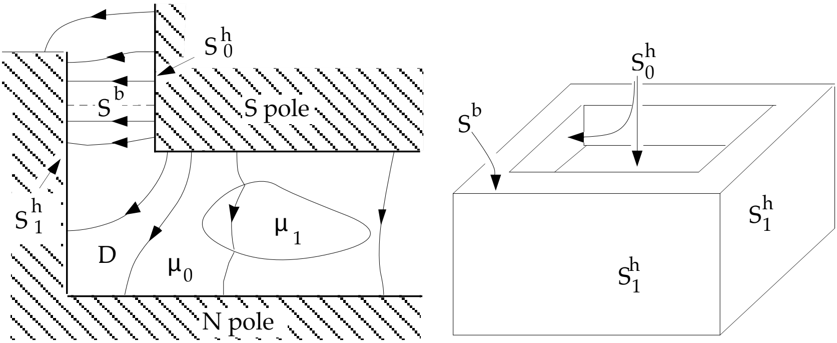


FIGURE 2.6. Left: Expected pattern of field lines inside the box, showing the existence of a horizontal plane on which $n \,.\, b = 0$, an annular part of which, called S^b, will close the box. Right: Persective view of the "computational domain" D thus delimited, and of its surface. One has $S = S^h \cup S^b$, and the "magnetic wall" S^h is in two parts, S^h_0 and S^h_1.

But one can go farther here, and restrict the domain of interest to the "central box" of Fig. 2.1, the experimental volume. Fig. 2.6 explains why: The air region $E_3 - M$ is almost cut in two by the magnetic circuit, and between the North and South poles of the electromagnet, there is an air gap in which the flux lines go straight from one magnetic wall to the opposite one, horizontally. So we can introduce there an artificial boundary (S^b in Fig. 2.6), horizontal, which one can assume is spanned by flux lines (this is only approximately true, but a legitimate approximation), and therefore

$$n \cdot b = 0 \text{ on } S^b.$$

Consequently, let us restrict our computational domain to the part of the inner box below the plane of S^b, and call D this region. Its boundary S is thus made of S^b and of the part of ∂M which bounds the inside of the box, which we shall denote by S^h. Hence our boundary conditions, which, combined with the strong form of the magnetostatics equation, lead to

(20) $\mathrm{rot}\, h = 0$ in D, (21) $n \times h = 0$ on S^h,

(22) $b = \mu\, h$ in D,

(23) $\mathrm{div}\, b = 0$ in D, (24) $n \cdot b = 0$ on S^b.

We note that S^h is in two parts, S^h_0 and S^h_1, corresponding to the two poles of the electromagnet.

This calls for a few remarks. First, let's not forget that materials of various permeabilities can be put inside D, so we must expect discontinuous fields, and weak formulations are still in order.

The second remark is about the symmetry of the box, and of its content. In the "Bath cube" experiment, four identical aluminum cubes (hence the nickname) were put inside the box, symmetrically disposed, so that it was possible to solve for only a quarter of the region, since the whole field is then symmetrical with respect to the vertical symmetry planes, hence $n \cdot b = 0$ there. The equations are thus the same, provided D and S^b are properly redefined: D as a quarter of the cavity, and S^b as a quarter of the former S^b plus the part of the symmetry planes inside the box. We shall do that in Chapter 6, but we may ignore the issue for the time being.

Third remark, the above display (20–24) does not say anything about the source of the field. That was j, the current density in the coil, which is now out of the picture. This lost information must be reintroduced into the formulation in some way.

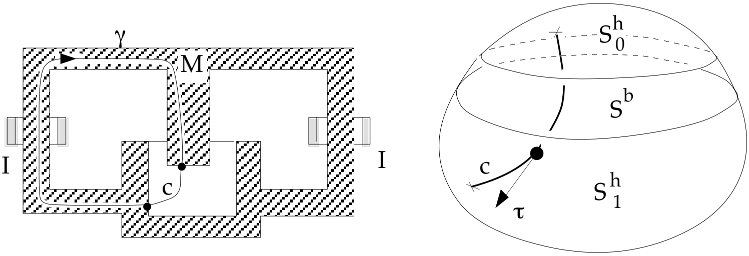


FIGURE 2.7. Left: Applying Ampère's theorem to the path γ shows that the mmf along c is approximately equal to the DC intensity I. Right: topological aspects of the situation.

Figure 2.7 suggests how it can be done. Consider a circuit γ which, except for the part c inside D that links opposite poles, is entirely contained in M. By Ampère's theorem, the circulation of h along γ is equal to I, the DC intensity in the coil.[15] But μ being very large in M, the field h is so small there that the circulation along γ is approximately equal to the circulation along the sub-path c. Since we already assumed $h = 0$ in M in this modelling, we consistently set

$$(25) \qquad \int_c \tau \cdot h = I,$$

where τ is the field of unit tangent vectors along c (Fig. 2.7). (**Exercise 2.5:** Show that any path c from S^h_0 to S^h_1 will give the same circulation.) Now, common sense says that (20–24) and (25) do uniquely determine the field, and the mathematical model we are building had better have this property (which we'll eventually see is the case).

There is another possibility: We could specify the magnetic flux F through the box instead, like this:

$$(26) \qquad \int_{S_1^h} n \cdot b = F.$$

(**Exercise 2.6:** Show that other surfaces than S^h_1 can be used in (26) with the same result. How would you characterize them?) Of course, F is not known here, but this is not important for a *linear* problem: Just solve with some value for F, get I, and scale. In fact, since we want to compute the *reluctance* of the system, which is by definition the ratio $R = I/F$, the flux is the objective of the computation if I is known, and the other way around. We may thus solve (20–24) (25) and then compute F, using some approximation of formula (26), or solve (20–24) (26), with an arbitrary nonzero value for F, and then compute I by (25). This alternative reflects the symmetry between b and h in the problem's formulation.

We shall return to this symmetry (Chapter 6). We now break it by playing the obvious move in the present situation, which is to introduce a *magnetic potential*.

2.4.2 Introduction of a magnetic potential

Indeed, since the field h we want must be curl-free, it is natural to look for it as the gradient of some function φ. The boundary condition $n \times h = 0$ on S^h is then satisfied by taking φ equal to a constant there. (This is general: Magnetic walls are equipotentials for φ in static contexts.) Since S^h is in two pieces, there are two such constants, one of which can be 0. The other one must then be equal to I, after (25).

[15] Notice how the equality of intensities in the energizing coils is necessary in this reasoning: Otherwise, we could not assume μ infinite in M without contradiction. This is a well-known difficulty of the theory of the transformer, which we shall ignore here.

All these considerations lead us to the definition of a class of *admissible* potentials: continuous piecewise smooth functions φ, which satisfy all the a priori requirements we have about φ (finite energy, being equal to 0 or I on S^h), and we shall select in this class *the* potential which solves the problem. This is, still grossly sketched, the *functional point of view:* Define a functional space of eligible candidates, characterize the right one by setting tests it will have to pass, and hence an *equation,* which one will have to solve, exactly or approximately.

To define admissible potentials, let's proceed by successive reductions. First, a broad enough class:

$$\Phi = \{\text{all } \varphi\text{'s continuous piecewise smooth (over the closure of D)}\}.$$

(If D was not, as here, bounded, we should add "such that $\int_D |\operatorname{grad} \varphi|^2$ is finite", in order to take care of the finite energy requirement. This is implicit in the present modelling, but should be kept in mind.) Next,

(27) $$\Phi^I = \{\text{all } \varphi \in \Phi : \varphi = 0 \text{ on } S^h_0 \text{ and } \varphi = I \text{ on } S^h_1\}$$

where I is just a real parameter for the moment. In particular, we shall have $\Phi^0 = \{\varphi \in \Phi : \varphi = 0 \text{ on } S^h\}$. If φ is in Φ^I for *some* value of I, it means that $n \times \operatorname{grad} \varphi = 0$ on S^h, and thus Eqs. (20) and (21) are satisfied by $h = \operatorname{grad} \varphi$, if φ is any of these potentials. Last, we select the given value of I, and now, if φ is in *this* Φ^I, (25) is satisfied.

Eligible potentials thus fulfill conditions (20), (21), and (25). To deal with the other conditions, we request b ($\equiv \mu \operatorname{grad} \varphi$) to satisfy (23) by using the weak solenoidality condition. But since *the set of test functions is left to our choice,* we may do better and also check (24), all in one stroke:

Proposition 2.3. *If* $\varphi \in \Phi^I$ *is such that*

(28) $$\int_D \mu \operatorname{grad} \varphi \cdot \operatorname{grad} \varphi' = 0 \quad \text{for all test functions } \varphi' \text{ in } \Phi^0,$$

then the field $b = \mu \operatorname{grad} \varphi$ *verifies* (23) *and* (24).

(Pay attention to the notational shift: Since from now on we shall have the eligible potentials on the one hand, and the test functions on the other hand, the latter will be denoted with a prime. This convention will be used throughout.)

Proof. Set $b = \mu \operatorname{grad} \varphi$. This is a piecewise continuous field. Since Φ^0 contains $C_0^\infty(D)$, we have $\operatorname{div} b = 0$ in the weak sense, as required. But since there are test functions in Φ^0 which do not belong to $C_0^\infty(D)$ (all those that do not vanish on S^b), the implications of (28) may not have been all derived. Starting from (28), and integrating by parts with formula

(9), we get

$$0 = \int_D \mathrm{b} \cdot \mathrm{grad}\, \varphi' = -\int_D \varphi' \,\mathrm{div}\, \mathrm{b} + \int_S \mathrm{n} \cdot \mathrm{b}\ \varphi' = \int_{S^b} \mathrm{n} \cdot \mathrm{b}\ \varphi' \ \ \forall\, \varphi' \in \Phi^0,$$

since $\mathrm{div}\, \mathrm{b} = 0$ a.e. and $\varphi' = 0$ on S^h *by our choice* of test functions. What is thus left is the following implication of (28):

$$\int_{S^b} \mathrm{n} \cdot \mathrm{b}\ \varphi' = 0 \quad \forall\, \varphi' \in \Phi^0,$$

which can be satisfied, since values of φ' are unconstrained on S^b, only by $\mathrm{n} \cdot \mathrm{b}$ vanishing on this part of the boundary. ◊

We are thus entitled to set a *problem:*

(29) *find* φ *in* Φ^I *such that* (28) *hold.*

This (mathematical) problem, more accurately described as an *equation,*[16] is *the weak formulation, in scalar potential,* of our (physical) problem. We just proved that if there is a solution, it will satisfy all the requirements of the modelling.

2.4.3 Uniqueness

No need to underline what this proof owes to that of Prop. 2.1. (Notice that the ideas of Remark 2.4 and Eq. (15) also have been exploited, to some extent.) But the serendipity by which Φ^0 happened to be the right space of test functions calls for an explanation, which Fig. 2.8 will suggest: In the linear space Φ, the Φ^Is form a family of *parallel* affine subspaces, and are thus all isomorphic with the vector subspace Φ^0. In particular the difference between two eligible potentials φ_1 and φ_2, being in Φ^0, qualifies as a test function.

Now, (29) can be construed as a system of linear equations, to be satisfied by φ, one equation for each test function engaged. Even though we are dealing here with infinite-dimensional spaces, and thus, so to speak, with an infinity of unknowns, the general rule of algebra that there should be "as many equations as unknowns" in a properly formed linear system is still in force: Fig. 2.8 shows that our choice of test functions obeys this

[16] In the more precise language of Appendix A, an equation is the problem consisting in finding all the values of the free variable in some predicate. Here the free variable is φ, and the predicate is (28); it consists of a *list* of subpredicates, indexed by the bound variable φ'. Note again the importance of the "for all" clause in (28) in this mechanism. Without it, we wouldn't have an equation, only nonsense.

rule automatically, thanks to the one-to-one correspondence between Φ^{I} and its parallel vector subspace.

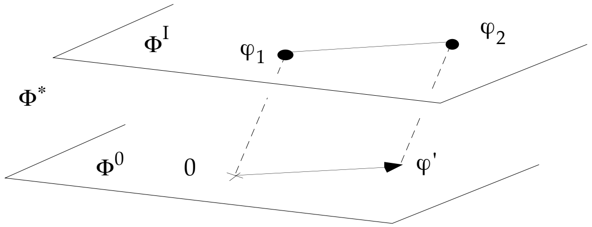


FIGURE 2.8. Geometry of the variational method. The "space" of the picture represents Φ, and the parallel "planes" represent Φ^{I} and Φ^{0}. The latter contains the origin. Dots and arrows signal points and vectors, respectively, in these infinite-dimensional spaces. Φ^* is an ad-hoc notation for the set union $\cup\{\Phi^{I} : I \in \mathbb{R}\}$, which does not fill out Φ.

This *proves* nothing yet, of course. But the heuristic principle thus suggested is of enormous value: *To find the weak form of a problem, set up the affine space of all a priori eligible solutions, then use the elements of the parallel vector subspace as test functions.*

This principle is quite flexible: "Eligible" depends on which equations and boundary conditions we can, and wish to, enforce a priori, and the others are automatically taken into account by weak formulation of the remaining requirements of the model (cf. Exer. 2.9). Here, we chose to enforce the equations relative to h (which is why this method can be depicted as "h-oriented"), but we might as well have focused on the equations relative to b, hence a b-oriented method (the opening move of it, of course, would be to introduce a *vector* potential, b = rot a). We'll do this in Chapter 6. There is also some leeway with the constant I, which was imposed here, but could have been left in charge of the weak formulation, as we shall see also.

As a first testimony of the power of the principle, let us prove this "uniqueness" result:

Proposition 2.4. *Problem* (29) *has at most one solution.*

Proof. Suppose there are two solutions φ_1 and φ_2. Then

$$(30) \qquad \int_D \mu \operatorname{grad}(\varphi_1 - \varphi_2) \cdot \operatorname{grad} \varphi' = 0 \quad \forall \ \varphi' \text{ in } \Phi^0.$$

But (see Fig. 2.8), $\varphi_1 - \varphi_2$ is one of the test functions, and for *that* one, (30) yields $\int_D \mu\ |\text{grad}(\varphi_1 - \varphi_2)|^2 = 0$, hence (cf. (18)) $\text{grad}(\varphi_1 - \varphi_2) = 0$, which means $\varphi_1 = \varphi_2$, since they coincide on S^h. ◊

Remark 2.7. This prompts the question, irrelevant here, but sensible in other circumstances, "What if S^h is empty?" Then, simply, the potential is not unique, but the field $h = \text{grad}\ \varphi$ is, which is generally what one is interested in. ◊

If this was linear algebra, Prop. 2.4 would solve the problem! For in finite dimension, *uniqueness forces existence,* as the old saying goes, when the number of equations and unknowns coincide. (**Exercise 2.7:** Why?) But here we deal with elements of an infinite-dimensional space, in which things are not that simple. Whether and when problem (29) has a solution, the *existence* issue, will be the concern of the next chapter. But before leaving the present one, something you may have been surprised to see de-emphasized:

2.4.4 Laplace, Poisson, Dirichlet, and Neumann

As a consequence of (23) and (24), the solution of (29) will satisfy

$$(31) \qquad - \text{div}(\mu\ \text{grad}\ \varphi) = 0 \quad \text{in } D,$$

(or at least, inside regions of regularity—but we shall stop reminding that all the time, from now on), and

$$(32) \qquad \varphi = 0 \ \text{ on } S^h{}_0, \quad \varphi = I \ \text{ on } S^h{}_1,$$

$$(33) \qquad \mu\ \partial_n \varphi = 0 \ \text{ on } S^b,$$

where $\partial_n\varphi$ is the notation in force here for the normal derivative of φ at the boundary, often denoted as $\partial\varphi/\partial n$. Equation (31) is an immediate generalization of the *Laplace equation* $\Delta\varphi = 0$, to which it reduces if $\mu = \mu_0$ in all D. The expression *Poisson problem* refers to (31) with a nonzero right-hand side, which we don't have here, but could easily handle (cf. (11) and Exer. 2.2). One calls (32) and (33) the *Dirichlet* and *Neumann boundary conditions,* respectively. Here the latter are homogeneous (right-hand side equal to 0), but non-homogeneous similar conditions can be accommodated by the above method, as we'll see later.

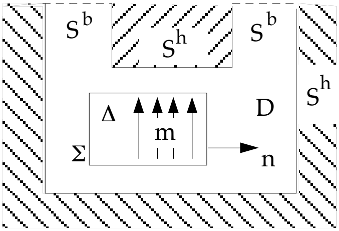
This Dirichlet vs Neumann opposition is classical and quite important, but here we should rather focus on the *fields* h and b than on the potentials, and the $n \times h$ vs $n \cdot b$ contrast is thus more topical. Also more important conceptually is the distinction between *essential* boundary conditions, like (32), which are built into the very definition of the set of admissible solutions, and *natural* conditions like (33), which are enforced by the weak formulation.

EXERCISES

Exercise 2.1 is on p. 34. Exers. 2.2 and 2.3 are on p. 44, and Exer. 2.4 p. 46. Exercises 2.5 and 2.6 are on p. 51, and Exer. 2.7 on p. 55.

Exercise 2.8. Suppose a part of D contains a permanent magnet, characterized by $b = \mu_0(h + m)$, where m is a given field, the rest of D being air. Show that the weak formulation, *find* $\varphi \in \Phi^I$ *such that*

(34) $$\int_D \mu_0 \operatorname{grad} \varphi \cdot \operatorname{grad} \varphi' = - \int_D \mu_0 m \cdot \operatorname{grad} \varphi' \quad \forall\, \varphi' \in \Phi^0,$$



is a correct interpretation of the problem. In case m is uniform over a region $\Delta \subset D$, and 0 outside (inset), show that (34) can be written

(35) $$\int_D \mu_0 \operatorname{grad} \varphi \cdot \operatorname{grad} \varphi' = \int_\Sigma n \cdot (\mu_0 m)\, \varphi' \quad \forall\, \varphi' \in \Phi^0,$$

where $\Sigma = \partial\Delta$.

Exercise 2.9. Consider the space $\Phi^* = \cup_I \Phi^I$ of Fig. 2.8. Each $\varphi \in \Phi^*$ belongs to one of the Φ^Is, so let’s define $\mathcal{J}$ as the map that assigns to φ the corresponding value of I. Let F be given. Show that if $\varphi \in \Phi^*$ satisfies

(36) $$\int_D \mu \operatorname{grad} \varphi \cdot \operatorname{grad} \varphi' = F\, \mathcal{J}(\varphi') \quad \forall\, \varphi' \in \Phi^*,$$

then $b = \mu \operatorname{grad} \varphi$ verifies (23), (24) and (26).

Exercise 2.10. Here the artificial boundary S^b has been placed in a position where it was known in advance that $n \cdot b = 0$. It may happen that $n \cdot b$ is thus known on some conveniently placed surfaces, but not null. What to do then?

HINTS

2.1. The definition doesn't say *bounded* regions. Recall that continuous functions are bounded on closed *bounded* (hence, compact) regions of a finite-dimensional space.

2.2. Just redo the proof of Prop. 2.1, reintroducing f at the right places. Observe the way a minus sign appears.

2.3. Cauchy–Schwarz. Observe (this is for experts) that a certain condition on the supports of the φ_ns should be satisfied.

2.4. Same as Exer. 2.2.

2.5. Build a circuit on which to apply Stokes.

2.6. Build a volume to which Ostrogradskii–Gauss may apply, part of its surface being S^h_1.

2.7. The question is, if $\mathbf{A}$ is an $n \times n$ matrix, and $\mathbf{b}$ an n-vector, why does uniqueness of $\mathbf{x}$ such that $\mathbf{Ax}=\mathbf{b}$ imply the existence of a solution, whatever the right-hand side?

2.8. All that matters is $\operatorname{div} b = 0$, where $b = \mu_0 (m + \operatorname{grad} \varphi)$, and the proof of Prop. 2.3 handles that. For (35) vs (34), apply (9) to Δ.

2.9. Φ^* being *larger* than Φ^0, the proof of Prop. 2.3 can be recycled in its entirety, hence (23) and (24). So concentrate on (26), using (9).

2.10. Imitate (35), which can be understood as describing a flux injection $n \cdot (\mu_0 m)$ on Σ.

SOLUTIONS

2.1. The constant 1 is not integrable over all E_3, so the restriction to bounded domains is certainly necessary. Now suppose f is smooth over each region R_i of a finite family. The way we understand "over", f is smooth, and in particular continuous, in a bounded domain D_i containing the closure of $R_i \cap D$, which is thus compact, so f is bounded there, hence integrable. Pieces being in finite number, f is integrable on D.

2.3. Since $\operatorname{supp}(\varphi)$ is compact, one can build the φ_ns so that there exists a compact K containing all the $\operatorname{supp}(\varphi_n)$. Then (Exer. 2.1) b is bounded on

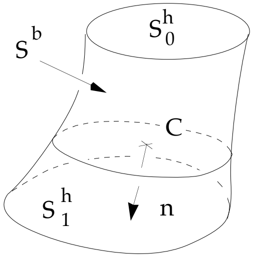
K, and hence, applying the Cauchy–Schwarz inequality,

$$|\int_D b \cdot \text{grad}(\varphi - \varphi_n)|^2 \le \int_D |b|^2 \int_D |\text{grad}(\varphi - \varphi_n)|^2$$

tends to zero.

2.5. Take c_1 and c_2 from S^h_0 to S^h_1, with the same orientation, and join the extremities by two paths lying in S^h_0 and S^h_1 respectively, in order to make a closed circuit. As rot h = 0, and by the Stokes theorem, the circulations along c_1 and c_2 are equal (those along the boundary links are 0 by (21)). One says that c_1 and c_2 are *homologous.* (The relation between them is an equivalence, called *relative homology modulo* S^h. We'll have more to say about this in Chapter 5. Cf. [GH].)

2.6. See the inset. Surface C is what is commonly called a "cut": Its boundary is entirely in S^b, and it separates D into two parts, each containing one piece of S^h. Moreover, C has an external orientation (provided by a normal field n), compatible with that of S^h_1. Now, as $n \cdot b = 0$ on S^b, the fluxes through C and S^h_1 are equal, by Ostrogradskii, since div b = 0 in D. All possible cuts of this kind will do, including S^h_1 and S^h_0, *but* the latter must be oriented the other way with respect to S. Again we have here an equivalence relation (relative homology, but now modulo S^b), and "cuts" are elements of a same class of surfaces, of which one says they are *homologous* (mod S^b). We'll return to this in Chapter 4, and again, more formally, in 5.2.5.



2.7. Because then **A** is regular.

2.8. $\int_D m \cdot \text{grad}\,\varphi' = \int_\Delta m \cdot \text{grad}\,\varphi' = -\int_\Delta \text{div}\, m\ \varphi' + \int_\Sigma n\,.\,m\varphi'$. If $\text{div}\, m \neq 0$, there is no special avantage to this formulation over (34), but otherwise (35) may be easier to implement in the subsequent finite element modelling. Be careful about the correct orientation of the normal on Σ when doing that. (You may worry about what happens when Δ touches S. This is a good question, but no more a simple exercise.)

2.9. By (9), and using the information brought by the proof of Prop. 2.3 ($\text{div}\, b = 0$, $n \cdot b = 0$ on S^b), plus $\varphi' = 0$ on S^h_0, (36) reduces to

$$(37) \qquad F\,\mathcal{J}(\varphi') = \int_{S^h} n \cdot b\ \varphi' \equiv \int_{S^h_1} n \cdot b\ \varphi' \quad \forall\, \varphi' \in \Phi^*,$$

and since the value of φ' on S^h_1 is precisely $\mathcal{J}(\varphi')$, by definition of $\mathcal{J}$, we have $F\,\mathcal{J}(\varphi') = (\int_{S^h} n \cdot b)\,\mathcal{J}(\varphi')$ for all φ', hence $\int_{S^h} n \cdot b = F$.

Now, φ being known, $\mathcal{J}(\varphi)$ has a definite value I, and the desired reluctance is $R = I/F$. This trick, by which the essential boundary constraint $\varphi = I$ on S^h_1 (condition (25)) has been exchanged for the natural boundary condition (37), is known as the *dualization* of the constraint (25).

2.10. Let g be the known value of $n \cdot b$ on S^b. Prolongate g to all S by setting $g = 0$ (or any value, it doesn't matter) on S^h. The relevant weak formulation is *find* $\varphi \in \Phi^I$ *such that*

$$\int_D \mu \operatorname{grad} \varphi \cdot \operatorname{grad} \varphi' = \int_S g\, \varphi' \quad \forall\, \varphi' \in \Phi^0.$$

Indeed, (9) yields $\int_S n \cdot b \ \varphi' = \int_S g\, \varphi' \ \ \forall\, \varphi' \in \Phi^0$, therefore $n \cdot b = g$ on S^b, where the values of φ' are free.

REFERENCES

[Ar] V. Arnold: **Équations différentielles ordinaires**, Mir (Moscow), 1974.

[Bo] A. Bossavit: "On the condition 'h normal to the wall' in magnetic field problems", **Int. J. Numer. Meth. Engng., 24** (1987), pp. 1541–1550.

[B5] A. Bossavit: "Results for benchmark problem 5, the Bath-cube experiment: An aluminium block in an alternating field", **COMPEL, 7,** 1–2 (1988), pp. 81–88.

[Br] F.E. Browder (ed.): **Mathematical developments arising from Hilbert Problems,** AMS (Providence, R.I.), 1976.

[CL] E.A. Coddington, N. Levinson: **Theory of Ordinary Differential Equations,** McGraw-Hill (New York), 1955.

[DB] J.A.M. Davidson, M.J. Balchin: "Experimental Verification of Network Method for Calculating Flux and Eddy-Current Distributions in Three Dimensions", **IEE Proc., 128, Pt. A** (1981), pp. 492–496.

[Fr] K.O. Friedrichs: **Advanced Ordinary Differential Equations**, Gordon and Breach (New York), 1965.

[GH] M.J. Greenberg, J.R. Harper: **Algebraic Topology, A First Course**, Benjamin/Cummings (Reading, MA), 1981.

[Hö] L. Hörmander: "On the interior regularity of the solutions of partial differential equations", **Comm. Pure & Appl. Math., 11** (1958), pp. 197–218.

[LS] W.D. Lakin, D.A. Sanchez: **Topics in Ordinary Differential Equations**, Prindle, Weber & Schmidt (Boston), 1970 (Dover edition, New York, 1982).

[Pe] J. Peetre: "A Proof of the Hypoellipticity of Formally Hypoelliptic Differential Operators", **Comm. Pure & Appl. Math., 14** (1961), pp. 737–744.

[T&] L.R. Turner, K. Davey, C.R.I. Emson, K. Miya, T. Nakata, A. Nicolas: "Problems and workshops for eddy current code comparison", **IEEE Trans., MAG-24,** 1 (1988), pp. 431–434.

The differential geometric point of view alluded to in Subsection 2.2.3 will probably soon gain popularity. Most introductions to differential geometry have a chapter on Electromagnetism, e.g.:

[Bu] W.L. Burke: **Applied Differential Geometry**, Cambridge University Press (Cambridge, U.K.), 1985.

[Sc] B. Schutz: **Geometrical methods of mathematical physics**, Cambridge University Press (Cambridge, U.K.), 1980.

[We] S. Weintraub: **Differential Forms, A Complement to Vector Calculus,** Academic Press (San Diego), 1997.

However, few book-size treatments are available so far:

[Ko] P.R. Kotiuga: **Hodge Decompositions and Computational Electromagnetics** (Thesis), Department of Electrical Engineering, McGill University (Montréal), 1984.

[BH] D. Baldomir, P. Hammond: **Geometry of Electromagnetic Systems,** Oxford U.P. (Oxford), 1996.

The idea by itself is not new, and several authors have devoted work to its promotion. See:

[Bn] F.H. Branin, Jr.: "The algebraic-topological basis for network analogies and the vector calculus", in **Symposium on Generalized Networks** (12–14 April 1966)**,** Polytechnic Institute of Brooklyn , 1966, pp. 453–491.

[BH] D. Bohm, B.J. Hiley, A.E.G. Stuart: "On a new mode of description in physics", **Int. J. Theoret. Phys., 3**, 3 (1970), pp. 171–183.

[To] E. Tonti: "On the Geometrical Structure of Electromagnetism", in **Gravitation, Electromagnetism and Geometrical Structures** (G. Ferrarese, ed.)**,** Pitagora (Bologna), 1996, pp. 281–308.

I do think this differential geometric approach, far from being merely an esthetically attractive alternative, is mandatory when it comes to the question (not addressed here) of force computation in deformable materials. Cf.:

A. Bossavit: "Differential forms and the computation of fields and forces in Electromagnetism", **Europ. J. Mech., B/Fluids, 10**, 5 (1991), pp. 474–488.

A. Bossavit: "Edge-element Computation of The Force Field in Deformable Bodies", **IEEE Trans., MAG-28,** 2 (1992), pp. 1263–1266.

A. Bossavit: "On local computation of the force field in deformable bodies", **Int. J. Applied Electromagnetics in Materials**, **2**, 4 (1992), pp. 333–343.

CHAPTER 3

Solving for the Scalar Magnetic Potential

3.1 THE "VARIATIONAL" FORMULATION

We now treat the problem we arrived at for what it is, an *equation*, to be studied and solved as such: Given a bounded domain D, a number I (the mmf), and a function μ (the permeability), subject to the conditions $0 < \mu_0 \leq \mu \leq \mu_1$ of Eq. (18) in Chapter 2,

$$\text{(1)}\quad \left|\begin{array}{l} \textit{find } \varphi \in \Phi^I = \{\varphi \in \Phi : \ \varphi = 0 \text{ on } S^h{}_0, \ \varphi = I \text{ on } S^h{}_1\} \textit{ such that} \\ \quad \int_D \mu \operatorname{grad} \varphi \cdot \operatorname{grad} \varphi' = 0 \quad \forall\ \varphi' \in \Phi^0. \end{array}\right.$$

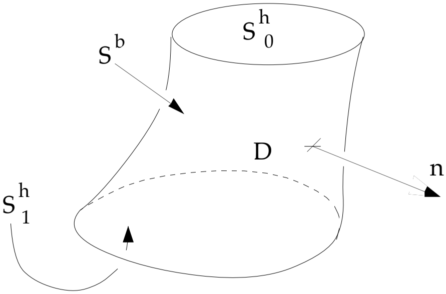


FIGURE 3.1. The situtation, reduced to its meaningful geometrical elements.

All potentials φ and test functions φ' belong to the encompassing linear space Φ of piecewise smooth functions on D (cf. 2.4.2), and the geometrical elements of this formulation, surface $S = S^h \cup S^b$, partition $S^h = S^h{}_0 \cup S^h{}_1$ of the "magnetic wall" S^h (Fig. 3.1), are all that we abstract from the concrete situation we had at the beginning of Chapter 2. We note that the magnetic energy (or rather, coenergy, cf. Remark 2.6) of

$h = \text{grad}\ \varphi$, that is,

$$F(\varphi) = \tfrac{1}{2} \int_D \mu\ |\text{grad}\ \varphi|^2,$$

is finite for all elements of Φ. The function F, the type of which is *FIELD → REAL*, and more precisely, $\Phi \to \mathbb{R}$, is called the *(co)energy functional.*

Remark 3.1. The use of the quaint term "functional" (due to Hadamard), not as an adjective here but as a somewhat redundant synonym for "function", serves as a reminder that the argument of F is not a simple real- or vector-valued variable, but a point in a space of infinite dimension, representative of a field. This is part of the "functional" point of view advocated here: One *may* treat complex objects like fields as mere "points" in a properly defined functional space. ◊

Function F is quadratic with respect to φ, so this is an analogue, in infinite dimension, of what is called a *quadratic form* in linear algebra. Quadratic forms have associated polar forms. Here, by analogy, we define the *polar form* of F as $\mathcal{F}(\varphi, \psi) = \int_D \mu\ \text{grad}\ \varphi \cdot \text{grad}\ \psi$, a bilinear function of two arguments, that reduces to F, up to a factor 2, when both arguments take the same value.

The left-hand side of (1) is thus $\mathcal{F}(\varphi, \varphi')$. This cannot be devoid of significance, and will show us the way: In spite of the dimension being infinite, let us try to apply to the problem at hand the body of knowledge about quadratic forms. There is in particular the following trick, in which only the linearity properties are used, not the particular way F was defined: For any real λ,

(2) $$0 \le F(\varphi + \lambda\psi) = F(\varphi) + \lambda\ \mathcal{F}(\varphi, \psi) + \lambda^2\ F(\psi) \quad \forall\ \psi \in \Phi.$$

One may derive from this, for instance, the Cauchy–Schwarz inequality, by noticing that the discriminant of this binomial function of λ must be nonnegative, and hence

$$\mathcal{F}(\varphi, \psi) \le 2\ [F(\varphi)]^{1/2}\ [F(\psi)]^{1/2},$$

with equality only if $\psi = a\varphi + b$, with a and b real, $a \ge 0$. Here we shall use (2) for a slightly different purpose:

Proposition 3.1. *Problem* (1) *is equivalent to*

(3) *Find* $\varphi \in \Phi^I$ *such that* $F(\varphi) \le F(\psi)\ \ \forall\ \psi \in \Phi^I$,

the coenergy minimization *problem.*

Proof. Look again at Fig. 2.8, and at Fig. 3.2 below. If φ solves (3), then $F(\varphi) \leq F(\varphi + \lambda\varphi')$ for all φ' in Φ^0, hence $\lambda\,\mathcal{F}(\varphi, \varphi') + \lambda^2 F(\varphi') \geq 0$ for all $\lambda \in \mathbb{R}$, which implies (the discriminant, again[1]) that $\mathcal{F}(\varphi, \varphi') = 0$ for all φ' in Φ^0, which is (1). Conversely, if φ solves (1), and ψ belongs to Φ^I, then (cf. (2)) $F(\psi) = F(\varphi) + \mathcal{F}(\varphi, \psi - \varphi) + F(\psi - \varphi) = F(\varphi) + F(\psi - \varphi) \geq F(\varphi)$, since $\psi - \varphi \in \Phi^0$ and $F(\psi - \varphi) \geq 0$. ◊

This confirms our intuitive expectation that the physical potential should be the one, among all eligible potentials, that minimizes the coenergy. Problem (3) is called the *variational form* of the problem. In the tradition of mathematical physics, a problem has been cast in variational form when it has been reduced to the minimization of some function subject to some definite conditions, called "constraints". The constraint, here, is that φ must belong to the affine subspace Φ^I (an *affine constraint*, therefore). Such problems in the past were the concern of the calculus of variations, which explains the terminology. Nowadays, Problem (1) is often described as being "in variational form", but this is an abuse of language, for such a weak formulation does not necessarily correspond to a minimization problem: In harmonic-regime high-frequency problems, for instance, a complex-valued functional is stationarized, not minimized. For the sake of definiteness, I'll refer to (1) as "the weak form" and to (3) as "the variational form".

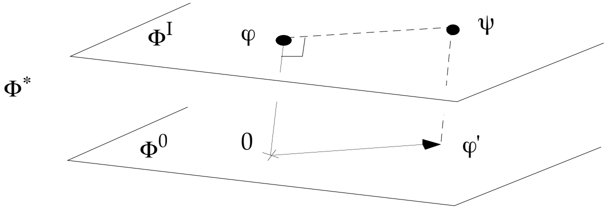


FIGURE 3.2. Geometry of the variational method ($\psi = \varphi + \varphi'$).

Conversely, however, variational problems with affine constraints have as a rule a weak form, which can be derived by consideration of the *directional derivative*[2] of F at point φ. By definition, the latter is the linear map

(4) $$\psi \rightarrow \lim_{\lambda \rightarrow 0} [F(\varphi + \lambda\psi) - F(\varphi)]/\lambda.$$

If φ yields the minimum, the directional derivative of F should vanish

[1] Alternatively, first divide by λ, then let λ go to 0.

[2] Known as the *Gâteaux derivative*.

at φ, for all directions that satisfy the constraint. The condition obtained that way is called the *Euler equation* of the variational problem.

Here, (4) is the map $\psi \rightarrow \mathcal{F}(\varphi, \psi)$, after (2). Therefore, Problem (1), which expresses the cancellation of this derivative in all directions parallel to Φ^0, is the Euler equation of the coenergy minimization problem (3).

Exercise 3.1. Find variational forms for Problems (2.34) and (2.36).

In the space Φ^* of the last chapter (Exer. 2.9), which is visualized as ordinary space in Fig. 3.2, we may define a *norm,* $\|\varphi\|_\mu = (2F(\varphi))^{1/2} \equiv [\int_D \mu\ |\text{grad}\,\varphi|^2]^{1/2}$, hence a notion of distance: The *distance in energy* of two potentials is $d_\mu(\varphi, \psi) = \|\varphi - \psi\|_\mu \equiv [\int_D \mu\ |\text{grad}(\varphi - \psi)|^2]^{1/2}$. The variational problem can then be described as the search for this potential in Φ^I that is closest to the origin, in energy: in other words, the *projection* of the origin on Φ^I.

Moreover, this norm stems from a scalar product, which is here, by definition, $(\varphi, \psi)_\mu = \int_D \mu \,\text{grad}\,\varphi \cdot \text{grad}\,\psi$ ($\equiv \mathcal{F}(\varphi, \psi)$, the polar form), with $\|\varphi\|_\mu = [(\varphi, \varphi)_\mu]^{1/2}$. The weak form also then takes on a geometrical interpretation: It says that vector φ is orthogonal to Φ^0, which amounts to saying (Fig. 3.2) that point φ is the *orthogonal projection* of the origin on Φ^I. The relation we have found while proving Prop. 3.1,

$$(5) \qquad F(\psi) = F(\varphi) + F(\psi - \varphi) \quad \forall\, \psi \in \Phi^I,$$

if φ is the solution, then appears as nothing but the Pythagoras theorem, in a functional space of infinite dimension.

Exercise 3.2. Why the reference to Φ^*, and not to Φ ?

This is our first encounter with a *functional space:* an affine space, the elements of which can usually be interpreted as functions or vector fields, equipped with a notion of distance. When, as here, this distance comes from a scalar product on the associated vector space, we have a *pre-Hilbertian* space. (Why "pre" will soon be explained.) The existence of this metric structure (scalar product, distance) then allows one to speak with validity of the "closeness" of two fields, of their orthogonality, of converging sequences, of the continuity of various mappings, and so forth. For instance (and just for familiarization, for this is a trivial result), if we call $\varphi(I)$ the solution of (1) or (3), considered as a function of the mmf I, we have

Proposition 3.2. *The mapping* $I \rightarrow \varphi(I)$ *is continuous in the energy metric.*

Proof. By (1), $\varphi(I) = I\,\varphi(1)$, hence $\|\varphi(I)\|_\mu = |I|\ \|\varphi(1)\|_\mu$, that is, $\|\varphi(I)\|_\mu \leq$

$C\,|I|$ for all I, where C is a constant, thus satisfying the criterion for continuity of linear operators. ◊

Exercise 3.3. Show that $\mathcal{J}$ (notation of Exer. 2.9) is continuous on Φ^*.

As for the functional point of view, also heralded before, we now have a good illustration of it: Having built a functional space of eligible potentials, we search for a distinguished one, here the orthogonal projection of the origin on Φ^I.

Remark 3.2. Once we have this solution φ, then, by the integration by parts formula, $\int_D \mu\,|\text{grad}\,\varphi|^2 = \int_S n\cdot b\ \varphi = I\int_{S^h_1} n\cdot b = I^2/R$, by definition of the reluctance R (cf. 2.4.1). So, finding the magnetic coenergy will give access to R. We'll return to this in Chapter 4 (Subsection 4.1.3). ◊

3.2 EXISTENCE OF A SOLUTION

After this promising commencement, the bad news: Problems (1) or (3) *may fail to have a solution.*

3.2.1 Trying to find one

Call d the distance of the origin to Φ^I, that is, $d = \inf\{||\psi||_\mu :\ \psi\in\Phi^I\}$. For each integer n, there certainly exists some φ_n in Φ^I such that $||\varphi_n||_\mu \le d + 1/n$. (Otherwise, d would be lower than the infimum.) Moreover, $d = \lim_{n\to\infty} ||\varphi_n||_\mu$. One says that the φ_ns form a *minimizing sequence,* which we may expect to converge towards a limit φ. If so, this limit will be the solution.

Indeed, by developing $||\varphi_n \pm \varphi_m||_\mu^2 = \int_D \mu\,|\text{grad}(\varphi_n \pm \varphi_m)|^2$, we have

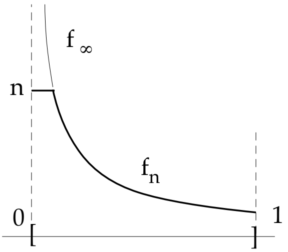
$$||\varphi_n - \varphi_m||_\mu^2 + ||\varphi_n + \varphi_m||_\mu^2 = 2(||\varphi_n||_\mu^2 + ||\varphi_m||_\mu^2)\,.$$

The point $(\varphi_n + \varphi_m)/2$ belongs to Φ^I, so its distance to 0 is no smaller than d; therefore $||\varphi_n + \varphi_m||_\mu^2 \ge 4d^2$, and hence,

$$(6)\qquad ||\varphi_n - \varphi_m||_\mu^2 \le 2(||\varphi_n||_\mu^2 + ||\varphi_m||_\mu^2) - 4d^2.$$

Now, let n and m tend to infinity; the right-hand side tends to 0, so $||\varphi_n - \varphi_m||_\mu$ tends to 0; this qualifies $\{\varphi_n :\ n\in\mathbb{N}\}$ as a *Cauchy sequence,* which a converging sequence must be (cf. Appendix A, Subsection A.4.1).

But this necessary condition is not sufficient. Just as the set of rational numbers does not contain all the limits of its Cauchy sequences, functional spaces which do not contain the limits of their own abound. For instance (inset), in the space of piecewise smooth functions over $[0, 1]$, equipped with the norm $\|f\| = \int_0^1 |f(x)|\, dx$, the sequence $f_n = x \to \inf(n, 1/\sqrt{x})$ is Cauchy (**Exercise 3.4**: prove it), but the would-be limit $x \to 1/\sqrt{x}$ is *not* piecewise smooth. Hence the necessity of the following definition:



Definition 3.1. *A metric space* X *is* complete *if all Cauchy sequences in* X *converge towards an element of* X.

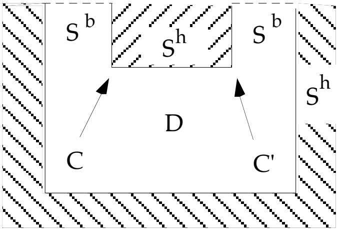
In particular, a complete normed space is called a *Banach space*, and a complete pre-Hilbertian space is called a *Hilbert space*.

If our underlying space Φ^* was complete (and then each slice Φ^I, being closed by Exer. 3.3, would be complete), the above reasoning would thus establish the existence of a solution to (1) or (3), which we already know is unique. But Φ^* is not complete, as counterexamples built on the same principle as in Exer. 3.4 will show. What this points to, however, is a failure of our *method*. Conceivably, a piecewise smooth solution could exist in spite of our inability to prove its existence a priori by this minimizing sequence approach. One might even be tempted to say, "Never mind, we know this solution exists on physical grounds, and we shall be content with an approximation (that is, an element of high enough rank of some minimizing sequence). Moreover, didn't we *prove*, with this Weyl lemma, that φ would be smooth in homogeneous regions? If so the present space Φ, though not complete, is rich enough. So let's proceed and focus on finding a usable approximation."

Such a stand would not be tenable. First, there is a logical fallacy: smoothness was proved, in Chapter 1, but in case the solution *exists*, which is what we want to assess. Besides, you don't prove something "on physical grounds". Rather, modelling sets up a correspondence between a segment of reality and a mathematical framework, by which some empirical *facts* have mathematical *predicates* as counterparts. The truth of such predicates must be proved *within the model*, and failure to achieve that just invalidates *the modelling*. So the responsibility of asserting the existence of a solution to (1) or (3), within this mathematical framework, is ours.

Alas, not only can't we prove piecewise smoothness a priori, but we can build counterexamples, that correspond to quite realistic situations. We shall display one.

3.2.2 Φ^* is too small



Refer to Fig. 2.6 (recalled in inset), and imagine the system as so long in the z–direction that all field lines are in the x–y plane, which makes a 2D modelling feasible. It is well known that the field will be infinite at the tips of the "re-entrant corners" C and C' with such geometry. (This is the same phenomenon as the "spike effect" in electrostatics.) By doing Exercises 5 and 6, you should be able to see why: An analogue of Problems (1) and (3), in an appropriately simplified two-dimensional setting, can be solved in closed form, and its solution exhibits a mild singularity at the origin (which corresponds to corner C). The potential is well-behaved (cf. Fig. 3.6, p. 89), but its gradient becomes infinite at the origin, in spite of the magnetic coenergy[3] being bounded.

This is an idealization, but it points to an unacceptable weakness of our modelling: The restriction to piecewise smooth potentials, which seemed quite warranted, bars the existence of such mild singularities,[4] whereas physics requires they be accounted for, as something that can happen. Our space is too small: The frame is too narrow.

Of course, we could blame this failure on too strict a definition of smoothness, and revise the latter in the light of new data, contriving to accept mildly singular fields as "smooth" according to some new, looser definition. But first, this kind of "monster-barring" [La] would lead to even more technical concepts and (likely) to something more esoteric than the radical solution we shall eventually adopt. Moreover, it might be only the beginning of an endless process: One may easily imagine how fractal-like boundaries, for instance, could later be invoked to invalidate our attempts to deal with corners.

The radical (and right) solution is *completion:* Having a non-complete functional space, immerse it into a larger, complete space. Then the above method works: The solution exists, in the completed space, as the limit of a minimizing sequence. (All of them will yield the same limit.)

[3] In such a 2D modelling, the μ-norm corresponds to the (co)energy contributed by the region of space lying between two horizontal planes, one unit of length apart.

[4] Precisely: the singularity at 0 makes it impossible to extend φ to a domain that would contain the origin and where its gradient would be finite, which is required by 1-smoothness "over" D, as we defined it.

3.2.3 Completing Φ^*

Completion logically belongs to the mathematical Appendix of this book, but the idea is so important, and so germane to what physicists do spontaneously when they define "generalized solutions" to problems which have no "classical" ones, that it may be worthwhile to discuss it here.

First, note this is not the same thing as *closure*. Indeed, if A is a part of a metric space $\{X, d\}$, sequences which fail to converge in A may converge to an element of its closure, so if X is complete, the completion of A will be its closure $\overline{A}$. But there, A is already immersed in a pre-existing metric space. If such an encompassing complete space does not yet exist, we can't proceed that way.

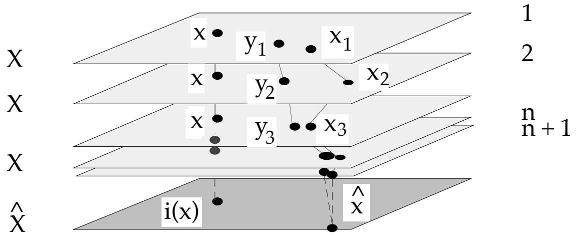


FIGURE 3.3. The idea of completion. $\hat{X}$ is an abstract space: its elements are of a different type that those of X. But there is a natural injection of X into $\hat{X}$, so the process can be seen, very informally, as "plugging the holes" in X.

The idea (Fig. 3.3) conforms with the usual method for building new mathematical objects from old ones: define equivalence relations (cf. A.1.6), and take equivalence *classes*. This is how, one will remember, rational numbers are built from pairs of integers, and real numbers from sequences of rationals. Completion is quite analogous to the construction of $\mathbb{R}$ in this respect. Suppose $\{X, d\}$ is a metric space, that is, a set X equipped with a distance d. Let $X°$ be the set of all Cauchy sequences in X. Two elements of $X°$, say $x° = \{x_1, x_2, \ldots, x_n, \ldots\}$ and $y° = \{\ldots, y_n, \ldots\}$, will be deemed equivalent if $\lim_{n \to \infty} d(x_n, y_n) = 0$ (Fig. 3.3). That is easily seen to be an equivalence relation, under the hypothesis that we are dealing with Cauchy sequences. We thus consider the quotient $\hat{X}$, we give it a distance, $\hat{d}(\hat{x}, \hat{y}) = \lim_{n \to \infty} d(x_n, y_n)$, where $\{x_n\}$ and $\{y_n\}$ are representatives of the classes $\hat{x}$ and $\hat{y}$, and we define the newly found metric space $\{\hat{X}, \hat{d}\}$ as *the completion of* X. This is a bold move, for the elements of $\hat{X}$, being sets of sequences of elements of X, seem of a completely different nature than those of X. But there is a natural way to inject X into $\hat{X}$: to $x \in X$, associate the class $\hat{x} = i(x)$ of the constant sequence $x^\circ = \{x, x, \dots, x, \dots\}$. This way, $\hat{X}$ appears as an extension of X (note that $\hat{d}$ restricts to d on the image i(X) of X under the injection i). Moreover, as proved in A.4.1, $\{\hat{X}, \hat{d}\}$ is complete, and X, or rather its image i(X), is dense in $\hat{X}$.

This mechanism does not guarantee that the completion of a functional space will be a functional space: its elements being equivalence classes of sequences of functions, some of these classes might not be identifiable[5] with any classically defined function. As a rule, one must invoke other mathematical theories to establish the functional nature of the elements of the completion—when such is the case.

The classical example is $L^2(D)$, the prototypal Hilbert space: $L^2(D)$ is defined as the completion of[6] $C_0^\infty(D)$ with respect to the norm $||f|| = (\int_D |f|^2)^{1/2}$. A central result of Lebesgue integration theory, then, is that $L^2(D)$ coincides with the space of square-integrable functions over D, or rather, of equivalence *classes* of such functions, with respect to the "a.e. =" relation (equality except on a negligible set). If this sounds complex, it's because it really is . . . (see Appendix A, Subsection A.4.2). Fortunately, this complexity can be circumscribed, and once in possession of $L^2(D)$, and of its analogue $\mathbb{L}^2(D)$ (square integrable vector fields), completion is an easy task, as we shall see later.

Completion corresponds to a very natural idea in physics. Many problems are idealizations. For instance, there is no such thing in nature as a sharp corner, but the sharp corner idealization helps understand what happens near a surface with high curvature. In this respect, the whole *family* of solutions, parameterized by curvature, contains information that one solution for a finite curvature would fail to give. This information is summarized by the singular solution, which belongs to the completion, because an element of the completion *is*, in the sense we have seen, a sequence of smooth solutions.

[5] This is no hair-splitting: For instance, the completion of $C_0^\infty(E_2)$ with respect to the norm $\varphi \to (\int_{E_2} |\text{grad } \varphi|^2)^{1/2}$ is *not* a functional space, not even a space of distributions [DL]. Still, this *Beppo Levi space* is home to electric or magnetic potentials in 2D problems. This reflects the intrinsic difficulty of dimension 2, for in 3D, Beppo Levi's space is functional, being continuously injectable in the space $L^6(E_3)$ of functions with integrable sixth power. We'll see that in Chapter 7.

[6] Note that, since a space is dense in its completion, spaces in which $C_0^\infty(D)$ is dense, with respect to the same quadratic norm, have the same completion. So we would obtain the same result, $L^2(D)$, by starting from $C^1(\overline{D})$, or $C^0(\overline{D})$, or for that matter, from the space of piecewise smooth functions over D.

A little more abstractly, suppose the problem has been cast in the form $Ax = b$, where b symbolizes the data, x the solution, and A some mapping of type $SOLUTION \rightarrow DATA$. Solving the problem means, at a high enough level of abstraction, finding[7] the inverse A^{-1}, which may not be defined for some values of b (those corresponding to sharp corners, let's say, for illustration). But if there is a solution x_n for each element b_n of some sequence that converges toward b, it's legitimate to define the limit $x = \lim_{n \rightarrow \infty} x_n$ as the solution, if there is such a limit, and *if there isn't, to invent one.* That's the essence of completion. Moreover, attributing to Ax the value b, whereas A did not make sense, a priori, for the *generalized solution* x, constitutes a prolongation of A beyond its initial domain, a thing which goes along with completion (cf. A.2.3). Physicists made much mileage out of this idea of a generalized solution, as the eventual limit of a parameterized family, before the concepts of modern functional analysis (complete spaces, distributions, etc.) were elaborated in order to give it status.

Summing up: We now attribute the symbol Φ^* to the completion of the space of piecewise smooth functions in D, null on $S^h{}_0$ and equal to some constant on $S^h{}_1$, with respect to the norm $||\varphi||_\mu = [\int_D \mu \ |\text{grad } \varphi|^2]^{1/2}$. Same renaming for Φ^I (which is now the closure of the previous one in Φ^*). Equation (1), or Problem (3), has now a (unique) solution. The next item in order[8] is to *solve* for it.

3.3 DISCRETIZATION

But what do we mean by that? Solving an equation means being able to answer specific questions about its solution with controllable accuracy, whichever way. A century ago, or even more recently in the pre-computer era, the only way was to represent the solution "in closed form", or as the sum of a series, thus allowing a numerical evaluation with help of formulas and tables. Computers changed this: They forced us to work from the outset with *finite* representations. Eligible fields and solutions must

[7] An unpleasantly imprecise word. What is required, actually, is some *representation* of the inverse, by a formula, a series, an algorithm . . . anything that can give *effective* access to the solution.

[8] Whether Problem (3) is well posed (cf. Note 1.16) raises other issues, which we temporarily bypass, as to the continuous dependence of φ on data: on I (Prop. 3.2 gave the answer), on μ (cf. Exers. 3.17 and 3.19), on the dimensions and shape of the domain (Exers. 3.18 and 3.20).

therefore be parameterized, with a perhaps very large, but finite, number of parameters.

3.3.1 The Ritz–Galerkin method

Suppose we have a finite catalog of elements of Φ, $\{\lambda^i : i \in \mathcal{J}\}$, often called *trial functions*, where $\mathcal{J}$ is a (finite) set of indices. Each λ^i must be a simple function, one which can be handled in closed form. If we can find a family of real parameters $\{\boldsymbol{\varphi}_i : i \in \mathcal{J}\}$ such that $\sum_i \boldsymbol{\varphi}_i \lambda^i$ is an approximation of the solution, this will be enough to answer questions the modelling was meant to address, provided the approximation is good enough, because all the data-processing will be done via the λ^is. The parameters $\boldsymbol{\varphi}_i$ (set in bold face) are called the *degrees of freedom* (abbreviated as DoFs or DoF, as the case may be) of the field they generate. We shall denote by $\boldsymbol{\varphi}$, bold also, the family $\boldsymbol{\varphi} = \{\boldsymbol{\varphi}_i : i \in \mathcal{J}\}$.

The Ritz–Galerkin idea consists in restricting the search for a field of least energy to those of the form $\sum_{i \in \mathcal{J}} \boldsymbol{\varphi}_i \lambda^i$ that belong to Φ^I. The catalog of trial functions is then known as a *Galerkin basis*. (We shall say that it defines an *approximation method*, and use the subscript m to denote all things connected with it, when necessary; most often, the m will be understood.) This is well in the line of the above constructive method for proving existence, for successive enlargements of the Galerkin basis will generate a sequence with *decreasing* energy, and if, moreover, this is a minimizing sequence, the day is won.

To implement this, let us introduce some notation: Φ_m is the finite dimensional space of linear combinations of functions of the catalog, that is to say, the space spanned by the λ^is, and we define

$$(7) \qquad \Phi^I_m = \Phi_m \cap \Phi^I, \qquad \Phi^0_m = \Phi_m \cap \Phi^0.$$

The approximate problem is thus:

$$(8) \qquad \textit{Find } \varphi_m \in \Phi^I_m \textit{ such that } \mathrm{F}(\varphi_m) \le \mathrm{F}(\psi) \ \ \forall\, \psi \in \Phi^I_m .$$

This problem has a solution (by the compactness argument of A.2.3), since we are considering here a positive definite quadratic form on a *finite*-dimensional space. If in addition we assume that Φ^I_m and Φ^0_m are parallel, just as Φ^I and Φ^0 were in Fig. 3.2 (this is not automatic, and depends on a sensible choice of trial functions), then (8) is equivalent, by exactly the same reasoning we made earlier, to the following Euler equation:

(9) *Find* $\varphi_m \in \Phi^{\mathrm{I}}{}_m$ *such that* $\int_D \mu \operatorname{grad} \varphi_m \cdot \operatorname{grad} \varphi' = 0 \;\; \forall\; \varphi' \in \Phi^0{}_m$.

This parallelism condition is usually easy to achieve: It is enough that some specific combination $\varphi^{\mathrm{I}}{}_m = \sum_{\mathrm{i}} \boldsymbol{\varphi}_{\mathrm{i}}{}^{\mathrm{I}} \lambda^{\mathrm{i}}$ satisfy $\varphi^{\mathrm{I}}{}_m = 0$ on $S^h{}_0$ and $\varphi^{\mathrm{I}}{}_m = \mathrm{I}$ on $S^h{}_1$. If necessary, such a function will be built on purpose and added to the list of trial functions. Now all functions of $\Phi^{\mathrm{I}}{}_m$ are of the form $\varphi^{\mathrm{I}}{}_m + \varphi$ with $\varphi \in \Phi^0{}_m$, which we can write in compact form like this:

(10) $\Phi^{\mathrm{I}}{}_m = \varphi^{\mathrm{I}}{}_m + \Phi^0{}_m$.

In words: $\Phi^{\mathrm{I}}{}_m$ is the *translate* of $\Phi^0{}_m$ by vector $\varphi^{\mathrm{I}}{}_m$ (Fig. 3.4).

It would now be easy to show that (9) is a *regular linear system*. The argument relies on uniqueness and on the equality between the number of unknowns (which are the degrees of freedom) and the number of equations in (9), which is the dimension of $\Phi^0{}_m$ (cf. Exer. 2.7). We defer this, however, as well as close examination of the properties of this linear system, till we have made a specific choice of trial functions.

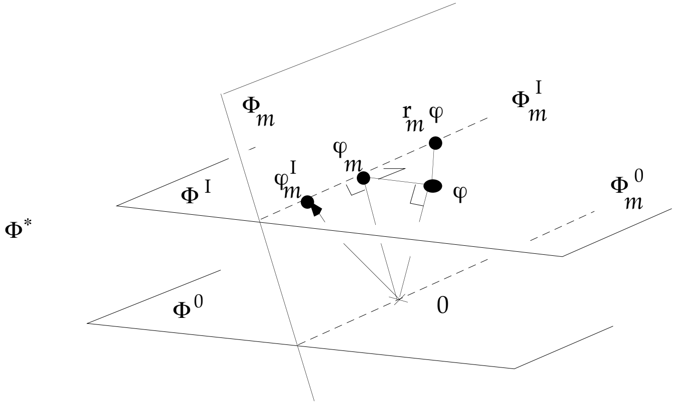


FIGURE 3.4. Geometry of the Ritz–Galerkin method.

Problem (9) is called the "discrete formulation", as opposed to the "continuous formulation" (1). Both are in weak form, but (9) is obviously "weaker", since there are fewer test functions. In particular, the weak solenoidality of $b = \mu \operatorname{grad} \varphi$ has been destroyed by restricting to a *finite* set of test functions. The span of such a set cannot be dense in Φ^0, so the proof of Prop. 2.3 is not available, and we can't expect Eqs. (2.23) and (2.24) in Chapter 2 (about $\operatorname{div} b = 0$ and $n \cdot b = 0$) to hold for $b_m = \mu \operatorname{grad} \varphi_m$. Still, something must be preserved, which we shall call, for lack of a better term, "*m*-weak solenoidality[9] of b" and "*m*-weak enforcement of the $n \cdot b$ boundary condition" on S^b. This also will wait (till the next chapter).

Meanwhile, it's interesting to examine the geometry of the situation (Fig. 3.4). The figure suggests that φ_m, which is the projection of 0 on $\Phi^I{}_m$, is also the projection of φ (the exact solution) on $\Phi^I{}_m$. This is correct: To see it, just restrict the test functions in (1) to elements of $\Phi^0{}_m$, which we have assumed (cf. (7)) are contained in Φ^0, which gives

$$\int_D \mu \operatorname{grad} \varphi \,.\, \operatorname{grad} \varphi' = 0 \quad \forall\, \varphi' \in \Phi^0{}_m \,.$$

But by (9) we also have

$$\int_D \mu \operatorname{grad} \varphi_m \,.\, \operatorname{grad} \varphi' = 0 \quad \forall\, \varphi' \in \Phi^0{}_m \,,$$

therefore, by difference,

$$(11) \qquad \int_D \mu \operatorname{grad}(\varphi_m - \varphi) \cdot \operatorname{grad} \varphi' = 0 \quad \forall\, \varphi' \in \Phi^0{}_m \,,$$

which expresses the observed orthogonality.

The figure also suggests a general method for error estimation. Let $r_m\varphi$ be an element of $\Phi^I{}_m$ that we would be able to associate with φ, as an approximation, *if* we knew it. Then we have, as read off the figure, and proved by setting $\varphi' = \varphi_m - r_m\varphi$ in (11),

$$||\varphi - \varphi_m\,||_\mu \leq ||\varphi - r_m\varphi\,||_\mu$$

($r_m\varphi$ is farther from φ, in energy, than φ_m is). So if we are able somehow to bound $||\varphi - r_m\varphi||_\mu$, an error bound on $\varphi_m - \varphi$ will ensue. The potential of the idea for error control is obvious, and we whall return to it in Chapter 4, with a specific Galerkin basis and a specific r_m.

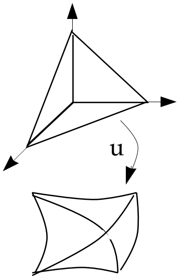
The Ritz–Galerkin method is of surprising efficiency. If trial functions are well designed, by someone who has good feeling for the real solution, a handful of them may be enough for good accuracy in estimating the functional. But it's difficult to give guidelines of general value in this respect, especially for three-dimensional problems. Besides, the computer changed the situation. We can afford many degrees of freedom nowadays (some modern codes use millions [We]) and can lavish machine time on the systematic design of Galerkin bases in a *problem-independent* way: This is what *finite elements* are about.

[9] The terminology is hesitant: Some say these equations are approximately satisfied "in the sense of weighted residuals", or "in the weak sense of finite elements", or even simply "in the weak sense", which may induce confusion. "Discrete" solenoidality might be used as a more palatable alternative to "m-weak" solenoidality.

3.3.2 Finite elements

So let be given a bounded domain $D \subset E_3$ with a piecewise smooth boundary S, and also inner boundaries, corresponding to material interfaces (discontinuity surfaces of μ, in our model problem).

A *finite element mesh* is a tessellation of D by volumes of various shapes, but arranged in such a way that two of them intersect, if they do, along a common face, edge, or node,[10] and never otherwise. We shall restrict here to tetrahedral meshes, where all volumes have six edges and four faces, but this is only for clarity. (In practice, hexahedral meshes are more popular.[11]) Note that a volume is not necessarily a straight tetrahedron, but may be the image of some "reference tetrahedron" by a smooth mapping u (inset).[12] This may be necessary to fit curved boundaries, or to cover infinite regions. Usually, one also arranges for material interfaces to be paved by faces of the mesh.



Exercise 3.7. Find all possible ways to mesh a cube by tetrahedra, under the condition that no new vertex is added.

Drafting a mesh for a given problem is a straightforward, if tedious, affair. But designing *mesh generators* is much more difficult, a scientific specialty [Ge] and an industry. We shall not touch either subject, and our only concern will be for the output of a mesh-generation process. The mesh is a complex data structure, which can be organized in many different ways, but the following elements are always present, more or less directly: (1) a list of nodes of the mesh, pointing to their locations; (2) a list of edges, faces, and volumes, with indirections allowing one to know which nodes are at the ends of this and that edge, etc.; (3) parameters describing the mapping of each volume to the reference one; (4) for each volume, parameters describing the material properties (for instance, the average value of μ, in our case).

For maximum simplicity in what follows, we assume that all volumes are straight tetrahedra. This can always be enforced, by distorting D to a polyhedron with plane faces, which is then chopped into tetrahedra.

[10] Or vertex. For some, "vertex" and "node" specialize in distinct meanings, vertices being the tips of the elementary volumes, and nodes the points that will support degrees of freedom. This distinction will not be made here.

[11] Most software systems offer various shapes, including tetrahedra and prisms, to be used in conjunction. This is required in practice for irregular regions.

[12] A more precise definition will be given in Chapter 7.

(This changes the model a little, of course, and adds some error to the approximation error inherent in the finite element method.)

We shall use the following simple description of the mesh: (1) four *sets*, denoted $\mathcal{N}$, $\mathcal{E}$, $\mathcal{F}$, $\mathcal{T}$, for nodes, edges, faces, and tetrahedra; (2) *incidence relations*, on which more below; (3) the *placement* of the mesh: this is a function $n \to x_n$, from $\mathcal{N}$ to $\overline{D}$, giving for each node n its position x_n in D or on S. In the case of straight tetrahedra, this is enough to determine the location of all *simplices* (the generic name for node, edge, face, etc.), and no other placement parameters are needed.

Thanks to this placement, one can confuse under a single expression, for example, "tetrahedron T", two conceptually different things: here the element T of $\mathcal{T}$, which is a mere label, and the tetrahedron T, a part of D, which is its image under the placement. It's a convenient and not too dangerous abuse,[13] which I'll commit freely, for all simplices. Symbols $\mathcal{F}(e)$, $\mathcal{N}(T)$, and other similar ones, will stand for, respectively, the subset of all faces that contain edge e, the subset of all nodes that are contained in tetrahedron T, and other similar subsets for various simplices. The purpose of the incidence relations, which we shall wait until Chapter 5 to describe in full detail, is to point to the faces of a given tetrahedron, the edges of a given face, etc., and thus to give full knowledge of subsets like $\mathcal{F}(e)$ or $\mathcal{N}(T)$. Finally, we shall denote by D_n the subdomain of D obtained by putting together all tetrahedra of the subset $\mathcal{T}(n)$, and use similar notation for D_e and D_f, calling D_s the *cluster* of tetrahedra *around* simplex s (Fig. 3.5). (No attempt is made to distinguish between open and closed clusters, as that will be clear from context.)

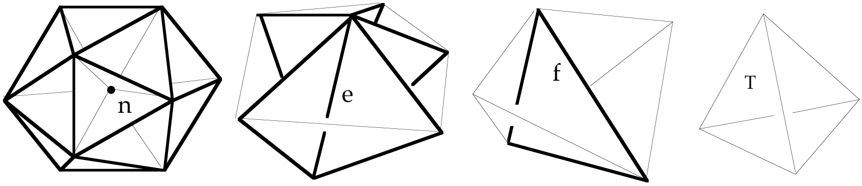


FIGURE 3.5. Clusters of tetrahedra around simplex s (s being, from left to right, node n, edge e, face f, and tetrahedron T). For better view, faces containing the simplex are supposed to be opaque, and others transparent.

Let's now recall the notion of barycentric coordinates. Four points x_1, x_2, x_3, x_4 in three-dimensional space are in *generic position* if the determinant $\det(x_2 - x_1, x_3 - x_1, x_4 - x_1)$ does not vanish. In that case, they

[13] Mathematicians use s for the simplex as an algebraic object and $|s|$ for its image.

form a tetrahedron. Four real numbers $\lambda^1, \lambda^2, \lambda^3, \lambda^4$ such that $\sum_i \lambda^i = 1$ determine a point x, the *barycenter* of the x_is for these *weights*, uniquely defined by

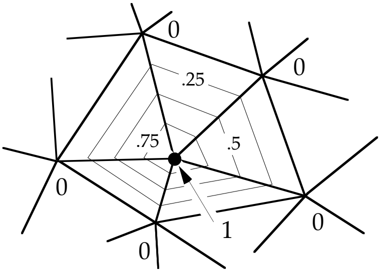
$$x - x_0 = \sum_{i=1,4} \lambda^i (x_i - x_0), \tag{12}$$

where x_0 is any origin (for instance, one of the x_is). Conversely, any point x has a unique representation of the form (12), and the weights λ^i, considered as four functions of x, are the *barycentric coordinates* of x in the *affine basis* provided by the four points. Note that x belongs to the tetrahedron if $\lambda^i(x) \geq 0$ for all i. The λ^is are affine functions of x.

Remark 3.3. Consequently, a function p which is polynomial with respect to the three Cartesian coordinates can be expressed as a polynomial expression $x \to P(\lambda^1(x), \ldots, \lambda^4(x))$ of the barycentric coordinates, where P is another polynomial, *of the same maximum degree* as p, with four variables. This possibility is often used, usually without warning. ◊

Now, consider our paving of $\overline{D}$ by tetrahedra. To each node n of the mesh, let us attribute a function, defined as follows: Its value at point x is 0 if the cluster D_n does not contain x, and if it does, it is the barycentric coordinate of x with respect to n, in the affine basis provided by the tetrahedron to which x belongs. (There is no ambiguity in that, because if x belongs to a simplex s, and thereby, to all tetrahedra of the cluster of s, its barycentric coordinates with respect to vertices of s are all the same, whatever the tetrahedron one considers to reckon them.) We shall reattribute to this *nodal function* the symbol λ^n. Note that, by construction, $\lambda^n(x) \geq 0$, its support is $\overline{D}_n$, its domain is $\overline{D}$ (but doesn't go beyond), and

$$\sum_{n \in \mathcal{N}} \lambda^n(x) = 1 \text{ for all } x \in \overline{D}. \tag{13}$$



The λ^ns themselves are often called "barycentric coordinates", though they coincide with the previous λ^is only for the nodes around x. This abuse is harmless, but I'll stick to "nodal functions", notwithstanding.

A shorter way to describe them is to say: λ^n is the only *piecewise affine* function[14] that takes the value 1 at node n and 0 at all other nodes. The inset shows the pattern of level lines of λ^n in the 2D case (triangulation

[14] Meaning: affine by restriction to each tetrahedron. I will use "*mesh-wise*" in such cases: mesh-wise affine, mesh-wise quadratic, etc. (this is not standard terminology).

of a plane domain D). It is easy from this to imagine the graph of the corresponding function, and to understand why the λ^ns are often called "hat functions".

Exercise 3.8. Prove that the hat functions are linearly independent.

Exercise 3.9. Compute the average of λ^n over (1) an edge e, (2) a face f, (3) a tetrahedron τ, all containing n.

Remark 3.4. Two things are essential in this construction: (1) each λ^n is supported on the cluster of n, (2) they form a *partition of unity* over D, i.e., $\sum_{n \in \mathcal{N}} \lambda^n = 1$, relation (13). The affine character is secondary, and is lost in case of curved tetrahedra.[15] But it considerably simplifies the programming, in conjunction with Remark 3.3, as we'll see. ◊

Well, that's all: *The finite element method is the Ritz–Galerkin method, the basis functions being a partition of unity associated with a mesh,* as above.

There are many ways to devise such a partition of unity, and the use of barycentric functions is only the simplest. When one refers to "a" finite element, it's this whole procedure one has in mind, not only the analytical expression of the basis functions. However, the latter suffices in many cases. Here, for instance, the restrictions of the λ^ns to individual tetrahedra are affine functions, that is, polynomials of maximum degree 1 of the Cartesian coordinates (one calls them "P^1 elements" for this reason), and this is enough characterization.[16]

Let us give another example, which demonstrates the power of this notation. What are "P^2 elements"? This means functions with small support, like the above λ^ns, which restrict to each tetrahedron as a second-degree polynomial, and therefore (Remark 3.3) are in the span of the products $\lambda^n\lambda^m$. This is enough to point to the partition of unity, for the set $\{\lambda^n\lambda^m : n \in \mathcal{N}, m \in \mathcal{N}\}$ is perfect in this respect: we do have

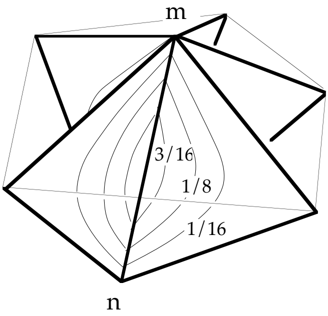
$$\sum_{n,\, m \in \mathcal{N}} \lambda^n \lambda^m = \sum_{n \in \mathcal{N}} [\lambda^n (\sum_{m \in \mathcal{N}} \lambda^m)] = \sum_{n \in \mathcal{N}} \lambda^n = 1$$

after (13), and the support of $\lambda^n \lambda^m$ is either the cluster of n, if n = m, or the cluster of edge n to m, if n and m are neighbors (the inset, next page,

[15] What is affine, then, is the "pull-back" of λ^n onto the reference tetrahedron. For this notion, push a little forward (Note 7.9).

[16] There is in finite element theory a traditional distinction between "basis functions", like the λ^n, and "shape functions", which are their restrictions to mesh volumes. As one sees here, shape functions are more simply characterized. Theory, on the other hand, is easier in terms of basis functions.

shows the level lines of $\lambda^n \lambda^m$). Note how the coefficients $\boldsymbol{\varphi}_{n\,m}$ in the expansion $\varphi = \sum_{n,\,m} \boldsymbol{\varphi}_{n\,m} \lambda^n \lambda^m$ are determined by the values of φ at the nodes and the mid-edges (Exer. 3.11).



Exercise 3.10. Compute the averages of $\lambda^n \lambda^m$ and $\lambda^n \lambda^m \lambda^\ell$ on a tetrahedron, in all cases, $n \neq m$, $n = m$, etc.

Exercise 3.11. Devise a set of P^2 functions w_{mn} such that $w_{mn} = 1$ at the middle of edge $\{m, n\}$, or at node n if $n = m$, and 0 at all other nodes and mid-edges.

Exercise 3.12 (Gaussian quadrature formulas). The average of an affine function over a tetrahedron is the average of its nodal values. The average of a *quadratic* function is a *weighted* average of its nodal and mid-edge values. Which weights? What about triangles?

Finite elements with degrees of freedom attached to specific points (cf. Note 10), like the P^1 and P^2 elements, are called *Lagrangian* [CR]. There are other varieties, built on hexahedra or other shapes, or with derivatives as DoFs (those are *Hermitian* elements), and so forth. Refer to specialized books such as [Ci]. There are also vector-valued finite elements, to which we shall return in Chapters 5 and 6.

3.3.3 The linear system

Generated by these basis elements, the finite dimensional subspace Φ_m contains all functions of the form

$$(14) \qquad \varphi = \sum_{n \in \mathcal{N}} \boldsymbol{\varphi}_n \lambda^n.$$

There is one degree of freedom $\boldsymbol{\varphi}_n$ for each node n, equal to the value of φ at node n. The family $\boldsymbol{\varphi} = \{\boldsymbol{\varphi}_n : n \in \mathcal{N}\}$ can be construed as a vector of an N-dimensional space, where $N = \#\mathcal{N}$ is the number of nodes in the mesh. We shall denote this vector space by $\boldsymbol{\Phi}_m$ (and drop the m, which can be done without any risk of confusion while we are dealing with *one* mesh at a time). Of course Φ_m and $\boldsymbol{\Phi}$ are isomorphic, but they are objects of different kinds, and we shall keep the difference in mind. To stress it, let us call p_m the injective map from $\boldsymbol{\Phi}$ into Φ defined by (14), which sends $\boldsymbol{\varphi}$ to $\varphi_m = p_m(\boldsymbol{\varphi})$. Then, $\Phi_m = p_m(\boldsymbol{\Phi})$. Similar notation will be used throughout, with capitals for spaces, and boldface connoting degrees of freedom and the vector spaces they span. In particular, we shall denote with bold parentheses the Euclidean scalar product of two elements of $\boldsymbol{\Phi}$, like this:

(15) $$(\boldsymbol{\varphi}, \boldsymbol{\varphi}') = \sum_{n \in \mathcal{N}} \boldsymbol{\varphi}_n \boldsymbol{\varphi}'_n.$$

To introduce Φ^I_m, first call $\mathcal{N}(S^h)$ the set of all boundary nodes that belong to S^h, including those on the frontier between S^h and S^b. Formally, $\mathcal{N}(S^h) = \{n \in \mathcal{N} : x_n \in \mathrm{cl}(S^h)\}$, where cl stands for the closure relative to S. Let $\mathcal{N}(S^h_0)$ and $\mathcal{N}(S^h_1)$ similarly be defined. Then, define

(16) $$\boldsymbol{\Phi}^I = \{\boldsymbol{\varphi} \in \boldsymbol{\Phi} : \boldsymbol{\varphi}_n = 0 \text{ if } n \in \mathcal{N}(S^h_0),\ \boldsymbol{\varphi}_n = I \text{ if } n \in \mathcal{N}(S^h_1)\}$$

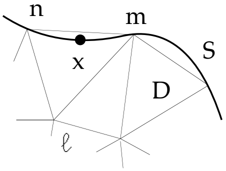
and, similarly, $\boldsymbol{\Phi}^0$, two parallel subspaces of $\boldsymbol{\Phi}$. Finally, let us set

(17) $$\Phi^0_m = p_m(\boldsymbol{\Phi}^0), \quad \Phi^I_m = p_m(\boldsymbol{\Phi}^I).$$

Relation (10), $\Phi^I_m = \varphi^I_m + \Phi^0_m$, has a counterpart here. Let us construct $\boldsymbol{\varphi}^1$, a special vector, with all components $\boldsymbol{\varphi}^1_n = 0$ except for $n \in \mathcal{N}(S^h_1)$, where they are set to 1. Then, with $\boldsymbol{\varphi}^I$ defined as $\boldsymbol{\varphi}^I = I\,\boldsymbol{\varphi}^1$,

(18) $$\boldsymbol{\Phi}^I = \boldsymbol{\varphi}^I + \boldsymbol{\Phi}^0.$$

Remark 3.5. If you try to check (7) at this stage, you will see that it fails if the faces at the boundary do not fit it exactly. Cf. the inset: a piecewise affine function that vanishes at n and m, but not at ℓ, cannot be zero at x. Because of this tiny difference, Φ^I_m is not contained in Φ^I, and applying the geometrical reasonings suggested by Fig. 3.4 would be a "variational crime", in the sense of Strang and Fix [SF]. This (jocular) charge should not deter anyone from using a mesh similar to the one in inset in case of a curved boundary. This is perfectly right! What is not, and would constitute the crime, would be to apply the simple convergence proof that will follow to such a situation, which calls for more cumbersome treatment. Thanks to our decision to deform D into a polyhedron before meshing, we do have $\Phi^I_m = \Phi_m \cap \Phi^I$ and $\Phi^0_m = \Phi_m \cap \Phi^0$, as announced in (7). But this will not be effectively used before we address convergence and error analysis, and what immediately follows does not depend on the truth of these assertions. ◊



We want now to interpret problem (9), that is,

(9') *find* $\varphi_m \in \Phi^I_m$ *such that* $\int_D \mu \operatorname{grad} \varphi_m \cdot \operatorname{grad} \varphi' = 0 \quad \forall\, \varphi' \in \Phi^0_m$,

in algebraic terms. Since $\Phi^I_m = p_m(\boldsymbol{\Phi}^I)$, this is a linear system with respect to the "free" degrees of freedom, that is, those not constrained by (16), which are the nodes of the subset $\mathcal{N}_0 = \mathcal{N} - \mathcal{N}(\mathrm{S^h})$. On the other hand, since $\Phi^0{}_m$ is parallel to $\Phi^{\mathrm{I}}{}_m$, there are as many equations as unknowns in (9'). Our aim is to rewrite (9') in terms of the degrees of freedom. For this, let us set, for any two nodes n and m,

$$(19) \qquad \mathbf{M}_{\mathrm{n\,m}} = \int_{\mathrm{D}} \mu \operatorname{grad} \lambda^{\mathrm{n}} \cdot \operatorname{grad} \lambda^{\mathrm{m}},$$

and form the symmetric matrix $\mathbf{M}$, indexed on $\mathcal{N} \times \mathcal{N}$, of which this is the entry at row n and column m. Then (just write φ_m and φ' as in (14), and expand), (9') is equivalent to

$$(9'') \qquad \textit{find } \boldsymbol{\varphi} \in \boldsymbol{\Phi}^{\mathrm{I}}{}_m \textit{ such that } \quad (\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi}') = 0 \quad \forall\, \boldsymbol{\varphi}' \in \boldsymbol{\Phi}^0,$$

via the correspondence $\varphi_m = \mathrm{p}_m(\boldsymbol{\varphi})$. As a matter of course (we did that twice already), this is equivalent to the variational problem

$$(9''') \qquad \textit{find } \boldsymbol{\varphi} \in \boldsymbol{\Phi}^{\mathrm{I}} \textit{ such that } \ \mathbf{F}(\boldsymbol{\varphi}) \leq \mathbf{F}(\boldsymbol{\psi}) \quad \forall\, \boldsymbol{\psi} \in \boldsymbol{\Phi}^{\mathrm{I}},$$

where $\mathbf{F}(\boldsymbol{\varphi}) = \frac{1}{2}(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi})$.

$\mathbf{M}$ is traditionally dubbed the *stiffness matrix* of the problem, because of the origins of the finite elements method: In mechanics, the analogue of our $\boldsymbol{\varphi}$ is most often a displacement vector, and $\mathbf{F}(\boldsymbol{\varphi})$ is deformation energy, so $\mathbf{M}\boldsymbol{\varphi}$ is a force vector, and a force-to-displacement ratio is a stiffness. (One could make a case for *admittance* matrix, in our context.)

As a last step, let us write $\mathbf{M}$ in block form, by partitioning the indexing set $\mathcal{N}$ as[17] $\mathcal{N} = \mathcal{N}_0 + \mathcal{N}(\mathrm{S^h})$. With ad-hoc but obvious notation,

$$\mathbf{M} = \begin{vmatrix} {}^{00}\mathbf{M} & {}^{01}\mathbf{M} \\ {}^{10}\mathbf{M} & {}^{11}\mathbf{M} \end{vmatrix},$$

where the submatrix ${}^{00}\mathbf{M}$ is indexed over $\mathcal{N}_0$ and thus operates in the subspace $\boldsymbol{\Phi}^0$ of genuine unknowns (those not constrained by essential boundary conditions). We also write vectors in block form, $\boldsymbol{\varphi} = \{{}^0\boldsymbol{\varphi}, {}^1\boldsymbol{\varphi}\}$, and $\boldsymbol{\varphi}^{\mathrm{I}} = \{\mathbf{0}, {}^1\boldsymbol{\varphi}^{\mathrm{I}}\}$. Thanks to this and to (18), we see that (9'') is equivalent to *find* ${}^0\boldsymbol{\varphi} \in \boldsymbol{\Phi}^0$ *such that*

$$(20) \qquad {}^{00}\mathbf{M}\,{}^0\boldsymbol{\varphi} = -\,{}^{01}\mathbf{M}\,{}^1\boldsymbol{\varphi}^{\mathrm{I}},$$

at last a standard linear system, since the right-hand side $-\,{}^{01}\mathbf{M}\,{}^1\boldsymbol{\varphi}^{\mathrm{I}}$ is known.

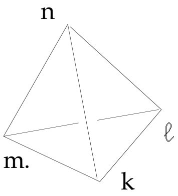
[17] The union sign $\cup$ is replaced by + when sets are disjoint, as here.

3.3.4 "Assembly", matrix properties

Methods to effectively *solve* this linear system are beyond our scope. One may refer to many excellent handbooks for this, among which are [Cr, GL, Gv, Va]. The choice of methods, however, strongly depends on the structure and properties of ${}^{00}\mathbf{M}$, so we need a few indications on this. And of course, we must address the practical problem of computing the entries (19).

Let us say from the outset that it's not a good idea to concentrate on the "useful" matrix ${}^{00}\mathbf{M}$ of (20), thus forgetting about $\mathbf{M}$, for two reasons. First, the properties of $\mathbf{M}$ are simpler to discover, and those of its *principal* submatrices (i.e., diagonal sub-blocks), like ${}^{00}\mathbf{M}$, easily follow. Next, the boundary conditions one wishes to consider may change during the study of a given problem, thus changing the set $\mathcal{N}_0$. Finally, as we shall see in the next chapter, some data one wishes to access require the knowledge of all $\mathbf{M}$.

The first concern is for the computation of the entries of $\mathbf{M}$. With ∇ standing for grad for shortness, let us define (cf. (19))



$$\mathbf{M}^{\mathrm{T}}_{n\,m} = \int_{\mathrm{T}} \mu\, \nabla\lambda^n \cdot \nabla\lambda^m,$$

so that $\mathbf{M}_{n\,m} = \sum_{\mathrm{T}\in\mathcal{T}} \mathbf{M}^{\mathrm{T}}_{n\,m}$. If one replaces μ by its average $\mu(\mathrm{T})$ over the tetrahedron, this is easy to compute, as we now show.

Call $\{k, \ell, m, n\}$ the vertices of T, and suppose for definiteness they are placed as shown in inset, vectors[18] $k\ell$, km, kn forming a positively oriented frame. Notice that $|k\ell \times km|/2$ is the area of face $\{k, \ell, m\}$ and $1/|\nabla\lambda^n|$ the height of the tetrahedron above that face, which results in $\nabla\lambda^n = (k\ell \times km)/(6\,\mathrm{vol}(\mathrm{T}))$. Now,

$$\text{(21)}\qquad \int_{\mathrm{T}} \nabla\lambda^n \cdot \nabla\lambda^m = \frac{1}{36\,\mathrm{vol}(\mathrm{T})}\,(k\ell \times km)\cdot(\ell n \times \ell k\,)$$

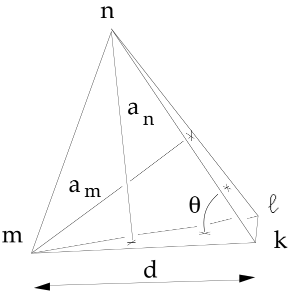
$$= \frac{1}{36\,\mathrm{vol}(\mathrm{T})}\,[\,(k\ell \cdot \ell n)\,(\ell k \cdot km) + |k\ell|^2\; km \cdot \ell n]$$

by a well-known formula[19] (which shows, incidentally, that the result is insensitive to the orientation of T). There is another expression for (21), known as the *cotangent formula*, which gives useful insight. The dot product of $\nabla\lambda^n$ and $\nabla\lambda^m$ is $-(1/a_m)(1/a_n)\cos\theta$, with the notation given

[18] A symbol like km denotes the vector from point x_k (the location of node k) to point x_m. This is a gross abuse of notation, but an innocuous one.

[19] $(a \times b)\cdot(c \times d) = (a\cdot c)\,(b\cdot d) - (a\cdot d)\,(b\cdot c)$.

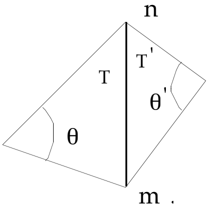
in inset (a for "altitude"), and vol(T) = $a_n d\ |k\ell|/6$. But $a_m = d \sin\theta$, where d is the distance from m to the line supporting $k\ell$, and thus,



$$\int_T \nabla\lambda^n\cdot\nabla\lambda^m$$
$$= -\,a_n\ |km|\ |k\ell|\ \cos\theta/(6\ a_m a_n)$$
$$= -\ \frac{1}{6}\ |k\ell|\ \cot\theta.$$

Adding all contributions, we thus have, with ad-hoc and obvious notation,

(22) $$\mathbf{M}_{nm} = \int_D \mu\,\nabla\lambda^n\cdot\nabla\lambda^m = -\ \frac{1}{6}\ \sum_{T\in\mathcal{T}(\{n,\,m\})} \mu(T)\ |k\ell|_T \cot\theta_T\,.$$



In dimension 2 (notation in inset), this reduces to

(23) $$\mathbf{M}_{nm} = -\ \frac{1}{2}\ (\,\mu(T)\cot\theta + \mu(T')\cot\theta'\,).$$

Remark 3.6. Diagonal entries $\mathbf{M}_{nn}$ cannot be obtained by this formula, but since $\nabla\lambda^n = -\sum_{m\neq n}\nabla\lambda^m$, one has $\sum_m \mathbf{M}_{nm} = \sum_m \int_D \mu\ \mathrm{grad}\ \lambda^n\cdot\mathrm{grad}\ \lambda^m = \int_D \mu\ \mathrm{grad}\ \lambda^n\cdot\mathrm{grad}(\sum_m \lambda^m) = 0$, hence $\mathbf{M}_{nn} = -\sum_{m\neq n}\mathbf{M}_{nm}$, and also $\mathbf{M}^T_{nn} = -\sum_{m\neq n}\mathbf{M}^T_{nm}$ by summing over T instead of D. So (22) is enough. ◊

Since the data structure gives access to the node locations, and hence to the components of the edge vectors, the simplest programming is via (21), which requires no more than coding a handful of determinants (the volume itself is $|\det(k\ell, km, kn)|/6$). This will be the basic subroutine for the assembly program. Running it for all pairs of nodes gives the "elementary matrix" $\mathbf{M}^T$ and then $\mathbf{M} = \sum_{T\in\mathcal{T}}\mathbf{M}^T$ by looping over the tetrahedra. This is the *assembly* process, by which the matrix is constructed from the mesh data structure.

This way, only terms which do contribute to the matrix are evaluated. A priori, of course, $\mathbf{M}_{nm} = 0$ for pairs of nodes which are not linked by a common edge, that is, most of them: $\mathbf{M}$ is *sparse*, that is to say, has a small percentage of nonzero entries. This has consequences, also, on the way these entries are stored (the precise coding of the assembly depends on options taken at this level) and on the algorithms for solving the linear system [BR, GL].

This sparsity is perhaps the most important property of finite-element matrices. (The Galerkin method generates full matrices, unless the supports of the basis functions are small, which is precisely what finite elements achieve.) Other properties we now list are not specific to finite elements, but depend on the "partition of unity" feature (the equality $\sum_{n \in \mathcal{N}} \lambda^n = 1$). For shortness, $\mathbf{1}$ will denote the vector of $\mathbf{\Phi}$ all components of which are equal to 1, and $\vee\{\boldsymbol{\varphi}, \boldsymbol{\psi}, \ldots\}$ the *span* of a family of vectors $\boldsymbol{\varphi}$, $\boldsymbol{\psi}$, . . . (cf. A.2.2).

Proposition 3.3. $\mathbf{M}$ *is symmetric, and nonnegative definite, that is* (cf. Section B.1),

$$(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi}) \geq 0 \ \ \forall \, \boldsymbol{\varphi}.$$

Proof. $(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi}) = \|\mathrm{p}(\boldsymbol{\varphi})\|_{\mu}^{2} \equiv \int_D \mu \ |\mathrm{grad}\ \mathrm{p}(\boldsymbol{\varphi})|^2 \geq 0.$ ◊

Proposition 3.4. $\ker(\mathbf{M}) = \vee\{\mathbf{1}\}$.

Proof. We already know that $\mathbf{M1} = 0$ (Remark 3.6). Conversely, if $\mathbf{M}\boldsymbol{\varphi} = 0$, then $(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi}) = 0$, hence grad $\mathrm{p}(\boldsymbol{\varphi}) = 0$, hence $\mathrm{p}(\boldsymbol{\varphi}) = \mathrm{c}$, a constant, and $\sum_n (\boldsymbol{\varphi}_n - \mathrm{c}) \lambda^n = 0$, hence $\boldsymbol{\varphi} = \mathrm{c}\,\mathbf{1}$, if the λ^ns are independent, which we know is the case for hat functions, by Exer. 3.8. ◊

Exercise 3.13. If $\mathbf{M}$ is symmetric and nonnegative definite, show that $\ker(\mathbf{M})$, which is defined as $\{\boldsymbol{\varphi} : \mathbf{M}\boldsymbol{\varphi} = 0\}$, is equal to $\{\boldsymbol{\varphi} : (\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi}) = 0\}$.

Proposition 3.5. *Apart from* $\mathbf{M}$ *itself, all principal submatrices of* $\mathbf{M}$ *are positive definite* (cf. Section B.1), *and hence regular.*

Proof. Let $\mathcal{N}_0$ be a part of $\mathcal{N}$, and consider a block partitioning of $\mathbf{M}$ on the basis of the $\mathcal{N} = \mathcal{N}_0 + (\mathcal{N} - \mathcal{N}_0)$ partitioning of the node set. To avoid vertical displays, let us write this $\mathbf{M} = \{\{{}^{00}\mathbf{M}, {}^{01}\mathbf{M}\}, \{{}^{10}\mathbf{M}, {}^{11}\mathbf{M}\}\}$, by rows of blocks, according to the standard convention. Then ${}^{00}\mathbf{M}$ is (by definition) a *principal* submatrix of $\mathbf{M}$. Suppose there is a vector ${}^{0}\boldsymbol{\varphi}$, supported on $\mathcal{N}_0$, such that $({}^{00}\mathbf{M}\,{}^{0}\boldsymbol{\varphi}, {}^{0}\boldsymbol{\varphi}) = 0$, and build from it a vector $\boldsymbol{\varphi}$ supported on all $\mathcal{N}$ by attributing the value 0 to all DoFs in $\mathcal{N} - \mathcal{N}_0$. Then $(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi}) = 0$, which we know implies $\boldsymbol{\varphi} = \mathrm{c}\,\mathbf{1}$. But if $\mathcal{N}_0 \neq \mathcal{N}$, then some components of $\boldsymbol{\varphi}$ vanish, hence $\mathrm{c} = 0$, and ${}^{0}\boldsymbol{\varphi} = \mathbf{0}$. Then, by the result of Exercise 3.13, ${}^{00}\mathbf{M}$ is regular. ◊

A particular case is when $\mathcal{N}_0$ reduces to *one* node n, showing that $\mathbf{M}_{nn} > 0$. So all diagonal coefficients of $\mathbf{M}$ are positive, and the sum of entries of a same row, or column, is zero, by Prop. 3.4.

Now, a property which is more closely linked with the use of barycentric functions. Formula (22) shows that in case of acute dihedral angles, all off-diagonal entries are nonpositive. Symmetric positive definite matrices with ≤ 0 off-diagonal coefficients are called *Stieltjes matrices* [Va] and are important because of the following property:

Proposition 3.6. *If* $\mathbf{A}$ *is a Stieltjes matrix, all entries of its inverse are nonnegative.*[20]

Proof. Let's agree to write $\mathbf{v} \ge 0$ if no component of vector $\mathbf{v}$ is negative. Let $\mathbf{u}$ be the solution of the linear system $\mathbf{A}\,\mathbf{u} = \mathbf{b}$, and suppose $\mathbf{b} \ge 0$. Write $\mathbf{u}$ in the form $\mathbf{u} = \mathbf{u}^+ - \mathbf{u}^-$, with $\mathbf{u}^+ \ge 0$ and $\mathbf{u}^- \ge 0$, and remark that $\mathbf{u}^+_n \mathbf{u}^-_n = 0$ for all $n \in \mathcal{N}_0$ (if we continue to call $\mathcal{N}_0$ the indexing set of $\mathbf{S}$, $\mathbf{u}$, and $\mathbf{b}$). Now,

$$0 \le (\mathbf{b}, \mathbf{u}^-) = (\mathbf{A}(\mathbf{u}^+ - \mathbf{u}^-), \mathbf{u}^-) = (\mathbf{A}\mathbf{u}^+, \mathbf{u}^-) - (\mathbf{A}\mathbf{u}^-, \mathbf{u}^-).$$

But $(\mathbf{A}\mathbf{u}^+, \mathbf{u}^-) \le 0$, because only off-diagonal entries of $\mathbf{A}$ contribute to this scalar product, so $(\mathbf{A}\mathbf{u}^-, \mathbf{u}^-) = 0$, hence $\mathbf{u}^- = 0$. Thus $\mathbf{A}^{-1}\mathbf{b}$ has no negative components if $\mathbf{b}$ has none, hence the result. ◊

This applies to our system matrix ${}^{00}\mathbf{M}$ in the case where no dihedral angle is obtuse, with interesting consequences that we shall discover in the next chapter.

EXERCISES

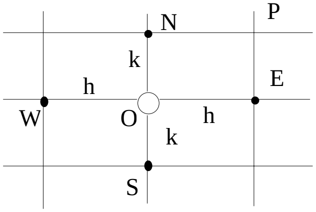
Texts for Exercises 3.1 to 3.4 are on pp. 64 and 66.

Exercise 3.5. Reinforce your knowledge of the following facts: The real and imaginary parts of a function which is holomorphic in some domain of the complex plane are harmonic. Conformal transformations preserve harmonicity.

Exercise 3.6. In the plane $\{x, y\}$, find a function φ which is harmonic in the domain $\{\{x, y\} : x < 0$ **or** $y < 0\}$ and null on the axes $y = 0$ and $x = 0$. Take its restriction to the domain D obtained by clipping the regions $y \le -1$ and $x \le -1$. Examine the singularity of φ at 0. Is $\nabla\varphi$ square-integrable in D ? Show that this is the idealization of a situation which can happen physically.

Exercise 3.7 is on p. 74. Exers. 3.8 to 3.12 are on pp. 77 to 79, and Exer. 3.13 on p. 83.

[20] Matrices with this property are called "monotone". (They are akin to "M-matrices" [BP, Jo, Na]. Beware the terminological confusion around this concept.) Notice that at least one term on each row of the inverse must be positive.

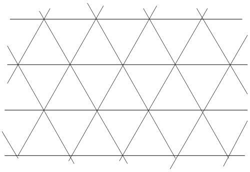


Exercise 3.14. In the *finite differences* method, potential values at the nodes of a so-called "orthogonal grid" are the unknowns, and equations are obtained via local Taylor expansions of the unknown potential [Va]. For instance (inset), if φ must satisfy $-\Delta\varphi = 0$, the values of φ at a node O and at neighboring nodes E, N, W, S, will approximately satisfy

$$(24) \qquad \varphi_O = [k^2(\varphi_E + \varphi_W) + h^2(\varphi_N + \varphi_S)]/2(k^2 + h^2).$$

There is one equation of this kind for each node like O, and altogether they form a linear system similar to (20), when one takes into account boundary nodal values. Describe this method as a special case of the finite element method.

Exercise 3.15. The method of finite differences does not adapt easily to domains with complicated boundaries, and the finite element method has a decisive advantage in this respect. However, it's intuitive that inside physically homogeneous regions (constant coefficients), one should use meshes as "regular" (that is, uniform and repetitive, crystal-like) as one can devise. For instance, in 2D, equilateral triangles (inset) are a good idea. As far as tetrahedral elements are concerned, do we have the equivalent of this in 3D? Can one pave space with regular tetrahedra?



Exercise 3.16. The next best thing to a regular tetrahedron is an *isosceles* tetrahedron, one for which opposite edges are equal, two by two [Co]. Find an isosceles tetrahedron that will pave.

Exercise 3.17. One expects the reluctance of a circuit to decrease when the permeability increases anywhere inside. Show that, indeed, if $\mu_2 \geq \mu_1$ a.e. in D in our model problem, the corresponding reluctances satisfy $R_2 \leq R_1$.

Exercise 3.18. One expects the reluctance to decrease when all the dimensions of the device increase proportionally. Prove it.

Exercise 3.19. Study the continuity of φ with respect to μ in the model problem.

Exercise 3.20 (research project). Study the continuity of φ with respect to the shape of the domain.

HINTS

3.1. The question amounts to finding a quadratic functional whose directional derivative at point φ would be

$$\varphi' \rightarrow \int_D \mu_0 \operatorname{grad} \varphi \cdot \operatorname{grad} \varphi' + \mu_0 \int_D \mathrm{m} \cdot \operatorname{grad} \varphi',$$

and we know the answer when $\mathrm{m} = 0$. What remains to be found is a function of φ (obviously, a linear function, since m does not depend on φ) having $\varphi' \rightarrow \mu_0 \int_D \mathrm{m} \cdot \operatorname{grad} \varphi'$ as its directional derivative.

3.2. On Φ, $\| \ \|_\mu$ is not a norm: Properties $\|\lambda\, \varphi\|_\mu = |\lambda|\, \|\varphi\|_\mu$, $\|\varphi + \psi\|_\mu \leq \|\varphi\|_\mu + \|\psi\|_\mu$ and $\|\varphi\|_\mu \leq 0$ do hold, but $\|\varphi\|_\mu = 0$ does not entail $\varphi = 0$. We have only a *semi-norm* there. How can that be cured?

3.3. The goal is to find a constant C such that $|\mathcal{J}(\psi)| \leq C\, \|\psi\|_\mu$, where ψ is any member of Φ^*, not one that satisfies (1) necessarily. But on the other hand, the solution of (1) which has the same mmf as ψ is a good reference, after Fig. 3.2 and the proof of Prop. 3.1.

3.6. This is a simple exercise in conformal transformations. First find a harmonic function in a half-plane that vanishes on the boundary, then map the half-plane onto the desired region.

3.7. Call "small diagonals" and "large diagonals" the segments joining two vertices, depending on whether they belong to the cube's surface or not. Show that at most one large diagonal can exist in the mesh. If there is one, show that at least three inner faces must have it as an edge. (Beware, it's a challenging exercise.)

3.8. Look at the nodal values of $\sum_n \boldsymbol{\alpha}_n \lambda^n$.

3.9. In particular, $(\int_T \lambda^n)/\mathrm{vol}(\mathrm{T}) = 1/4$, where vol denotes the volume, and the general case is, obviously, $(\int_s \lambda^n)/\mathrm{meas}(s) = 1/(p + 1)$ for a simplex s of dimension p, where meas for "measure" stands for length, area, etc. To say "all sums $\int_s \lambda^n$ for $n \in \mathcal{N}(s)$ are equal, and they add to meas(s), by (13)" is a fine symmetry argument, but why this equality? It stems from general results on change of variables in integration—but rather try a pedestrian and straightforward "calculus proof".

3.10. Probably the simplest way is to use the calculus proof to compute $\int_s (\lambda^n)^2$, then the symmetry argument for $m \neq n$.

3.11. Up to the factor 4, $\lambda^n \lambda^m$ is right. As for $\lambda^n \lambda^n$, look at its behavior along a typical edge $\{n, m\}$, and rectify at mid-edge.

3.12. Combine Exers. 3.10 and 3.11.

3.13. The same trick as in Prop. 3.1.

3.14. Grid cells must be cut in two, so if point P for instance is linked with O by an edge, making them "neighbors" on the finite element mesh, one expects a nonzero entry in the stiffness matrix at row O and column P, which is *not* the case of the finite-difference scheme (24). Explaining why this term vanishes is the key. It has to do with the right angle, obviously.

3.15. If paving was possible, tetrahedra around a given edge would join without leaving any gap, so the dihedral angle would have to be $2\pi/n$ for some integer n. Is that so?

3.17. Suppose $I = 1$ for simplicity. Then $R_1^{-1} = \inf(\int_D \mu_1 \,|\, \nabla\varphi \,|^2 : \ \varphi \in \Phi^1\}$. Replace φ by φ_2, the solution for $\mu = \mu_2$.

3.18. Map the problem concerning the enlarged region onto the reference one, and see how this affects μ.

3.19. Consider two problems corresponding to permeabilities μ_1 and μ_2, all other things being equal. Denote the respective solutions by φ_1 and φ_2. Let $\|\varphi\|_1$ or $\|\varphi\|_2$ and $(\varphi, \varphi')_1$ or $(\varphi, \varphi')_2$ stand for $\|\varphi\|_\mu$ and $(\varphi, \varphi')_\mu$, depending on the value of μ. One has

$$(25) \qquad \int_D \mu_i \operatorname{grad} \varphi_i \cdot \operatorname{grad} \varphi' = 0 \quad \forall\ \varphi' \in \Phi^0, \text{ for } i = 1, 2.$$

Choose appropriate test functions, combine both equations, and apply the Cauchy–Schwarz inequality.

3.20. If the deformation is a homeomorphism, the same mapping trick as in Exer. 3.18 reduces the problem to analyzing the dependence with respect to μ, with a new twist, however, for μ will become a tensor. You will have to work out a theory to cover this case first.

SOLUTIONS

3.1. Let $\mathcal{M}(\varphi) = \int_D m \cdot \operatorname{grad} \varphi$. Since $\mathcal{M}(\varphi + \lambda\varphi') = \mathcal{M}(\varphi) + \lambda\mathcal{M}(\varphi')$, the directional derivative of $\mathcal{M}$ is $\varphi' \to \lim_{\lambda \to 0}(\mathcal{M}(\varphi + \lambda\varphi') - \mathcal{M}(\varphi))/\lambda$, that is, $\varphi' \to \int_D m \cdot \operatorname{grad} \varphi'$, the same formally[21] as $\mathcal{M}$ itself. This holds for all linear functionals, so we shall not have to do it again. The variational forms of (2.34) and (2.36) thus consist in minimizing the functionals

$$\varphi \to \tfrac{1}{2}\int_D \mu_0 \,|\operatorname{grad}\varphi|^2 + \mu_0\int_D m\cdot\operatorname{grad}\varphi \quad \text{on } \Phi^I ,$$

$$\varphi \to \tfrac{1}{2}\int_D \mu \,|\operatorname{grad}\varphi|^2 - F\,\mathcal{J}(\varphi) \qquad\qquad \text{on } \Phi^*,$$

respectively.

3.2. On Φ, $||\varphi||_\mu = 0$ implies a constant value of φ, but not $\varphi = 0$, so $||\ ||_\mu$ is not a norm, whereas its restriction to Φ^* is one. This hardly matters, anyway, since two potentials which differ by an additive constant have the same physical meaning. So another possibility would be for us to define the quotient $\Phi/\mathbb{R}$ of Φ by the constants, call that $\dot\Phi$, and give it the norm $||\dot\varphi||_\mu = \inf\{c \in \mathbb{R} :\ ||\varphi + c||_\mu\}$, where φ is a member of the class $\dot\varphi \in \dot\Phi$. Much trouble, I'd say, for little advantage, at least for the time being. Later, we'll see that what happens here is a general fact, which has to do with *gauging*: It's *equivalence classes* of potentials, not potentials themselves, that are physically meaningful, so this passage to the quotient I have been dodging here will have to be confronted.

3.3. Take $\psi \in \Phi^*$, and let $I = \mathcal{J}(\psi)$. Then (Fig. 3.2) $||\varphi(I)||_\mu \leq ||\psi||_\mu$. Using Prop. 3.2, we thus have

$$|\mathcal{J}(\psi)| = |I| = [||\varphi(1)||_\mu]^{-1}\, ||\varphi(I)||_\mu \leq [||\varphi(1)||_\mu]^{-1}\, ||\psi||_\mu .$$

3.4. If $m > n$, $\int |f_n - f_m| = \int_{[1/m,\ 1/n]} dx/\sqrt{x} = 2/\sqrt{n} - 2/\sqrt{m} < 2/\sqrt{n}$ tends to zero.

3.5. Let $f(z) = P(x, y) + i\, Q(x, y)$. *Holomorphy* of f inside D means differentiability in the *complex* field $\mathbb{C}$, that is, for all $z \in D$, existence of a complex number $\partial f(z)$ such that $f(z + dz) = f(z) + \partial f(z)\, dz + o(dz)$ for all dz in $\mathbb{C}$. Cauchy conditions for holomorphy are $\partial_x P = \partial_y Q$ and $\partial_y P = -\partial_x Q$, so $\partial_{xx} P = \partial_{xy} Q = \partial_{yx} Q = -\,\partial_{yy} P$, hence $\Delta P = 0$, and the same for Q. In dimension 2, *conformal mappings* (those which preserve angles, but not distances) are realized by holomorphic maps from $\mathbb{C}$ to $\mathbb{C}$, and holomorphy is preserved by composition.

3.6. A harmonic function in the upper half-plane $y > 0$ which vanishes for $y = 0$ is $\{x, y\} \to y$, the function denoted Im (for imaginary part). The *fan map* $g = z \to i\, z^{3/2}$ sends the upper half-plane to the domain

[21] But not conceptually. The argument of $\mathcal{M}$ is a point in an *affine* space, whereas φ', in the expression of the directional derivative, is an element of the associated *vector* space.

$\{\{x, y\} : x < 0$ **or** $y < 0\}$. Composition of Im and g^{-1} yields the desired function (cf. Fig. 3.6), better expressed in polar coordinates:

$$\varphi(r, \theta) = r^{2/3} \sin((2\theta - \pi)/3).$$

(Note that φ cannot be extended to the whole plane. Note also that it is not piecewise k-smooth for $k > 0$, in the sense we adopted in Chapter 2.) Its gradient is infinite at the origin, where its modulus behaves like $r^{-1/3}$. Since $\int_0^R (r^{-1/3})^2\, r\, dr = \int_0^R r^{1/3}\, dr$ converges, this is a potential with finite associated magnetic (co)energy.

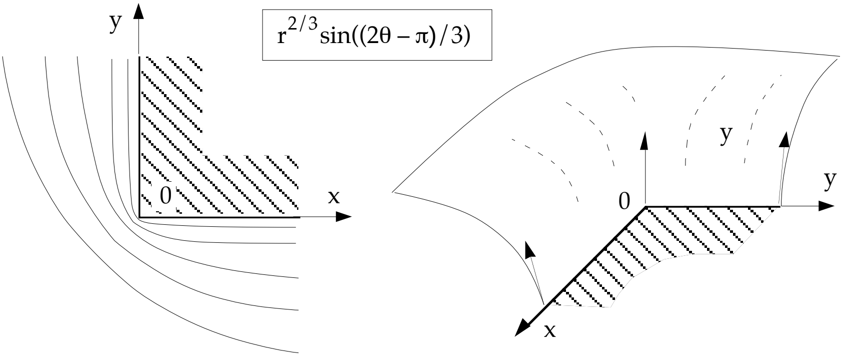


FIGURE 3.6. The function $f(r, \theta) = r^{2/3} \sin((2\theta - \pi)/3)$ of plane polar coordinates, for $\pi/2 \le \theta \le 2\pi$. Left: level lines. Right: perspective view of the graph.

Now imagine one of the level surfaces of φ (a cylinder along 0z) is lined up by some perfectly permeable material. This potential is then the solution of a two-dimensional analogue[22] of our model problem, in which the system would be infinite in the z–direction.

3.7. Two large diagonals would cut at the center, so there can't be more than one. At least three faces must hinge on it, since dihedral angles are less than π, and at most six, corresponding to the six possible vertices. So, see how to leave out one, two, or three of these. Fig. 3.7 gives the result. Alternatively, one may consider whether opposite faces are cut by parallel

[22] A genuinely three-dimensional example would be more demonstrative. See [Gr] for the (more difficult) techniques by which such examples can be constructed.

or anti-parallel small diagonals. (This is meaningful when one thinks of stacking cubes in order to make a tetrahedral mesh.) Whichever way, it's pretty difficult to prove the enumeration complete!

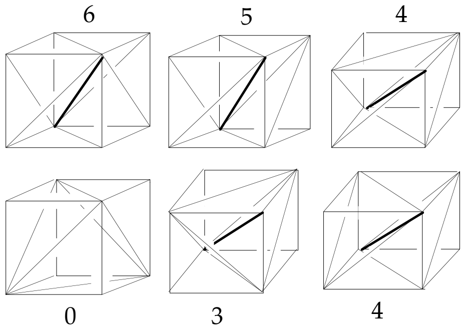


FIGURE 3.7. All ways to mesh a cube, depending on the number of inner faces that have a large diagonal as one of their edges.

3.8. $\sum_n \boldsymbol{\alpha}_n \lambda^n(x_m) = \boldsymbol{\alpha}_m$, by construction of the nodal functions, so if $\sum_n \boldsymbol{\alpha}_n \lambda^n = 0$, then all $\boldsymbol{\alpha}_n$s vanish.

3.9. One may invoke the general result about "change of variables" in integration, $\int_{u(D)} f\, J_u = \int_D u^* f$, where u^*f is the pull-back $x \to f(u(x))$ and J_u the Jacobian of the mapping u, for there is an affine map from T to itself that swaps n and m, λ^n and λ^m, which is volume preserving ($J_u = 1$). A much more elementary but safer alternative is, in Cartesian coordinates: Place the basis of tetrahedron T in plane x–y, and let n be the off-plane node, at height h. Then $\lambda^n(x, y, z) = z/h$. If A is the area of the basis, then $\int_T \lambda^n = A \int_0^h dz\, (1 - z/h)^2\, z/h = hA/12$, hence $\int_T \lambda^n = \mathrm{vol}(T)/4$. Same thing for a triangle: basis on x axis, height h, etc. You may prefer a proof by recurrence on the dimension. Anyway, once in possession of these basic symmetry results, further computations (cf. Exer. 3.10) simplify considerably.

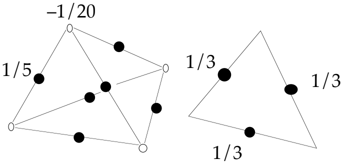
3.10. First compute $I_{nn} = \int_T (\lambda^n)^2 = A/h^2 \int_0^h dz\, (h - z)^2\, z^2/h^2 = \mathrm{vol}(T)/10$, and similarly, obtain the equality $I_{nnn} = \int_T (\lambda^n)^3 = \mathrm{vol}(T)/20$. Then $I_{nn} = \int_T \lambda^n (1 - \sum_{m \neq n} \lambda^m) = \mathrm{vol}(T)/4 - 3\, I_{nm}$), hence $I_{nm} = \mathrm{vol}(T)/20$. And so on. The general formula,

$$\int_T (\lambda^\ell)^i\, (\lambda^m)^j\, (\lambda^n)^k = 6\ \mathrm{vol}(T)\ i!\ j!\ k!\ /(i+j+k+3)!$$

(cf. [Sf]), may save you time someday. (Thanks to (13) and Remark 3.3, this is enough to sum any polynomial over T.) The analogue on faces [SF] is $\int_f (\lambda^m)^i (\lambda^n)^j = 2\ \text{area}(f)\ i!\ j!/(i + j + 2)!$.

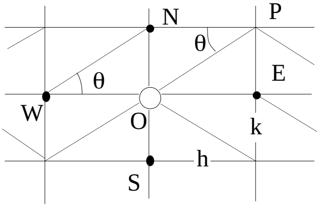
3.11. On [0, 1], the function $w = x \to 2x^2 - x$ behaves as requested, that is, $w(0) = w(1/2) = 0$ and $w(1) = 1$. Therefore, $w_{nm} = 4\,\lambda^n\lambda^m$ and $w_{nn} = \lambda^n(2\lambda^n - 1)$.

3.12. Weights $1/5$ at mid-edges and $-1/20$ at nodes. For the triangle, amusingly, $1/3$ at mid-edges and 0 at nodes (inset).



3.13. For all $\boldsymbol{\psi}$, $0 \le (\mathbf{M}(\boldsymbol{\varphi} + \lambda\boldsymbol{\psi}), \boldsymbol{\varphi} + \lambda\boldsymbol{\psi}) = 2\lambda\,(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\psi}) + \lambda^2\,(\mathbf{M}\boldsymbol{\psi}, \boldsymbol{\psi})$, so $(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\psi}) = 0$ for all $\boldsymbol{\psi}$, which implies $\mathbf{M}\boldsymbol{\varphi} = 0$. ◊

3.14. Use the cotangent formula (23). There are two right angles in front of edge OP, hence the nullity of the coefficient A_{OP}. Formula (24) comes immediately in the case of node pairs like O–N, O–E, etc., by using (23) and the relation $\tan\theta = k/h$, where θ is the angle shown in the inset. Note how (23) gives the same value for matrix entries corresponding to such pairs whatever the diagonal along which one has cut the rectangular cell. (Further study: Consider orthogonal, but not uniformly spaced, grids. Generalize to dimension 3.)



3.15. No, the regular tetrahedron is not a "space-filling" solid. Its dihedral angle, easily computed, is about 70°32′, hence a mismatch. See [Ka] or [Si] on such issues.

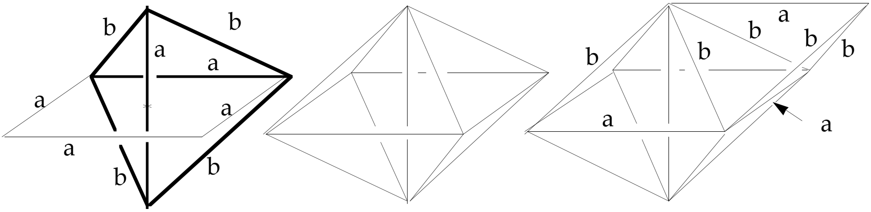


FIGURE 3.8. Left: The tetrahedron (in thick lines). Middle: Assembly of four copies of it into an octahedron, by rotation around the middle vertical pole. Right: Sticking a fifth copy to the upper right flank. A sixth copy will be attached to the lower left flank in the same way, hence a paving parallellepiped.

3.16. Cf. Fig. 3.8. Length b should equal $a\sqrt{2/3}$. There is numerical evidence [MP] that such tetrahedra yield better accuracy in some computations than the standard "cubic" grid, subdivided as Fig. 3.7 suggests. (A suitable combination of regular octahedra and tetrahedra, by which one *can* pave [Ka], may also be interesting in this respect.)

3.17. Call φ_1 and φ_2 the solutions corresponding to μ_1 and μ_2. Then $R_1^{-1} = \inf\{\int_D \mu_1 |\nabla\varphi|^2 : \varphi \in \Phi^1\} \le \int_D \mu_1 |\nabla\varphi_2|^2 \le \int_D \mu_2 |\nabla\varphi_2|^2 = R_2^{-1}$.

3.18. With respect to some origin, map D to D_λ by $x \to \lambda x$, with $\lambda > 0$, and assign to D_λ the permeability μ_λ defined by $\mu_\lambda(\lambda x) = \mu(x)$. If φ is an admissible potential for the problem on D, then φ_λ, similarly defined by $\varphi_\lambda(\lambda x) = \varphi(x)$, is one for the problem on D_λ. Changing variables, one sees that $\int_{D_\lambda} \mu_\lambda |\nabla\varphi_\lambda|^2 = \int_D \lambda\mu\, |\nabla\varphi|^2$, so it all goes as if μ had been multiplied by λ (in vacuum, too!). Hence the result by Exer. 3.17.

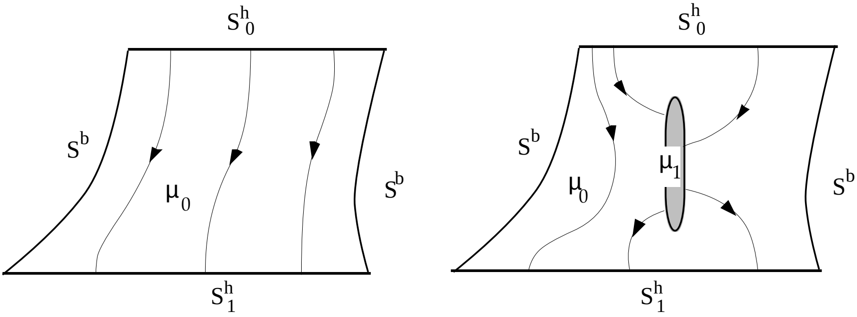


FIGURE 3.9. Exercise 3.19. How the presence in the domain under study of a highly permeable part ($\mu_1 >> \mu_0$), even of very small relative volume, is enough to distort the field. (Two-dimensional drawing, for clarity. In the case of Fig. 3.1, a similar effect would be achieved by putting a high-μ thin sheet inside D.)

3.19. Since both φ_1 and φ_2 belong to Φ^I, one can set $\varphi' = \varphi_1 - \varphi_2$ in both equations (25), and subtract, which yields

$$(\varphi_1, \varphi_1 - \varphi_2)_1 + (\varphi_2, \varphi_2 - \varphi_1)_2 = 0.$$

Therefore,

$$\|\varphi_1 - \varphi_2\|_1^2 = \int_D (\mu_1 - \mu_2)\, \nabla\varphi_2 \cdot \nabla(\varphi_2 - \varphi_1) \equiv \int_D \mu_1(1 - \mu_2/\mu_1)\, \nabla\varphi_2 \cdot \nabla(\varphi_2 - \varphi_1)$$

hence $\|\varphi_1 - \varphi_2\|_1 \le C(\mu)\, \|\varphi_2\|_1$ by Cauchy–Schwarz, where $C(\mu)$ is an upper bound for $|1 - \mu_2/\mu_1|$ over D. Hence the continuity with respect to μ (a small uniform variation of μ entails a small change of the solution), but only, as mathematicians say, “in the L^∞ norm”. The result cannot be improved in this respect: A large variation of μ, even concentrated on a small part of the domain, can change the solution completely, as Fig. 3.9 suggests.

REFERENCES

[BP] A. Berman, R.J. Plemmons: **Nonnegative Matrices in the Mathematical Sciences,** Academic Press (New York), 1979.

[BR] J.R. Bunch, D.J. Rose: **Sparse Matrix Computations,** Academic Press (New York), 1976.

[Ci] P.G. Ciarlet: **The Finite Element Method for Elliptic Problems**, North-Holland (Amsterdam), 1978.

[CR] P.G. Ciarlet, P.A. Raviart: “General Lagrange and Hermite interpolation in $\mathbb{R}^n$ with applications to finite element methods”, **Arch. Rat. Mech. Anal., 46** (1972), pp. 177–199.

[Cr] P.G. Ciarlet: **Introduction à l'analyse numérique matricielle et à la programmation**, Masson (Paris), 1982.

[Co] N.A. Court: **Modern pure solid geometry,** Chelsea (Bronx, NY), 1964. (First edition, Macmillan, 1935.)

[DL] J. Deny, J.L. Lions: "Les espaces du type de Beppo Levi", **Ann. Inst. Fourier**, **5** (1953–1954), pp. 305–370.

[Ge] P.L. George: **Génération automatique de maillages. Applications aux méthodes d'éléments finis,** Masson (Paris), 1990.

[GL] A. George, J.W. Liu: **Computer Solution of Large Sparse Positive Definite Systems**, Prentice-Hall (Englewood Cliffs, NJ), 1981.

[Gv] G.H. Golub, C.F. Van Loan: **Matrix Computations**, North Oxford Academic (Oxford) & Johns Hopkins U.P. (Baltimore), 1983.

[Gr] P. Grisvard: **Elliptic Problems in Nonsmooth Domains,** Pitman (Boston), 1985.

[Jo] C.R. Johnson: "Inverse M-matrices", **Lin. Alg. & Appl., 47** (1982), pp. 195–216.

[Ka] J. Kappraff: **Connections**: The Geometric Bridge Between Art and Science, McGraw-Hill (New York), 1991.

[La] I. Lakatos: **Proofs and Refutations,** Cambridge U.P. (Cambridge), 1976.

[MP] P. Monk, K. Parrott, A. Le Hyaric: “Analysis of Finite Element Time Domain Methods in Electromagnetic Scattering”, Int. report 96/25, Oxford University Computing Laboratory (Oxford, U.K.), 1996.

[Na] R. Nabben: "Z-Matrices and Inverse Z-Matrices", **Lin. Alg. & Appl., 256** (1997), pp. 31–48.

[Sf] P.P. Silvester, R. Ferrari: **Finite Elements for Electrical Engineers,** Cambridge University Press (Cambridge), 1991.

[Si] J. Sivardière: **La symétrie en Mathématiques, Physique et Chimie**, Presses Universitaires de Grenoble (Grenoble), 1995.

[SF] G. Strang, G.J. Fix: **An Analysis of the Finite Element Method,** Prentice-Hall (Englewood Cliffs, NJ), 1973.

[Va] R.S. Varga: **Matrix Iterative Analysis**, Prentice-Hall (Englewood Cliffs, NJ), 1962.

[We] T. Weiland: "Time Domain Electromagnetic Field Computation with Finite Difference Methods", **Int. Journal of Numerical Modelling, 9** (1996), pp. 295–319.

CHAPTER 4

The Approximate Scalar Potential: Properties and Shortcomings

The question now on our agenda: Assuming we have solved the linear system (20) in Chapter 3, and thus obtained the vector $\boldsymbol{\varphi}$ of nodal potentials, to what extent is the approximate solution $\varphi_m = \sum_{n \in \mathcal{N}} \boldsymbol{\varphi}_n \lambda^n$ satisfactory as a representation of the field? On the side of h, all goes well: Setting $h_m = \text{grad } \varphi_m$, we have $\text{rot } h_m = 0$ as well as $n \times h_m = 0$ on S^h and $\int_c \tau \cdot h_m = I$, all that by construction. Errors are on the side of b: We lose solenoidality of $b_m = \mu h_m$, since not all test functions have been retained. Some measure of flux conservation still holds, however, and we'll see in which precise sense. When the mesh is refined, we expect to recover $\text{div } b = 0$ "at the limit"; this is the issue of *convergence*. But how *fast* do h_m and b_m converge toward h and b, and how *far* apart are they in energy? These are related questions, but the latter is more difficult and will not be resolved before Chapter 6. Last, there is a property of the true solution, expressed by the so-called *maximum principle*, which may be preserved to some extent, provided the mesh is carefully devised, and Voronoi–Delaunay meshes seem to be adequate in this respect.

4.1 THE "*m*-WEAK" PROPERTIES

Last chapter, we defined "discrete" or "*m*-weak" solenoidality as the property

$$(1) \qquad \int_D b \,.\, \text{grad } \varphi' = 0 \quad \forall\, \varphi' \in \Phi^0_{\ m} ,$$

that is, with nodal finite elements, $\int_D b \cdot \text{grad } \lambda^n = 0$ for all n in the subset $\mathcal{N}_0 = \mathcal{N} - \mathcal{N}(S^h)$. We shall dub *active* the nodes of this set, which includes inner nodes (i.e., not on S) and surface nodes interior to S^b, but not those at the boundary common to S^b and S^h, where nodal values are imposed. Active nodes are those which bear an unknown degree of freedom, and each of them corresponds to a row of the submatrix ${}^{00}\mathbf{M}$ of Eq. (3.20).

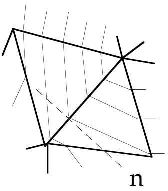
4.1.1 Flux losses

Condition (1) is much less stringent than weak solenoidality, so what is left of $\operatorname{div} b = 0$? In the worst case, nothing: h_m is constant inside a tetrahedron, so if μ varies, then $\operatorname{div}(\mu\, h_m) = \nabla\mu \cdot h_m$, which has no reason to vanish. But this is easily cured: Replace μ, either *before* the computation of **M** by (3.19), or at the stage we consider now, by a mesh-wise constant function $\bar{\mu}$, equal to $(\int_T \mu)/\mathrm{vol}(T)$ on tetrahedron T. Then $b_m = \bar{\mu}\, h_m$, being mesh-wise constant, is solenoidal inside each T. Can we expect its normal jumps $[n \cdot b_m]_f$, which are a priori constant over each face of the mesh, to vanish, all of them? By no means, because that would make b_m divergence-free and enforce $n \cdot b_m = 0$ on S^b. Thereby, all the equations of the continuous model would be satisfied by the discrete model (in the case of a mesh-wise constant μ), which would then yield the right solution, and such miracles are not to be expected. Jumps of $n \cdot b_m$ don't vanish, so there is a "loss of induction flux", equal to the integral of this jump, at each inter-element boundary.

This prediction is confirmed by comparatively counting these lost fluxes and the equations. We have N nodes, E edges, F faces and T tetrahedra. By a famous result in topology to which we shall return, one has

$$N - E + F - T = \chi,$$

where χ, the *Euler–Poincaré* constant, which a priori seems to depend on *m*, is actually determined by the global topological properties of D. It's a small integer, typically 1 for simple domains. (**Exercise 4.1:** Compute χ for a single tetrahedron and for a meshed cube.) Assume a typical mesh, made by first generating hexahedra, then chopping them into six tetrahedra each (cf. Exer. 3.7). Then $T \sim 6N$, and $F \sim 12N$, since each tetrahedron has four faces that, most of them, belong to two tetrahedra, hence $E \sim 7N$. Having about N degrees of freedom, we can't satisfy $F \sim 12N$ constraints—but wait, are there really F lost fluxes to cancel? No, because the fluxes through faces of a same tetrahedron add to zero, since $\operatorname{div} b_m = 0$ inside. So there are about $F - T$ *independent* lost fluxes to consider. Still, this is about $E - N$, much larger than N.



We must therefore accept nonzero jumps of $n \cdot b$ through faces as a weakness inherent in the method: b_m is not, as one says, “div-conformal”. (The latter expression does not mean “solenoidal”: A field is *div-conformal* when its normal component is continuous through all surfaces, which is a weaker condition.) So the approximate solution fields will *not* satisfy the “law of tangents” of (2.5), and indeed, in two-dimensional simulations, it's fairly common to see flux lines that behave “wrongly”, as shown in the inset, staying on the same side of the normal to an edge when going from one triangle to the next.[1] We are used to that nowadays, knowing this is the price to pay for having a *finite* system of equations to solve instead of the *infinite* system that Problem (3.1) represented, and we can rely on the assurance that with refinement of the mesh, such non-physical behavior of flux lines will disappear (a proof to this effect will come). But in the early days of the finite element method, this feature was met with harsh criticism, touching off a controversy, some echoes of which can be found in [CS, EF, FE].

Remark 4.1. A nonzero jump of $n \cdot b$ at element interfaces is equivalent to the presence of a magnetic charge density $[n \cdot b]$ there. Thus, b_m can be described as the induction field that would appear if these fictitious charges were really present, in addition to the external sources of the field. ◊

Let us therefore try to assess the damage as regards these inevitable flux losses, or spurious charges. This will depend on an interpretation of each component of the vector $\mathbf{M}\boldsymbol{\varphi}$, as follows.

First, let us consider any DoF vector $\boldsymbol{\varphi}$, *not* related to the solution. We form the mesh-wise affine function[2] $\varphi = p_m(\boldsymbol{\varphi}) = \sum_{n \in \mathcal{N}} \boldsymbol{\varphi}_n \lambda^n$, and the vector field $b = \bar{\mu}$ grad φ. Note that b is mesh-wise constant. By the very definition of $\mathbf{M}$,

$$(\mathbf{M}\boldsymbol{\varphi})_n \equiv \sum_{m \in \mathcal{N}} \mathbf{M}_{nm}\, \boldsymbol{\varphi}_m = \int_D b \cdot \operatorname{grad} \lambda^n.$$

Integrating by parts on each tetrahedron, and summing up, one obtains (sorry for the clash of n's):

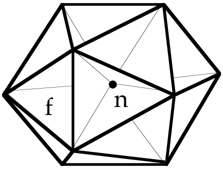
$$(2) \qquad (\mathbf{M}\boldsymbol{\varphi})_n = \sum_{f \in \mathcal{F}(n)} \int_f [n \cdot b]\, \lambda^n = \tfrac{1}{3} \sum_{f \in \mathcal{F}(n)} \int_f [n \cdot b]\,,$$

[1] Curiously, the bending of flux lines across mesh edges, which is just as “unphysical” in a homogeneous region, was not regarded as scandalous when the normal was properly crossed by flux lines.

[2] The notation p_m has been introduced in Subsection 3.3.3.

because the jump is constant over f, and the average of λ^n over f is equal to 1/3 (Exer. 3.9). For further reference, let's give a name to $(\mathbf{M}\boldsymbol{\varphi})_n$, the component of $\mathbf{M}\boldsymbol{\varphi}$ at node n, and call it the *flux loss at*, or *about* node n, as regards b. After (2), this loss is one-third of the sum of flux losses at faces that have n as common node (face set $\mathcal{F}(n)$).

There is another interpretation of this flux loss, for which it will be convenient to distinguish inner nodes and surface nodes. If n is an interior node, the faces of $\mathcal{F}(n)$ are the "inner faces" of the cluster (the opaque ones in the inset drawing). Since b is mesh-wise constant, $\int_{\partial T} n \cdot b = 0$ for all tetrahedra. The sum of these terms over all tetrahedra of the cluster is also the sum of the outward flux through the cluster's boundary and of the inner flux losses. Therefore, the flux loss at n is one third of the flux entering its cluster. The same argument works if n belongs to the surface: Then n lies on the polyhedral boundary of its own cluster (Fig. 4.1), and the flux loss at n is one-third of the flux entering the "polyhedral cap" of n, as sketched in Fig. 4.1, right.



Now, consider the case when $\boldsymbol{\varphi}$ is the solution of the discrete problem, Eq. (3.20). Row n of this linear system corresponds, equivalently, to

(3) $$\int_{D_n} b_m \cdot \operatorname{grad} \lambda^n = 0,$$ (3') $$(\mathbf{M}\boldsymbol{\varphi})_n \equiv \sum_{m \in \mathcal{N}(n)} \mathbf{M}_{n\,m}\, \boldsymbol{\varphi}_m = 0.$$

So discrete solenoidality entails the cancellation of all flux losses of b_m at active nodes. Therefore, by what precedes, the *fluxes of* b_m *through the surface of the cluster of each inner node* and *through the cap of each active boundary node* must vanish.

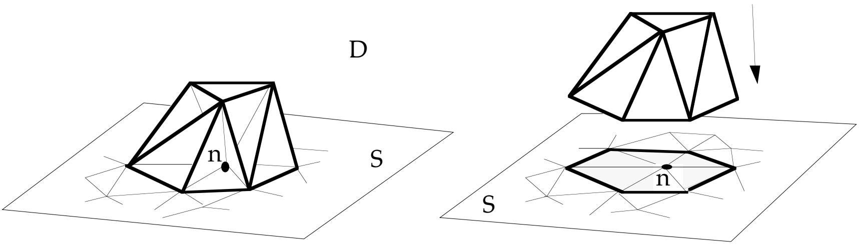


FIGURE 4.1. Left: Cluster of tetrahedra around a boundary node (D here is above the triangulated plane). Inner faces of the cluster are opaque; others are transparent. Right: Dissecting the cluster's boundary into a "patch" of boundary triangles around n, and a "cap" of inside faces.

Note that we *don't* find $n \cdot b_m = 0$ on faces of S^b. Actually, we had no reason to entertain such hopes, since there are about twice[3] as many faces as nodes on S^b, thus not enough equations to cancel all these fluxes.

It's very tempting to try and combine these results by merging clusters of different nodes into larger clusters, and to say, "Well, just as the flux of the true solution b is null for all closed surfaces inside D (remember, this is the integral interpretation of weak solenoidality), the flux of b_m through polyhedral surfaces made of mesh-faces will vanish." Fine guesswork . . . but wrong, as the following exercise will show.

Exercise 4.2. In dimension 2, suppose b_m satisfies (3) over a domain that contains the two "extended clusters" of Fig. 4.2. Show that the flux through Σ_1 or Σ_2 does not vanish.

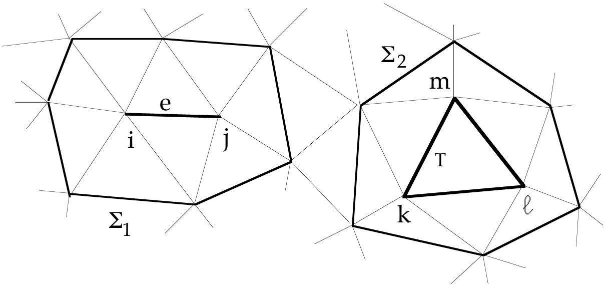


FIGURE 4.2. Part of a 2D mesh m by which a discretely solenoidal induction b_m has been computed. Although the flux of b_m through cluster boundaries vanishes, it does not on the boundaries Σ_1 or Σ_2 of extended clusters, that is, unions of clusters of node subsets such as {i, j} or {k, ℓ, m}.

And yet there is something correct in this intuition. But we need relatively sophisticated new concepts to develop this point.

4.1.2 The dual mesh, and which fluxes are conserved

First, the *barycentric refinement* of a simplicial mesh m. This is a new simplicial mesh, which we shall denote by $m/2$, obtained as suggested by Fig. 4.3: Add one node at each mid-edge and at the center of gravity of each face and each tetrahedron, subdivide, and add edges and faces as

[3] It's Euler–Poincaré again, for surfaces this time: $N - E + F = \chi$, and $2E \sim 3F$. Things may seem different in dimension 2, where there are almost as many boundary nodes as boundary *edges* (the difference is the number of connected components of S^h). Still, even when these numbers coincide, there is no reason to expect fluxes at boundary edges to be exactly zero.

required. Visualizing it in dimension 2 is easy (Fig. 4.3), but it takes some imagination in three dimensions.

Next, the *dual* mesh (the primitive one then being referred to as the *primal* mesh). The dual mesh is not a simplicial mesh, but what can be called a "cellular" tessellation, the cells being polyhedra, polyhedral surfaces, broken lines (Fig. 4.4), and points. The 3-cells, one for each primal node, are clusters of tetrahedra around n, but tetrahedra of the subdivided mesh $m/2$, not of m. Such a shrunk cluster (see Fig. 4.3 for one in dimension 2) is informally called a "box". Fig. 4.4 shows a part of the box around n, the part that intersects tetrahedron T. Two-cells are associated with edges: The 2-cell of edge e is the union of all faces of $m/2$ that contain the midpoint of e, but none of its extremities (Fig. 4.4). Note that it's not a *plane* polygon (though its parts within each tetrahedron are plane). One-cells, associated with faces, are unions of the two segments which join the barycenter of a face to those of the two tetrahedra flanking it. And 0-cells, the nodes of the dual mesh, are centers of gravity of the primal tetrahedra.

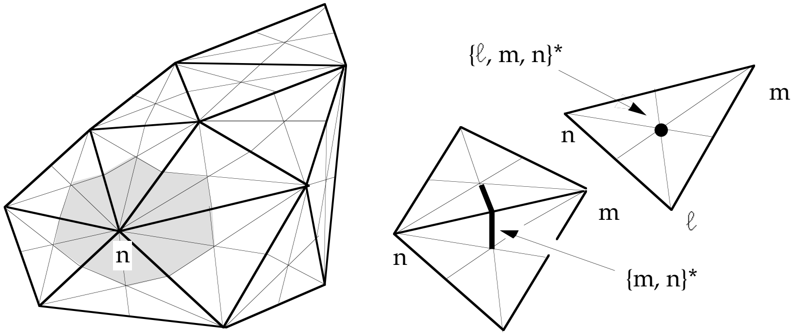


FIGURE 4.3. Barycentric refinement of a 2D mesh. Thick edges are those of the primitive mesh. Shaded, the "box", or "dual 2-cell" around node n. Right: dual cells (for 3D examples, see Fig. 4.4).

There is thus a perfect duality between the two meshes, p-simplices of m being in one-to-one correspondence with (d – p)-cells of the dual mesh, where d is the spatial dimension. We may denote the dual mesh by m^*, and play on this notation: the box around n can be denoted n* (but we'll call it B_n), the corolla of small faces around edge e is e*, etc. Note that—this is clear in Fig. 4.3, but valid in all dimensions—dual p-cells intersect along dual q-cells, with $q < p$. In particular, the common boundary of two adjacent boxes is the dual 2-cell of the edge joining their nodes. (**Exercise 4.3:** When do two dual 2-cells intersect at a point? Along a line?)

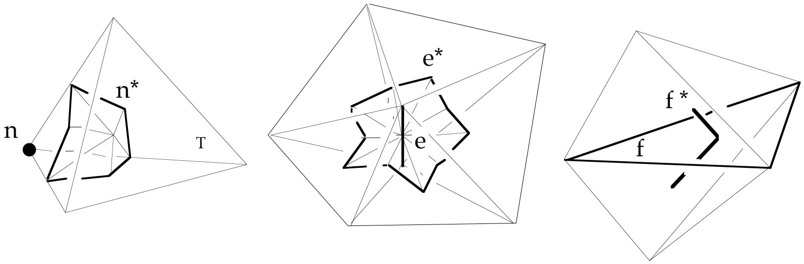


FIGURE 4.4. Cells of the dual mesh m^*: All dual p-cells are unions of p-simplices of the barycentric refinement of m. From left to right, a part of the box n^*, a 2-cell e^* around edge e, and a 1-cell f^* through face f.

Finally, for further use, let us define something we shall call, for shortness, *m^*-surfaces:* Surfaces, with or without boundary, made of dual 2-cells. Box surfaces are m^*-surfaces (after removing the part that may lie in S), and the surface of a union of boxes is one, too. Similarly, of course, we have *m^*-points* (nodes of the dual mesh, i.e., centers of tetrahedra), *m^*-lines* (made of dual 1-cells), and *m^*-volumes* (unions of boxes).

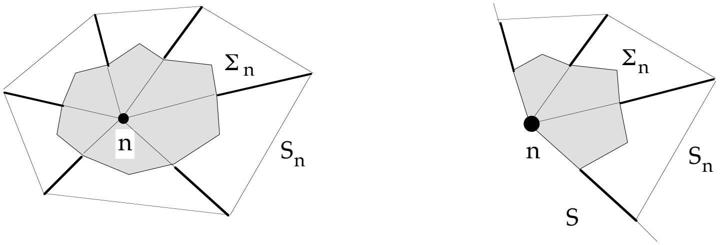


FIGURE 4.5. Comparing the fluxes through Σ_n and S_n. (Beware, this is a 2D representation, in which surfaces Σ_n, S_n, S appear as lines.)

With this, we can refine our interpretation of $\mathbf{M}\boldsymbol{\varphi}$. Again, given some DoF vector $\boldsymbol{\varphi}$, form $\varphi = p_m(\boldsymbol{\varphi})$ and $b = \overline{\mu}$ grad φ. Then,

Proposition 4.1. *The term* $(\mathbf{M}\boldsymbol{\varphi})_n = \int_D b \cdot \operatorname{grad} \lambda^n$, *or flux loss of* b *about* n, *is the inward flux of* b *across the surface of the box around* n, *if* n *is an inner node, and across the "m^*-cap" of Fig. 4.5, if* n *is a surface node.*

Proof. Since b is divergence-free inside tetrahedra, the difference between its fluxes through the cluster surface S_n and through the box surface Σ_n is due to flux leaks at the parts of inner faces of the cluster which are in the shell between Σ_n and S_n (thick lines in the 2D drawing). But for each such face, exactly two-thirds of its area are there (in dimension 2, and on the drawing, one-half of the edge length). And since jumps of $n \cdot b$ are constant over each face, the total of these flux leaks in the shell is thus two-thirds of the total of flux leaks in the cluster. The remaining third is thus the flux through Σ_n. Same reasoning if n belongs to the boundary, Σ_n now being the "m^*-cap" over n (smaller than the cap of Fig. 4.1) made of dual 2-cells. ◊

Remark 4.2. One cannot overstress the importance of having a *barycentric* subdivision to get this result; the uniformity of the ratio 2/3 between areas was essential. ◊

In other words, $(\mathbf{M}\,\varphi)_n$ is the part of the flux of b entering box B_n that comes from other boxes, and the entries of **M** govern fluxes between boxes in a very simple way: B_m gives $\mathbf{M}_{n\,m}(\varphi_m - \varphi_n)$ to B_n. The finite element method thus appears as simple bookkeeping of induction flux exchanges between boxes, or "finite volumes", in which the computational domain has been partitioned. Exercise 4.4 at the end is an invitation to follow up on this idea.

With this, we can return to the characterization of b_m: Its flux through the box surface of an inner node, or through the small cap of an active surface node, vanishes. Since inter-box boundaries are always *inside* tetrahedra, where b_m is solenoidal, there are no flux leaks at such interfaces, so we may aggregate boxes, and the flux entering such an aggregate is the sum of flux losses at all nodes inside it. Therefore, the flux of b_m will vanish across polyhedral surfaces of two kinds: m^*-surfaces that enclose one or several boxes around inner nodes, and "extended m^*-caps", covering one or several S^b-nodes.

To make sense out of this, let's compare the present "discrete" situation to the "continuous" one. In the latter, the flux of the true solution b is null for all closed surfaces inside D which enclose a volume (this is the integral interpretation of weak solenoidality), and also, since $n \cdot b = 0$ on S^b, for all surfaces with boundary which, with the help of S^b, enclose a volume. So it goes exactly the same in the discrete situation, except that surfaces must be made of dual 2-cells. All this cries out for the introduction of new definitions, if only for ease of expression:

Definition 4.2. *Let* C *be a part of* D. *A surface in* $\overline{D}$ *is* closed modulo C *if its boundary is in* C. *A surface* Σ bounds modulo C *if there is a volume* Ω *contained in* $\overline{D}$ *such that* $\partial\Omega - \Sigma$ *is in* C.

Same concepts[4] for a line σ, which is *closed mod* C if its end-points are in C, and *bounds mod* C if there is a surface Γ such that $\partial\Gamma - \sigma$ is in C.

[4] Lines and surfaces can be in several pieces, but trying to formalize that, via the concept of *singular chain*, would lead us too early and too far into *homology* [GH, HW]. (Don't confuse the present notion of closedness with the topological one; cf. A.2.3.)

Figure 4.6 gives a few examples. (As one sees, what is often called informally a "cutting surface" or "cut" is a closed surface (mod. something) that doesn't bound. Cf. Exer. 2.6.)

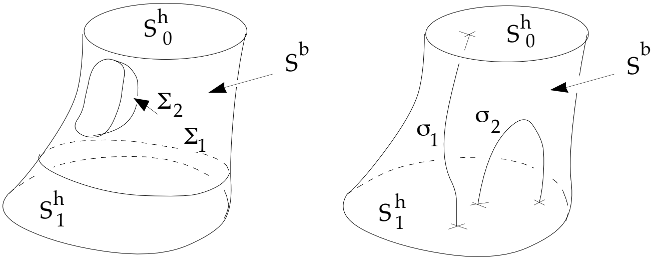


FIGURE 4.6. Left: Surfaces Σ_1 and Σ_2 are both closed modulo S^b, but only the latter bounds mod S^b. Right: Lines σ_1 and σ_2 are both closed modulo S^h, but only the latter bounds mod S^h.

Thus equipped, we can reformulate our findings. The conditions about b, in the continuous model, were that $\int_\Sigma n \cdot b = 0$ for all surfaces Σ which bound modulo S^b. In contrast, and in full recovery from our earlier fiasco, we have obtained this:

Proposition 4.2. *The discretized field satisfies* $\int_\Sigma n \cdot b_m = 0$ *for all* m^**-surfaces* Σ *which bound modulo* S^b.

Exercise 4.5. What if there are flux sources inside the domain (cf. Exers. 2.8 and 3.1) ?

4.1.3 The flux through S^h

The aim of our modelling was supposed to be the reluctance $R = I/F$, so we need the induction flux F through the domain in terms of $\boldsymbol{\varphi}$.

Following the hint of Remark 3.2, one has

$$\int_D \mu \ |\text{grad}\ \varphi|^2 = \int_D b \cdot \text{grad}\ \varphi = \int_S n \,.\, b\,\varphi = I \int_{S^h_1} n \cdot b = I\,F,$$

if φ is the exact solution. Since $(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi})$ is by construction the best approximation we have for $\int_D \mu \ |\text{grad}\ \varphi|^2$, the best estimate for F is the mesh-dependent ratio $F_m = (\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi})/I$, hence an approximation R_m of the reluctance in our model problem: $R_m = I^2/(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi})$, of which we may remark it is *by default*—from below—since $(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi})$, being obtained by minimizing on too small a space, exceeds the infimum of $\int_D \mu \ |\text{grad}\ \varphi|^2$. We'll return to this in Chapter 6.

Developing $(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi})$, we see that

$$F_m = I^{-1} \sum_{n \in \mathcal{N}} (\mathbf{M}\boldsymbol{\varphi})_n \, \boldsymbol{\varphi}_n = \sum \{ (\mathbf{M}\boldsymbol{\varphi})_n : n \in \mathcal{N}(S^h_1)\},$$

since $(\mathbf{M}\boldsymbol{\varphi})_n = 0$ for all nodes $n \in \mathcal{N}_0$ and $\boldsymbol{\varphi}_n = 0$ or I for n in $\mathcal{N}(S^h_0)$ or $\mathcal{N}(S^h_1)$. The flux is therefore approximated by the sum of flux losses at points of S^h_1, which is equal, after Prop. 4.1 (aggregate the boxes of all points of S^h_1), to the flux at the surface Σ of Fig. 4.7 (seen as a broken line in this 2D sketch), obtained by merging the small caps (m^*-caps) of all nodes of S^h_1.

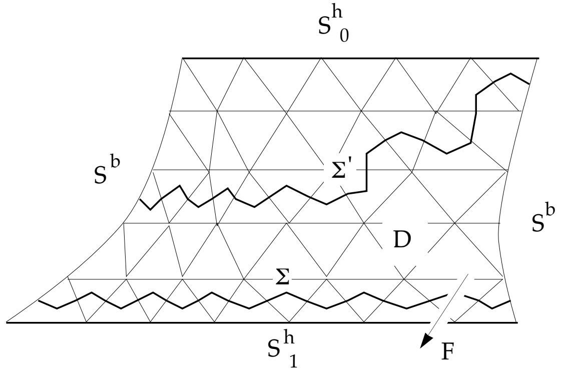


FIGURE 4.7. Where to compute the flux of b_m.

Exercise 4.6. What about m^*-surfaces like Σ' in Fig. 4.7? Show that the flux of b_m is the same through all of those which are homologous to Σ, in the sense of Exer. 2.6.

As one sees, the "variationally correct" approximation of the flux is not what a naive approach would suggest, that is, $\int_{S^h_1} \mu \, \partial_n(p_m(\boldsymbol{\varphi}))$, but the same integral taken over Σ (or any m^*-surface homologous to it) instead of S^h_1. One should not, in consequence, use $\mu \, \partial_n(p_m(\boldsymbol{\varphi}))$ as approximation for the normal induction, but treat the latter, for all purposes, as a surface distribution $\varphi' \to \int_S \mu \, \partial_n\varphi \, \varphi'$, and use the scalar product $(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi}')$ as approximation of this integral, with $\boldsymbol{\varphi}'_n = \varphi'(x_n)$. (Cf. Exer. 4.8.)

Exercise 4.7. In which sense is $\mathbf{M}$ a discrete analogue of the differential operator $-\operatorname{div}(\mu \operatorname{grad}\,)$?

Exercise 4.8. Suppose one solves a problem similar to our model problem, but with φ, for some reason, equal to a given boundary data φ^h on S^h. Write down the best estimate of the functional $\int_D \mu \, |\operatorname{grad} \varphi|^2$ in terms of the DoFs on S^h. (This way of expressing the energy inside a region in terms of boundary values of the field is a very useful procedure, known as "static condensation" in mechanics. Can you find a better denomination, more germane to electromagnetism?)

4.2 THE MAXIMUM PRINCIPLE

If $\Delta\varphi = 0$ in some domain, φ cannot reach its maximum or minimum elsewhere than on the *boundary* of the domain. This is the *maximum principle* for harmonic functions. A similar property holds for the magnetic potential, and some of it may be retained at the discrete level.

4.2.1 Discrete maximum principle

Let's recall the proof. First (an easy assignment):

Exercise 4.9. Show that the function $y \to (4\pi\ |x - y|)^{-1}$ is harmonic in $E_3 - \{x\}$.

Next, suppose $\text{div}(\text{grad}\ \varphi) = 0$ in some domain D. Consider two spheres $S(x, r)$ and $S(x, R)$ contained in D, both centered at x, with radiuses $r < R$, and call v_x the function $y \to (4\pi)^{-1}(1/|x - y| - 1/R)$. Integrate by parts in the domain O between the two spheres, which can be done in two ways. First, lifting the grad off v_x,

$$\int_O \text{grad}\ \varphi \cdot \text{grad}\ v_x = \int_{\partial O} n\ .\ \text{grad}\ \varphi\ \ v_x$$

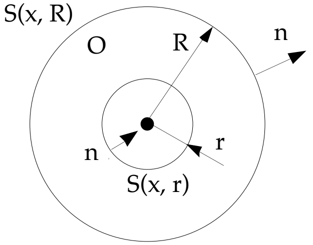
$$= \frac{1}{4\pi}\,(r^{-1} - R^{-1})\int_{S(x,\ r)} n \cdot \text{grad}\ \varphi =$$

$$= -\frac{1}{4\pi}\,(r^{-1} - R^{-1})\int_{B(x,\ r)} \text{div}(\text{grad}\ \varphi) = 0,$$

and then, the other way around,

$$\int_O \text{grad}\ \varphi \cdot \text{grad}\ v_x =$$

$$\frac{1}{4\pi}\,[R^{-2}\int_{S(x,\ R)} \varphi - r^{-2}\int_{S(x,\ r)} \varphi].$$



Letting r tend to 0, one finds, finally,

$$(4) \qquad \varphi(x) = [\int_{S(x,\ R)} \varphi]/(4\pi R^2),$$

a useful representation formula, which says that $\varphi(x)$ is equal to the average of φ over the sphere $S(x, R)$. Of course, φ cannot be extremal at x without contradicting this. Hence the maximum principle.

In the case of a non-unifom, but positive permeability, the magnetic potential enjoys a similar property,[5] but the proof ([GT], Chapter 3) is no

[5] This is the basis of Earnshaw's famous result [Ea]: "A charged particle in empty space cannot remain in stable equilibrium under electrostatic forces alone, or alternatively there can be no maximum or minimum of the potential at points free of charge density." (Quoted from [Sc].)

longer elementary. It relies on the intuitive idea that if there was an isolated maximum at x, normal fluxes $\mu\, n \cdot \text{grad}\, \varphi$ would all be negative on the surface of a small sphere centered at x, thus contradicting flux conservation.

It would be highly unphysical and quite embarrassing indeed if a similar property did not hold for the computed discrete potential, that is, if the potential could surpass I, or be negative, inside D or on S^b. *This won't happen if all extra-diagonal terms of* **M** *are nonpositive*. Indeed, one can interpret Eq. (3') as the discretized counterpart of (4), as follows:

$$(4') \qquad \boldsymbol{\varphi}_n = \sum \{ m \in \mathcal{N}(n), m \neq n : (-\mathbf{M}_{nm}/\mathbf{M}_{nn})\, \boldsymbol{\varphi}_m \},$$

showing how $\boldsymbol{\varphi}_n$ is the weighted average of neighboring nodal values. The sum of weights $\sum_{m \neq n} - \mathbf{M}_{nm}/\mathbf{M}_{nn}$ is always equal to 1 (cf. Remark 3.6), but what counts here is the positivity of each of them: If all weights are positive in (4'), then $\boldsymbol{\varphi}_n$ is strictly contained in the interval formed by the minimum and the maximum values of DoFs $\boldsymbol{\varphi}_m$ around it, and thereby,

Proposition 4.3 ("discrete maximum principle"). *If no extra-diagonal entry of* **M** *is positive, the maximum of the approximate potential* φ_m *on any cluster* D_n *is reached on its boundary.*

As an immediate corollary, the extrema of φ_m on $\overline{D}$ are reached only on the boundary, which is the discrete version of the principle recalled at the beginning of this section.

Exercise 4.10. Prove $0 \leq \boldsymbol{\varphi}_n \leq I$ directly from Eq. (3.20), $^{00}\mathbf{M}\,^{0}\boldsymbol{\varphi} = -\,^{01}\mathbf{M}\,^{1}\boldsymbol{\varphi}^{I}$, by using Prop. 3.6.

Nonpositivity of extra-diagonal terms thus appearing as a desirable property, when does it hold, and how can it be achieved? An "acute" mesh (one with no dihedral angle larger than 90°) is enough, as we remarked earlier. But this is a sufficient, not a necessary, condition. What constitutes a necessary and sufficient condition in this respect does not seem to be known, although there is a way in which Voronoi–Delaunay[6] meshes satisfy this requirement. Such meshes have stirred interest in computational electromagnetism [C&, SD, Z&], perhaps because of their apparent (but still not well understood) connection with "network methods" [He], or "finite volume methods" (cf. [Va], p. 191). So let us digress on this for a while.

[6] "Delaunay: This is the French transliteration of the name of Boris Nikolaevitch Delone, who got his surname from an Irish ancestor called Deloney, who was among the mercenaries left in Russia after the Napoleonic invasion of 1812." (J. Conway, `//www.forum.swarthmore.edu/news.archives/geometry.software.dynamic/article49.html`, 15 12 1994.)

4.2.2 Voronoi–Delaunay tessellations and meshes

In the plane or space (d = 2 or 3 is the dimension), consider a finite set $\mathcal{N}$ of points, and let domain D, to be meshed, be the interior of their convex[7] hull. The *Voronoi cell* V_n of n is the closed convex set

$$V_n = \{x \in \overline{D} : \ |x - x_n| \le |x - x_m| \ \ \forall m \in \mathcal{N} - \{n\}\}$$

made of points not farther to node n than to any other node. Fig. 4.8 gives an example, with 11 nodes and as many Voronoi cells (the polygons with irregular shapes).

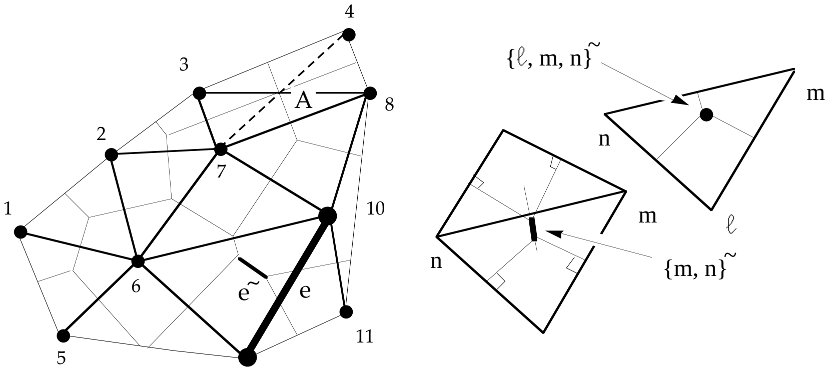


FIGURE 4.8. Voronoi-Delaunay mesh in dimension 2. Note the ambiguity about point A and how it is resolved by arbitrarily preferring {3, 8} to {2, 4}. A dual q-cell and the associated (d – q)-simplex are supported by orthogonal and complementary affine subspaces, but do not necessarily encounter each other (like here the edge e and its dual $e^{\sim}$).

These are "d-cells", if one refers to their dimension. One can also define "q-cells" by taking all non-empty intersections of d-cells, two by two (q = d – 1), three by three (q = d – 2), etc. Points of these q-cells are closer to d – q + 1 of the original nodes. For instance, the (d – 1)-cell associated with nodes n and m is

$$V_{n,m} = \{x \in \overline{D} : \ |x - x_n| = |x - x_m| \le |x_n - x_k| \ \ \forall k \in \mathcal{N} - \{n, m\}\},$$

and x_n and x_m are its nearest neighbors among nodes. Generically, p Voronoi cells intersect, if they do, as a convex set of dimension d + 1 – p. There are exceptions: cf. nodes 3 and 8 in Fig. 4.8, where a small segment

[7] Convexity is important, and special precautions must be taken when generating VD meshes for nonconvex regions (cf. [Ge], [We]).

near point A would be made of points closer to 3 and 8, if only 4 and 7 were a little farther away. For simplicity in this description, we shall assume that such degenerate cases are absent (although they can be a nuisance in practice). Voronoi d-cells and their intersections form a cellular tessellation of the domain.

Now let us associate to each of these q-cells the p nodes that define it, p = d + 1 – q, that is, nodes which are nearest neighbors to all points of the q-cell. They form a p-simplex, which we shall call a *Dirichlet simplex*. The *Voronoi–Delaunay* (VD) mesh is the simplicial mesh thus obtained. In spite of its being derived from the Voronoi paving, we shall consider the simplicial VD mesh as *primal*, and the system of Voronoi cells as its *dual* cellular mesh, and denote by $\tilde{s}$ the Voronoi cell that corresponds to the primal simplex s. As an example, Fig. 4.8 displays the Voronoi cell $\tilde{e}$ of edge e.

Analogies between this dual and the barycentric one are obvious. From the combinatorial point of view, they are even the same: the dual cells s^* and $\tilde{s}$ are defined by the same set of primal nodes. But the shapes of the cells differ widely. Contrary to m^*-cells, Voronoi cells are all convex and lie in a definite affine subspace (of dimension q for q-cells). Compare Figs. 4.4 and 4.9. On the other hand, barycentric duals always intersect their primal associates, whereas Voronoi cells may lie some distance away from their mates (case of e and $\tilde{e}$, Fig. 4.8).

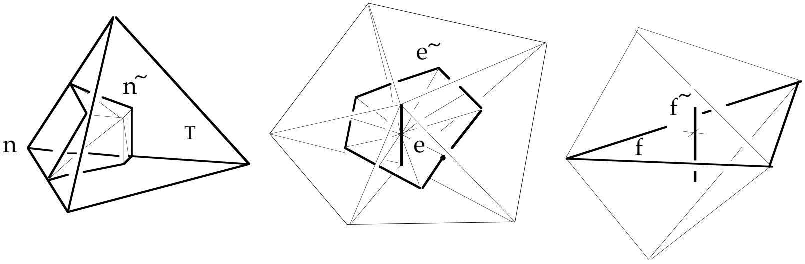
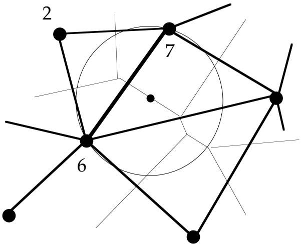


FIGURE 4.9. Cells of the Voronoi dual mesh $\tilde{m}$. Compare with Fig. 4.4.

Such VD meshes have remarkable properties. For instance, this, which is an almost immediate consequence of the construction principle:

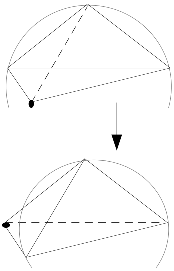
Proposition 4.4. *For each Dirichlet simplex, there is a sphere that contains it and its lower dimensional faces, but no other simplex.*

Proof. Take a sphere centered at one of the points of the dual Voronoi cell, of radius equal to the distance to one of the nodes of the simplex



(inset). Other nodes of the simplex, being equidistant, are on the sphere, and all remaining nodes, being farther away, are outside. ◊

This "sphere property", or circle property in 2D, happens to be characteristic (**Exercise 4.11:** Give an argument to this effect), and is a key-element in incremental VD mesh construction: To add a new vertex to what is already a VD mesh, gather the d-simplices the circumscribed sphere of which contains this vertex, hence a polytope, which is subdivided by joining its vertices to the new vertex. The new mesh is still a VD one [SI, Wa, We].



In dimension 2, any triangulation can be transformed into a VD one by successively swapping diagonals of quadrilaterals formed by adjacent triangles. Why this works locally is clear: Angles of quadrilaterals add to 2π, so one of the two diagonals has opposite angles which sum up to less than π, and a swap will enforce the circle property, as shown in the inset. The difficulty is to prove the *finiteness* of the sequence of swaps [Ch]. (It's due to the swaps decreasing the total area of circumscribed circles [Ni].) Let's finally mention the "maxmin angle property": In dimension 2 again, the VD mesh is the[8] one, among all triangulations with the same node set, that maximizes the smallest angle [RS, SI, Si]. It also minimizes the energy of the finite element solution [RS].

According to [Hr], what we call nowadays a Voronoi cell was introduced in two dimensions by Dirichlet [Di] and in n dimensions by Voronoi [Vo]. (M. Senechal [Se] also gives priority to Dirichlet.) Then came Delaunay [De]. Other names are in use: "Thiessen polygons" among meteorologists[9] [CR], "Dirichlet domains", "Brillouin zones", "Wigner–Seitz cells", etc., among crystallographers.

4.2.3 VD meshes and Stieltjes matrices

Now let's come back to the question of nonpositive off-diagonal coefficients. There is an apparently favorable situation in dimension 2:

[8] When there is a *unique* one, of course, which is the generic case, but there are obvious exceptions (Fig. 4.8 shows one).

Proposition 4.5. *Assume* μ *constant in* D. *For all inner nodes* n, $\mathbf{M}_{mn} \leq 0$ *for all* $m \neq n$.

Proof. Edge mn is flanked by two triangles T and T' (Fig. 4.10), and the opposite angles add up to less than 180°, thanks to the circle property. By the cotangent formula (3.23), $-\mathbf{M}_{mn}$ is proportional to $\cot\theta + \cot\theta'$, which is ≥ 0 if $\theta + \theta' \leq \pi$. ◊

Alas, this leaves many loose strands. First, obtuse angles at boundary triangles (there is one in Fig. 4.8, triangle {8, 10, 11}). But if the objective is to enforce the discrete maximum principle, only inner nodes are involved, and anyway, one may add nodes at the boundary and, if necessary, remesh (which is a local, inexpensive process). Second, and more serious, the condition of uniformity of μ is overly restrictive, and although the cure is of the same kind (add nodes at discontinuity interfaces to force acute angles), further research is needed in this direction.

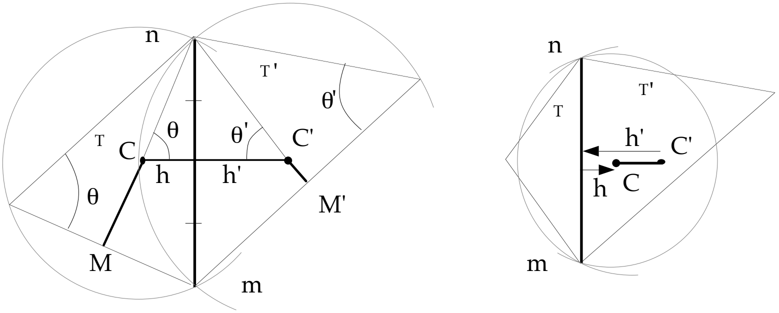


FIGURE 4.10. Proof of (5). M and M' are the mid-edges.

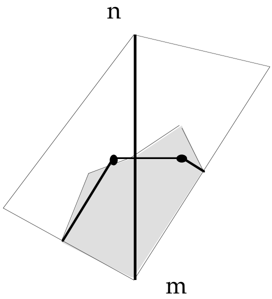
Proposition 4.5 can be proven in a different and instructive way (Fig. 4.10). By the cotangent formula, and the obvious angular relation of Fig. 4.10 (where C and C' are the circumcenters), one has

$$-\mathbf{M}_{mn} = \tfrac{1}{2}\,(\mu(T)\cot\theta + \mu(T')\cot\theta') = (h\,\mu(T) + h'\,\mu(T'))/\,|mn|, \tag{5}$$

[9] Conventionally, the conditions reported by a meteorological station (temperature, hygrometry, etc.) are supposed to hold in the whole "Thiessen polygon" around that station. As explained in [CH], "Stations are always being added, deleted, moved, or temporarily dropped from the network when they fail to report for short periods of time (missing data)", hence the necessity to frequently solve the typical problem: having a Voronoi–Delaunay mesh, add or delete a node, and recalculate the boundaries. Recursive application of this procedure is the standard Watson–Bowyer algorithm for VD mesh generation [Wa], much improved recently [SI] by making it resistant to roundoff errors. Fine displays of VD meshes can be found in [We].

where $|mn|$ is the length of edge $\{m, n\}$, and h, h' are to be counted algebraically, in the direction of the outward normal (thus, $h \leq 0$ if the circumcenter C is outside T, as on Fig. 4.10, right part). This way, in the case where μ is uniform, $\mathbf{M}_{mn} = -\mu\,(h + h')/|mn|$, negative indeed if the circle condition is satisfied.

This quantity happens to be the flux of μ grad λ^n out of the Voronoi cell of node m (**Exercise 4.12:** Prove this, under some precise assumption). This coincidence is explained in inset: Although the Voronoi cell and the barycentric box don't coincide, the flux through their boundaries is the same, because $\mu\,\nabla\lambda^n$ is divergence-free in the region in between. But beware: Not only does this argument break down when there is an obtuse angle (cf. Exer. 4.12), but it doesn't extend to dimension 3, where circumcenter and gravity center of a face do not coincide.



Still, there is some seduction in a formula such as (5), and it *has* a three-dimensional analogue. Look again at Fig. 4.9, middle. The formulas

$$\tilde{\mathbf{M}}_{mn} = -[\textstyle\sum_F\ (\text{area}(F)\ \mu(F)]/|mn|, \quad \tilde{\mathbf{M}}_{nn} = -\textstyle\sum_{m \in \mathcal{N}} \tilde{\mathbf{M}}_{mn},$$

where F is an ad-hoc index for the small triangles of the dual cell $\widetilde{\{m, n\}}$, do provide negative exchange coefficients between n and m, and hence a matrix with Stieltjes principal submatrices. This is a quite interesting discretization method, but not the finite element one, and $\tilde{\mathbf{M}} \neq \mathbf{M}$.

Exercise 4.13. Interpret this "finite volume" method in terms of fluxes through Voronoi cells.

4.3 CONVERGENCE AND ERROR ANALYSIS

We now consider a family $\mathcal{M}$ of tetrahedral meshes of a bounded spatial domain D. Does φ_m converge toward φ, in the sense that $||\varphi_m - \varphi||_\mu$ tends to zero, when m . . . when m does *what*, exactly? The difficulty is mathematical, not semantic: We need some structure[10] on the set $\mathcal{M}$ to validly talk about convergence and limit.

[10] The right concept is that of *filter* [Ca]. But it would be pure folly to smuggle that into an elementary course.

The first idea that comes to mind in this respect is to gauge the "coarseness" of m, as follows. Let us denote by $\gamma_n(m)$, or simply γ_n, the maximum distance between x_n and a point of its cluster D_n. Call *grain* of the mesh, denoted $\gamma(m)$ or simply γ, the least upper bound of the γ_ns, which is also the maximum distance between two points which belong to the same tetrahedron T, or maximum *diameter* of the T's.

Now, the statement to prove would seem to be, in the time-honored ε–δ tradition of calculus, "Given $\varepsilon > 0$, there exists $\delta > 0$ such that, if $\gamma(m) \leq \delta$, then $||\varphi_m - \varphi||_\mu \leq \varepsilon$." Unfortunately, this is plainly *false*. There are straight counter-examples of meshes of arbitrary small grain for which the energy of the computed field stays above the energy minimum by a finite amount: Obtuse angles, larger and larger, do the trick [BA].

What we may expect, however, and which turns out to be true, is the validity of the above statement if the family of meshes is *restricted* by some qualifying conditions. "Acuteness", for instance, defined as the absence of obtuse dihedral angles between any two adjacent faces, happens to work: The statement "Given $\varepsilon > 0$, there exists $\delta > 0$ such that, *if m is acute*, and if $\gamma(m) \leq \delta$, then $||\varphi_m - \varphi||_\mu \leq \varepsilon$" is true (we'll prove it).

Such *convergence* results are essential, because it would make no sense to use the Galerkin method in their absence. But in practice, they are not enough: We should like to know which kind of mesh to build to obtain a prescribed accuracy. Knowing how the above δ depends on ε would be ideal: Given ε, make a mesh the grain of which is lower than $\delta(\varepsilon)$. No such general results are known, however, and we shall have to be content with *asymptotic estimates* of the following kind:

$$(6) \qquad ||\varphi_m - \varphi||_\mu \leq C\,\gamma(m)^\alpha,$$

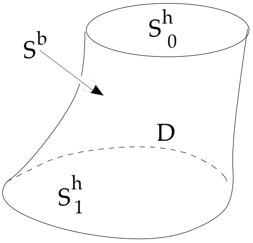
where α is a known positive exponent and C a constant[11] which depends on the true solution φ, but not on the mesh. Again, C cannot be known in advance in general, but (6) tells how *fast* the error will decrease when the grain tends to 0, and this is quite useful. One usually concentrates on the exponent α, which depends on the shape functions. Typically, $\alpha = 1$ for the P^1 elements.

We shall first present the general method by which estimates like (6) can be obtained, then address the question of which restrictions to force on $\mathcal{M}$ in order to make them valid.

[11] From now on, all C's in error estimates will be constants of this kind, not necessarily the same at different places, which may depend on φ (via its derivatives of order 2 and higher, as we shall see), but not on the mesh.

4.3.1 Interpolation error and approximation error

Let's develop an idea that was only suggested in Section 3.3.



First, a definition: Given a family $\mathcal{M}$ of meshes, an *interpolation procedure* is a similarly indexed family of linear mappings $r_m : \mathcal{U} \rightarrow \Phi_m$, where $\mathcal{U}$ is *dense* in Φ^*. Let's give an example immediately: $\mathcal{U}$ is made of all continuous functions over D that vanish on S^h_0, and its m-interpolate is

$$r_m u = \sum_{n \in \mathcal{N}} u(x_n)\, \lambda^n.$$

In other words, u is sampled at nodes, and linearly interpolated in between. This explains why $\mathcal{U}$ cannot coincide with Φ^* (the complete space), which contains non-continuous functions, for which nodal values may not make sense. In fact, for technical reasons that soon will be obvious, we further restrict $\mathcal{U}$ to piecewise 2-smooth functions that vanish on S^h_0. The energy distance $||u - r_m u||_\mu = [\int_D \mu\ |\text{grad}(u - r_m u)|^2]^{1/2}$ between a function and its interpolate is called the *interpolation error*.

Next, let's suppose we know something of the same form as (6) about the interpolation error,

$$(7) \qquad || u - r_m u ||_\mu \leq C(u)\, \gamma\, (m)^\alpha.$$

Then, two things may happen. If the true solution φ is a member of the class $\mathcal{U}$, the remark of Section 3.3 (cf. Fig. 3.4) about the approximation error $||\varphi_m - \varphi||_\mu$ being lower[12] than the interpolation error $||r_m \varphi - \varphi||_\mu$ immediately yields (6), with $C = C(\varphi)$. So we may conclude that for meshes of the family for which (7) holds, $r_m \varphi$ converges in energy toward φ when the grain tends to 0. Moreover, the speed of the convergence is what the interpolation procedure provides.

This situation does not present itself all the time, however, because the solution may not be smooth enough to belong to $\mathcal{U}$ (cf. the example of Exer. 3.6). But the density of $\mathcal{U}$ still allows us to conclude: Given ε, there exists some $u \in \mathcal{U}$ such that $||u - \varphi||_\mu \leq \varepsilon / 2$, and since

$$||\varphi_m - \varphi||_\mu \leq ||r_m u - \varphi||_\mu \leq ||r_m u - u||_\mu + ||u - \varphi||_\mu,$$

[12] Note that $r_m u \in \Phi^I$ if $u \in \Phi^I$ with the present interpolation procedure, if one turns a blind eye to possible variational crimes at the boundary. This is important in asserting that $||\varphi_m - \varphi\, ||_\mu \leq ||r_m \varphi - \varphi\, ||_\mu$.

$||r_m u - u||_\mu$ will be smaller that the still unspent half-epsilon for $\gamma(m)$ small enough, hence the convergence. The convergence *speed*, however, is no longer under control. This is not a practical difficulty, because singularities of φ are usually located at predictable places (corners, spikes), and special precautions about the mesh (pre-emptive refinement, or special shape functions) can be taken there.

4.3.2 Taming the interpolation error: Zlamal's condition

We may therefore concentrate on the interpolation error. By the very definition of hat functions, one has

$$(8) \qquad \textstyle\sum_{n \in \mathcal{N}} \lambda^n(x)\ (x_n - x) = 0 \quad \forall\, x \in D,$$

which makes sense as a weighted sum of *vectors* $x_n - x$.

Let u be an element of $\mathcal{U}$. We'll make use of its second-order Taylor expansion about x, in integral form, as follows:

$$(9) \qquad u(y) = u(x) + \nabla u(x) \cdot (y - x) + \tfrac{1}{2}\, A_u(x, y)(y - x) \cdot (y - x)$$

where, denoting $\partial^2 u$ the matrix of second derivatives of u,

$$A_u(x, y) = \textstyle\int_0^1 (1 - t^2)\ \partial^2 u(x + t(y - x))\, dt,$$

a symmetric matrix that smoothly depends on x and y. Note that $\nabla u(x)$ is treated as a vector in (9), and that A_u acts on vector $y - x$.

Remark 4.3. The validity of formula (9) is restricted to pairs of points $\{x, y\}$ which are linked by a segment entirely contained in D. Not to be bothered by this, we assume u has a smooth extension to the convex hull of D. Anyway, only values of $A_u(x, y)$ for points x and y close to each other will matter. ◊

It is intuitive that the distance between u and its interpolate $r_m u$ should depend on the grain in some way. Our purpose is to show that if the mesh is "well behaved", in a precise sense to be discovered, the quadratic semi-norm, which differs only in an inessential way from the energy one,

$$||u - r_m u|| = [\textstyle\int_D |\nabla(u - r_m u)|^2]^{1/2},$$

is bounded by $\gamma(m)$, up to a multiplicative constant that depends on u (via its derivatives of order 2).

By Taylor's formula (9), we have, for all node locations x_n,

$$u(x_n) = u(x) + \nabla u(x) \cdot (x_n - x) + \tfrac{1}{2}\, A_u(x, x_n)(x_n - x) \cdot (x_n - x).$$

Multiplying this by λ^n, then using (8) and (9), we see that

$$r_m u(x) = u(x) + \sum_{n \in \mathcal{N}} \lambda^n(x)\ \alpha_n(x),$$

where

$$\alpha_n(x) = \tfrac{1}{2}\, A_u(x, x_n)(x_n - x) \cdot (x_n - x). \tag{10}$$

Therefore,

$$\nabla(r_m u - u) = \sum_{n \in \mathcal{N}} \lambda^n \nabla\alpha_n + \sum_{n \in \mathcal{N}} \alpha_n \nabla\lambda^n. \tag{11}$$

On D_n, after (10), $|\nabla\alpha_n|$ is bounded by γ_n, up to a multiplicative constant. Fields $\lambda^n \nabla\alpha_n$ are thus uniformly bounded by $C\gamma$ on D, and the first term on the right in (11) is on the order of $\gamma(m)$. The one term we may worry about is therefore $\sum_{n \in \mathcal{N}} \alpha_n \nabla\lambda^n$.

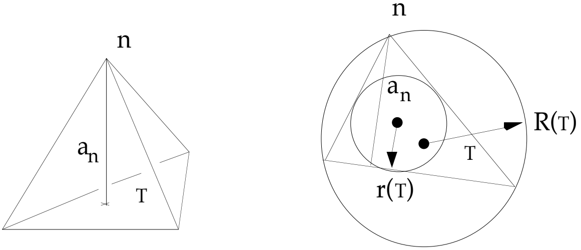


FIGURE 4.11. The norm $|\nabla\lambda^n|$ is $1/a_n$, where a_n is the length of the altitude drawn from node n to the opposite face. Right (in 2D for clarity, but this generalizes without problem), the ratio a_n/γ is always larger than $r(T)/R(T)$, hence "Zlamal's condition" [Zl]: $R(T)/r(T) \le C$, for all triangles and all meshes in the family. It amounts to the same as requiring that the smallest angle (or in 3D, the smallest dihedral angle) be bounded from below.

And worry we should, for $\nabla\lambda^n$ can become very large: Its amplitude within tetrahedron T is the inverse of the distance a_n from node n to the opposite face in T (Fig. 4.11), so there is no necessary link between $|\nabla\lambda^n|$ and γ_n. But it's not difficult to establish such a link if the mesh behaves. Remarking that, for x in T (refer to Fig. 4.11, right, for the notation),

$$|x_n - x|\, |\nabla\lambda^n(x)| \le \gamma_n / a_n \le R(T)/r(T),$$

we are led to introduce the dimensionless number

$$A(m) = \sup_{T \in \mathcal{T}(m)} R(T)/r(T)$$

(maximum ratio of radii of the circumscribed and inscribed spheres), which measures the global "angle acuteness" of the mesh, as explained under Fig. 4.11. Then

$$|\sum_{n \in \mathcal{N}} \alpha_n(x)\ \nabla\lambda^n(x)| \leq C \sum_{n \in \mathcal{N}} \gamma_n |x_n - x|\ |\nabla\lambda^n(x)|$$

$$\leq CA(m)\,\gamma(m).$$

So then, the second term on the right in (11) also is in $\gamma(m)$, and we have, whatever x,

$$(12) \qquad |\nabla(r_m u - u)(x)| \leq C\, A(m)\, \gamma(m).$$

Integrating (12) over the bounded domain D, we find that

$$(13) \qquad \sup_m A(m) < \infty \Rightarrow ||u - r_m u|| \leq C(D, u)\, \gamma(m).$$

Hence our first result: For a family of meshes with bounded acuteness, there is convergence when the grain tends to 0, and the exponent α in (7) is equal to 1.

The existence of such an upper bound for acuteness is equivalent to Zlamal's famous "angle condition" (Fig. 4.11): All angles, for all meshes in the family, should be bounded from below by a fixed, positive amount.

4.3.3 Taming the interpolation error: Flatness

But the rough way by which we obtained estimates suggests that Zlamal's condition may be too strong. Actually, very early in the practice of finite elements, it was clear from experience that acute angles were not necessarily "bad". Besides, as we saw with Exercises 3.14 and 3.15, there is a way to recover finite difference schemes from finite element approximations by using a regular orthogonal mesh and by halving rectangles, in dimension 2, or dissecting hexahedra into tetrahedra, in dimension 3. The approximation theory for such finite difference schemes was well-established at the beginning of the finite element era, and it showed no trace of an "acute angle condition", which strongly suggested that Zlamal's condition was too strong indeed.

It took some time, however (in spite of an early remark by Synge [Sy]), before the condition was properly relaxed, and replaced, in dimension 2, by a *maximum* angle condition, with counter-examples showing that obtuse angles were effectively detrimental [BA], and a more accurate condition (less intuitive than Zlamal's criterion, unfortunately) was formulated [Ja]. The current state of the art can be summarized informally like this: "If the error is too large, or if convergence rate seems poor, don't blame it on acute angles. Watch out for *obtuse* angles instead."

To make this more precise, let us try and improve on the above estimate, as follows. We have, using $\lambda^n = 1 - \sum_{m\neq n}\lambda^m$ to pass from line 2 to line 3 of the following string of equalities, and imbedding $\mathcal{E}$ in $\mathcal{N}\times\mathcal{N}$ the obvious way,

$$\begin{aligned}
\|\textstyle\sum_{n\in\mathcal{N}}\alpha_n\nabla\lambda^n\|^2 &= \textstyle\int_D |\sum_{n\in\mathcal{N}}\alpha_n\nabla\lambda^n|^2 \\
&= \textstyle\sum_{n\in\mathcal{N},\, m\in\mathcal{N}}\int_D \alpha_n\alpha_m\,\nabla\lambda^n\cdot\nabla\lambda^m \\
&= \textstyle\sum_{n\neq m}\int_D \alpha_n\alpha_m\;\nabla\lambda^n\cdot\nabla\lambda^m - \sum_{n\neq m}\int_D \alpha_n{}^2\,\nabla\lambda^n\cdot\nabla\lambda^m \\
&= \textstyle\sum_{n\neq m}\int_D \alpha_n(\alpha_m-\alpha_n)\,\nabla\lambda^n\cdot\nabla\lambda^m \\
&= -\textstyle\sum_{\{m,\, n\}\in\mathcal{E}}\int_D (\alpha_m-\alpha_n)^2\,\nabla\lambda^n\cdot\nabla\lambda^m \\
&= -\textstyle\sum_{T\in\mathcal{T}}\int_T\sum_{\{m,\, n\}\in\mathcal{E}(T)}(\alpha_m-\alpha_n)^2\,\nabla\lambda^n\cdot\nabla\lambda^m.
\end{aligned}$$

As $\alpha_m - \alpha_n = \frac{1}{2}\,A_u(x_m - x)\cdot(x_m - x) - \frac{1}{2}\,A_u(x_n - x)\cdot(x_n - x)$ (up to terms of higher order), one has, with the same degree of approximation, $\alpha_m - \alpha_n \sim \frac{1}{2}\,A_u(x_m - x_n)\cdot(x_n + x_m - 2x)$, hence the estimate

$$|\alpha_m - \alpha_n| \leq C\gamma_n\,|x_m - x_n|.$$

Let us therefore introduce the dimensionless quantity

$$\text{(14)}\qquad F(m) = \sup_{T\in\mathcal{T}}[\textstyle\int_T\sum_{\{m,\, n\}\in\mathcal{E}(T)}|x_n - x_m|^2\,|\nabla\lambda^n\cdot\nabla\lambda^m|\,/\mathrm{vol}(T)]^{1/2}$$

and call it "flatness" of the mesh. Then,

$$\|\textstyle\sum_{n\in\mathcal{N}}\alpha_n\nabla\lambda^n\|^2 \leq C\,\gamma(m)^2\,F(m)^2\sum_{T\in\mathcal{T}}\mathrm{vol}(T),$$

so we may conclude:

$$\text{(15)}\qquad \sup_m F(m) < \infty \Rightarrow \|u - r_m u\| \leq C(D,\, u)\,\gamma(m).$$

By limiting flatness, thus, one makes sure the interpolation error will tend to zero with the mesh grain, just as one did by limiting acuteness, via Zlamal's criterion.

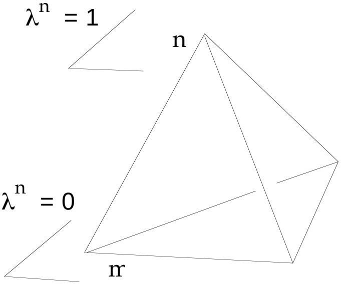
But flatness, in this respect, is a better criterion than acuteness, for controlling the latter amounted to bounding $|\nabla\lambda^n|\;|\nabla\lambda^m|$ instead of $|\nabla\lambda^n\cdot\nabla\lambda^m|$. When evaluating flatness, near orthogonality of $\nabla\lambda^n$ and

$\nabla\lambda^{\mathrm{m}}$ is acknowledged as a favorable factor (cf. the cotangent formula), which the acuteness criterion ignores.

Still, what we have in (14) is only an algebraic figure of merit, not yet formally linked with the shape of the tetrahedra. We now prove that a ban on obtuse angles does limit flatness (and hence, since (15) applies, is a sufficient, though not necessary condition for convergence).

Lemma 4.1. *One has* $\mathrm{mk}\cdot\nabla\lambda^{\mathrm{n}} = 0$ *unless* $\mathrm{n} = \mathrm{k}$ *or* m, *and* $\mathrm{mn}\cdot\nabla\lambda^{\mathrm{n}} = 1$.

Proof. The circulation of $\nabla\lambda^{\mathrm{n}}$ along edge {m, k} is $\mathrm{mk}\cdot\nabla\lambda^{\mathrm{n}}$. On the other hand, the circulation of $\nabla\lambda^{\mathrm{n}}$ along any line is the difference of values of λ^{n} at its ends (inset), and λ^{n} vanishes at all nodes except n, where it takes the value 1. ◊



Proposition 4.6. *In spatial dimension* d, *one has*

$$-\sum_{\{\mathrm{m},\,\mathrm{n}\}\in\mathcal{E}(\mathrm{T})} |\mathrm{x}_{\mathrm{n}}-\mathrm{x}_{\mathrm{m}}|^2\, \nabla\lambda^{\mathrm{n}}\cdot\nabla\lambda^{\mathrm{m}} = \mathrm{d}. \tag{17}$$

Proof. The proof is easy if d = 2, and we sketch it for d = 3. Start from the following identity (Jacobi's),

$$\mathrm{kn}\times(\mathrm{k}\ell\times\mathrm{km}) + \mathrm{k}\ell\times(\mathrm{km}\times\mathrm{kn}) + \mathrm{km}\times(\mathrm{kn}\times\mathrm{k}\ell) = 0,$$

which entails $\mathrm{kn}\times\nabla\lambda^{\mathrm{n}} + \mathrm{k}\ell\times\nabla\lambda^{\ell} + \mathrm{km}\times\nabla\lambda^{\mathrm{m}} = 0$. Square this, using the identity $(\mathrm{a}\times\mathrm{b})\cdot(\mathrm{c}\times\mathrm{d}) = \mathrm{a}\,.\,\mathrm{c}\ \mathrm{b}\,.\,\mathrm{d} - \mathrm{a}\,.\,\mathrm{d}\ \mathrm{b}\,.\,\mathrm{c}$, and Lemma 4.1. Square terms give things like $|\mathrm{kn}|^2\,|\nabla\lambda^{\mathrm{n}}|^2 - 1$ and rectangle terms contribute factors like

$$2\,(\mathrm{kn}\times\nabla\lambda^{\mathrm{n}})\cdot(\mathrm{km}\times\nabla\lambda^{\mathrm{m}}) = 2\,\mathrm{kn}\cdot\mathrm{km}\ \nabla\lambda^{\mathrm{n}}\cdot\nabla\lambda^{\mathrm{m}}$$

$$= (\mathrm{mn}^2 - \mathrm{km}^2 - \mathrm{kn}^2)\,\nabla\lambda^{\mathrm{n}}\cdot\nabla\lambda^{\mathrm{m}}.$$

Adding all that yields (17). ◊

Corollary. *If all dihedral angles are acute, then*

$$\|\mathrm{u} - \mathrm{r}_m\mathrm{u}\| \le \mathrm{C}(\mathrm{D},\mathrm{u})\,\gamma\,(m).$$

Proof. Then all coefficients $\nabla\lambda^{\mathrm{n}}\cdot\nabla\lambda^{\mathrm{m}}$ are negative, and therefore

$$\|\sum_{\mathrm{n}\in\mathcal{N}}\alpha_{\mathrm{n}}\nabla\lambda^{\mathrm{n}}\|^2 \le -\mathrm{C}\sum_{\mathrm{T}\in\mathcal{T}}\sum_{\{\mathrm{m},\,\mathrm{n}\}\in\mathcal{E}(\mathrm{T})}\int_{\mathrm{T}}(\alpha_{\mathrm{m}}-\alpha_{\mathrm{n}})^2\,\nabla\lambda^{\mathrm{n}}\cdot\nabla\lambda^{\mathrm{m}}$$

$$\le -\mathrm{C}\,\gamma\,(m)^2\sum_{\mathrm{T}\in\mathcal{T}}\mathrm{vol}(\mathrm{T})\sum_{\{\mathrm{m},\,\mathrm{n}\}\in\mathcal{E}(\mathrm{T})}|\mathrm{x}_{\mathrm{n}}-\mathrm{x}_{\mathrm{m}}|^2\,\nabla\lambda^{\mathrm{n}}\cdot\nabla\lambda^{\mathrm{m}}$$

$$\leq \mathrm{dC}\,\gamma\,(m)^2 \textstyle\sum_{\mathrm{T}\in\mathcal{T}} \mathrm{vol}(\mathrm{T}) \equiv \mathrm{dC}\,\gamma\,(m)^2\ \mathrm{vol}(\mathrm{D}).\quad \Diamond$$

Thus, acute meshes have all virtues: well-behaved interpolation error, and enforcement of the discrete maximum principle. A pity they are so difficult to produce!

Anyway, there is something disappointing in all these error estimates and convergence criteria: Nowhere did we find any indication on how to compute *upper bounds* on the approximation error, either a priori (but let's not dream) or a posteriori. So the simple and so important question, "how far apart are φ and φ_m ?" is still unanswered. We'll find a way in this direction in Chapter 6. But before that, we need improved mathematical equipment.

EXERCISES

Exercises 4.1, 4.2, and 4.3 are on pp. 96, 99, and 100, respectively.

Exercise 4.4. Suppose the problem consists in finding φ in Φ^0 such that $\int_{\mathrm{D}} \mu \text{ grad } \varphi \cdot \text{grad } \varphi' = \int_{\mathrm{D}} \mathrm{f}\, \varphi' + \int_{\mathrm{S}} \mathrm{g}\, \varphi' \quad \forall\ \varphi' \in \Phi^0$. Use the "flux accounting" idea to find the discrete equations directly.

Exercises 4.5 to 4.9 are on pp. 103 to 105, Exer. 4.10 is on p. 106, and Exers. 4.11 to 4.13 are on pp. 109 to 111.

HINTS

4.1. Cf. Exer. 3.7 for the cube. But you will see the result does not depend on which way the cube is partitioned.

4.2. Use (2), with obvious changes to cater for dimension 2: 1/2 instead of 1/3, etc., and sum over the nodes of the subset list. The most effective way to proceed may be to label all edge sides, as in Fig. 4.12 below, and to work out the algebraic relations implied by (3) between outward fluxes. Remember that fluxes across the whole boundary of an element must vanish.

4.3. Edges must be neighbors, but this can happen in two ways: by sharing a node (then both belong to some face), or not (then both belong to some tetrahedron). Draw the parts of the 2-cells inside a single tetrahedron in order to visualize the intersection.

4.4. Data f and g are flux loss densities, so their integrals over boxes, or in the case of g, the part of the box's boundary that lies in S, balance inter-box exchanges.

4.5. Apply Prop. 4.2 to $b - \mu_0 m$.

4.6. Σ_1 does not bound modulo S^b, but the union of *two* such surfaces does, hence the equality of fluxes through each of them.

4.7. Since $(\mathbf{M}\boldsymbol{\varphi})_n$ has the dimensions of a flux, whereas $-\operatorname{div}(\mu \operatorname{grad} \varphi)$ is a flux *density*, introduce the volume of the box B_n.

4.8. Call ${}^h\boldsymbol{\varphi}$ the vector of nodal values on S^h, and use the same block forms as in Section 3.3.3, with $\boldsymbol{\varphi} = \{{}^0\boldsymbol{\varphi}, {}^h\boldsymbol{\varphi}\}$. Observe that $\int_D \mu\,|\operatorname{grad} \varphi|^2 = \int_S \mu\,\partial_n\varphi\,\varphi^h = \int_{S^h} \mu\,P\varphi^h\,\varphi^h$, where P is a certain linear operator. The aim of the exercise is thus to work out the discrete analogue of this operator. How does it relate with the reluctance of the region, "as seen from the boundary"?

4.9. Work on $x \to 1/|x|$, and remember $\operatorname{div}(\varphi\,u) = \varphi \operatorname{div} u + \nabla\varphi \cdot u$.

4.10. Show that ${}^0\boldsymbol{\varphi} \geq 0$ first, then work on the translate $\boldsymbol{\varphi} - I$ to show that ${}^0\boldsymbol{\varphi} \leq I$, by the same method.

4.11. Suppose a non-Delaunay tetrahedron is in the mesh. Show that its circumscribed sphere must contain other nodes.

4.12. Note that the flux density through CC' is $1/|mn|$. Treat separately segments CC' and MC, $M'C'$, of the cell boundary. Beware the obtuse angle.

4.13. Same as Exer. 4.12, if all circumcenters are inside tetrahedra.

SOLUTIONS

4.1. Tetrahedron: $\chi = 4 - 6 + 4 - 1 = 1$. Cube without inner edge: $\chi = 8 - 18 + 16 - 5 = 1$. With one: $\chi = 8 - 19 + 18 - 6 = 1$.

4.2. The following pedestrian solution works convincingly. Call F_k the outgoing flux at edge k, with the labelling of Fig. 4.12. One has $F_2 + F_{11} + F_{12} = 0$, and eight other similar relations. Add them all, which results in $\sum_{1 \leq k \leq 27} F_k = 0$. On the other hand, relation (2) expressed at nodes i and j implies $\sum_{8 \leq k \leq 17} F_k + F_{26} + F_{27} = 0$ and $\sum_{18 \leq k \leq 25} F_k + F_{26} + F_{27} = 0$ respectively. Add these, and subtract the previous one, hence $2\,(F_{26} + F_{27}) = \sum_{1 \leq k \leq 7} F_k$, which is the flux exiting from Σ_1. But $F_{26} + F_{27}$ is one of the "flux leaks", which has no reason to vanish. (If that accidentally happened, a slight perturbation in the nodal positions would restore the generic situation.) The case on the right is similar: The flux exiting from Σ_2 is the sum of flux leaks at the perimeter of T, that is, $\sum_{1 \le k \le 6} F_k$. Subtract $F_1 + F_2 + F_3$, which is 0, and what remains, $F_4 + F_5 + F_6$, again does not vanish, generically.

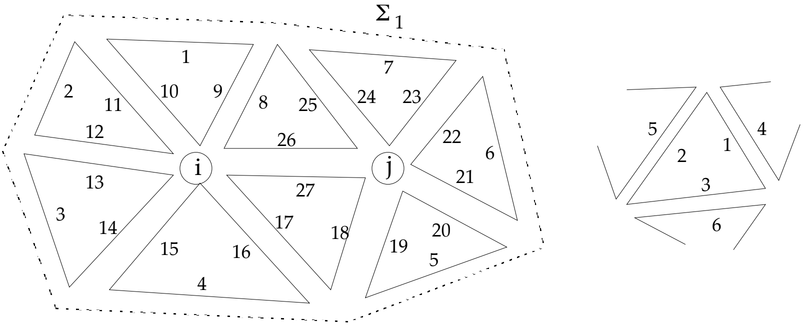


FIGURE 4.12. Ad-hoc edge numberings for Exer. 4.2.

4.3. Figure 4.13.

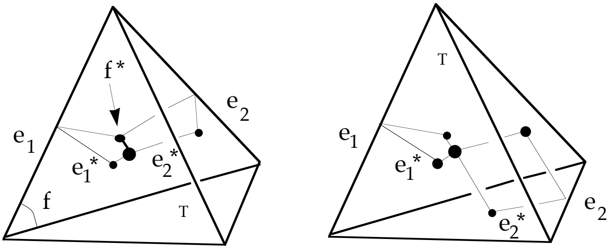


FIGURE 4.13. How dual 2-cells can intersect. Left: Edges e_1 and e_2 have a common node, and define face f. Then e_1^* and e_2^* intersect along f*. Right: they belong to the same tetrahedron, but without any common node. Then $e_1^* \cap e_2^*$ is the barycenter of T.

4.4. Define $\mathbf{f}_n = \int_{B_n} f$ and $\mathbf{g}_n = \int_{S \cap \partial B_n} g$. Then box B_n receives $(\mathbf{M}\boldsymbol{\varphi})_n$ from adjacent boxes and loses $\mathbf{f}_n + \mathbf{g}_n$. The linear system to solve is thus $\mathbf{M}\boldsymbol{\varphi} = \mathbf{f} + \mathbf{g}$. The equations are identical if f and g are approximated by mesh-wise affine functions.

4.7. Let us denote by $\mathbf{V}_n$ the volume of the box B_n and by $\mathbf{V}$ the diagonal matrix of the $\mathbf{V}_n$'s. Since $(\mathbf{M}\boldsymbol{\varphi})_n$ is the "flux loss at n", the term $(\mathbf{M}\boldsymbol{\varphi})_n/\mathbf{V}_n$ can be dubbed the "flux loss *density* about n". Since, on the other hand, $\int_{\Sigma_n} n \,.\, b = \int_{\Sigma_n} \mu \, n \cdot \text{grad}\, \varphi = \int_{B_n} \text{div}(\mu \text{ grad}\, \varphi)$, this flux-loss density is

$-\operatorname{div}(\mu \operatorname{grad} \varphi)$ in the continuous case. The matrix equivalent of $-\operatorname{div}(\mu \operatorname{grad}\ )$, therefore, is not $\mathbf{M}$ but $\mathbf{V}^{-1}\mathbf{M}$.

4.8. Let's do things formally: Φ^h is the functional space $\{\varphi \in \Phi : \varphi = \varphi^h$ on $S^h\}$, and φ satisfies $\int_D \mu \operatorname{grad} \varphi \cdot \operatorname{grad} \varphi' = 0 \ \forall \varphi' \in \Phi^0$. This solution linearly depends on the data φ^h, hence a linear map $P : \varphi^h \to \mu\, \partial_n \varphi$. Now, integrating by parts,[13] $\int_D \mu\, |\operatorname{grad} \varphi|^2 = \int_{S^h} P\varphi^h\, \varphi^h$. The best estimate of this is $(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi})$, as evaluated by taking account of the relation ${}^{00}\mathbf{M}\,{}^{0}\boldsymbol{\varphi} + {}^{01}\mathbf{M}\,{}^{h}\boldsymbol{\varphi} = 0$. Therefore,

$$(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi}) = ({}^{10}\mathbf{M}\,{}^{0}\boldsymbol{\varphi} + {}^{11}\mathbf{M}\,{}^{h}\boldsymbol{\varphi}, {}^{h}\boldsymbol{\varphi}) = ([{}^{11}\mathbf{M} - {}^{10}\mathbf{M}({}^{00}\mathbf{M})^{-1}\,{}^{01}\mathbf{M}]\,{}^{h}\boldsymbol{\varphi}, {}^{h}\boldsymbol{\varphi})$$

$$= (\mathbf{P}\,{}^{h}\boldsymbol{\varphi}, {}^{h}\boldsymbol{\varphi}).$$

The variationally correct approximation of P, accordingly, is found to be $\mathbf{P} = {}^{11}\mathbf{M} - {}^{10}\mathbf{M}({}^{00}\mathbf{M})^{-1}\,{}^{01}\mathbf{M}$. In case of one mmf I, reluctance R is related with the magnetic coenergy by $\int_D \mu\, |\operatorname{grad} \varphi|^2 = I^2/R$. Matrix $\mathbf{P}$ is thus the inverse of a "multipolar inductance", by which the region can been treated as an inductive circuit element, in some higher-level modelling.

4.9. First, $\operatorname{grad}(x \to |x|^2) = x \to 2x$, and therefore, $\operatorname{grad}(x \to |x|^\alpha) = x \to \alpha\, |x|^{\alpha - 2}\, x$. In particular (and of constant usefulness), the gradient of $x \to |x|$ is $x \to x/|x|$ and $\operatorname{grad}(x \to 1/|x|) = x \to -x/|x|^3$. Now, $\operatorname{div}(x \to x) = 3$, by Exer. 1.3, thus

$$\Delta(x \to 1/|x|) = x \to [-3/|x|^3 + 3x \cdot x/|x|^5] \equiv 0 \text{ if } x \neq 0.$$

4.10. All entries of ${}^{01}\mathbf{M}$ are off-diagonal in $\mathbf{M}$, so $-\,{}^{01}\mathbf{M}\,{}^{1}\boldsymbol{\varphi}^{I} \geq 0$, in the notation of Chapter 3. The principal submatrix ${}^{00}\mathbf{M}$ is Stieltjes, hence ${}^{0}\boldsymbol{\varphi} = -\,({}^{00}\mathbf{M})^{-1}\,{}^{01}\mathbf{M}\,{}^{1}\boldsymbol{\varphi}^{I} \geq 0$. Now invert the roles of boundaries S^h_0 and S^h_1, setting $\boldsymbol{\varphi} - I = 0$ on S^h_1 and $-I$ on S^h_0, hence all potentials $\leq I$ by the same reasoning.

4.11. If a tetrahedron is retained in the VD mesh, it's because there is a set of points which are closer to its vertices than to all other nodes, and this set contains the center of the circumscribed sphere. So this center is closer to another node if the tetrahedron was not Delaunay to start with. (One may apply the same reasoning to all simplices, not only those of maximum dimension: Introduce the "mediator set" of a subset of nodes, as the set of points equidistant to all of them, and center circumscribed spheres on points of this set.)

[13] All this implicitly assumes some regularity, but the reader is encouraged to ignore such issues, which will be addressed in earnest in Chapter 7.

4.12. Since $\mu\nabla\lambda^n = 0$ out of T ∪ T', the flux to consider is through the broken line MCC'M'. But since $\nabla\lambda^n$ is parallel to MC and M'C', what remains is the flux through CC', which is $\mu\,(h + h')/\,|mn|$. The whole thing breaks down if one of the circumcenters, C for instance, is outside its triangle, for then $\nabla\lambda^n$ is not parallel to the part of MC lying inside T'.

4.13. If all circumcenters are inside tetrahedra, this is the box-method, relative to Voronoi boxes.

REFERENCES

[BA] I. Babuška, A.K. Aziz: "On the angle condition in the finite element method", **SIAM. J. Numer. Anal., 13**, 2 (1976), pp. 214–226.

[Ca] H. Cartan: "Théorie des filtres", **C.R. Acad. Sc. Paris, 205** (1937), pp. 595–598.

[C&] Z.J. Cendes, D. Shenton, H. Shahnasser: "Magnetic Field Computation Using Delaunay Triangulation and Complementary Finite Element Methods", **IEEE Trans, MAG-19**, 6 (1983), pp. 2551–2554.

[CS] M.V.K. Chari, P. Silvester: "Finite Element Analysis of Magnetically Saturated D–C Machines", **IEEE Trans., PAS-90,** 5 (1971), pp. 2362–2372.

[Ch] C. Cherfils, F. Hermeline: "Diagonal swap procedures and characterization of Delaunay triangulations", **M2AN, 24**, 5 (1990), pp. 613–625.

[CR] A.K. Cline, R.L. Renka: "A storage-efficient method for construction of a Thiessen triangulation", **Rocky Mountain J. Math., 14,** 1 (1984), pp. 119–139.

[CH] T.E. Croley II, H.C. Hartmann: "Resolving Thiessen Polygons", **J. Hydrology, 76** (1985), pp. 363–379.

[De] B. Delaunay: "Sur la sphère vide", **Bull. Acad. Sci. URSS, Class. Sci. Math. Nat**. (1934), pp. 793–800.

[Di] G.L. Dirichlet: "Über die Reduction der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen", **J. Reine Angew. Math., 40** (1850), p. 209.

[Ea] S. Earnshaw (1805–1888): "On the Nature of the Molecular Forces which Regulate the Constitution of the Luminiferous Ether", **Trans. Cambridge Phil. Soc., 7** (1842), pp. 97–114.

[EF] E.A. Erdelyi, E.F. Fuchs: "Fields in Electrical Devices Containing Soft Nonlinear Materials", **IEEE Trans., MAG-10,** 4 (1974), pp. 1103–1108.

[FE] E.F. Fuchs, E.A. Erdelyi: "Nonlinear Theory of Turboalternators—Pt. I", **IEEE PES Summer Meeting**, San Francisco, July 9–14 (1972), pp. 583–599.

[Ge] P.L. George: **Génération automatique de maillages. Applications aux méthodes d'éléments finis,** Masson (Paris), 1990.

[GT] D. Gilbarg, N.S. Trudinger: **Elliptic Partial Differential Equations of Second Order,** Springer-Verlag (Berlin), 1977.

[GH] M.J. Greenberg, J.R. Harper: **Algebraic Topology, A First Course**, Benjamin/ Cummings (Reading, MA), 1981.

[He] B. Heinrich: **Finite Difference Methods on Irregular Networks**, Akademie-Verlag (Berlin), 1987.

[Hr] F. Hermeline: "Triangulation automatique d'un polyèdre en dimension N", **RAIRO Analyse Numérique, 16**, 3 (1982), pp. 211–242.

[HW] P.J. Hilton, S. Wylie: **Homology Theory,** An Introduction to Algebraic Topology, Cambridge U.P. (Cambridge), 1965.

[Ja] P. Jamet: "Estimations d'erreur pour des éléments finis droits presque dégénérés", **RAIRO Anal. Numer., 20** (1976), pp. 43–61.

[Ni] G.M. Nielson: "A criterion for computing affine invariant triangulations", **IMACS 88**, pp. 560–562.

[RS] S. Rippa, B. Schiff: "Minimum energy triangulations for elliptic problems", **Comp. Meth. Appl. Mech. Engng., 84** (1990), pp. 257–274.

[SD] Y. Saito, S. Ikeguchi, S. Hayano: "An Efficient Computation of Saturable Magnetic Field Problem Using Orthogonal Discretization", **IEEE Trans., MAG-24**, 6 (1988), pp. 3138–3140.

[Sc] W.T. Scott: "Who Was Earnshaw?", **Am. J. Phys., 27**, 4 (1959), pp. 418–419.

[Se] M. Senechal: **Crystalline Symmetries,** An informal mathematical introduction, Adam Hilger (Bristol), 1990.

[Si] R. Sibson: "Locally equiangular triangulations", **The Computer Journal, 21,** 3 (1978), pp. 243–245.

[SI] K. Sugihara, M. Iri: "A robust topology-oriented incremental algorithm for Voronoi diagrams", **Int. J. Comp. Geometry & Appl., 4,** 2 (1994), pp. 179–228.

[Sy] J. L. Synge: **The Hypercircle in Mathematical Physics**, Cambridge U.P. (Cambridge), 1957.

[Va] R.S. Varga: **Matrix Iterative Analysis**, Prentice-Hall (Englewood Cliffs, NJ), 1962.

[Vo] G. Voronoi: "Nouvelles applications des paramètres continus à la théorie des formes quadratiques: recherches sur les paralléloèdres primitifs", **J. Reine Angew. Math., 134** (1908), pp. 198–287.

[Wa] D.F. Watson: "Computing the *n*-dimensional Delaunay tessellation with application to Voronoi polytopes", **Comp. J., 24,** 2 (1981), pp. 167–72.

[We] N.P. Weatherhill: "A method for generating irregular computational grids in multiply connected domains", **Int. J. Numer. Meth. Fluids, 8** (1988), pp. 181–97.

[Z&] Zhou Jian-ming, Shao Ke-ran, Zhou Ke-ding, Zhan Qiong-hua: "Computing constrained triangulation and Delaunay triangulation: A new algorithm", **IEEE Trans**., **MAG-26**, 2 (1990), pp. 694–7.

[Zl] M. Zlamal: "On the finite element method", **Numer. Math., 12** (1968), pp. 394–409.

CHAPTER 5

Whitney Elements

We now leave the first part of this book, devoted to the study of the "div-side" of the modelling of Chapter 2, and will turn to the "curl-side". As we need a more encompassing viewpoint to survey this enlarged landscape, we shall use more sophisticated mathematical tools. Hence this transition chapter. First, we enlarge the mathematical framework, studying the three fundamental operators grad, rot, div, from the functional point of view, thus making visible a rich structure, which happens to be the right functional framework for Maxwell's equations. Then, we present a family of geometrical objects introduced around 1957 by Whitney (Hassler Whitney, 1907–1989, one of the masters of differential geometry), known as "Whitney (differential) forms" [W h]. They constitute a *discrete* realization of the previous structure, and therefore, the right framework in which to develop a finite element discretization of electromagnetic theory. (This is why I call them "Whitney elements" here, rather than "Whitney forms".) Finally, now-popular "tree and cotree" techniques are addressed.

5.1 A FUNCTIONAL FRAMEWORK

In Section 3.2, when we had to complete the space of potentials, we saw a connection between the physically natural idea of "generalized solution" of an equation Ax = b, and the prolongation of operator A beyond its initial domain of definition. This will now be systematized, and applied to the classical differential operators grad, rot, and div. The idea is extremely simple: We take the *closures* of the graphs of all operators in sight, thus finding extensions of them with good properties. But the proofs along the way can be quite involved, so they are placed in such positions as to make it easy to ignore them at first reading. Familiarity with the Hilbert spaces $L^2(D)$ and $\mathbb{L}^2(D)$ is now assumed. (Cf. 3.2.3 and Appendix A, Section A.4.)

5.1.1 The "weak" grad, rot, and div

Let D be a regular bounded domain of E_3 and denote, as in Chapter 2, $C^\infty(\overline{D})$ and $\mathbb{C}^\infty(\overline{D})$ the spaces of restrictions to D of smooth functions or fields with compact support in E_3. All three components of a smooth field b have partial derivatives at all points of $\overline{D}$ and its boundary, hence a function $f = \text{div } b$ that belongs to $C^\infty(\overline{D})$, and hence a linear operator, denoted div, the standard, or "strong" one.

We found it not so convenient a tool, back in Chapter 2. For instance, if $\{b_n\}$ is a sequence of smooth solenoidal vector fields of finite energy which converge in energy toward a field b, we expect b to be solenoidal as well. And yet, we would have "no right" to say that, because "div b = 0" doesn't make sense if b is not smooth! A silly situation, from which we escaped thanks to the notion of weak formulation, but there is a more direct approach, as suggested by the very idea of completion, as follows. Set $f_n = \text{div } b_n$, not necessarily zero for more generality. Suppose that $\lim_{n \to \infty} b_n = b$ and $\lim_{n \to \infty} f_n = f$ in $\mathbb{L}^2(D)$ and $L^2(D)$ respectively. Why not *decree* that div b does exist, as a scalar field, and is equal to f, thus enlarging the domain of div? This is quite in the spirit of generalized solutions. By doing that for all similar sequences, we may expect the extension[1] of div thus obtained to be free of the inadequacies of the strong divergence.

So let us denote by DIV the graph in $\mathbb{L}^2(D) \times L^2(D)$ of the strong divergence, i.e., the set of pairs $\{b, f\} \in \mathbb{C}^\infty(\overline{D}) \times C^\infty(\overline{D})$ such that $\text{div } b = f$. The recipe just described—enlarge the graph in order to include limits of related pairs $\{b_n, f_n\}$ —simply consists in taking its *closure,* denoted $\overline{\text{DIV}}$, in $\mathbb{L}^2(D) \times L^2(D)$. Hence a new operator, that we shall provisionally denote ${}^w\text{div}$ and call the *weak* divergence, for reasons which will be obvious in a moment.

But . . . *does* this subset $\overline{\text{DIV}}$ define a function? Is it a *functional* graph? Conceivably, two sequences $\{b_n, f_n\}$ and $\{\tilde{b}_n, \tilde{f}_n\}$ could converge toward the same b but different f's, hence a multivalued extension of div. We say that a linear operator (or, for that matter, any function) is *closable* if such mishaps cannot occur, i.e., if the closure of its graph is functional.[2] And indeed,

Proposition 5.1. $\overline{\text{DIV}}$ *is a functional graph.*

Proof. Otherwise, there would exist a nonzero f such that $\{0, f\}$ be in the

[1] See Appendix A, Subsection A.1.2, for the basic notions about relations, functional or not, their graphs, their restrictions, their extensions, etc., and A.3.2 for metric-related notions.

[2] Of course, *closed* operators are those with a closed graph. Cf. A.4.4.

closure of DIV. Let then $\{b_n, f_n\} \in$ DIV go to $\{0, f\}$ in the sense of the $\mathbb{L}^2(D) \times L^2(D)$ norm. For all test functions $\varphi' \in C_0^\infty(D)$, one would have

$$(1) \qquad \int_D b_n \cdot \operatorname{grad} \varphi' = -\int_D \operatorname{div} b_n \; \varphi',$$

and since the two sides tend to 0 and $\int_D f\, \varphi'$ respectively, by continuity of the scalar product, this implies $\int_D f\, \varphi' = 0 \;\; \forall\, \varphi' \in C_0^\infty(D)$, hence $f = 0$ by density of $C_0^\infty(D)$ in $L^2(D)$ (cf. A.2.3), which proves the point. ◊

The relation of this procedure with the weak formulation of the equation div b = f is now patent, which stirs us to try and prove the following clincher result:

Proposition 5.2. *The closure of* DIV *coincides with the set* ^{w}DIV *of pairs* $\{b, f\}$ *in* $\mathbb{L}^2(D) \times L^2(D)$ *such that*

$$(2) \qquad \int_D b \cdot \operatorname{grad} \varphi' + \int_D f\; \varphi' = 0 \quad \forall\, \varphi' \in C_0^\infty(D).$$

The proof will qualify ^{w}div as the proper generalization of the "weak divergence" of Eq. (2.11). (The residual and a bit awkward restriction to "piecewise smooth" fields, more or less forced upon us in Chapter 2, has now been lifted for good.) However, it's surprisingly difficult, so let me postpone it for a moment, in order to keep the main ideas in focus.

So—provisionally accepting Prop. 5.2 as valid—what we called earlier "weak solenoidality" corresponds to ^{w}div b = 0, the fact for a field to have a null weak divergence in the present sense, and this justifies the terminology.

Let's generalize. First, give names to the graphs of the strong operators:

GRAD, the graph of grad : $C^\infty(\overline{D}) \rightarrow \mathbb{C}^\infty(\overline{D})$ in $L^2(D) \times \mathbb{L}^2(D)$,

ROT, the graph of rot : $\mathbb{C}^\infty(\overline{D}) \rightarrow \mathbb{C}^\infty(\overline{D})$ in $\mathbb{L}^2(D) \times \mathbb{L}^2(D)$,

DIV, the graph of div : $\mathbb{C}^\infty(\overline{D}) \rightarrow C^\infty(\overline{D})$ in $\mathbb{L}^2(D) \times L^2(D)$,

then define the weak operators ^{w}grad, ^{w}rot, and ^{w}div via their graphs, which are the closures $\overline{\mathrm{GRAD}}$, $\overline{\mathrm{ROT}}$, and $\overline{\mathrm{DIV}}$ of the former ones. (**Exercise 5.1:** Imitate the proof of Prop. 5.1 to show that $\overline{\mathrm{GRAD}}$ and $\overline{\mathrm{ROT}}$ are functional.) Note that functions may have a weak gradient without being differentiable in the classical sense (**Exercise 5.2**: Provide examples), and fields have a weak curl or a weak divergence in spite of their components not being differentiable at places.

We shall denote the domains of these weak operators by $L^2_{\mathrm{grad}}(D)$, $\mathbb{L}^2_{\mathrm{rot}}(D)$, and $\mathbb{L}^2_{\mathrm{div}}(D)$. You may see them defined as follows, in the literature:

$$L^2_{\text{grad}}(D) = \{\varphi \in L^2(D) : \ \text{grad}\ \varphi \in \mathbb{L}^2(D)\},$$

$$\mathbb{L}^2_{\text{rot}}(D) = \{u \in \mathbb{L}^2(D) : \ \text{rot}\ u \in \mathbb{L}^2(D)\},$$

$$\mathbb{L}^2_{\text{div}}(D) = \{u \in \mathbb{L}^2(D) : \ \text{div}\ u \in L^2(D)\}.$$

In such cases, grad, rot, and div are understood in the weak sense; they are actually what we denote here ^{w}grad, ^{w}rot, ^{w}div. Thus stretching the scope of the notation is so convenient that we'll practice it systematically: *From now on, when* grad, rot, div *appear somewhere, it will be understood that their weak extensions* ^{w}grad, ^{w}rot, ^{w}div *are meant.*

5.1.2 New functional spaces: L^2_{grad}, $\mathbb{L}^2_{\text{rot}}$, $\mathbb{L}^2_{\text{div}}$

Up to this point, $L^2_{\text{grad}}(D)$, $\mathbb{L}^2_{\text{rot}}(D)$, and $\mathbb{L}^2_{\text{div}}(D)$ have been mere subspaces of $L^2(D)$, $\mathbb{L}^2(D)$, and $\mathbb{L}^2(D)$. Beware, they are *not* closed, contrary to the graphs! They are dense in L^2 or $\mathbb{L}^2$, actually, since they contain all smooth functions or fields. So they are not complete with respect to the scalar product of L^2 or $\mathbb{L}^2$. We can turn them into Hilbert spaces on their own right by endowing them with new scalar products, as follows:

$$((\varphi, \varphi')) = \int_D \varphi \cdot \varphi' + \int_D \text{grad}\ \varphi \cdot \text{grad}\ \varphi' \quad \text{for } L^2_{\text{grad}}(D),$$

$$((u, u')) = \int_D u \cdot u' + \int_D \text{rot}\ u \cdot \text{rot}\ u' \qquad \text{for } \mathbb{L}^2_{\text{rot}}(D),$$

$$((u, u')) = \int_D u \cdot u' + \int_D \text{div}\ u\ \ \text{div}\ u' \qquad \text{for } \mathbb{L}^2_{\text{div}}(D),$$

where of course grad, rot, and div are the weak ones.

Let us then set, for instance (the two other lines can be treated in parallel fashion)

$$(3) \qquad |||\varphi||| = (\int_D |\,\text{grad}\ \varphi\,|^2 + \int_D |\,\varphi\,|^2)^{1/2}$$

(we reserve the notation $\|\ \|$ for the L^2 norm). This is called the *graph norm,* because (cf. A.1.2) it combines the norms of both elements of the pair $\{\varphi, \text{grad}\ \varphi\}$, which spans $\overline{\text{GRAD}}$. With this norm, $L^2_{\text{grad}}(D)$ is complete, and hence a Hilbert space, since its Cauchy sequences $\{\varphi_n\}$ are in one-to-one correspondence with sequences $\{\varphi_n, \text{grad}\ \varphi_n\}$ belonging to the graph. Moreover, grad is continuous from the new normed space to $\mathbb{L}^2(D)$, since $\|\text{grad}\ \varphi\| \le |||\varphi|||$ by construction.

This can be seen as the real achievement of the whole procedure: *By putting the graph norms on the domains of the weak operators* ^{w}grad,

${}^{w}\mathrm{rot}$, ${}^{w}\mathrm{div}$, *we obtain Hilbert spaces on which these operators are continuous.* From now on, $L^2_{\mathrm{grad}}(D)$, $\mathbb{L}^2_{\mathrm{rot}}(D)$, $\mathbb{L}^2_{\mathrm{div}}(D)$ will thus be understood as these Hilbert spaces, duly furnished with the graph norm.

Remark 5.1. The Hilbert space $\{L^2_{\mathrm{grad}}(D), |||\ |||\}$ is the *Sobolev space* usually denoted $H^1(D)$. We shun this standard notation here for the sake of uniform treatment of grad, rot, and div, which is reason enough to so depart from tradition.[3] ◊

This graph closing is quite similar to the "hole plugging" of Chapter 3, where we invoked completion the first time. However, the link between the foregoing procedure and completion is much stronger than a mere analogy. There is a way in which what we have just done *is* completion of a space, followed by an application of the principle of extension by continuity of A.4.1.

Let's show this by discussing the case of grad. Start from the space $C^\infty(\overline{D})$, and put on it the norm (3). Now, the strong grad is continuous from the new normed space $\{C^\infty(\overline{D}), |||\ |||\}$ into $\mathbb{L}^2(D)$. Let us take the completion of $C^\infty(\overline{D})$ with respect to the $|||\ |||$ norm. Limits of Cauchy sequences of pairs belonging to GRAD span its closure, $\overline{\mathrm{GRAD}}$, so there is no problem this time in identifying this completion with a functional space: The completion is in one-to-one correspondence with $\overline{\mathrm{GRAD}}$, and therefore—since the latter is a *functional* graph, as we know (Exer. 5.1)—with its projection on $L^2(D)$. The completion is thus (identifiable with) a subspace of $L^2(D)$, which is the classical $H^1(D)$ —our $L^2_{\mathrm{grad}}(D)$.

By construction, $C^\infty(\overline{D})$ is dense in $H^1(D)$, so we are in a position to apply this principle of extension by continuity of Appendix A (Theorem A.4): A uniformly continuous mapping f_U from a metric space X to a *complete* metric space Y, the domain of which is not all of X but only a dense subset U, can be extended by continuity to a map from all X to Y. Here, f_U is the strong gradient, U and X are $C^\infty(\overline{D})$ and $H^1(D)$, and Y is $\mathbb{L}^2(D)$. The extension of the strong gradient thus obtained, on the one hand, and the weak gradient, on the other hand, are the same operator, because their graphs coincide, by construction.

Let us now update these integration by parts formulas which were so useful up to now. This will essentially rely on the fact that $L^2_{\mathrm{grad}}(D)$, etc., are *complete* spaces, and justify the work invested in various completions up to this point.

First, over all space, the formulas

[3] Some Sobolev diehards insist that $\mathbb{L}^2_{\mathrm{rot}}(D)$ and $\mathbb{L}^2_{\mathrm{div}}(D)$ be denoted H(rot; D) and H(div; D) respectively, which would be just fine . . . if only they also used H(grad; D) for $H^1(D)$.

$$\int_{E_3} \operatorname{div} u \ \varphi = -\int_{E_3} u \cdot \operatorname{grad} \varphi \quad \forall\, \varphi \in L^2_{\text{grad}}(E_3),\ u \in \mathbb{L}^2_{\text{div}}(E_3),$$

$$\int_{E_3} \operatorname{rot} u \cdot v = \int_{E_3} u \cdot \operatorname{rot} v \quad \forall\, u,\ v \in \mathbb{L}^2_{\text{rot}}(E_3),$$

are valid. We know they are when φ is in $C_0^\infty(E_3)$ and u and v both in $\mathbb{C}_0^\infty(E_3)$. So if $\{u_n\}$ and $\{v_n\}$ are Cauchy sequences converging to u and v, then $\int \operatorname{rot} u_n \cdot v_m = \int u_n \cdot \operatorname{rot} v_m$ for all n and m, and one may pass to the limit, first with respect to n, then to m, by continuity (that is, if one insists on rigor, by applying the principle of extension by continuity to the linear continuous maps thus defined).

When D is not all E_3, however, there are surface terms in these formulas, for smooth fields:

$$(4) \qquad \int_D \operatorname{div} u \ \varphi = -\int_D u \cdot \operatorname{grad} \varphi + \int_S n \cdot u \ \varphi,$$

$$(5) \qquad \int_D \operatorname{rot} u \cdot v = \int_D u \cdot \operatorname{rot} v + \int_S n \times u \cdot v,$$

and the extension by continuity becomes a very delicate affair because of the difficulty to give sense to the limits of the *restrictions* to the boundary of fields or functions that form a Cauchy sequence. Cf. A.4.2 for a glimpse of this problem of "traces". We shall go in painful detail over only a *part* of it in Chapter 7. But one can do much mileage with the following idea, which borrows from the "distribution" point of view. Suppose u has a weak divergence in $L^2(D)$. Then, for a smooth φ, the expression $N_u(\varphi) = \int_D \operatorname{div} u \ \varphi + \int_D u \cdot \operatorname{grad} \varphi$ makes sense and vanishes for $\varphi \in C_0^\infty(D)$. It means[4] that, for a smooth φ that doesn't necessarily vanish on S, the values of $N_u(\varphi)$ depend on the *boundary* values of φ only. Therefore, we have there a linear map, $\varphi \rightarrow N_u(\varphi)$, which actually depends on the restriction φ_S of φ to S. In other words, this is a distribution defined on S (the required sequential continuity is obvious, in the case of a smooth surface). If u is smooth, $n \cdot u$ makes sense, and this distribution is seen to be the map $\varphi \rightarrow \int_S n \cdot u \ \varphi$, so we are entitled to identify it with the function $n \cdot u$. Now, moving backwards, we *define* $n \cdot u$, for u in $\mathbb{L}^2_{\text{div}}(D)$, *as* precisely this distribution. A similar approach gives sense to $n \times u$ in (5).

It's in this sense that fields of $\mathbb{L}^2_{\text{rot}}(D)$ and $\mathbb{L}^2_{\text{div}}(D)$ have well-defined tangential parts and normal parts, respectively, on the boundary, which may fail to be functions or fields in the usual sense of these words, but make sense as distributions, and reduce to the standard interpretation of $n \times u$ and $n \cdot u$ in case of regularity (of u *and* of S, of course). Note that, in contrast, $n \times u$ and $n \cdot u$ *don't* make sense for $u \in \mathbb{L}^2(D)$ if u has no more

[4] Some cheating occurs here. See [LM] for a genuine proof.

regularity than that: Square-summable fields have no traces on boundaries!

Although I cut a few corners, I hope the foregoing was reason enough for you to *use* formulas (4) and (5), in both directions, without undue apprehension, provided of course $\varphi \in L^2_{\text{grad}}(D)$ and $u \in \mathbb{L}^2_{\text{div}}(D)$, as regards (4), and both u and v belong in $\mathbb{L}^2_{\text{rot}}(D)$, as regards (5). For extra security, however, note that we can always decide that, *by definition,* $n \cdot u = 0$ means "$\int_D \operatorname{div} u \ \varphi = - \int_D u \cdot \operatorname{grad} \varphi$ for all φ in $L^2_{\text{grad}}(D)$". This doesn't go further than the weak formulation we adopted in Chapter 3: There it was for all φ in $C^\infty(\overline{D})$, but since the latter is dense in $L^2(D)$, the present interpretation is simply the application of the principle of extension by continuity. Same remark for $n \times u$, so from now on, we'll agree that

$$n \cdot u = 0 \text{ on } S \Leftrightarrow \int_D \operatorname{div} u \ \varphi + \int_D u \cdot \operatorname{grad} \varphi = 0 \quad \forall\, \varphi \in L^2_{\text{grad}}(D),$$

$$n \times u = 0 \text{ on } S \Leftrightarrow \int_D \operatorname{rot} u \cdot v - \int_D u \cdot \operatorname{rot} v = 0 \quad \forall\, v \in \mathbb{L}^2_{\text{rot}}(D),$$

with obvious adaptations in case we want such equalities on a part of S only.

Exercise 5.3. Show that the subspace $\{b \in \mathbb{L}^2_{\text{div}}(D) : \operatorname{div} b = 0\}$ is closed, not only with respect to the graph norm (which is trivial) but in the $\mathbb{L}^2$ norm as well. Same thing for $\{h \in \mathbb{L}^2_{\text{rot}}(D) : \operatorname{rot} h = 0\}$. Same question for $\{b \in \mathbb{L}^2_{\text{div}}(D) : n \cdot b = 0\}$ and for $\{h \in \mathbb{L}^2_{\text{rot}}(D) : n \times h = 0\}$.

5.1.3 Proof of Proposition 5.2

Before moving on, let's give the deferred proof.

By (1), ${}^w\mathrm{DIV}$ contains DIV, and is closed, as the orthogonal of the subspace $\{\{\operatorname{grad} \varphi', \varphi'\} : \varphi' \in C_0^\infty(D)\}$. So if it were strictly larger than $\overline{\mathrm{DIV}}$, there would exist[5] a pair $\{b, f\}$ in $\mathbb{L}^2(D) \times L^2(D)$ satisfying both (2) (p. 127) and

$$(6) \qquad \int_D b \cdot b' + \int_D f \ \operatorname{div} b' = 0 \quad \forall\, b' \in \mathbb{C}^\infty(\overline{D}),$$

which expresses orthogonality to DIV. Taking $b' = \operatorname{grad} \varphi'$ in (6), the same φ' as in (2), we see that $\int_D f\,(\varphi' - \Delta\varphi') = 0 \ \forall\, \varphi' \in C_0^\infty(D)$. *If* f was smooth, this would imply $f = \Delta f$ in D, and therefore, $0 = \int_D (f - \Delta f)\, f = \int_D |f|^2 + \int_D |\nabla f|^2$, hence $f = 0$, then $b = 0$ by (6) and density. The idea of the proof is to smooth out f by convolution before applying this trick. So

[5] By the projection theorem of A.4.3, applied in the Hilbert space $X = {}^w\mathrm{DIV}$ (with the scalar product induced by $\mathbb{L}^2(D) \times L^2(D)$) in the case where C is the closure of DIV.

let $f_n = \rho_n * f$, where ρ_n is a sequence of mollifiers, as in A.2.3, but let's *not* restrict the f_ns to D yet. Set $\delta_n = \sup(|x| : \rho_n(x) \neq 0)$ and $D_n = \{x \in D : d(x, E_3 - D) > \delta_n\}$. Notice that δ_n tends to zero, so the domain D_n grows as n increases, to eventually fill D. Now select a fixed φ' with support inside D_n, and note that $\rho_n * \varphi'$ has its support in D, so $\varphi'_n = \rho_n * \varphi'$ belongs to $C_0^\infty(D)$. After this preparation, we have, by using the Fubini theorem and the possibility of permuting $*$ and Δ,

$$\int_{E_3} f_n (\varphi' - \Delta\varphi') = \int_{E_3} (f * \rho_n)(\varphi' - \Delta\varphi') = \int_{E_3} f\,(\rho_n * (\varphi' - \Delta\varphi'))$$
$$= \int_{E_3} f\,(\varphi'_n - \Delta\varphi'_n) = \int_D f\,(\varphi'_n - \Delta\varphi'_n) = 0,$$

which shows that $f_n = \Delta f_n$ in D_n, and hence, $f_n = 0$ in D_n by the previous argument. The limit f of the f_ns must therefore be 0. ◊

5.1.4 Extending the Poincaré lemma

The three differential operators grad, rot, and div should not only be treated in parallel, but also as an integrated whole, which as one knows has a strong structure: curls of gradients vanish, curls are divergence-free, and to some extent, these properties have reciprocals. It's important to check whether such structure persists when we pass to the weak extensions.

To be more precise, we'll say that a domain of E_3 is *contractible* if it is simply connected with a connected boundary.[6] A classical result of Poincaré (cf. A.3.3) asserts that, in such a domain, a smooth curl-free [resp. div-free] field is a gradient [resp. a curl]. Is that still true if we replace the strong operators by the weak ones?

To better discuss such issues, let us introduce some vocabulary. A family of vector spaces $X^0, \ldots, X^n$ (all on the same scalar field) and of linear maps A^p from X^{p-1} to X^p, $p = 1, \ldots, d$, forms an *exact sequence at the level of* X^p if $\mathrm{cod}(A^p) = \ker(A^{p+1})$ in case $1 \le p \le d - 1$, if A^1 is injective in case $p = 0$, and if A^d is surjective in case $p = d$. An *exact* sequence is one which is exact at all levels. It's customary to discuss sequences with help of diagrams of this form:

[6] Because it can then be contracted onto one of its points by continuous deformation. "Connected" means in one piece, "simply connected" that any closed path can be contracted to a point by continuous deformation. (This is not the case, for instance, for the inside of a torus, which is connected but not simply connected. On the other hand, the space between two nested spheres forms a simply connected domain, but one whose boundary is not connected.)

$$\{0\} \rightarrow X^0 \xrightarrow{A^1} X^1 \xrightarrow{A^2} \ldots \rightarrow X^{d-1} \xrightarrow{A^d} X^d \rightarrow \{0\},$$

where $\{0\}$ is the space of dimension 0. In such diagrams, arrows are labeled with operators and the image, by any of these operators, of the space left to its arrow, is in the kernel of the next operator on the right.

The *Poincaré lemma* just evoked can then be stated as follows: For a contractible domain, the sequence

$$\{0\} \rightarrow C^\infty(\overline{D}) \xrightarrow{\text{grad}} \mathbb{C}^\infty(\overline{D}) \xrightarrow{\text{rot}} \mathbb{C}^\infty(\overline{D}) \xrightarrow{\text{div}} C^\infty(\overline{D}) \rightarrow \{0\}$$

is exact at levels 1 and 2 (at all levels[7] from 1 to $d-1$, in dimension d). For a regular bounded[8] contractible domain, we expect the following sequence, where grad, rot, and div are now the weak operators, to have the same structural property:

$$\{0\} \rightarrow L^2_{\text{grad}}(D) \xrightarrow{\text{grad}} \mathbb{L}^2_{\text{rot}}(D) \xrightarrow{\text{rot}} \mathbb{L}^2_{\text{div}}(D) \xrightarrow{\text{div}} L^2(D) \rightarrow \{0\}. \tag{7}$$

This is true, but the proof calls on some difficult technical results. Let us sketch it for level 1. By definition of the strong curl,

$$\ker(\text{rot}\,;\ \mathbb{C}^\infty(\overline{D})) = \{h : \textstyle\int_D h \cdot \text{rot}\, a' = 0 \quad \forall\, a' \in \mathbb{C}_0^\infty(D)\}.$$

If D is contractible, the left-hand side is $\text{grad}(C^\infty(\overline{D}))$, by the Poincaré lemma, so

$$\text{grad}(C^\infty(\overline{D})) = \mathbb{C}^\infty(\overline{D}) \cap (\text{rot}(\mathbb{C}_0^\infty(D))^\perp.$$

Taking the closures of both sides, we find that

$$\overline{\text{grad}(C^\infty(\overline{D}))} = \ker(\text{rot}\,;\ \mathbb{L}^2_{\text{rot}}(D)).$$

It means that if $\text{rot}\, h = 0$ in the weak sense, there is a sequence of functions φ_n, smooth over D, such that $h = \lim_{n \to \infty} \text{grad}\, \varphi_n$. Now suppose D bounded. One may impose $\int_D \varphi_n = 0$, and by a variant of the Poincaré

[7] At level 0, $\text{grad}\, \varphi = 0$ does not imply $\varphi = 0$ but only φ equal to some constant, unless $D = E_d$. At level d, and if $D = E_d$ this time, $\text{div}\, u = f$ implies $\int f = 0$, so not all f's qualify.

[8] One should be cautious with unbounded domains. For instance, as we shall have to worry about in Chapter 7, the image $\text{grad}(L^2_{\text{grad}}(E_3))$ is not closed, and thus does not fill out the kernel of rot. The closure of this image is $\ker(\text{rot})$, however, which is enough for the purposes of cohomology (see infra).

inequality (see Exercises 5.11 and 5.12, at the end of the chapter), one has then $\|\varphi_n\| \leq c(D)\ \|\text{grad } \varphi_n\|$, where $c(D)$ only depends on D. The φ_n's thus form a Cauchy sequence. Let φ be its limit. Then $h = \text{grad } \varphi$, since grad is closed. Hence the result: $\ker(\text{rot} \ ; \ \mathbb{L}^2_{\text{rot}}) = \text{grad}(L^2_{\text{grad}})$. For a similar proof at level 2, aiming at $\ker(\text{div} \ ; \ \mathbb{L}^2_{\text{div}}) = \text{rot}(\mathbb{L}^2_{\text{rot}})$, begin with Exercise 5.13, and use the "Coulomb gauge" ($\text{div } a = 0$).

As one sees, all this is difficult to establish with rigor, but the foundations are solid, and the results are easily summarized: At least in the case of bounded regular domains, *all structural properties of the sequence of operators* grad → rot → div *carry over to their weak extensions.*

5.1.5 "Maxwell's house"

We'll take this remark quite seriously and *base further study of models derived from Maxwell equations on the systematic exploitation of these structural properties* (an ambitious working program, to which the present book can only begin to contribute). Figure 5.1 should help convey the idea.

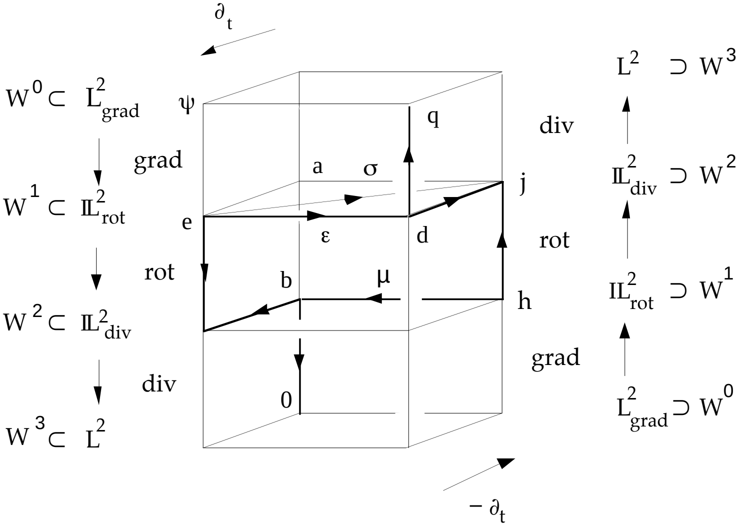

∂_t
$W^0 \subset L^2_{\text{grad}}$
ψ
grad
$W^1 \subset \mathbb{L}^2_{\text{rot}}$
e
rot
$W^2 \subset \mathbb{L}^2_{\text{div}}$
div
$W^3 \subset L^2$
q
a
σ
j
ε
d
b
μ
h
0
div
rot
grad
$L^2 \supset W^3$
$\mathbb{L}^2_{\text{div}} \supset W^2$
$\mathbb{L}^2_{\text{rot}} \supset W^1$
$L^2_{\text{grad}} \supset W^0$
$-\partial_t$

FIGURE 5.1. The functional framework for Maxwell's equations. Note how Ohm's law spoils the otherwise perfect symmetry of the structure.

The structure depicted by Fig. 5.1 is made of four copies of the sequence (7), placed vertically. The two on the left go downwards, the two on the right go upwards, which reflects the symmetry of the Maxwell equations. We need two such "pillars" on each side, linked by the time-derivative, to account for time-dependence. The four pillars are connected by horizontal beams, which link entities related by constitutive laws. This is like a building, in which as we'll see Maxwell's equations are well at home: "Maxwell's house", let's say.

Joints between pillars and beams make as many niches for electromagnetic-related entities. For instance, magnetic field, being associated with lines (dimension 1) is at level 1 on the right, whereas b, associated with surfaces (dimension 2), is at level 2 on the left, at the right position to be in front of h. Note how the equations can be read off the diagram. Ampère's relation, for instance, is obtained by gathering at level 2, right, back, the outcomes of the arrow actions on nearby fields: $-\partial_t d$ comes from the front and rot h from downstairs, and they add up to j. All aspects of the diagram shoud be as easy to understand, except the leftmost and rightmost columns. These concern the finite dimensional spaces W^p of *Whitney elements* announced in the introduction, which we now address.

5.2 THE WHITNEY COMPLEX

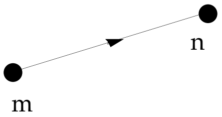
Let us start back from the notion of finite element mesh of Chapter 3: Given a regular bounded domain $D \subset E_3$, with a piecewise smooth boundary S, a *simplicial mesh* is a tessellation of D by tetrahedra, subject to the condition that any two of them may intersect along a common face, edge or node, but in no other way. We denote by $\mathcal{N}, \mathcal{E}, \mathcal{F}, \mathcal{T}$ (nodes, edges, faces, and tetrahedra, respectively) the sets of simplices of dimension 0 to 3 thus obtained,[9] and by m the mesh itself. (The possibility of having curved tetrahedra is recalled, but will not be used explicitly in this section, which means that D is assumed to be a polyhedron.)

Besides the list of nodes and of their positions, the mesh data structure also contains *incidence matrices,* saying which node belongs to which edge, which edge bounds which face, and so on. Moreover, there is a notion of orientation of the simplices, which was downplayed up to now. In short, an edge, face, etc., is not only a two-node, three-node, etc., subset of $\mathcal{N}$, but such a set *plus* an *orientation* of the simplex it subtends. Let's define these concepts (cf. A.2.5 for more details).

[9] Note that if a simplex s belongs to the mesh, all simplices that form the boundary ∂s also belong. Moreover, each simplex appears only once. (This restriction may be lifted to advantage in some circumstances, for instance when "doubling" nodes or edges, as we'll do without formality in Chapter 6.) The structure thus defined is called a *simplicial complex.*

5.2.1 Oriented simplices

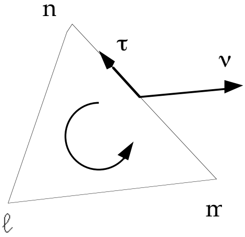
An edge {m, n} of the mesh is oriented when, standing at a point of e, one knows which way is "forward" and which way is "backward". This amounts to distinguishing two classes of vectors along the line that supports e, and to select one of these classes as the "forward" (or positively oriented) one. To denote the orientation without too much fuss, we'll make the convention that edge e = {m, n} is oriented from m to n. All edges of the mesh are oriented, and the opposite edge {n, m} is not supposed to belong to $\mathcal{E}$ if e does.



Now we define the so-called *incidence numbers* $\mathbf{G}_{e\,n} = 1$, $\mathbf{G}_{e\,m} = -1$, and $\mathbf{G}_{e\,k} = 0$ for nodes k other than n and m. They form a rectangular matrix **G**, with $\mathcal{N}$ and $\mathcal{E}$ as column set and row set, which describes how edges connect to nodes. (See A.2.2 for the use of boldface.)

Faces also are oriented, and we shall adopt a similar convention to give the list of nodes that define one and its orientation, all in one stroke: A face f = {ℓ, m, n} has three vertices, which are nodes ℓ, m, and n; we regard even permutations of nodes, {m, n, ℓ} and {n, ℓ, m}, as being the same face, and odd permutations as defining the oppositely oriented face, which is not supposed to belong to $\mathcal{F}$ if f does. This does orient the face, for when sitting at a point of f, one knows what it means to "turn left" (i.e., clockwise) or to "turn right". In more precise terms, vectors ℓm and ℓn, for instance, form a reference frame in the plane supporting f. Given two independent vectors v_1 and v_2 at a point of the face, lying in its plane, one may form the determinant of their coordinates with respect to this basis. Its sign, + or −, tells whether v_2 is to the left or to the right with respect to v_1. Observe that v_1 and v_2 also form a frame, so this sign comparison defines an equivalence relation with two classes, *positively oriented* and *negatively oriented* frames. The positive ones include {ℓm, ℓn}, and also of course {mn, mℓ} and {nℓ, nm}.

An orientation of f induces an orientation of its boundary: A tangent vector τ along the boundary is positively oriented if {ν, τ} is a direct frame, where ν is any outgoing vector[10] in the plane of f, originating from the same point as τ (inset). Thus, with respect to the orientation of the face, an edge may "run along", like



[10] No ambiguity on that: In the plane of f, the boundary is a closed curve that separates two regions of the plane, so "outwards" is well defined. Same remark for the surface of a tetrahedron (Fig. 5.2).

e = {m, n}, when its orientation matches the orientation of the boundary, or "run counter" when it doesn't.

We can now define the incidence number $\mathbf{R}_{f\,e}$: it's + 1 if e runs along the boundary, –1 otherwise, and of course 0 if e is not one of the edges of f. Hence a matrix **R**, indexed over $\mathcal{E}$ and $\mathcal{F}$.

A matrix **D**, indexed over $\mathcal{F}$ and $\mathcal{T}$, is similarly defined: $\mathbf{D}_{T\,f} = \pm 1$ if face f bounds tetrahedron T, the sign depending on whether the orientations of f and of the boundary of T match or not. This makes sense only after the tetrahedron T itself has been oriented, and our convention will be that if T = {k, ℓ, m, n}, the vectors kℓ, km, and kn, in this order, define a positive frame. (Beware: {ℓ, m, n, k} has the opposite orientation, so it does not belong to $\mathcal{T}$ if T does.) The orientation of T may or may not match the usual orientation of space (as given by the corkscrew rule): these are independent things (Fig. 5.2).

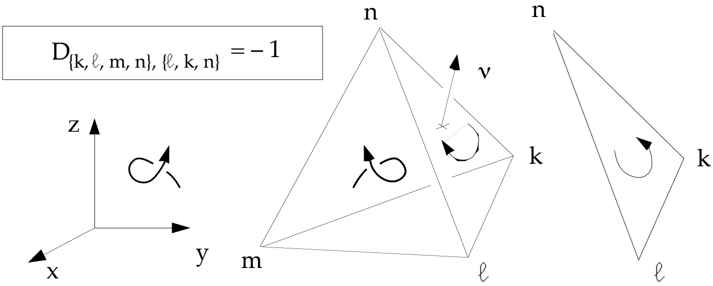
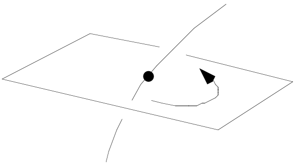


FIGURE 5.2. Left: Standard orientation of space. Right: The tetrahedron T = {k, ℓ, m, n}, "placed" this way in E_3, has "counter-corkscrew" orientation. See how, thanks to the existence of a canonical "crossing direction" (here inside-out, materialized by the outgoing vector ν), this orientation induces one on the boundary of the tetrahedron, which here happens to be opposite to the orientation of f = {k, n, ℓ}. Concepts and graphic conventions come from [VW] and [Sc], via [Bu].

Remark 5.2. The orientation of faces is often casually defined by providing each face with its own normal vector, which is what we did earlier when we had to consider crossing directions. This is all right if the ambient space E_3 has been oriented, which is what we assume as a rule (the standard orientation is that of Fig. 5.2, left). In that case, the normal vector and the ambient orientation join forces to orient the face. But there are two distinct concepts of orientation here. What we have described above is *inner* orientation, which is intrinsic and does not depend on the simplex being embedded in a larger space. In contrast, giving a crossing direction[11] for a surface is *outer,* or *external* orientation. More generally, when a manifold (line, surface, . . .) is immersed in a space of higher



dimension, an outer orientation of the tangent space at a point is by definition an inner orientation of its complement. (Outer orienting a line is thus the same as giving a way to turn around it, cf. the inset.) So *if* the encompassing space is oriented, outer orientation of the tangent space at a point of the manifold determines its inner orientation, and the other way around. (Cf. A.2.2 for more detail.) It's better not to depend on the orientation of E_3, however, so let it be clear that faces have inner orientations, like edges and tetrahedra. ◊

Remark 5.3. For consistency, one is now tempted to attribute an orientation to nodes as well, which is easy to do: just assign a sign, + 1 or – 1, to each node, and for each node n with "orientation" – 1, change the sign of all entries of column n in the above **G**. Implicitly, we have been orienting all nodes the same way (+ 1) up to now, and we'll continue to do so, but all proofs below are easily adapted to the general case. ◊

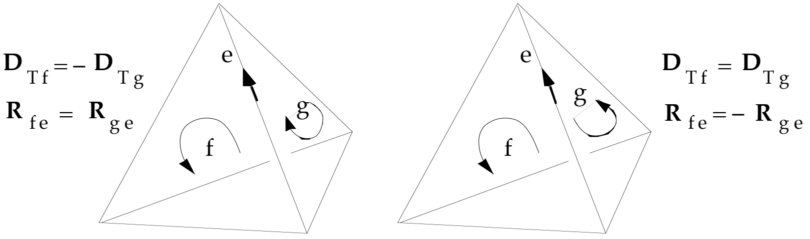

$\mathbf{D}_{Tf} = -\mathbf{D}_{Tg}$
$\mathbf{R}_{fe} = \mathbf{R}_{ge}$
e
g
f
e
g
f
$\mathbf{D}_{Tf} = \mathbf{D}_{Tg}$
$\mathbf{R}_{fe} = -\mathbf{R}_{ge}$

FIGURE 5.3. Opposition of incidence numbers, leading to **DR** = 0, whatever the orientations.

Next point:

Proposition 5.3. *One has* **DR** = 0, **RG** = 0. (Does that ring a bell?)

Proof. For $e \in \mathcal{E}$ and $T \in \mathcal{T}$, the {T, e}-entry of **DR** is $\sum_{f \in \mathcal{F}} \mathbf{D}_{Tf} \mathbf{R}_{fe}$. The only nonzero terms are for faces that both contain e and bound T, which means that e is an edge of T, and there are exactly two faces f and g of T hinging on e (Fig. 5.3). If $\mathbf{D}_{Tg} = \mathbf{D}_{Tf}$, their boundaries are oriented in such a way that e must run along one and counter the other, so $\mathbf{R}_{ge} = -\mathbf{R}_{fe}$, and the sum is zero. If $\mathbf{D}_{Tg} = -\mathbf{D}_{Tf}$, the opposite happens, $\mathbf{R}_{ge} = \mathbf{R}_{fe}$, with the same final result. The proof of **RG** = 0 is similar. ◊

[11] Which doesn't require a *normal* vector, for any outgoing vector will do; cf. Fig. 5.2, middle.

Let us finally mention some facts about the dual mesh *m** of 4.1.2, obtained by barycentric subdivision and reassembly (although we shall not make use of it as such). Each dual cell s* inherits from s an *outer* orientation (and hence, an inner one if space E_3 is oriented). Incidence relations between dual cells are described by the same matrices **G**, **R**, **D**, only transposed: $\mathbf{R}_{f\,e} \neq 0$, for instance, means that the bent edge f* is part of the boundary of the skew face e* (cf. Fig. 4.4), and so on.

5.2.2 Whitney elements

Now, we assign a function or a vector field to all simplices of the mesh. For definiteness, assume the ususal orientation of space, although concepts and results do not actually depend on it.

For notational consistency, we make a change with respect to Chapter 3, which consists of denoting by w_n the continuous, piecewise affine function, equal to 1 at n and to 0 at other nodes, that was there called λ^n. The w stands for "Whitney", and as we shall see, the hat function λ^n, now w_n, is the Whitney element of lowest "degree", this word referring not to the degree of w_n as a polynomial, but to the dimension of the simplices it is associated with (the nodes). We shall have Whitney elements associated with edges, faces, and tetrahedra as well, and the notation for them, as uniform as we can manage, will be w_e, w_f, and w_T. Recall the identity

(8) $$\sum_{n \in \mathcal{N}} w_n = 1$$

over D. We shall denote by W^0 the span of the w_ns (that was, in Chapter 3, space Φ_m). Finite-dimensional spaces W^p will presently be defined also, for p = 1, 2, 3. They all depend on *m*, and should therefore rather be denoted by $W^0(m)$, or $W^0_{\ m}$, but the index *m* can safely be understood and is omitted in what follows.

Next, degree, 1. To edge e = {m, n}, let us associate the vector field

(9) $$w_e = w_m \nabla w_n - w_n \nabla w_m$$

(cf. Fig. 5.4, left), and denote by W^1 the finite-dimensional space generated by the w_es. Similarly, W^2 will be the span of the w_fs, one per face f = {ℓ, m, n}, with

(10) $$w_f = 2(w_\ell \nabla w_m \times \nabla w_n + w_m \nabla w_n \times \nabla w_\ell + w_n \nabla w_\ell \times \nabla w_m)$$

(cf. Fig. 5.4, right). Last, W^3 is generated by functions w_T, one for each tetrahedron T, equal to $1/\mathrm{vol}(\mathrm{T})$ on T and 0 elsewhere. (Its analytical expression in the style of (9) and (10), which one may guess as an exercise, is of little importance.)

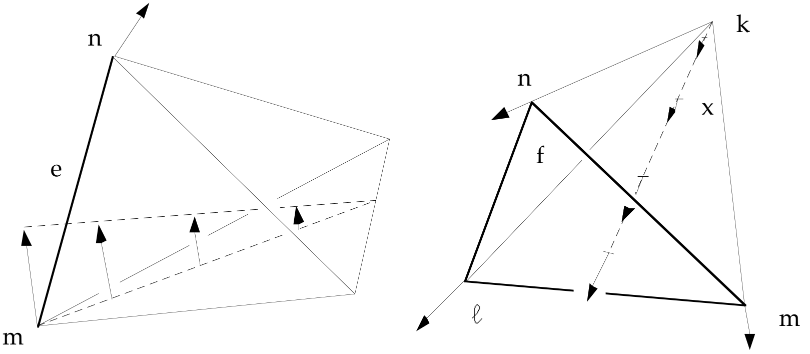


FIGURE 5.4. Left: The "edge element", or Whitney element of degree 1 associated with edge $e = \{m, n\}$, here shown on a single tetrahedron with e as one of its edges. Right: The "face element", or Whitney element of degree 2 associated with face $f = \{\ell, m, n\}$, here shown on a single tetrahedron with f as one of its faces. The arrows suggest how the vector fields w_e and w_f, as defined in (9) and (10), behave. At point m on the left, for instance, $w_e = \nabla w_n$, after (9), and this vector is orthogonal to the face opposite m. At point m on the right, $w_f = \nabla w_n \times \nabla w_\ell$, after (10); this vector is orthogonal to both ∇w_n and ∇w_ℓ, and hence parallel to the planes that support faces $\{\ell, m, k\}$ and $\{k, m, n\}$, that is, to their intersection, which is edge $\{k, m\}$.

Thus, to each simplex s is attached a field, scalar- or vector-valued. These fields are the Whitney elements. (The proper name is "Whitney forms" in the context in which they were introduced [Wh].) We'll review their main properties, all easy to prove. First,

- The value of w_n at node n is 1 (and 0 at other nodes),
- The circulation of w_e along edge e is 1,
- The flux of w_f across face f is 1,
- The integral of w_T over tetrahedron T is 1,

(and also, in each case, 0 for other simplices).

For degree 0, we already knew that, and for degree 3, it it so by way of definition. Let us prove the point for degree 1, i.e., about the circulation of w_e. Since the tangent vector τ is equal to $mn/|mn|$, one has, with help of Lemma 4.1,

$$\int_e \tau \cdot (w_m \nabla w_n) = mn \cdot \nabla w_n \, (\int_e w_m)/|mn| = (\int_e w_m)/|mn|,$$

and hence

$$\int_e \tau \cdot w_e = \int_e \tau \cdot (w_m \nabla w_n - w_n \nabla w_m) = \int_e (w_m + w_n)/\,|mn|\, = 1,$$

since $w_m + w_n = 1$ on edge $\{m, n\}$.

The reader will easily treat the case of faces. (Doing Exercise 5.4 before may help.) Note the convoluted way in which *orientation* of the ambient space intervenes (in the definition of both the cross product and the crossing direction), without influencing the final result, in spite of what one may have feared.

Exercise 5.4. Review Exer. 3.9, showing that the volume of a tetrahedron $\mathrm{T} = \{k, \ell, m, n\}$ is $\mathrm{vol}(\mathrm{T}) = 4 \int_{\mathrm{T}} w_n$. Show that the area of face $\{k, \ell, m\}$ is 3 $\mathrm{vol}(\mathrm{T})\ |\nabla w_n|$, the length of vector $\{k, \ell\}$ is $6\ \mathrm{vol}(\mathrm{T})\ |\nabla w_m \times \nabla w_n|$, and that $6\ \mathrm{vol}(\mathrm{T}) \det(\nabla w_k, \nabla w_\ell, \nabla w_m) = 1$.

Exercise 5.5. Compute $\int_{\mathrm{T}} w_e \cdot w_{e'}$, according to the respective positions of edges e and e', in terms of the scalar products $\nabla w_n \cdot \nabla w_m$.

Exercise 5.6. Show that field (9) is of the form $x \rightarrow a \times x + b$ in a given tetrahedron, where a and b are three-component vectors, vector a being parallel to the edge opposite $\{m, n\}$. Show that the field (10) is of the form $x \rightarrow \alpha\, x + b$ (where now $\alpha \in \mathbb{R}$).

A second group of properties concerns the continuity, or lack thereof, of the w's across faces of the mesh. Function w_n is continuous. For the field w_e, it's more involved. Let us consider two tetrahedra with face $\{\ell, m, n\}$ in common, and let x be a point of this face. Then the vector field ∇w_n is not continuous at x, since w_n is not differentiable. But on the other hand, the tangential part (cf. Fig. 2.5) of ∇w_n on face $\{\ell, m, n\}$ changes in a continuous way when one crosses the face from one tetrahedron to its neighbor; indeed, it only depends on the values of w_n on this face, whatever the tetrahedron one considers. As this goes the same for ∇w_m, and for all faces of the mesh, one may conclude that the tangential part of w_e is continuous across faces. Similar reasoning shows that the *normal* part of w_f is continuous across faces. As for w_{T}, it is just discontinuous.

Thanks to these continuity properties, W^0 is contained in L^2_{grad}, W^1 in $\mathbb{L}^2_{\mathrm{rot}}$, and W^2 in $\mathbb{L}^2_{\mathrm{div}}$. The W^p are of finite dimension. They *can therefore play the role of Galerkin approximation spaces* for the latter functional spaces. We knew that as far as W^0 is concerned. For $p = 1$ or 2, however, this calls for an unconventional interpretation of the degrees of freedom. Take h in W^1, for instance. Then, by definition,

$$(11) \qquad h = \sum_{e \in \mathcal{E}} \mathbf{h}_e \, w_e,$$

where each $\mathbf{h}_e$ (set in boldface) is a scalar coefficient. As the circulation of w_e is 1 along edge e and 0 along others, the circulation of h along edge e is the degree of freedom $\mathbf{h}_e$. So the DoFs are associated with *edges* of the mesh, not with *nodes*, which is the main novelty with respect to classical finite elements. In the same way, if $b \in W^2$, one has $b = \sum_{f \in \mathcal{F}} \mathbf{b}_f w_f$, and the $\mathbf{b}_f$s are the fluxes of b through faces. So in this case, degrees of freedom sit at faces. Last, there is one DoF for each tetrahedron in the case of functions belonging to W^3.

Remark 5.4. So the w_es (as well as other Whitney elements) are linearly independent, for $h = 0$ in (11) implies $\mathbf{h}_e = 0$ for all e (cf. Exer. 3.8). ◊

The convergence properties of Whitney elements are quite similar to those we already know as regards W^0. Let φ be a smooth function, and set $\varphi_m = \sum_{n \in \mathcal{N}} \boldsymbol{\varphi}_n w_n$, where $\boldsymbol{\varphi}_n$ is the value of φ at node n. (This is the *m*-interpolate of Subsection 4.3.1, with adapted notation.) When the mesh is refined, so that the grain tends to zero, while avoiding "asymptotic flattening" of the simplices, φ_m converges towards φ in $L^2_{\text{grad}}(D)$, as we proved in Chapter 4. In the same way, if h is a smooth vector field, if $\mathbf{h}_e$ is the circulation of h along edge e, and if one sets $h_m = r_m h = \sum_{e \in \mathcal{E}} \mathbf{h}_e w_e$, then h_m converges to h in $\mathbb{L}^2_{\text{rot}}(D)$. Same thing for $b_m = \sum_{f \in \mathcal{F}} \mathbf{b}_f w_f$, where $\mathbf{b}_f$ is the flux of b through f, with convergence with respect to the norm of $\mathbb{L}^2_{\text{div}}(D)$. See [Do] for proofs.

5.2.3 Combinatorial properties of the complex

The properties we have noticed (nature of the degrees of freedom, continuity, convergence) concerned spaces W^p as taken one by one, for different values of p. But there is more: properties of the structure made by all the W^ps when taken together, or "Whitney complex", which are even more remarkable. These structural properties are what makes possible a discretization of the structure of Fig. 5.1 *as a whole*. First:

Proposition 5.4. *The following inclusions hold:*

$$(12) \qquad \text{grad}(W^0) \subset W^1, \;\; \text{rot}(W^1) \subset W^2, \;\; \text{div}(W^2) \subset W^3.$$

Proof. Let us consider node m. If $\mathbf{G}_{e\,n} \neq 0$, either $e = \{m, n\}$ or $e = \{n, m\}$, but in both cases, $\mathbf{G}_{e\,m} w_e = w_n \nabla w_m - w_m \nabla w_n$, by definition of the incidence numbers $\mathbf{G}_{e\,n}$. Therefore,

$$\sum_{e \in \mathcal{E}} \mathbf{G}_{e\,m} w_e = \sum_{n \in \mathcal{N}} (w_n \nabla w_m - w_m \nabla w_n)$$
$$= (\sum_{n \in \mathcal{N}} w_n)\, \nabla w_m - w_m\, \nabla(\sum_{n \in \mathcal{N}} w_n) \equiv \nabla w_m$$

since $\sum_{n \in \mathcal{N}} w_n \equiv 1$, hence grad $w_m \in W^1$, and hence the first inclusion by linearity. Similarly, for $e = \{m, n\}$, one has rot $w_e = 2\, \nabla w_m \times \nabla w_n = \sum_{f \in \mathcal{F}} \mathbf{R}_{f\,e}\, w_f$, hence rot $w_e \in W^2$, and div $w_f = \sum_{T \in \mathcal{T}} \mathbf{D}_{T\,f}\, w_T$, that is to say, div $w_f \in W^3$ (**Exercise 5.7**: prove all this). Hence (12). ◊

This result has the following important implication. If one sets $h =$ grad φ, where $\varphi = \sum_{n \in \mathcal{N}} \boldsymbol{\varphi}_n w_n$ is an element of W^0, this field h can also be expressed as in (11), the edge DoF being $\mathbf{h}_{\{m,\, n\}} = \boldsymbol{\varphi}_n - \boldsymbol{\varphi}_m$. Let $\mathbf{h}$ be the vector of the $\mathbf{h}_e$s (of length E, the number of edges), and $\boldsymbol{\varphi}$ the vector of the $\boldsymbol{\varphi}_n$s (of length N, the number of nodes). Then $\mathbf{h} = \mathbf{G}\boldsymbol{\varphi}$, where $\mathbf{G}$ is the above E × N incidence matrix, which thus appears as a discrete analogue of the gradient operator, via the correspondence between the potential φ [resp. the field h] and the associated **vector**[12] of DoFs $\boldsymbol{\varphi}$ [resp. $\mathbf{h}$].

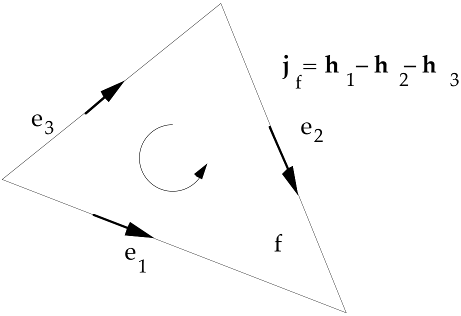


FIGURE 5.5. Computing $\mathbf{j}_f$ (flux through face f of field j = rot h), from the DoFs of h. Here, $\mathbf{h}_i$ is the circulation of h along edge e_i, the orientation of the e_is with respect to f is indicated by the arrows, and the terms of the incidence matrix $\mathbf{R}$ are $\mathbf{R}_{f\,e1} = 1$, $\mathbf{R}_{f\,e2} = \mathbf{R}_{f\,e3} = -1$.

Similarly (Fig. 5.5), if $h = \sum_{e \in \mathcal{E}} \mathbf{h}_e\, w_e$, then $j \equiv$ rot $h = \sum_{f \in \mathcal{F}} \mathbf{j}_f\, w_f$, where the $\mathbf{j}_f$ s form the components of **vector** $\mathbf{j} = \mathbf{R}\mathbf{h}$, of dimension F (the number of faces). Last, one has div $b = \sum_{T \in \mathcal{T}} \boldsymbol{\psi}_T\, w_T$, where $\boldsymbol{\psi} = \mathbf{D}\,\mathbf{b}$, when $b = \sum_{f \in \mathcal{F}} \mathbf{b}_f\, w_f$. Matrices $\mathbf{R}$ and $\mathbf{D}$, of respective dimensions F × E and T × F (where T is the number of tetrahedra), thus correspond to the curl and the divergence. We now understand the equalities $\mathbf{D}\,\mathbf{R} = 0$ and $\mathbf{R}\,\mathbf{G} = 0$: They are the discrete counterparts of the differential relations div(rot .) = 0 and rot(grad .) = 0.

[12] See A.2.2 about this use of boldface for DoF-vectors. This is only an attention-catching device, which will not be used throughout.

We'll denote by $\mathbf{W}^p$, $p = 0$ to 3, the spaces $\mathbb{R}^{\mathcal{N}}$, $\mathbb{R}^{\mathcal{E}}$, $\mathbb{R}^{\mathcal{F}}$, $\mathbb{R}^{\mathcal{T}}$, isomorphic to the Cartesian products $\mathbb{R}^N$, $\mathbb{R}^E$, etc. These spaces are isomorphic, for a given *m*, to the $W^p{}_m$, but conceptually distinct from them. We can summarize all our findings by the following sketch, called a *commutative diagram*,[13] which describes the structure of Whitney element spaces:

(13)
$$\begin{tikzcd}
W^0 \arrow[r, "\mathrm{grad}"] \arrow[d, dash] & W^1 \arrow[r, "\mathrm{rot}"] \arrow[d, dash] & W^2 \arrow[r, "\mathrm{div}"] \arrow[d, dash] & W^3 \arrow[d, dash] \\
\mathbf{W}^0 \arrow[r, "\mathbf{G}"'] & \mathbf{W}^1 \arrow[r, "\mathbf{R}"'] & \mathbf{W}^2 \arrow[r, "\mathbf{D}"'] & \mathbf{W}^3
\end{tikzcd} .$$

Graphic conventions should be obvious, once it is understood that vertical arrows denote isomorphisms.

Whether the top and bottom sequences in (13) are exact is then a natural question. The answer depends on the topology of D.

Proposition 5.5. *If the set–union of all tetrahedra in the mesh is contractible, one has the following identities:*

$$W^1 \cap \ker(\mathrm{rot}) = \mathrm{grad}\ W^0, \qquad W^2 \cap \ker(\mathrm{div}) = \mathrm{rot}\ W^1,$$

in addition to (12).

Proof. Let h be an element of W^1 such that $\mathrm{rot}\ h = 0$. Then (D being simply connected) there exists a function φ such that $h = \mathrm{grad}\ \varphi$. The $\boldsymbol{\varphi}_n$s being the values of φ at nodes, let us form $k = \mathrm{grad}(\sum_{n \in \mathcal{N}} \boldsymbol{\varphi}_n w_n)$. Then $k \in W^1$ by the first inclusion of Prop. 5.4, and its DoFs are those of h by construction, so $h = k \in \mathrm{grad}\ W^0$. As for the second equality, take an element b of W^2 such that $\mathrm{div}\ b = 0$. There exists[14] a vector field a such that $b = \mathrm{rot}\ a$. The $\mathbf{a}_e$s being the circulations of a along the edges, let us form $c = \mathrm{rot}(\sum_{e \in \mathcal{E}} \mathbf{a}_e\ w_e)$. Then $c \in W^2$ by the second inclusion, and its DoFs are those of b by construction, hence $b \equiv c \in \mathrm{rot}\ W^1$. ◊

[13] In practice, it means that, by following a path on the diagram, and by composing the operators encountered along the way, the operator thus obtained depends only on the points of departure and arrival. Allowed paths are along the arrows (in the direction indicated) and along unarrowed segments (in both directions).

[14] Beware, "D simply connected" is not enough for that, and the hypothesis "S connected" cannot be forgotten. For instance, if $D = \{x \in E_3 : 1 < |x| < 2\}$, the field $\mathrm{grad}(x \rightarrow 1/|x|)$ is divergence-free, since the function $x \rightarrow 1/|x|$ is harmonic, but is not a curl, since its flux across the closed surface $\{x : |x| = 1\}$ does not vanish.

So the image fills the kernel at both middle positions of diagram (13) if D is topologically trivial, i.e., contractible. But things are more interesting the other way around, for if the sequences in (13) *fail* to be exact at one of these positions or both, this tells something about the topology of D. For instance, the existence of curl-free fields that are not gradients implies the presence of one or more "loops" in the domain (as for a torus, which has one such loop). Solenoidal fields which are not curls can't exist unless there is a "hole", as when D is the volume between two nested spheres. The sequences are thus an algebraic tool by which the topology of D can be explored (and this of course was Whitney's concern). Although topological difficulties are avoided in this book as a rule, the reader may be interested by the information on all this contained in the next subsection.

The main interest of the Whitney complex from our point of view lies elsewhere, however. Propositions 5.4 and 5.5 justify the replacement of each pillar of Maxwell's house, Fig. 5.1, by one of the isomorphic sequences of (13). Hence, for a given mesh, a "discrete" building, or "Maxwell–Whitney house", in which we'll try to embed any problem at hand, thus obtaining a modelling in finite terms. It is already clear that the two "vertical" equations, $-\partial_t \mathrm{d} + \mathrm{rot\ h} = \mathrm{j}$ and $\partial_t \mathrm{b} + \mathrm{rot\ e} = 0$, will be discretized as $-\partial_t \mathbf{d} + \mathbf{R}\ \mathbf{h} = \mathbf{j}$ and $\partial_t \mathbf{b} + \mathbf{R}\ \mathbf{e} = 0$. The difficulty, therefore, lies in the discretization of the constitutive laws. This will be our main concern in Chapters 6 to 9.

5.2.4 Topological properties of the complex

(This subsection is independent, and can be skipped.) In the case of a contractible domain, we just proved the sequence

$$\{0\} \rightarrow \mathrm{W}^0 \xrightarrow{\mathrm{grad}} \mathrm{W}^1 \xrightarrow{\mathrm{rot}} \mathrm{W}^2 \xrightarrow{\mathrm{div}} \mathrm{W}^3 \rightarrow \{0\}$$

exact at all levels except 0. As we knew beforehand, the following sequence has the same property, in the case of a regular bounded domain:

$$(14) \qquad \{0\} \rightarrow \mathrm{L}^2_{\mathrm{grad}} \xrightarrow{\mathrm{grad}} \mathbb{L}^2_{\mathrm{rot}} \xrightarrow{\mathrm{rot}} \mathbb{L}^2_{\mathrm{div}} \xrightarrow{\mathrm{div}} \mathrm{L}^2 \rightarrow \{0\}.$$

This is no coincidence, as we shall verify for two particular cases where D is not contractible.

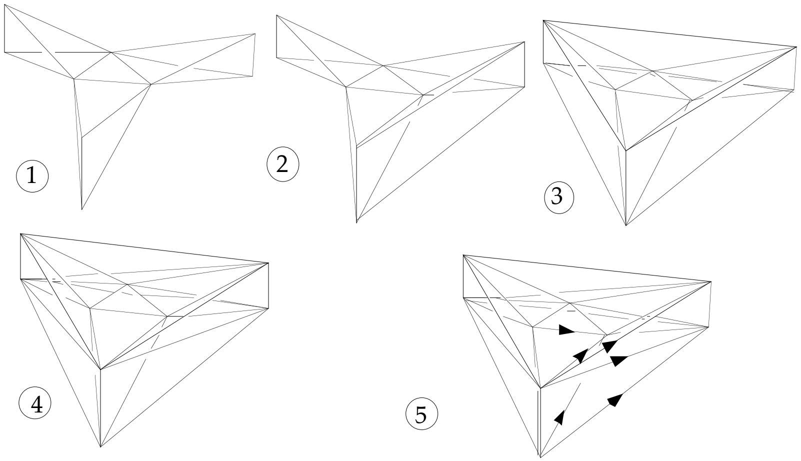


FIGURE 5.6. Steps in the construction of a "simplicial torus": Join three tetrahedra around a triangle (1), add a pyramid (2), then two others (3), in order to form a solid ring, then cut each pyramid in two tetrahedra (4). The toric polyhedron thus obtained comprises 9 tetrahedra, 27 faces, 27 edges, and 9 vertices ($\chi = 0$). In (5), how to assign DoFs to edges in order to get a curl-free field in W^1 which is not a gradient.

Let's first consider the case of the mesh[15] of a torus, Fig. 5.6. Let us assign the DoF $\mathbf{h}_e = 0$ to all edges, except the six shown in Fig. 5.6, for which $\mathbf{h}_e = 1$ (the arrows mark orientation). One obtains this way an element h of W^1 which is curl-free (this can be checked by summing the $\mathbf{h}_e$s along the perimeters of all faces, hence $\mathbf{R}\,\mathbf{h} = 0$), but is certainly not a gradient, since its circulation does not vanish along some closed circuits, such as the one formed by the boundary of the empty central triangle, for instance.

So grad W^0 is strictly contained in ker(rot ; W^1). The quotient[16] ker(rot ; W^1)/grad(W^0) then does not reduce to 0, and it's easy to see its dimension is 1 in the present case. In the general case, this dimension is called the *Betti number of dimension* 1 of the mesh. This number measures the lack of exactitude at level 1 of the Whitney sequence. Let's denote it $b_1(m)$, or just b_1.

Now, if one considers the sequence (14) relative to this toric volume, one sees the same lack of exactitude. Indeed, curl-free fields in this torus which are not gradients can all be obtained by adding some gradient to a

[15] A very coarse mesh, but this doesn't matter: Properties proved this way are mesh-independent.

[16] This notion is discussed in A.1.6 and A.2.2.

multiple of the just constructed special field. The dimension of the quotient $\ker(\text{rot} ; \, \mathbb{L}^2_{\text{rot}}) \, / \, \text{grad}(L^2_{\text{grad}})$ is therefore equal to 1.

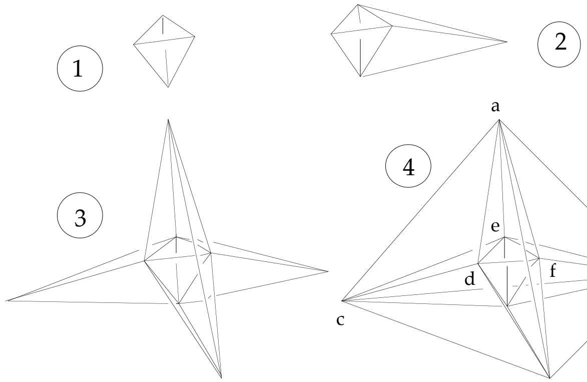


FIGURE 5.7. Steps in the construction of a "hollow tetrahedron": To the faces of a regular tetrahedron (1), stick four tetrahedral spikes (2–3), then six tetrahedra that share an edge with the central one (4), and finally the four tetrahedra necessary to fill up the flanks. Remove the central tetrahedron. The hollow solid that remains comprises 14 tetrahedra, 32 faces, 24 edges, and 8 vertices ($\chi = 2$).

A similar phenomenon can be observed in the case of the hollow tetrahedron of Fig. 5.7. By assigning the DoF 0 to all faces except {a, c, e}, {a, d, e}, and {d, e, f}, which are given the DoF 1, one obtains a solenoidal field b in W^2 (add fluxes through faces for all tetrahedra, hence $\mathbf{D}\,\mathbf{b} = 0$), but not the curl of any field, since its flux through some closed surfaces, such as for instance the boundary of the inner tetrahedron, does not vanish. This time, $\text{rot}\, W^1$ is strictly contained in $\ker(\text{div} ; W^2)$, and the dimension of the quotient $\ker(\text{div} ; W^2) / \text{rot}(W^1)$ is 1. In the general case, this dimension is called the *Betti number of dimension* 2 of the mesh (denoted b_2) and measures the lack of exactitude of the sequence at level 2.

These departures from exactitude thus appear as *global topological properties* of the meshed domain. From what precedes, one can guess that the Betti numbers b_1 and b_2 are respectively the numbers of "loops" and "holes" in D, and do not depend on the mesh. The foregoing observations thus suggest that the Whitney sequence is a tool of algebraic and combinatorial nature that is able to convey topological information.

Indeed, this sequence is one of the constructions of *algebraic topology*, the part of mathematics that is concerned with associating algebraic objects (invariant by homeomorphism) to topological spaces, in order to study topology by the methods of algebra. Thus, for instance, what we said about loops and holes actually goes the other way: These intuitive notions receive a proper definition by considering basis elements of some quotient spaces, the dimensions of which are the Betti numbers, as in the two foregoing examples. Algebraic topology offers several constructions of this kind. One is *homology*, which we used extensively up to now without being formal about it (but see next subsection). Another is *cohomology*, which roughly speaking consists in setting up sequences similar to (13) or (14). For instance, the grad–rot–div sequence is the three-dimensional case of *de Rham's cohomology* (for which, as we saw, it's unimportant whether strong or weak operators are meant, at least for regular bounded domains). The Whitney sequence thus appears as a kind of *discretized* cohomology, lending itself to (combinatorial) computations, a definite advantage over de Rham's one, and this is why it was developed [Wh].

Though following this direction would be of the utmost interest, this is not the place, and anyway, the only result of topology we really need is the following one. Having defined the Betti numbers by $b_p = \dim(\ker(d\,;\,W^p)/dW^{p-1})$, $p = 1$ to 3, where d stands for grad, rot, div, according to the value of p, and b_0 as the dimension of the kernel ker(grad) in W^0 (equal to the number of connected components of D), one proves these numbers are *topological invariants*, meaning they depend on D up to homeomorphism, but not on the mesh. The integer $\chi = b_0 - b_1 + b_2 - b_3$ is called the *Euler–Poincaré constant* of the domain. By the very definition of the b_ps, one has the already met *Euler–Poincaré formula:*

$$(15) \qquad N - E + F - T = \chi(D),$$

where N, E, F, T are the numbers of simplices of all kinds, as previously defined. The constant χ is typically equal to 0, 1 or 2 (cf. Figs. 5.6 and 5.7). When the meshed region is bounded and contractible, $\chi = 1$. A similar formula holds of course in all dimensions (we had use for the two-dimensional one already), and not only for domains of E_d, but for all topological spaces that admit of simplicial meshes.

Exercise 5.8. In dimension 2, prove by direct counting that $N - E + F$ is the same for meshes m and $m/2$ (Subsection 4.1.2).

5.2.5 Metric properties of the complex

All that precedes was of *combinatorial* character. Matrices **G**, **R**, **D** encompass all the knowledge on the topology of the mesh, but say nothing of *metric* properties: lengths, angles, areas, etc. To take these into account, we introduce the following "mass matrices".

Let α be a function over D, strictly positive. (For our needs here, it will be one of the coefficients ε, μ, etc., or its inverse.) We denote by $\mathbf{M}_p(\alpha)$, $p = 0, 1, 2, 3$, the square matrices of size $N \times N$, $E \times E$, $F \times F$, $T \times T$, whose entries are

$$(16) \qquad (\mathbf{M}_p(\alpha))_{s\,s'} = \int_D \alpha\, w_s \cdot w_{s'} \quad \text{if } p = 1 \text{ or } 2,$$

$$= \int_D \alpha\ w_s\ w_{s'} \quad \text{if } p = 0 \text{ or } 3,$$

where s and s' are two simplices of dimension p. The $\mathbf{M}_p$s are called *mass matrices* because one of them ($\mathbf{M}_1$) is found in the same position as the mass matrix of a vibrating mechanical system when one sets up the numerical scheme for computing the modes of a resonating cavity, as we'll see in Chapter 9.

Note that in the first line, the coefficient α can be replaced by a symmetrical tensor of Cartesian components $\alpha_{i\,j}$:

$$(\mathbf{M}_p(\alpha))_{s\,s'} = \int_D \sum_{i,\, j\, =\, 1,\, 2,\, 3} \alpha_{ij}\ w^i_s\, w^j_{s'}\,.$$

This makes possible the consideration of *anisotropic* materials.

5.3 TREES AND COTREES

In the practice of computation, the need arises to sort out the curl-free fields among fields in W^1 and (though less often) the solenoidal fields among fields in W^2.

Why is that a problem? Aren't curl-free fields sufficiently characterized, in terms of the DoF vector $\mathbf{h}$, by $\mathbf{Rh} = 0$? They are, but this is an *implicit* characterization, by algebraic constraints on $\mathbf{h}$. That such vectors be of the form $\mathbf{h} = \mathbf{G}\boldsymbol{\varphi}$, at least in the contractible case, often helps, because there are no constraints on $\boldsymbol{\varphi}$. (It *did* help in Chapters 2 and 3, where we treated the equation rot h = 0 by the introduction of a magnetic potential.) But one may ask for more and better: an *explicit* representation of the subspace $\{\mathbf{h} \in \mathbf{W}^1 : \mathbf{Rh} = 0\}$ by way of a *basis* for it, that is, some family $\{\mathbf{h}^1, \mathbf{h}^2, \ldots, \mathbf{h}^{N-1}\}$ of independent DoF vectors[17] that would generate $\ker(\mathbf{R})$. A similar problem arises in relation with gauging: One may wish to select a basis of independent vectors $\{\mathbf{a}^1, \ldots, \mathbf{a}^A\}$ in $\mathbf{W}^1$ the span of which is the codomain $\mathbf{R}\mathbf{W}^1$ (equal to $\{\mathbf{b} \in \mathbf{W}^2 : \mathbf{Db} = 0\}$ in the contractible

[17] It should be clear that their number will be N – 1 (where N in the number of nodes), in the contractible case.

case), for this singles out a unique $a \in W^1$ such that $b = \text{rot } a$, given a solenoidal b in W^2. This is what "trees" and "cotrees" are about.

Exercise 5.9. Show that, if D is contractible, the dimension A of $\ker(\mathbf{D})$ is $E - N + 1$.

In Chapters 6 and 8, and in Appendix C, we shall have several examples of use of such techniques, which are popular nowadays [AR, Fu, GT, Ke, PR, RR, T&, . . .]. Alas, due to their origins in circuit–graphs theory [Ha], their intimate connection with *homology* is generally overlooked, which is a pity. So this may be the right time to disclose a few elements of homology, at least those necessary to understand trees and cotrees.

5.3.1 Homology

The basic concept is that of "chain". Call $\mathcal{S}_p$ the sets of p-simplices of the mesh. A p-*chain* c is then simply the assignment to each $s \in \mathcal{S}_p$ of a number $\mathbf{c}_s$, i.e., a family of numbers indexed on $\mathcal{S}_p$. This is conveniently denoted by $c = \sum_{s \in \mathcal{S}_p} \mathbf{c}_s\, s$. (Note, as usual, the one-to-one correspondence between the chain c and the **vector** $\mathbf{c} = \{\mathbf{c}_s : s \in \mathcal{S}_p\}$.)

This may sound more abstract than it really is: Think for instance—in connection with our model problem—about a path of edges of the mesh, going from S^h_0 to S^h_1. It's an oriented line, so each edge runs either "along" or "counter" this orientation (cf. p. 137), hence a number $\mathbf{c}_e = \pm 1$ for each edge of the path. Assigning the number 0 to all other edges of the mesh, we do have a 1-chain. This makes precise the fuzzy notion of "circuit" by which, obviously, we mean more than the supporting line: A circuit is a line *plus* a way to run along it; so when the line is made of oriented edges, we need to tell the proper direction along each edge, which is precisely what the chain coefficients do.

In dimension 2, the concept is just as useful to make precise the notion of "polyhedral surface composed of faces of the mesh" that we repeatedly invoked. (Think about it.) What we had to call up to now, in a rather clumsy way, m^*-lines and m^*-surfaces, are just chains over the *dual* simplices, with $p = 1$ or 2.

Note that chains encompass more than that. Rendering the concept of a circuit "run k times", for instance, is obvious: a 1-chain with coefficients ± k. A collection of m-paths ("open circuits" composed of edges of the mesh), not necessarily connected, also is a chain, and so on. Non-integer coefficients make less intuitive sense, of course (although one can think of various useful interpretations in electromagnetism). Indeed, there are several versions of homology, depending on which kind of numbers the coefficients $\mathbf{c}_s$ are allowed to be. Most often, they are taken as relative integers. But there is some gain in simplicity in assuming real-valued coefficients, as we shall do here.

One can add chains ($c + c'$ is the chain $\sum_{s \in \mathcal{S}} (\mathbf{c}_s + \mathbf{c}'_s)\, s$) and multiply a chain by a scalar. The set of all p-chains, that we shall denote by[18] $W_p(D)$, is thus a vector space.[19]

Next concept: The *boundary* operator ∂. This is a linear map, which assigns a $(p-1)$-chain ∂c to any p-chain c. By linearity, $\partial c \equiv \partial(\sum_{s \in \mathcal{S}} \mathbf{c}_s\, s) = \sum_{s \in \mathcal{S}} \mathbf{c}_s\, \partial s$. To fully specify ∂, therefore, we need only state what the boundary ∂s is for any single simplex s. By definition,

$$\partial e = \sum_{n \in \mathcal{N}} \mathbf{G}_{e\,n}\, n, \quad \partial f = \sum_{e \in \mathcal{E}} \mathbf{R}_{f\,e}\, e, \quad \partial T = \sum_{f \in \mathcal{F}} \mathbf{D}_{T\,f}\, f.$$

This makes perfect sense: The boundary of $e = \{m, n\}$, for instance, is thus the chain $n - m$ (assuming all nodes have orientation $+1$), or if one prefers, the 0-chain c with $\mathbf{c}_n = 1$, $\mathbf{c}_m = -1$, and $\mathbf{c}_k = 0$ for other nodes. The ∂ of a 0-chain we can define for thoroughness as a special (and unique) "(–1)-chain" denoted 0. (It doesn't matter much.) Be aware that a boundary is more than the topological boundary, just as a chain is more than the set–union of simplices supporting it.

We'll say a chain c is *closed* if $\partial c = 0$. (One often says, a bit improperly, that its boundary is "empty".) Closed chains are rather called *cycles*, in standard texts, but the word "closed" is convenient to make contact with our observations of Chapter 4 (cf. Fig. 4.6). Notice that ∂ is represented by a matrix: $\mathbf{G}^t$, $\mathbf{R}^t$, $\mathbf{D}^t$, depending on the dimension p. Observe also how the contents of Proposition 5.3 (cf. Fig. 5.3) can now elegantly be summarized: $\partial\partial = 0$, "the boundary of a boundary is empty".

A p-chain c is *a boundary* if there is a $(p+1)$-chain γ such that $c = \partial\sigma$. Boundaries are cycles, of course. But not all cycles are boundaries . . . and from there we might go into topology again. See a specialized book (e.g., [HW]) for the way Betti numbers can be redefined, as dimensions of the quotients[20] $\ker(\partial;\ W_p)/\partial W_{p+1}$. That this technique and the foregoing

[18] It should rather be $W_p(\mathcal{S}_p(D))$, where $\mathcal{S}_p(D)$ denotes the set of p-simplices of the mesh m of D, but this is too heavy notation, and I hope the few abuses of this kind that follow will be harmless.

[19] If the case of $\mathbb{Z}$-valued chain-coefficients, $W_p(D)$, more classically denoted by $C_p(D)$ in algebraic topology, is only what algebraists call a *module* (the structure which is to a ring, here $\mathbb{Z}$, what a vector space is to a field).

[20] Two p-chains c and c' are *homologous* modulo Δ if $\partial(c - c')$ is a chain on Δ, i.e., $\partial(c - c') \in W_p(\Delta)$. The classes of this equivalence relation are called [relative] *homology classes* mod Δ. These are the formal definitions of the notions evoked in Exercises 2.5 and 2.6, at least for paths and surfaces made of simplices of the mesh. (Lifting this restriction is not difficult; this is the concern of *singular homology* [GH].)

one, based on Whitney fields and the d operator, thus yield the same topological information, is one of the great *duality* features of algebraic topology. Rather than being formal about that, let's just point to the following fact: fields and 1-chains of W^1 and W_1 are in duality via the formula

$$(17) \qquad <h, c> = \sum_{e \in \mathcal{E}} \mathbf{h}_e \mathbf{c}_e \equiv (\mathbf{h}, \mathbf{c}),$$

which stems from the basic property of edge elements, $\int_e \tau \cdot w_\varepsilon = \delta_{e\,\varepsilon}$. (Observe how (17) generalizes the concept of circulation of h along a path c.) What is meant by "in duality" is that $<h, c> = 0 \ \ \forall$ c implies h $= 0$ and the other way around. One should understand from this how the concepts of "curl-free field which is not a gradient" and "1-cycle which is not a 1-boundary" are dual, and be able to generalize to $p > 1$. It's also illuminating to think of the Stokes theorem as the statement $<\mathrm{d}h, c> = <h, \partial c>$ for all $h \in W^p$ and $c \in W_{p+1}$, and to remark that the matrix representations of d and ∂ are transposed of each other. The duality (17) is also the key to an explanation of *why* the Whitney elements have the form they have ((9) and (10)); see [B1] on this.

To be really useful, all these notions need to be "relativized", the same concepts being redefined "modulo something", as follows. Suppose our simplices are those of the mesh of a domain D, and let Δ be a closed part of D which is itself a union of simplices of the mesh. (Often, Δ will be the surface of the domain, or a part of it, but it's not the only possibility; Δ might correspond, in some magnetostatics problems, to regions inside D occupied by bodies of high permeability, taken as infinite in the modelling.) Let us denote by $W_p(\Delta)$ the set of chains over Δ : all p-chains over D whose coefficients are all 0, except for p-simplices belonging to Δ. Now we say that a chain $c \in W_p(D)$ is *closed mod* Δ if $\partial c \in W_p(\Delta)$. A p-chain c *bounds mod* Δ if there exists a $(p + 1)$-chain γ such that $c - \partial\gamma \in W_p(\Delta)$. (Rather than puzzling over these definitions, look again at Fig. 4.6, take Δ as S^b or S^h as the case may be, and imagine the various paths and surfaces as made of edges and faces of the mesh.) Chains closed mod Δ, or boundaries mod Δ, are also called *relative* cycles or boundaries (meaning, relative to Δ).

5.3.2 Trees, co-edges

Now we have enough to introduce trees. To well understand the two definitions that follow, ignore the bracketed parts first, then think again about the "relative" version:

Definition 5.1. *A set* S^T *of* p-*simplices of the mesh* m *such that* $W_p(S^T)$ *does not contain any cycle* [*mod* Δ], *except the null one, is called a* tree *of dimension* p, *or* p-tree [mod Δ].

Definition 5.2. *A* p-*tree is a* spanning tree [mod Δ] *if there is no strictly larger* p-*tree* [mod Δ] *containing it. The set of all left-over simplices [not belonging to* Δ] *is called the associated* tree complement [mod Δ], *or* cotree [mod Δ]. *Its elements are th*e co-simplices *with respect to this tree.*

To grasp this, take $p = 1$ and an empty Δ. Then $\partial c = 0$ means that vector $\mathbf{G}^t\mathbf{c}$, of length N, which is a linear combination of *rows* of the matrix **G**, vanishes. Algebraically, therefore, extracting a spanning tree is equivalent to finding a *maximal set of independent rows* of **G** (or **R**, or **D**), which amounts to looking for a submatrix of maximal rank[21]—a standard problem in linear algebra, all the more easy than matrix entries are integers. When $\Delta \neq \emptyset$, *relative* trees are obtained by the same procedure, but after removal of all rows and columns corresponding to simplices that belong to Δ.

Other rows, those corresponding to co-edges, are thus expressible as linear combinations of the previous ones, and form a basis for $\ker(\mathbf{G}^t)$. Co-edges thus furnish a basis for 1-cycles, in the sense that, given a co-edge e, there is a unique way to assign an integer $\mathbf{c}_\varepsilon$ to each edge ε of the tree in order to get a closed 1-chain: $\partial(e + \sum_{\varepsilon \in \mathcal{E}^T} \mathbf{c}_\varepsilon\, \varepsilon) = 0$, where $\mathcal{E}^T$ denotes the set of tree edges. In less formal language, one says that each co-edge "closes a circuit" in conjunction with edges of the tree.

In the general case $\Delta \neq \emptyset$, we have only $\partial(e + \sum_{\varepsilon \in \mathcal{E}^T} \mathbf{c}_\varepsilon\, \varepsilon) \in W_0(\Delta)$, still with uniqueness of the $\mathbf{c}_\varepsilon$s. In words, the co-edge e closes a circuit, in conjunction with edges of the tree, *if passage through* Δ *is allowed*. (The part of the circuit within Δ is *not* uniquely determined.)

Figure 5.8, which shows a spanning tree in a two-dimensional mesh, relative to a part of it, should help understand all this. Three kinds of co-edges are shown, each with its associated circuit. For co-edges like a, the circuit doesn't pass through Δ, contrary to what happens for co-edges of the same type as b. Co-edges like c are special in that the cycles they generate do *not* bound, which reveals the existence of a loop in the meshed region. (All of this is valid in three dimensions, too.)

These notions, here explained for $p = 1$, have obvious counterparts for all simplex dimensions. For $p = 2$, a tree would be a maximal set of faces that doesn't generate closed surfaces (2-cycles). Again, any extra face

[21] And hence, spanning the same range as the original matrix. We'll return to the graph-theoretical origin of the expression "spanning tree" in a moment.

would generate one, and this surface may not bound, owing to the presence of a hole in D.

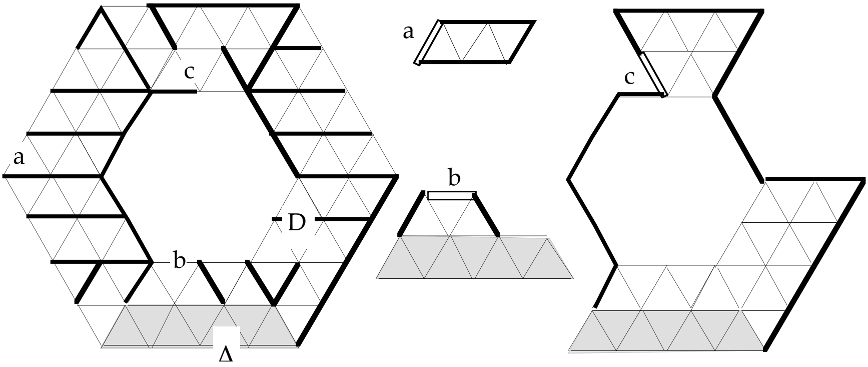


FIGURE 5.8. Notions of relative tree and co-edge. The tree on the left, with thick edges, is relative to the shaded region, Δ. On the right, closed chains mod Δ generated by three typical co-edges a, b, c.

How to *use* trees and cotrees will be explained by way of examples in Chapter 8 and Appendix C, but a few general indications can be given at this stage.

Suppose D contractible, and let $\mathcal{E}^T$ be a spanning tree of edges. For each co-edge e, let us build a DoF-vector $\mathbf{a}^e$ by setting $\mathbf{a}^{\mathbf{e}}_{\varepsilon} = \mathbf{c}_{\boldsymbol{\varepsilon}}$ (cf. p. 153) for all edges $\varepsilon \neq e$ in its circuit, and $\mathbf{a}^e_e = 1$. These (independent) **vector**s form a basis for $\ker(\mathbf{G}^t)$. We know already what $\mathbf{G}^t\mathbf{a}^{\mathbf{e}} = 0$ implies about the corresponding vector fields $a^e \in W^1$: *m*-weak solenoidality. So this is a kind of "discrete Coulomb gauge" imposed on the vector fields $\{a^e : e \in \mathcal{E} - \mathcal{E}^T\}$, the curls of which will span $\ker(\text{div} ; W^2)$. On the other hand, thanks to the general algebraic relation $\mathbf{W}^1 = \text{cod}(\mathbf{G}) \oplus \ker(\mathbf{G}^t)$ (cf. Appendix B), the DoF-vectors $\mathbf{h}^e$, with $e \in \mathcal{E}^T$, such that $\mathbf{h}^e_{\varepsilon} = 0$ for all edges $\varepsilon \neq e$ and $\mathbf{h}^e_e = 1$ form a basis for $\mathbf{GW}^0$, and hence the corresponding vector fields form a basis for grad W^0. Spanning trees of edges thus resolve the two problems mentioned at the beginning of this section.

In the general case, however, this will not work satisfactorily: We may not get enough $\mathbf{h}^e$s to span ker(rot), if there exist curl-free fields which are not gradients, and too many $\mathbf{a}^e$s, for there can exist nonzero fields in W^1 which are simultaneously curl-free and *m*-weakly solenoidal. Providing a solution to this problem in its full generality exceeds the scope of this book, but what to do is intuitively obvious. Look at Fig. 5.8. To obtain all curl-free fields, one should add to the tree *one* (because there is one loop in this case) of the co-edges of the same class as c, the circuit of which doesn't bound. The augmented tree we get this way can be called[22] a "belted tree", the "belt" being this non-bounding circuit,[23] and the loop co-edge acting as the belt "fastener". Note that, thanks to this added edge, the circuits of all remaining co-edges do bound. (The circuits of other co-edges homologous to c pass by the belt fastener.)

5.3.3 Trees and graphs

If $p = 1$, and if Δ is empty, we may consider nodes and edges as forming a graph; a spanning tree then appears to be a maximal subgraph "without loops". (As one easily sees, maximality implies that such a tree must "visit" all nodes, hence "spanning".) But the distinction between closed chains and boundaries is lost in the graph-theoretic context, so there is no straightforward way to build a belted tree via graph-oriented algorithms, whereas this problem is easily solved in algebraic terms. In the case of edges, for example, the belted tree corresponds to a basis of $\mathrm{cod}(\mathbf{R}^t)$, and the tree to a basis of $\ker(\mathbf{G}^t)$, which both are found by the same kind of algebraic manipulations (extract a matrix of maximal rank).

The other case where graphs are relevant is when $p = d - 1$, where d is the dimension of space. For instance, if $d = 3$, a spanning tree of faces, in the sense of Def. 5.2, can be described as a spanning tree of the graph the nodes of which are tetrahedra (beware!), and the arrows, the faces of the mesh. This is easy to understand, by duality, for this graph is nothing else than the standard nodes-to-edges graph of the *dual* mesh, the incidence matrix of which is $\mathbf{D}^t$. Otherwise, the case $1 < p < d - 1$ is not explainable in terms of graphs.[24]

However, $d = 2$ in many applications, which partly explains why the irrelevance of graph theory may have been overlooked. (The esthetic appeal of graphs also probably played a role.) In dimension 2, some problems about belted trees can even be solved in terms of graphs.

Figure 5.9 offers an example. Start from a spanning tree (in the graph-theoretic sense) on the surface of a torus. Form the *dual* subgraph, by

[22] Of course this "belted tree" is no longer a tree in the strict sense, so this is doubtful teminology. But we face a dilemma here. The right concepts are those of homology, not of graph theory, but the vocabulary of the latter has already prevailed, and it's too late to go against the grain. The oxymoron "belted tree" is a compromise, trying at once to refer to the familiar concept of tree and to mark its inadequacy.

[23] The received name for a belt is *homology cycle* of dimension 1.

[24] Even if $p = 1$ or d, the (indispensable) notion of *relative* tree is quite awkward in a graph-theoretic framework.

joining all centers between adjacent triangles which are not separated by an edge of the primal graph. The dual subgraph is not a tree, only a "bounding-circuit free" maximal subgraph, that is, a belted tree, for it contains two circuits that do not bound, or belts. Now, the two co-edges of the "primal" spanning tree which are crossed by the dual belt-fasteners are special in that the circuits they close do not bound on the torus (they are representative of the two homology classes of cycles, i.e., classes of cycles that don't bound, cf. Notes 20 and 23). So if we add them to the primal tree, as belt fasteners, we do have a belted tree.

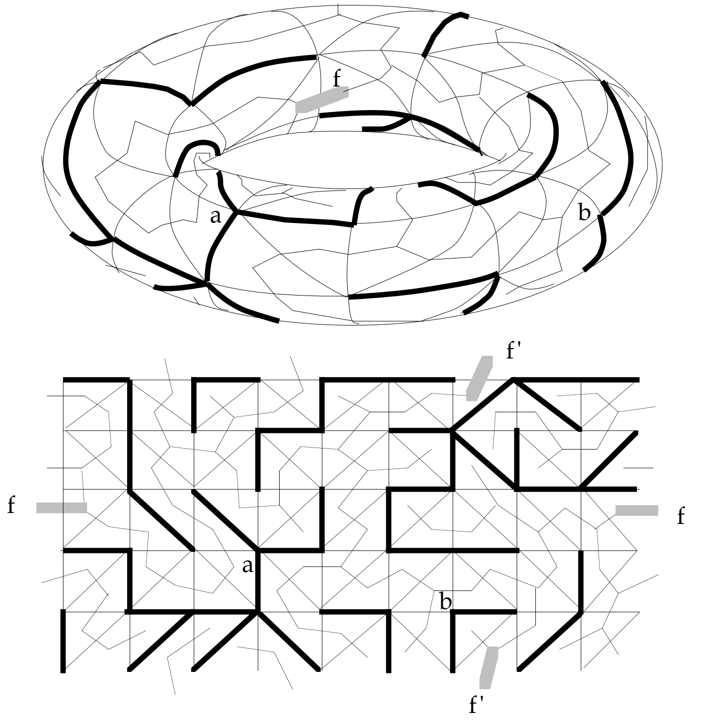


FIGURE 5.9. Spanning tree on the surface of a torus (in thick lines) and its dual, which is no more a genuine tree but a "belted tree" (co-edges not drawn). Points a and b should help make the correspondence between the spatial view and the plane diagram (which is an unfolding of the torus surface, after suitable cutting ; nodes on opposite sides should be identified). The two "belt fasteners" f and f ' are drawn in thick lines (f ' can't be seen in the top view).

Techniques of this kind are useful for problems of eddy-currents on thin conductive sheets [B2, T&]. But the nice illustrations by graphs should not hide their essentially algebraic nature: Tree and cotree methods really belong to *homology*.

EXERCISES

Exercises 5.1 and 5.2 are on p. 127, Exer. 5.3 on p. 131. Exercises 5.4 to 5.6 are on p. 141, Exer. 5.7 p. 143, Exer. 5.8 p. 148, and Exer. 5.9 p. 150.

Exercise 5.10. Compute all the terms of $\mathbf{M}_p$, as defined in (16), when $\alpha = 1$, for all p.

Exercise 5.11. Inquire about the "Poincaré inequality" (and preferably, devise your own proof): *If* D *is a bounded domain of* E_d, *there exists a constant* c(D) *such that*

$$\int_D |\varphi|^2 \leq c(D) \int_D |\text{grad}\, \varphi|^2$$

for all functions $\varphi \in C_0^\infty(D)$.

Exercise 5.12. In the previous exercise, the point of having φ vanish on the boundary is to provide a "reference value" for φ, to which one might otherwise add any constant (and hence, give an arbitrary large norm) without changing the gradient. This reference value may as well be the average of φ over the domain, that is, $\overline{\varphi} = (\int_D \varphi)/\text{vol}(D)$. So prove the existence of a constant c(D) such that

$$(\int_D |\varphi - \overline{\varphi}|^2)^{1/2} \leq c(D)\, (\int_D |\text{grad}\, \varphi|^2)^{1/2}$$

for all functions $\varphi \in C^\infty(\overline{D})$. (This is the "Poincaré–Friedrichs (or Poincaré–Wirtinger) inequality".)

Exercise 5.13. Show that, for a smooth field $a = \{a^1, a^2, a^3\}$,

$$\text{rot rot } a = \text{grad div } a - \Delta a,$$

where $\Delta a = \{\Delta a^1, \Delta a^2, \Delta a^3\}$. Use this to prove that, if a has bounded support,

$$(18) \qquad \int (\text{div } a)^2 + \int |\text{rot } a|^2 = \sum_{i=1,2,3} \int |\text{grad } a^i|^2.$$

where integrals are over all space.

HINTS

5.2. In dimension $d = 1$, for $D =]-1, 1[$, the function $x \to 1 - |x|$. For $d > 1$ and $D = \{x : |x| < 1\}$, aim at a function of $|x|$ with a singularity at 0, and not too fast growth there. Case $d = 2$ will appear special.

5.3. Of course the kernels are closed in the stronger norm, as pre-images of the closed set $\{0\}$, but one cannot employ this argument about the $\mathbb{L}^2$ norm, in which rot and div are not continuous, only closed. Use (2), and its analogue for rot.

5.4. See the cotangent formula of 3.3.4 and Lemma 4.1. The latter is especially useful (if applied with a measure of creative laziness).

5.5. In terms of the nodes-to-edges incidence matrix elements, one has

$$w_e = \mathbf{G}_{m\,e}\, w_m \nabla w_n + \mathbf{G}_{n\,e}\, w_n \nabla w_m.$$

Develop, and use Exer. 3.10. Set $g^{ij} = \nabla w_i \cdot \nabla w_j$. (The analogy with the metric coefficients g_{ij} of Riemannian geometry is not accidental.)

5.6. First show that $x \cdot \nabla w_n - w_n(x)$ is a constant inside each tetrahedron (Lemma 4.1). Then develop $(\nabla w_m \times \nabla w_n) \times x$.

5.8. Look at Fig. 4.3 and express the numbers N', E', F' relative to the refined mesh in terms of N, E, F.

5.9. Use Proposition 5.5, second part first, then first part.

5.11. Begin with $d = 1$. Then $D =]a, b[$, and $\varphi(x) = \int_a^x \partial\varphi(\xi)\, d\xi$. Use Cauchy-Schwarz, then sum with respect to x. For $d > 1$, note that, in the arrowed notation where "$X \to Y$" means "all functions from X to Y", the functional space $\mathbb{R} \times \ldots$ [d times] $\ldots \times \mathbb{R} \to \mathbb{R}$ can be identified with $\mathbb{R} \to (\mathbb{R} \times \ldots$ [d − 1 times] $\ldots \times \mathbb{R} \to \mathbb{R})$.

5.12. In dimension 1 first, $\varphi(y) - \varphi(x) = \int_x^y \partial\varphi(\xi)\, d\xi$, hence $\varphi(x) - \varphi(y) \leq C\, \|\partial\varphi\|$, for all pairs of points $\{x, y\}$ in $[a, b]$. Integrate with respect to y to get $\varphi(x) - \overline{\varphi} \leq C\, \|\partial\varphi\|$, then invoke Cauchy-Schwarz. Adapt this to d dimensions as in the previous case.

5.13. In Cartesian coordinates, $(\text{rot rot } a)^i = \sum_j \partial_j(\partial_i a^j - \partial_j a^i)$. For (18), integrate by parts.

SOLUTIONS

5.1. If $\{0, u\}$ is in the closure of GRAD, i.e., is the limit of some sequence $\{\psi_n, \text{grad } \psi_n\}$ the terms of which belong to GRAD, then $\int_D \psi_n \text{ div } j' = -\int_D \text{grad } \psi_n \cdot j'$ for all j' in $\mathbb{C}_0^\infty(D)$, hence $\int_D u \cdot j' = 0 \;\; \forall\, j' \in \mathbb{C}_0^\infty(D)$ at the limit, and hence $u = 0$. If $\{0, u\}$ is in the closure of ROT, i.e., the limit of some $\{a_n, \text{rot } a_n\}$ of ROT, then $\int_D a_n \cdot \text{rot } h' = -\int_D \text{rot } a_n \cdot h'$ for all h' in $\mathbb{C}_0^\infty(D)$, hence $\int_D u \cdot h' = 0 \;\; \forall\, h' \in \mathbb{C}_0^\infty(D)$ at the limit, and $u = 0$.

5.2. On $D = \{x : |x| < 1\}$, functions of the form $x \to |x|^{-\alpha}$ foot the bill, if $\alpha > 0$ (in order to have a singularity at 0), $\int_0^1 r^{-2\alpha}\, dr < \infty$ (for the function to be square-summable) and $\int_0^1 r^{d - 1 - 2(1 + \alpha)}\, dr < \infty$ (for its gradient to be square-summable). This happens for $0 < \alpha < 1/2$ and $1 + \alpha < d/2$, the latter constraint being redundant if $d > 2$. For $d = 2$, look at the function $x \to |x| \log|x|$.

5.3. After (2), "$\int_D b \cdot \operatorname{grad} \varphi' = 0 \;\; \forall\, \varphi' \in C_0^\infty(D)$" characterizes elements of ker(div), and if a sequence of fields $\{b_n\}$ which all satisfy this predicate converges to some b in $\mathbb{L}^2(D)$, this also holds for b, by continuity of the scalar product. Same argument for fields b such that $n \cdot b = 0$, since they are characterized by $\int_D \operatorname{div} b\ \varphi' + \int_D b \cdot \operatorname{grad} \varphi' = 0 \;\; \forall\, \varphi' \in L^2_{\text{grad}}(D)$, and for fields such that $n \times h = 0$, by using the similar formula in rot.

5.4. Let h be the height of node n above the plane of f (observe how "above" makes sense if space is oriented, as well as f). Then $\mathrm{vol}(\textsc{t}) = h\ \mathrm{area}(\{k, \ell, m\})/3 = \mathrm{area}(\{k, \ell, m\})/(3|\nabla w_n|)$. There are many ways to derive these relations, but the most illuminating is to remark that 3×3 matrices such as $(\nabla w_k, \nabla w_\ell, \nabla w_m)$ and $(nk, n\ell, nm)$ are inverses, by Lemma 4.1, and that $\mathrm{vol}(\textsc{t}) = \det(nk, nl, nm)/6$, that $\mathrm{area}(\{k, \ell, m\}) = \det(k\ell, km)/2$, etc.

5.5. Up to obvious sign changes, there are only three cases:

(a) $e = e' = \{m, n\}$: $(12\, \mathrm{vol}(\textsc{t})/5!\,)\, (g^{mn} + g^{mm} - g^{nm}\},$

(b) $e = \{m, n\},\ e' = \{m, \ell\}$: $(6\, \mathrm{vol}(\textsc{t})/5!\,)(2\, g^{n\ell} - g^{m\ell} - g^{nm} + g^{mm}),$

(c) $e = \{k, \ell\},\ e' = \{m, n\}$: $(6\, \mathrm{vol}(\textsc{t})/5!\,)(g^{\ell n} - g^{kn} - g^{\ell m} + g^{km}).$

5.6. $(\nabla w_m \times \nabla w_n) \times x = x \cdot \nabla w_m\, \nabla w_n - x \cdot \nabla w_n\, \nabla w_m = w_m\, \nabla w_n - w_n\, \nabla w_m + b,$ where b is some vector, hence $w_{\{m,\, n\}} = (\nabla w_m \times \nabla w_n) \times x + b$. Now, observe (cf. 3.3.4 and Lemma 4.1) that

$$\nabla w_m \times \nabla w_n = (kn \times k\ell) \times \nabla w_n / 6\, \mathrm{vol}(\textsc{t}) = k\ell / 6\, \mathrm{vol}(\textsc{t}).$$

Using this, one has the following alternative form for the face element:

$$w_f = 2(w_\ell\, k\ell + w_m\, km + w_n\, kn)/6\, \mathrm{vol}(\textsc{t}),$$

hence the desired result (place the origin at node k).

5.7. If $e = \{m, n\}$ and $R_{f\,e} \neq 0$, then $f = \{\ell, m, n\}$ or $\{\ell, n, m\}$, for some ℓ. In both cases,

$$R_{f\,e}\, w_f = w_f = 2(w_\ell\, \nabla w_m \times \nabla w_n + w_m\, \nabla w_n \times \nabla w_\ell + w_n\, \nabla w_\ell \times \nabla w_m).$$

Therefore, summing over all faces,

$$\sum_{f \in \mathcal{F}} R_{f\,e}\, w_f = 2 \sum_{\ell \in \mathcal{N}} (w_\ell \,\nabla w_m \times \nabla w_n + \dots) = 2\, \nabla w_m \times \nabla w_n.$$

For the divergence, just notice that $\operatorname{div} w_f = 2(\nabla w_\ell \cdot \nabla w_m \times \nabla w_n + \dots) = 6 \det(\nabla w_\ell\,, \nabla w_m\,, \nabla w_n) = w_T$ if $T = \{k, \ell, m, n\}$. Two compensating changes of sign occur if $T = \{\ell, k, m, n\}$, the other orientation.

5.8. $F' = 4F$, $E' = 2E + 3F$, $N' = N + E$, hence $N' - E' + F' = N + E - (2E + 3F) + 4F \equiv N - E + F$.

5.9. Since $\ker(\operatorname{div}\,;\ W^2) = \operatorname{rot} W^1$, its dimension is the dimension of W^1, which is E, minus the dimension of $\ker(\operatorname{rot};\ W^1) \equiv \operatorname{grad} W^0$. The latter is the dimension of W^0, i.e., N, minus the dimension of $\ker(\operatorname{grad};\ W^0)$, which is 1. Project: Practice with this in the general case, to see how the Betti numbers come to slightly modify these dimensions (but not their asymptotic behavior when the mesh is refined).

5.11. Since D is bounded, it is contained in a set of the form $P =]a, b[\times \mathbb{R} \times \dots \times \mathbb{R}$. Extending by 0, outside D, the functions of $C_0^\infty(D)$, one identifies the latter space with $C_0^\infty(P)$, which is isomorphic to $C_0^\infty([a, b]\,;\, C_0^\infty(\mathbb{R}^{d-1}))$. Then, if $x = \{x^1, \dots, x^d\} \in P$, one has

$$\varphi(x) = \int_a^{x^1} \partial_1\varphi(\xi, x^2, \dots, x^d)\, d\xi,$$

where $\partial_1\varphi$ is the partial derivative with respect to the variable x^1. By the Cauchy-Schwarz inequality,

$$|\varphi(x)|^2 \leq (x - a)^{1/2} \int_a^{x^1} |\partial_1\varphi(\xi, \dots)|^2\, d\xi \leq (b - a)^{1/2} \int_a^b |\nabla\varphi(\xi, \dots)|^2\, d\xi,$$

and hence, by Fubini,

$$\int_P |\varphi(x)|^2\, dx \leq (b - a)^{1/2} \int_P dx^1 \dots dx^d \int_a^b |\nabla\varphi(\xi, \dots)|^2\, d\xi$$
$$= (b - a)^{1/2} \int_a^b dx^1 \int_P |\nabla\varphi(\xi, \dots)|^2\, d\xi\, dx^2 \dots dx^d,$$

hence $c(D) \leq (b - a)^{3/2}$. Of course, this is an upper bound, not the "best" value of $c(D)$, which can be obtained, but by very different methods.

5.13. First,

$$(\operatorname{rot}\operatorname{rot} a - \operatorname{grad}\operatorname{div} a)^i = \sum_j [\partial_j(\partial_i a^j - \partial_j a^i) - \partial_i(\partial_j a^j)]$$
$$= \sum_j [\partial_j(\partial_i a^j - \partial_j a^i) - \partial_j(\partial_i a^j)] = -\sum_j \partial_{j\,j}\, a^i.$$

Then $\int (\operatorname{div} a)^2 + \int |\operatorname{rot} a|^2 = \sum_i \int -\Delta a^i\, a^i = \sum_i \int |\operatorname{grad} a^i|^2$. (Further study: What of a domain D with surface S? Try to cast the surface terms that then appear in coordinate-free form, by using adequate curvature operators.)

REFERENCES and Bibliographical Comments

Whitney elements were rediscovered by numerical analysts beginning in 1975 in relation to the search for "mixed" finite elements. In particular, the edge element (9) appeared in [Nd], where it was described in terms of its "shape functions" $x \to a(T) \times x + b(T)$ in the tetrahedron T (cf. Exer. 5.6), where $a(T)$ and $b(T)$ are three-dimensional vectors. (There are thus six degrees of freedom per tetrahedron, in linear invertible correspondence with the six edge circulations.) Similarly [Nd], the face element's shape functions (the 2D version of which first appeared in [RT]) are $x \to \alpha(T)x + b(T)$, with $\alpha(T) \in \mathbb{R}$ (Exer. 5.6). The obvious advantage of representations such as (9) or (10) over shape functions is to explicitly provide a *basis* for W^1 or W^2. The presentation by shape functions, however, seems preferable for vector-valued elements of polynomial degree higher than one (cf. [Ne]), whose description in Whitney's style, with basis functions clearly related to geometrical elements of the simplicial complex, is still an open problem, as is, for that matter, the classification of "tangentially[25] continuous" vectorial elements proposed up to now by various authors [Cr, MH, vW, WC, . . .]. See [YT] for a recent attack on the problem. Face elements also were independently rediscovered, in relation with diffraction modelling [SW]. (The latter authors have the face flux *density* equal to 1, instead of the total flux.)

[AR] L. Albanese, G. Rubinacci: "Magnetostatic field computations in terms of two-component vector potentials", **Int. J. Numer. Meth. Engnrg., 29** (1990), pp. 515–532.

[Ba] I. Babuška: "Error bounds in the finite element method", **Numer. Math., 16** (1971), pp. 322–333.

[B1] A. Bossavit: "A new rationale for edge-elements", **Int. Compumag Society Newsletter, 1,** 3 (1995), pp. 3–6.

[B2] A. Bossavit: "Eddy currents on thin shells", in **8th Int. Workshop on Electric and Magnetic Fields** (Proceedings, Liège, 6-9 May 1996), A.I.M. (31 Rue St-Gilles, Liège), 1996, pp. 453–458.

[Br] F. Brezzi: "On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers", **RAIRO, Anal. Num., 8**, R2 (1974), pp. 129–151.

[Bu] W.L. Burke: **Applied Differential Geometry**, Cambridge University Press (Cambridge, UK), 1985.

[Cr] C.W. Crowley, P.P. Silvester, H. Hurwitz: "Covariant projection elements for 3D vector field problems", **IEEE Trans., MAG-24**, 1 (1988), pp. 397–400.

[25] I don't mean to condone this dubious terminology, for it's "o*nly* tangentially continuous" that one is supposed to understand: $\mathbb{P}^1$ elements *are* "tangentially continuous", and don't qualify, precisely because they are *also* "normally continuous". Anyway, this misses the point: The point is the structural property $\text{grad}\, W^0 = \ker(\text{rot}\,;\, W^1)$, for simply- connected domains. This is the rule of the game, for whoever wants to propose new elements. Substitutes for the simple but too coarse tetrahedral first-degree edge-elements should be looked for, but *in conjunction with companion scalar-valued nodal elements,* in order to satisfy this structural rule. The discussion of spurious modes at the end of Chapter 9 should make that clear. (A more general and more complex compatibility condition between elements, the "Ladyzhenskaya-Babuška-Brezzi (LBB) condition", or "inf–sup condition", invented by mixed elements researchers [Ba, Br], happens to be implied by the above structural rule.)

[Do] J. Dodziuk: "Finite-Difference Approach to the Hodge Theory of Harmonic Forms", **Amer. J. Math., 98,** 1 (1976), pp. 79–104.

[Fu] K. Fujiwara: "3-D magnetic field computation using edge elements", in **Proc. 5th Int. IGTE Symposium,** IGTE (26 Kopernikusgasse, Graz, Austria), 1992, pp. 185–212.

[GT] N.A. Golias, T.D. Tsiboukis: "Magnetostatics with Edge Elements: A Numerical Investigation in the Choice of the Tree", **IEEE Trans., MAG-30,** 5 (1994), pp. 2877–2880.

[GH] M.J. Greenberg, J.R. Harper: **Algebraic Topology, A First Course**, Benjamin/Cummings (Reading, MA), 1981.

[Ha] H.W. Hale: "A Logic for Identifying the Trees of a Graph", **AIEE Trans.** (June 1961), pp. 195–198.

[HW] P.J. Hilton, S. Wylie: **Homology Theory, An Introduction to Algebraic Topology,** Cambridge University Press (Cambridge), 1965.

[Ke] L. Kettunen, K. Forsman, D. Levine, W. Gropp: "Volume integral equations in nonlinear 3-D magnetostatics", **Int. J. Numer. Meth. Engng., 38** (1995), pp. 2655–2675.

[LM] J.L. Lions, E. Magenes: **Problèmes aux limites non homogènes et applications**, Vols. 1–2, Dunod (Paris), 1968.

[MH] G. Mur, A.T. de Hoop: "A Finite-Element Method for Computing Three-Dimensional Electromagnetic Fields in Inhomogeneous Media", **IEEE Trans., MAG-19**, 6 (1985), pp. 2188–2191.

[Nd] J.C. Nedelec: "Mixed finite elements in $\mathbb{R}^3$", **Numer. Math., 35** (1980), pp. 315–341.

[Ne] J.C. Nedelec: "A new family of mixed finite elements in $\mathbb{R}^3$", **Numer. Math., 50** (1986), pp. 57–81.

[PR] K. Preis, I. Bardi, O. Biro, C. Magele, G. Vrisk, K.R. Richter: "Different Finite Element Formulations of 3D Magnetostatic Fields", **IEEE Trans., MAG-28,** 2 (1992), pp. 1056–1059.

[RT] P.A. Raviart, J.M. Thomas: "A mixed finite element method for second order elliptic problems", in **Mathematical Aspects of Finite Element Methods** (A. Dold, B. Eckmann, eds.), Springer-Verlag (New York), 1977.

[RR] Z. Ren, A. Razek: "Boundary Edge Elements and Spanning Tree Technique in Three-Dimensional Electromagnetic Field Computation", **Int. J. Numer. Meth. Engng., 36** (1993), pp. 2877–2893.

[SW] D.H. Schaubert, D.R. Wilton, W. Glisson: "A Tetrahedral Modeling Method for Electromagnetic Scattering by Arbitrarily Shaped Inhomogeneous Dielectric Bodies", **IEEE Trans., AP-32**, 1 (1984), pp. 77–85.

[Sc] J.A. Schouten: **Tensor analysis for physicists**, Dover (New York), 1989. (First edition, Clarendon, Oxford, 1951.)

[T&] H. Tsuboi, T. Asahara, F. Kobayashi, T. Misaki: "Eddy Current Analysis on Thin Conducting Plate by an Integral Equation Method Using Edge Elements", **IEEE Trans., MAG-33**, 2 (1997), pp. 1346–1349.

[VW] O. Veblen, J.H.C. Whitehead, **The Foundations of Differential Geometry**, Cambridge, 1932.

[vW] J.S. Van Welij: "Calculation of Eddy Currents in Terms of H on Hexahedra", **IEEE Trans., MAG-21**, 6 (1985), pp. 2239–2241.

[Wh] H. Whitney: **Geometric Integration Theory,** Princeton U.P. (Princeton), 1957.

[WC] S.H. Wong, Z.J. Cendes: "Combined Finite Element-Modal Solution of Three-Dimensional Eddy-Current Problems", **IEEE Trans., MAG-24**, 6 (1988), pp. 2685–2687.

[YT] T.V. Yioultsis, T.D. Tsiboukis: "The Mystery and Magic of Whitney Elements—An Insight in their Properties and Construction", **Int. Compumag Society Newsletter, 3,** 3 (1996), pp. 6–13.

CHAPTER 6

The "curl side": Complementarity

We return to the model problem of Chapter 2, the magnetostatics version of the "Bath cube" setup. This time, the two mirror symmetries with respect to vertical planes are taken into account (Fig. 6.1): The field b is obviously invariant with respect to both reflections, hence $n \cdot b = 0$ on these planes.[1]

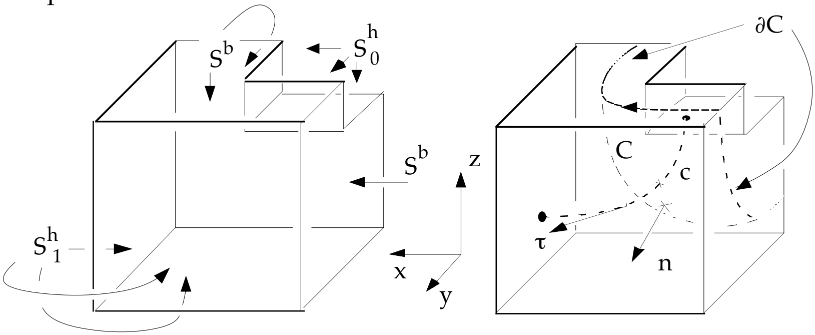


FIGURE 6.1. Notations for solving the Bath-cube magnetostatics problem in a quarter of the cavity. Symmetry planes bear the boundary condition $n \cdot b = 0$, hence contributing to the S^b boundary. The "link" c must go from the upper pole (here, part $S^h{}_0$ of the boundary S^h) to the lower pole (part $S^h{}_1$ of S^h), whereas the "cut" C should separate the two poles, while having its own boundary ∂C inside S^b. (In the jargon of Fig. 4.6, c and C are "closed mod S^h [resp. mod S^b] and non-bounding".)

Denoting by D the domain thus defined, and by S its boundary, with S^h the "magnetic boundary" and S^b the part of S at the top of the cavity (cf. Fig. 2.6) and in symmetry planes, we must have:

[1] We could reduce further to an eighth, by taking into account the invariance with respect to a 90° rotation around the z-axis. But the symmetry group thus obtained is not Abelian, a feature which considerably complicates the exploitation of symmetry. Cf. Refs. [B1, B2] of Appendix A.

(1) $\mathrm{rot}\, h = 0$ in D, (3) $\mathrm{div}\, b = 0$ in D,

(2) $n \times h = 0$ on S^h (4) $n \cdot b = 0$ on S^b,

(5) $b = \mu\, h$ in D.

The problem is made well-posed by imposing one of the two conditions

(6) $\int_c \tau \cdot h = I$, (7) $\int_C n \cdot b = F$,

as we saw[2] in 2.4.1. These equations are the same as (2.20)–(2.26), and little changed in the former model due to geometrical symmetry, except for one thing: The flux F in (7) is now the flux through one-quarter of the device, and the computed reluctance will be relative to a quarter as well.

The layout of Eqs. (1–7) underlines a symmetry of another kind, which will be our main concern in this chapter: the symmetry of the magnetostatic equations *with respect to the* b–h *interchange*. This can be made even more patent by setting the problem as follows: We look for *pairs* {F, I} for which problem (1–7) has a solution. By linearity, they all lie on the characteristic line I = RF in the F–I plane, and the problem thus consists in finding R. (This remark, though of moderate interest in the present linear case, is the key to nonlinear generalization [B3].)

6.1 A SYMMETRICAL VARIATIONAL FORMULATION

We shall strive to preserve this symmetry in the search for a variational formulation. The functional point of view continues to prevail: We look for the solution as (first step) an element of some predefined functional space that (second step) can be characterized as the minimizer of some easily interpretable, energy-related quantity.

6.1.1 Spaces of admissible fields

By "the" solution, now, we mean the *pair* {h, b}. What are the eligible fields, a priori? Both h and b will certainly belong to $\mathbb{L}^2(D)$, since magnetic energy is finite. Moreover, rot h = 0 and div b = 0, and if we had to generalize what we do to cases where a given current density j

[2] The relative arbitrariness in the choice of c and C is reminded (cf. Exers. 2.5, 2.6 and Fig. 4.6). In precise language, integrals (6) and (7) depend on the *homology classes* of c and C, mod S^h and mod S^b, respectively.

exists in the cavity, rot h ≡ j would be square-integrable, since Joule dissipation must remain finite. This points to $\mathbb{L}^2_{\text{rot}}(D)$ as the space in which to look for h. Symmetrically, b will belong to $\mathbb{L}^2_{\text{div}}(D)$.

We can do better, by anticipating a little on the discretization process yet to come. Some a priori constraints on the solution are easy to enforce at the discretized level (those are the "essential" or "Dirichlet-like" conditions mentioned in 2.4.4), and it pays to take them into account from the onset in the definition of admissible fields, since this reduces the scope of the search. Boundary conditions (2) and (4) are of this kind. So let us define (recall we are now using the *weak* grad, rot, div)

$$\mathbb{H} = \{h \in \mathbb{L}^2_{\text{rot}}(D) : \ \text{rot}\, h = 0, \ n \times h = 0 \ \text{ on } S^h\},$$

$$\mathbb{B} = \{b \in \mathbb{L}^2_{\text{div}}(D) : \ \text{div}\, b = 0, \ n \cdot b = 0 \ \text{ on } S^b\}.$$

These are closed subspaces of $\mathbb{L}^2(D)$, after Exer. 5.3. We note that fields of the form h = grad φ belong to $\mathbb{H}$ if φ is a potential which assumes constant values on both parts of the magnetic boundary S^h (not necessarily the same constant on each), and we recycle the symbol Φ to denote the space of such potentials:

$$\Phi = \{\varphi \in L^2_{\text{grad}}(D) : \ n \times \text{grad}\, \varphi = 0 \ \text{ on } S^h\}.$$

(It's not exactly the same as the earlier Φ, beware: φ is a constant on S^h_0, but not necessarily the constant 0.) Similarly, on the side of b, let's introduce

$$A = \{a \in \mathbb{L}^2_{\text{rot}}(D) : \ n \cdot \text{rot}\, a = 0 \ \text{ on } S^b\},$$

and remark that fields of the form b = rot a, with a ∈ A, belong to $\mathbb{B}$. Moreover, if D is contractible (no loops, no holes[3]) then, by the Poincaré lemma (understood in its extended version of 5.1.4),

$$(8) \qquad \mathbb{H} = \text{grad}\, \Phi, \qquad\qquad \mathbb{B} = \text{rot}\, A,$$

instead of mere inclusions.

Conditions (6) and (7) also can be enforced a priori. Let's define linear functionals $\mathcal{J} : \mathbb{H} \to \mathbb{R}$ and $\mathcal{F} : \mathbb{B} \to \mathbb{R}$ as follows. First, if h and b are smooth, set

$$\mathcal{J}(h) = \int_c \tau \cdot h, \qquad\qquad \mathcal{F}(b) = \int_C n \cdot b,$$

[3] The reader who suspects that (8) may hold in spite of the existence of loops or holes in D, for more complex geometries, is right. Only *relative* loops and holes (mod S^h and S^b) are harmful.

Then, let us prove the following:

Proposition 6.1. *$\mathcal{J}$ and $\mathcal{F}$ have extensions,* continuous *with respect to the metric of* $\mathbb{L}^2(D)$, *to* $\mathbb{H}$ *and* $\mathbb{B}$.

Proof. Let φ^1 be a smooth function assuming the values 0 on $S^h{}_0$ and 1 on $S^h{}_1$. For $b \in \mathbb{B}$, then, $\int_D b \cdot \operatorname{grad} \varphi^1 = \int_S n \cdot b \ \ \varphi^1 = \int_{S^h_1} n \cdot b = \mathcal{F}(b)$, as we saw with Exer. 2.6, and the map $b \rightarrow \int_D b \cdot \operatorname{grad} \varphi^1$, which is $\mathbb{L}^2$-continuous, is thus the announced extension. The proof for $\mathcal{J}$ is a bit more involved. (Doing now Exercise 6.6 on p. 187, as preparation, may help.) Pick a smooth vector field a^1 such that $n \cdot \operatorname{rot} a^1 = 0$ on S^b and $\int_C n \cdot \operatorname{rot} a^1 = 1$. For $h \equiv \operatorname{grad} \varphi \in \mathbb{H}$, then, $\int_D h \cdot \operatorname{rot} a^1 = - \int_S n \times h \cdot a^1 = - \int_S n \times \operatorname{grad} \varphi \cdot a^1 = \int_S n \times a^1 \cdot \operatorname{grad} \varphi = - \int_S \operatorname{div}(n \times a^1) \ \varphi = - \int_S n \cdot \operatorname{rot} a^1 \ \varphi = - \int_{S^h} n \cdot \operatorname{rot} a^1 \ \varphi = \mathcal{J}(h) \int_C n \cdot \operatorname{rot} a^1 = \mathcal{J}(h)$. Again, the continuity of $h \rightarrow \int_D h \cdot \operatorname{rot} a^1$ proves the point. From now on, we let $\mathcal{J}$ and $\mathcal{F}$ denote the extended continuous functionals. ◊

Remark 6.1. Integrals such as $\mathcal{J}$ and $\mathcal{F}$ stand no chance of being $\mathbb{L}^2(D)$-continuous if one tries to enlarge their domains beyond $\mathbb{H}$ and $\mathbb{B}$. The conditions $\operatorname{rot} h = 0$ and $\operatorname{div} b = 0$ are necessary. ◊

As a corollary, subspaces

$$\mathbb{H}^I = \{h \in \mathbb{H} : \int_c \tau \cdot h = I\}, \qquad \mathbb{B}^F = \{b \in \mathbb{B} : \int_C n \cdot b = F\},$$

are closed. By (8), we have $\mathbb{H}^I = \operatorname{grad} \Phi^I$ and $\mathbb{B}^F = \operatorname{rot} A^F$, where $\Phi^I = \{\varphi \in \Phi : \int_c \tau \cdot \operatorname{grad} \varphi = I\}$ and $A^F = \{a \in A : \int_C n \cdot \operatorname{rot} a = F\}$, the pre-images of $\mathbb{H}^I$ and $\mathbb{B}^F$.

Note these are not vector subspaces, but *affine* subspaces of $\mathbb{H}$, $\mathbb{B}$, etc., unless $I = 0$ or $F = 0$. We consistently denote by $\mathbb{H}^0$, $\mathbb{B}^0$, Φ^0, A^0 the subspaces that would be obtained in the latter case. $\mathbb{H}^I$ is *parallel,* the usual way, to $\mathbb{H}^0$, and so forth. The following lemma will be important:

Lemma 6.1. *Spaces* $\mathbb{H}^0$ *and* $\mathbb{B}$ *are* orthogonal *in* $\mathbb{L}^2(D)$, *i.e.,* $\int_D h \cdot b = 0$ *if* $h \in \mathbb{H}^0$ *and* $b \in \mathbb{B}$. *Similarly,* $\mathbb{H}$ *and* $\mathbb{B}^0$ *are orthogonal.*

Proof. Let $h \in \mathbb{H}$ and $b \in \mathbb{B}$. Then $h = \nabla\varphi$. Let φ_0 and φ_1 be the values of φ on both parts of S^h. One has $\int_D h \cdot b = \int_D b \cdot \nabla\varphi = - \int_D \varphi \ \operatorname{div} b + \int_S \varphi \ n \cdot b = \int_{S^h} \varphi \ n \cdot b$ after (4), and this is equal to $(\varphi_1 - \varphi_0) \int_{S^h} n \cdot b$, that is, to the product $(\int_c \tau \cdot h)(\int_C n \cdot b)$. Now if $h \in \mathbb{H}^0$, or $b \in \mathbb{B}^0$, one of the factors vanishes. ◊

Remark 6.2. There is more, actually: With the simple topology we have here, both pairs are ortho-*complements* in $\mathbb{L}^2(D)$, which amounts to saying that any square-integrable field u can be written as $u = h + b$, with $h \in \mathbb{H}^0$ and $b \in \mathbb{B}$, or with $h \in \mathbb{H}$ and $b \in \mathbb{B}^0$, i.e., as the sum of a curl-free field and a solenoidal field. These are *Helmholtz decompositions.* We won't need a thorough treatment of them, but the paradigm is important, and will recur. ◊

The present state of affairs is summarized by Fig. 6.2, in which one may recognize a part of the front of Maxwell's building of Fig. 5.1. Note how all "vertical" relations have been taken care of, in advance, by the very choice of functional spaces. Only the "horizontal" condition (5), which expresses the constitutive laws of materials inside D, remains to be dealt with.

$$\begin{tikzcd}
\Phi^I \ni \varphi \arrow[d, "\mathrm{grad}"'] & & 0 \\
\mathbb{H}^I \ni h \arrow[d, "\mathrm{rot}"'] \arrow[r, dash] & b = \mu h \arrow[r, dash] & b \in \mathbb{B}^F \arrow[u, "\mathrm{div}"'] \\
0 & & a \in A^F \arrow[u, "\mathrm{rot}"']
\end{tikzcd}$$

FIGURE 6.2. Structure of the magnetostatics problem. (This is a "Tonti diagram" [To]. Similar graphic conventions have been independently proposed by many researchers, Roth in particular [Rt]. See [Bw] for some history.)

6.1.2 Variational characterization of the solution

The following result will make the horizontal connection.

Proposition 6.2. *The problem which consists in finding* $h \in \mathbb{H}^I$ *and* $b \in \mathbb{B}^F$ *such that*

$$(9) \qquad \int_D \mu^{-1} |b - \mu h|^2 \le \int_D \mu^{-1} |b' - \mu h'|^2 \quad \forall\, h' \in \mathbb{H}^I, \ \forall\, b' \in \mathbb{B}^F,$$

has a unique solution, which is the solution of (1–7) *when the latter exists.*

Proof. The proof of Lemma 6.1 shows that $\int_D h' \cdot b' = I\ F$. Therefore,

$$(10) \qquad \int_D \mu^{-1} |b' - \mu h'|^2 = \int_D \mu^{-1} |b'|^2 + \int_D \mu\, |h'|^2 - 2\ I\ F$$

for all pairs $\{h', b'\} \in \mathbb{H}^I \times \mathbb{B}^F$. This means that problem (9) splits in two *independent* minimization problems:

$$(11) \qquad \textit{find } h \in \mathbb{H}^I \textit{ such that } \int_D \mu\, |h|^2 \textit{ is minimum}$$

and

(12) $\quad$ *find* $b \in \mathbb{B}^F$ *such that* $\int_D \mu^{-1} |b|^2$ *is minimum.*

Both problems have a unique solution by the Hilbertian projection theorem (because $\mathbb{H}^I$ and $\mathbb{B}^F$ are closed convex sets). The minima are necessarily of the form $S\,I^2$ and $R\,F^2$, where R and S are positive constants. Then (10) shows that

$$\int_D \mu^{-1} |b - \mu h|^2 = R\,F^2 + S\,I^2 - 2\,I\,F \geq 0$$

whatever F and I. If (1–7) has a solution for a nontrivial pair {I, F}, the left-hand side vanishes, which implies $RS = 1$ (look at the discriminant), $I = RF$, and $b = \mu h$. We call R the *reluctance* of D under the prevailing boundary conditions. ◊

We note that, for a given nonzero F, there is always a value of I such that the left-hand side of (9) vanishes, namely $I = RF$, so this proves the existence in (1–7) for the right value of I/F. However, the point of Prop. 6.2 is not to prove again the existence of a solution to the magnetostatics problem, but to introduce a new variational characterization of it. The quantity at the right-hand side of (9) is an "error in constitutive law" as regards the pair {h', b'}. Thus, compliance with such a law amounts to looking for a couple of fields that minimize the discrepancy, among those that satisfy all other required conditions. This old and esthetically attractive idea, which generalizes to *monotone* nonlinear constitutive laws, thanks to the theory of convex functions in duality [Fe, Ro], seems to date back to Moreau [Mo, Ny], and has been increasingly popular ever since (and rediscovered), in the "computational magnetics" community [R§] and others [OR, LL]. It works for *time-dependent* problems just as well [B2, A&].

There are several equivalent ways to formulate problems (11) and (12). One consists of writing the associated Euler equations, or weak formulations:

(13) $\quad$ *find* $h \in \mathbb{H}^I$ *such that* $\int_D \mu\; h \cdot h' = 0 \quad \forall\; h' \in \mathbb{H}^0$,

(14) $\quad$ *find* $b \in \mathbb{B}^F$ *such that* $\int_D \mu^{-1} b \cdot b' = 0 \quad \forall\; b' \in \mathbb{B}^0$.

Another consists of using *potentials,* thanks to (8):

(15) $\quad$ *find* $\varphi \in \Phi^I$ *minimizing* $\int_D \mu\; |\mathrm{grad}\, \varphi|^2$,

(16) $\quad$ *find* $a \in A^F$ *minimizing* $\int_D \mu^{-1}\; |\mathrm{rot}\, a|^2$.

Now, of course, neither φ nor a need be unique. Equivalent weak formulations are

(17) *find* $\varphi \in \Phi^{\mathrm{I}}$ *such that* $\int_D \mu \operatorname{grad} \varphi \cdot \operatorname{grad} \varphi' = 0 \quad \forall\, \varphi' \in \Phi^0$,

(18) *find* $\mathrm{a} \in \mathrm{A}^{\mathrm{F}}$ *such that* $\int_D \mu^{-1} \operatorname{rot} \mathrm{a} \cdot \operatorname{rot} \mathrm{a}' = 0 \quad \forall\, \mathrm{a}' \in \mathrm{A}^0$.

Dualizing the constraints (6) and (7), as in Exer. 2.9, is also an option, which leads to

(11') *find* $\mathrm{h} \in \mathbb{H}$ *such that* $\int_D \mu\, |\mathrm{h}|^2 - 2\,\mathrm{F}\,\mathcal{J}(\mathrm{h})$ *is minimum,*

(12') *find* $\mathrm{b} \in \mathbb{B}$ *such that* $\int_D \mu^{-1}\, |\mathrm{b}|^2 - 2\,\mathrm{I}\,\mathcal{F}(\mathrm{b})$ *is minimum,*

with again associated Euler equations and equivalent formulations with potentials $\varphi \in \Phi$ and $\mathrm{a} \in \mathrm{A}$, similar to (15)–(18).

6.1.3 Complementarity, hypercircle

At this stage, we have recovered the variational formulation in φ, Eq. (15), and derived the other one, on the "curl side", in a strictly symmetrical way. This is an encouragement to proceed in a similarly parallel fashion at the discrete level.

So let Φ_m be the subspace of mesh-wise affine functions in Φ (recall they are constant over both parts of S^{h}). Likewise, let's have Φ^{I}_m and Φ^0_m as Galerkin subspaces for Φ^{I} and Φ^0 (with, again, the now-standard caveat about variational crimes and polyhedral domains). As in Chapter 3, we replace Problem (15) by the approximation

(19) *find* $\varphi_m \in \Phi^{\mathrm{I}}_m$ *minimizing* $\int_D \mu\, |\operatorname{grad} \varphi|^2$,

which is equivalent to

(20) *find* $\varphi_m \in \Phi^{\mathrm{I}}_m$ *such that* $\int_D \mu \operatorname{grad} \varphi_m \cdot \operatorname{grad} \varphi' = 0 \quad \forall\, \varphi' \in \Phi^0$.

There is a tiny difference in the present treatment, however: Observe that the solution is not unique! An additive constant has yet to be chosen, which can be achieved by, for instance, imposing $\boldsymbol{\varphi}_{\mathrm{n}} = 0$ for those nodes n that lie in $\mathrm{S}^{\mathrm{h}}_0$, and this is most often done, without even thinking about it. So did we in Chapter 3. But this time, I *do* want to call attention on this "gauge-fixing" procedure, trivial as it is in this case.

Since the minimization in (19) is performed on a smaller space than in (17), the minimum achieved is an upper estimate of the true one: $\mathrm{I}^2/\underline{\mathrm{R}}_m$

instead of I^2/R, with $I^2/\underline{R}_m \geq I^2/R$, hence $\underline{R}_m \leq R$. Solving (19) or (20) thus yields a *lower bound* for the reluctance.

Now, it would be nice to have an *upper* bound as well! If we could perform the same kind of Galerkin approximation on the "vector potential" version of the problem, (16) or (18), we would indeed have one: For if $A^F{}_m$ is some finite dimensional subspace of A^F, solving either the quadratic optimization problem,

$$\text{(21)} \qquad \textit{find } a_m \in A^F{}_m \textit{ minimizing } \int_D \mu^{-1} \, |\operatorname{rot} a|^2,$$

in terms of still to be defined degrees of freedom, or the associated linear system,

$$\text{(22)} \qquad \textit{find } a_m \in A^F{}_m \textit{ such that } \int_D \mu^{-1} \operatorname{rot} a_m \cdot \operatorname{rot} a' = 0 \ \ \forall \ a' \in A^0{}_m,$$

will yield an upper bound $\bar{R}_m F^2$ to RF^2. Hence the bilateral estimate

$$\text{(23)} \qquad \underline{R}_m \leq R \leq \bar{R}_m,$$

obviously a very desirable outcome. This is *complementarity*, as usually referred to in the literature [Fr, HP, PF].

There is more to it. Suppose one has solved both problems (19) and (21), and let us set

$$E(b_m, h_m) = \int_D \mu^{-1} \, |b_m - \mu \, h_m|^2 \equiv \int_D \mu^{-1} \, |\operatorname{rot} a_m - \mu \operatorname{grad} \varphi_m|^2.$$

Be well aware that I and F are unrelated, a priori (we'll return to this later), so even for the exact solutions h and b of (11) and (12), the error in constitutive law $E(b, h)$ does not vanish. Thus $E(b_m, h_m)$, obtained by minimizing on finite dimensional subspaces, will be even larger. Now,

Proposition 6.3. *One has*

$$E(b_m, h_m) = \int_D \mu^{-1} \, |b_m - b + b - \mu \, h + \mu(h - h_m)|^2$$

$$\text{(24)} \qquad = \int_D \mu^{-1} \, |b_m - b|^2 + \int_D \mu^{-1} \, |b - \mu \, h|^2 + \int_D \mu \, |h - h_m|^2,$$

Proof. Develop the first line and observe that all rectangle terms vanish, because of a priori orthogonality relations which Fig. 6.3 should help visualize. Indeed, $b_m - b$ and $h_m - h$ belong to $\operatorname{rot} A^0 \equiv \mathbb{B}^0$ and $\operatorname{grad} \Phi^0 \equiv \mathbb{H}^0$, which are orthogonal by Lemma 6.1. This disposes of the term $\int_D \mu^{-1} (b_m - b) \cdot (\mu \, (h - h_m))$. As for $\int_D \mu^{-1} (b - \mu \, h) \cdot (\mu \, (h - h_m)$, this is equal to $\int_D (b - \mu \, h) \cdot \operatorname{grad} \psi$ for some $\psi \in \Phi^0$, which vanishes because both b and μh are solenoidal. Same kind of argument for the third term $\int_D (b_m - b) \cdot (\mu^{-1} b - h)$, because $\operatorname{rot} h = 0$ and $\operatorname{rot}(\mu^{-1} b) = 0$. ◊

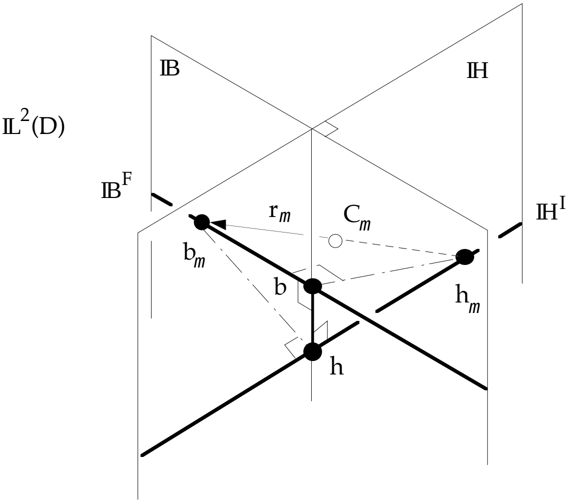


FIGURE 6.3. The geometry of complementarity. All right angles in sight, marked by carets, do correspond to orthogonality properties, in the sense of the scalar product of $\mathbb{L}^2(D)$, which result from Lemma 6.1 and from variational characterizations. (All $\mathbb{B}^F$s [resp. all $\mathbb{H}^I$s] are orthogonal to $\mathbb{H}$ [resp. to $\mathbb{B}$].) Note how h, h_m, b, b_m, all stand at the same distance r_m from $C_m = (h_m + b_m)/2$, on a common "hypercircle", and how the equality (24) can be read off the picture. For readability, this is drawn as if one had $\mu = 1$, but all geometric relations stay valid if all symbols h and $\mathbb{H}$ [resp. b and $\mathbb{B}$] are replaced by $\mu^{1/2}\,h$ and $\mu^{1/2}\,\mathbb{H}$ [resp. by $\mu^{-1/2}\,b$ and $\mu^{-1/2}\,\mathbb{B}$].

Since $E(b_m, h_m) = \int_D \mu^{-1}\,|b_m - \mu\,h_m|^2 \equiv \sum_{T\,\in\,\mathcal{T}} \int_T \mu^{-1}\,|b_m - \mu\,h_m|^2$ is a readily computable quantity, (24) stops the gap we had to deplore in Chapter 4: *At last, we have a posteriori bounds* on the approximation errors[4] for both h_m and b_m, which appear in first and third position at the right-hand side of (24). All it requires is to solve for *both* potentials φ and a, by some Galerkin method. Of course, the smaller the middle term $E(b, h)$, the sharper the bounds, so I and F should not be taken at random. For efficiency, one may set I, get φ_m, evaluate the flux F by the methods of Chapter 4 (cf. Subsection 4.1.3, especially Fig. 4.7), and finally, compute a_m for *this* value of the flux.

Exercise 6.1. Show that this procedure is actually optimal, giving the sharpest bound for a given I.

[4] Global bounds, not local: It is not necessarily true that $E_T(b_m, h_m)$, that is, $\int_T \mu^{-1}\,|b_m - \mu\,h_m|^2$, is an upper bound for $\int_T \mu^{-1}\,|b - b_m|^2$ and $\int_T \mu\,|h - h_m|^2$. Still, it's obviously a good idea to look for tetrahedra T with relatively high E_T, and to refine the mesh at such locations. Cf. Appendix C.

Remark 6.3. As Fig. 6.3 suggests, the radius r_m of the hypercircle is given by $[E(b_m, h_m)]^{1/2}/2$. This information allows one to get bilateral bounds on other quantities than the reluctance. Suppose some quantity of interest is a linear continuous functional of (say) h, L(h). There is a Riesz vector h^L for this functional (cf. A.4.3), such that $L(h) = \int_D \mu\, h^L \cdot h$. What is output is $L(h_m)$. But one has $|L(h) - L(h_m)| = |\int_D \mu\, h^L \cdot (h - h_m)| \le ||h^L||_\mu ||h - h_m||_\mu \le r_m\, ||h^L||_\mu$, hence the bounds. There is a way to express the value of the potential at a point x as such a functional [Gr]. Hence the possibility of *pointwise* bilateral estimates for the magnetic potentials. This was known long ago (see bibliographical comments at the end), but seems rarely applied nowadays, and some revival of the subject would perhaps lead to interesting applications. ◊

6.1.4 Constrained linear systems

With such incentives, it becomes almost mandatory to implement the vector potential method. All it takes is some Galerkin space $A^F{}_m$, and since the unknown a is *vector*-valued, whereas φ was *scalar*-valued, let's pretend we don't know about edge elements and try this: Assign a *vector-valued* DoF $\underline{\mathbf{a}}_n$ to each node, and look for the vector field a as a linear combination

$$a = \sum_{n \in \mathcal{N}} \underline{\mathbf{a}}_n w_n.$$

(We shall have to refer to the space spanned by such fields later, so let us name it $\mathbb{P}^1{}_m$, on the model of the P^1 of Chapter 3, the "blackboard" style reminding us that each DoF is a vector.) Now (21) is a quadratic optimization problem in terms of the Cartesian components of the vector DoFs. The difficulty is, these degrees of freedom are not *independent,* because a must belong to $A^F{}_m$. As such, it should first satisfy $n \cdot \text{rot}\, a = 0$ on S^b, that is, on each face f of the mesh that belongs to S^b. Assume again flat faces, for simplicity, and let n_f be the normal to f. Remembering that $\mathcal{N}(f)$ denotes the subset of nodes that belong to f, we have, on face f,

$$n \cdot \text{rot}\, a = \sum_{v \in \mathcal{N}(f)} n_f \cdot \text{rot}(\underline{\mathbf{a}}_v\, w_v) = \sum_{v \in \mathcal{N}(f)} n_f \cdot (\nabla w_v \times \underline{\mathbf{a}}_v)$$
$$\equiv \sum_{v \in \mathcal{N}(f)} (n_f \times \nabla w_v) \cdot \underline{\mathbf{a}}_v\,,$$

since $n \times \nabla w_v$ vanishes for all other nodes v (Fig. 6.4), hence the linear constraints to be verified by the DoFs:

$$\sum_{v \in \mathcal{N}(f)} (n_f \times \nabla w_v) \cdot \underline{\mathbf{a}}_v = 0$$

for each face f in S^b. The condition on the flux, $\int_C n \cdot \text{rot } a = F$, will also yield such a constraint. Taken all together, these constraints can be expressed as $\mathbf{L}\,\mathbf{a} = F\,\mathbf{k}$, where $\mathbf{L}$ is some rectangular matrix and $\mathbf{k}$ a fixed vector. (We shall be more explicit later about $\mathbf{L}$ and $\mathbf{k}$. Just note that entries of $\mathbf{L}$ are not especially simple, not integers at any rate, and frame-dependent.)

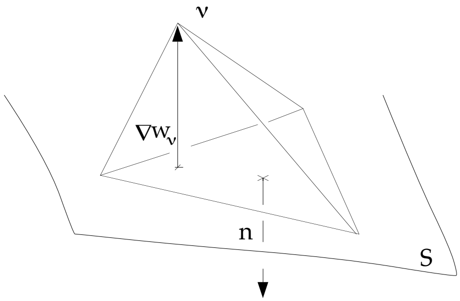


FIGURE 6.4. All surface fields $n \times \nabla w_\nu$ vanish, except when node ν belongs to the boundary.

To sum up: The quadratic optimization problem (21) is more complex than it would appear. Sure, the quantity to be minimized is a quadratic form in terms of the vector $\mathbf{a}$ of DoFs (beware, this is a vector of dimension 3N, if there are N nodes):

$$\int_D \mu^{-1}\, |\text{rot } a|^2 = (\mathbf{M}\,\mathbf{a},\, \mathbf{a}),$$

if we denote by $\mathbf{M}$ the associated symmetric matrix. But $(\mathbf{M}\mathbf{a},\, \mathbf{a})$ should be minimized *under the constraint* $\mathbf{L}\,\mathbf{a} = F\,\mathbf{k}$, so the components of $\mathbf{a}$ are not *independent* unknowns.

Problems of this kind, that we may dub *constrained linear systems*, happen all the time in numerical modelling, and there are essentially two methods to deal with them. Both succeed in removing the constraints, but one does so by increasing the number of unknowns, the other one by decreasing it.

The first method[5] consists in introducing Lagrange multipliers: minimize the Lagrangian $(\mathbf{M}\,\mathbf{a},\, \mathbf{a}) + 2(\boldsymbol{\lambda}, \mathbf{L}\,\mathbf{a})$ with respect to $\mathbf{a}$, without constraints, and adjust the vector of multipliers $\boldsymbol{\lambda}$ in order to enforce $\mathbf{L}\,\mathbf{a} = F\,\mathbf{k}$. This amounts to solving the following augmented linear system, in block form:

[5] Often referred to as the "dualization of constraints", a systematic application of the trick we encountered earlier in Exer. 2.9.

$$(25) \qquad \begin{vmatrix} \mathbf{M} & \mathbf{L}^t \\ \mathbf{L} & 0 \end{vmatrix} \begin{vmatrix} \mathbf{a} \\ \boldsymbol{\lambda} \end{vmatrix} = \mathrm{F} \begin{vmatrix} 0 \\ \mathbf{k} \end{vmatrix}.$$

This is what is called, according to a rather dubious but already entrenched terminology, a *mixed system*. It's a standard (unconstrained) symmetric linear system, but deprived of properties such as positive definiteness, so solving (25) is a challenge for which classical matrix analysis did not prepare us. See Appendix B for a few directions.

Exercise 6.2. What is the physical interpretation of the components of $\boldsymbol{\lambda}$?

The second method consists in expressing all unknowns in terms of a well-chosen set of independent variables. These may or may not coincide with a subset of the original unknowns. Most often they do, and picking the independent ones is so easy and so natural that one is not even aware of doing it. This is the case with Problem (19) or (20) above. Set in terms of $\boldsymbol{\varphi}$, (19) is actually a constrained linear system, the constraints on $\boldsymbol{\varphi}$ being as follows: (1) all $\boldsymbol{\varphi}_n$ for n in S^h_0 equal to some constant, (2) all those for n in S^h_1 equal to some other constant, and (3) the difference between these constants being equal to I. This can be compactly written as $\mathbf{L}\boldsymbol{\varphi} = \mathrm{I}\,\mathbf{k}$, just as before (with, of course, a different $\mathbf{L}$ and a different $\mathbf{k}$), and one could imagine using Lagrange multipliers. But it's much easier to set $\boldsymbol{\varphi}_n = 0$ for all n in S^h_0 (one will recognize the previous "gauge fixing" in action there), $\boldsymbol{\varphi}_n = \mathrm{I}$ for all n in S^h_1, and to solve with respect to other, obviously independent, nodal DoFs. So this is an example of a constrained linear system for which the sifting of dependent variables is straightforward.

Unfortunately, such is not the case with (21) or (22). The only recourse is to extract from $\mathbf{L}$ a submatrix of maximum rank (there do exist algorithms for this purpose [AE]), and thus to select independent variables to solve for. But this is a costly process. So the vector potential approach to magnetostatics looks unappealing, a priori.

6.2 SOLVING THE MAGNETOSTATICS PROBLEM

Should we then renounce the benefits of complementarity? No, thanks to edge elements.

6.2.1 Embedding the problem in Maxwell–Whitney's house

In fact, after Chapter 5, it's hard *not* to think of edge elements in the present context. We know, by Fig. 6.2, how the present problem fits within Maxwell's "continuous" building, so all we have to do is *embed* it in the relevant part of the Maxwell–Whitney "discrete" building. To make this formal, let us introduce a few definitions: Just as Φ^I_m above is the intersection $W^0_m \cap \Phi^I$, let us set $\mathbb{H}^I_m = W^1_m \cap \mathbb{H}^I$, as well as $\mathbb{B}^F_m = W^2_m \cap \mathbb{B}^F$ and—now committing ourselves to a definite approximation space— $A^F_m = W^1_m \cap A^F$. This is a sensible move, since elements of $\mathbb{H}^I_m$ and $\mathbb{B}^F_m$ have the kind of continuity required from h and b respectively, and the representations by potentials work nicely: indeed, $\mathrm{grad}\, \Phi^I_m = \mathbb{H}^I_m$ and $\mathrm{rot}\, A^F_m = \mathbb{B}^F_m$.

Any pair $\{h, b\}$ taken in $\mathbb{H}^I_m \times \mathbb{B}^F_m$ will thus satisfy all equations (1–7) except $b = \mu\, h$, the constitutive law. The latter cannot hold, since W^1_m and W^2_m are different spaces (and by the no-free-lunch principle we had to apply once already: *all* equations cannot *exactly* be satisfied when we discretize). Hence the question mark in the center of Fig. 6.5.

$$\begin{tikzcd}
\Phi^I \cap W^0 \arrow[d, "\mathrm{grad}"] & \varphi \arrow[d] & & 0 & W^3 \\
\mathbb{H}^I \cap W^1 \arrow[d, "\mathrm{rot}"] & h \arrow[r, dash] \arrow[d] & ? \arrow[r, dash] & b \arrow[u, "\mathrm{div}"] & W^2 \cap \mathbb{B}^F \arrow[u] \\
W^2 & 0 & & a \arrow[u, "\mathrm{rot}"] & W^1 \cap A^F \arrow[u]
\end{tikzcd}$$

FIGURE 6.5. Two copies of the top sequence of (5.13), put vertically, one downwards, one upwards, and appropriately clipped, form a framework in which Fig. 6.2 can be embedded, except for the horizontal relation, which must be relaxed.

So we settle for the next best thing, which is to minimize the error in constitutive law: *find* $h_m \in \mathbb{H}^I_m$ *and* $b_m \in \mathbb{B}^F_m$ *such that*

$$\int_D \mu^{-1}\, |b_m - \mu\, h_m|^2 \leq \int_D \mu^{-1}\, |b' - \mu\, h'|^2 \quad \forall\, h' \in \mathbb{H}^I_m, \quad \forall\, b' \in \mathbb{B}^F,$$

and for exactly the same reasons as in Prop. 6.2, this splits into a pair of independent problems:

(26) $\quad$ *find* $\mathrm{h}_m \in \mathbb{H}^{\mathrm{I}}{}_m$ *minimizing* $\int_{\mathrm{D}} \mu\ |\mathrm{h}|^2,$

(27) $\quad$ *find* $\mathrm{b}_m \in \mathbb{B}^{\mathrm{F}}{}_m$ *minimizing* $\int_{\mathrm{D}} \mu^{-1}\ |\mathrm{b}|^2.$

To investigate the algebraic nature of these problems, let us express them in terms of degrees of freedom. (Recall the generic notation $\mathcal{E}(\ldots)$ or $\mathcal{F}(\ldots)$ for the sets of edges or faces that belong to some region of space or to some geometric element (surface, line . . .) whose name stands inside the parentheses.) We assume the link c and the cut C of Fig. 6.1 are unions of edges and faces of the mesh. By analogy with the incidence numbers $\mathbf{R}_{\mathrm{f\,e}}$, let us have $\mathbf{R}_{\mathrm{c\,e}} = \pm 1$ if edge e belongs to c, the sign being plus if their orientations match, and $\mathbf{R}_{\mathrm{c\,e}} = 0$ otherwise. Similarly, let $\mathbf{D}_{\mathrm{C\,f}} = \pm 1$ if face f belongs to C, with the same sign convention, and 0 otherwise. We may now define

$$\mathbf{H}^{\mathrm{I}} = \{\mathbf{h} \in \mathbf{W}^1 : \ \mathbf{h}_{\mathrm{e}} = 0\ \ \forall\, \mathrm{e} \in \mathcal{E}(\mathrm{S}^{\mathrm{h}}),\ \textstyle\sum_{\mathrm{e}\in\mathcal{E}(\mathrm{c})} \mathbf{R}_{\mathrm{c\,e}}\, \mathbf{h}_{\mathrm{e}} = \mathrm{I}\},$$

$$\mathbf{B}^{\mathrm{F}} = \{\mathbf{b} \in \mathbf{W}^2 : \ \mathbf{b}_{\mathrm{f}} = 0\ \ \forall\, \mathrm{f} \in \mathcal{F}(\mathrm{S}^{\mathrm{b}}),\ \textstyle\sum_{\mathrm{f}\in\mathcal{F}(\mathrm{C})} \mathbf{D}_{\mathrm{C\,f}}\, \mathbf{b}_{\mathrm{f}} = \mathrm{F}\},$$

and using the mass matrices of Chapter 5, problems (26) and (27) are equivalent to

(28) $\quad$ *find* $\mathbf{h} \in \mathbf{H}^{\mathrm{I}}$ *such that* $(\mathbf{M}_1(\mu)\, \mathbf{h}, \mathbf{h})$ *is minimum,*

(29) $\quad$ *find* $\mathbf{b} \in \mathbf{B}^{\mathrm{F}}$ *such that* $(\mathbf{M}_2(\mu^{-1})\, \mathbf{b}, \mathbf{b})$ *is minimum,*

two "constrained linear systems", according to the foregoing terminology. Solving both will give the bilateral estimate (23) of the reluctance.

As constrained linear systems, problems (28) and (29) can be attacked by both general strategies: (1) introduce Lagrange multipliers, hence the so-called "mixed" formulations (and though not very popular, some of them have been tried; see, e.g., [PT]), or (2) select independent variables. For this second strategy, there are again two variants. Independent variables can just be picked among the original ones, and this is what spanning tree extraction techniques permit (cf. Fig. 6.6), hence numerical methods in terms of h and b directly. Both have been considered [Ke]. (The one in b seems less robust [Ke], and it would be worthwhile to understand why.) The second variant corresponds to the introduction of potentials: node or edge variables that help represent the above h and b, while automatically taking constraints into account.

Thus treated, (26) and (27) become

(30) $\quad$ *find* $\varphi_m \in \Phi^{\mathrm{I}}{}_m$ *such that* $\int_{\mathrm{D}} \mu\ |\mathrm{grad}\, \varphi|^2$ *is minimum,*

(31) $\quad$ *find* $\mathrm{a}_m \in \mathrm{A}^{\mathrm{F}}{}_m$ *such that* $\int_{\mathrm{D}} \mu^{-1} \,|\mathrm{rot\,a}|^2$ *is minimum,*

and in terms of degrees of freedom,

(32) $\quad$ *find* $\boldsymbol{\varphi} \in \boldsymbol{\Phi}^{\mathrm{I}}$ *such that* $(\mathbf{M}_1(\mu)\,\mathbf{G}\boldsymbol{\varphi}, \mathbf{G}\boldsymbol{\varphi})$ *is minimum,*

(33) $\quad$ *find* $\mathbf{a} \in \mathbf{A}^{\mathrm{F}}$ *such that* $(\mathbf{M}_2(\mu^{-1})\,\mathbf{R}\mathbf{a}, \mathbf{R}\mathbf{a})$ *is minimum,*

where $\boldsymbol{\Phi}^{\mathrm{I}} = \{\boldsymbol{\varphi} \in \mathbf{W}^0 : \mathbf{L}\boldsymbol{\varphi} = \mathrm{I}\,\mathbf{k}\}$ has already been described, for system (32) is nothing else than the scalar potential method of Chapter 3. The novelty is (33), which realizes (at last!) the discretization "on the curl side" of Eqs. (2.20) to (2.26).

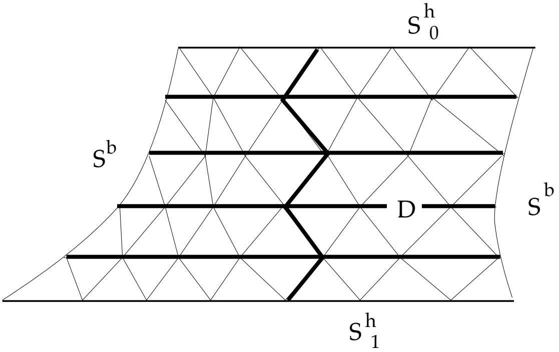


FIGURE 6.6. Tree of edges for the model problem in dimension 2. (This is actually a "belted tree mod S^h" in the language of Section 5.3.) Note how all edge circulations of h are determined by those of the tree edges (in thick lines), thanks to rot h = 0 (cf. Fig. 5.5) and $\mathrm{n} \times \mathrm{h} = 0$ on S^h.

So let us look at $\mathbf{A}^{\mathrm{F}}$, or equivalently, at the associated space of fields, $\mathrm{A}^{\mathrm{F}}{}_m$. Elements of the latter are subject to the conditions $\mathrm{n} \cdot \mathrm{rot\,a} = 0$ over each face f in S^b and $\int_C \mathrm{n} \cdot \mathrm{rot\,a} = \mathrm{F}$. Since rot a is mesh-wise constant, $\mathrm{n} \cdot \mathrm{rot\,a} = 0$ over f results in one relation between the $\mathbf{a}_e$s, namely $\sum_{e \in \mathcal{E}} \mathbf{R}_{f\,e}\, \mathbf{a}_e = 0$, which involves only the DoFs of the three edges that bound f (the trick of Fig. 5.5, again). The condition on the flux through C results in a similar relation: $\sum_{e \in \mathcal{E}(\partial C)} \mathbf{R}_{\partial C\, e}\, \mathbf{a}_e = \mathrm{F}$, where $\mathbf{R}_{\partial C\, e} = \pm 1$, depending on the orientation of edge e relative to ∂C. Taken all together, the constraints can thus be expressed as $\mathbf{L\,a} = \mathrm{F}\,\mathbf{k}$, just as in the case of nodal vectorial elements, but **L** is now much simpler, with entries ± 1 or 0 that are obtained by simply looking at how edges are oriented. This is a considerable improvement.

Still, these constraints are in the way, and whether we can get rid of them simply is the litmus test that will, if passed, establish the superiority of edge elements in the vector potential approach.

6.2.2 Dealing with the constraints

Indeed, the constraints can be removed by the following method. First construct a special DoF vector $\mathbf{a}^F$, such that the associated field $a^F = \sum_{e \in \mathcal{E}} \mathbf{a}^F_e w_e$ belong to A^F. This will be done in the next paragraph. Then, instead of minimizing over all A^F_m, we shall do it over fields of the form $a^F + \sum \mathbf{a}_e w_e$, where the index e spans $\mathcal{E} - \mathcal{E}(S^b)$, i.e., without any restriction on the $\mathbf{a}_e$s, but excluding edges of S^b. Edge DoFs in this summation are now *independent*. The final version of the problem will thus be: *find* $\mathbf{a} \in \tilde{\mathbf{A}}^F$ *such that* $(\mathbf{M}_2(\mu^{-1})\, \mathbf{R}\mathbf{a}, \mathbf{R}\mathbf{a})$ *be minimum*, where $\tilde{\mathbf{A}}^F$ is the subset $\{\mathbf{a} \in \mathbf{W}^1 : \mathbf{a}_e = \mathbf{a}^F_e \ \forall\ e \in \mathcal{E}(S^b)\}$, which amounts to solving a linear system with respect to the DoFs of the "inner edges" (those not in S^b).

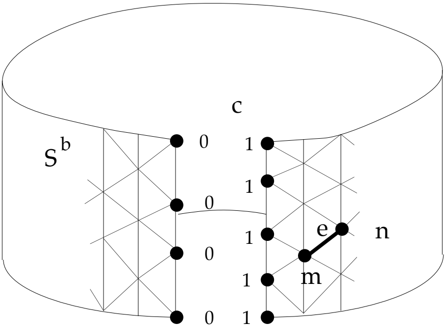


FIGURE 6.7. One step in the construction of $\mathbf{a}^F$: assigning scalar values to nodes on the boundary S^b, after having doubled the nodes on the "cut" c.

The construction of $\mathbf{a}^F$ proceeds as follows (Fig. 6.7). Consider the mesh of surface S^b, as induced by m. Make a (one-dimensional) "cut" c by following a path of edges from top to bottom. Double the nodes along c. Assign scalar values $\mathbf{v}_n$ to nodes, arbitrary values, except for the pairs of nodes along c, that should receive zeroes and ones, on the pattern suggested by Fig. 6.7. Then, assign to all edges $e = \{m, n\}$ the value $\mathbf{a}^F_e = (\mathbf{v}_n - \mathbf{v}_m)F$ if $e \in \mathcal{E}(S^b)$, $\mathbf{a}^F_e = 0$ otherwise. The recipe works because, whatever $a = \sum\{e \in \mathcal{E} : \mathbf{a}_e w_e\}$ in A^F_m, there is a modified field $\tilde{a} = \sum_{e \in \mathcal{E}} \tilde{\mathbf{a}}_e w_e$, with $\tilde{\mathbf{a}}_e = \mathbf{a}^F_e$ if $e \in \mathcal{E}(S^b)$ (hence $\tilde{\mathbf{a}} \in \tilde{\mathbf{A}}$), *that has the same curl* as a (this is the point of the above construction). Thus, one minimizes in (31) over a space $\tilde{A}^F_m$ strictly smaller than A^F_m, but with $\operatorname{rot} A^F_m = \operatorname{rot} \tilde{A}^F_m$ ($\equiv \mathbb{B}^F_m$), so the same b is reached.

Remark 6.4. As the kernel $\ker(\mathbf{R}\,;\, \tilde{\mathbf{A}}) = \{\mathbf{a} \in \tilde{\mathbf{A}} : \mathbf{R}\mathbf{a} = 0\}$ does not reduce to 0, this final version of the vector potential approach does not give a unique a, although a unique b is obtained. The associated linear system is therefore singular. We shall return to this apparent difficulty. ◊

6.2.3 "*m*-weak" properties

We should now proceed as in Section 4.1, and answer questions about the quality of the approximation provided by a_m : We are satisfied that $b_m = \text{rot}\ a_m$ is solenoidal and that $n \cdot b_m = 0$ on the S^b boundary, but what is left of the "weak irrotationality" of $h_m = \mu^{-1} b_m$? And so forth.

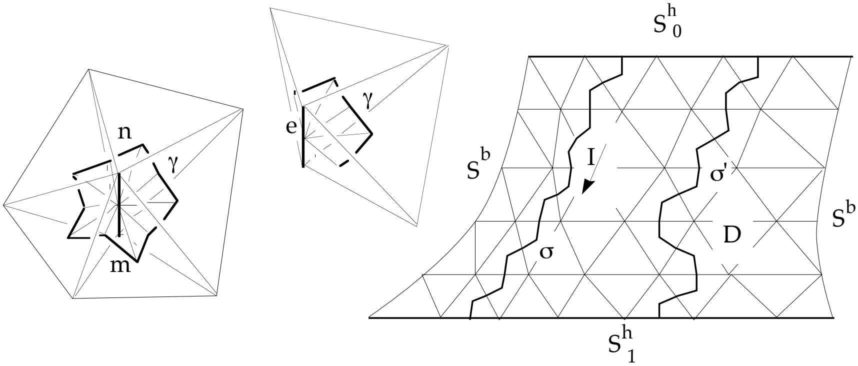


FIGURE 6.8. "*m*-weak" properties of the vector potential solution. Left: The circulation of $h_m = \mu^{-1}$ rot a_m is zero along the circuit γ that joins barycenters around any inner edge e = {m, n}. Middle: If e belongs to S^h, the circulation is null along the open path γ. Right: The "variationally correct" mmf is obtained by taking the circulation of the computed h_m along the "*m**-line" σ joining S^h_0 and S^h_1, or along any homologous *m**-line σ'.

It would be tedious to go through all this again, however, and there is more fun in *guessing* the results, thanks to the analogies that Tonti's diagrams so strongly suggest. So we can expect with confidence the following statements to be true:

- For any DoF vector **a**, the term $(\mathbf{R}^t \mathbf{M}_2(\mu^{-1})\ \mathbf{R}\mathbf{a})_e \equiv \int_D h \cdot \text{rot}\ w_e$, where $h = \mu^{-1}\ \text{rot}(\sum_{e \in \mathcal{E}} \mathbf{a}_e w_e)$, is the circulation of h along the smallest closed *m**-line (closed mod S, in the case of surface edges) that surrounds edge e (cf. Fig. 6.8).

- The circulation of the computed field, h_m, vanishes along all *m**-lines which bound modulo S^h (cf. Fig. 4.6, right).

- The circulation of h_m is equal to I along all *m**-paths similar to σ (or its homologue σ') on Fig. 6.8.

Exercise 6.3. Prove all this.

There is more to say about complementarity. In particular, there is an obvious problem of duplication of efforts: Computations of φ and a are independent, though their results are closely related. No use is made of the knowledge of φ when solving for a, which seems like a waste, since as one rightly suspects, and as the following discussion will make clear, solving for a is much more expensive than solving for φ. There is a way to save on this effort, which is explained in detail in Appendix C. We now address the more urgent question of whether edge elements are really mandatory in the vector potential approach.

6.3 WHY NOT STANDARD ELEMENTS ?

As we saw, edge elements are able to solve a problem that was quite difficult with nodal vectorial elements, namely, to obtain a vector potential formulation in terms of explicit *independent* degrees of freedom. This is a good point, but not the whole issue, for difficult does not mean impossible, and if bilateral estimates, or any other consequence of the hypercircle trick, are the objective, one must concede that it *can* be achieved with standard nodal elements, scalar-valued for φ, vector-valued for a. All that is required is approximation "from inside" (within the functional space), Galerkin style. It's thus a gain in simplicity or in accuracy, or both, that we may expect of edge elements.

To make a fair comparison, let us suppose that, after proper selection of independent variables, the vector potential approach with nodal elements consists in looking for a in some subspace of $\mathbb{P}^1_m$, that we shall denote $\tilde{A}^F_m(\mathbb{P}^1)$. Let us similarly rename $\tilde{A}^F_m(W^1)$ the above $\tilde{A}^F_m$, to stress its relationship with edge elements. In both methods, the quantity $\int_D \mu^{-1} |\text{rot a}|^2$ is minimized, but on different subspaces of $\mathbb{L}^2_{\text{rot}}(D)$, which results in linear systems of the same form $\mathbf{Ma} = \mathbf{b}$, but with different $\mathbf{M}$ and $\mathbf{b}$, and a different interpretation for the components of $\mathbf{a}$. Let us call $\mathbf{M}_P$ and $\mathbf{M}_W$, respectively, the matrix $\mathbf{M}$ in the case of the $\mathbb{P}^1$ and the W^1 approximation. Both are symmetric and (a priori) nonnegative definite. We shall find the nodal vectorial approach inferior on several counts. But first . . .

6.3.1 An apparent advantage: $\mathbf{M}_P$ is regular

Indeed, let $a \in \tilde{A}^F_m(\mathbb{P}^1)$ be such that $\text{rot a} = 0$, which is equivalent to $\mathbf{Ma} = 0$. Then $a = \text{grad}\ \psi$. Since a is piecewise linear with respect to the mesh m a*nd* continuous (in all its three scalar components), φ is piecewise quadratic *and* differentiable, two hardly compatible conditions. The space P^2 of piecewise quadratic functions on m is generated by the products $\mathrm{w}_\mathrm{n}\,\mathrm{w}_\mathrm{m}$, where n and m span the nodal set $\mathcal{N}$. Therefore, $\operatorname{grad}\psi = \sum_{\mathrm{m,n}\,\in\mathcal{N}}\alpha_{\mathrm{mn}}\operatorname{grad}(\mathrm{w}_\mathrm{m}\mathrm{w}_\mathrm{n})$, and since the products $\mathrm{w}_\mathrm{n}\,\mathrm{w}_\mathrm{m}$ are not differentiable, the normal component of this field has a nonzero jump across all faces. (This jump is affine with respect to coordinates over a face.) Demanding that all these jumps be 0 is a condition that considerably constrains the α's (in practice, only globally quadratic ψ, as opposed to mesh-wise quadratic, will comply). Consequently, the kernel of rot in $\tilde{\mathrm{A}}^{\mathrm{F}}{}_{m}(\mathbb{P}^1)$ will be of very low dimension, and as a rule reduced to 0, because of additional constraints imposed by the boundary conditions. So unless the mesh is very special, one may expect a regular $\mathbf{M}_\mathrm{P}$, whereas the matrix $\mathbf{M}_\mathrm{W}$ is singular, since W^1 contains gradients (cf. Prop. 5.4, asserting that $\mathrm{W}^1 \supset \operatorname{grad}\mathrm{W}^0$).

Good news? We'll see. Let us now look at the weak points of the nodal vectorial method.

6.3.2 Accuracy is downgraded

When working with the same mesh, accuracy is downgraded with nodal elements, because $\operatorname{rot}(\tilde{\mathrm{A}}^{\mathrm{F}}{}_{m}(\mathbb{P}^1)) \subset \operatorname{rot}(\tilde{\mathrm{A}}^{\mathrm{F}}{}_{m}(\mathrm{W}^1))$, with as a rule a *strict* inclusion. Minimizing over a smaller space will thus yield a less accurate upper bound in the case of nodal elements, for a given mesh m. (Let's omit the subscript m for what follows, as far as possible. Recall that N, E, F, T refer to the number of nodes, etc., in the mesh.)

The inclusion results from this:

Proposition 6.4. *For a given mesh* m, *any field* $\mathrm{u}\in\mathbb{P}^1$ *is sum of some field in* W^1 *and of the gradient of some piecewise quadratic function, i.e.,*

$$\mathbb{P}^1 \subset \mathrm{W}^1 + \operatorname{grad}\mathrm{P}^2.$$

Proof. Given $\mathrm{u}\in\mathbb{P}^1$, set $\mathbf{u}_\mathrm{e} = \int_\mathrm{e} \tau\cdot\mathrm{u}$, for all $\mathrm{e}\in\mathcal{E}$, and let $\mathrm{v} = \sum_{\mathrm{e}\,\in\mathcal{E}}\mathbf{u}_\mathrm{e}\,\mathrm{w}_\mathrm{e}$ be the field in W^1 which has these circulations as edge DoFs. Then, both u and v being linear with respect to coordinates, $\operatorname{rot}(\mathrm{u}-\mathrm{v})$ is piecewise constant. But its fluxes through faces are 0, by construction (again, see Fig. 5.5), so it vanishes. Hence $\mathrm{u} = \mathrm{v} + \nabla\varphi$, where φ is such that $\nabla\varphi$ be piecewise linear, that is, $\varphi\in\mathrm{P}^2$. ◊

As a corollary, $\operatorname{rot}\mathbb{P}^1 \subset \operatorname{rot}\mathrm{W}^1$, hence the inclusion, as far as (but this is what we assumed for fairness) $\tilde{\mathrm{A}}^{\mathrm{F}}{}_{m}(\mathbb{P}^1) \subset \mathrm{A}^\mathrm{F}$. As a rule, this is strict inclusion, because the dimension of $\operatorname{rot}\mathbb{P}^1$ cannot exceed that of $\mathbb{P}^1$, which is 3N (three scalar DoFs per node), whereas the dimension of rot W^1 is (approximately the same as) that of the quotient $W^1/\mathrm{grad}(W^0)$, which is E – N + 1, that is, 5 to 6 N, depending on the mesh, as we know (cf. 4.1.1).

Exercise 6.4. Check that $\mathbb{P}^1$ is *not* contained in W^1.

Exercise 6.5. Show that $W^1 \subset \mathbb{P}^1 + \mathrm{grad}\, P^2$ does *not* hold.

6.3.3 The "effective" conditioning of the final matrix is worsened

The importance of the *condition number* of a matrix, that is, the ratio of its extreme eigenvalues, is well known. This number determines to a large extent the speed of convergence of iterative methods of solution, and the numerical accuracy of direct methods.

This is true, that is, in the case of *regular* matrices. But if a symmetric nonnegative definite matrix $\mathbf{M}$ is singular, this does not preclude the use of iterative methods to solve $\mathbf{Ma} = \mathbf{f}$. All that is required is that $\mathbf{f}$ be in the range of $\mathbf{M}$, so that $(\mathbf{Ma}, \mathbf{a}) - 2(\mathbf{f}, \mathbf{a})$ have a finite lower bound. Then any "descent" method (i.e., one that tries to minimize this function by decreasing its value at each iteration) will yield a minimizing sequence $\mathbf{u}_k$ that may not converge, but *does converge modulo* ker($\mathbf{M}$), and this may be just enough. To be definite, suppose the matrix $\mathbf{M}$ is a principal submatrix of $\mathbf{R}^t\mathbf{M}(\mu^{-1})\mathbf{R}$, as was shown to be the case with edge elements. The quadratic form to be minimized is indeed bounded from below, and the desired convergence is that of $\mathbf{Ra}_k$, not $\mathbf{a}_k$. By working out the simple example of the iterative method $\mathbf{u}_{n+1} = \mathbf{u}_n - \rho\,(\mathbf{Ma}_n - \mathbf{f})$, which is easy in the basis of eigenvectors of $\mathbf{M}$, one will see that what counts, as far as convergence modulo ker($\mathbf{M}$) is concerned, is the ratio of extreme strictly positive eigenvalues. Let us call this *effective* conditioning, denoted by $\kappa(\mathbf{M})$.

We now show that $\kappa(\mathbf{M}_P) >> \kappa(\mathbf{M}_W)$, thus scoring an important point for edge elements. This will be quite technical, unfortunately.

Both matrices can be construed as approximations of the "curl–curl" operator, $\mathrm{rot}(\mu^{-1}\,\mathrm{rot}\,)$, or rather, of the associated boundary-value problem. Their higher eigenvalues have similar asymptotic behavior, when the mesh is refined. (I shall not attempt to prove this, which is difficult,[6] but it can easily be checked for meshes with a regular, repetitive layout.) So we should compare the first positive eigenvalue $\lambda_1(\mathbf{M}_W)$ with its homologue $\lambda_1(\mathbf{M}_P)$. As we noticed, $\mathbf{M}_P$ is regular, in general. But zero is an eigenvalue of the curl–curl operator, and as a rule, spectral elements of an operator are approximated (when the mesh is repeatedly refined while

[6] The difficulty lies in *stating* the claim with both precision and generality.

keeping flatness under control, which we informally denote as "$m \to 0$") by those of its discrete matrix counterpart, which *does not contain* 0. *Therefore,* when $m \to 0$, $\lambda_1(\mathbf{M}_P)$ tends to 0. But for $\mathbf{M}_W$, the situation is quite different: This matrix is singular, since it contains the vectors $\mathbf{a} = \mathbf{G}\psi$ (with $\psi_n \neq 0$ for all n not in S^b). No need in consequence for the eigenvalue 0 to be approximated "from the right", as was the case for $\mathbf{M}_P$.

And indeed, $\lim_{m \to 0} \lambda_1(\mathbf{M}_W) > 0$. This can be seen by applying the Rayleigh quotients theory, according to which

$$\lambda_1(\mathbf{M}_W) = \inf\{(\mathbf{M}_W\, \mathbf{a}, \mathbf{a}) : |\mathbf{a}| = 1,\ (\mathbf{a}, \mathbf{a}') = 0\ \forall\, \mathbf{a}' \in \ker(\mathbf{M}_W)\}.$$

In terms of the associated vector fields, this orthogonality condition means

$$(34) \qquad \int_D \mathrm{a} \cdot \operatorname{grad} \psi' = 0 \quad \forall\, \psi' \in \Psi^0 \cap W^0_m,$$

where $\Psi^0 = \{\psi \in L^2_{\mathrm{grad}}(D) : \psi = 0 \text{ on } S^b\}$. Let $\mathrm{a}_1(m)$ be the field whose DoFs form the eigenvector $\mathbf{a}_1$ corresponding to λ_1, and a_1 its limit when $m \to 0$. Equation (34) holds for a_1. Therefore (take the projections on W^0_m of ψ' in (34), and pass to the limit),

$$\int_D \mathrm{a}_1 \cdot \operatorname{grad} \psi' = 0 \quad \forall\, \psi' \in \Psi^0,$$

and a_1 is thus divergence-free. Hence

$$\lim_{m \to 0} \lambda_1(\mathbf{M}_W) = \inf\{\int_D \sigma^{-1} |\operatorname{rot} \mathrm{a}|^2 : \mathrm{a} \in \tilde{A}^0,\ \operatorname{div} \mathrm{a} = 0,\ \int_D |\mathrm{a}|^2 = 1\},$$

and this Rayleigh quotient is strictly positive.

So we may conclude that $\kappa(\mathbf{M}_W)/\kappa(\mathbf{M}_P)$ tends to 0: Effective conditioning is asymptotically better with edge elements.

6.3.4 Yes, but ...

Is the case over? Not yet, because the defendants have still some arguments to voice. Ease in setting up boundary conditions? Yes, but think of all these standard finite element packages around. Reusing them will save much effort. Bad conditioning? Yes, but asymptotically so, and we don't go to the limit in practice; we make the best mesh we can, within limits imposed by computing resources; this mesh may not be the same for both methods, since the number of degrees of freedom will be different, so the comparison may well be of merely academic interest. Same thing about the relative loss of accuracy: Given the same resources, we may use more refined meshes in the case of nodal elements, since the number of degrees of freedom is lower, apparently. After all, the number of edges E is much higher than 3N, isn't it?

Let us count again. Assume the mesh is first done with bricks, each of these being further divided into five tetrahedra (cf. Exer. 3.7). Thus, $T \approx 5N$. Then $F \approx 10\,N$ (four faces for each tetrahedron, shared by two), and the Euler–Poincaré formula, that is, as we know,

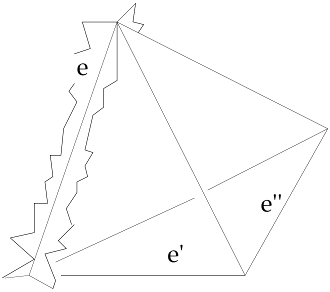
$$N - E + F - T = \chi(D),$$

shows that $E \approx 6N$. So indeed, the number of DoFs in the edge element method (about $E \approx 6N$) will be twice as high as in the nodal vectorial one ($\approx 3N$).

These figures, which ignore boundary conditions, are quite approximate (cf. [Ko] for precise counting). The ratios are valid for big meshes only. Still, the conclusion is neat: Tetrahedral edge elements generate more degrees of freedom than classical elements.

But is that really topical? The most meaningful number, from the point of view of data storage and CPU time, is not the size of the matrix, but the number of its nonzero entries. It happens this number is *smaller*, for a given mesh, for $\mathbf{M}_W$ than for $\mathbf{M}_P$, against pessimistic expectations.

For let us count the average number of entries on a given row of $\mathbf{M}_P$: This equals the number of DoFs that may interact with a given nodal one, that is, if we denote by v_i the basis vectors in a Cartesian frame, the number of couples $\{m, j\}$ for which $\int_D \mu^{-1} \operatorname{rot}(v_i w_n) \cdot \operatorname{rot}(v_j w_m) \neq 0$, for a given $\{n, i\}$. But $\operatorname{rot}(v_i w_n) = -\, v_i \times \operatorname{grad} w_n$, so this term vanishes if the supports of w_n and w_m don't intersect, so two DoFs may interact only when they belong to the same node, or to nodes linked by an edge. As each node is linked with about 12 neighbors, there are 38 extra-diagonal non-vanishing terms on a row of $\mathbf{M}_P$, on average.



As for $\mathbf{M}_W$, the number of extra-diagonal non-vanishing terms on the row of edge e is the number of edges e' for which the integral $\int_D \mu^{-1} \operatorname{rot} w_e \cdot \operatorname{rot} w_{e'}$ differs from 0, that is, edges belonging to a tetrahedron that contains e. On the average, e belongs to 5 tetrahedra (because it belongs to 5 faces, since $F \approx 5E/3$, each face having 3 edges). This edge thus has 15 "neighbors" (inset): 10 edges which share a node with e, and 5 opposite in their common tetrahedron. So there are about 15 extra-diagonal nonzero entries per row, that is, about 90 N terms of this kind in $\mathbf{M}_W$, against $3 \times 38 = 114$ N in $\mathbf{M}_P$, a sizable advantage in favor of edge elements.

This rather satisfying conclusion should not mask the obvious problem with *all* vector potential methods, whatever the finite elements: a large number of DoFs, relatively. In the same conditions as previously, the φ method only generates 12N off-diagonal nonzero terms. In Appendix C, we shall see how some savings are possible, but only to some extent. Complementarity has its price.

6.3.5 Conclusion

Focusing on magnetostatics as I do in this book has obvious shortcomings, but also the advantage of delimiting a narrow field in which theory can deploy itself, and any fundamental observation in this limited area has all chances to be valid for magnetics in general. This seems to be the case of the nodal elements vs edge elements debate. But of course, theory will not carry the day alone, and numerical experience is essential. We have a lot of that already. It began with M. Barton's thesis [Br] (see an account in [BC]), whose conclusions deserve a quote:

> When the novel use of tangentially continuous edge-elements for the representation of magnetic vector potential was first undertaken, there was reason to believe it would result in an interesting new way of computing magnetostatic field distributions. There was only hope that it would result in a significant improvement in the state-of-the-art for such computations. As it has turned out, however, the new algorithm has significantly out-performed the classical technique in every test posed. The use of elements possessing only tangential continuity of the magnetic vector potential allows a great many more degrees of freedom to be employed for a given mesh as compared to the classical formulation; and these degrees of freedom result in a global coefficient matrix no larger than that obtained from the smaller number of degrees of freedom of the other method. (. . .) It has been demonstrated that the conjugate gradient method for solving sets of linear equations is well-defined and convergent for symmetric but underdetermined sets of equations such as those generated by the new algorithm. As predicted by this conclusion, the linear equations have been successfully solved for all test problems, and the new method has required significantly fewer iterations to converge in almost all cases than the classical algorithm.

One may also quote this [P&], about scattering computations:

> Experience with the 3D node-based code has been much more discouraging: many problems of moderate rank failed to converge within N iterations (. . .). The

Whitney elements were the only formulation displaying robust convergence with diagonal PBCG[7] iterative solution.

Countless objections have been raised against edge elements. The most potent one is that W^1, contrary to IP^1, does not contain globally linear vector fields, like for instance $x \rightarrow x$, and thus lack "first-order completeness". This is both true and irrelevant. In magnetostatics, where the object of attention is not the unknown a but its curl, we already disposed of the objection with Prop. 6.4. But even in eddy-current computations (where, as we'll see in Chapter 8, edge elements are natural approximants for the field h), it does no good to enlarge W^1 to the space spanned by the $w_n \nabla w_m$ (which does include linear fields). See [B4, DB, Mk].

The debate on this and other issues relative to the edge elements vs nodal elements contest winds its way and will probably go on for long, but it would be tedious to dwell on it. As one says, those who ignore history are bound to live it anew. A lot remains to be done, however: research on higher-order edge elements (to the extent that [Ne] does not close the subject), error analysis [Mk, MS, Ts], edge elements on other element forms than tetrahedra, such as prisms, pyramids, etc. [D&].

The problem of singularity addressed in Remark 6.4 is crucial, in practice, when there is a distributed source-term, as is the case in magnetostatics. If the discretization of the right-hand side is properly done, one will obtain a system of the form $\mathbf{R}^t\mathbf{M}_2(\sigma^{-1})\mathbf{R}\,\mathbf{a} = \mathbf{R}^t\mathbf{k}$, where $\mathbf{a}$ is the DoF-vector of the vector potential, and $\mathbf{k}$ a given vector, and in this case the right-hand side $\mathbf{R}^t\mathbf{k}$ is in the range of $\mathbf{R}^t\mathbf{M}_2(\sigma^{-1})\mathbf{R}$. But otherwise, the system has no solution, which the behavior of iterative algorithms tells vehemently (drift, as evoked in 6.3.3, slowed convergence, if not divergence). This seems to be the reason for the difficulties encountered by some, which can thus easily be avoided by making sure that the right-hand side behaves [Re]. There, again, tree–cotree techniques may come to the rescue.

Finally, let us brush very briefly on the issue of singularity. Should the edge-element formulation in a be "gauged", that is, should the process of selecting *independent* variables be pushed further, to the point of having only *non-redundant* DoFs? This can be done by extracting spanning trees. Such gauging is necessary with nodal elements, but not with edge elements. Experiments by Ren [Re] confirm this, which was already suggested by Barton's work.

[7] This refers to the "preconditioned biconjugate gradient" algorithm [Ja].

EXERCISES

See p. 171 for Exercise 6.1, p. 174 for Exer. 6.2, p. 179 for Exer. 6.3, and p. 174 for Exers. 6.4 and 6.5.

Exercise 6.6. Let a smooth surface S be equipped with a field of normals. Given a smooth function φ and a smooth vector field u, we may define the restriction of φ to S, or *trace* φ_S, the *tangential part* u_S of u (that is, the surface field of orthogonal projections of vector $u(x)$ onto the tangent plane at x, where x spans S, as in Fig. 2.5), and the *normal component* $n \cdot u$ of u. For smooth functions and tangential fields living on S, like φ_S and u_S, define operators grad_S, rot_S, and div_S in a sensible way, and examine their relationships, including integration-by-parts formulas.

HINTS

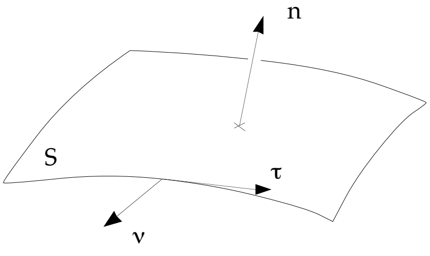
6.1. Notice that this approach amounts to solving (11) and (12′).

6.2. Their physical *dimension* is the key. Note that components of $\mathbf{L\,a}$ are induction fluxes, and $(\mathbf{M\,a}, \mathbf{a})$ has the dimension of energy.

6.3. For a tetrahedron T which contains $e = \{m, n\}$, integrate by parts the contribution $\int_T h \cdot \text{rot}\, w_e$, hence a weighted sum of the jump $[n \times h]$ over ∂T. Check that faces opposite n or m contribute nothing to this integral. As for faces f which contain e, relate $\int_f [n \times h] \cdot w_e$ with the circulation of $[h]$ along the median. Use $\text{rot}\, h_m = 0$ *inside* each tetrahedron to derive the conclusion.

6.4. Compute the divergence of $u = \sum_{n \in \mathcal{N}} \underline{\mathbf{u}}_n w_n$.

6.5. Take the curls.



6.6. Obviously, $\text{grad}_S\, \varphi_S$ must be defined as $(\text{grad}\, \varphi)_S$, and $\text{rot}_S u_S$ as $n \cdot \text{rot}\, u$, when φ and u live in 3D space, for consistency. (Work in x–y–z coordinates when S is the plane $z = 0$ to plainly see that.) Verify that these are indeed *surface* operators, that is, they only depend on the traces on S of fields they act on. Define div_S by Ostrogradskii–Gauss (in order to have a usable integration by parts formula on S), and observe its kinship with rot_S. You'll see that a second integration-by-parts formula is wanted. Use notation as suggested in inset.

SOLUTIONS

6.1. It's equivalent to the two-stage optimization

$$\inf\{F \in \mathbb{R} : \ \inf\{h' \in \mathbb{H}^I, \ b' \in \mathbb{B}^F : \ E(b', \ h')\}\}$$

which indeed aims at the lowest error in constitutive law, given I.

6.2. Flux × mmf = energy, so all components of $\boldsymbol{\lambda}$ are magnetomotive forces. One of them is the driving mmf I applied between $S^h{}_0$ and $S^h{}_1$. All others are associated with faces which pave S^b, and assume the exact value necessary to cancel the induction flux through each of these faces.

6.3. Cf. Fig. 6.9.

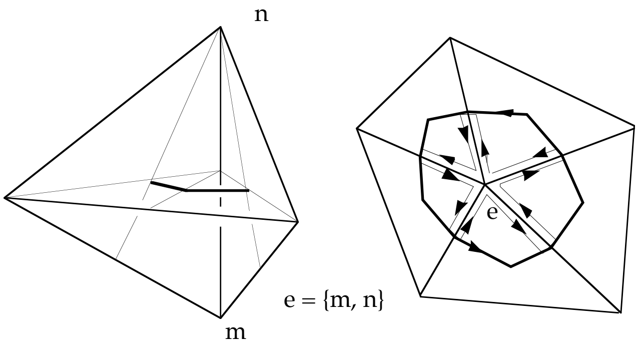


FIGURE 6.9. Same circulation of h along the equatorial circuit joining the barycenters of faces and volumes around e (i.e., the boundary of the dual cell e*, cf. Fig. 4.4) and along the star-shaped circuit that radiates from edge e to the centers of faces containing e. The latter circulation is equal to $\int h \cdot \mathrm{rot}\, w_e$.

6.4. Elements of W^1 are divergence-free inside tetrahedra, whereas $\mathrm{div}\, u = \sum_{n \in \mathcal{N}} \underline{\mathbf{u}}_n \cdot \nabla w_n$ has no reason to vanish.

6.6. By Stokes, $n \cdot \mathrm{rot}\, u$ at point x is the limit $(\int_\gamma \tau \cdot u)/\mathrm{area}(\omega)$, where γ is the boundary of a shrinking surface domain ω enclosing x (see the inset in the "hints" section). Since $\tau \cdot u = \tau \cdot u_S$, only the tangential part u_S is involved. (Remark that only the orientation of γ matters here. The normal field serves to provide it, in association with the orientation of ambient space.) Note that $\nu = - n \times \tau$ is a surface field of outward unit normals with respect to ω. Since $\tau \cdot u_S = n \times \nu \cdot u_S = - \nu \cdot \ n \times u_S$, the circulation $\int_\gamma \tau \cdot u$ is also the outgoing flux (*along* the surface) of $- n \times u_S$, which suggests to define the surface divergence as

$$\mathrm{div}_S u_S = - \mathrm{rot}_S(n \times u), \tag{35}$$

also a surface operator by the same argument. The formula

$$\int_S u_S \cdot \mathrm{grad}_S\, \varphi_S = -\int_S \varphi_S\, \mathrm{div}_S u_S + \int_{\partial S} \varphi_S\, \nu \cdot u_S$$

(where ν now refers to the rim of S), is proved by the same technique as in dimension 3, but (35) suggests to also introduce a rot operator acting on *scalar* surface fields, as follows:

$$\mathrm{rot}_S\, \varphi_S = -\, n \times \mathrm{grad}_S\, \varphi_S,$$

hence the formula

$$\int_S \varphi_S\, \mathrm{rot}_S u_S = \int_S u_S \cdot \mathrm{rot}_S \varphi_S - \int_{\partial S} \varphi_S\, \tau \cdot u_S,$$

the nice symmetry of which compensates for the slight inconvenience[8] of overloading the symbol rot_S.

REFERENCES and Bibliographical Comments

The search for complementarity, a standing concern in computational electromagnetism [HP, HT], is related to the old "hypercircle" idea of Prager and Synge [Sy]. In very general terms, this method consisted in partitioning the set of equations and boundary conditions into two parts, thus defining two orthogonal subsets in the solution space, the solution being at their intersection. Finding two approximations within each of these subsets allowed one (by the equivalent of the Pythagoras theorem in the infinite dimensional solution space, as was done above in 6.3.1) to find the center and radius of a "hypercircle" containing the unknown solution, and hence the bounds (not only on quadratic quantities such as the reluctance, but on linear functionals and, even better, on pointwise quantities [Gr]). After a period of keen interest, the idea was partly forgotten, then revisited or rediscovered by several authors [DM, HP, LL, Ny, R&, Va, . . .], Noble in particular [No], who is credited for it by some (cf. Rall [Ra], or Arthurs [An], who also devoted a book to the subject [Ar]). Thanks to the Whitney elements technology, we may nowadays adopt a different (more symmetrical) partitioning of equations than the one performed by Synge on the Laplace problem. (The one exposed here was first proposed in [B1].) On pointwise estimates, which seem to stem from Friedrichs [Fr], see [Ba], [Co], [Ma], [St], [Sn].

[8] The risk of confusion, not to be lightly dismissed, will be alleviated by careful definition of the types of the fields involved: The first rot is *VECTOR* → *SCALAR* (fields), the other one is *SCALAR* → *VECTOR*. Various devices have been proposed to stress the distinction, including the opposition rot vs Rot, but they don't seem to make things more mnemonic. A. Di Carlo has proposed to denote the second operator by "grot", which would solve this terminological difficulty.

[A&] R. Albanese, R. Fresa, R. Martone, G. Rubinacci: "An Error Based Approach to the Solution of Full Maxwell Equations", **IEEE Trans., MAG-30**, 5 (1994), pp. 2969–2971.

[AE] P. Alfeld, D.J. Eyre: "The Exact Analysis of Sparse Rectangular Linear Systems", **ACM TOMS, 17,** 4 (1991), pp. 502–518.

[An] N. Anderson, A.M. Arthurs, P.D. Robinson: "Pairs of Complementary Variational Principles", **J. Inst. Math. & Applications, 5** (1969), pp. 422–431.

[Ar] A.M. Arthurs: **Complementary Variational Principles,** Clarendon Press (Oxford), 1970.

[Br] M.L. Barton: **Tangentially Continuous Vector Finite Elements for Non-linear 3-D Magnetic Field Problems**, Ph.D. Thesis, Carnegie-Mellon University (Pittsburgh), 1987.

[BC] M.L. Barton, Z.J. Cendes: "New vector finite elements for three dimensional magnetic fields computations", **J. Appl. Phys., 61**, 8 (1987), pp. 3919–3921.

[Ba] N.M. Basu: "On an application of the new methods of the calculus of variations to some problems in the theory of elasticity", **Phil. Mag., 7,** 10 (1930), pp. 886–896.

[B1] A. Bossavit: "Bilateral Bounds for Reluctance in Magnetostatics", in **Numerical Field Calculation in Electrical Engineering** (Proc. 3d Int. IGTE Symp.), IGTE (26 Kopernikusgasse, Graz, Austria), 1988, pp. 151–156.

[B2] A. Bossavit: "A Numerical Approach to Transient 3D Non-linear Eddy-current Problems", **Int. J. Applied Electromagnetics in Materials**, **1**, 1 (1990), pp. 65–75.

[B3] A. Bossavit: "Complementarity in Non-Linear Magnetostatics: Bilateral Bounds on the Flux-Current Characteristic", **COMPEL, 11**, 1 (1992), pp. 9–12.

[B4] A. Bossavit: "A new rationale for edge-elements", **Int. Compumag Society Newsletter, 1,** 3 (1995), pp. 3–6.

[Bw] K. Bowden: "On general physical systems theories", **Int. J. General Systems, 18** (1990), pp. 61–79.

[Co] Ph. Cooperman: "An extension of the method of Trefftz for finding local bounds on the solutions of boundary value problems, and on their derivatives", **Quart. Appl. Math., 10,** 4 (1953), pp. 359–373.

[DM] Ph. Destuynder, B. Métivet: "Estimation explicite de l'erreur pour une méthode d'éléments finis non conforme", **C.R. Acad. Sci. Paris, Série I** (1996), pp. 1081–1086.

[DB] D.C. Dibben, R. Metaxas: "A Comparison of the Errors Obtained with Whitney and Linear Edge Elements", **IEEE Trans., MAG-33**, 2 (1997), pp. 1524–1527.

[D&] P. Dular, J.-Y. Hody, A. Nicolet, A. Genon, W. Legros: "Mixed Finite Elements Associated with a Collection of Tetrahedra, Hexahedra and Prisms", **IEEE Trans., MAG-30**, 5 (1994), pp. 2980–2983.

[Fe] W. Fenchel: "On Conjugate Convex Functions", **Canadian J. Math., 1** (1949), pp. 23–27.

[Fr] R.L. Ferrari: "Complementary variational formulations for eddy-current problems using the field variables E and H directly", **IEE Proc**., **Pt. A**, **132**, 4 (1985), pp. 157–164.

[Fr] K.O. Friedrichs: "Ein Verfahren der Variationsrechnung, das Minimum eines Integrals als das Maximum eines Anderen Ausdruckes darzustellen", **Nachrichten der Ges. d. Wiss. zu Göttingen** (1929), p. 212.

[Gr] H.J. Greenberg: "The determination of upper and lower bounds for the solution of the Dirichlet problem", **J. Math. Phys., 27** (1948), pp. 161–182.

[HP] P. Hammond, J. Penman: "Calculation of inductance and capacitance by means of dual energy principles", **Proc. IEE, 123**, 6 (1976), pp. 554–559.

[HT] P. Hammond, T.D. Tsiboukis: "Dual finite-elements calculations for static electric and magnetic fields", **IEE Proc., 130, Pt. A,** 3 (1983), pp. 105–111.

[Ja] D.A.H. Jacobs: "Generalization of the Conjugate Gradient Method for Solving Non-symmetric and Complex Systems of Algebraic Equations", CEGB Report RD/L/N70/80 (1980).

[Ke] L. Kettunen, K. Forsman, D. Levine, W. Gropp: "Volume integral equations in nonlinear 3-D magnetostatics", **Int. J. Numer. Meth. Engng., 38** (1995), pp. 2655–2675.

[Ko] P.R. Kotiuga: "Analysis of finite-element matrices arising from discretizations of helicity functionals", **J. Appl. Phys., 67**, 9 (1990), pp. 5815–5817.

[LL] P. Ladévèze, D. Leguillon: "Error estimate procedure in the finite element method and applications", **SIAM J. Numer. Anal., 20,** 3 (1983), pp. 485–509.

[Ma] C.G. Maple: "The Dirichlet problem: Bounds at a point for the solution and its derivative", **Quart. Appl. Math., 8,** 3 (1950), pp. 213–228.

[Mk] P. Monk: "A finite element method for approximating the time-harmonic Maxwell equations", **Numer. Math., 63** (1992), pp. 243–261.

[MS] P. Monk, E. Süli: "A convergence analysis of Yee's scheme on nonuniform grids", **SIAM J. Numer. Anal., 31**, 2 (1994), pp. 393–412.

[Mo] J.J. Moreau: **Fonctionnelles convexes**, Séminaire Leray, Collège de France, Paris (1966).

[Ny] B. Nayroles: "Quelques applications variationnelles de la théorie des fonctions duales à la mécanique des solides", **J. de Mécanique, 10**, 2 (1971), pp. 263–289.

[Ne] J.C. Nedelec: "A new family of mixed finite elements in $\mathbb{R}^3$ ", **Numer. Math., 50** (1986), pp. 57–81.

[No] B. Noble: "Complementary variational principles for boundary value problems I: Basic principles with an application to ordinary differential equations", MRC Technical Summary Report N° 473, Nov. 1964.

[OR] J.T. Oden, J.N. Reddy: "On Dual Complementary Variational Principles in Mathematical Physics", **Int. J. Engng. Sci., 12** (1974), pp. 1–29.

[P&] J. Parker, R.D. Ferraro, P.C. Liewer: "Comparing 3D Finite Element Formulations Modeling Scattering from a Conducting Sphere", **IEEE Trans., MAG-29**, 2 (1993), pp. 1646–1649.

[PF] J. Penman, J.R. Fraser: "Dual and Complementary Energy Methods in Electromagnetism", **IEEE Trans., MAG-19**, 6 (1983), pp. 2311–2316.

[PT] Y. Perréal, P. Trouvé: "Les méthodes variationnelles pour l'analyse et l'approximation des équations de la physique mathématique: Les méthodes mixtes-hybrides conservatives", **Revue Technique Thomson-CSF, 23**, 2 (1991), pp. 391–468.

[Ra] L.B. Rall: "On Complementary Variational Principles", **J. Math. Anal. & Appl., 14** (1966), pp. 174–184.

[Re] Z. Ren, in A. Bossavit, P. Chaussecourte (eds.): **The TEAM Workshop in Aix-les-Bains,** July 7–8 1994, EdF, Dpt MMN (1 Av. Gal de Gaulle, 92141 Clamart), 1994.

[R§] J. Rikabi, C.F. Bryant, E.M. Freeman: "An Error-Based Approach to Complementary Formulations of Static Field Solutions", **Int. J. Numer. Meth. Engnrg., 26** (1988), pp. 1963–1987.

[Ro] R.T. Rockafellar: **Convex Analysis**, Princeton U.P. (Princeton), 1970.

[Rt] J.P. Roth: "An application of algebraic topology: Kron's method of tearing", **Quart. Appl. Math., 17,** 1 (1959), pp. 1–24.

[St] H. Stumpf: "Über punktweise Eingrenzung in der Elastizitätstheorie. I", **Bull. Acad. Polonaise Sc., Sér. Sc. Techniques, 16,** 7 (1968), pp. 329–344, 569–584.

[Sn] J.L. Synge: "Pointwise bounds for the solutions of certain boundary-value problems", **Proc. Roy. Soc. London, A 208** (1951), pp. 170–175.

[Sy] J.L. Synge: **The Hypercircle in Mathematical Physics,** Cambridge University Press (Cambridge, UK), 1957.

[To] E. Tonti: "On the mathematical structure of a large class of physical theories", **Lincei, Rend. Sc. fis. mat. e nat., 52,** 1 (1972), pp. 51–56.

[Ts] I.A. Tsukerman: "Node and Edge Element Approximation of Discontinuous Fields and Potentials", **IEEE Trans., MAG-29,** 6 (1993), pp. 2368–2370.

[Va] M.N. Vaĭnberg: **Variational Methods for the Study of Nonlinear Operator Equations,** Holden Day (San Francisco), 1963. (Russian edition: Moscow, 1956.)

CHAPTER 7

Infinite Domains

In Chapter 2, we managed to reduce the problem to a bounded region. This is not always possible. We shall address here a typical magnetostatics modelling, for which the computational domain is a priori the whole space. This will be an opportunity to introduce the technique of "finite elements and boundary elements in association", which is essential to the treatment of all "open space" problems, static or not, and will be applied to eddy-currents modelling in Chapter 8.

7.1 ANOTHER MODEL PROBLEM

Figure 7.1 describes the configuration we shall study: an electromagnet, with its load.

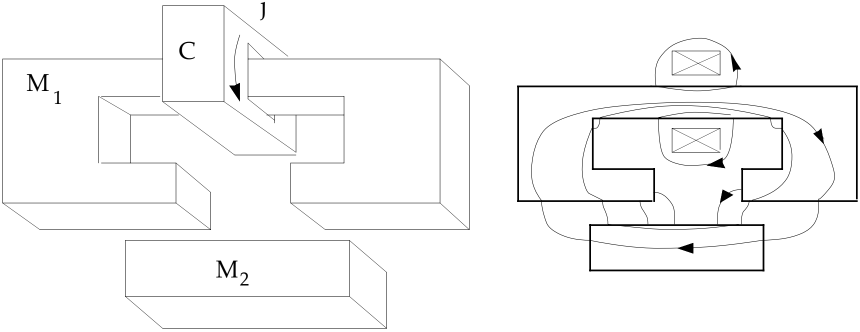


FIGURE 7.1. Left: Electromagnet, with its ferromagnetic core M_1, its coil C, and a load M_2. When a continuous current j is fed into C, the load is attracted upwards. Right: Typical flux lines, in a vertical cross-section.

The working principle of such devices is well known: A direct current creates a permanent magnetic field. Channeled by an almost closed magnetic circuit, the induction flux closes through a ferromagnetic piece (here M_2), to which the load is attached. Force lines (cf. Fig. 7.1, right) "tend to shorten", and the load is thus lifted towards the poles of the electromagnet, upwards in the present case, whatever the sense of the current.

This "shortening of field lines" is of course a very naive explanation, but a useful one all the same, and rigorous analysis does support it, as follows. Suppose for definiteness the electromagnet is fixed, the load M_2 being free to move vertically. Let u denote its position on the vertical axis (oriented upwards). For a given u, some field {h, b} settles, the magnetic coenergy[1] of which is $W_u(h) = \frac{1}{2}\int \mu\ |h|^2$. Thanks to the virtual work principle, one may prove that the lifting force upon M_2 is[2] $f = \partial_u W$. Now, let us verify that $W_u(h)$ *increases* with u (whence the sign of f), by the reasoning that follows.

First, μ is large in M. Let us take it as infinite. In that case, h = 0 in M. By Ampère's theorem, the circulation of h along a field line is equal to the intensity in the coil, call it I. If H is the magnitude of the field in the air gap, of width d, one thus has Hd ~ I, since only the part of the field line contained in the air gap contributes to the circulation. The air gap volume being proportional to d, the coenergy is thus proportional to $\mu_0 H^2\ d/2$, that is $\mu_0\, I^2/2d$, so it increases when d decreases. Hence the direction of the force: upwards, indeed.

This reasoning is quite useful, and moreover, it suggests how to pass from qualitative to quantitative statements: If one is able to compute the field with enough accuracy to plot W as a function of u, one will have the force as a function of u by simple differentiation[3]. So, if the mechanical characteristics of the system are known (masses, inertia tensors, restoring

[1] As already pointed out (Remark 2.6), one should carefully distinguish between a *function* of the field which, by its evaluation, yields the energy, and the numerical *result* of such an evaluation (the real number one refers to when one speaks of the energy stored in the field). Energy can be obtained in two different ways, by evaluating either the function $b \to V(b) \equiv \frac{1}{2}\int \mu^{-1}|b|^2$ or the function $h \to W(h) \equiv \frac{1}{2}\int \mu\ |h|^2$, and to tell them apart, one calls them *energy* and *coenergy function(al)s* respectively. Making this distinction is essential, because the results of both evaluations coincide in the linear case only (and otherwise the correct value of the energy is obtained by computing V, not W). In problems with motion, V and W are parameterized by the configuration of the system, here denoted by u.

[2] Force is also given by $f = -\partial_u V$, as can be seen by differentiating with respect to u the equality $V_u(b_u) + W_u(h_u) = \int b_u \cdot h_u$, which holds when b_u and h_u are the induction and the field that effectively settle in configuration u.

[3] A better, more sophisticated approach, is available [Co], by which $\partial_u W$ instead of W_u is obtained via a finite-element computation. It consists essentially of differentiating the elementary matrices with respect to u *before* proceeding to their assembly, and solving the linear system thus obtained.

forces, etc.), one may study its dynamics. This opens the way to a whole realm of applications: electromagnets, of course, but also electromagnetic "actuators" of various kinds (motors, linear or rotatory, launchers, etc.).

The underlying problem, in all such applications, is thus: Knowing the geometry of the system at hand, and the values of μ, compute the magnetic field, with its coenergy (or its energy) as a by-product. The principal difficulty then comes from the nonlinearity of the b–h relation (to say nothing of hysteresis).

Without really addressing this difficulty, let us only recall that nonlinear problems are generally solved by successive approximations, Newton–Raphson style, and thus imply the solution of a sequence of linear problems. In the present case, each of these problems assumes the form

(1) $\quad \mathrm{rot}\, h = j, \qquad$ (2) $\quad \mathrm{div}\, b = 0,$

(3) $$b = \mu\, h,$$

in all space, with $\mu = \mu_0$ outside M, and μ function of the position x (via the values of the field at this point at the previous iterations) inside M. This points to (1–3), the *magnetostatics model* in all space, as the basic problem.

Apart from this spatial variation of μ, and the explicit presence of the source-term j, the problem is quite alike the "Bath cube" one of Chapter 2. The only really new element is the non-boundedness of the domain, which will allow us to concentrate on that.

7.2 FORMULATION

According to the functional point of view, we look for a precise formulation of (1–3): *find* h *and* b *in . . . such that . . .*, etc. The first item on the agenda is thus to delimit the field of investigation: exactly in which functional space are we looking for b and h ? Once these spaces of "admissible fields" are identified, one realizes that Problem (1–3) has a *variational* formulation, which means it can be expressed in the form *find* h *and* b *which minimize*, etc. All that is left to do is then to replace spaces of admissible fields by "large enough" *finite* dimensional subspaces in order to obtain approximate formulations open to algebraic treatment on a computer. Let us thus try to find this "variational framework", as one says, in which problem (1–3) makes sense.

7.2.1 Functional spaces

First, set $\mathbb{H} = \mathbb{L}^2_{\text{rot}}(E_3)$ and $\mathbb{B} = \mathbb{L}^2_{\text{div}}(E_3)$, and put irrotational and solenoidal fields apart:

(4) $$\mathbb{H}^0 = \{h \in \mathbb{H} : \text{ rot } h = 0\}, \quad \mathbb{B}^0 = \{b \in \mathbb{B} : \text{ div } b = 0\}.$$

Then,

Proposition 7.1. *Subspaces* $\mathbb{H}^0$ *and* $\mathbb{B}^0$ *are ortho-complementary in* $\mathbb{L}^2(E_3)$:

(5) $$\mathbb{L}^2(E_3) = \mathbb{H}^0 \oplus \mathbb{B}^0.$$

In other words, any vector u *of* $\mathbb{L}^2(E_3)$ *can be written, in a unique way, as* $u = h + b$, *with* $\text{rot } h = 0$, $\text{div } b = 0$, *and* $\int_{E_3} h \cdot b = 0$.

Proof. First, let u be a smooth field with bounded support, form div u and rot u, and set

$$\varphi(x) = -\frac{1}{4\pi}\int_{E_3} \frac{(\text{div } u)(y)}{|x-y|}\, dy, \quad a(x) = \frac{1}{4\pi}\int_{E_3} \frac{(\text{rot } u)(y)}{|x-y|}\, dy.$$

Then, differentiating under the integral signs, and applying the formula rot rot = grad div – Δ, one obtains that

(6) $$u = \text{grad } \varphi + \text{rot } a,$$

which is called the *Helmholtz*[4] *decomposition* of u, a standard result. (Beware, neither φ nor a has bounded support.) Setting h = grad φ and b = rot a, one sees that $\int_{E_3} b \cdot h = 0$, as announced. If instead of considering a single field u we look at functional spaces wholesale, (6) is equivalent to

(6') $$\mathbb{C}_0^\infty(E_3) = -\text{ grad}(\text{div}(\mathbb{C}_0^\infty(E_3))) \oplus \text{rot}(\text{rot}(\mathbb{C}_0^\infty(E_3))),$$

which we can write $\mathbb{C}_0^\infty(E_3) = \mathcal{H}^0 \oplus \mathcal{B}^0$, where $\mathcal{H}^0$ and $\mathcal{B}^0$ are subspaces —not *closed* subspaces—of $\mathbb{L}^2(E_3)$ composed of smooth curl-free and divergence-free fields, respectively. Now call $\mathbb{H}^0$ and $\mathbb{B}^0$ the *closures* in $\mathbb{L}^2(E_3)$ of $\mathcal{H}^0$ and $\mathcal{B}^0$. For each pair $h \in \mathbb{H}^0$ and $b \in \mathbb{B}^0$, we thus have sequences of smooth fields $\{h_n \in \mathcal{H}^0 : n \in \mathbb{N}\}$ and $\{b_n \in \mathcal{B}^0 : n \in \mathbb{N}\}$ which converge towards h and b, while satisfying $\text{rot } h_n = 0$, $\text{div } b_n = 0$, and $\int b_n \cdot h_n = 0$. All these properties "pass to the limit" by continuity. For instance (to do it only once), $\text{div } b_n = 0$ means $\int b_n \cdot \text{grad } \varphi = 0$ for all φ in $C_0^\infty(E_3)$, hence, by continuity of the scalar product, $\int b \cdot \text{grad } \varphi = 0$ for all these test functions, which is weak solenoidality, and which we are writing

[4] Due to Stokes (1849), actually, according to [Hu], p. 147.

$\operatorname{div} b = 0$ since we decided to adopt the "weak" interpretation for differential operators in Chapter 5. Last step: Since $\mathbb{C}_0^\infty(E_3)$ is dense in $\mathbb{L}^2(E_3)$ by construction of the latter, $\mathbb{H}^0 \oplus \mathbb{B}^0$ fills out all $\mathbb{L}^2(E_3)$, hence (5). ◊

But can one go further and have (6) for all, not only smooth, square-integrable fields? Yes, if one accepts having φ and a in the right functional spaces, those obtained by completion. Take $h \in \mathbb{H}^0$. By the foregoing, there is a sequence $\varphi_n \in C_0^\infty(E_3)$ such that $\operatorname{grad} \varphi_n$ converges to h. Thus, the sequence $\{\varphi_n\}$ is Cauchy with respect to the norm $\|\varphi\| = [\int_{E_3} |\operatorname{grad} \varphi|^2]^{1/2}$. Now *complete* $C_0^\infty(E_3)$ with respect to this norm: There comes a complete space, which we shall denote Φ, an extension-by-continuity of grad (called, again,[5] the *weak* gradient), and we do have $h = \operatorname{grad} \varphi$, with $\varphi \in \Phi$.

A sleight of hand? Yes, in a sense, since elements of Φ are abstract objects (equivalence classes of Cauchy sequences of functions), but not really, because one can identify this space Φ with a subspace[6] of a functional space, the Sobolev space $L^6(E_3)$. This is an immediate consequence of the inequality $\int_{E_3} |\varphi|^6 \leq C \int_{E_3} |\operatorname{grad} \varphi|^2$, a proof of which (not simple) can be found in [Br], p. 162. So as regards irrotational fields, we have $\mathbb{H}^0 = \operatorname{grad} \Phi$, where Φ is a well-defined functional space. This is the *Beppo Levi space* alluded to in Chapter 3, Note 5.

The representation (6) of a field u, now with $\varphi \in \Phi$ and $a \in A$, extends the Helmholtz decomposition, just as the Poincaré lemma was extended in Chapter 5. (Note that a is not unique, but φ is, because $\|\varphi\|$ is a norm, since there are no constants in $C_0^\infty(E_3)$.)

Let now $j \in \mathbb{L}^2(E_3)$ be given, with bounded support, and $\operatorname{div} j = 0$. The field $h^j = \operatorname{rot} a^j$, where a^j is given by the integral

$$a^j(x) = \frac{1}{4\pi} \int_{E_3} \frac{j(y)}{|x-y|} \, dy,$$

is in[7] $\mathbb{L}^2_{\text{rot}}(E_3)$ and satisfies $\operatorname{rot} h^j = j$. Let us set $\mathbb{H}^j = h^j + \mathbb{H}^0$.

[5] In spite of its domain being larger than the closure of the strong gradient in $L^2(E_3) \times \mathbb{L}^2(E_3)$. No Poincaré inequality is available in the present case; hence the two methods of extension of the differential operators examined in Subsection 5.1.2 are no longer equivalent.

[6] There is a more precise characterization of Φ as the space of functions φ such that $\operatorname{grad} \varphi \in \mathbb{L}^2(E_3)$ and $\int (1 + |x|^2)^{-1/2} |\varphi(x)|^2 \, dx < \infty$.

[7] This is not totally obvious, and j having a bounded support (or at least, a support of finite volume) plays a decisive role there. (Show that $\operatorname{rot} \operatorname{rot} a^j = j$, then compute $\int \operatorname{rot} a^j \cdot \operatorname{rot} a^j = \int \operatorname{rot} \operatorname{rot} a^j \cdot a^j$, etc.)

7.2.2 Variational formulations

Now we have all the ingredients required to make the problem "well posed":

Proposition 7.2. *Let there be given* $\mathrm{j} \in \mathbb{L}^2(\mathrm{E}_3)$*, with bounded support and* $\operatorname{div} \mathrm{j} = 0$*, and a function* μ *such that* $\mu_1 \geq \mu(\mathrm{x}) \geq \mu_0$ *a.e. in* E_3*. The problem* find $\mathrm{h} \in \mathbb{H}^{\mathrm{j}}$ and $\mathrm{b} \in \mathbb{B}^0$ such that[8]

$$(7) \qquad \int_{E_3} \mu^{-1} \, |\mathrm{b} - \mu \mathrm{h}|^2 \leq \int_{E_3} \mu^{-1} \, |\mathrm{b}' - \mu \mathrm{h}'|^2 \quad \forall\, \mathrm{h}' \in \mathbb{H}^{\mathrm{j}}, \ \forall\, \mathrm{b}' \in \mathbb{B}^0,$$

has a unique solution, which satisfies (1–3).

Proof. If $\mathrm{h}' \in \mathbb{H}^{\mathrm{j}}$, then $\mathrm{h}' - \mathrm{h}^{\mathrm{j}} \in \mathbb{H}^0$. Thus, $\int_{E_3} \mathrm{h}' \cdot \mathrm{b}' = \int_{E_3} \mathrm{h}^{\mathrm{j}} \cdot \mathrm{b}'$ for each pair $\{\mathrm{h}', \mathrm{b}'\} \in \mathbb{H}^{\mathrm{j}} \times \mathbb{B}^0$, after (5), and therefore,

$$\int_{E_3} \mu^{-1} \, |\mathrm{b}' - \mu\, \mathrm{h}'|^2 = \int_{E_3} \mu^{-1} \, |\mathrm{b}'|^2 + \int_{E_3} \mu \, |\mathrm{h}'|^2 - 2 \int_{E_3} \mathrm{h}^{\mathrm{j}} \cdot \mathrm{b}',$$

so that Problem (7) is equivalent to the following pair of *independent* optimization problems, taken together:

$$(8) \qquad \textit{find } \mathrm{h} \in \mathbb{H}^{\mathrm{j}} \textit{ such that } \ W(\mathrm{h}) \leq W(\mathrm{h}') \quad \forall\, \mathrm{h}' \in \mathbb{H}^{\mathrm{j}},$$

where $W(\mathrm{h}) = \frac{1}{2} \int_{E_3} \mu \, |\mathrm{h}|^2$ is the magnetic coenergy introduced earlier, and

$$\textit{find } \mathrm{b} \in \mathbb{B}^0 \textit{ such that}$$

$$(9) \qquad V(\mathrm{b}) - \int_{E_3} \mathrm{h}^{\mathrm{j}} \cdot \mathrm{b} \leq V(\mathrm{b}') - \int_{E_3} \mathrm{h}^{\mathrm{j}} \cdot \mathrm{b}' \quad \forall\, \mathrm{b}' \in \mathbb{B}^0,$$

where $V(\mathrm{b}) = \frac{1}{2} \int_{E_3} \mu^{-1} \, |\mathrm{b}|^2$ is the magnetic energy. These functionals being continuous on $\mathbb{L}^2(\mathrm{E}_3)$, with coercive[9] quadratic parts, both (8) and (9) have a unique solution. The Euler equations of (8) and (9) being

$$(10) \qquad \textit{find } \mathrm{h} \in \mathbb{H}^{\mathrm{j}} \textit{ such that } \int_{E_3} \mu\, \mathrm{h} \cdot \mathrm{h}' = 0 \quad \forall\, \mathrm{h}' \in \mathbb{H}^0,$$

$$(11) \qquad \textit{find } \mathrm{b} \in \mathbb{B}^0 \textit{ such that } \int_{E_3} \mu^{-1} \mathrm{b} \cdot \mathrm{b}' = \int_{E_3} \mathrm{h}^{\mathrm{j}} \cdot \mathrm{b}' \quad \forall\, \mathrm{b}' \in \mathbb{B}^0,$$

the pair $\{\mathrm{h}, \mathrm{b}\}$ thus found satisfies $\int_{E_3} \mu \, |\mathrm{h}|^2 = \int_{E_3} \mu\, \mathrm{h} \cdot \mathrm{h}^{\mathrm{j}}$ and $\int_{E_3} \mu^{-1} \, |\mathrm{b}|^2 =$

[8] The quantity on the right-hand side of (7) is again the "error in constitutive law" of Chapter 6. It measures the failure of the pair $\{\mathrm{h}', \mathrm{b}'\}$ to satisfy the behavior law $\mathrm{b}' = \mu \mathrm{h}'$, and to minimize it over $\mathbb{H}^{\mathrm{j}} \times \mathbb{B}^0$ amounts to looking, among the pairs which obey other equations (here, $\operatorname{rot} \mathrm{h} = \mathrm{j}$ and $\operatorname{div} \mathrm{b} = 0$), for the one that, by minimizing the error (and actually, cancelling it), best obeys (and actually, exactly obeys) the constitutive law.

[9] This is said (cf. A.4.3) of a quadratic functional $\mathrm{u} \rightarrow (\mathrm{Au}, \mathrm{u})$ over a Hilbert space U for which exists $\alpha > 0$ such that $(\mathrm{Au}, \mathrm{u}) \geq \alpha \, ||\mathrm{u}||^2$ for all u. By the Lax–Milgram lemma, the equation $\mathrm{Au} = \mathrm{f}$ has then a unique solution, which is the minimizer of the functional $\mathrm{u} \rightarrow \frac{1}{2} (\mathrm{Au}, \mathrm{u}) - (\mathrm{f}, \mathrm{u})$.

$\int_{E_3} h^j \cdot b$ (set $h' = h - h^j$ and $b' = b$), whence

$$\int_{E_3} \mu^{-1} |b - \mu h|^2 = \int_{E_3} \mu^{-1} |b|^2 + \int_{E_3} \mu \, |h|^2 - 2\int_{E_3} h^j \cdot b = 0,$$

and therefore $b = \mu h$. ◊

Problems (10) and (11) are of the now-familiar kind of "constrained linear problems", and it is natural to try to solve them via "unconstrained" representations of the affine spaces $\mathbb{H}^j$ and $\mathbb{B}^0$. Since, as we saw earlier,

(12) $\mathbb{H}^j = h^j + \text{grad } \Phi, \qquad \mathbb{B}^0 = \text{rot } A,$

(8) and (9) amount to

(13) *find* $\varphi \in \Phi$ *such that* $W(\varphi) \leq W(\varphi') \quad \forall \varphi' \in \Phi,$

where $W(\varphi) = \frac{1}{2} \int_{E_3} \mu \, |h^j + \text{grad } \varphi|^2$ (magnetic coenergy again, but now considered as a function of φ, as signaled by the slight notational variation), and

(14) *find* $a \in A$ *such that* $V(a) - \int_{E_3} j \cdot a \leq V(a') - \int_{E_3} j \cdot a' \quad \forall a' \in A,$

where $V(a) = \frac{1}{2} \int_{E_3} \mu^{-1} \, |\text{rot } a|^2$ (magnetic energy, as a function of a). The Euler equations of (13) and (14) are

(15) *find* $\varphi \in \Phi$ *such that* $\int_{E_3} \mu \, (h^j + \text{grad } \varphi) \cdot \text{grad } \varphi' = 0 \quad \forall \varphi' \in \Phi,$

(16) *find* $a \in A$ *such that* $\int_{E_3} \mu^{-1} \text{rot } a \cdot \text{rot } a' = \int_{E_3} j \cdot a' \quad \forall a' \in A.$

Being assured of existence and uniqueness for h and b solutions of (8) and (9), we know that (15) has a unique solution φ (the only $\varphi \in \Phi$ such that $h^j + \text{grad } \varphi = h$) and that (16) has a family of solutions all of which verify $\text{rot } a = b$. One might as well, of course, study (15) and (16) ab initio: The mapping $\varphi' \rightarrow \int_{E_3} \mu \, h^j \cdot \text{grad } \varphi'$ is continuous on Φ (since h^j is $\mathbb{L}^2$, and μ is bounded), therefore (15) has a unique solution, by the Lax–Milgram lemma. (For (16), it's less direct, because one must apply the lemma to the *quotient* of A by the kernel of rot.)

7.3 DISCRETIZATION

Whatever the selected formulation, the problem concerns the whole space, which cannot be meshed with a *finite* number of bounded elements. There are essentially three ways to deal with this difficulty. I shall be very brief and allusive about the first two, but this should not imply that they are less important.

7.3.1 First method: "artificial boundary, at a distance"

Start from a mesh $\{\mathcal{N}, \mathcal{E}, \mathcal{F}, \mathcal{T}\}$ of a bounded region of space $\hat{D}$, bounded by surface $\hat{S}$. Denote by $\mathcal{N}(\hat{S})$, etc., as before, the sets of nodes, etc., contained in $\hat{S}$. Then consider a smooth injective mapping u, called a *placement* of $\hat{D}$ in E_3, built in such a way that the image $D = u(\hat{D})$, bounded by $S = u(\hat{S})$, cover the region of interest as well as "enough" space around it (Fig. 7.2). Define "u-adapted" Whitney elements by setting ${}^u w_n(x) = w_n(u^{-1}(x))$ for node n, then ${}^u w_e(x) = {}^u w_m(x)\ \nabla\ {}^u w_n(x) - {}^u w_n(x)\ \nabla\, {}^u w_m(x)$ for edge $e = \{m, n\}$, etc.[10] Then set

$$W^0_{\ m} = \{\varphi : \ \varphi = \textstyle\sum_{n \in \mathcal{N} - \mathcal{N}(\hat{S})} \boldsymbol{\varphi}_n\, {}^u w_n\},$$

$$W^1_{\ m} = \{h : \ h = \textstyle\sum_{e \in \mathcal{E} - \mathcal{E}(\hat{S})} \mathbf{h}_e\, {}^u w_e\},$$

etc., where the $\boldsymbol{\varphi}_n$, $\mathbf{h}_e$, etc., assume real values. (The notation *m* now refers to both the mesh and its placement.)

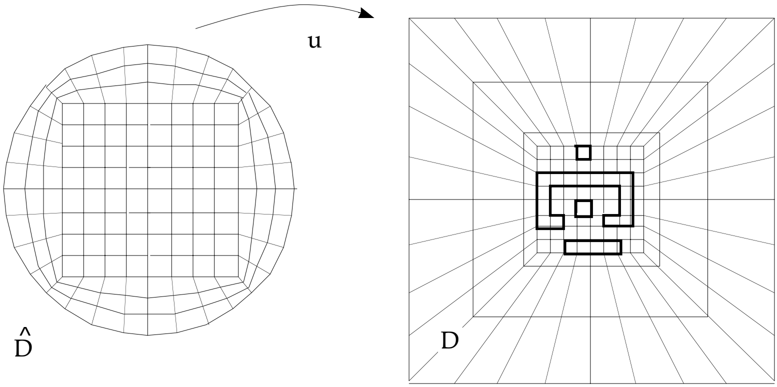


FIGURE 7.2. Illustrating the idea of "placement" of a reference mesh, for a problem similar to the one of Fig. 7.1. What we see is actually a 3D "macro-mesh", the "bricks" of which, once placed in order to accommodate the material interfaces, will be subdivided as required.

[10] This is the precise definition of finite elements on "curved tetrahedra", informally introduced in Chapter 3. In all generality, ${}^u f$ defined by ${}^u f(x) = f(u(x))$ is the *push-forward* of f by u, and f is the *pull-back* of ${}^u f$. So ${}^u w_n$ comes from w_n by push-forward.

With only tiny variations, we may carry on with the former notational system: $\boldsymbol{\varphi}$, $\mathbf{h}$, etc., are the vectors of degrees of freedom, $\boldsymbol{\Phi} = \mathrm{I\!R}^{\mathcal{N} - \mathcal{N}(\hat{\mathrm{S}})}$, $\mathbf{A} = \mathrm{I\!R}^{\mathcal{E} - \mathcal{E}(\hat{\mathrm{S}})}$, etc., are the spaces they generate (isomorphic to $\mathrm{W}^0{}_m$, $\mathrm{W}^1{}_m$, etc.), and the incidence matrices $\mathbf{G}$, etc., are such that $\mathbf{h} = \mathbf{G}\boldsymbol{\varphi}$ if $\mathrm{h} = \mathrm{grad}\ \varphi$, etc. Observe that $\mathrm{W}^0{}_m \subset \Phi$ and $\mathrm{W}^1{}_m \subset \mathrm{A}$, thanks to our having set to zero the degrees of freedom of surface nodes and edges. The intersections $\Phi_m = \mathrm{W}^0{}_m \cap \Phi$ and $\mathrm{A}_m = \mathrm{W}^1{}_m \cap \mathrm{A}$ are thus Galerkin approximation subspaces for Φ and A. The approximations of (15) and (16) determined by m and u are then

(17) *find* $\varphi_m \in \Phi_m$ *such that* $\int_{\mathrm{D}} \mu\,(\mathrm{h}^{\mathrm{j}} + \mathrm{grad}\ \varphi_m) \cdot \mathrm{grad}\ \varphi' = 0 \quad \forall\, \varphi' \in \Phi_m$,

(18) *find* $\mathrm{a}_m \in \mathrm{A}_m$ *such that* $\int_{\mathrm{D}} \mu^{-1}\,\mathrm{rot}\ \mathrm{a}_m \cdot \mathrm{rot}\ \mathrm{a}' = \int_{\mathrm{D}} \mathrm{j} \cdot \mathrm{a}' \quad \forall\, \mathrm{a}' \in \mathrm{A}_m$.

In order to set these linear systems in standard form, let us redefine the "mass matrices" of Chapter 5 as follows (p is the dimension of the simplices):

$$(\mathbf{M}_{\mathrm{p}}(\alpha))_{\mathrm{s\,s'}} = \int_{\mathrm{D}} \alpha\ {}^{\mathrm{u}}\mathrm{w}_{\mathrm{s}} \cdot {}^{\mathrm{u}}\mathrm{w}_{\mathrm{s'}} \quad \text{if } \mathrm{p} = 1 \text{ or } 2,$$

$$= \int_{\mathrm{D}} \alpha\ {}^{\mathrm{u}}\mathrm{w}_{\mathrm{s}}\ {}^{\mathrm{u}}\mathrm{w}_{\mathrm{s'}} \quad \text{if } \mathrm{p} = 0 \text{ or } 3,$$

the indices s and $\mathrm{s'}$ being restricted to the sets of "internal" simplices (those not in $\hat{\mathrm{S}}$). Then (17) and (18) can be rewritten as

(19) $\mathbf{G}^{\mathrm{t}}\mathbf{M}_1(\mu)\,(\mathbf{G}\boldsymbol{\varphi} + \mathbf{h}^{\mathrm{j}}) = 0,$ (20) $\mathbf{R}^{\mathrm{t}}\mathbf{M}_2(\mu^{-1})\mathbf{R}\ \mathbf{a} = \mathbf{j},$

where vectors $\mathbf{h}^{\mathrm{j}}$ and $\mathbf{j}$ are defined by $\mathbf{h}^{\mathrm{j}}{}_{\mathrm{e}} = \int_{\mathrm{e}} \tau \cdot \mathrm{h}^{\mathrm{j}}$ and $\mathbf{j}_{\mathrm{f}} = \int_{\mathrm{f}} \mathrm{n} \cdot \mathrm{j}$.

7.3.2 Second method: "infinite elements"

This method makes use of a more sophisticated placement u, but otherwise coincides with the first one. The difference is that here u maps $\hat{\mathrm{D}}$ onto the *whole* space E_3, all points of the boundary $\hat{\mathrm{S}}$ being sent to infinity. The elements of the mesh which are immediately under $\hat{\mathrm{S}}$ are then sent onto regions of infinite volume, called infinite elements. See, e.g., [BM] for a construction of such a mapping. It is often convenient, in this respect, to use a geometric inversion with respect to some point [IM, LS].

The literature on infinite elements is huge. For a bibliography and a comparative study, from a practical viewpoint, cf. C. Emson's contribution (in English . . .) to [B &].

7.3.3 Third method: "finite elements and integral method in association"

There, in contrast, D is made as small as possible, including the region of interest while still having a boundary of simple shape. Applying the first method would then amount to neglecting the field outside D, which is not acceptable. However, the far field is not of primary interest by itself: All that matters is its contribution to the energy or the coenergy, as the following informal approach will suggest.

Let D, bounded by surface S, be such that $\mu \equiv \mu_0$ outside D. Then[11]

$$\inf\{W(\mathrm{h}^{\mathrm{j}} + \operatorname{grad}\varphi) :\ \varphi \in \Phi\}$$
$$= \inf\{\tfrac{1}{2}\textstyle\int_{\mathrm{D}} \mu\ |\,\mathrm{h}^{\mathrm{j}} + \operatorname{grad}\varphi\,|^2 + W_{\mathrm{ext}}(\mathrm{j}, \varphi_{|\mathrm{S}}) :\ \varphi \in \Phi\},$$

where, by way of definition,

$$\text{(21)} \qquad \mathrm{W}_{\mathrm{ext}}(\mathrm{j}, \varphi_{\mathrm{S}}) = \inf\{\tfrac{1}{2}\textstyle\int_{\mathrm{E}_3 - \mathrm{D}} \mu_0\ |\,\mathrm{h}^{\mathrm{j}} + \operatorname{grad}\varphi\,|^2 :\ \varphi \in \Phi,\ \varphi_{|\mathrm{S}} = \varphi_{\mathrm{S}}\}.$$

This "exterior coenergy" term only depends on the *boundary* values of the potential φ. So, in order to be able to solve Problem (15) by meshing region D only, it would be sufficient to know some approximation of $W_{\mathrm{ext}}(\mathrm{j}, \varphi_{\mathrm{S}})$ as a function of nodal values of φ on S (which are *not* set to 0 here, in contrast with the first method). For the same reason, having an approximation of the functional $\mathrm{V}_{\mathrm{ext}}(\mathrm{a}) - \int_{\mathrm{E}_3 - \mathrm{D}} \mathrm{j} \cdot \mathrm{a}$ (where $\mathrm{V}_{\mathrm{ext}}(\mathrm{a}) = \frac{1}{2}\int_{\mathrm{E}_3 - \mathrm{D}} \mu_0^{-1}\,|\operatorname{rot}\mathrm{a}\,|^2$) in terms of the circulations of a along edges of S would allow one to solve (16) without having to mesh the outer region.

Now let φ be the potential (outside D) that minimizes (21). One has $\varphi_{|\mathrm{S}} = \varphi_{\mathrm{S}}$ and

$$\textstyle\int_{\mathrm{E}_3 - \mathrm{D}} \mu_0\ (\mathrm{h}^{\mathrm{j}} + \operatorname{grad}\varphi) \cdot \operatorname{grad}\varphi' = 0 \quad \forall\, \varphi' \in \Phi \ \text{ such that } \ \varphi'_{|\mathrm{S}} = 0.$$

So $\operatorname{div}(\mathrm{h}^{\mathrm{j}} + \operatorname{grad}\varphi) = 0$ outside D. Since $\operatorname{div}\mathrm{h}^{\mathrm{j}} = 0$, by construction, $\Delta\varphi = 0$ outside D, thus φ is solution to the "exterior Dirichlet problem":

$$\Delta\varphi = 0 \ \text{ outside D}, \quad \varphi_{|\mathrm{S}} = \varphi_{\mathrm{S}}.$$

On the other hand, thanks to the integration by parts formula,

$$\mathrm{W}_{\mathrm{ext}}(\mathrm{j}, \varphi_{\mathrm{S}}) = \tfrac{1}{2}\textstyle\int_{\mathrm{E}_3 - \mathrm{D}} \mu_0\ |\,\mathrm{h}^{\mathrm{j}} + \operatorname{grad}\varphi\,|^2$$
$$= \tfrac{1}{2}\textstyle\int_{\mathrm{E}_3 - \mathrm{D}} \mu_0\ |\,\mathrm{h}^{\mathrm{j}}|^2 + \tfrac{1}{2}\ \mu_0 \int_{\mathrm{S}}\ \mathrm{n} \cdot (2\,\mathrm{h}^{\mathrm{j}} + \operatorname{grad}\varphi)\ \varphi_{\mathrm{S}}$$

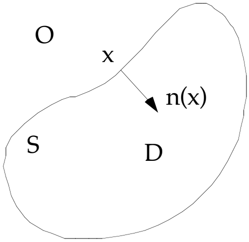
[11] For a while, we need to distinguish $\varphi_{|\mathrm{S}}$, the restriction of φ to S, and φ_{S}, which will denote some function defined on S.

$$= \tfrac{1}{2}\int_{E_3 - D} \mu_0 \,|h^j|^2 + \tfrac{1}{2}\,\mu_0 \int_S (2\,n \cdot h^j + P\varphi_S)\;\varphi_S\,,$$

where the normal n is oriented towards D (beware!) and P the operator that maps φ_S to the normal derivative $n \cdot \operatorname{grad} \varphi$. So the problem would be solved if P was known, or at least if one could compute an approximation of $\int_S P\,\varphi_S\,\varphi_S$ in terms of the DoF $\boldsymbol{\varphi}_n$, where n spans $\mathcal{N}(S)$. We shall therefore look for a matrix approximation of this "Dirichlet-to-Neumann" operator (also called "Poincaré–Steklov" operator [AL], or else "capacity" [DL], because of its interpretation in electrostatics). This operator is a quite delicate object to handle, and we'll have to spend some time on its precise definition and its properties. The reader who feels the foregoing overview was enough (despite, or perhaps thanks to, many abuses) may skip what follows and proceed to Subsection 7.4.4.

7.4 THE "DIRICHLET-TO-NEUMANN" MAP

So let D be a regular bounded domain, inside a closed surface S. We shall denote by O (for "outside") the complement of $D \cup S$, which is also the interior of $E_3 - D$. Domains D and O have S as common boundary, and the field of normals to S is taken as outgoing from O (not from D).



7.4.1 The functional space of "traces"

The theory, unfortunately, is more demanding than anything we have done up to now, and the time has come to introduce something which could be evaded till this point: *traces* of functions in L^2_{grad}, and the Sobolev space of traces, $H^{1/2}(S)$.

Smooth functions φ over D have restrictions $\varphi_{|S}$ to S, which are piecewise smooth functions. Let us denote by γ the mapping $\varphi \to \varphi_{|S}$. We shall base our approach on the following lemma:

Lemma 7.1. *There exists a constant* $C(D)$, *depending only on* D, *such that, for all functions* φ *smooth over* D,

$$(22) \qquad \int_S |\gamma\,\varphi|^2 \le C(D)\;\left[\int_D |\varphi|^2 + \int_D |\operatorname{grad} \varphi|^2\right].$$

This is technical, but not overly difficult if one accepts cutting a few corners, and Exer. 7.6 will suggest an approach to this result. Our purpose is to extend to $L^2_{\text{grad}}(D)$ this operator γ, by using the prolongation principle of A.4.1.

Pick some $\varphi \in L^2_{\text{grad}}(D)$. By definition of the latter space, there exists a Cauchy sequence of smooth functions $\varphi_n \in C_0^\infty(E_3)$ which, once restricted to D, converge towards φ in the sense of $L^2_{\text{grad}}(D)$, and thus $\{\varphi_n\}$ and $\{\text{grad}\ \varphi_n\}$ are Cauchy sequences in $L^2(D)$ and $\mathbb{L}^2(D)$, respectively. Then, by an immediate corollary of (22), $\{\gamma\ \varphi_n\}$ also is a Cauchy sequence and therefore converges towards a limit in $L^2(S)$, which we define as $\gamma\ \varphi$ and call the *trace* of φ. By (22), γ is a continuous linear map from $L^2_{\text{grad}}(D)$ into $L^2(S)$.

Its image, however, has no reason to be all of $L^2(S)$, and constitutes only a dense[12] subspace, which we shall call $T(S)$. Let us provide $T(S)$ with the so-called *quotient norm*, as follows. Pick some φ_S in $T(S)$. There is, by definition of $T(S)$, at least one function φ the trace of which is φ_S, so the pre-image of φ_S is a non-empty affine space, that we may denote $\Phi_D(\varphi_S)$. Now, let us set

$$(23) \qquad [|\varphi_S|] = \inf\{\, [\textstyle\int_D |\varphi|^2 + \int_D |\text{grad}\ \varphi|^2]^{1/2} : \varphi \in \Phi_D(\varphi_S)\,\}.$$

This defines a *norm* $[|\ \ |]$ on $T(S)$, with respect to which γ keeps being continuous, since $[|\gamma\ \varphi|] = [|\ \varphi_S|] \le |||\varphi|||$, where $|||\ \ |||$ denotes the $L^2_{\text{grad}}(D)$-norm. Now,

Definition 7.1. *The normed space* $\{T(S),\ [|\ \ |]\}$ *is denoted* $H^{1/2}(S)$.

Why this name, that we shall explain, but let's just accept this notation for the moment. More importantly,

Proposition 7.3. $H^{1/2}(S)$ *is a Hilbert space.*

Proof. Note that $\Phi_D(\varphi_S)$ is closed in $L^2_{\text{grad}}(D)$, as the pre-image of φ_S by the continuous map γ. Since $L^2_{\text{grad}}(D)$ is complete by definition, the infimum in (23) is reached at a (unique) point φ, which is the projection of 0 on $\Phi_D(\varphi_S)$. Let us call $\pi(\varphi_S)$ this special element, and remark that $\pi\,\gamma\ \varphi = \varphi$ and $\gamma\ \pi\ \varphi_S = \varphi_S$ (operator π is called a "lifting" from S to D). Now let us set $[(\varphi_S, \psi_S)] = \int_D \pi\varphi_S\ \pi\psi_S + \int_D \text{grad}\ \pi\varphi_S \cdot \text{grad}\ \pi\psi_S$, thus defining a scalar product on $T(S)$. Since the norm associated with this scalar product is precisely $[|\ \ |]$, the norm of $H^{1/2}(S)$, the latter is pre-Hilbertian, and since its Cauchy sequences lift to Cauchy sequences of $L^2_{\text{grad}}(D)$, which is complete, they converge, so $H^{1/2}(S)$ is complete, and thus a Hilbert space. ◊

Note that, after (22), $||\varphi_S|| \le [|\varphi_S|]$, where $||\ \ ||$ denotes the $L^2(S)$-norm. Therefore, sequences which converge for $[|\ \ |]$ also converge for $||\ \ ||$, and the identity mapping $\varphi_S \to \varphi_S$ is continuous from $H^{1/2}(S)$ into $L^2(S)$. One says that the new norm $[|\ \ |]$ is *stronger*[13] than $||\ \ ||$, and that there is

[12] Because it contains the restrictions of smooth functions, which are dense in $L^2(S)$.

"topological inclusion" of $H^{1/2}(S)$ into $L^2(S)$, not only the mere "algebraic" inclusion of a set (here $T(S)$) into another set (here $L^2(S)$). Note that the norm couldn't be made any stronger (authorize fewer converging sequences) without breaching the continuity of γ : It's the *strongest* norm with respect to which γ stays continuous.

Now why this name, $H^{1/2}$? This notation pertains to the theory of Sobolev spaces [Ad, Br, Yo]: On any "measured differentiable manifold" X, which S is, there exists a whole family of functional spaces, denoted $H^s(X)$, which includes $L^2(X)$ for $s = 0$, and the one I have been calling here $L^2_{\text{grad}}(X)$, for the sake of notational consistency, for $s = 1$. They form a kind of ladder, H^s being topologically included in H^t, for $t < s$. It happens that our trace space is midway from $L^2(S) \equiv H^0(S)$ to $H^1(S)$ in this hierarchy, hence its name. Giving sense to "midway" and proving the point is not easy, but not important either, since we *know,* having provided a definition, what $H^{1/2}(S)$ is. So let's just accept the name as a dedicated symbol.

Remark 7.1. Now that we have a linear continuous map γ from $L^2_{\text{grad}}(D)$ to $H^{1/2}(S)$, which is by construction surjective, but certainly not injective, since functions supported inside D map to 0, a natural question arises: What is $\ker(\gamma)$? This closed subspace, which is traditionally denoted $H^1_0(D)$, obviously contains $C_0^\infty(D)$. Less obviously, and we'll admit this result [LM], $\ker(\gamma) \equiv H^1_0(D)$ is the closure of $C_0^\infty(D)$ in $L^2_{\text{grad}}(D)$. ◊

Exercise 7.1. Remember that $C_0^\infty(D)$ is dense in $L^2(D)$, which is its completion. How can the closure of $C_0^\infty(D)$ then be *smaller* than $L^2_{\text{grad}}(D)$, which is already smaller than $L^2(D)$?

7.4.2 The interior Dirichlet-to-Neumann map

Next step, let's have a closer look at the lifting π. Finding the minimizer in (23) is a variational problem, which has an associated Euler equation. The latter is *find* $\varphi \in \Phi_D(\varphi_S)$ *such that*

$$\int_D \varphi\, \varphi' + \int_D \operatorname{grad} \varphi \cdot \operatorname{grad} \varphi' = 0 \quad \forall\, \varphi' \in \ker(\gamma), \tag{24}$$

since $\ker(\gamma)$ is the vector subspace parallel to $\Phi_D(\varphi_S)$. This implies, by specializing to smooth test functions,

$$\int_D \varphi\, \varphi' + \int_D \operatorname{grad} \varphi \cdot \operatorname{grad} \varphi' = 0 \quad \forall\, \varphi' \in C_0^\infty(D),$$

[13] Because it is "more demanding", letting fewer sequences converge, having more closed sets or open sets (all these things are equivalent).

which means that $\operatorname{div}(\operatorname{grad}\varphi) = \varphi$ in the weak sense. So (24) is the weak formulation of a boundary value problem, *find* φ *such that*

$$(25) \qquad -\Delta\varphi + \varphi = 0, \quad \gamma\,\varphi = \varphi_S,$$

which gives a nice[14] interpretation of the lifting: π maps the Dirichlet data φ_S to the solution of (25).

Now in case this solution is smooth over[15] D, let us denote by P the "Dirichlet-to-Neumann" linear map $\varphi_S \to \partial_n\varphi$. The divergence integration by parts formula gives

$$(26) \qquad \int_D \varphi\,\psi + \int_D \operatorname{grad}\varphi \cdot \operatorname{grad}\psi = \int_S \partial_n\varphi\,\psi = \int_S \mathrm{P}\varphi_S\,\psi$$

for any smooth function ψ, and hence an *explicit* formula for the scalar product in $H^{1/2}(S)$,

$$(27) \qquad [(\varphi_S\,,\,\psi_S)] = \int_S \mathrm{P}\varphi_S\,\psi_S\,,$$

when φ_S is smooth enough for P to make sense. Here, $\mathrm{P}\varphi_S$ is a function, defined on S. But from the point of view which we have so often adopted, a function is known by its effect on test functions, so we may identify $\mathrm{P}\varphi_S$ and the linear map $\psi_S \to \int_S \mathrm{P}\,\varphi_S\psi_S$. The latter map, in turn, being continuous on $H^{1/2}(S)$, constitutes an element of the space *dual* to $H^{1/2}(S)$, which is denoted by $H^{-1/2}(S)$ (again the question "why $-1/2$?" arises and will be answered, to some extent, in a moment). Now, (27) shows that the map $\varphi_S \to (\psi_S \to [(\varphi_S\,,\,\psi_S)])$, which sends $\varphi_S \in H^{1/2}(S)$ to an element of $H^{-1/2}(S)$, is an extension, a prolongation of the operator P. Quite naturally, we denote by P this extended map, and call it the *Dirichlet-to-Neumann operator*, even though the image $\mathrm{P}\varphi_S$ may fail to be a function.

All these identifications suggest some notational conventions. It is customary to denote $<f, v>$ the scalar which results from applying an element f of the dual space V' to an element v of V. (In case of ambiguities, the more precise notation $<f, v>_{V', V}$ may help.) We then have

$$(28) \qquad [(\varphi_S, \psi_S)] = <\mathrm{P}\varphi_S, \psi_S> = \int_S \mathrm{P}\varphi_S\,\psi_S = \int_S \partial_n(\pi\varphi_S)\,\psi_S$$

[14] And useful: This is the paradigm of a classical approach to boundary-value problems by Hilbertian methods [LM].

[15] Smoothness inside D is assured by a variant of this Weyl lemma we mentioned in Chapter 2. But smoothness *over* D, in the sense of Chapter 2, Subsection 2.2.1, is not warranted, even if S and φ_S are piecewise smooth.

when the latter terms make sense. So we shall abuse the notation and use whichever form is most convenient.

Note that $\sup\{|\langle P\varphi_S, \psi_S\rangle| / [|\psi_S|] : \psi_S \in H^{1/2}(S)\} = [|\varphi_S|]$ after (28). Thus, P is isometric. Moreover, its restriction to $L^2(S)$ is self-adjoint (since $\int_S P\varphi_S\, \psi_S = \int_S \varphi_S\, P\psi_S$) and positive definite (for $\int_S P\varphi_S\, \varphi_S \geq 0$, with equality for $\varphi_S = 0$ only).

As for the terminology, $H^{-1/2}(S)$ is of course one of the Sobolev spaces, which one proves is isomorphic to the dual of $H^{1/2}(S)$. It is made of distributions, not of functions, and contains $L^2(S)$. It may come as a surprise that $H^{1/2}(S)$ can be, like all Hilbert spaces, isomorphic with its dual $H^{-1/2}(S)$, and at the same time, can be identified with a subspace of it, via a continuous injection. But the latter is not *bi*-continuous (continuous in both directions), so there is no contradiction in that.

7.4.3 The exterior Dirichlet-to-Neumann map

For the exterior region, the approach is strictly the same, except for the basic functional space and for the absence of the term φ in the analogue of (25). We'll go much faster, directing attention only to the differences with respect to the previous case.

So let Φ_O be the space[16] of restrictions to $O \equiv E_3 - D$ of elements of Φ. Fitted with the norm $\varphi \rightarrow (\int_O |\text{grad}\, \varphi|^2)^{1/2}$ (which *is* a norm, since O is connected), it becomes a Hilbert space (larger than $L^2_{\text{grad}}(O)$, this time; it's one of these small but irreducible technical differences that force one to do the same work twice, as we are doing here). We denote by $\Phi_O(\varphi_S)$ the affine subspaces of Φ_O of the form $\{\varphi \in \Phi_O : \gamma\, \varphi = \varphi_S\}$, where φ_S is a given function belonging to $H^{1/2}(S)$. The subspaces $\Phi_O(\varphi_S)$ are not empty (there exists a function of $L^2_{\text{grad}}(E_3)$, thus of Φ_O, the trace of which is φ_S), and are closed by continuity of the trace mapping.

Proposition 7.4. *Let* $\varphi_S \in H^{1/2}(S)$ *be given. There exists* $\varphi \in \Phi_O(\varphi_S)$, *unique, such that*

$$(29) \qquad \int_O |\text{grad}\, \varphi|^2 \leq \int_O |\text{grad}\, \varphi'|^2 \quad \forall\, \varphi' \in \Phi_O(\varphi_S).$$

Proof. This specifies φ as the projection of the origin on the affine subspace $\Phi_O(\varphi_S)$, which is non-empty and closed, hence existence and uniqueness for φ. ◊

We note that the Euler equation of the variational problem (29) is

[16] It can be defined, alternatively, as the completion of $C_0^\infty(E_3)$ with respect to the norm $\varphi \rightarrow (\int_O |\text{grad}\, \varphi|^2)^{1/2}$.

(30) $\int_O \operatorname{grad} \varphi \cdot \operatorname{grad} \varphi' = 0 \quad \forall\, \varphi' \in \Phi_O(0).$

Hence, taking smooth compactly supported test functions and integrating by parts, $\Delta\varphi = 0$ in O. The function φ is thus the *harmonic continuation* of φ_S to O, i.e., the solution of the "exterior Dirichlet problem":

$$\Delta\varphi = 0 \text{ in } O, \quad \varphi_{|S} = \varphi_S.$$

(A "condition at infinity" is implicitly provided by the inclusion $\varphi \in \Phi$: Although functions of Φ do not necessarily vanish at infinity (**Exercise 7.2:** Find such a freak), smooth functions of Φ do.) Consider now the normal derivative $\partial_n\varphi$ of φ on S, and denote by P the linear operator $\varphi \to \partial_n\varphi$. By the same arguments as before, P extends to an isometry from $H^{1/2}(S)$ to its dual $H^{-1/2}(S)$. Only the scalar product differs, and

(31) $<P\varphi_S, \psi_S> = \int_S \partial_n\varphi\, \psi_S = \int_O \operatorname{grad} \varphi \cdot \operatorname{grad} \psi,$

where ψ is the harmonic continuation of ψ_S, hence self-adjointness and positivity. Note that both operators, exterior and interior, realize isometries between $H^{1/2}(S)$ and its dual, but for *different* norms on $H^{1/2}(S)$.

7.4.4 Integral representation

We now explain how the knowledge of P is equivalent to solving a particular integral equation on surface S. Subsection 7.4.5 will deal with the discretization of this equation. This will give us a linear system, to be solved as a prelude to solving the magnetostatics problem, to which we shall return in Section 7.5.

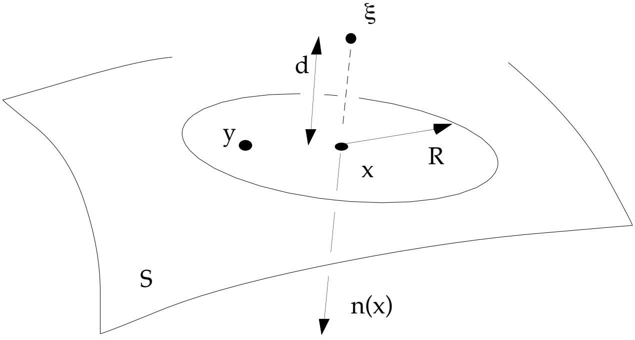


FIGURE 7.3. Notations. $S_R(x)$, or $S_R(\xi)$, is the disk of radius R drawn on S.

Let us introduce some notation: ξ will denote a point in space, and $x(\xi)$, or simply x, its projection on S (Fig. 7.3). If S is regular, the mapping $\xi \rightarrow x$ is well defined in some neighborhood of S. Let $d = |\xi - x|$ and $S_R(\xi) = \{y \in S : |y - x(\xi)| < R\}$. One has (this is trivial, but important):

$$(32) \qquad \int_{S_R(\xi)} |\xi - y|^{-1}\, dy \leq C\, R,$$

where dy is the measure of areas on S, and C a constant depending on S but not on ξ. Last, we denote by $n(\xi)$ the vector at ξ parallel to $n(x)$. One thus obtains in the neighborhood of S a vector field, still denoted by n, which extends the field of normals. (Field lines of n are orthogonal to S.)

Remark 7.2. If S is only piecewise smooth, as we assumed from the beginning, the properties to be established below stay valid at all points of regularity of the surface (those in the neighborhood of which S has a tangent plane, and bounded principal curvatures). ◊

Now let q be a function defined on S, taken as smooth in a first approach (continuous is enough), and let us consider its potential φ:

$$(33) \qquad \varphi(x) = \frac{1}{4\pi}\int_S \frac{q(y)}{|x-y|}\, dy.$$

One may interpret q as an *auxiliary magnetic charge density* and φ as a magnetic potential (called “single layer potential”), from which derives a magnetic field $h = \operatorname{grad} \varphi$. If $x \notin S$, the integral converges in an obvious way. But moreover, q being bounded, it also converges when $x \in S$ (study the contribution to the integral of a small disk centered at x, and invoke (32)). Finally, the function φ is continuous (thanks to the uniform bound (32), again), null at infinity, and harmonic outside S. Let us call K the operator $q \rightarrow \varphi_{|S}$.

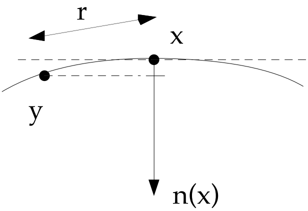
Next we study the field $h = \operatorname{grad} \varphi$. By differentiation under the summation sign, one finds

$$\operatorname{grad} \varphi = x \rightarrow \frac{1}{4\pi}\int_S q(y)\, \frac{y-x}{|x-y|^3}\, dy,$$

and this time the convergence of the integral when $x \in S$ is by no means certain. On the other hand, the real-valued integral

$$(34) \qquad (Hq)(x) = \frac{1}{4\pi}\int_S q(y)\, n(x) \cdot \frac{y-x}{|x-y|^3}\, dy$$

does converge when $x \in S$: For if x is a point of regularity of S, R a



positive real value, and $|q|_R$ an upper bound for $q(y)$ on the set $S_R(x)$, the contribution of $S_R(x)$ to the integral is bounded by

$$(35) \qquad \frac{|q|_R}{4\pi}\int_{S_R(x)} n(x)\cdot\frac{y-x}{|x-y|^3}\,dy,$$

a quantity which tends to 0 when R tends to 0 (take polar coordinates originating at x, remark by looking at the inset that $|x - y| \sim r$ and $n(x)\cdot(y - x) \sim r^2$). Hence another integral operator H, of the same type as K.

The form of Hq could suggest that it is the restriction to S of the function $n \cdot \operatorname{grad} \varphi$ (compact notation for $\xi \to n(\xi)\cdot \operatorname{grad} \varphi(\xi)$). But this is not so, for $n \cdot \operatorname{grad} \varphi$, contrary to φ, is *not* continuous across S, but has a jump, equal to q, as we presently see. (Recall the definition of the jump as

$$[n\cdot \operatorname{grad} \varphi]_S = n_+ \cdot (\operatorname{grad} \varphi)_+ + n_- \cdot (\operatorname{grad} \varphi)_-,$$

with the notation of Fig. 7.4. This can be denoted by $[\partial\varphi/\partial n]_S$, or better, $[\partial_n\varphi]_S$.)

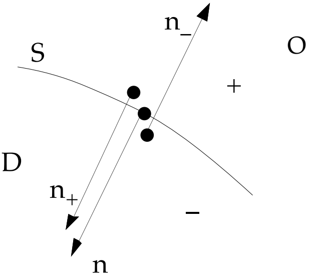


FIGURE 7.4. Notations for Proposition 7.5.

Proposition 7.5. *Let* φ *be the function defined in* (33). *One has*

$$[n\cdot \operatorname{grad} \varphi]_S \equiv n_+ \cdot (\operatorname{grad} \varphi)_+ + n_- \cdot (\operatorname{grad} \varphi)_- = q,$$

$$n_+ \cdot (\operatorname{grad} \varphi)_+ - n_- \cdot (\operatorname{grad} \varphi)_- = 2\,Hq.$$

Proof. All these are well-defined functions, for φ is C^∞ outside S. Let d and R be fixed. Let us sit at point $\xi = x + \alpha\, d\, n(x)$, where α is meant to eventually converge to 0, and let $\beta = |\alpha|^{1/2}$. The contribution of the set $S - S_{\beta R}(x)$ to the integral

$$n(\xi) \cdot (\text{grad}\ \varphi)(\xi) = (4\pi)^{-1} \int_S dy\ q(y)\ |\xi - y|^{-3}\, n(\xi) \cdot (y - \xi)$$

has a well-defined limit (namely, $(Hq)(x)$) when ξ tends to x. So let us examine, according to a standard technique in singular integral computations, the contribution of $S_{\beta R}(x)$, whose limit will depend on the sign of α. Up to terms in $o(\alpha)$, this is

$$(4\pi)^{-1}\, q(x)\, n(x) \cdot \int_{S_{\beta R}} dy\ |\xi - y|^{-3}\, (y - \xi) \approx$$
$$(4\pi)^{-1}\, q(x)\, \alpha\, d \int_0^{\beta R} 2\pi\, r\, dr\, (r^2 + \alpha^2 d^2)^{-3/2}\ .$$

Studying this integral is a classical exercise:[17] Its limit is $\pm\ q(x)/2$ according to whether α tends to 0 from above or from below. The limit of $n(\xi)$, in the same circumstances, is n_+ or n_-. Thus $n_\pm \cdot (\text{grad}\ \varphi)_\pm = q/2 \pm Hq$. Hence, by addition and subtraction, the announced equalities. ◊

Here follows a first implication of Prop. 7.5. Let q' be another charge density, and φ' its potential. According to the divergence integration by parts formula, one has

$$(36) \qquad \int_{E_3} \text{grad}\ \varphi \cdot \text{grad}\ \varphi' = \int_D \text{grad}\ \varphi \cdot \text{grad}\ \varphi' + \int_O \text{grad}\ \varphi \cdot \text{grad}\ \varphi'$$
$$= \int_S \varphi\, [\partial_n \varphi] = \int_S \varphi\, q' = \int_S Kq\, q',$$

and in particular, $\int_{E_3} |\text{grad}\ \varphi|^2 = \int_S Kq\, q$. The operator K is thus self-adjoint and (strictly) positive definite on its domain, which we restricted up to now to regular functions. Note the formal similarity with P, exterior or interior: K is a bilateral Dirichlet-to-Neumann operator, so to speak.

This suggests the following extension of the definition of K, which is a variant of the "extension by continuity" of A.4.1. Suppose $q \in H^{-1/2}(S)$ given. The problem *find* $\varphi \in \Phi$ *such that*

$$\int_{E_3} \text{grad}\ \varphi \cdot \text{grad}\ \varphi' = \int_S q\, \varphi' \quad \forall\, \varphi' \in \Phi$$

is well posed, since $\varphi' \in H^{1/2}(S)$, with continuity of the trace mapping, and the map $q \to \varphi_S$ is therefore continuous from $H^{-1/2}(S)$ into $H^{1/2}(S)$. As it constitutes an extension of K, it is only natural to also denote it by K. (The K thus extended is an isometry between $H^{-1/2}(S)$ and $H^{1/2}(S)$, again for a different norm than in the previous two cases.) Then, after (36),

$$\int_{E_3} \text{grad}\ \varphi \cdot \text{grad}\ \varphi' = \langle q, Kq' \rangle,$$

[17] This is the same computation one does in electrostatics when studying the field due to a uniform plane layer of electric charge.

hence $\int_{E_3} \operatorname{grad} \varphi \cdot \operatorname{grad} \varphi' = \int_S q \, Kq'$ when $q \in L^2(S)$, so that K, now considered as an operator from $L^2(S)$ into itself, is self-adjoint and positive definite.

A second implication is the following formula, which explicitly gives the exterior normal derivative of φ in terms of the charge q:

$$\partial_n\varphi(x) = \tfrac{1}{2}\, q(x) + \frac{1}{4\pi}\int_S q(y)\, n(x) \cdot \frac{y-x}{|x-y|^3}\, dy,$$

that is,

(37) $\partial_n\varphi = (1/2 + H)\, q.$

Since the mapping $q \to \partial_n\varphi$ is linear continuous from $H^{-1/2}(S)$ into itself, and since $Hq = \partial_n\varphi - q/2$ when q is regular, after (37), we may extend the operator H to $H^{-1/2}(S)$ in the present case, too.

We may now, at last, give an explicit form to the operator P. (Let’s revert to our usual convention that φ_S denotes the trace of φ.) Since $\partial_n\varphi = P\varphi_S$, by definition, and $\varphi_S = Kq$, one has $PK = 1/2 + H$, after (37), hence the result we were after:

(38) $P = (1/2 + H)\, K^{-1}.$

Yet we are not through, far from it, for (38) must be discretized, and this cannot be done simply by replacing the operators H and K by their matrix equivalents **H** and **K**, whatever they are (we’ll soon give them): The matrix $(1/2 + \mathbf{H})\, \mathbf{K}^{-1}$ thus obtained would not be symmetrical, contrary to our wishes.

7.4.5 Discretization

In this subsection, we simply write φ for φ_S, and Φ for the space $H^{1/2}(S)$. Let Q denote the space $H^{-1/2}(S)$ where q lives. After (38), written in weak form, operator P is such that

(39) $\langle P\varphi, \varphi'\rangle = \langle q, \varphi'\rangle/2 + \langle Hq, \varphi'\rangle \quad \forall\, \varphi' \in \Phi$

for each couple $\{\varphi, q\}$ linked by the relation

(40) $\langle \varphi, q'\rangle = \langle Kq, q'\rangle \quad \forall\, q' \in Q.$

A mesh m of D, and thus of S, being defined, let us denote by Φ_m and Q_m mesh-dependent approximation spaces for Φ and Q, to be constructed. We have a natural choice for Φ_m already: the trace on S of $W^0_m(D)$.

The choice of Q_m, for which we have no obvious rationale yet, is deferred for a while.

On the sight of (39) and (40), a discretization principle suggests itself: We look for P_m, an operator of type $W^0{}_m(S) \to W^0{}_m(S)$, that should be symmetrical like P and such that

(41) $$<P_m\varphi, \varphi'> = <q, \varphi'>/2 + <Hq, \varphi'> \quad \forall\, \varphi' \in \Phi_m$$

for all couples $\{\varphi, q\} \in \Phi_m \times Q_m$ linked by the relation

(42) $$<\varphi, q'> = <Kq, q'> \quad \forall\, q' \in Q_m.$$

The representations $\varphi = \sum_{n \in \mathcal{N}(S)} \boldsymbol{\varphi}_n w_n$ and $q = \sum_{i \in \mathcal{J}} \mathbf{q}_i \zeta_i$ (where the set $\mathcal{J}$ and the basis functions ζ_i have not yet been described), define isomorphisms between the spaces Φ_m and Q_m and the corresponding spaces $\boldsymbol{\Phi}$ and $\mathbf{Q}$ of vectors of DoFs. Let us denote by $\mathbf{B}, \mathbf{H}, \mathbf{K}$ the matrices defined as follows, which correspond to the various brackets in (41) and (42):

(43) $$\mathbf{B}_{n\,i} = \int_S dx\, w_n(x)\, \zeta_i(x),$$

(44) $$\mathbf{K}_{i\,j} = (4\pi)^{-1} \iint_S dx\, dy\, (|y - x|)^{-1} \zeta_i(x)\, \zeta_j(y),$$

(45) $$\mathbf{H}_{n\,i} = (4\pi)^{-1} \iint_S dx\, dy\, (|y - x|)^{-3}\, n(x) \cdot (y - x)\, \zeta_i(y)\, w_n(x).$$

According to the foregoing discretization principle, we look for the symmetric matrix $\mathbf{P}$ (of order $N(S)$, the number of nodes of the mesh on S), such that

(46) $$(\mathbf{P}\,\boldsymbol{\varphi}, \boldsymbol{\varphi}') = (\mathbf{B}\,\mathbf{q}, \boldsymbol{\varphi}')\,/2 + (\mathbf{H}\,\mathbf{q}, \boldsymbol{\varphi}') \quad \forall\, \boldsymbol{\varphi}' \in \boldsymbol{\Phi},$$

for all couples $\{\boldsymbol{\varphi}, \mathbf{q}\} \in \boldsymbol{\Phi} \times \mathbf{Q}$ linked by

(47) $$(\mathbf{B}^t\, \boldsymbol{\varphi}, \mathbf{q}') = (\mathbf{K}\,\mathbf{q}, \mathbf{q}') \qquad \forall\, \mathbf{q}' \in \mathbf{Q}$$

(the bold parentheses denote scalar products in finite dimension, as in 4.1.1). Since (47) amounts to $\mathbf{q} = \mathbf{K}^{-1}\mathbf{B}^t\,\boldsymbol{\varphi}$, we have

(48) $$\mathbf{P} = \mathrm{sym}((\mathbf{B}/2 + \mathbf{H})\, \mathbf{K}^{-1}\, \mathbf{B}^t)$$

after (46), with t for “transpose” and sym for “symmetric part”.

This leaves the selection of “basis charge distributions” ζ_i to be performed. An obvious criterion for such a choice is the eventual simplicity of the computation of double integrals in (43–45), and from this point of view, taking ζ_i constant on each triangle is natural: thus $\mathcal{J}$ will be the set of surface triangles, and one will define ζ_i for $i \in \mathcal{J}$ as the characteristic function of triangle i (equal to one over it and to 0 elsewhere), divided by the area. This was the solution retained for the Trifou eddy-current code (cf. p. 225), and although not totally satisfactory (cf. Exer. 7.3), it does make the computation of double integrals simple.

Simple does not mean trivial however, and care is required for terms of **K**, which are of the form

$$\mathrm{K}_{\mathrm{TT'}} = \tfrac{1}{4\pi} \int_{\mathrm{T}} \mathrm{d}y \int_{\mathrm{T'}} \mathrm{d}x \, |x - y|^{-1},$$

where T and T' are two non-intersecting triangles in generic position in 3-space. The internal integral is computed analytically, and the outer one is approximated by a quadrature formula, whose sophistication must increase when triangles T and T' are close to each other. Any programmer with experience on integral or semi-integral methods of some kind has had, at least once in her life, to implement this computation, and knows it's a tough task. Unfortunately, the details of such implementations are rarely published (more out of modesty than a desire to protect shop secrets). Some indications can be gleaned from [AR, Cl, R&].

Remark 7.3. As anticipated earlier, the "naive" discretization of (38), yielding $\mathbf{P} = (1/2 + \mathbf{H})\mathbf{K}^{-1}$, would be inconsistent (the dimension of $\mathbf{K}$ is not what is expected for $\mathbf{P}$, that is, N(S)). But the more sophisticated expression $(\mathbf{B}/2 + \mathbf{H})\,\mathbf{K}^{-1}\,\mathbf{B}^t$ would not do, either, since this matrix is not symmetric, and the symmetrization in (48), to which we were led in a natural way, is mandatory. ◊

Exercise 7.3. Show, by a counter-example, that matrix $\mathbf{P}$ may happen to be singular with the above choice for the ζ_i.

7.5 BACK TO MAGNETOSTATICS

We may now finalize the description of the "finite elements and integral method in association" method of 7.3.3, in the case when the unknown is the scalar potential. Let Φ(D) be the space of restrictions to D of the scalar potentials in space Φ. We denote again by φ_S the trace of φ on S. Thanks to the operator P, the Euler equation (15) is equivalent to the following problem: *find* $\varphi \in \Phi(\mathrm{D})$ *such that*

$$(49) \qquad \int_{\mathrm{D}} \mu\,(\mathrm{h}^{\mathrm{j}} + \mathrm{grad}\,\varphi) \cdot \mathrm{grad}\,\varphi' + \mu_0 \int_{\mathrm{S}} (\mathrm{n} \cdot \mathrm{h}^{\mathrm{j}} + \mathrm{P}\varphi_{\mathrm{S}})\,\varphi'_{\mathrm{S}} = 0 \quad \forall\, \varphi' \in \Phi(\mathrm{D}).$$

Let m be a mesh of D. Then $\Phi_m(\mathrm{D}) = \{\varphi : \varphi = \sum_{\mathrm{n} \in \mathcal{N}} \boldsymbol{\varphi}_{\mathrm{n}} \mathrm{w}_{\mathrm{n}}\}$ is the natural approximation space for Φ(D). Hence the following approximation of (49), *find* $\varphi \in \Phi_m(\mathrm{D})$ *such that, for all* $\varphi' \in \Phi_m(\mathrm{D})$,

$$(50)\qquad \int_D \mu\,(h^j + \text{grad}\,\varphi)\cdot \text{grad}\,\varphi' + \mu_0 \int_S (n\cdot h^j + P\varphi_S)\,\varphi'_S = 0.$$

Let $\boldsymbol{\varphi}$ be the vector of degrees of freedom (one for each node, including this time those contained in S), and $\boldsymbol{\Phi} = \mathbb{R}^{\mathcal{N}}$ (there are N nodes). We denote

$$\boldsymbol{\eta}^j_n = \int_D \mu\, h^j \cdot \text{grad}\, w_n + \mu_0 \int_S n \cdot h^j\, w_n,$$

$\boldsymbol{\eta}^j = \{\boldsymbol{\eta}^j_n : n \in \mathcal{N}\}$, and let $\mathbf{G}$, $\mathbf{R}$, $\mathbf{M}_1(\mu)$ be the same matrices as in Chapter 5. Still denoting by $\mathbf{P}$ the extension to $\boldsymbol{\Phi}$ (obtained by filling-in with zeroes) of the matrix $\mathbf{P}$ of (48), we finally get the following approximation for (50):

$$(51)\qquad (\mathbf{G}^t\mathbf{M}_1(\mu)\mathbf{G} + \mathbf{P})\boldsymbol{\varphi} + \boldsymbol{\eta}^j = 0.$$

Although the matrix $\mathbf{P}$ of (38) is full, the linear system (51) is reasonably sparse, because $\mathbf{P}$ only concerns the "S part" of vector $\boldsymbol{\varphi}$.

Remark 7.4. The linear system is indeed an *approximation* of (50), and not its interpretation in terms of degrees of freedom, for $(\mathbf{P}\boldsymbol{\varphi}, \boldsymbol{\varphi}')$ is just an approximation of $\int_S P\varphi_S\,\varphi'_S$ on the subspace $\Phi_m(D)$, not its restriction, as in the Galerkin method. (This is another example of "variational crime".) ◊

Remark 7.5. There are other routes to the discretization of P. Still using magnetic charges (which is a classic approach, cf. [Tz]), one could place them differently, not on S but inside D [MW]. One might, for example [Ma], locate a point charge just beneath each node of S. (The link between $\mathbf{q}$ and $\boldsymbol{\varphi}$ would then be established by collocation, that is to say, by enforcing the equality between φ and the potential of q at nodes.[18]) Another approach [B2] stems from the remark that interior and exterior Dirichlet-to-Neumann maps (call them P_{int} and P_{ext}) add to something which is easily obtained in discrete form, because of the relation $(P_{\text{int}} + P_{\text{ext}})\varphi = q = K^{-1}\,\varphi$. Since, in the present context, we must mesh D anyway, a natural discretization $\mathbf{P}_{\text{int}}$ of P_{int} is available, thanks to the "static condensation" trick of Exer. 4.8: One minimizes the quantity $\int_D |\,\text{grad}(\sum_{n\in\mathcal{N}(D)} \varphi_n w_n)|^2$ with respect to the *inner* node values $\boldsymbol{\varphi}_n$, hence a quadratic form with respect to the vector $\boldsymbol{\varphi}$ (of surface node DoFs), the matrix of which is $\mathbf{P}_{\text{int}}$. A reasoning similar to the one we did around (46–47) then suggests $\mathbf{B}\,\mathbf{K}^{-1}\,\mathbf{B}^t$ as the correct discretization of K^{-1}, hence finally $\mathbf{P} \equiv \mathbf{P}_{\text{ext}} = \mathbf{B}\,\mathbf{K}^{-1}\,\mathbf{B}^t - \mathbf{P}_{\text{int}}$ (which ensures the symmetry of $\mathbf{P}_{\text{ext}}$, but does not eliminate the difficulty evoked in Exer. 7.3). And (lest we

[18] See, e.g., [KP, ZK]. These authors' method does provide a symmetric $\mathbf{P}$, but has other drawbacks. Cf. [B2] for a discussion of this point.

forget . . .) for some simple shapes of S (the sphere, for example), P is known in closed form, as the sum of a series. ◊

A similar theory can be developed "on the curl side" [B2, B3]: Start from problem (16), introduce the exterior energy $\frac{1}{2}\int_{E_3 - D} \mu^{-1} |\operatorname{rot} a|^2$, then the operator $\mathbb{P} = a_S \to n \times \operatorname{rot} a$, where a satisfies $\operatorname{rot}\operatorname{rot} a = 0$ outside D. Instead ot the auxiliary charge q, one has an auxiliary current density borne by S. See [RR] for an application of this technique.

EXERCISES

Exercises 7.1 to 7.3 are on pp. 205, 208, and 214, respectively.

Exercise 7.4. Show that (in spite of Exer. 7.3) the matrix of system (51) is regular.

Exercise 7.5. Given a smooth function q with bounded support, its *Newtonian potential* φ is

$$\varphi(x) = \frac{1}{4\pi}\int_{E_3} \frac{q(y)}{|x-y|}\,dy.$$

Show that $-\Delta\varphi = q$.

Exercise 7.6. Prove (22). Show that one can reduce the problem to the case where D is a half-space and S a plane, with φ compactly supported. Take Cartesian coordinates for which $D = \{x : x^1 \geq 0\}$, and use Fubini to show that the problem can be reduced to studying functions of one real variable t (here x^1) with values in a functional space X (here, $L^2(E_2)$). Work on functions $u \in C^1([0, 1] ; X)$ and bound $\|u(0)\|^2$ by $\int_0^1 \|u(t)\|^2\,dt + \int_0^1 \|\partial_t u(t)\|^2\,dt$, using Cauchy–Schwarz.

HINTS

7.1. This is a poorly wrapped paradox, almost a mere play on words: Make sure you understand that "closure" means different things in the text of the exercise and in the sentence just before. (My apologies if you felt insulted.)

7.2. Take a sequence of points x_n in E_3 tending to infinity, and have φ supported by small neighborhoods of these points. It's easy to enforce $\int_{E_3} |\operatorname{grad}\varphi|^2 < \infty$, although φ(x) does not tend to zero when |x| tends to infinity, by construction.

7.5. Since differential operators commute with the convolution product, the problem reduces to showing that $-\Delta(x \rightarrow 1/|x|) = 4\pi\delta_0$, where δ_0 is the Dirac mass at the origin. Half of it was solved with Exer. 4.9. The delicate point is the computation of the divergence *in the sense of distributions* of the field $x \rightarrow -x/|x|^3$ (refer to A.1.9 for the arrowed notation).

SOLUTIONS

7.1. $C_0^\infty(D)$, which is dense in $L^2(D)$ with respect to the L^2-norm, is indeed dense also in the subspace $L^2_{\text{grad}}(D)$, *with respect to this same norm.* But with respect to the *stronger* norm put on $L^2_{\text{grad}}(D)$, which is the point, it's not, simply because there are *fewer* Cauchy sequences for this norm, so their limits form a smaller space, namely, $H^1_0(D)$. In short, the stronger the norm, the smaller the closure.

7.2. Let $\varphi_n = y \rightarrow n\,(1 - n^4\,|y - x_n|^+)$, where $+$ denotes the positive part of an expression. Then $\int |\operatorname{grad} \varphi_n|^2$ is in n^{-2}, so $\varphi = \sum_n \varphi_n$ is in the Beppo Levi space, but $\varphi(x_n) = n$ for n large enough, and doesn't vanish at infinity.

7.3. See Ref. [CC].

7.5. The same computation as in Exer. 4.9 would give

$$(*) \qquad \operatorname{div}(x \rightarrow x/|x|^3) = x \rightarrow 3/|x|^3 - 3\,x \cdot x/|x|^5 \equiv 0,$$

if it were not for the singularity at the origin. What was obtained there is only the "function part" of the *distribution* $\operatorname{div}(x \rightarrow x/|x|^3)$, which is therefore concentrated at the origin. To find it, apply Ostrogradskii to a sphere of radius r centered at the origin, which gives the correct result, $\operatorname{div}(x \rightarrow x/|x|^3) = -4\pi\,\delta_0$. Then, denoting by χ the kernel $x \rightarrow 1/(4\pi|x|)$, one has $\varphi = \chi * q$, and hence, $-\Delta\varphi = -\Delta(\chi * q) = (-\Delta\chi) * q = \delta_0 * q = q$.

REFERENCES

[Ad] R. Adams: **Sobolev Spaces**, Academic Press (Orlando, FL), 1975.

[AL] B.I. Agoshkov, B.I. Lebedev: "Operatory Puancare-Steklova i metody razdeleniya oblasti v variatsionnyx zadachax", in **Vychislitel'nye Protsessy i Sistemy**, Nauka (Moscow), 1985, pp. 173–227.

[AR] R. Albanese, G. Rubinacci: "Integral formulation for 3d eddy-current computation using edge elements", **IEE Proc., 135, Pt. A,** 7 (1988), pp. 457–463.

[B&] A. Bossavit, C. Emson, I. Mayergoyz: **Méthodes numériques en électromagnétisme**, Eyrolles (Paris), 1991.

[B2] A. Bossavit: "Mixed Methods and the Marriage Between 'Mixed' Finite Elements and Boundary Elements", **Numer. Meth. for PDEs, 7** (1991), pp. 347–362.

[B3] A. Bossavit: "On various representations of fields by potentials and their use in boundary integral methods", **COMPEL, 9** (1990), Supplement A, pp. 31–36.

[Br] H. Brezis: **Analyse fonctionnelle**, Masson (Paris), 1983.

[BM] X. Brunotte, G. Meunier, J.F. Imhoff: "Finite element solution of unbounded problems using transformations", **IEEE Trans., MAG-30**, 5 (1994), pp. 2964–2967.

[CC] P. Chaussecourte, C. Chavant: "Discretization of the Poincaré Steklov operator and condition number of the corresponding matrix in the Trifou FEM/BIEM aproach to solving 3D eddy current equations", in A. Kost, K. Richter (eds.): **Confererence record, Compumag 1995** (TU Berlin, 1995), pp. 678–679.

[Cl] C.J. Collie: "Magnetic fields and potentials of linearly varying current or magnetisation in a plane bounded region", in **Proc. Compumag Conference on the Computation of Magnetic fields**, St. Catherine's College, Oxford, March 31 to April 2, 1976, The Rutherford Laboratory (Chilton, Didcot, Oxon, U.K.), 1976, pp. 86–95.

[Co] J.L. Coulomb: "A methodology for the determination of global electromechanical quantities from a finite element analysis and its application to the evaluation of magnetic forces, torques and stiffness", **IEEE Trans., MAG-19,** 6 (1983), pp. 2514-2519.

[DL] R. Dautray, J.L. Lions (eds.): **Analyse mathématique et calcul numérique pour les sciences et les techniques**, t. 2, Masson (Paris), 1985.

[Hu] M. Hulin, N. Hulin, D. Perrin: **Équations de Maxwell, Ondes Électromagnétiques**, Dunod (Paris), 1992.

[IM] J.F. Imhoff, G. Meunier, J.C. Sabonnadière: "Finite element modeling of open boundary problems", **IEEE Trans., MAG-26**, 2 (1990), pp. 588–591.

[KP] A. Kanarachos, C. Provatidis: "On the symmetrization of the BEM formulation", **Comp. Meth. Appl. Mech. Engng., 71** (1988), pp. 151–165.

[LM] J.L. Lions, E. Magenes: **Problèmes aux limites non homogènes et applications**, Vols. 1–2, Dunod (Paris), 1968.

[LS] H. Li, S. Saigal: "Mapped infinite elements for 3-D vector potential magnetic problems", **Int. J. Numer. Meth. Engng., 37,** 2 (1994), pp. 343–356.

[Ma] I. Mayergoyz: "Boundary Galerkin's approach to the calculation of eddy currents in homogeneous conductors", **J. Appl. Phys., 55,** 6 (1984), pp. 2192–2194.

[MW] B.H. McDonald, A. Wexler: "Finite element solution of unbounded field problems", **IEEE Trans., MTT-20**, 12 (1972), pp. 841–347.

[R&] Z. Ren, F. Bouillault, A. Razek, J.C. Vérité: "An efficient semi-analytical integration procedure in three dimensional boundary integral method", **COMPEL, 7,** 4 (1988), pp. 195–205.

[RR] Z. Ren, A. Razek: "Boundary Edge Elements and Spanning Tree Technique in Three-Dimensional Electromagnetic Field Computation", **Int. J. Numer. Meth. Engng., 36** (1993), pp. 2877–2693.

[Tz] E. Trefftz: "Ein gegenstück zum Ritzschen Verfahren", in **Proc. 2nd Int. Congr. Appl. Mech.** (Zürich), 1926, pp. 131–137.

[Yo] K. Yosida: **Functional Analysis**, Springer-Verlag (Berlin), 1965.

[ZK] O.C. Zienkiewicz, D.W. Kelly, P. Bettess: "The Coupling of Finite Element and Boundary Solution Procedures", **Int. J. Numer. Meth. Engng., 11** (1977), pp. 355–376.

CHAPTER 8

Eddy-current Problems

We now move beyond magnetostatics to tackle a non-stationary model. The starting point is again Maxwell's system with Ohm's law:

(1) $-\partial_t d + \mathrm{rot}\, h = j,$ (2) $\partial_t b + \mathrm{rot}\, e = 0,$

(3) $d = \varepsilon\, e,$ (4) $j = j^g + \sigma e,$ (5) $b = \mu\, h,$

where j^g is a given current density. In almost all of this chapter, we suppose j^g "harmonic", that is, of the form[1]

(6) $j^g(t) = \mathrm{Re}[\mathrm{J}^g \exp(i\omega t)],$

where J^g is a *complex*-valued vector field, and we'll look for all fields in similar form: $h(t) = \mathrm{Re}[\mathrm{H} \exp(i\omega t)]$, $e(t) = \mathrm{Re}[\mathrm{E} \exp(i\omega t)]$, etc. The functional spaces where these fields roam will still be denoted by $\mathbb{L}^2$, $\mathbb{L}^2_{\mathrm{rot}}$, etc., but it should be clearly understood that *complexified* vector spaces are meant (see A.4.3). By convention, for two complex vectors $\mathrm{U} = u_R + iu_I$ and $\mathrm{V} = v_R + iv_I$, one has $\mathrm{U} \cdot \mathrm{V} = u_R \cdot v_R - u_I \cdot v_I + i(u_I \cdot v_R + u_R \cdot v_I)$, the Hermitian scalar product being $\mathrm{U} \cdot \mathrm{V}^*$, where the star denotes complex conjugation, and the norm being given by $|\mathrm{U}|^2 = \mathrm{U} \cdot \mathrm{U}^*$, not by $\mathrm{U} \cdot \mathrm{U}$. Note that an expression such as $(\mathrm{rot}\, \mathrm{U})^2$ should thus be understood as $\mathrm{rot}\, \mathrm{U} \cdot \mathrm{rot}\, \mathrm{U}$, not as $|\mathrm{rot}\, \mathrm{U}|^2$.

The form (6) of the given current is extremely common in electrotechnical applications, where one deals with alternating currents at a well-defined frequency f. The constant $\omega = 2\pi f$ is called *angular frequency*.

Under these conditions, system (1–5) becomes

(7) $-\, i\omega \mathrm{D} + \mathrm{rot}\, \mathrm{H} = \mathrm{J},$ (8) $i\omega\, \mathrm{B} + \mathrm{rot}\, \mathrm{E} = 0,$

(9) $\mathrm{D} = \varepsilon\, \mathrm{E},$ (10) $\mathrm{J} = \mathrm{J}^g + \sigma \mathrm{E},$ (11) $\mathrm{B} = \mu\, \mathrm{H}.$

[1] There are other possibilities, such as $\mathrm{Im}[\mathrm{J}^g \exp(i\omega t)]$, or $\mathrm{Re}[\sqrt{2}\ \mathrm{J}^g \exp(i\omega t)]$, etc. What matters is consistency in such uses.

8.1 THE MODEL IN H

The model under study in the present chapter is further characterized by a few simplifications, the most noteworthy being the neglect, for reasons we now indicate, of the term $-\mathrm{i}\omega\mathrm{D}$ in (7).

8.1.1 A typical problem

Figure 8.1 ([Na], pp. 209–247) shows a case study, typical of computations one may have to do when designing induction heating systems, among other examples: An induction coil, fed with alternative current, induces currents (called "eddy currents" or, in many national traditions, "Foucault currents") in an aluminum plate, and one wants to compute them.

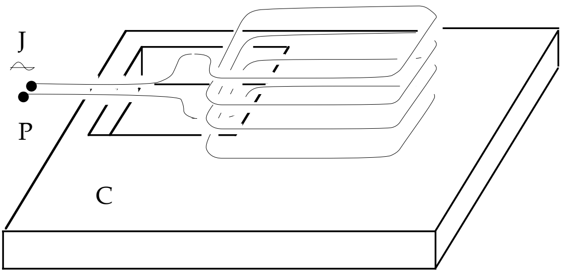


FIGURE 8.1. The real situation ("Problem 7" of the TEAM Workshop [Na]): compute eddy currents induced in the "passive conductor" C by an inductive coil, or "active conductor", which carries low-frequency alternating current. (The coil has many more loops than represented here, and occupies volume I of Fig. 8.2 below.) The problem is genuinely three-dimensional (no meaningful 2D modelling). Although the pieces are in minimal number and of simple shape, and the constitutive laws all linear, it's only during the 1980s that computations of similar complexity became commonplace.

Computing the field inside the coil, while taking its fine structure into account, is a pratical impossiblity, but is also unnecessary, so one can replace the situation of Fig. 8.1 by the following, idealized one, where the inducting current density is supposed to be given in some region I, of the same shape as the coil (Fig. 8.2).

This equivalent distribution of currents in I is easily computed, hence the source current J^g of (11), with I as its support. (One takes as given a mean current density, with small scale spatial variations averaged out.) If $\{\mathrm{E}, \mathrm{H}\}$ is the solution of (7–11), H is then a correct approximation to the actual magnetic field (inside I, as well), because the same currents are at stake (up to small variations near I) in both situations. However, the field E, as given by the same equations, has not much to do with the actual electric field, since in particular the way the coil is linked to the power supply is not considered. (There is, for instance, a high electric field between the connections, in the immediate vicinity of point P of Fig. 8.1, a fact which of course cannot be discovered by solving the problem of Fig. 8.2.)

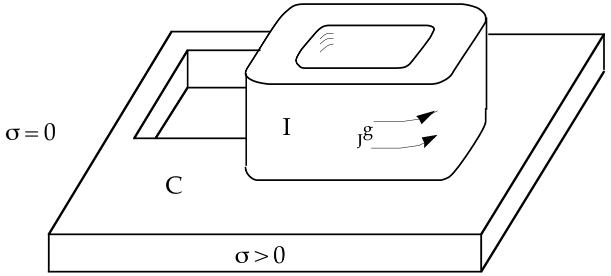


FIGURE 8.2. The modelled imaginary situation: Subregion I (for "inductor") is the support of a known alternative current, above a conductive plate C.

These considerations explain why emphasis will lie, in what follows, on the magnetic field H, and not on E (which we shall rapidly eliminate from the equations).

8.1.2 Dropping the displacement-currents term

Let us now introduce the main simplification, often described as the "low-frequency approximation", which consists in neglecting the term of "Maxwell displacement currents", that is $i\omega\,\mathrm{D}$, in (7). By rewriting (7), (9), and (10) in the form $\mathrm{rot}\ \mathrm{H} = \mathrm{J}^g + \sigma\mathrm{E} + i\omega\ \varepsilon\mathrm{E}$, one sees that this term is negligible in the conductor inasmuch as the ratio $\varepsilon\,\omega/\sigma$ can be considered small. In the air, where $\sigma = 0$, everything goes as if these displacement currents were added to the source-current J^g, and the approximation is justified if the ratio $\|i\omega\mathrm{D}\|/\|\mathrm{J}^g\|$ is small ($\|\ \|$ being some convenient norm). In many cases, the induced currents $\mathrm{J} = \sigma\,\mathrm{E}$ is of the same order of magnitude as the source current J^g, and the electric field is of the same order of magnitude outside and inside the conductor. If so, the ratio of $i\omega\mathrm{D}$ to J^g is also on the order of $\varepsilon\,\omega/\sigma$. *The magnitude of the ratio* $\varepsilon\,\omega/\sigma$ *is thus often a good indicator of the validity of the low-frequency approximation.* In the case of induction heating at industrial frequencies, for instance $\omega = 100\ \pi$, the magnitude of σ being about 5×10^6, and $\varepsilon = \varepsilon_0 \cong 1/(36\pi \times 10^9)$, the ratio $\varepsilon\,\omega/\sigma$ drops to about 5×10^{-16}, and one cannot seriously object to the neglect of $-\,i\omega\mathrm{D}$. Much higher in the spectrum, in simulations related to some medical practices such as hyperthermia [GF], where frequencies are on the order of 10 to 50 MHz, the conductivity of tissues on the order of 0.1 to 1, and with $\varepsilon \sim 10$ to $90\ \varepsilon_0$ [K&], one still gets a ratio $\varepsilon\,\omega\,/\sigma$ lower than 0.3, and the low-frequency approximation may still be acceptable, depending on the intended use. "Low frequency" is a very relative concept.

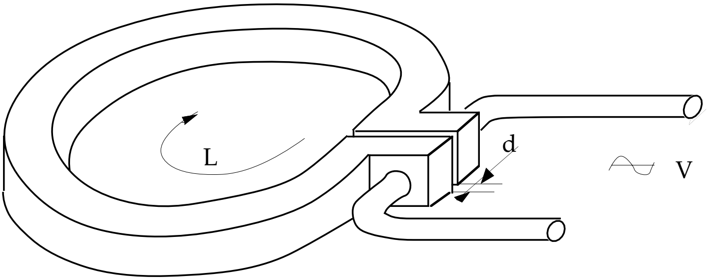


FIGURE 8.3. A case where capacitive effects may not be negligible (d << L).

But there are circumstances in which the electric field outside conductors is far larger than inside, and these are as many special cases. Consider Fig. 8.3, for instance, where L is the length of the loop and d the width of the gap. A simple computation (based on the relation $V \sim d\,|\,\mathrm{E}\,|$) shows that the ratio of $\varepsilon\,\omega\mathrm{E}$ *in the gap* to the current density *in the conductor* is on the order of $\varepsilon\,\omega\,L/d\sigma$, and thus may cease to be negligible when the ratio L/d gets large. This simply amounts to saying that the *capacitance* C of this gap, which is in ε/d, cannot be ignored in the computation when its product by the *resistance* R, which is in L/σ, reaches the order of ω^{-1} (recall that RC, whose dimension is that of a time interval, is precisely the time constant of a circuit of resistance R and capacitance C). One can assert in general that dropping $i\omega\mathrm{D}$ from the equations amounts to neglecting *capacitive effects*. This is a legitimate approximation when the energy of the electromagnetic field is mainly stored in the "magnetic" compartment, as opposed to the "electric" one, in the language of Chapter 1.

One then has, instead of (7), $\mathrm{rot}\ \mathrm{H} = \mathrm{J}$. Consequently, $\mathrm{div}\ \mathrm{J} = 0$, and if the supports of J^g and σ are disjoint, one must assume $\mathrm{div}\ \mathrm{J}^g = 0$, after (11).

To sum up, we are interested in the family of problems that Fig. 8.4 depicts: a bounded conductor, connected, a given harmonic current, with bounded support, not encountering the conductor. (The last hypothesis is sensible in view of Fig. 8.1, but is not valid in all conceivable situations.) The objective is to determine the field H, from which the eddy currents of density $\mathrm{J} = \mathrm{rot}\,\mathrm{H}$ in the conductor will be derived.

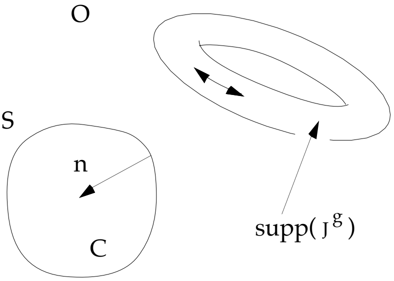


FIGURE 8.4. The theoretical situation. Note the convention about the normal unitary field on S, here taken as outgoing with respect to the "outer domain" $O = E_3 - C$.

From the mathematical point of view now, let thus C be a bounded domain of space, S its surface (Fig. 8.4), and J^g a given complex-valued field, such that $\mathrm{supp}(\mathrm{J}^g) \cap C = \varnothing$, and $\mathrm{div}\ \mathrm{J}^g = 0$. Again, we denote by O the domain which complements $C \cup S$, that is to say, the topological interior of $E_3 - C$ (take notice that O contains $\mathrm{supp}(\mathrm{J}^g)$). Domains C and O have S as common boundary, and the field of normals to S is taken as outgoing with respect to O. Conductivity σ and permeability μ being given, with $\mathrm{supp}(\sigma) = C$, and $\sigma_1 \geq \sigma(x) \geq \sigma_0 > 0$ on C, where σ_0 and σ_1 are two constants, as well as $\mu(x) \geq \mu_0$ on C and $\mu(x) = \mu_0$ in O, one looks for $\mathrm{H} \in \mathbb{L}^2_{\mathrm{rot}}(E_3)$, complex-valued, such that

$$(12) \qquad i\omega\mu\,\mathrm{H} + \mathrm{rot}\,\mathrm{E} = 0, \quad \mathrm{J} = \mathrm{J}^g + \sigma\mathrm{E}, \quad \mathrm{rot}\,\mathrm{H} = \mathrm{J}.$$

8.1.3 The problem in H, in the harmonic regime

Let us set, as we did up to now, $\mathbb{H} = \mathbb{L}^2_{\mathrm{rot}}(E_3)$ (complex), and

$$\mathbb{H}^g = \{\mathrm{H} \in \mathbb{H} : \ \mathrm{rot}\,\mathrm{H} = \mathrm{J}^g \ \text{ in } O\},$$

$$\mathbb{H}^0 = \{\mathrm{H} \in \mathbb{H} : \ \mathrm{rot}\,\mathrm{H} = 0 \ \text{ in } O\}.$$

We shall look for H in $\mathbb{H}^g$. One has $\mathbb{H}^g = \mathrm{H}^g + \mathbb{H}^0$, with, as in magnetostatics (cf. the h^j of Chapter 7), $\mathrm{H}^g = \mathrm{rot}\,\mathrm{A}^g$, where

$$\mathrm{A}^{g}(x) = \frac{1}{4\pi}\int_{E_3} \frac{\mathrm{J}^{d}(y)}{|x-y|}\,dy.$$

It all goes as if the source of the field was H^g, that is, the magnetic field that would settle in the presence of the inductor alone in space. The difference $\tilde{\mathrm{H}} = \mathrm{H} - \mathrm{H}^g$ between effective field and source field is called *reaction field*.

Let us seek a weak formulation. From the first Eq. (12), and using the curl integration by parts formula, one has

$$0 = \int_{E_3} (i\omega\,\mu\,\mathrm{H} + \mathrm{rot}\,\mathrm{E})\cdot \mathrm{H}' = i\omega \int_{E_3} \mu\,\mathrm{H}\cdot \mathrm{H}' + \int_{E_3} \mathrm{E}\cdot \mathrm{rot}\,\mathrm{H}' \quad \forall\, \mathrm{H}' \in \mathbb{H}^0.$$

As $\mathrm{rot}\,\mathrm{H}' = 0$ *outside* C (this is the key point), one may eliminate E by using the other two equations (12): for $\mathrm{E} = \sigma^{-1}(\mathrm{J} - \mathrm{J}^g) \equiv \sigma^{-1}\mathrm{J}$ in C, and thus $\mathrm{E} = \sigma^{-1}\,\mathrm{rot}\,\mathrm{H}$. We finally arrive at the following prescription: *find* $\mathrm{H} \in \mathbb{H}^g$ *such that*

$$(13)\qquad \int_{E_3} i\,\omega\,\mu\,\mathrm{H}\cdot \mathrm{H}' + \int_C \sigma^{-1}\,\mathrm{rot}\,\mathrm{H}\cdot \mathrm{rot}\,\mathrm{H}' = 0 \quad \forall\, \mathrm{H}' \in \mathbb{H}^0.$$

Proposition 8.1. *If* $\mathrm{H}^g \in \mathbb{L}^2_{\mathrm{rot}}(E_3)$, *problem* (13) *has a unique solution* H, *and the mapping* $\mathrm{H}^g \to \mathrm{H}$ *is continuous from* $\mathbb{L}^2(E_3)$ *into* $\mathbb{H}$.

Proof. Let us look for H in the form $\tilde{\mathrm{H}} + \mathrm{H}^g$. After multiplication of both sides by $1 - i$, the problem (13) takes the form $a(\tilde{\mathrm{H}}, \mathrm{H}') = L(\mathrm{H}')$, where L is continuous on $\mathbb{H}$, and

$$a(\tilde{\mathrm{H}}, \mathrm{H}') = \int_{E_3} \omega\,\mu\,\tilde{\mathrm{H}}\cdot \mathrm{H}' + \int_C \sigma^{-1}\,\mathrm{rot}\,\tilde{\mathrm{H}}\cdot \mathrm{rot}\,\mathrm{H}'$$
$$+\, i\,\Big(\int_{E_3} \omega\,\mu\,\tilde{\mathrm{H}}\cdot \mathrm{H}' - \int_C \sigma^{-1}\,\mathrm{rot}\,\tilde{\mathrm{H}}\cdot \mathrm{rot}\,\mathrm{H}'\Big).$$

As one sees, $\mathrm{Re}[a(\tilde{\mathrm{H}}, \tilde{\mathrm{H}}^*) \geq C\,(\int_{E_3} |\tilde{\mathrm{H}}|^2 + \int_C |\mathrm{rot}\,\tilde{\mathrm{H}}|^2)$ for some positive constant C. This is the property of *coercivity* on $\mathbb{H}^0$ under which Lax–Milgram's lemma of A.4.3 applies, in the complex case, hence the result. ◊

Remark 8.1. Problem (13) amounts to looking for the point of stationarity ("critical point") of the complex quantity

$$\mathcal{Z}(\mathrm{H}) = i\omega \int_{E_3} \mu\,\mathrm{H}^2 + \int_C \sigma^{-1}\,(\mathrm{rot}\,\mathrm{H})^2,$$

when H spans $\mathbb{H}^g$. (There is a tight relationship between $\mathcal{Z}$ and what is called the *impedance* of the system.) So it's not a variational problem in the strict sense, and the minimization approach of former chapters is no longer available. By contrast, this emphasizes the importance of Lax–Milgram's lemma. ◊

The electric field has thus disappeared from this formulation. One easily retrieves it in C, where $E = \sigma^{-1} \text{rot}\, H$. One may find it outside C by solving a static problem[2], formally similar to magnetostatics in region O, but this is rarely called for. (And anyway, this outside field would be fictitious, as already pointed out.)

To simplify, and to better emphasize the basic ideas, we first consider, in Section 8.2 below, the case when the passive conductor is contractible (i.e., simply connected with a connected boundary, cf. A.2.3), as in Fig. 8.4. It's obviously too strong a hypothesis (it doesn't hold in the above case), but the purpose is to focus on the treatement of the outer region, by the same method as for "open space" magnetostatics in Chapter 7. In Section 8.3, we'll reintroduce loops, but forfeit infinite domains, thus separately treating the two main difficulties in eddy-current computation.

8.2 INFINITE DOMAINS: "TRIFOU"

The key idea of the method to be presented now, already largely unveiled by the treatment of open-space magnetostatics of Chapter 7, is to reduce the computational domain to the conductor, in order not to discretize the air region[3] around. The method, implemented under the code name "Trifou", was promoted by J.C. Vérité and the author from 1980 onwards [B1, BV, BV'], and provided at the time the first solution of general applicability to the three-dimensional eddy-currents problem.

8.2.1 Reduction to a problem on C

We tackle problem (13), assuming C contractible (no current loops, no non-conductive hole inside C). In that case, the outside region O also is contractible.

[2] Namely the following problem: rot $E = -i\omega\mu\, H$, $D = \varepsilon_0\, E$, div $D = Q$ in O, with $n \times E$ known on the boundary S, where Q is the density of electric charge outside C. The difficulty is that the latter is not known in region I, for lack of information on the fine structure of the inductor. One may assume $Q = 0$ with acceptable accuracy if the objective is to obtain E near C (hence in particular the surface charge on S, which is $\varepsilon_0\, n \cdot E$). Such information may be of interest in order to appraise the magnitude of capacitive effects.

[3] It's not always advisable thus to reduce the computational domain D to the passive conductor C. It's done here for the sake of maximum simplicity. But "leaving some air" around C may be a good idea, for instance, in the presence of small air gaps, conductors of complex geometry, and so forth. Methods for such cases will be examined in Section 8.3.

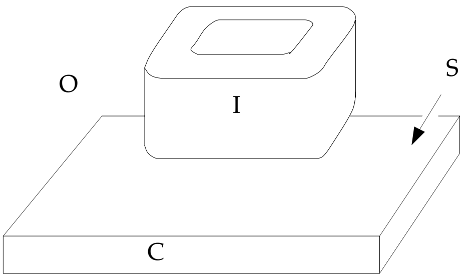


FIGURE 8.5. Model problem for the study of the hybrid approach in "Trifou", finite elements in the conductor C, and integral method over its boundary S to take the far field into account. The support of the given current density J^g is the inductor I. Contrary to Fig. 8.2, C here is loop-free, and we restrict consideration to this case to separate the difficulties. Section 8.3 will address loops (but shun the far-field effect).

Let's keep the notations of Chapter 7: Φ is the space of magnetic potentials (the Beppo Levi space of 7.2.1), Φ_O is composed of the restrictions to O of elements of Φ, and the set of elements of Φ_O that have in common the trace ϕ_S is denoted $\Phi_O(\phi_S)$. Let Φ^{00} stand[4] for the subspace $\{\phi \in \Phi : \phi = 0 \text{ on } C\}$. Set

$$\mathbb{K}^g = \{\mathrm{H} \in \mathbb{H}^g : \int_{E_3} \mathrm{H} \cdot \operatorname{grad} \phi' = 0 \quad \forall\, \phi' \in \Phi^{00}\}$$

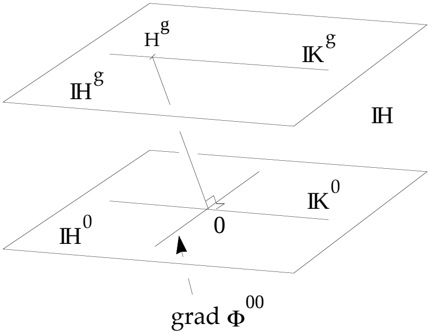
(the support of the integrand reduces to O, in fact), and

$$\mathbb{K}^0 = \{\mathrm{H} \in \mathbb{H}^0 : \int_{E_3} \mathrm{H} \cdot \operatorname{grad} \phi' = 0 \quad \forall\, \phi' \in \Phi^{00}\}.$$

We note that

$$(14) \qquad \mathbb{H}^0 = \mathbb{K}^0 \oplus \operatorname{grad} \Phi^{00},$$

by construction, and that $\mathbb{K}^g = \mathrm{H}^g + \mathbb{K}^0$, for H^g is orthogonal to grad Φ^{00}, since div $\mathrm{H}^g = 0$. (Cf. the inset drawing.)



By their very definition, the elements of $\mathbb{K}^g$ and of the parallel subspace $\mathbb{K}^0$ satisfy div $\mathrm{H} = 0$ in O. This property is shared by the required solution, since div $\mathrm{H} = \mu_0^{-1}$ div $\mathrm{B} = 0$ in O. One may therefore expect to find this solution in $\mathbb{K}^g$. Which is indeed the case:

[4] The notation Φ^0 is reserved for an analogous, but slightly larger space (see below).

Proposition 8.2. *The solution* H *of Problem* (13) *lies in* $\mathbb{K}^g$.

Proof. By letting $\text{H}' = \text{grad}\, \phi'$ in (13), where ϕ' roams in Φ^{00}, one gets

$$0 = i\omega \int_{E_3} \mu\, \text{H} \cdot \text{grad}\, \phi' = i\omega\, \mu_0 \int_{E_3} \text{H} \cdot \text{grad}\, \phi' \quad \forall\, \phi' \in \Phi^{00},$$

and hence $\text{H} \in \mathbb{K}^g$. ◊

To find the point of stationarity of $\text{H} \rightarrow \mathcal{Z}(\text{H})$ in $\mathbb{H}^g$, it is thus enough to look for it in $\mathbb{K}^g$, and to check that no other, spurious critical point is in the way. Indeed,

Corollary of Prop. 8.2. *Problem* (13) *is equivalent to* find $\text{H} \in \mathbb{K}^g$ such that

$$\text{(15)} \qquad i\omega \int_{E_3} \mu\, \text{H} \cdot \text{H}' + \int_C \sigma^{-1}\, \text{rot}\, \text{H} \cdot \text{rot}\, \text{H}' = 0 \quad \forall\, \text{H}' \in \mathbb{K}^0,$$

since this is the Euler equation for the search of critical points of $\mathcal{Z}$ in the affine subspace $\mathbb{K}^g$, and it has at most one solution.

Our effort, now, will concentrate on showing that Problem (15) is in fact "posed on C", meaning that a field in $\mathbb{K}^g$ (or in $\mathbb{K}^0$) is entirely determined by its restriction to C. I expect this to be obvious "on physical grounds", but this doesn't make the proofs any shorter. We are embarked on a long journey, till the end of 8.2.3. The operator P of Chapter 7 will play a prominent part in these developments.

Remark 8.2. Set $\Phi^0 = \{\phi \in \Phi : \text{grad}\, \phi = 0 \text{ on } C\}$. By introducing as before the orthogonal subspaces $\mathbb{K}^{g0}$ and $\mathbb{K}^{00}$, one would have $\mathbb{H}^0 = \mathbb{K}^{00} \oplus \text{grad}\, \Phi^0$, and one could proceed with the same kind of reduction, with $\mathbb{K}^{g0}$ strictly contained in $\mathbb{K}^g$. This may look like an advantage, but in practice, it makes little difference. ◊

Exercise 8.1. Show that $\int_S n \cdot \text{H} = 0$, and prove the analogue of Prop. 8.2 in the context suggested by Remark 8.2.

8.2.2 The space HΦ, isomorphic to $\mathbb{K}^g$

Let now HΦ (treated as a single symbol) stand for the vector space of pairs $\{\text{H}, \phi_S\}$, where H is a field supported on C and ϕ_S a surface potential "associated with" H, in the precise following sense[5]:

$$\text{(16)} \qquad H\Phi = \{\{\text{H}, \phi_S\} \in \mathbb{L}^2_{\text{rot}}(C) \times H^{1/2}(S) : \text{H}_S = \text{grad}_S\, \phi_S\},$$

where grad_S denotes the surface gradient. Note that the projection of HΦ on the first factor of the Cartesian product $\mathbb{L}^2_{\text{rot}}(C) \times H^{1/2}(S)$ is not

[5] Refer to Fig. 2.5 for H_S, the tangential part of H.

$\mathbb{L}^2_{\mathrm{rot}}(C)$ in its entirety, for there are constraints that H must satisfy, in particular $n \cdot \mathrm{rot}\, \mathrm{H} = 0$ on S. On the other hand, Φ_S may be any function in $H^{1/2}(S)$. One provides $H\Phi$ with its natural Hilbertian norm, induced by the norm of the encompassing Cartesian product. Then,

Proposition 8.3. $H\Phi$ *is isomorphic to* $\mathbb{K}^g$ *and* $\mathbb{K}^0$.

Proof. Since C is simply connected, and $\mathrm{rot}\, \mathrm{H}^g = 0$ in C, there exists $\Phi^g \in L^2_{\mathrm{grad}}(C)$ such that $\mathrm{H}^g = \mathrm{grad}\, \Phi^g$ in C. Let us still denote by Φ^g the harmonic continuation of this function outside C. Now, take $\mathrm{H} \in \mathbb{K}^g$. One has $\mathrm{rot}\, \mathrm{H} = \mathrm{J}^g = \mathrm{rot}(\mathrm{H}^g - \mathrm{grad}\, \Phi^g)$ in O. Since O is simply connected, there exists a unique Φ in $BL(O)$ such that the equality $\mathrm{H} = \mathrm{H}^g + \mathrm{grad}(\Phi - \Phi^g)$ hold in O. By restricting H and Φ to C and S, a map from H to the pair $\{\mathrm{H}, \Phi_S\}$ of $H\Phi$ is therefore defined. Conversely, such a pair $\{\mathrm{H}(C), \Phi_S\}$ being given, let Φ be the exterior harmonic continuation of Φ_S. Set H equal to $\mathrm{H}(C)$ in C and to $\mathrm{H}^g + \mathrm{grad}(\Phi - \Phi^g)$ outside C. This enforces $\mathrm{H} \in \mathbb{L}^2_{\mathrm{rot}}(E_3)$, because both tangential traces $\mathrm{H}(C)_S$ and $\mathrm{H}^g_S + \mathrm{grad}_S(\Phi_S - \Phi^g_S) \equiv \mathrm{grad}_S \Phi_S$ coincide, after (16). In O, $\mathrm{rot}\, \mathrm{H} = \mathrm{J}^g$ and $\mathrm{div}\, \mathrm{H} = 0$ by construction, whence $\mathrm{H} \in \mathbb{K}^g$. Moreover, the one-to-one correspondence thus established (in a way that would apply as well to $\mathbb{K}^0$, just consider the special case $\mathrm{J}^g = 0$) is an isometry. Hence the announced isomorphisms: For if H is the difference between two elements of $\mathbb{K}^g$, then $\mathrm{H} = \mathrm{grad}\, \Phi$ outside C, and

$$\int_{E_3} |\mathrm{H}|^2 + \int_C |\mathrm{rot}\, \mathrm{H}|^2 = \int_C |\mathrm{H}|^2 + \int_C |\mathrm{rot}\, \mathrm{H}|^2 + \int_O |\mathrm{grad}\, \Phi|^2$$
$$= \int_C |\mathrm{H}|^2 + \int_C |\mathrm{rot}\, \mathrm{H}|^2 + \int_S P\Phi_S\, \Phi_S,$$

which is indeed the square of the norm of the pair $\{\mathrm{H}, \Phi_S\}$ in the product $\mathbb{L}^2_{\mathrm{rot}}(C) \times H^{1/2}(S)$. ◊

Remark 8.3. The isomorphism depends of course on Φ^g, which in turn depends on J^g, up to an additive constant. ◊

8.2.3 Reformulation of the problem in HΦ

So, let Φ^g_S be the function on S associated with J^g that specifies the above isomorphism. One has

$$\mathcal{Z}(\mathrm{H}) = i\omega \int_{E_3} \mu\, \mathrm{H}^2 + \int_C \sigma^{-1} (\mathrm{rot}\, \mathrm{H})^2$$
$$= i\omega \int_C \mu\, \mathrm{H}^2 + \int_C \sigma^{-1} (\mathrm{rot}\, \mathrm{H})^2 + i\omega\, \mu_0 \int_O (\mathrm{H}^g + \mathrm{grad}(\Phi - \Phi^g))^2,$$

and we are looking for the critical point of this function in $H\Phi$. Since, thanks to the properties of P,

$$\int_O (\mathrm{H}^g + \mathrm{grad}(\Phi - \Phi^g))^2 = \int_O (\mathrm{H}^g - \mathrm{grad}\,\Phi^g)^2 + 2\int_S \mathrm{n}\cdot \mathrm{H}^g\ \Phi_S$$
$$+ \int_S \mathrm{P}\Phi_S\,\Phi_S - 2\int_S \mathrm{P}\Phi^g{}_S\,\Phi_S\,,$$

Problem (15), equivalent to the initial problem (13) under our assumptions, amounts to finding the critical point of the function $\tilde{\mathcal{Z}}$ (equal to $\mathcal{Z}$ up to a constant) thus defined:

$$\tilde{\mathcal{Z}}(\{\mathrm{H}, \Phi_S\}) = i\omega\,[\int_C \mu\,\mathrm{H}^2 + \mu_0 \int_S \mathrm{P}\Phi_S\,\Phi_S] + \int_C \sigma^{-1}\,(\mathrm{rot}\ \mathrm{H})^2$$
$$+ 2\,i\omega\ \mu_0 \int_S (\mathrm{n}\cdot\mathrm{H}^g - \mathrm{P}\Phi^g{}_S)\ \Phi_S,$$

whence, by taking the Euler equation, the following result (index S is understood in Φ_S, Φ'_S and $\Phi^g{}_S$):

Proposition 8.4. *When* C *is contractible,* (13) *is equivalent to* find $\{\mathrm{H}, \Phi\}$ in $H\Phi$ such that

$$(17)\qquad i\omega\,[\int_C \mu\,\mathrm{H}\cdot\mathrm{H}' + \mu_0 \int_S \mathrm{P}\Phi\,\Phi'] + \int_C \sigma^{-1}\,\mathrm{rot}\,\mathrm{H}\cdot\mathrm{rot}\,\mathrm{H}'$$
$$= i\omega\ \mu_0 \int_S (\mathrm{P}\Phi^g - \mathrm{n}\cdot\mathrm{H}^g)\ \Phi' \quad \forall\ \{\mathrm{H}', \Phi'\} \in H\Phi.$$

This is the final weak formulation, on which "Trifou" was based. The pending issue is how to discretize it. Clearly, H and H' in (17) will be represented by edge elements, and Φ and Φ' by surface nodal elements (these are compatible representations, thanks to the structural properties of the Whitney complex). The discretization of terms $\int_S \mathrm{P}\Phi\,\Phi'$ and $\int_S \mathrm{P}\Phi^g\,\Phi'$ by Galerkin's method will make use of the matrix **P** of Chapter 7 (Subsection 7.4.5, Eq. (48)).

8.2.4 Final discrete formulation

Let $\mathrm{K} = \{\mathbf{H}_e : \ e \in \mathcal{E}^0(C);\ \ \mathbf{\Phi}_n :\ n \in \mathcal{N}(S)\}$ be the vector of all degrees of freedom (complex valued): one DoF $\mathbf{H}_e$ for each edge *inside* C (that is, not contained in S) and one DoF $\mathbf{\Phi}_n$ for each surface node. The expression of H in C is thus

$$\mathrm{H} = \sum_{e \in \mathcal{E}^0(C)} \mathbf{H}_e\, w_e + \sum_{n \in \mathcal{N}(S)} \mathbf{\Phi}_n\, \mathrm{grad}\ w_n.$$

(This is an element of $W^1{}_m(C)$, thanks to the inclusion $\mathrm{grad}\ W^0 \subset W^1$.) Then (17) becomes

$$(18)\qquad i\omega\ (\mathbf{M}(\mu) + \mu_0\,\mathbf{P})\,\mathrm{K} + \mathbf{N}(\sigma)\,\mathrm{K} = i\omega\ \mu_0\,(\mathbf{P}\mathbf{\Phi}^g - \mathrm{L}^g),$$

with obvious notations, except for L^g, defined by $\mathrm{L}^g{}_m = \int_S \mathrm{n}\cdot\mathrm{H}^g\ w_m$ for all

$m \in \mathcal{N}(S)$, and other components null. (Beware, though $\mathbf{M}$ is very close to the mass-matrix $\mathbf{M}_1(\mu)$ of the mesh of C, it's not quite the same, just as $\mathbf{N}(\sigma)$ is not quite $\mathbf{R}^t\mathbf{M}_2(\sigma^{-1})\mathbf{R}$.) As in Chapter 7, $\mathbf{P}$ bears only on the "Φ part" of vector K, a priori, but is bordered by zeroes in order to give sense to (18). (Same remark about $\mathbf{P}\Phi^g$.) Matrices $\mathbf{M}$, $\mathbf{P}$, and $\mathbf{N}$ are symmetric, but because of the factor i, the matrix of the linear system (18) is not Hermitian. In spite of this, the conjugate gradient method, the convergence of which is only guaranteed in the Hermitian case, in theory, works fairly well in practice, with proper conditioning.

Computing the right-hand side of (18) is straightforward: H^g is known by the Biot and Savart formula, and the vector Φ^g of nodal values is derived from the circulations of H^g along the edges of S.

8.2.5 Transient regimes

From (18) to a scheme for the temporal evolution problem is but a short trip, compared to what has been done, so let's show the way, even though this is beyond the objectives of this chapter. Now the given current j^g is real-valued and time-dependent, and the field h is given at time 0, with $\operatorname{div}(\mu h(0)) = 0$. The DoFs are real-valued again. The evolution scheme, discrete in space but not yet in time, is

$$\partial_t[(\mathbf{M}(\mu) + \mu_0\,\mathbf{P})\,\mathbf{k}] + \mathbf{N}(\sigma)\,\mathbf{k} = \mu_0\,\partial_t(\mathbf{P}\boldsymbol{\varphi}^g - \mathbf{l}^g) \tag{19}$$

($\mathbf{k}(0)$ given by the initial conditions, and $\mathbf{l}$ similar to L, but real-valued). Over a temporal interval $[0, T]$, with a time step δt, one may treat this by a Crank–Nicolson scheme [CN] : $\mathbf{k}^0 = \mathbf{k}(0)$, then, for each integer m from 0 to $T/\delta t - 1$,

$$\begin{aligned}(\mathbf{M}(\mu) + \mu_0\,\mathbf{P})\,(\mathbf{k}^{m+1} - \mathbf{k}^m) + \delta t\ \mathbf{N}(\sigma)\,(\mathbf{k}^{m+1} + \mathbf{k}^m)/2 \\ = \mu_0\,\mathbf{P}\,[\boldsymbol{\varphi}^g((m+1)\,\delta t) - \boldsymbol{\varphi}^g(m\delta t)] + \mathbf{l}^g((m+1)\,\delta t) - \mathbf{l}^g(m\,\delta t).\end{aligned} \tag{20}$$

If $\mathbf{k}^m$ is known, this is a linear system with respect to the unknown $\mathbf{k}^{m+1}$. (Actually, better take $\mathbf{k}^{m+1/2} \equiv (\mathbf{k}^{m+1} + \mathbf{k}^m)/2$ as unknown. Then $\mathbf{k}^{m+1} - \mathbf{k}^m = 2(\mathbf{k}^{m+1/2} - \mathbf{k}^m)$. Unconditionally stable as it may be, the Crank–Nicolson scheme may suffer from numerical oscillations ("weak instability"), to which the sequence of the $\mathbf{k}^{m+1/2}$s is less sensitive than the $\mathbf{k}^m$s.)

This scheme can easily be adapted to the case of a nonlinear b–h law. See [Bo] for details.

Remark 8.4. When j^g is sinusoidal, using this scheme is a viable alternative to solving (18) directly. One should use an informed guess of the solution as initial condition, and monitor the time average over a period, $(2\pi)^{-1}\omega\int_{t-2\pi/\omega}^{t} \exp(i\omega\ s)\ \mathbf{k}(s)\ ds$, duly approximated by a sum of the kind $N^{-1} \sum_{j=1,\,N} \exp(i\omega(m + 1/2 - j)\delta t)\ \mathbf{k}^{m+1/2-j}$, where $N\delta t = 2\pi/\omega$, the period. This will converge, relatively rapidly (no more than three or four periods, in practice) towards the solution κ of (18). Such a "time domain" approach to the harmonic problem can thus be conceived as another iterative scheme to solve (18). Fast Fourier Transform techniques make the calculation of time averages quite efficient. ◊

8.3 BOUNDED DOMAINS: TREES, H–Φ

Now we change tack. Let D be a simply connected domain containing the region of interest (here the conductor C and its immediate neighborhood, Fig. 8.6), the inductor I, magnetic parts where $\mu \neq \mu_0$, if such exist, and suppose either that D is big enough so that one can assume a zero field beyond, or that $n \times h = 0$ on ∂D for physical reasons (as in the case of a cavity with ferromagnetic walls, used as a shield to confine the field inside). We thus forget about the far field and concentrate on difficulties linked with the *degeneracy* of the eddy-current equations in regions where $\sigma = 0$.

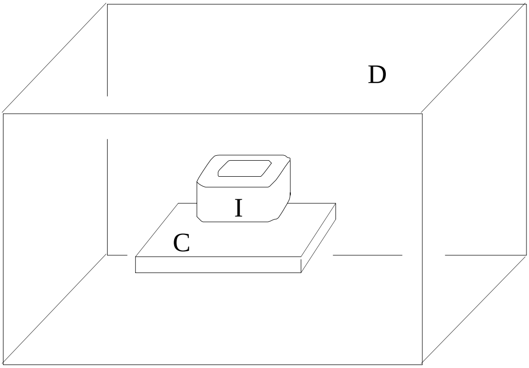


FIGURE 8.6. Computational domain D, containing the region of interest, and large enough for the boundary condition $n \times h = 0$ on ∂D to be acceptable. (In practice, the size of the elements would be graded in such a case, the farther from C the bigger, the same as with Fig. 7.2.)

8.3.1 A constrained linear system

Let m be a simplicial mesh of D. Functional spaces, a bit different now, are $\mathbb{H} = \{\mathrm{H} \in \mathbb{L}^2_{\mathrm{rot}}(\mathrm{D}) : \mathrm{n} \times \mathrm{H} = 0 \text{ on } \partial\mathrm{D}\}$ and

(21) $\mathbb{H}^g = \{\mathrm{H} \in \mathbb{H} : \mathrm{rot}\, \mathrm{H} = \mathrm{J}^g \text{ in } \mathrm{D} - \mathrm{C}\}$,

(22) $\mathbb{H}^0 = \{\mathrm{H} \in \mathbb{H} : \mathrm{rot}\, \mathrm{H} = 0 \text{ in } \mathrm{D} - \mathrm{C}\}$.

Now we know the paradigm well, and we can state the problem to solve without further ado: *find* $\mathrm{H} \in \mathbb{H}^g$ *such that*

(23) $\int_{\mathrm{D}} i\,\omega\,\mu\,\mathrm{H} \cdot \mathrm{H}' + \int_{\mathrm{C}} \sigma^{-1}\, \mathrm{rot}\, \mathrm{H} \,.\, \mathrm{rot}\, \mathrm{H}' = 0 \quad \forall\, \mathrm{H}' \in \mathbb{H}^0$.

As we intend to enforce null boundary conditions on the boundary of D, let us remove from $\mathcal{N}$, $\mathcal{E}$, $\mathcal{F}$ the boundary simplices, as we did earlier in 7.3.1, and for convenience, still call $\mathcal{N}$, $\mathcal{E}$, $\mathcal{F}$, $\mathcal{T}$ the simplicial sets of this "peeled out" mesh. Apart from this modification, the notation concerning the spaces W^{p} and the incidence matrices is the same as before. In particular, W^1_m is the span—with *complex* coefficients, this time—of Whitney edge elements w_{e}, for all e in $\mathcal{E}$. As this amounts to have null circulations along the boundary edges, a field in W^1_m can be prolongated by 0 to all space, the result being tangentially continuous and therefore an element of $\mathbb{L}^2_{\mathrm{rot}}(\mathrm{E}_3)$. So we can identify W^1_m with a subspace of $\mathbb{L}^2_{\mathrm{rot}}(\mathrm{E}_3)$. Let us set $\mathbb{H}_m = \mathrm{W}^1_m$, denote by $\mathrm{I}\mathbf{H}$ the isomorphic finite-dimensional space $\mathbb{C}^{\mathcal{E}}$, composed of all vectors $\mathrm{H} = \{\mathrm{H}_{\mathrm{e}} : \mathrm{e} \in \mathcal{E}\}$, and call $\mathrm{E} = \#(\mathcal{E})$ the number of inner edges. For $\mathbf{u}$ and $\mathbf{u}'$ both in $\mathrm{I}\mathbf{H}$, we set

$$\begin{aligned}(\mathbf{u}, \mathbf{u}') &= \sum_{\mathrm{e} \in \mathcal{E}} \mathbf{u}_{\mathrm{e}} \cdot \mathbf{u}'_{\mathrm{e}} \\ &\equiv \sum_{\mathrm{e} \in \mathcal{E}} (\mathrm{Re}[\mathbf{u}_{\mathrm{e}}] + i\,\mathrm{Im}[\mathbf{u}_{\mathrm{e}}]) \cdot (\mathrm{Re}[\mathbf{u}'_{\mathrm{e}}] + i\,\mathrm{Im}[\mathbf{u}'_{\mathrm{e}}]).\end{aligned}$$

(Again, beware: This is not the Hermitian scalar product.) This way, an integral of the form $\int_{\mathrm{D}} \alpha\, \mathrm{u} \cdot \mathrm{u}'$, where α is a function on D (such as μ, for instance), possibly complex-valued, is equal to $(\mathbf{M}_1(\alpha)\, \mathbf{u}, \mathbf{u}')$ when $\mathrm{u} = \sum_{\mathrm{e} \in \mathcal{E}} \mathbf{u}_{\mathrm{e}} \mathrm{w}_{\mathrm{e}}$ and $\mathrm{u}' = \sum_{\mathrm{e} \in \mathcal{E}} \mathbf{u}'_{\mathrm{e}} \mathrm{w}_{\mathrm{e}}$.

We know from experience the eventual form of the discretized problem: It will be *find* $\mathrm{H} \in \mathbb{H}^g_m$ *such that*

(24) $\int_{\mathrm{D}} i\,\omega\,\mu\,\mathrm{H} \cdot \mathrm{H}' + \int_{\mathrm{C}} \sigma^{-1}\, \mathrm{rot}\, \mathrm{H} \,.\, \mathrm{rot}\, \mathrm{H}' = 0 \quad \forall\, \mathrm{H}' \in \mathbb{H}^0_m$,

where $\mathbb{H}^g_m$ and $\mathbb{H}^0_m$ are parallel subspaces of $\mathbb{H}_m$.

But these subspaces must be constructed with some care. One cannot simply set $\mathbb{H}^g{}_m = W^1{}_m \cap \mathbb{H}^g \equiv \{\textsc{h} \in W^1{}_m : \operatorname{rot} \textsc{h} = \textsc{j}^g \text{ in } D - C\}$, because $\textsc{j}^g$ has no reason to be mesh-wise constant (which $\operatorname{rot} \textsc{h}$ is, if $\textsc{h} \in W^1{}_m$), although this happens frequently. Failing that, $W^1{}_m \cap \mathbb{H}^g$ may very well reduce to $\{0\}$. So we need to find a subspace of $W^1{}_m$ that closely approximate $W^1{}_m \cap \mathbb{H}^g$. For this, let $\mathcal{F}_C \subset \mathcal{F}$ be the subset of "conductive" faces, i.e., those *inside* C, faces of its boundary being excluded. Let us set

$$\mathbb{H}^g{}_m = \{\textsc{h} \in W^1{}_m : \textstyle\int_f n \cdot \operatorname{rot} \textsc{h} = \int_f n \cdot \textsc{j}^g \quad \forall\, f \notin \mathcal{F}_C\}.$$

This time, $\mathbb{H}^g{}_m$ is in $\mathbb{H}_m \equiv W^1{}_m$, and the isomorphic space $\boldsymbol{\mathbb{H}}^g$ is characterized by

$$(25) \qquad \boldsymbol{\mathbb{H}}^g = \{\mathbf{\textsc{h}} \in \boldsymbol{\mathbb{H}} : (\mathbf{R}\, \mathbf{\textsc{h}})_f = \mathbf{\textsc{j}}_f^g \quad \forall\, f \notin \mathcal{F}_C\},$$

where $\mathbf{\textsc{j}}_f^g$ is the intensity $\int_f n \cdot \textsc{j}^g$ through face f and $\mathbf{R}$ the edge-to-faces incidence matrix. As in Chapter 6, we shall abridge all this as follows: $\boldsymbol{\mathbb{H}}^g = \{\mathbf{\textsc{h}} \in \boldsymbol{\mathbb{H}} : \mathbf{L}\mathbf{\textsc{h}} = \mathbf{\textsc{l}}^g\}$, where $\mathbf{L}$ is a submatrix of $\mathbf{R}$, the dimensions of which are $(F - F_C) \times E$ (F inner faces in D, minus the F_C conductive faces) and $\mathbf{\textsc{l}}^g$ a known vector. Denoting by $\boldsymbol{\mathbb{H}}^0$ the kernel of $\mathbf{L}$ in $\boldsymbol{\mathbb{H}}$, we may now reformulate (24) like this: *find* $\mathbf{\textsc{h}} \in \boldsymbol{\mathbb{H}}^g$ *such that*

$$(26) \qquad i\,\omega\, (\mathbf{M}_1(\mu)\, \mathbf{\textsc{h}}, \mathbf{\textsc{h}}') + (\mathbf{M}_2(\sigma^{-1})\, \mathbf{R}\mathbf{\textsc{h}}, \mathbf{R}\mathbf{\textsc{h}}') = 0 \quad \forall\, \mathbf{\textsc{h}}' \in \boldsymbol{\mathbb{H}}^0.$$

This is, as in Chapter 6, a constrained linear system, which can be rewritten as follows:

$$(27) \qquad \begin{vmatrix} i\,\omega\, \mathbf{M}_1(\mu) + \mathbf{R}^t \mathbf{M}_2(\sigma^{-1})\mathbf{R} & \mathbf{L}^t \\ \mathbf{L} & 0 \end{vmatrix} \begin{vmatrix} \mathbf{\textsc{h}} \\ \mathbf{\textsc{v}} \end{vmatrix} = \begin{vmatrix} 0 \\ \mathbf{\textsc{l}}^g \end{vmatrix},$$

where the dimension of the vector-valued Lagrange multiplier $\mathbf{\textsc{v}}$ is $F - F_C$.

Exercise 8.2. Find a physical interpretation for the $\mathbf{\textsc{v}}_f$s.

One can very well tackle the system (27) as it stands (cf. Appendix B). But for the same reasons as in Chapter 6, one may prefer to use a representation of $\boldsymbol{\mathbb{H}}^g$ and $\boldsymbol{\mathbb{H}}^0$ in terms of *independent* degrees of freedom. There are two main ways to do that.

8.3.2 The tree method

We suppose C contractible for a while. Let $\mathcal{N}_C$, $\mathcal{E}_C$, $\mathcal{F}_C$ denote the sets of nodes, edges, and faces inside C. Let us form a spanning tree $\mathcal{E}^T$ (cf. 5.3.2) for the mesh of the closure of $D - C$. (Some edges of the interface ∂C will belong to $\mathcal{E}^T$, and form a spanning tree for this surface.) Let's recall that for each edge of $\mathcal{E} - \mathcal{E}_C - \mathcal{E}^T$, or co-edge, one can build a closed chain over $D - C$ by adding to it some edges of $\mathcal{E}^T$, with uniquely determined coefficients. The idea is to select as independent DoFs the mmf's along the E_C edges inside C and the E^T edges of the tree. Then, the mmf's along the co-edges can be retrieved by using (25), as explained in the next paragraph. This vector of independent DoFs will be denoted u. The corresponding vector space, isomorphic to $\mathbb{C}^{E_C + E^T}$, is denoted $\mathbf{U}$.

For each edge e of $\mathcal{E}_C \cup \mathcal{E}^T$, set $\mathrm{H}_e = \mathrm{u}_e$. Any other edge $e \in \mathcal{E}$ is a co-edge, and is thus the closing edge of a circuit all other edges of which come from $\mathcal{E}^T$, by construction. Call $C(e) \subset \mathcal{E}^T$ the set composed of these other edges, and $\mathbf{c}_\varepsilon$ the chain-coefficient assigned to edge $\varepsilon \in C(e)$ by the procedure of 5.3.2. The circuit they form bounds a two-sided[6] and hence orientable polyhedral surface Σ_e, formed of faces of $\mathcal{F} - \mathcal{F}_C$. Each of the two possible fields of normals on Σ_e orients its boundary $\partial\Sigma_e$, as we saw in Chapter 5. Let n be the one for which e and $\partial\Sigma_e$ are oriented the same way. Now, assign the value

$$(28) \qquad \mathrm{H}_e = \int_{\Sigma_e} n \cdot \mathrm{j}^g + \sum\nolimits_{\varepsilon \in C(e)} \mathbf{c}_\varepsilon \, \mathrm{u}_\varepsilon,$$

as DoF to co-edge e. This completes the mmf vector H, hence a field $\mathrm{h} = \sum_{e \in \mathcal{E}} \mathrm{H}_e \, w_e$, associated with u. We'll denote this correspondence by $\mathrm{h} = f(\mathrm{u}, \mathrm{j}^g)$. (Beware: u is a **vector**, h is a field.) Let now

$$(29) \qquad \mathbb{K}_m{}^g = \{\mathrm{h} : \mathrm{h} = \sum\nolimits_{e \in \mathcal{E}} \mathrm{H}_e \, w_e\} \equiv \{f(\mathrm{u}, \mathrm{j}^g) : \mathrm{u} \in \mathbf{U}\}$$

be the span of these fields in $\mathbb{H}_m$, and $\mathbb{K}_m{}^0$ be the parallel subspace, which is obtained by exactly the same construction, but with $\mathrm{j}^g = 0$. Thus constructed, $\mathbb{K}_m{}^g$ and $\mathbb{K}_m{}^0$ coincide with $\mathbb{H}_m{}^g$ and $\mathbb{H}_m{}^0$.

As a bonus, the above construction gives an approximation $\mathrm{h}^g{}_m \in W^1{}_m$ of the source-field h^g: the field corresponding to $\mathrm{u} = 0$, that is $\mathrm{h}^g{}_m = f(\mathbf{0}, \mathrm{j}^g)$. Note that $\mathbb{K}^g{}_m = \mathrm{h}^g{}_m + \mathbb{K}^0{}_m$, and hence $\mathbb{H}^g{}_m = \mathrm{h}^g{}_m + \mathbb{H}^0{}_m$ as well, that is to say,

$$(30) \qquad \mathbb{H}^g{}_m = \{\mathrm{h}^g{}_m + \mathrm{h} : \mathrm{h} \in \mathbb{H}^0{}_m\}.$$

[6] This is *not* supposed to be obvious (but please read on, and return to the present note at leisure). The circuit based on a co-edge can be a knot of arbitrary complexity, so it's not so clear that it always bounds an orientable and non-self-intersecting surface. But this is true, being a theorem in knot theory. Such a surface, called a *Seifert surface* (cf., e.g., [Ro]), always exists [Se], however tangled the knot may be. See Fig. 8.9 in Exercise 8.3 at the end.

Since $\mathbb{K}^g{}_m$ and $\mathbb{K}^0{}_m$ coincide with $\mathbb{H}^g{}_m$ and $\mathbb{H}^0{}_m$, all we have to do now is throw into (24) the expressions $\mathrm{h} = \mathrm{f}(\mathrm{u}, \mathrm{j}^g)$ and $\mathrm{h}' = \mathrm{f}(\mathrm{u}', 0)$ in order to obtain a linear system in terms of u, the form of which is

$$(i\omega\, \mathbf{M} + \mathbf{N})\, \mathrm{u} = \mathrm{l}^g, \tag{31}$$

with $\mathbf{M}$ and $\mathbf{N}$ symmetrical, non-negative definite, and $\mathbf{M} + \mathbf{N}$ regular. But this time $\mathbf{M}$ and $\mathbf{N}$ largely differ from $\mathbf{M}_1(\mu)$ and $\mathbf{R}^t\, \mathbf{M}_2(\sigma^{-1})\mathbf{R}$ (only the blocks relative to the edges of $\mathcal{E}_C$ coincide), and overall, their conditioning greatly depends on the tree construction. (To each spanning tree corresponds a particular basis for the space $\mathbb{H}_m{}^0$.) Not all trees are thus equivalent in this respect, and finding methods that generate good spanning trees is a current research subject.

The matrix $i\omega\, \mathbf{M} + \mathbf{N}$ is not Hermitian, and this raises specific algorithmic problems. So here begins the *numerical* work (to say nothing of the *programming* work, which is far from run-of-the-mill), but we shall stop there, because the *modelling* work is done—at least in the case when C is contractible.

So how can the technique be generalized to the non-contractible case? If there are only "holes", i.e., if C is simply connected but with a non-connected boundary, no difficulty: Just build a spanning tree for each connected component of D – C. The problem is with "loops". Suppose for definiteness there is a single loop in C, as in Fig. 8.6. Then, by a deep but intuitively obvious result of topology ("Alexander's duality", cf. [GH]), there is also one loop in D – C. There are now two kinds of co-edges, depending on whether the circuits they close surround the conductive loop or not. (Note that those which do surround the loop do *not* bound a polyhedral surface of the kind discussed above, that is, made of faces in $\mathcal{F} - \mathcal{F}_C$, and this is what characterizes them.) Next, select *one* of these loop co-edges, and add it to the initial tree, thus obtaining a "belted tree". Thanks to this added edge, the circuits of all remaining co-edges do bound, as we noticed in 5.3.2. Obviously (by Ampère), the DoF of the belt fastener is the intensity in the current loop. There is one additional DoF of this kind for each current loop. With this, the key result ($\mathbb{K}_m{}^g$ and $\mathbb{K}_m{}^0$ coincide with $\mathbb{H}_m{}^g$ and $\mathbb{H}_m{}^0$) stays valid, and everything elses goes unchanged.

8.3.3 The H–Φ method

The H–Φ method is "edge elements and nodal elements in association" and stems from the second way to obtain a set of independent degrees of freedom. With the previous method, the DoFs were all magnetomotive forces, those along the selected edges. Now, we'll have two different kinds of DoF: Besides the mmf's along edges inside C, there are others, associated with the nodes in the air and on the conductor's surface, which can be interpreted as nodal values of the magnetic scalar potential, as we shall see. Again, let us first treat the contractible case.

Let $\mathcal{E}_C$, as above, be the subset of edges *inside* C, that is, entirely contained, apart from the extremities, in the interior of C. The set $\mathcal{N} - \mathcal{N}_C$ is composed of the nodes which are neither in int(C), nor in ∂D. Let E_C be the number of edges in $\mathcal{E}_C$ and N_0 the number of nodes in $\mathcal{N} - \mathcal{N}_C$. Last, call **U** (isomorphic to $\mathbb{C}^{E_C + N_0}$) the space of vectors $\mathrm{u} \equiv \{\mathbf{H}, \mathbf{\Phi}\} = \{\mathbf{H}_e : e \in \mathcal{E}_C, \mathbf{\Phi}_n : n \in \mathcal{N} - \mathcal{N}_C\}$, where the degrees of freedom $\mathbf{H}_e$ and $\mathbf{\Phi}_n$ are now unconstrained complex numbers. Let at last $\mathbb{K}^0_m$ be the space of vector fields of the form

$$\text{(32)} \qquad \mathrm{H} = \sum_{e \in \mathcal{E}_C} \mathbf{H}_e\, w_e + \sum_{n \in \mathcal{N} - \mathcal{N}_C} \mathbf{\Phi}_n \operatorname{grad} w_n.$$

Proposition 8.5. $\mathbb{K}^0_m$ *is isomorphic to* **U**.

Proof. This amounts to saying that degrees of freedom are independent, that is to say, $\mathrm{H} = 0$ in (32) implies all $\mathbf{H}_e$ and $\mathbf{\Phi}_n$ are zero. We know this is the case of the $\mathbf{H}_e$'s, by restriction to C (cf. Remark 5.2). As for the $\mathbf{\Phi}_n$'s, $0 = \sum_n \mathbf{\Phi}_n \operatorname{grad} w_n \equiv \operatorname{grad}(\sum_n \mathbf{\Phi}_n w_n)$ implies $\sum_n \mathbf{\Phi}_n w_n$ equal to a constant in the only connected component of D – C, a constant which is the value of this potential on ∂D, that is, 0. Again we know (cf. Exer. 3.8) that all $\mathbf{\Phi}_n$'s must vanish in this case. ◊

Proposition 8.6. *If* C *is contractible,* $\mathbb{K}^0_m = \mathbb{H}^0_m \equiv \{\mathrm{H} \in W^1_m : \operatorname{rot} \mathrm{H} = 0$ out of C$\}$.

Proof. After (32), one has $\operatorname{rot} \mathrm{H} = \sum_{e \in \mathcal{E}_C} \mathbf{H}_e \operatorname{rot} w_e$, and $\operatorname{supp}(\operatorname{rot} w_e)$ is contained in the closure of C, so $\operatorname{rot} \mathrm{H} = 0$ out of C. Conversely, if $\mathrm{H} \in W^1_m$ and if $\operatorname{rot} \mathrm{H} = 0$ in D – C, which is simply connected, there exists a linear combination Φ of the w_n, for $n \in \mathcal{N} - \mathcal{N}_C$, such that $\mathrm{H} = \operatorname{grad} \Phi$ in D – C, hence (32). ◊

Now let $\mathrm{H}^g_m \in W^1_m$ be an approximation of the source field. Again, $\mathbb{H}^g_m = \mathrm{H}^g_m + \mathbb{H}^0_m$, and we can "suffix everything with *m*", hence the desired Galerkin approximation for problem (23), the same, formally, as (24): *find* $\mathrm{H} \in \mathbb{H}^g_m$ *such that*

$$\text{(33)} \qquad \int_D i\,\omega\,\mu\, \mathrm{H} \cdot \mathrm{H}' + \int_C \sigma^{-1} \operatorname{rot} \mathrm{H} \,.\, \operatorname{rot} \mathrm{H}' = 0 \quad \forall\, \mathrm{H}' \in \mathbb{H}^0_m.$$

This is, in an obvious way, a linear system with respect to the unknowns $\mathbf{H}_e$ and $\mathbf{\Phi}_n$, the form of which is similar to (31).

To build $\mathrm{H}^g{}_m$, two techniques are possible. The first one consists of first computing H^g by the Biot and Savart formula, then evaluate the circulations $\mathbf{H}_e{}^g$ of H^g along all edges inside D. (For edges on the boundary, one sets $\mathbf{H}_e{}^g = 0$, which does introduce some error, but compatible with the desired accuracy,[7] if the mesh was well designed.) One then sets $\mathrm{H}^g{}_m = \sum_{e \in \mathcal{E}} \mathbf{H}^g{}_e \, w_e$.

However, this does not warrant rot $\mathrm{H}^g{}_m = 0$ where $\mathrm{J}^g = 0$, as it should be, for the Biot and Savart integral is computed with some error, and the sum of circulations of H^g along the three edges of a face where no current flows may come out nonzero, and this numerical error can be important when the edges in question happen to be close to inductor parts. This is a serious setback.

Hence the idea of again using the spanning tree method, which automatically enforces these relations. But contrary to the previous section, it's not necessary to deal with all the outside mesh to this effect. One will treat only a submesh, as small as possible, covering the support of J^g. Of the set $\mathcal{E}^g$ of edges of this submesh, one extracts a spanning tree $\mathcal{E}^t$, and one attributes a DoF to each co-edge the same way as in (28). One finally sets $\mathrm{H}^g{}_m = \sum_{e \in \mathcal{E}^g - \mathcal{E}^t} \mathbf{H}_e{}^g \, w_e$.

8.3.4 Cuts

Difficulties with the non-contractible case are about the same as in Subsection 8.3.2. Holes are no problem: Just pick one node n inside each non-conductive cavity within C and set $\mathbf{\Phi}_n = 0$ for this node. But the "loop problem" arises again, for if D – C is not simply connected, $\mathrm{IK}^0{}_m$ is strictly included in $\mathrm{IH}^0{}_m$: Missing are the fields H that, although curl-free in D – C, are only *local* gradients, or if one prefers, gradients of *multivalued* potentials Φ.

Hence the concept of "cuts", that is, for each current-loop, a kind of associated Seifert surface (cf. Note 6 and Exer. 8.3), formed of faces of the mesh. (Figure 8.7 will be more efficient than any definition to suggest what a cut is, but still, recall the formal definition of 4.1.2: a surface in $\overline{\mathrm{D}}$, closed mod C, that doesn't bound mod C.) One then doubles the nodal DoF for each node of this surface (Fig. 8.7, right): to n_+ is assigned the DoF $\mathbf{\Phi}^*{}_{n_+} = \mathbf{\Phi}_n$, and to n_- the nodal value $\mathbf{\Phi}^*{}_{n_-} = \mathbf{\Phi}_n + J$, where J is the loop-intensity. Let us denote by $\mathcal{N}^*$ the system of nodes thus obtained,

[7] If this is not the case, one may always resort to solving the magnetostatics problem, rot $\mathrm{H}^g = \mathrm{J}^g$ and div $\mathrm{H}^g = 0$ in D, with the same boundary conditions.

and set $\Phi = \sum_{n \in \mathcal{N}^*} \boldsymbol{\Phi}^*_n w_n$, where the $w_{n\pm}$ are supported on one side of Σ, as suggested on Fig. 8.7: Then Φ is multivalued in $D - C$, and the fields $H = \sum_{e \in \mathcal{E}_C} \mathbf{H}_e w_e + \sum_{n \in \mathcal{N}^*} \boldsymbol{\Phi}^*_n \operatorname{grad} w_n$ do fill out IH^0_m. It all goes again as if there was one extra DoF (the unknown intensity J) for each current loop.

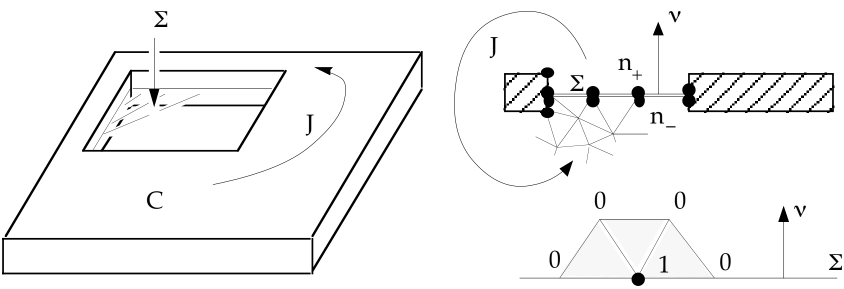


FIGURE 8.7. Cutting surface Σ, and doubling of nodes in Σ. The loop intensity J is also the circulation of the magnetic field along a circuit crossing Σ (along with the normal ν) and is therefore equal to the jump of Φ. Bottom right: support of the nodal function w_{n+}.

The big difficulty with this method is the construction of cuts. Several algorithms have been proposed [Br, HS, LR, VB], all more or less flawed because of a faulty definition of cuts. All these early works assumed, explicitly or not, that cuts must constitute a system of orientable (i.e., two-sided) surfaces, having their boundaries on ∂C ("closed modulo C but non-bounding", in the parlance of Chapter 4)—which is all right up to now—but also, *such that the complement in* D *of their set union with* C, that is, what remains of air after cuts have been removed, *be simply connected*. And this makes the definition defective, as Fig. 8.8 suffices to show: The complement of the set union of C and of the Seifert surface Σ is *not* simply connected, but the three components of Σ do qualify as cuts notwithstanding, for the magnetic potential Φ is effectively single-valued outside $C \cup \Sigma$. See the controversy in the IEE Journal [B&] triggered by the publication of [VB] for a discussion of this point. What cuts should do is make every curl-free field equal to a gradient in the outer region minus cuts. Credit is due to Kotiuga and co-workers [Ko] for the first correct definition of a cut, a constructive algorithm, and an implementation [GK].

It cannot be assessed, as the time this is written, whether this method is preferable to the "belted tree" approach. This is the matter of ongoing research [K&]. Let us, however, acknowledge that the problem of "knotted loops" is really marginal. Even common loops are infrequent in everyday work, because good modelling, taking symmetries into account, often allows one to dodge the difficulties they might otherwise raise. Eddy-current codes that were implemented, years ago, with cut-submodules based on some of the above-mentioned premature methods, on which current research is trying to improve, still work superbly in their respective domains of validity [BT, RL]

FIGURE 8.8. (Look first at Fig. 8.9 for the way Σ, here in three components, is constructed.) If some parts of the conductor (here made of four distinct connected components) are knotted or linked, it may happen that the complement of C and Σ is *not* simply connected, although the cuts Σ do play their role, which is to forbid multivalued potentials in the outside region. (Apply Ampère to circuits γ_1 and γ_2.)

The adaptation of the previous ideas to evolutionary regimes is straightforward, along the lines of 8.2.5.

8.4 SUMMING UP

What is special with eddy-current problems, and explains the almost unclassifiable variety of methods to deal with them, is the difference in the nature of the equations in the conductor and in the air. From the mathematical point of view, we have "parabolic equations" in conductors, "elliptic equations" in the air. For sure, passing in complex representation makes them elliptic all over, but different operators apply in the two main categories of regions: the "curl–curl" operator in the conductors, the "div–grad" operator in the air. We saw in Chapter 6 how intricate the relations between these two basic differential operators could be, marked by deep symmetries as well as striking differences. In the eddy-current problem, they coexist and must be compatible in some way at the common boundary. No wonder there is such a variety of methods for eddy currents! A few general ideas emerge, however:

1. Inside conductors, edge elements are the natural choice.

2. Outside, the magnetic potential is the most natural representation.

3. The structural relation $\text{grad } W^0 \subset W^1$ makes the two previous approaches compatible, be it in air volumes (H–Φ method) or on air–conductor interfaces (the "Trifou" hybrid method).

4. Two types of numerical treatment of the magnetic potential are available: integral equations (as we used to precompute the Dirichlet-to-Neumann map) and finite elements.

5. Multivaluedness of the magnetic potential, in the case of current loops, compounds difficulties, but tree and cotree methods offer solutions.

All these ideas can be combined. For instance, one may associate the H–Φ method in a bounded domain, thus dealing with current loops, and the integral equation method to account for the far field, provided the computational domain is simply connected. (There are cases when the latter restriction is still too much, however. It can be lifted thanks to the notion of *vectorial* Dirichlet-to-Neumann operator (end of Chapter 7). Cf. [Ve] for the early theory, and [RR] for an implementation, and an application to the problem of Fig. 8.2.) The variety of associations thus made possible is far from being exhausted, as witnessed by an abundant production of research papers (cf. especially **IEEE Trans. on Magnetism**, **IEE Proc., Series A**, **COMPEL**).

EXERCISES

Exercises 8.1 and 8.2 are on p. 227 and p. 233.

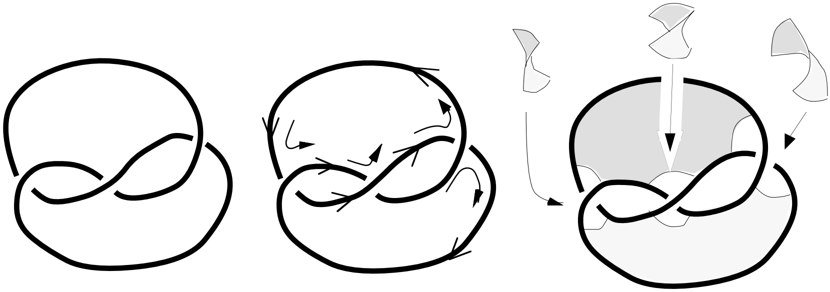


FIGURE 8.9. A knot, its Seifert surface (which has two distinct faces, as one can see, and is thus orientable), and the construction method.

Exercise 8.3. Figure 8.9 explains how to build a Seifert surface (as defined in Note 6) for a relatively simple knot (the procedure makes sense for *links*, too, as on Fig. 8.8, right): Work on an orthogonal plane projection of the knot, and select an arbitrary orientation along it; start from a point and follow the selected direction, never crossing at an apparent intersection, but instead, jumping forward to another part of the knot, and going on in the right direction, as suggested by the middle of Fig. 8.9; do this as many times as possible, thus obtaining as many pieces of the Seifert surface; these pieces are then seamed together by "flaps", as explained on the right of the figure, in order to obtain a single, two-sided surface. Practice on some knots of your own design. Show (by a series of drawings) that the surface thus obtained in the case of Fig. 8.9 is homeomorphic to a punctured torus.

Exercise 8.4. Show that if a circuit γ bounds a surface Σ entirely contained in a current-free region, then $\int_\gamma \tau \cdot h = 0$. The converse happens to be true. For instance, the circulation of h along γ_1 on Fig. 8.8 is null whatever the current in C. Show that γ_1 is indeed the boundary of a surface which does not encounter C.

Exercise 8.5. Can you devise a continuous current distribution on a torus, the way Fig. 8.10 suggests, such that the magnetic field outside be zero, though the induction flux through a surface bounded by γ is not?

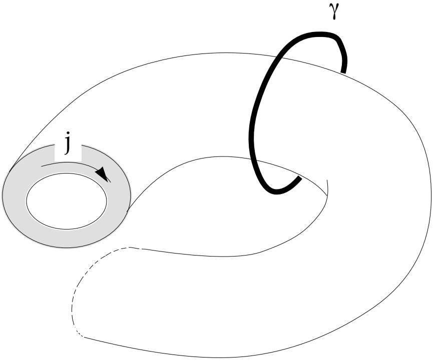


FIGURE 8.10. Such a current distribution can correspond to a null magnetic field in the outside region. Part of the hollow torus is cut off for better view. The conductive section is shaded. Circuit γ will be relevant to the next Exercise.

Exercise 8.6. In the situation of Fig. 8.10, suppose the *intensity* of the current distribution changes in time (but not the shape of its spatial distribution). Assuming γ is a conductive thread, will it support an induced current? Do you see a paradox in this? Can you solve it *within* eddy current theory (that is, without invoking retardation effects and the like)?

Exercise 8.7. The original Bath cube problem is described by Fig. 8.11. Differences with respect to the static version are: The mmf I is alternative, at angular frequency ω, and an aluminum cube is placed inside each quarter of the cavity, away from the bottom and from the walls. Do the modelling (continuous and discrete formulation).

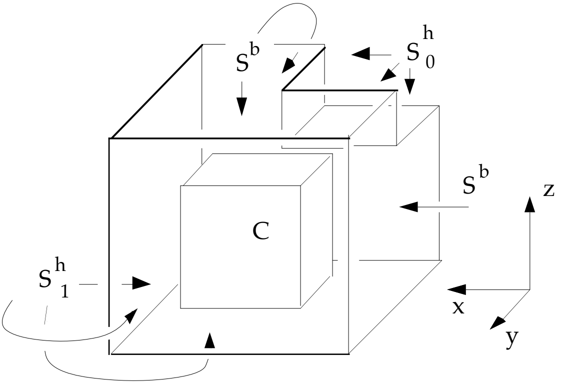
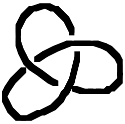


FIGURE 8.11. The Bath-cube problem (one quarter of the cavity, with conductive cube C included).

HINTS

8.3. You may have difficulties with some drawings, for example if you sketch the trefoil knot of Fig. 8.9 like this:  (it's the same knot). In that case, imagine the knot as projected on the surface of a sphere of large radius, instead of on the plane.

8.4. Stokes.

8.5. Call D the domain occupied by the torus of Fig. 8.10, including the inner bore, and C the conductive part. Build a smooth solenoidal field a, curl-free out of C, and make sure the circulation of a along γ is non-zero. This means a = grad ψ outside D, but with a multivalued ψ, so use a cut. Similar thing in D – C. Extend a to C so that a is curl-conformal, then rectify a to make it sinusoidal *in all space.* Then take b = rot a, h = μ_0^{-1} b, and j = rot h.

8.6. Let j be the stationary current distribution of Exer. 8.4, h the corresponding field, and assume a current density J(t) j, hence an induction field μ_0 J(t) h. Its flux through the loop γ is not null, and changes with time, so there is an emf along γ, by Faraday's law, which may drive a current in the conductive loop. Now, to quote from [PK], "The commonly asked question is: 'We know that charges in the loop move in response to an emf produced by a changing magnetic field; but since there is no magnetic field outside the solenoid, how do the charges in the loop know they must move?'." What of it?

8.7. Use the h–Φ method: edge elements in C, nodal elements for Φ in $D - C$, scalar DoFs for Φ on ∂C. No constraints on Φ on surface S^b. On $S^h{}_0$ and $S^h{}_1$, set $\Phi = 0$ and $\Phi = I$ respectively, where I is the imposed mmf between the poles.

SOLUTIONS

8.3. Figure 8.12.

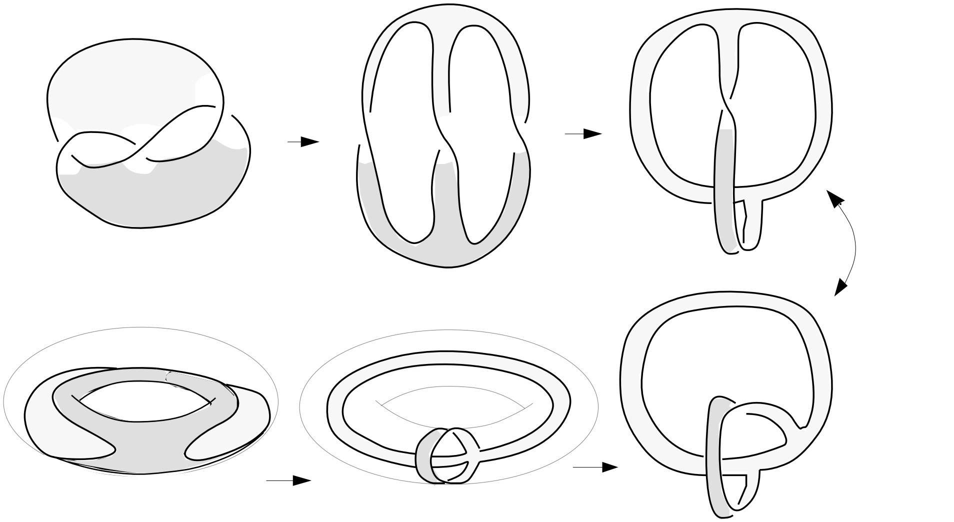


FIGURE 8.12. Homeomorphism between a punctured torus and the Seifert surface of a trefoil knot.

8.4. Figure 8.13 displays an orientable surface (a punctured torus, again) with γ_1 as its boundary. Since $\mathrm{rot}\, h = 0$ outside C, one has $\int_{\gamma_1} \tau \cdot h = \int_\Sigma n \cdot \mathrm{rot}\, h = 0$, whatever the intensity in C.

8.5. Let Σ_{ext} be a "cut" in $E_3 - D$, and $O_{ext} = E_3 - D - \Sigma_{ext}$. Let ψ_{ext} be the minimizer of $\int_{O_{ext}} |\mathrm{grad}\, \psi|^2$ in $\{\psi \in BL(O) : [\psi]_{\Sigma_{ext}} = F_{ext}\}$, with $F_{ext} \neq 0$. Similarly, select $\psi_{ext} \in \mathrm{arginf}\{\int_B |\mathrm{grad}\, \psi|^2 : \psi' \in BL(O_{int}) : [\psi']_{\Sigma_{int}} = F_{int}\}$ (these minimizers differ by an additive constant), where Σ_{int} is a cut inside $D - C$. Set $a = \mathrm{grad}\, \psi_i$ in O_i, with $i = ext$ or int. Extend a to a smooth field $\tilde{a}$ in $D - C$, with null jumps $[n \times \tilde{a}]_{\partial C}$ on the air–conductor interface. Now $\tilde{a} \in \mathbb{L}^2_{\text{rot}}(E_3)$ and its circulation along γ is nonzero, but $\tilde{a}$ is not solenoidal. Set $a = \tilde{a} + \text{grad}\,\tilde{\psi}$, where $\tilde{\psi}$ is the unique element of $BL(E_3)$ such that $\int_{E_3} (\tilde{a} + \text{grad}\,\tilde{\psi}) \cdot \text{grad}\,\psi' = 0 \;\; \forall\, \psi' \in BL(E_3)$. Now $\text{div}\, a = 0$ *in all space*, while its curl hasn't changed. Take $h = \mu_0^{-1} \,\text{rot}\, a$ and $j = \text{rot}\, h$. The field created by j is h, and has the required properties. By playing on the values of F_{ext} and F_{int}, one may control the coil intensity. The *pratical* realization is another issue, but is possible in principle: One may always imagine bundles of thin separate conductors coiled around the inner torus, along the small circles, and fed by small batteries.

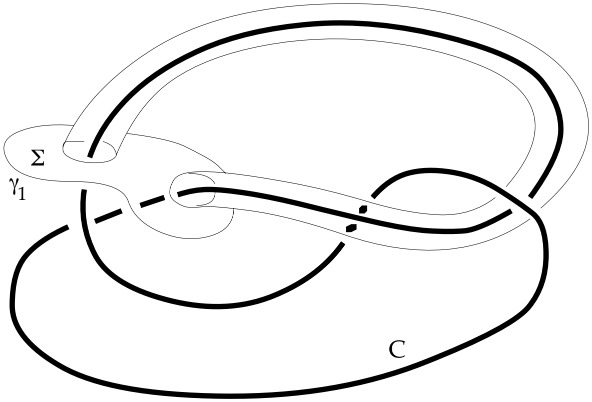


FIGURE 8.13. Circuit γ_1 does bound an orientable surface contained in the current-free region. The conductor C is the black knot.

8.6. Did you feel confused by this argument? Rightly so, because the expression, "emf produced by a changing magnetic field" *is* subtly confusing. The causal phenomenon by which changes of magnetic field generate electric fields is ruled by the equations $\text{rot}\, e = -\,\partial_t b$, $\text{div}(\varepsilon_0 e) = 0$ outside C, with tangential e known on ∂C, so it has an inherently *non-local* character. Changes of magnetic field in some region (here, domain D) thus produce emf's away from this region, including at places where the magnetic field is zero and stays zero. (This is no more paradoxical than the fact that changes in electric charge can modify the electric field in regions where there is no charge.) So there is an induced current in the loop, and its value can be predicted by eddy-current theory alone. No need to argue about the "physical unreality" of the situation" (*all* eddy current modellings are "unreal" to a comparable degree!) and to add irrelevant considerations on the way the solenoid is energized, on finite propagation speeds, and so forth [PK, Te].

This problem is relevant to discussions of the Aharonov–Bohm effect (cf. Remark A.2). Most papers on the subject assume a straight, infinite solenoid, instead of the above toroidal one, which makes some analytical computations easier, but also needlessly raises side issues.

REFERENCES

[BT] K.J. Binns, P.J. Lawrenson, C.W. Trowbridge: **The Analytic and Numerical Solution of Electric and Magnetic Fields**, Wiley (Chichester), 1992.

[B1] A. Bossavit: "On Finite Elements for the Electricity Equation", in **The Mathematics of Finite Elements** (J.R. Whiteman, ed.), Academic Press (London), 1982, pp. 85–92.

[BV] A. Bossavit, J.C. Vérité: "A Mixed FEM-BIEM Method to Solve Eddy-Current Problems", **IEEE Trans., MAG-18**, 2 (1982), pp. 431–435.

[BV'] A. Bossavit, J.C. Vérité: "The TRIFOU Code: Solving the 3-D Eddy-Currents Problem by Using H as State Variable", **IEEE Trans., MAG-19**, 6 (1983), pp. 2465–2470.

[B&] A. Bossavit, P.R. Kotiuga, A. Vourdas, K.J. Binns: Correspondence on "Magnetostatics with scalar potentials in multiply connected regions", **IEE Proc., 136, Pt. A,** 5 (1989), pp. 260–261, **137, Pt. A,** 4 (1989), pp. 231–232.

[Bo] A. Bossavit: "A Numerical Approach to Transient 3D Non-linear Eddy-current Problems", **Int. J. Applied Electromagnetics in Materials**, **1**, 1 (1990), pp. 65–75.

[Br] M.L. Brown: "Scalar Potentials in Multiply Connected Domains", **Int. J. Numer. Meth. Engng., 20** (1984), pp. 665–680.

[CN] J. Crank, P. Nicolson: in **Proc. Camb. Phil. Soc. math. phys. Sci, 43** (1947), p. 50.

[GF] O.P. Gandhi, J.F. DeFord, H. Kanai: "Impedence Method for Calculation of Power Deposition Patterns in Magnetically Induced Hyperthermia", **IEEE Trans., BME-31**, 10 (1984), pp. 644–651.

[GH] M.J. Greenberg, J.R. Harper: **Algebraic Topology, A First Course**, Benjamin/ Cummings (Reading, MA), 1981.

[GK] P.W. Gross, P.R. Kotiuga: "A Challenge for Magnetic Scalar Potential Formulations of 3-D Eddy Current Problems: Multiply Connected Cuts in Multiply Connected Regions which Necessarily Leave the Cut Complement Multiply Connected", in **Electric and Magnetic Fields** (A. Nicolet & R. Beulmans, eds.), Plenum Press (New York, London), 1995, pp. 1–20.

[HS] C.S. Harrold, J. Simkin: "Cutting multiply connected domains", **IEEE Trans., MAG-21**, 6 (1985), pp. 2495–2498.

[K&] Y. Kanai, T. Tsukamoto, Y. Saitoh, M. Miyakawa, T. Kashiwa: "Analysis of a Hyperthermic Treatment using a Reentrant Resonant Cavity Applicator for a Heterogeneous Model with Blood Flow", **IEEE Trans., MAG-33**, 2 (1997), pp. 2175–2178.

[K&] L. Kettunen, K. Forsman, A. Bossavit: "Formulation of the eddy current problem in multiply connected regions in terms of h", **Int. J. Numer. Meth. Engng., 41,** 5 (1998)., pp. 935–954.

[Ko] P.R. Kotiuga: "An algorithm to make cuts for magnetic scalar potentials in tetrahedral meshes based on the finite element method", **IEEE Trans., MAG-25**, 5 (1989), pp. 4129–4131.

[LR] P.J. Leonard, D. Rodger: "A new method for cutting the magnetic scalar potential in multiply connected eddy current problems", **IEEE Trans., MAG-25**, 5 (1989), pp. 4132–4134.

[RL] P.J. Leonard, H.C. Lai, R.J. Hill-Cottingham, D. Rodger: "Automatic Implementation of Cuts in Multiply Connected Magnetic Scalar Regions for 3D Eddy Current Models", **IEEE Trans., MAG-**29, 2 (1993), pp. 1368–1377.

[Na] T. Nakata (ed.): **3-D Electromagnetic Field Analysis** (Proc. Int. Symp. & TEAM Workshop, Okayama, Sept. 1989), James and James (London), 1990 (Supplement A to Vol. 9 of the Journal **COMPEL**).

[PK] R. Protheroe, D. Koks: "The transient magnetic field outside an infinite solenoid", **Am. J. Phys., 64**, 11 (1996), pp. 1389–1393.

[RR] Z. Ren, A. Razek: "Boundary Edge Elements and Spanning Tree Technique in Three-Dimensional Electromagnetic Field Computation", **Int. J. Numer. Meth. Engng., 36** (1993), pp. 2877–2893.

[Ro] D. Rolfsen: **Knots and Links**, Publish or Perish, Inc. (Wilmington, DE 19801, USA), 1976.

[Se] H. Seifert: "Über das Geschlecht von Knoten", **Math. Ann., 110** (1934), pp. 571–592.

[Te] J.D. Templin: "Exact solution to the field equations in the case of an ideal, infinite solenoid", **Am. J. Phys., 63**, 10 (1995), pp. 916–920.

[Ve] J.C. Vérité: "Calculation of multivalued potentials in exterior regions", **IEEE Trans., MAG-23**, 3 (1987), pp. 1881–1887.

[VB] A. Vourdas, K.J. Binns: "Magnetostatics with scalar potentials in multiply connected regions", **IEE Proc., 136, Pt. A,** 2 (1989), pp. 49–54.

CHAPTER 9

Maxwell's Model in Harmonic Regime

9.1 A CONCRETE PROBLEM: THE MICROWAVE OVEN

9.1.1 Modelling

Our last model, about the microwave oven, is typical of the class of time-harmonic problem *with* displacement currents taken into account, in bounded regions.

Such an oven is a cavity enclosed in metallic walls, containing an antenna and something that must be heated, called the "load" (Fig. 9.1). One may model the antenna by a current density j^g, periodic in time (the typical frequency is 2450 MHz), hence $j^g(t) = \text{Re}[\text{J}^g \exp(i\omega t)]$, the support of J^g being a part of the cavity. Note that this current has no reason to be divergence-free. The average power necessary to sustain it, which will be retrieved in part as thermal power in the load, is $-\frac{1}{2}\text{Re}[\int \text{J}^g \cdot \text{E}^*]$. The load occupies a part of the cavity and is characterized by *complex*-valued coefficients ε and μ, for reasons we shall explain.

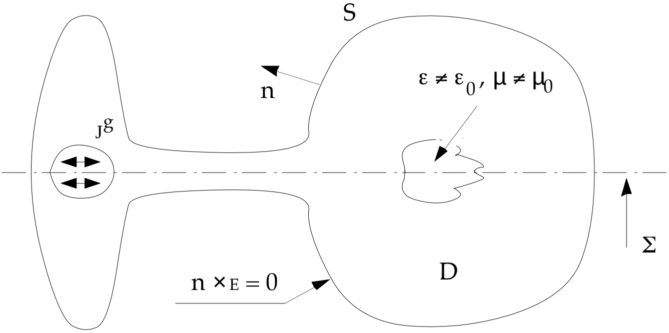


FIGURE 9.1. Notations for the microwave oven problem.

The conductivity of metallic walls is high enough to assume $\mathrm{E} = 0$ there, so the equations are

(1) $\quad - i\omega\, \mathrm{D} + \mathrm{rot}\, \mathrm{H} = \mathrm{J}^{g} + \sigma\, \mathrm{E}$ in D,

(2) $\quad i\omega\, \mathrm{B} + \mathrm{rot}\, \mathrm{E} = 0$ in D,

(3) $\quad n \times \mathrm{E} = 0$ on S.

The load, as a rule, is an aqueous material with high permittivity, hence a strong polarization in presence of an electric field. Moreover, because of the inertia of dipoles, the alignment of the polarization vector p on the electric field is not instantaneous, as we assumed in Chapter 1. If one sticks to the hypothesis of linearity of the constitutive law, then $p(t) = \int^{t} f(t - s)\, e(s)\, ds$, where the function f is a characteristic of the medium. After Fourier transformation, this becomes

$$\mathrm{P}(\omega\,) = \sqrt{2\pi}\, \mathrm{F}(\omega)\, \mathrm{E}(\omega),$$

(F is the *transfer function* of the polarized medium), hence $\mathrm{D}(\omega) = \varepsilon_0\, \mathrm{E}(\omega) + \mathrm{P}(\omega) \equiv \varepsilon\, \mathrm{E}(\omega)$, with $\varepsilon(\omega) = \varepsilon_0\, \mathrm{E} + \sqrt{2\pi}\, \mathrm{F}(\omega)$, a complex and frequency-dependent permittivity. It is customary to set $\varepsilon = \varepsilon' - i\, \varepsilon''$, with real ε' and ε''. It all goes (transfer $i\, \varepsilon''$ to the right-hand side in (1)) as if one had a real permittivity ε, a conductivity $\varepsilon''\omega$ (in addition to the normal ohmic conductivity—but the latter can always be accounted for by the ε'' term, by adding σ/ω to it), and a current density $j(t) = \mathrm{Re}[\mathrm{J} \exp(i\omega t)]$, where $\mathrm{J} = \omega\, \varepsilon''\mathrm{E}$. The reason for the minus sign can be seen by doing the following computation, where $T = 2\pi/\omega$ is the period:

$$\tfrac{1}{T} \int_{t-T}^{t} ds \int_D j(s) \cdot e(s) = \tfrac{1}{T} \int_{t-T}^{t} ds \int_D \mathrm{Re}[\mathrm{J} \exp(i\omega t)] \cdot \mathrm{Re}[\mathrm{E} \exp(i\omega t)]$$

$$= \mathrm{Re}[\mathrm{J} \cdot \mathrm{E}^*]/2 \equiv \omega\, \varepsilon''\, |\mathrm{E}|^2/2,$$

since this quantity, which is the thermal power yielded to the EM compartment (cf. Chapter 1, Section 3), now agrees in sign with ε''.

Of course, f cannot directly be measured, just theorized about (cf. [Jo]). But ε' and ε'' can (cf. Fig. 1.4).[1]

[1] As real and imaginary parts of the Fourier transform of one and the same function, ε' and ε'' are not independent (they are "Hilbert transforms" of each other). One could thus, in theory, derive one from the other, provided one of them is known over *all* the spectrum, and with sufficient accuracy. This is of course impossible in practice, and ε' and ε'' are independently measured (as real and imaginary parts of the impedance of a sample) on an appropriate frequency range.

For the sake of symmetry and generality, let's also write $\mu = \mu' - i\,\mu''$, hence the definitive form of the equations:

$$(4) \qquad -\, i\omega\, \varepsilon\, E + \operatorname{rot} H = J^g, \quad i\omega\, \mu\, H + \operatorname{rot} E = 0 \text{ in } D, \quad n \times E = 0 \text{ on } S,$$

with $\varepsilon = \varepsilon' - i\,\varepsilon''$ and $\mu = \mu' - i\,\mu''$. They will normally be coupled with the heat equation, the source-term being the average thermic power $i\omega\, \varepsilon'' |E|^2/2$ (plus $i\omega\, \mu'' |H|^2/2$, if this term exists). The parameter ε, temperature-dependent, will therefore change during the heating.

9.1.2 Position of the problem

We want a variational formulation of (4), for J^g given in $\mathbb{L}^2_{\text{div}}(D)$, where the unknown will be the field E, after elimination of H. Let $\mathbb{E}(D)$ denote the (complex) space $\mathbb{L}^2_{\text{rot}}(D)$, and

$$\mathbb{E}^0(D) = \{E \in \mathbb{E}(D) : \; n \times E = 0 \;\text{ on } S\}.$$

The scalar product of two complex *vectors* U and V is as in Chapter 8 (*no* conjugation on the right), but we shall adopt a space-saving notational device, as follows: If U and V are two complex *fields*, we denote by $(U, V)_D$, or simply (U, V), the expression $\int_D U(x) \cdot V(x)\, dx$ (which, let's stress it again, is *not* the Hermitian scalar product).

A precise formulation of (4) is then: *find* $E \in \mathbb{E}^0(D)$ *such that*

$$(5) \qquad (i\omega\, \varepsilon\, E, E') + ((i\omega\, \mu)^{-1} \operatorname{rot} E, \operatorname{rot} E') = -\,(J^g, E') \quad \forall\, E' \in \mathbb{E}^0(D).$$

Unfortunately, the existence question is not trivial, because the bilinear form $a(E, E')$ on the left-hand side of (5) is not coercive. Indeed,

$$\operatorname{Re}[a(E, E^*)] = \omega \int_D \varepsilon''\ |E|^2 + \omega^{-1} \int_D \mu''/|\mu|^2\ |\operatorname{rot} E|^2,$$

which vanishes if the support of E does not overlap with those of ε'' and μ'', and

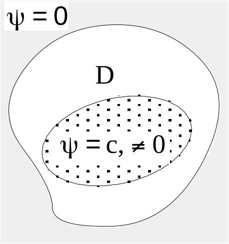
$$\operatorname{Im}[a(E, E^*)] = \omega \int_D \varepsilon'\ |E|^2 - \omega^{-1} \int_D \mu'/|\mu|^2\ |\operatorname{rot} E|^2$$

has no definite sign (and no premultiplication by a scalar will cure that).

But the restriction to a *bounded* domain (finite volume is enough, actually) introduces some compactness which makes up for this lack of coercivity, at least for non-singular values of ω, thanks to the Fredholm alternative.

9.2 THE "CONTINUOUS" PROBLEM

9.2.1 Existence



Let's first prove an auxiliary result. Let D be a regular bounded domain in E_3, with boundary S. For simplicity (but this is not essential), assume D simply connected. Let Ψ^0 be the space of restrictions to D of functions ψ of $L^2_{\text{grad}}(E_3)$ for which $\text{grad}\ \psi = 0$ outside D. (If S is connected, they belong to the Sobolev space $H^1_0(D)$, but otherwise, it's a slightly larger space, for ψ can be a nonzero constant on some parts of S, as shown in inset.) Call V the following closed subspace of $\mathbb{L}^2(D)$:

$$V = \{v \in \mathbb{L}^2(D) : \ (v, \text{grad}\ \psi') = 0 \quad \forall\ \psi' \in \Psi^0\}.$$

Proposition 9.1. *Let* J *be given in* $\mathbb{L}^2(E_3)$, *with* $\text{div}\ J = 0$ *and* $\text{supp}(J) \subset D$. *Suppose* $\mu' \geq \mu_0$ *in* D. *There exists a unique* $A \in \mathbb{E}^0(D)$ *such that*

$$(6) \qquad (\mu^{-1}\ \text{rot}\ A, \text{rot}\ A') = (J, A') \quad \forall\ A' \in \mathbb{E}^0$$

as well as $\varepsilon A \in V$, *and the map* $G = J \rightarrow \varepsilon A$ *is* compact *in* V.

Before giving the proof, note that such a field A verifies

$$(6') \qquad \text{rot}(\mu^{-1}\ \text{rot}\ A) = J, \ \ \text{div}\ \varepsilon A = 0 \ \text{ in } D, \ \ n \times A = 0 \text{ on } S,$$

but these conditions are not enough to determine it, unless S is connected. In that case, $V = \{v \in \mathbb{L}^2(D) : \ \text{div}\ v = 0\}$. But otherwise, V is a strictly smaller subspace, characterized by $\int n \cdot v = 0$ on each connected component of S, hence as many similar conditions[2] on A, to be appended to the "strong formulation" (6). Note also that $J \in V$, under the hypotheses of the statement, so G does operate from V to V.

Proof of Prop. 1. Uniqueness holds, because the kernel of rot in $\mathbb{E}^0$ is precisely $\text{grad}\ \Psi^0$ (this is why Ψ^0 was defined this way). The proof will consist in showing that one passes from J to εA by composing continuous maps, one of which at least is compact.

Set $U = \chi * J$, with $\chi = x \rightarrow 1/(4\pi\ |x|)$, and take its restriction ${}^D U$ to D. The map $J \rightarrow {}^D U$ thus defined is compact in $\mathbb{L}^2(D)$ [Yo]. Therefore, the map $J \rightarrow \text{rot}\ U \in \mathbb{L}^2(E_3)$ is compact, too, for if $\{J_n\}$ is a sequence of V

[2] This is one of the advantages of weak formulations: They foster thoroughness, by reminding one of conditions which one might have overlooked in the first place.

such that $\mathrm{J}_n \rightharpoonup \mathrm{J}$ (weak convergence—cf. A.4.3), then $\mathrm{U}_n \rightharpoonup \mathrm{U}$ by continuity, and hence (dot-multiply by a test field A' and integrate by parts) $\mathrm{rot}\,\mathrm{U}_n \rightharpoonup \mathrm{rot}\,\mathrm{U}$. Moreover, $\int |\mathrm{rot}\,\mathrm{U}_n|^2 = \int_D \mathrm{J}_n \cdot \mathrm{U}_n{}^*$, and ${}^D\mathrm{U}_n$ tends to ${}^D\mathrm{U}$, so the norm of $\mathrm{rot}\,\mathrm{U}_n$ converges towards that of $\mathrm{rot}\,\mathrm{U}$, therefore $\mathrm{rot}\,\mathrm{U}_n$ tends to $\mathrm{rot}\,\mathrm{U}$. Setting $\mathrm{H} = \mathrm{rot}\,{}^D\mathrm{U}$, one has thus proved the compactness of the map $\mathrm{J} \rightarrow \mathrm{H}$.

Now let $\Phi \in L^2_{\mathrm{grad}}(D)$ be such that

$$(\mu\,(\mathrm{H} + \mathrm{grad}\,\Phi), \mathrm{grad}\,\Phi') = 0 \quad \forall\, \Phi' \in L^2_{\mathrm{grad}}(D).$$

The solution of this problem is unique up to an additive constant only, but $\mathrm{grad}\,\Phi$ is unique, and the map $\mathrm{H} \rightarrow \mathrm{grad}\,\Phi$ is continuous. Let us set $\mathrm{B} = \mu(\mathrm{H} + \mathrm{grad}\,\Phi)$. Then $\mathrm{div}\,\mathrm{B} = 0$ and $n \cdot \mathrm{B} = 0$, so the prolongation by 0 of B outside D is divergence-free. If one sets $\mathrm{A}_0 = {}^D(\mathrm{rot}(\chi * \mathrm{B}))$ —again, the restriction to D —then $\mathrm{rot}\,\mathrm{A}_0 = \mathrm{B}$, and the mapping $\mathrm{H} \rightarrow \mathrm{A}_0$ is continuous.

Notice that the tangential trace of A_0 is a gradient, by the Stokes theorem, for the flux of B through a closed circuit drawn on S vanishes, since $n \cdot \mathrm{rot}\,\mathrm{A}_0 = n \cdot \mathrm{B} = 0$. For this reason, the set of the $\Psi \in L^2_{\mathrm{grad}}(D)$ for which $n \times (\mathrm{A}_0 + \mathrm{grad}\,\Psi) = 0$ is not empty, and there is one among them (unique up to an additive constant) for which

$$(\varepsilon\,(\mathrm{A}_0 + \mathrm{grad}\,\Psi), \mathrm{grad}\,\Psi') = 0 \quad \forall\, \Psi' \in \Psi^0.$$

Then $\mathrm{A} = \mathrm{A}_0 + \mathrm{grad}\,\Psi$ is the desired solution, and $\mathrm{grad}\,\Psi$ continuously depends on A_0, with respect to the norm of $\mathbb{L}^2(D)$. The map $\mathrm{J} \rightarrow \mathrm{A}$ is therefore compact, hence the compactness of the operator $G = \mathrm{J} \rightarrow \varepsilon\mathrm{A}$, whose domain is the subspace V. ◊

Let's call *singular* (or *resonating)* the nonzero values of ω for which the homogeneous problem associated with (5) has a nontrivial solution, i.e., $\mathrm{E} \neq 0$ such that

$$(7) \qquad (i\omega\,\varepsilon\,\mathrm{E}, \mathrm{E}') + ((i\omega\,\mu)^{-1}\,\mathrm{rot}\,\mathrm{E}, \mathrm{rot}\,\mathrm{E}') = 0 \quad \forall\, \mathrm{E}' \in \mathbb{E}^0.$$

Such an E verifies $\varepsilon\,\mathrm{E} \in V$ (take $\mathrm{E}' \in \mathrm{grad}\,\Psi^0$) as well as $\mathrm{rot}(\mu^{-1}\,\mathrm{rot}\,\mathrm{E}) = \omega^2\,\varepsilon\,\mathrm{E}$ (integrate by parts). In other words, $\varepsilon\,\mathrm{E} = \omega^2 G\,\varepsilon\,\mathrm{E}$. Thus, $\varepsilon\,\mathrm{E}$ is an eigenvector of G, corresponding to the eigenvalue ω^{-2}. (One says that the pair $\{\mathrm{E}, \mathrm{H}\}$, where $\mathrm{H} = -\,(\mathrm{rot}\,\mathrm{E})/i\omega\,\mu$, is an "eigenmode" of the cavity, for the angular frequency ω.) By Fredholm's theory, there exists a denumerable infinity of eigenvalues for G, each with finite multiplicity, and not clustering anywhere except at the origin in the complex plane.[3] The singular values are thus the square roots of the inverses of the eigenvalues

[3] Owing to uniqueness in (6), 0 is not an eigenvalue of G.

of G. (A priori, eigenvalues are complex, unless both ε and μ are real.)

Theorem 9.1. *For each non-singular value of* ω, *problem* (5) *is well posed, i.e., has a unique solution* E, *and the map* $\mathrm{J}^g \rightarrow \mathrm{E}$ *is continuous from* $\mathbb{L}^2(D)$ *into* $\mathbb{E}(D)$.

Proof. Since ω is not singular, uniqueness holds. Let's look for a solution of the form $\mathrm{E} = -\,i\omega\,\mathrm{A} - \operatorname{grad}\psi$, with $\mathrm{A} \in \mathbb{E}^0$, $\varepsilon\,\mathrm{A} \in V$, and $\psi \in \Psi^0$. Set $\mathrm{E}' = \operatorname{grad}\psi'$ in (5), with $\psi' \in \Psi^0$. This yields

$$(i\omega\,\varepsilon\,(\mathrm{A} + \operatorname{grad}\psi), \operatorname{grad}\psi') = (\mathrm{J}^g, \operatorname{grad}\psi') \quad \forall\, \psi' \in \Psi^0,$$

and hence, since $\varepsilon\mathrm{A}$ is orthogonal to all $\operatorname{grad}\psi'$,

$$(8) \qquad (i\omega\,\varepsilon\,\operatorname{grad}\psi, \operatorname{grad}\psi') = (\mathrm{J}^g, \operatorname{grad}\psi') \quad \forall\, \psi' \in \Psi^0,$$

a well-posed problem in Ψ^0, hence the continuity of $\mathrm{J}^g \rightarrow \operatorname{grad}\psi$ in $\mathbb{L}^2(D)$. This leaves A to be determined. After (5), one must have

$$(\mu^{-1}\operatorname{rot}\mathrm{A}, \operatorname{rot}\mathrm{A}') = (\mathrm{J}^g + i\omega\,\varepsilon\,\mathrm{E}, \mathrm{A}') \quad \forall\, \mathrm{A}' \in \mathbb{E}^0$$

$$= (\mathrm{J}^g - i\omega\,\varepsilon\,\operatorname{grad}\psi, \mathrm{A}') + \omega^2\,(\varepsilon\mathrm{A}, \mathrm{A}') \quad \forall\, \mathrm{A}' \in \mathbb{E}^0.$$

But this is the Fredholm equation of the second kind,

$$(1 - \omega^2 G)\,\varepsilon\,\mathrm{A} = G\,(\mathrm{J}^g - i\omega\,\varepsilon\,\operatorname{grad}\psi),$$

hence A by the Fredholm alternative, if ω is not a singular value, and provided $\mathrm{J}^g - i\omega\,\varepsilon\,\operatorname{grad}\psi \in V$ —which is what (8) asserts. ◊

9.2.2 Uniqueness

Hence the question: Are there singular values? For an empty cavity ($\mu = \mu_0$ and $\varepsilon = \varepsilon_0$), or with lossless materials (μ and ε real and positive), *yes*, since all eigenvalues of G are then real and positive. If $\omega \neq 0$ is one of them and $\mathrm{E} = \mathrm{e}_R$ a nonzero associated real solution of (7) (there *is* a *real* one), then $\mathrm{H} = i\,\mathrm{h}_I$, with real h_I. The existence of such a solution means that a time-periodic electromagnetic field, of the form $e(x, t) = \mathrm{Re}[\mathrm{E}(x)\exp(i\omega t)] \equiv e_R(x)\cos\omega t$ and $h(x, t) = -\,h_I(x)\sin\omega t$ can exist forever in the cavity, without any power expense, and also of course without loss.

To verify this point, let's start from the equations $i\omega\,\varepsilon\mathrm{E} - \operatorname{rot}\mathrm{H} = 0$ and $i\omega\,\mu\,\mathrm{H} + \operatorname{rot}\mathrm{E} = 0$, dot-multiply by E and H, add, and integrate over D: by the curl integration-by-parts formula, and because of $n \times \mathrm{E} = 0$, this gives $\int_D \varepsilon\ \mathrm{E}^2 + \int_D \mu\ \mathrm{H}^2 = 0$, that is, since $\mathrm{E} = \mathrm{e}_R$ and $\mathrm{H} = \mathrm{i}\ \mathrm{h}_I$,

$$\int_D \varepsilon\ |\mathrm{e}_R|^2 = \int_D \mu\ |\mathrm{h}_I|^2.$$

But the energy contained in D at time t, which is (cf. Chapter 1)

$$\begin{aligned} W(t) &= \tfrac{1}{2} \int_D (\varepsilon\ |\mathrm{e}(t)|^2 + \mu\ |\mathrm{h}(t)|^2) \\ &= \tfrac{1}{2} \int_D \varepsilon\ |\mathrm{e}_R|^2 \cos^2 \omega t + \tfrac{1}{2} \int_D \mu\ |\mathrm{h}_I|^2 \sin^2 \omega t \\ &\equiv \tfrac{1}{2} \int_D \varepsilon\ |\mathrm{e}_R|^2 \equiv \tfrac{1}{2} \int_D \mu\ |\mathrm{h}_I|^2, \end{aligned}$$

is indeed constant, and is the sum of two periodic terms of identical amplitudes, "electric energy" and "magnetic energy" in the cavity, the former vanishing at each half-period and the latter a quarter-period later.

Such a behavior seems unlikely in the case of a loaded cavity, since the energy of the field decreases in the process of yielding heat to the region $C = \mathrm{supp}(\varepsilon'') \cup \mathrm{supp}(\mu'')$. How can this physical intuition be translated into a proof? It's easy to see that E and H vanish in C: Setting $\mathrm{E}' = \mathrm{E}^*$ in (7), one gets

$$\omega \int_D \varepsilon''\ |\mathrm{E}|^2 + \mathrm{i}\omega \int_D \varepsilon'\ |\mathrm{E}|^2 + \int_D (\omega\ |\mu|^2)^{-1} \mu''\ |\mathrm{rot}\ \mathrm{E}|^2$$
$$+ \int_D (\mathrm{i}\omega\ |\mu|^2)^{-1} \mu'\ |\mathrm{rot}\ \mathrm{E}|^2 = 0,$$

hence, taking the real part, $\mathrm{E} = 0$ or $\mathrm{rot}\ \mathrm{E} = 0$ on C, hence $\mathrm{H} = 0$, too, and therefore $\mathrm{E} = 0$ after (2). But what opposes the existence of a nonzero mode E that would be supported in the complementary region D – C, what one may call an "air mode" ?

If such an air mode existed, both $n \times \mathrm{E}$ and $n \times \mathrm{H}$ would vanish on ∂C, which is impossible if E and H are to satisfy Maxwell's equations in region D – C. We shall prove this by way of a mathematical argument, here encapsulated as a context-independent statement, cast in non-dimensional form:

Proposition 9.2. *Let* Ω *be a regular domain of* E_3. *Let* u *and* v *satisfy*

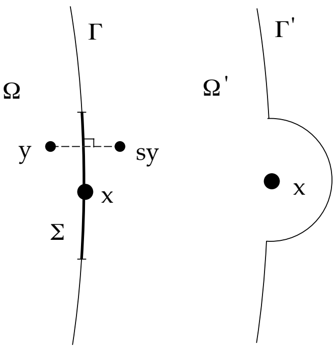
(13) $\qquad \mathrm{i}\ \mathrm{rot}\ u = v, \qquad - \mathrm{i}\ \mathrm{rot}\ v = u \ \text{ in } \ \Omega,$

(14) $\qquad n \times u = 0, \qquad n \times v = 0 \ \text{ on } \ \Sigma,$

where Σ *is a part of* Γ *with a smooth boundary and a non-empty interior (relative to* Σ*). Then* u *and* v *vanish in all* Ω.

Proof. (Though reduced to the bare bones by many oversimplifications, the proof will be long.) Both u and v satisfy $\text{div}\, u = 0$ and $-\Delta u = u$ in Ω, after (13). It is known (cf., e.g., [Yo]), that every u which satisfies $(\Delta + 1)u = 0$ in some open set is *analytic* there. This is akin to the Weyl lemma discussed in 2.2.1, and as mentioned there, this result of "analytic ellipticity" is valid also for div(a grad) + b, where a and b are smooth. If we could prove that all derivatives of all components of u and v vanish at some point, u and v would then have to be 0 in all Ω, by analyticity.

The problem is, we can prove this fact, but only for boundary points such as x (inset), which is in the relative interior of Σ, not inside Ω. So we need to expand Ω in the vicinity of x, as suggested by the inset, and to make some continuation of the equations to this expanded domain Ω' in a way which preserves analytic ellipticity. To do this, first straighten out Σ around x by an appropriate diffeomorphism, then consider the mirror symmetry s with respect to the plane where Σ now lies. Pulling back this operation to Ω gives a kind of warped reflexion with respect to Σ, still denoted by s. Let us define continuations of u and v to the enlarged domain Ω' by setting $\tilde{u}(sy) = -\, su(y)$ and $\tilde{v}(sy) = s_* v(y)$, where s_* is the mapping induced by s on vectors (cf. A.3.4, p. 301). Now the extensions $\tilde{u}$ and $\tilde{v}$ satisfy in Ω' a system similar to (13–14), with some smooth coefficients added, the solutions of which are similarly analytic.



The point about vanishing derivatives remains to be proven. This is done by working in an appropriate coordinate system $y \rightarrow \{y^1, y^2, y^3\}$ with the above point x at the origin, y^1 and y^2 charting Σ, and y^3 along the normal. In this system, $u = \{u^1, u^2, u^3\}$, and $u^j(x^1, x^2, 0\} = 0$, for $j = 1, 2$, and the same for v. Derivatives $\partial_i u^j$ vanish for all j and $i = 1$ or 2 at point x by hypothesis. One has also $\partial_1 u^3 = -\, i\, \partial_1(\partial_1 v^2 - \partial_2 v^1) = 0$ on Σ, and in the same way, $\partial_2 v^3 = 0$, $\partial_1 u^3 = 0$, $\partial_2 u^3 = 0$. Since $\text{rot}\, u = \{\partial_2 u^3 - \partial_3 u^2, \partial_3 u^1 - \partial_1 u^3, \partial_1 u^2 - \partial_2 u^1\} \equiv \{-\, \partial_3 u^2, \partial_3 u^1, 0\}$ is also 0 at x, we have $\partial_3 u^j(x) = 0$ for $j = 1$ and 2. The last derivative to consider is $\partial_3 u^3 \equiv \text{div}\, u - (\partial_1 u^1 + \partial_2 u^2) = 0$, since $\text{div}\, u = \text{div}\, \tilde{u} = 0$ at point x. This disposes of first-order derivatives.

What has just been proved for a generic point of Σ implies that all first-order derivatives of u and v vanish on Σ. Thus, differentiating (13) with respect to the ith coordinate, we see that $\partial_i u$ and $\partial_i v$ satisfy

(13–14), so the previous reasoning can be applied to derivatives of u and v of all order: They all vanish at x, and now the analyticity argument works, and yields the announced conclusion. ◊

The anti-air-mode result now comes by setting $\Omega = D - C$ and $\Sigma = \partial C$, and by choosing a system of units in which $\varepsilon_0 \mu_0 \omega^2 = 1$. (Another corollary is that one cannot simultaneously impose $n \times H$ and $n \times E$ on a part of nonzero area of the cavity boundary.)

We may thus conclude that no resonant modes exist if there is a charge in a microwave oven, and that Maxwell equations have a unique solution in that case, assuming the tangential part of one of the fields H or E is specified at every point of the boundary.

9.2.3 More general weak formulations

Thus satisfied that (4), or better, its weak form (5), has a unique solution, we shall try to get an approximation of it by finite elements. But first, it's a good idea to generalize (5) a little, as regards source terms and boundary conditions, without bothering too much about the physical meaning of such generalizations.

First, let's exploit the geometrical symmetry, by assuming the source current J^g is symmetrically placed with respect to plane Σ. Then, for $x \in \Sigma$, one has $H(x) = - s_* H(x)$, where s is the mirror reflection with respect to Σ, and s_* the induced mapping on vectors. The boundary condition to apply is therefore, if n denotes the normal as usual,

$$(9) \qquad n \times H = 0 \text{ on } \Sigma.$$

Therefore, our customary splitting of the boundary into complementary parts is in order: $S = S^e \cup S^h$, with $n \times E = 0$ and $n \times H = 0$, respectively, on S^e and S^h.

Second generalization: Introduce a right-hand side K^g in the second equation (4). This term does not correspond to anything physical (it would be a magnetic charge current, if such a thing existed), but still it pops up in a natural way in some modellings. For instance, if the field is decomposed as $H^g + \tilde{H}$, where H^g is a known field, a term $K^g = - i\omega \mu H^g$ appears on the right-hand side of the equation relative to the reaction field $\tilde{H}$.

This suggests also preserving the possibility of *non-homogeneous* boundary conditions: $n \times E = n \times E^g$ in (4) and $n \times H = n \times H^g$ in (9), where E^g and H^g are given fields, of which only the tangential part on S will matter. For instance, in the case of Fig. 9.1, one might wish to limit the computation to the oven proper, that is, the rightmost part of the cavity. Indeed, the middle part is a waveguide, in which the *shape* of the electric field, if not its amplitude, is known in advance. Hence the source–term E^g, up to a multiplicative factor.

All this considered, we shall treat a general situation characterized by the following elements: a regular bounded domain D limited by S, the latter being partitioned as $S = S^e \cup S^h$, and given fields J^g, K^g (in $\mathbb{L}^2_{\mathrm{div}}(D)$), H^g, E^g (in $\mathbb{L}^2_{\mathrm{rot}}(D)$). We denote, with the dependence on D understood from now on, $\mathbb{E} = \mathbb{L}^2_{\mathrm{rot}}(D)$ and also $\mathbb{H} = \mathbb{L}^2_{\mathrm{rot}}(D)$, then

$$\mathbb{E}^g = \{\mathrm{E} \in \mathbb{E} : n \times \mathrm{E} = n \times \mathrm{E}^g \text{ on } S^e\},$$

$$\mathbb{E}^0 = \{\mathrm{E} \in \mathbb{E} : n \times \mathrm{E} = 0 \text{ on } S^e\},$$

$$\mathbb{H}^g = \{\mathrm{H} \in \mathbb{H} : n \times \mathrm{H} = n \times \mathrm{H}^g \text{ on } S^h\},$$

$$\mathbb{H}^0 = \{\mathrm{H} \in \mathbb{H} : n \times \mathrm{H} = 0 \text{ on } S^h\},$$

and we set the following two problems:

$$\textit{find } \mathrm{E} \in \mathbb{E}^g \textit{ such that } (i\omega\,\varepsilon\,\mathrm{E}, \mathrm{E}') + ((i\omega\,\mu)^{-1} \operatorname{rot} \mathrm{E}, \operatorname{rot} \mathrm{E}') = - (\mathrm{J}^g, \mathrm{E}') + ((i\omega\,\mu)^{-1} \mathrm{K}^g, \operatorname{rot} \mathrm{E}') + \int_S n \times \mathrm{H}^g \cdot \mathrm{E}' \quad \forall\, \mathrm{E}' \in \mathbb{E}^0, \tag{10}$$

$$\textit{find } \mathrm{H} \in \mathbb{H}^g \textit{ such that } (i\omega\,\mu\,\mathrm{H}, \mathrm{H}') + ((i\omega\,\varepsilon)^{-1} \operatorname{rot} \mathrm{H}, \operatorname{rot} \mathrm{H}') = (\mathrm{K}^g, \mathrm{H}') + ((i\omega\,\varepsilon)^{-1} \mathrm{J}^g, \operatorname{rot} \mathrm{H}') - \int_S n \times \mathrm{E}^g \cdot \mathrm{H}' \quad \forall\, \mathrm{H}' \in \mathbb{H}^0. \tag{11}$$

(Beware, there is no relation, a priori, between K^g and H^g, or between J^g and E^g.)

Integrating by parts, one easily sees that *each* weak formulation (10) or (11) solves the following "strong" problem (compare with (4)):

$$\left|\begin{array}{l} -\, i\omega\,\varepsilon\,\mathrm{E} + \operatorname{rot} \mathrm{H} = \mathrm{J}^g \text{ in } D, \; n \times \mathrm{E} = n \times \mathrm{E}^g \text{ on } S^e, \\ i\omega\,\mu\,\mathrm{H} + \operatorname{rot} \mathrm{E} = \mathrm{K}^g \text{ in } D, \; n \times \mathrm{H} = n \times \mathrm{H}^g \text{ on } S^h. \end{array}\right. \tag{12}$$

One may therefore solve (12), approximately, by discretizing either (10) or (11). But (and this is *complementarity*, again!) one obtains solutions which slightly differ, and yield complementary information about the exact solution.

9.3 THE "DISCRETE" PROBLEM

9.3.1 Finite elements for (10) or (11)

Let $m = \{\mathcal{N}, \mathcal{E}, \mathcal{F}, \mathcal{T}\}$ be a mesh of D, and $W^1_m(D)$ (or, for shortness, W^1_m) the edge-element subspace of $\mathbb{L}^2_{\mathrm{rot}}(D)$ of Chapter 5. The idea is to *restrict* the formulations (10) and (11) to this subspace.

For this, let us denote by $\mathcal{E}^e$ [resp. $\mathcal{E}^h$] the subset of $\mathcal{E}$ formed by edges that belong to S^e [resp. to S^h], and let $\mathbb{E}_m$ and $\mathbb{H}_m$ be two copies of W^1_m. Set

$$\mathbb{E}^g_m = \{E \in \mathbb{E}_m : \int_e \tau \,.\, E = \int_e \tau \,.\, E^g \ \ \forall\, e \in \mathcal{E}^e\},$$

$$\mathbb{E}^0_m = \{E \in \mathbb{E}_m : \int_e \tau \,.\, E = 0 \ \ \forall\, e \in \mathcal{E}^e\},$$

$$\mathbb{H}^g_m = \{H \in \mathbb{H}_m : \int_e \tau \,.\, H = \int_e \tau \,.\, H^g \ \ \forall\, e \in \mathcal{E}^h\},$$

$$\mathbb{H}^0_m = \{H \in \mathbb{H}_m : \int_e \tau \,.\, H = 0 \ \ \forall\, e \in \mathcal{E}^h\}.$$

All we have to do now is to index by m all the spaces that appear in (10) and (11) in order to obtain *approximate* weak formulations for both problems:

$$\textit{find } E \in \mathbb{E}^g_m \textit{ such that } \ (i\omega\,\varepsilon\, E, E') + ((i\omega\,\mu)^{-1} \operatorname{rot} E, \operatorname{rot} E') =$$

$$(13) \qquad - (J^g, E') + ((i\omega\,\mu)^{-1} K^g, \operatorname{rot} E') + \int_S n \times H^g \cdot E' \ \ \forall\, E' \in \mathbb{E}^0_m,$$

$$\textit{find } H \in \mathbb{H}^g_m \textit{ such that } \ (i\omega\,\mu\, H, H') + ((i\omega\,\varepsilon)^{-1} \operatorname{rot} H, \operatorname{rot} H') =$$

$$(14) \qquad (K^g, H') + ((i\omega\,\varepsilon)^{-1} J^g, \operatorname{rot} H') - \int_S n \times E^g \cdot H' \ \ \forall\, H' \in \mathbb{H}^0_m.$$

These are linear systems, with a finite number of equations: The choice $E' = w_e$, at the right-hand side, for all edges e not contained in $\mathcal{E}^e$ [resp. $H' = w_e$, for all e not in $\mathcal{E}^h$] does give one equation for each unknown edge circulation.

By the usual notational shift, we shall cast these equations in matrix form. Let $E = \sum_{e \in \mathcal{E}} \mathbf{E}_e\, w_e$ and $H = \sum_{e \in \mathcal{E}} \mathbf{H}_e\, w_e$ be the required fields of $\mathbb{E}^g_m$ and $\mathbb{H}^g_m$, and denote $\mathbf{E} = \{\mathbf{E}_e : e \in \mathcal{E}\}$[4] and $\mathbf{H} = \{\mathbf{H}_e : e \in \mathcal{E}\}$ the DoF vectors. They span vector spaces $\boldsymbol{\mathbb{E}}_m$ and $\boldsymbol{\mathbb{H}}_m$, isomorphic to $\mathbb{C}^E$, where E is the number of edges in m. Remind that, for $\mathbf{U}$ and $\mathbf{U}'$ both in $\mathbb{C}^E$, one denotes

$$\mathbf{(}\mathbf{U}, \mathbf{U}'\mathbf{)} = \sum_{e \in \mathcal{E}} \mathbf{U}_e \cdot \mathbf{U}'_e \equiv \sum_{e \in \mathcal{E}} (\mathrm{Re}[\mathbf{U}_e] + i\,\mathrm{Im}[\mathbf{U}_e]) \cdot (\mathrm{Re}[\mathbf{U}'_e] + i\,\mathrm{Im}[\mathbf{U}'_e]).$$

[4] In memoriam G.P.

By sheer imitation of what precedes, let us set

$$\mathbb{E}^g_m = \{\mathrm{E} \in \mathbb{E}_m : \mathrm{E}_e = \int_e \tau \, . \mathrm{E}^g \;\; \forall\, e \in \mathcal{E}^e\},$$

$$\mathbb{E}^0_m = \{\mathrm{E} \in \mathbb{E}_m : \mathrm{E}_e = 0 \;\; \forall\, e \in \mathcal{E}^e\},$$

$$\mathbb{H}^g_m = \{\mathrm{H} \in \mathbb{H}_m : \mathrm{H}_e = \int_e \tau \, . \mathrm{H}^g \;\; \forall\, e \in \mathcal{E}^h\},$$

$$\mathbb{H}^0_m = \{\mathrm{H} \in \mathbb{H}_m : \mathrm{H}_e = 0 \;\; \forall\, e \in \mathcal{E}^h\},$$

and also, for ease in the expression of the right-hand sides,

$$\mathrm{F}^g_e = -\,(\mathrm{J}^g, \mathrm{w}_e) + ((i\omega\,\mu)^{-1}\mathrm{K}^g, \mathrm{rot}\, \mathrm{w}_e) + \int_S n \times \mathrm{H}^g \cdot \mathrm{w}_e,$$

$$\mathrm{G}^g_e = (\mathrm{K}^g, \mathrm{w}_e) + ((i\omega\,\varepsilon)^{-1}\mathrm{J}^g, \mathrm{rot}\, \mathrm{w}_e) - \int_S n \times \mathrm{E}^g \cdot \mathrm{w}_e,$$

and $\mathrm{F}^g = \{\mathrm{F}^g_e : e \in \mathcal{E}\}$, as well as $\mathrm{G}^g = \{\mathrm{G}^g_e : e \in \mathcal{E}\}$.

Using the matrices $\mathbf{M}_1(\mu)$, $\mathbf{R}$, etc., of Chapter 5, one may restate the two problems as follows:

find $\mathrm{E} \in \mathbb{E}^g_m$ *such that*

$$(15) \qquad (i\omega\, \mathbf{M}_1(\varepsilon)\, \mathrm{E}, \mathrm{E}') + (i\omega)^{-1}(\mathbf{R}^t\, \mathbf{M}_2(\mu^{-1})\, \mathbf{R}\, \mathrm{E}, \mathrm{E}') = (\mathrm{F}^g, \mathrm{E}') \;\; \forall\, \mathrm{E}' \in \mathbb{E}^0_m,$$

find $\mathrm{H} \in \mathbb{H}^g_m$ *such that*

$$(16) \qquad (i\omega\, \mathbf{M}_1(\mu)\, \mathrm{H}, \mathrm{H}') + (i\omega)^{-1}(\mathbf{R}^t\, \mathbf{M}_2(\varepsilon^{-1})\, \mathbf{R}\, \mathrm{H}, \mathrm{H}') = (\mathrm{G}^g, \mathrm{H}') \;\; \forall\, \mathrm{H}' \in \mathbb{H}^0_m.$$

Since $\mathbb{E}^g_m$ and $\mathbb{E}^0_m$ [resp. $\mathbb{H}^g_m$ and $\mathbb{H}^0_m$] are parallel by their very definition, they have same dimension, so there are *as many equations as unknowns* in (15) [resp. in (16)]. All that is left to do is to assess the regularity, in the algebraic sense, of the corresponding matrices.

9.3.2 Discrete models

To this effect, let's write the unknown vector E in the form $\mathrm{E} = {}^0\mathrm{E} + {}^1\mathrm{E}$, the 0's corresponding to edges of $\mathcal{E} - \mathcal{E}^e$ and the 1's to those of $\mathcal{E}^e$, so ${}^1\mathrm{E}$ is a known item. If $\mathbf{K}$ is some $E \times E$ matrix, the identity

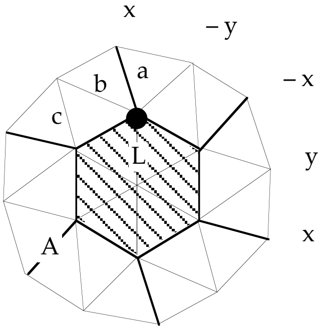
$$(\mathbf{K}\,({}^0\mathrm{E} + {}^1\mathrm{E}), {}^0\mathrm{E} + {}^1\mathrm{E}) = ({}^{00}\mathbf{K}\,{}^0\mathrm{E}, {}^0\mathrm{E}) + ({}^{01}\mathbf{K}\,{}^1\mathrm{E}, {}^0\mathrm{E})$$
$$\ldots + ({}^{10}\mathbf{K}\,{}^0\mathrm{E}, {}^1\mathrm{E}) + ({}^{11}\mathbf{K}\,{}^1\mathrm{E}, {}^1\mathrm{E})$$

defines a partition of $\mathbf{K}$ in blocks of dimensions $E - E^e$ and E^e, where E^e is the number of edges in $\mathcal{E}^e$. Set, for simplicity, $\mathbf{K}_{\varepsilon\mu}(\omega) = i\omega\; \mathbf{M}_1(\varepsilon) + (i\omega)^{-1}\, \mathbf{R}^t\, \mathbf{M}_2(\mu^{-1})\mathbf{R}$, and let ${}^{00}\mathbf{K}_{\varepsilon\mu}(\omega)$, ${}^{01}\mathbf{K}_{\varepsilon\mu}(\omega)$, etc., be the corresponding blocks. Same thing for the matrix $\mathbf{K}_{\mu\varepsilon}(\omega) = i\omega\, \mathbf{M}_1(\mu) + (i\omega)^{-1}\, \mathbf{R}^t\, \mathbf{M}_2(\varepsilon^{-1})\, \mathbf{R}$, partitioned as ${}^{00}\mathbf{K}_{\mu\varepsilon}(\omega)$, etc. Then (15) and (16) can be rewritten as

$$(17) \qquad {}^{00}\mathbf{K}_{\varepsilon\mu}(\omega)\ {}^{0}\mathrm{E} = {}^{0}\mathrm{F}^{g} - {}^{01}\mathbf{K}_{\varepsilon\mu}(\omega)\ {}^{1}\mathrm{E},$$

$$(18) \qquad {}^{00}\mathbf{K}_{\mu\varepsilon}(\omega)\ {}^{0}\mathrm{H} = {}^{0}\mathrm{G}^{g} - {}^{01}\mathbf{K}_{\mu\varepsilon}(\omega)\ {}^{1}\mathrm{H},$$

two systems of orders $E - E^e$ and $E - E^h$, respectively. The values of ω (not the same for both) for which they are singular are approximations of the above singular values.



The question arises again: In the case of a loaded cavity, can there be *real* such values, i.e., discrete air modes? Unfortunately, no proof similar to the previous one seems possible for the discretized version of the problem, and the following counter-example shakes hopes of finding one without some qualifying assumptions, which remain to be found. Consider the stiffness matrix of the air region, including its boundary, and denote by $\mathbf{u}$ the vector of degrees of freedom for all edges except those on the boundary of the load, which form vector $\mathbf{v}$. Write the discrete eigenvalue problem as $\mathbf{Au} + \mathbf{B}^t\,\mathbf{v} = \lambda\mathbf{u}$, $\mathbf{Bu} + \mathbf{Cv} = \lambda\mathbf{v}$. The question is: Can one exclude solutions with $\lambda \neq 0$ but $\mathbf{v} = 0$, i.e., such that $\mathbf{Au} = \lambda\mathbf{u}$ and $\mathbf{Bu} = 0$? For the mesh in inset, and in 2D, where the equation reduces to $-\Delta\varphi = \lambda\varphi$, one cannot. Let a, b, c, off-diagonal coefficients, repeat by six-fold symmetry. Let $\mathbf{u}$ be such that $\mathbf{Au} = \lambda\mathbf{u}$, with $\lambda \neq 0$, and $\mathbf{u}$ antisymmetric, that is, such that the degrees of freedom x, y, etc., alternate as suggested. (There *is* such an eigenvector.) Now, the row of $\mathbf{Au} = \lambda\mathbf{u}$ corresponding to the marked node yields $(b - c)y + ax = 0$, hence $\mathbf{Bu} = 0$.

Discrete air modes thus cannot so easily be dismissed. Do they appear? This would destroy well-posedness. But even their existence for meshes "close", in some sense, to the actual one would be enough to create difficulties, when solving $(-\,\omega^2\,\mathbf{A} + \mathbf{B})\ \mathbf{u} = \mathbf{f}$, that is, in the frequential approach. Note that, fortunately, alternative "time domain" approaches exist (Remark 8.4), not prone to such difficulties [DM].

Note finally that problems (10) and (11) were equivalent, but (17) and (18) are not. They yield *complementary* views of the solution: (17) gives E (approximately, of course), hence $\mathrm{B} = (\mathrm{K}^g - \mathrm{rot}\ \mathrm{E})/i\omega$, whereas (18) gives H, whence $\mathrm{D} = (\mathrm{rot}\ \mathrm{H} - \mathrm{J}^g)/i\omega$. These four fields satisfy Maxwell equations *exactly*. But the constitutive relations $\mathrm{B} = \mu\mathrm{H}$ and $\mathrm{D} = \varepsilon\mathrm{E}$ are not rigorously satisfied, and the magnitude of the discrepancy is a good gauge of the accuracy achieved in this dual calculation. See [PB] for a development of this idea, with applications to adaptive mesh refinement in particular.

9.3.3 The question of spurious modes

Let's end with a topic that much intrigued the microwave community during the past 20 years (cf. [KT] for a review).

Let's go back to (12), but for an *empty* cavity ($\varepsilon = \varepsilon_0$ and $\mu = \mu_0$ all over) and *not* excited by outside sources (so E^g and H^g are zero). The equations reduce to

$$(19) \qquad -\, i\omega\, \varepsilon_0\, \mathrm{E} + \mathrm{rot}\, \mathrm{H} = 0 \ \text{ in } D, \ \ n \times \mathrm{E} = 0 \ \text{ on } S^e,$$

$$(20) \qquad i\omega\, \mu_0\, \mathrm{H} + \mathrm{rot}\, \mathrm{E} = 0 \ \text{ in } D, \ \ n \times \mathrm{H} = 0 \ \text{ on } S^h,$$

and they have nonzero solutions for resonating values of ω, which in this case correspond to *all* the eigenvalues (since all are real). So, if $\omega \neq 0$ is such a value, one has $\mathrm{div}\, \mathrm{B} = 0$ and $\mathrm{div}\, \mathrm{D} = 0$, where $\mathrm{B} = \mu_0\, \mathrm{H}$ and $\mathrm{D} = \varepsilon_0 \mathrm{E}$: The electric and magnetic induction fields are solenoidal. (Note this is not the case when $\omega = 0$: There are solutions of the form $\mathrm{H} = \mathrm{grad}\, \Phi$ and $\mathrm{E} = \mathrm{grad}\, \Psi$, with Φ and Ψ non-harmonic.) Of course, whatever the method, one does not solve (19) and (20) with absolute accuracy, and one does not expect the relations $\mathrm{div}(\mu_0 \mathrm{H}) = 0$ and $\mathrm{div}(\varepsilon_0\, \mathrm{E}) = 0$ to hold true, but at least one may hope for the magnitudes[5] of these divergences to be small, and to get smaller and smaller when the mesh grain tends to zero under the usual anti-flattening restrictions. Yet, before the advent of edge elements, such was not the case; all meshes showed modes with sizable divergence, which had to be rejected as "non-physical". The discussion in Chapter 6 helps understand why the emergence of such "spurious modes" is a defect inherent in the use of classical node-based vector-valued elements, and indeed, using Whitney elements is a sufficient condition for such spurious modes not to appear, as we now show.

[5] Since $\mu\mathrm{H}$ is *not* normally continuous across faces, there is a problem of definition here, for the divergence of $\mu\mathrm{H}$ is a distribution, not a function. In order to assess the "magnitude of the divergence" of $\mu\mathrm{H}$, one should evaluate the norm of the mapping $\varphi' \to \int_D \mu\, \mathrm{H} \cdot \mathrm{grad}\, \varphi'$, that is

$$\sup\{\, |\textstyle\int_D \mu\mathrm{H} \cdot \mathrm{grad}\, \varphi'| \,/\, [\int_D |\varphi'|^2]^{1/2} : \ \varphi' \in L^2_{\mathrm{grad}}(D), \ \varphi' \neq 0\}.$$

In practice, a weighted sum of the "flux losses" at faces makes a good indicator.

The "continuous" spectral problem consists in finding the values of ω for which, all source data being zero, and μ and ε real positive, Problem (11), that is, *find* $\mathrm{H} \in \mathbb{H}^0$ *such that*

$$(\mathrm{i}\omega\, \mu\, \mathrm{H}, \mathrm{H}') + ((\mathrm{i}\omega\, \varepsilon)^{-1} \operatorname{rot} \mathrm{H}, \operatorname{rot} \mathrm{H}') = 0 \quad \forall\, \mathrm{H}' \in \mathbb{H}^0,$$

has a nonzero solution. (The situation with respect to (10) is symmetrical, as we have seen.) Let's consider some Galerkin approximation to this problem, by which one wants to *find* $\mathrm{H} \in \mathcal{H}_m$ *such that*

$$(\mathrm{i}\omega\, \mu\, \mathrm{H}, \mathrm{H}') + ((\mathrm{i}\omega\, \varepsilon)^{-1} \operatorname{rot} \mathrm{H}, \operatorname{rot} \mathrm{H}') = 0 \quad \forall\, \mathrm{H}' \in \mathcal{H}_m, \tag{21}$$

where $\mathcal{H}_m$ is a *finite*-dimensional subspace of $\mathbb{H}^0$. (This is $\mathbb{H}^0{}_m$, for the same mesh m, if edge elements are used.) This problem has no nonzero solution, except for a finite number of values of ω, corresponding to the eigenvalues of the matrix that represents, in some basis of $\mathcal{H}_m$, the bilinear form of the left-hand side in (21).

Now, consider the kernel of rot in the space $\mathcal{H}_m$. It's some subspace $\mathcal{K}_m$ which is, if one assumes a simply connected D, the image by grad of some finite-dimensional space $\mathcal{F}_m$, composed of functions which belong to $\mathrm{L}^2_{\mathrm{grad}}$. If, for some $\omega \neq 0$, (21) has a nonzero solution H, the latter verifies, a fortiori,

$$(\mu\, \mathrm{H}, \operatorname{grad} \Phi') = 0 \quad \forall\, \Phi' \in \mathcal{F}_m . \tag{22}$$

This is a familiar relation: We spent most of Chapter 4 studying its consequences, where we saw to which extent $\mu\, \mathrm{H}$ is satisfactory, as an "m-weakly solenoidal" field. This happens when $\mathcal{F}_m$ is a *good* approximation space for $\mathrm{L}^2_{\mathrm{grad}}$, that is, "big enough", in an intuitively clear sense.

When $\mathcal{H}_m$ is $\mathbb{H}^0{}_m$, that is, with edge elements, the subspace $\mathcal{F}_m$ is indeed big enough, as we saw in Chapter 5; thereby, spurious modes are effectively eliminated [Bo, PR, WP]. In contrast, the use of nodal elements entails spaces $\mathcal{F}_m$ of very small dimensions, possibly 0, as we saw in Chapter 6. In such a case, nothing warrants any kind of weak solenoidality of $\mu\,\mathrm{H}$, hence the occurrence of spurious modes, so often observed and deplored [KT] before the advent of edge elements.

> You stop, but that does not mean you have come to the end.
>
> P. AUSTER, *"In the Country of Last Things"*

REFERENCES

[Bo] A. Bossavit: “Solving Maxwell’s Equations in a Closed Cavity, and the Question of Spurious Modes”, **IEEE Trans., MAG-26,** 2 (1990), pp. 702–705.

[DM] D.C. Dibben, A.C. Metaxas: “Finite Element Time Domain Analysis of Multimode Applicators Using Edge Elements”, **J. Microwave Power & Electromagn. Energy, 29,** 4 (1994), pp. 242–251.

[Jo] A.K. Jonscher: “The ‘universal’ dielectric response”, **Nature**, 267 (23 June 1977), pp. 673–679.

[KT] A. Konrad, I.A. Tsukerman: “Comparison of high- and low-frequency electromagnetic field analysis”, **J. Phys. III France**, 3 (1993), pp. 363–371.

[PR] L. Pichon, A. Razek: “Analysis of Three-Dimensional Dielectric Loaded Cavities with Edge Elements”, **ACES Journal, 66,** 2 (1991), pp. 133–142.

[PB] L. Pichon, A. Bossavit: “A new variational formulation, free of spurious modes, for the problem of loaded cavities”, **IEEE Trans., MAG-29,** 2 (1993), pp. 1595–1600.

[St] M.A. Stuchly, S.S. Stuchly: “Dielectric Properties of Biological Substances —Tabulated”, **J. Microwave Power, 15**, 1 (1980), pp. 19–26.

[FS] K.P. Foster, H.P. Schwan: “Dielectric properties of tissues and biological materials: a critical review”, **Critical Reviews in Biomedical Engineering, 17**, 1 (1989), pp. 25–104.

[WP] M.-F. Wong, O. Picon, V. Fouad-Hanna: “Résolution par éléments finis d'arête des équations de Maxwell dans les problèmes de jonctions et cavités micro-ondes”, **J. Phys. III France** (1992), pp. 2083–2099.

[Yo] K. Yosida: **Functional Analysis**, Springer-Verlag (Berlin), 1965.

APPENDIX A

Mathematical Background

This is not a tutorial, just background material. It contains two interleaved texts. One is rather formal, with definitions, mainly, and occasionally, proofs, arranged in logical order: Things are defined in terms of primitive concepts and of previously defined things. The other part, where examples will be provided, is a commentary on why and how these notions and properties can be useful in computational electromagnetism. There, of course, one feels free to invoke not yet formally defined entities.

The treatment is neither exhaustive nor balanced. The space devoted to each notion does not necessarily reflect its intrinsic importance. Actually, most important notions will be familiar to the reader already and will cursorily be treated, just enough to provide a context for the ones I have chosen to emphasize: those that are (in my opinion) both important and generally underrated.

Most definitions in the formal part are implicit: When a new concept or object is introduced, its name is set in *italics*,[1] and the context provides a definition. The index should help locate such definitions, when needed.

A.1 BASIC NOTIONS

Notions we choose to consider as primitive, and that we shall not define, are those of set theory: sets, elements, subsets, equality, inclusion (symbols $=$, $\in$, and $\subseteq$), finite and infinite sets, and of logic: propositions, or "predicates", true or false. Basic notions that follow are defined in terms of primitive notions.

[1] Italics also serve to put emphasis on some words, according to standard practice. This should cause no confusion.

A.1.1 Sets, properties

If X is a set, $\mathcal{P}(X)$ will denote the set of all its parts (or *power set)*, and $\varnothing$ the empty set. Don't confuse elements, such as x, with one-element subsets, denoted $\{x\}$. The *Cartesian product* of sets X and Y is denoted by $X \times Y$. It is made of all *pairs* $\{x, y\}$, with $x \in X$ and $y \in Y$. The product $A \times B$ of two parts $A \subseteq X$ and $B \subseteq Y$ is the set of pairs $\{x, y\}$, with $x \in A$ and $y \in B$.

When speaking of pairs, order counts: x first, then y. If X and Y are different sets, no problem. But some confusion may occur when $X = Y$. If $x \neq y$, are $\{x, y\}$ and $\{y, x\}$ different elements of $X \times X$? Yes, of course, so $\{x, y\} \neq \{y, x\}$. But this same notation, $\{x, y\}$, is often used also for something else, namely, the subsets of X composed of two elements, and now, $\{x, y\}$ and $\{y, x\}$ point to the same object, an element of $\mathcal{P}(X)$. So we are dealing with a different concept here, that of *unordered* pair.

Some have tried to promote the use of a different word for unordered pairs ("couple", for instance) to stress the difference. But then it's difficult to remember which is which. So if you see $\{x, y\}$ at any place in this book, be it called couple or pair, assume the order counts, unless the context warns you otherwise. (Fortunately, the confusion is most often harmless.) The natural extension[2] of the pair concept is the *n-tuple,* $\{x_1, x_2, \ldots, x_n\}$. Unordered n-tuples are subsets containing n elements, all different.

Propositions and predicates are statements which can assume the value **true** or **false** (with a special face, because **true** and **false** are labels for the two elements of a special set, "Boolean algebra", to which we shall return). Some examples of predicates: $x \in X$, or $x \neq y$, or $A \subseteq B$, or else $x \in A$ **and** $y \notin B$, etc. (Again, **and** is a logical operation in Boolean algebra, about which we shall have more to say.) The difference between "proposition" and "predicate" is semantical: Predicates can contain variables, whose values may affect the truth value of the predicate. For instance, speaking of real numbers, $x > 0$ is a predicate, the truth value of which depends on the value of the free variable[3] x, whereas $2 < 1$ is just a proposition[4] (its value is **false**).

[2] It's recursively defined: a triple $\{x, y, z\}$ is the pair $\{\{x, y\}, z\}$, where the first element is itself a pair, and so forth. Of course, only *finite* strings can be formed this way, but we'll soon do better.

[3] "Free variables" are those whose value matters to the expression containing them, like x in $x^2 + 2x + 1$ (whose value depends on x), as opposed to "bound" or "dummy variables", like y in $\int f(y)\,dy$. I shall not attempt to be more rigorous (see [Ha] and the article "Symbolic logic" in [It]). Instead, I hope to convey some feeling for this by accumulating examples.

Properties—for example, positivity of real numbers, positive-definiteness of matrices, solenoidality of vector-fields, etc.—are predicates involving such objects. If p is a property, one denotes by $\{x \in X : p(x)\}$ the subset of X made of all elements for which this property holds true. For some immediate examples, consider a subset R of the Cartesian product X × Y. Its *section* by x is $R_x = \{y \in Y : \{x, y\} \in R\}$. (Beware it's a part of Y, not of X ! Cf. Fig. A.1.) Its *projection* on X is $p_X(R) = \{x \in X : R_x \neq \emptyset\}$. Any property thus defines a subset, and a subset A defines a property, which is $x \in A$. Since subsets and properties are thus identified, operations on sets translate into operations on properties: thus, for example, $\{x \in X : p(x)$ **and** $q(x)\} = \{x \in X : p(x)\} \cap \{x \in X : q(x)\}$, and the same with **or** and ∪.

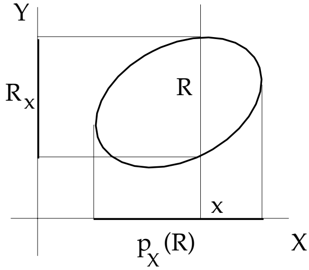


FIGURE A.1. Notions of section and projection.

A.1.2 Relations and functions

A *relation* is a triple $r = \{X, Y, R\}$, where X and Y are two sets and R a subset of X × Y, called the *graph* of the relation. Objects x and y such that $\{x, y\} \in R$ are said to be related (or linked) by r. There are various shorthands for the predicate $\{x, y\} \in R$, such as $r(x, y)$ or (more often) the so-called "infix" notation $x\ r\ y$. (Familiar examples of the latter are $x \leq y$, $u \perp v$, etc.) The *domain* and *codomain* of r are the projections $p_X R$ and $p_Y R$, thus denoted:

[4] Of course, when a sentence such as "x > 0" or "div b = 0" appears in a text the aim of which is not primarily mathematical, we assume this predicate has the value **true**. In fact, the author is usually telling us just that, but won't risk the ridicule of saying "I have just proven that the predicate 'div b = 0' is true". However, the occurrence in such texts of bits of formal reasoning, as for instance when discussing the truth of the statement, "if div b = 0, there exists some a such that b = rot a", clearly shows that maintaining the distinction between a predicate and the assertion that this predicate is true (which is of course, another predicate) is not only a formal game. Sometimes, it's the only way to settle an argument.

$$\mathrm{dom}(r) = \{x \in X : R_x \neq \varnothing\}, \quad \mathrm{cod}(r) = \{y \in Y : R_y \neq \varnothing\}.$$

So, informally, the domain dom(r) contains all those x of X that relate to some y in Y, and the codomain cod(r) is the symmetrical concept: all y's related to some x. (The codomain of r is also called its *range*.) The *inverse* of r is the relation $r^{-1} = \{Y, X, R\}$, and the domain of one is the codomain of the other.

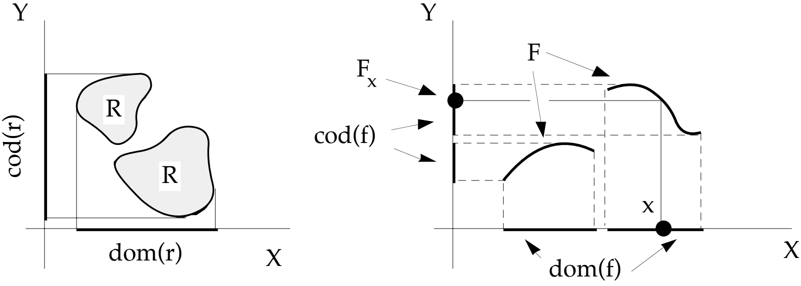


FIGURE A.2. Left: Domain and codomain. Right: A functional graph F.

The graph F of a relation f = {X, Y, F} is *functional* if each section F_x contains at most one element of Y (Fig. A.2, right). The relation f is then called a *function* "from X to Y", or a "Y-valued function on X". The function f is *surjective* if cod(f) = Y, *one-to-one* if its inverse f^{-1} (which is always defined, but only as a relation, a priori) is also a function, then called the *inverse function,* or *reciprocal* of f. The set of all functions from X into Y will be denoted by X → Y, and if f is such a function, we'll say that "the *type* of f is X → Y". The construct "$f \in X \to Y$" thus makes sense, under the convention that "→" has precedence over "∈", and I occasionally use it, but "f : X → Y" is the standard way to introduce a function of type X → Y.

Thus, all functions are a priori *partial*, that is, dom(f) may be strictly smaller than X. *Total* functions are those for which dom(f) = X. A total function is *injective* if it is one-to-one, surjective, as we just said, if cod(f) = Y, and *bijective* if both properties hold. Total functions are called *mappings* or *maps*, but again, excessive emphasis on such fine semantic distinctions is not very productive. Better take "function" and "mapping" as synonyms, and call attention on whether dom(f) = X or not, when necessary. Most functions in this book are partial.

Some relations (in the common sense of the word) between physical entities are better conceived as general relations than as functional ones. A good example is provided by "Bean's law", an idealization of what happens in a type-II superconductor when all currents flow parallel to some given direction. The scalar components j and e of the current density and the electric field along this direction are then supposed to be related as follows: if $e \neq 0$ at some point, then j at this point is equal to some characteristic value j_c, called the *critical current*, and the sign of j is that of e; if $e = 0$, then any value of j between $-j_c$ and j_c is possible, and which one actually occurs at any instant depends on the past evolution of e. This relatively complex prescription is elegantly summarized by a (non-functional) relation: The pair {e, j} must belong to the graph of Fig. A.3, left.

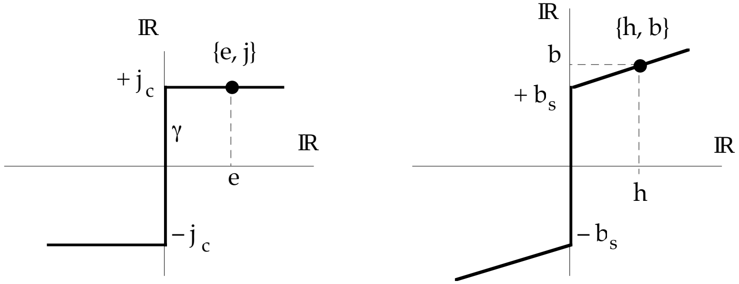


FIGURE A.3. Left: Bean's law, for type-II superconductors, expressed as a non-functional relation {IR, IR, γ}. Right: A similar idealization of the b–h characteristic of a "soft" ferromagnetic material.

The same trick is useful to express b–h constitutive laws in similar circumstances (horizontal currents, vertical magnetic field). If, as it happens for instance in induction heating simulations, one works over a large range of values of h (some 10^5 A/m, say), the hysteretic cycle is so narrow, relatively speaking, that one may as well ignore hysteresis. Hence the behavior depicted in Fig. A.3, right. Again, this b–h relationship is conveniently expressed by a non-functional relation, i.e., a graph.

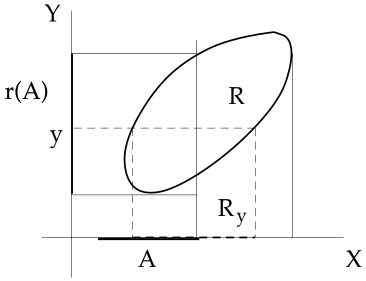


FIGURE A.4. Image of A under r = {X, Y, R}.

Let's proceed with concepts that are common to relations and functions. If A is a part of X, its *image* under r, denoted by $r(A)$, is

$$r(A) = \{y \in Y : R_y \cap A \neq \varnothing\}$$

(cf. Fig. A.4). Note that $r(X) = \mathrm{cod}(r)$. If $A = \{a\}$, a single element set, we write $r(a)$ instead of $r(\{a\})$, and call this set the image of a under r. Note that $r(x) = R_x$, hence the syntax $y \in r(x)$ as another way to say that x and y are r-related. For $B \subseteq Y$, the set $\{x \in X : R_x \cap B \neq \varnothing\}$, denoted $r^{-1}(B)$, is called the *inverse image* or *pre-image* of B, and $r^{-1}(Y) = \mathrm{dom}(r)$. Remark that $\mathrm{cod}(r) = \{y \in Y : r^{-1}(y) \neq \varnothing\}$. Relation $s = \{X, Y, S\}$ is *stronger* than relation r if $S \subseteq R$. This is obviously another relation,[5] between relations, which can logically be denoted by $s \subseteq r$.

We need mechanisms to build new relations from old ones.[6] Since relations are graphs, and hence sets, operations on sets apply to relations: given $r = \{X, Y, R\}$ and $s = \{X, Y, S\}$, one can form the new relation $\{X, Y, R \cap S\}$ (stronger than both r and s), which we denote by r **and** s. Similarly, r **or** $s = \{X, Y, R \cup S\}$ (weaker than both r and s). In the special case $S = A \times Y$, where A is a part of X, r **and** s is called the *restriction* of r to A (Fig. A.5). Its domain is $\mathrm{dom}(r) \cap A$. It's usually denoted by $r_{|A}$. If $r = s_{|A}$ for some $A \subseteq X$, one says that s is an *extension* of r.

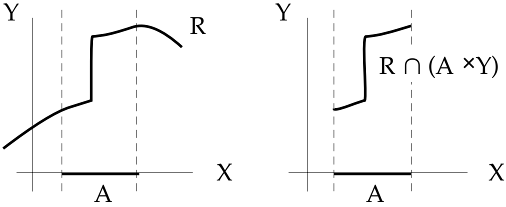


FIGURE A.5. Restricting relation $r = \{X, Y, R\}$ to A. An alternative definition of the restriction is $r_{|A} = r \circ \mathrm{id}(A)$, where $\mathrm{id}(A) = \{X, X, \Delta \cap (A \times X)\}$ (see the definition of Δ p. 270).

[5] Can you describe its graph? Is it an order (as defined below)? A total one or only a partial one?

[6] *Computer programming* is basically just that. A program is a function p defined on the set S of possible states of the machine; entering data selects some $s \in S$, and $p(s)$ is the final state, including output display; the game consists in building p from a set of basic "instructions", which are, as one may show, functions of type $S \to S$. So programming consists, indeed, in building new functions from old ones. (For serious developments on this, see a treatise on "functional programming", for instance [Hn].) Here, we need not go so far as formally presenting a programming language (the rare bits of programs that appear in this book should be self-explanatory), but some awareness of the underlying mechanisms may be useful.

Relations can also be composed, when their sets match properly: Given $r = \{X, Y, R\}$ and $s = \{Y, Z, S\}$, the *composition* of r and s, denoted by $s \circ r$, is

$$s \circ r = \{X, Z, \cup \{r^{-1}(y) \times s(y) : y \in Y\}\}.$$

This amounts to saying that $z \in (s \circ r)(x)$ if and only if there is at least one y such that $y \in r(x)$ and $z \in s(y)$ (Fig. A.6). The composition $g \circ f$ of two functions of respective types $X \to Y$ and $Y \to Z$ is a function of type $X \to Z$.

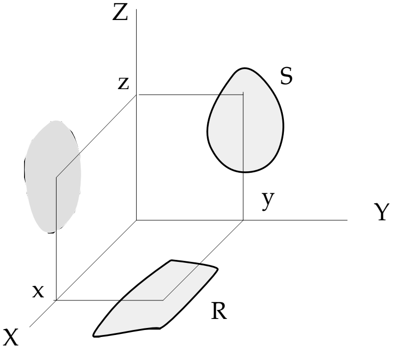


FIGURE A.6. Composing two relations. When {x, y} and {y, z} span R and S respectively, {x, z} spans the graph of the relation s o r.

A.1.3 Families

Family is just another name, more appropriate in some contexts, for "total function". If $\mathcal{J}$ is a set (finite or not), and X is a set, a *family of objects of type* X, *indexed by* $\mathcal{J}$, is a mapping from $\mathcal{J}$ into X, denoted by $\{x_i : i \in \mathcal{J}\}$. To each label i, taken in $\mathcal{J}$, thus corresponds an object x_i, of type X. It is convenient—although a bit confusing, perhaps—to denote the set of all such families by $X^{\mathcal{J}}$. The set $\mathcal{J}$ can be finite, in which case it may seem we have redefined n-tuples. Not so: for $\mathcal{J}$, finite or not, is *not* supposed to be ordered. So we really have a new concept[7] here.

The distinction may sometimes be useful. Think of the nodes of a finite element mesh. They form a set $\mathcal{N}$, usually finite. Let us call N the

[7] It's not "unordered n-tuple", either, which we chose earlier to interpret as an n-element subset of X. In the case of families, repetitions are allowed, and two labels i and j can point to the same object, $x_i = x_j$. A part Y of X can always be considered as a family, however, by indexing it over itself: $Y = \{x_y : y \in Y\}$, where $x_y = y$. So one should not worry too much about such fine distinctions.

number of nodes (that is, $N = \#\mathcal{N}$, if one wishes to use this convenient shorthand for the number of elements in a set). Suppose a real-valued "degree of freedom" (DoF) is assigned to each node. We thus have a family $\{\mathbf{u}_n : n \in \mathcal{N}\}$ of N real numbers, indexed over $\mathcal{N}$, that will be denoted by $\mathbf{u}$, in boldface. Here, X is the set $\mathbb{R}$ of real numbers, the index set $\mathcal{J}$ is $\mathcal{N}$, and $\mathbf{u}$ is thus a member of $\mathbb{R}^{\mathcal{N}}$, with the above notation. If you think, "This $\mathbf{u}$ is an N-dimensional real vector", you are right, for indeed, $\mathbb{R}^{\mathcal{N}}$ is a real vector space of dimension N. But you should resist the natural compulsion to say, "So, this is an element of $\mathbb{R}^N$ (the Cartesian product of $\mathbb{R}$ by itself, N times) and hence, an N-tuple". An N-element family is not an N-tuple, because no order among its members is implied. Nodes are labelled, not numbered. A family is less structured than an N-tuple, in this respect.

Remark A.1. Nodes are not *yet* numbered, that is. True, at some stage in the process of finite element modelling, a numbering scheme is introduced: When solving for the DoF, by using the Gauss–Seidel method for instance, there will be a first DoF, a second DoF, etc. But sound programming methodology demands that this numbering be deferred to the very stage where it becomes relevant and useful. Moreover, one may have to deal with several different numbering schemes for the same set of nodes—if only to test numbering schemes [KB] for efficiency (they affect the bandwidth of the matrices and the speed of iterative methods). Such a *numbering* will be a (bijective) mapping from $\mathcal{N}$ onto $[1, N]$, the segment of the first N nonzero integers, and it will assign to each N-member family an N-tuple, in a one-to-one way. So once a numbering is given, there is an identification (an isomorphism—see Note 11, p. 276) between $\mathbb{R}^{\mathcal{N}}$ and $\mathbb{R}^N$. But this is not a "canonical" identification, since it depends on the numbering. $\mathbb{R}^{\mathcal{N}}$ and $\mathbb{R}^N$ are definitely not the same object. ◊

A.1.4 Binary relations

We now look at the case where $Y = X$. A relation $r = \{X, X, R\}$, then called a *binary* relation in X, confers on X some structure, that X alone did not possess. Thus, the compound "X as equipped with the relation r" (that is, the pair $\{X, r\}$), is a new object, for which the notation $\{X, r\}$ is appropriate.

Two standard examples of binary relations, equivalence and order, will come to mind. Let us call the part $\Delta = \{\{x, y\} \in X \times X : x = y\}$ of $X \times X$ the *diagonal*, and the relation $\mathrm{id} = \{X, X, \Delta\}$ the *identity*. A relation r is *reflexive* if its graph contains Δ, that is, if $\mathrm{id} \subseteq r$, *symmetric* if $r^{-1} = r$, *antisymmetric* if $(r \textbf{ and } r^{-1}) \subseteq \mathrm{id}$, *transitive* if $(r \circ r) \subseteq r$. A reflexive, transitive, and symmetric (resp. antisymmetric) relation is an *equivalence* (resp. an *order,* or *ordering*). Cf. Fig. A.7.

Generic notation for equivalences and orders is $\equiv$ and $\leq$ (or $\subseteq$), and one uses expressions such as *lesser than, greater than,* etc., instead of symbols, occasionally. If $r = \{X, X, R\}$ is an order (e.g., the relation $\leq$ in $\mathbb{R}$), one calls $\{X, X, R - \Delta\}$ the *strict* associated relation (example: $<$ in $\mathbb{R}$ is thus associated with $\leq$), enunciated as *strictly lesser than*, etc. Such relations, for which the generic notation is $<$, are *not* orders (beware!), so one tends to avoid them; hence the use of contrived expressions, such as *nonnegative* for $\geq$, to avoid the ambiguous "positive" (is it $>$ or $\geq$?).

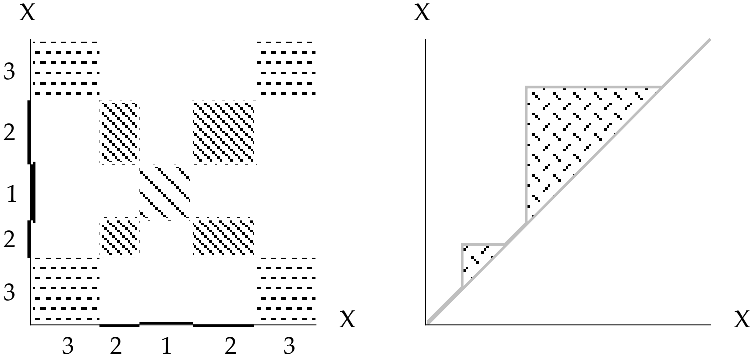


FIGURE A.7. Structure of the graph R (the shaded set) for an equivalence (left) and a partial order (right). Notice how R, on the left, splits into separate parts of the form $X_i \times X_i$ (i = 1, 2, 3 here, corresponding to three different shading textures), where each X_i is an equivalence class (see below, A.1.6). The shaded set includes Δ in both cases. See also Fig. A.8.

A.1.5 Orders

An order r is *total* if $r \cup r^{-1} = \{X, X, X \times X\}$, that is, for all pairs $\{x, y\}$, either $x \in r(y)$ or $y \in r(x)$. Total orders, like $\leq$ in $\mathbb{R}$ or $\mathbb{N}$, are also called *linear* orders, which makes intuitive sense: They line up things.

A nice standard example of partial order is divisibility in $\mathbb{N}$. (See also Note 5.) A more topical one, for us, is on the set $\mathcal{M}$ of all possible finite element meshes of a given region: A mesh *m'* is a refinement of a mesh *m* (one may say *m'* is *finer than m*) if each edge, face or volume of *m* is properly meshed by a suitable restriction of *m'*. Two different meshes may have a common refinement without any of them being finer than the other, so the order is only partial.

The supremum should not be confused with what is called a *maximal element* in A, that is, some x in A such that $x \leq y$ never holds, whatever y in A. In Fig. A.8, f and g are maximal in A, d and e are *minimal* (and a is minimal in X). Maximal elements are not necessarily unique, and may not exist at all (the open interval]0, 1[has none).

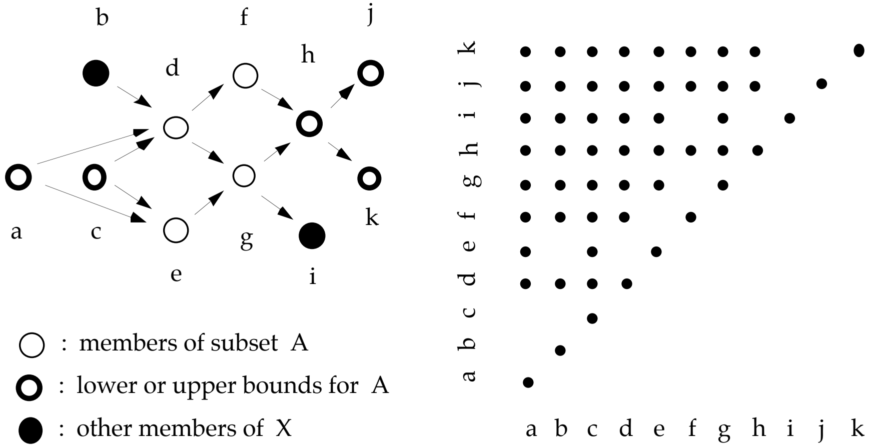
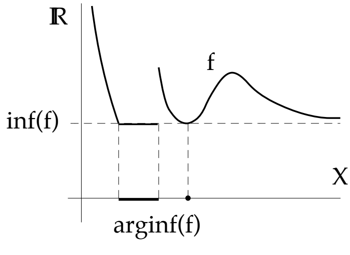


FIGURE A.8. A partial order on an 11-element set, X = {a, b, . . . , k}. Right: According to the graphic convention of Fig. A.7. Left: As a *sagittal* graph (i.e., with arrows), which is much more convenient in the case of transitive relations (because most arrows can be omitted, as for instance the one from a to f).

Notions of inf and sup pass to functions $f \in X \to Y$, when Y is ordered ($Y = \mathbb{R}$, most often). The infimum inf(f) of a function is the infimum inf(f(X)) of its image. More explicit notation, such as inf{f(x) : $x \in X$}, is generally used. The pre-image of inf(f(X)), that is to say, the subset of elements of X that realize the minimum of f, is denoted arginf(f), or arginf({f(x) : $x \in X$}) (but one will not bother with the double system of parentheses, usually). Note that arginf(f) is not an element of X but of $\mathcal{P}(X)$, which can be the empty set. One says that f "reaches its minimum" on arginf(f).



The index set $\mathcal{J}$ of a family may be ordered. If the order is total, a family $\{x_i : i \in \mathcal{J}\}$ is called a *sequence*. (The use of this word, in general, rather implies that $\mathcal{J} = \mathbb{N}$, or some subset of $\mathbb{N}$. If $\mathcal{J} = \mathbb{R}$, or some interval of $\mathbb{R}$, one will rather say something like *trajectory*.) This is the generalization of n-tuples, when $\mathcal{J}$ is infinite. If the order is only partial, we have a *generalized sequence*. The family of all finite element approximate solutions to a field problem is one.

Finally, a *minimizing sequence* for a function $f : X \to Y$ is a family $\{x_n : n \in \mathbb{N}\}$ of elements of X such that $\inf\{f(x_n) : n \in \mathbb{N}\} = \inf(f)$. The standard way to prove that $\text{arginf}(f)$ is not empty is to look for the limit of a minimizing sequence.

A.1.6 Equivalence classes, "gauging"

If r is an equivalence in X, the set $r(x)$ is called the *equivalence class* of x. Equivalence classes are disjoint, and their union is all of X (Fig. A.7, left). In other words, an equivalence relation generates a *partition* of X into equivalence classes (and the other way around: A partition induces an equivalence relation).

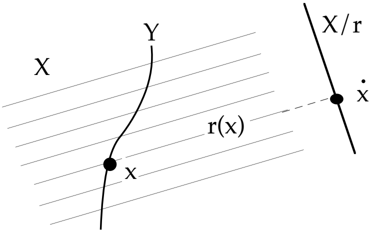


FIGURE A.9. Equivalence classes, quotient, representative section.

This provides one of the most powerful mechanisms for creating new objects in mathematics (and this is why the previous notions deserved emphasis). When objects of some kind are equivalent in some respect, it's often worthwhile to deal with them wholesale, by dumping all of them into an equivalence class, and treating the latter as a new, single object. If X is the initial set and r the equivalence relation, the set of classes is then called the *quotient* of X by r, with various denotations, such as X/r for instance. Don't confuse the quotient, elements of which are not of type X, with what one may call a *representative section* (Fig. A.9), which is a subset of X "transverse to the classes", so to speak, obtained by picking one element x (the *representative element*) in each equivalence class. The quotient is in one-to-one correspondence with each representative section. Among such sections, some may be more remarkable than others, depending on which structure X possesses.

Examples abound, and many appear in this book. Let us just mention the following, of special interest in electromagnetism. If some field b is divergence-free in some region of space, there may exist, under conditions which are not our present concern, a field a such that $b = \text{rot } a$, called a "vector potential" for b. Such a field is not unique (one may always add a gradient to it). Among vector potentials, the relation $\text{rot } a_1 = \text{rot } a_2$ is an equivalence, the classes of which are obviously in one-to-one relation with the b's.

The various representations of the electric field e provide a more involved example. From Faraday's law ($\partial_t b + \text{rot } e = 0$), and by using the above representation $b = \text{rot } a$, we have $e = -\,\partial_t a - \text{grad } \psi$, where ψ is called the "scalar (electric) potential". Calling A and Ψ the sets of suitable potentials a and ψ (they should satisfy some qualifying assumptions, which we need not give here explicitly), we have an equivalence relation in $A \times \Psi$: Two pairs $\{a_1, \psi_1\}$ and $\{a_2, \psi_2\}$ are equivalent if they correspond to the same electric field, i.e., if $\partial_t a_1 + \text{grad } \psi_1 = \partial_t a_2 + \text{grad } \psi_2$, over some specified span of time.

One may conceive all pairs $\{a, \psi\}$ in a given class as mere representations (all equivalent) of some electric field, but the mathematical point of view is bolder: The equivalence *class,* taken as a whole, *is* the same object as the electric field. Dealing with e (in numerical simulations, for instance), or dealing with the whole class of $\{a, \psi\}$s, is the same thing. But of course, a vector field and a class of pairs of fields are objects of very different nature, and doing the mental identification may not be easy. Hence the more conservative approach that consists in selecting among the members of an equivalence class some distinguished one, as representative of the class.

For instance, we may privilege among the pairs $\{a, \psi\}$ of a given class (i.e., a given e) the one for which $\text{div } a = 0$. (There is only one, if we work in the whole space. Otherwise, additional boundary conditions are needed to select a unique a.) Such a specification for selecting one member in each equivalence class is called a *gauging* procedure. (The previous one is "Coulomb gauge". Imposing $c^2\, \partial_t \psi + \text{div } a = 0$ is "Lorenz[8] gauge".) Now, one may feel more assertive in dealing with *the* pair $\{a, \psi\}$ as a representation of e. Things go sour, however, when one entertains the illusion that this pair would be more deserving, more "physical", than its siblings of the same class, that it would be the "right" one, in some way. Such futile concerns about gauging have delayed the implementation of 3D eddy-current codes for years in some institutes.

Remark A.2. The same delusion seems to be at the root of persistent misunderstanding about the Aharonov–Bohm effect. (Cf. [AB]: Interference experiments on electrons detect the existence of an induction flux inside an

[8] According to [NC], it's L. Lorenz, not H.A. Lorentz.

extremely thin tightly wound solenoid, in spite of the field being null outside it. This is paradoxical only when one insists on thinking of electrons as *localized* classical objects which, according to such a naive view, "have to pass" in the region where $b = 0$ and thus "cannot feel the influence" of b, whereas they could "feel" that of a.) The effect, some argue, points to the vector potential a rather than the induction b as the "most fundamental" descriptor[9] of the field. If the issue really is, "Of the two *mathematical* objects, b on the one hand, and the whole class of a's such that $b = \text{rot } a$ on the other hand, which one must be considered as 'the primitive concept' ?", one may conceivably take sides. This is a choice between two different *formalisms* for the *same* theory, since the two objects are in one-to-one correspondence and describe exactly the same physics (which is why no experiment can resolve such an issue). Undeniably, when it comes to quantum field–particle interactions, having a in the equations is more convenient. But since the equations are gauge-invariant, none of the representatives of the class is thus privileged, so there is nothing in the AB effect that would give arguments to consider the "Coulomb gauged", or the "Lorenz gauged" vector potential as the *physical* one. The frequent claim (cf., e.g., [Kn]) that AB would allow one to *measure* the Coulomb gauged vector potential is totally misleading. What can be measured, by determining the electron's phase shift, is the induction flux, from which of course the Coulomb gauged a is readily derived in the axisymmetric situation usually (and needlessly) assumed. ◊

A.1.7 Operations, structured sets

Operations are functions of type $X \to X$ (*unary* operations), $X \times X \to X$ (*binary* operations) etc., i.e., functional relations that map n-tuples of X to elements of X. The reader may wish to translate the standard concepts of *commutativity* and *associativity* of operations in terms of graphs of such relations.

Sets under consideration in specific questions are not naked, but structured, by relations and—mostly—operations. We saw how an equiva-

[9] In some cases, which verge on the tendentious, the observed interferences are said to be "due to the vector potential", and hence (the reader is subtly led to conclude, although it's never explicitly said) "not due" to the (magnetic induction) field. This is silly. The involved phase factor can be computed indifferently in terms of a (as $\exp[-iq/\hbar \int_\gamma \tau \cdot a]$, where γ is a loop around the solenoid), or b (as $\exp[-iq/\hbar \int_\Sigma n \cdot b]$, where Σ is some surface bounded by γ, and therefore, pierced by the solenoid). Both expressions yield the same value, by the Stokes theorem. The latter may be less *convenient*, in the thin-solenoid case usually discussed, for b is then a distribution, but this is a side issue. Anyhow, there is no need to postulate a "thin" solenoid to discuss the AB effect (cf. Exers. 8.5 and 8.6).

lence or an order could, already by itself, structure a set. But usually there is much more: On $\mathbb{R}$, for instance, there is order, addition, multiplication, division, and all these relations interact to give the set the structure we are used to. The same goes for all standard sets, such as $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{Z}$, $\mathbb{C}$, and so forth. So what is called $\mathbb{R}$ is in fact $\{\mathbb{R}, \leq, +, *, \ldots\}$, that is, the set in full gear, equipped with all its structuring relations, and this is where the concept of *type* is useful: A type is a structured set.[10]

Moreover, objects of different types may interact via other relations, hence an encompassing structure, informally called an *algebra*. (Appendix B gives a detailed example), which further stretches the notion of type. So when I say that x is "of type X", as has happened several times already, I mean not only that x is a member of the set X, but that x can enter in relation with other objects, belonging to X or to other sets, to all the extent allowed by the rules of the algebra. For instance, when X is $\mathbb{N}$, the integer n can be added to another integer m, can be multiplied by it, etc., but can also be added to a real number, serve as exponentiation factor, etc. The type of an object, in short, encompasses all one can do to it and with it.[11]

The practice of denoting the type and the underlying set the same way has its dark side: If one is fond of very compact symbols, as mathematicians are, some overloading is unavoidable, for the same symbol will have to represent different types. For example, $\mathbb{R}^3$ normally stands for the set of triples of real numbers. It is quite tempting to use this symbol also to denote structures which are isomorphic to $\mathbb{R}^3$, like the three-dimensional real vector space, or ordinary 3D space. This is highly questionable, for the operations allowed on triples, on 3-vectors, and on points of space are not the same. (We'll elaborate on that later. But already it is clear that points, as geometric objects, cannot be added or multiplied by scalars, the way vectors can.) Hence the occasional appearance in this book of long symbols like *POINT* or *VECTOR* to denote different types built upon the same underlying set (namely, $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$).

We'll work out the simple example of *Boolean algebra*, denoted B, to see how a few operations can give rich structure to even the less promising set, one with two elements, labeled **true** and **false**. Two relations will structure it. (The corresponding type is the one called *LOGICAL* in most programming languages.) The first operation is **not** = {B, B, NOT}, the

[10] An *isomorphism* between $\{X, r_1, r_2, \ldots\}$ and $\{X', r'_1, r'_2, \ldots\}$ is then a one-to-one map f such that $x\, r_i\, y \Leftrightarrow f(x)\, r'_i\, f(y)$ for all x, y, i, that is, a structure-preserving bijective map. If there is only one obvious sensible choice for f, one says the isomorphism is *canonical*.

[11] "Object", here, has the same sense as in "object-oriented programming" [Me].

graph of which, NOT, is depicted in Fig. A.10. It's the subset of B × B consisting of the two pairs {**true**, **false**} and {**false**, **true**}, out of four in all. This graph is functional, and the unary operation **not** it defines is indeed what was expected, turning **true** into **false**, and vice versa. The second function is **and** = {B × B, B, AND}, where the graph AND, functional again, now contains four elements out of the possible eight, as shown in Fig. A.10 (where the limits of the graphical representation we used till now become obvious; hence the preferred use of *tabular* representations for binary operations, as in Fig. A.11). We may now define the new function **or** by x **or** y = **not**(**not**(x), **not**(y)), and combine them in various ways.

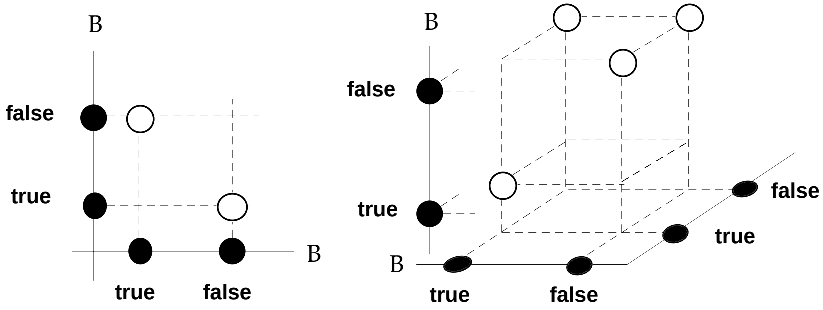


FIGURE A.10. Left: Graph of the "unary" function **not**. Right: Graph of the binary function **and**.

	T	T
T	T	F
F	F	F

and

	T	F
T	T	T
F	T	F

or

	T	F
T	F	T
F	T	F

xor

	T	F
T	T	F
F	T	T

⇒

FIGURE A.11. Some operations in Boolean algebra (**T** and **F** stand for **true** and **false**), in tabular representation. (Variable x spans rows of the table, y spans columns, and the entry at x—y is the truth value of x r y.) Note how the symmetry of relations, or lack thereof, is rendered.

A.1.8 Logic

Which leads us into propositional calculus, and logic. Propositions, and predicates more generally, can be seen as B-valued functions, whose domain is the set of all possible "well-formed" expressions (that is, all strings of symbols that conform to some specific grammar, which only logicians and programming-language designers take the trouble to make explicit). Since predicates are thus relations, the above building mechanisms apply to them. For instance, if p and q are predicates, p **and** q is one, too: Its value, according to the above definition of **and**, is **true** when both p and q assume the value **true**, and **false** otherwise. The algebraic structure of B allows a lot of similar constructions: **not** p, p **or** q, p **and** (q **or** r), etc.

One among these constructs, q **or not** p, whose value is **false** when p = **true** and q = **false**, and **true** in all other cases (Fig. A.11), is so frequently used that it deserves a special notation: $p \Rightarrow q$.[12] The abuse, and perhaps even the use of this, should be discouraged. Better use "if p, then q". The difference is only a matter of concrete syntax,[13] but it seems to matter much, and such a plain sentence is less prone to confusion than "$p \Rightarrow q$".

Two other important shorthands should be known, ∀ and ∃. They too allow new predicates to be built from old ones. Suppose p(x) is some predicate containing the variable x in which x is free. Then $p(x)\ \forall x$ is a new predicate, the value of which is **true** if and only if p(x) = **true** for all possible values of x, and now the variable x is *bound*.[14] Symmetrically, the truth value of $p(x)\ \exists x$ (or, with a more readable syntax, $\exists x : p(x)$) is **false** if and only if p(x) = **false** for all possible values of x. These symbols are dreaded by many engineers, and perhaps not without some reason, for their abuse in mathematical training during the 1970s has done much harm, worldwide. They should be used very sparingly, especially the latter, and there is alternative concrete syntax, such as p(x) **for all** x, or even "p(x) holds for all x" and "p(x) holds for some x", much closer to natural language, while still being unambiguous.

[12] So, be wary of the informal use of $p \Rightarrow q$, voiced as "p implies q". (See [Hr] for a nice discussion of this and similar issues.) The risk is high that "q is true" will be understood, which may be wrong. The safe use of this in reasoning demands that *two* different statements be proved: that $p \Rightarrow q$ is true (whatever the values of the free parameters in both p and q, which may affect their truth values) *and* that p is true. One can then conclude that q is true. Misuses of this basic and celebrated logical mechanism are too often seen. The most common mistake consists in carefully proving that $p \Rightarrow q$ holds, while overlooking that p can be false (for some values of the free variables that appear in it), and to go on believing that q has been proved.

[13] Abstract syntax deals with the deep structure of formal expressions (including programming languages). Concrete syntax is concerned with the choice of symbols, their position, how they are set, etc. See [Mr].

[14] One may fear some logical loophole here, but no worry: the definition of "free" and "bound" (they are antonyms) is recursive. If a variable appears at only one place, it's free, and the only way to bind two occurrences of the same variable is to invoke one among a limited list of binding mechanisms, including the use of the so-called "quantifiers" ∀ and ∃, as described above.

There is no reason to deny oneself the convenience of a shorthand, however—for instance, in constructs of the form "*Find* h *such that*

(p) $$i\omega \int \mu\, h \cdot h' + \int \sigma \operatorname{rot} h \cdot \operatorname{rot} h' = 0 \quad \forall\, h' \in \mathbb{H}.\text{"}$$

This one means that the equality $i\omega \int \mu\, h \cdot h' + \int \sigma \operatorname{rot} h \cdot \operatorname{rot} h' = 0$ should hold whatever the test field h', provided the latter is selected within the allowed class of such fields, which class is denoted by $\mathbb{H}$. The predicate (p), in which variable h' is bound whereas h is free, thus expresses a property of h, the unknown field, the property that characterizes h as the solution to the problem at hand. To say "*find* h *such that* $(\ldots) = 0$ *for all* h' *which belong to* $\mathbb{H}$" is the same prescription, only a little more verbose. But to omit the clause "$\forall$ h' . . . " or "for all h' . . . ", whatever the concrete syntax, would be a capital sin, turning the precise statement of a problem into gibberish. The issue is not tidiness, or lack thereof, but much more importantly, *meaning:* Without the clause "$\forall\, h' \in \mathbb{H}$", the problem is not posed at all.

A.1.9 A notation for functions

Before leaving this Section, I wish to explain an idiosyncrasy that you also may find convenient at times.

Many functions are defined via algebraic expressions. Take for instance the expression[15] $x^2 + 2x + 1$. The set $\{\{x, y\} \in \mathbb{R} \times \mathbb{R} : \ y = x^2 + 2x + 1\}$ defines a functional relation, f say. I find convenient, time and again, to write this

(e) $$f = x \rightarrow x^2 + 2x + 1,$$

which should be parsed as suggested by Fig. A.12 and understood as follows: "Let's name f the function that maps the real number x to the real number that results from evaluating the expression $x^2 + 2x + 1$." (Of course, the arrow should not be read as "tends to", according to the more standard convention, which I avoid, except in unambiguous constructs such as $\lim_{\varepsilon \to 0} \ldots$, etc.)

[15] An *expression* is just a combination of symbols that conforms to some definite syntax. In *algebraic* expressions, two kinds of symbols are allowed: variables or constants, of definite types (here, the real x and the integer 2), and relational symbols (here, + and the exponentiation) that belong to a specific algebra (here, standard arithmetic). To *evaluate* the expression consists in assigning to the variables definite values, and doing the computation according to the rules of the algebra. Note that the same expression could make sense in other algebras: x can be a matrix, for instance.

$$\boxed{f = \boxed{x \to \boxed{\boxed{x^2} + 2x + 1}}}$$

FIGURE A.12. Parsing (e), that is, finding its logical structure, here indicated by the hierarchy of nested boxes, as a syntactic analyzer would do, if instructed of precedence rules: Multiplication and exponentation take precedence over addition, the arrow is weaker than all operational symbols, and = is the weakest link of all.

On both sides of the equal sign in (e) we have the same mathematical object, a function, only differently tagged: by its name f on the left, and by the whole expression $x \to x^2 + 2x + 1$ on the right. (Variable x is bound in this expression: another example of the binding mechanism.) So the equal sign is quite legitimate at this place. On the other hand, the arithmetic expression $x^2 + 2x + 1$ (where x is a free variable) is *not* the function, and to write $f = x^2 + 2x + 1$ would be highly incorrect.[16]

Why not simply $f(x) = x^2 + 2x + 1$? This is a bit ambiguous, because it can also stand for the statement of an equality, unless you declare explicitly your intention to use it as a function definition. Hence the frequent use of a special symbol, $\hat{=}$ or :=, for "is defined as", like this: $f(x) \hat{=} x^2 + 2x + 1$. But if special symbol there must be, better choose the arrow, which puts emphasis on the right object, the defined one, which is f, not[17] $f(x)$. Also, the arrowed notation can be nested without limits, as the following example will show.

Given a function q on 3D-space, which may represent for instance an electric charge density, one may define its Newtonian *potential* (the electric potential, in that case, up to the factor ε_0) as follows:

$$\psi = x \to \frac{1}{4\pi} \int \frac{q(y}{|x - y|} \, dy,$$

where dy is the volume element and $|x - y|$ the distance between points x and y. (Observe that y is bound in the integral—yet another binding

[16] If you think I insist too much on such trivia, pay attention to the practice of physics journals: most often, $f(x)$ refers to a *function*, and f to its *value*. Mathematicians do exactly the opposite: f is the function, $f(x)$ its value at x. This schism is all the more detrimental to science in that it goes generally unnoticed.

[17] One might argue that $f(x) \hat{=} x^2 + 2x + 1$ needs some quantifier, such as perhaps $\forall$, to really define f. Actually, a quantifier has been designed for just that purpose: the λ of "lambda-calculus". Cf. [Kr].

mechanism—and x free, and how the arrow binds x.) Now, one may define a new function, of higher level: $G = q \to \psi$, that is, the operator (named after Green) that maps q to ψ. Instead of this two-step definition, we may, thanks to the arrowed notation, write

$$G = q \to (x \to \frac{1}{4\pi} \int \frac{q(y)}{|x - y|}\, dy\,),$$

in one stroke. (Parentheses force the correct parsing.) This is a precious shortcut at times, to be used sparingly, of course.

This is the first example we encounter of a function defined on a set whose elements are themselves functions, and which maps them to other functions. For clarity, such functions are called *operators* (especially when, as in the present case, the correspondence is linear). When the set they map to is $\mathbb{R}$, the word *functional* is used (cf. p. 62).

The arrowed notation is especially useful when variables and parameters occur together. Examine this:

$$\text{grad}(y \to \frac{1}{|x-y|}) = y \to \frac{x-y}{|x-y|^3}$$

(an equality between two vector fields, since $x - y$ is a vector). Here, x is the parameter, y the variable, and there is no ambiguity as to which gradient, with respect to x or to y, we mean.

Expressions other than arithmetic can be put on the right of the arrow: conditional expressions, and even whole programs. For example, we may have this

$$g = x \to \textbf{if}\ x \geq 0\ \textbf{then}\ x^2 + 2x + 1\ \textbf{else}\ 0,$$

$$h = x \to \textbf{if}\ x \geq 0\ \textbf{then}\ x^2 + 2x + 1.$$

The difference between g and the previously defined f (cf. (e)) is clear (they differ for $x \leq 0$), but what about h with respect to f ? They surely differ, since $\text{dom}(f) = \mathbb{R}$, whereas $\text{dom}(h) = \{x \in \mathbb{R} : x \geq 0\}$. Yet their defining expressions are the same. But h is the restriction of f to the positive half-line. As this example shows, the domain of a function should always be described with precision, for the expression or fomula or recipe for evaluating the function may well make sense *beyond* this domain.[18] We'll see this phenomenon recur when we study the differential operators grad, rot, div. *Different* operators will similarly be called

[18] Consider $f = x \to (x^2 - 2x + 1)/(x - 1)$, with $\text{dom}(f) = \mathbb{R} - \{1\}$.

"gradient", for instance, and will differ by the extent of their respective domains.

A.2 IMPORTANT STRUCTURES

A.2.1 Groups

Groups are important, not only because many mathematical structures like linear space, algebra, etc., are first and foremost groups, with added features, but as a key to *symmetry*.

A *group* is a set equipped with an associative binary operation, with a neutral element and for each element, an inverse. Examples: the group $\mathbb{Z}$ of relative integers, the regular matrices of some definite order, etc.

As these two examples show, the group operation may or may not be commutative, hence a notational schism. Commutative, or *Abelian*, groups, like $\mathbb{Z}$, are often denoted additively. But in the general case, the operation is called a product, denoted without any symbol, by simple juxtaposition, the neutral element is 1, and the inverse of g is g^{-1}.

A group G *acts* on a set X if for each $g \in G$ there is a map from X to X, that we shall denote by $\pi(g)$, such that $\pi(1)$ is the identity map, and $\pi(gh) = \pi(g) \circ \pi(h)$.[19] Observe, by taking $h = g^{-1}$, that $\pi(g)$ must be bijective, so $\pi(g)$ is a *permutation* of X. The set $\{\pi(g) : g \in G\}$ is thus a group of permutations,[20] the group law being composition of maps. Let's denote this set by $\pi(G)$.

The same abstract group can act in different ways on various related geometric objects: points, vectors, plane figures, functions, fields, tensors, etc. What counts with groups is their actions. Hence the importance of the related vocabulary, which we briefly sketch.

The action is *faithful*, or *effective*, if $\pi(g) = 1$ implies $g = 1$. (Informally, an action on X is effective if all group elements "do something" on X.) In that case, G and $\pi(G)$ are isomorphic, and $\pi(G)$ can be seen as a "concrete" realization of the "abstract" group G. This justifies writing gx, instead of $\pi(g)x$, for the image of x by $\pi(g)$. The *orbit* of x under the action of G is the set $\{gx : g \in G\}$ of transforms of x. Points x and y are in the same

[19] This is called an action *on the left*, or *left action*, as opposed to a *right action*, which would satisfy $\pi(gh) = \pi(h) \text{ o } \pi(g)$, the other possible convention. A non-Abelian group can act differently from the left and from the right, on the same set. All our group actions will be on the left.

[20] A subgroup of the "symmetric group" $S(X)$, which consists of *all* permutations on X, with composition as the group law.

orbit if there exists some group element g that transforms x into y. This is an equivalence relation, the classes of which are the orbits. If all points are thus equivalent, i.e., if there is a single orbit, one says the action is *transitive*. The *isotropy* group (or *stabilizer*, or *little group)* of x is the subgroup $G_x = \{g \in G : gx = x\}$ of elements of G that fix x. A transitive action is *regular* if there are no fixed points, that is, $G_x = 1$ for all x (where 1 denotes the trivial group, reduced to one element).

In the case of a regular action, X and G look very much alike, since they are in one-to-one correspondence. Can we go as far as saying they are identical? No, because the group has more structure than the set it acts upon. For a simple example, imagine a circle. No point is privileged on this circle, there is no mark to say “this is the starting point”. On the other hand, the group of planar rotations about a point (where there *is* a distinguished element, the identity transform) acts regularly on this circle. Indeed, the circle and this group (traditionally denoted SO_2) *can* be identified. But in order to *do* this identification, we must select a point of the circle and decide that it will be paired with the identity transform. The identification is not canonical, and there is no group structure on the circle before we have made such an identification.

The concept of *homogeneous space* subsumes these observations. It’s simply a set on which some group acts transitively and faithfully. If, moreover, the little group is trivial (regular action), the only difference between the homogeneous space X and the group G lies in the existence of a distinguished element in G, the identity. Selecting a point O in X (the origin) and then identifying gO with g —hence O in X with 1 in G —provides X with a group structure.

So when homogeneity is mentioned, ask what is supposed to be homogeneous (i.e., ask what the elements of X are) and ask about the group action. (As for isotropy and other words in tropy, it’s just a special kind of homogeneity, where the group has to do with rotations in some way.)

A.2.2 Linear spaces: V_n, A_n

I don’t want to be rude by recalling what a *vector space* (or *linear space)* is, just to stress that a vector space V is already a group (an Abelian one), with the notion of scalar[21] multiplication added, and appropriate axioms. The *span* $\vee \{v_i : i \in \mathcal{J}\}$ of a family of vectors of V is the set of all weighted sums $\sum_{i \in \mathcal{J}} \alpha^i v_i$, with scalar coefficients α^i only a *finite*

[21] Unless otherwise specified, the field of scalars is $\mathbb{R}$.

number of which are nonzero (otherwise there is nothing to give sense to the sum). This span, which is a vector space in its own right, is a *subspace* of V. A family is linearly *independent* if the equality $\sum_i \alpha^i v_i = 0$ forces all $\alpha^i = 0$. The highest number of vectors in a linearly independent family is the *dimension* of its span, if finite; otherwise we have an *infinite dimensional* subspace. The notion applies as a matter of course to the family of all vectors of V. If the dimension dim(V) of V is n, one may, by picking a basis (n independent vectors $e_1, \ldots, e_n$), write the generic vector v as $\mathbf{v}^1 e_1 + \ldots + \mathbf{v}^n e_n$, hence a one-to-one correspondence $v \leftrightarrow \{\mathbf{v}^1, \ldots, \mathbf{v}^n\}$ between v and the n-tuple of its *components*. So there is an isomorphism (non-canonical) between V and $\mathbb{R}^n$, which authorizes one to speak of *the* n-dimensional real vector space. That will be denoted V_n. Don't confuse V_n and $\mathbb{R}^n$, however, as already explained. In an attempt to maintain awareness of the difference between them, I use boldface for the components,[22] and call the n-tuple $\mathbf{v} = \{\mathbf{v}^1, \ldots, \mathbf{v}^n\}$ they form, not only a vector (which it is, as an element of the vector space $\mathbb{R}^n$), but a **vector**. Notation pertaining to $\mathbb{R}^n$ will as a rule be in boldface.

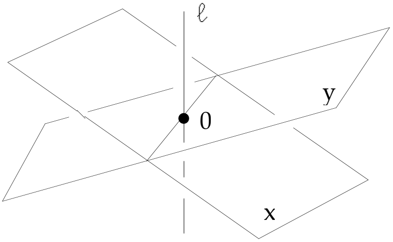
A relation $r = \{V, W, R\}$, where V and W are vector spaces is *linear* if the graph R is a vector space in its own right, that is, a subspace of the product V × W. If the graph is functional, we have a *linear map*. Linear maps $s : V \rightarrow W$ are thus characterized by $s(x + y) = s(x) + s(y)$ and $s(\lambda x) = \lambda s(x)$ for all factors. Note that dom(s) and cod(s) are subspaces of V and W.

Next, affine spaces. Intuitively, take V_n, forget about the origin, and what you have got is A_n, the n-dimensional affine space. But we are now equipped to say that more rigorously. A vector space V, considered as an additive group, acts on itself (now considered just as a set) by the mappings $\pi(v) = x \rightarrow x + v$, called *translations*. This action is transitive, because for any pair of points $\{x, y\}$, there is a vector v such that $y = x + v$, and regular, because $x + v \neq x$ if $v \neq 0$, whatever x. The structure formed by V as a set[23] equipped with this group action is called the *affine space* A *associated with* V. Each vector of V has thus become a point of A, but there is nothing special any longer with the vector 0, as a point in A.

More generally, an *affine space* A is a homogeneous space with respect to the action of some vector space V, considered as an additive group. By

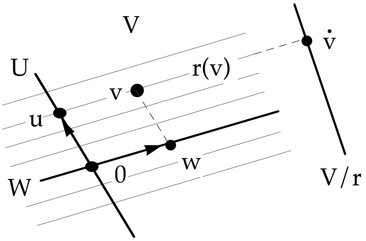
[22] At least, when such components can be interpreted as degrees of freedom, in the context of the finite element method. Our DoF-vectors are thus **vectors**. (Don't expect absolute consistency in the use of such conventions, however, as this can't be achieved.)

[23] Be well aware that V is first *stripped* of its operations, thus becoming a mere set, then *refurnished* with this group action, to eventually become something else, namely A.



selecting a point 0 in A to play origin, we can identify vector v of V with point $0 + v$ of A. But there may be no obvious choice for an origin. For example [Bu], having selected a point x and a line ℓ through x in 3D space, all planes passing through x and not containing ℓ form an affine space (inset). None of them is privileged, and the group action is not obvious.[24] For an easier example, consider a subspace W of some vector space V, and define an equivalence r by $u \; r \; v \Leftrightarrow v - u \in W$. Equivalence classes have an obvious affine structure (W acts on them regularly by $v \to v + w$) and are called *affine subspaces* of V, *parallel* to W. Of course, no point of an affine subspace qualifies more than any other as origin.

Remark A.3. The latter is not just any equivalence relation, but one which is compatible[25] with the linear structure: if $x \; r \; y$, then $\lambda x \; r \; \lambda y$, and $(x + z) \; r \; (y + z)$. This way, the quotient X/r is a vector space. Now if one wants to select a representative section, it makes sense to preserve this compatibility, by requesting this section to be a vector subspace U of V (inset, to be compared with Fig. A.9), which is said to be *complementary* with respect to W. Then each $v \in V$ can uniquely be written as $v = u + w$, with $u \in U$ and $w \in W$. Again, don't confuse the quotient V/r with the complement U, although they are isomorphic. ◊



Affine space is perhaps the most fundamental example of homogeneous space. From a philosophical standpoint, the fact that we chose to do almost all our applied physics in the framework provided by A_3 (plus, when needed, a time parameter) reflects the *observed* homogeneity of the space around us.

[24] Take a vector u parallel to ℓ, and two parallels ℓ' and ℓ'' to u, distinct from ℓ. They pierce plane x at x' and x''. The "translation" associated with $\{\lambda', \lambda''\} \in \mathbb{R}^2$ is the mapping $x \to \{$the plane determined by 0, $x' + \lambda' u$, $x'' + \lambda'' u\}$.

[25] Note the importance of this concept of compatibility between the various structures put on a same set.

What you can do on *VECTORS* may not be doable on *POINTS*. Indeed, the product λx is meaningless in an affine space: What makes sense is *barycenters*. The barycenter of points x and y with respective weights λ and $1 - \lambda$ is $x + \lambda(y - x)$. Generalizing to n points is easy. Affine independence, dimension of the affine space, and affine subspaces follow from the similar concepts as defined about the vector space. *Barycentric coordinates* could be discussed at this juncture, if this had not already been done in Chapter 3.[26]

Affine relations are characterized by affine graphs. If the graph is functional, we have an *affine map*. Affine maps on A_n are those that are linear with respect to the $(n + 1)$-**vector** of barycentric coordinates. *Affine subspaces* are the pre-images of affine maps. Affine subspaces of a vector space are of course defined as the affine subspaces of its associate. The sets of solutions of equations of the form $Lx = k$, where L is a linear map (from V_n to V_m, $m \leq n$) and k a vector, are affine subspaces, and those corresponding to the same L and different k's are parallel. The one corresponding to $k = 0$ (called *kernel* of L, denoted ker(L)) is the vector subspace parallel to them all.

If x is a point in affine space A, vectors of the form $y - x$ are called *vectors at* x. They form of course a vector space isomorphic with the associate V, called *tangent space at* x, denoted T_x. (In physics, elements of V are often called *free vectors*, as opposed to *bound vectors*, which are vectors "at" some point.) The tangent space to a curve or a surface that contains x is the subspace of T_x formed by vectors at x which are tangent to this curve or surface. Note that vector fields are maps of type *POINT* → *BOUND_VECTOR*, actually, with the restriction that the value of v at x, denoted v(x), is a vector at x. The distinction between this and a *POINT* → *FREE_VECTOR* map, which may seem pedantic when the point spans ordinary space, must obviously be maintained in the case of fields of tangent vectors to a surface.

A *convex set* in an affine space is a part C such that

$$(x \in C \text{ and } y \in C) \Leftrightarrow \lambda x + (1 - \lambda)y \in C \quad \forall \lambda \in [0, 1].$$

Affine subspaces are convex. The intersection of a family of convex sets is a convex set. The *convex hull* of a part K is the intersection of all convex sets containing K, and thus the smallest of such sets. It coincides with the union of all barycenters, with nonnegative weights, of pairs of points of K.

[26] One may—but it's a bit more awkward than the previous approach—define affine spaces ab initio, without first talking of vector spaces, by axiomatizing the properties of the *barycentric map*, which sends $\{x, y, \lambda\}$ to $\lambda x + (1 - \lambda)y$.

Let us finally discuss *orientation* of vector and affine spaces (cf. 5.2.3). This cannot be ignored, because of the prominent role played in electromagnetism by the cross product and the curl operator—both "sensitive to orientation", in a sense we shall discover later.

A *frame* in V_n is an (ordered) n-tuple of linearly *independent* vectors. Select a basis (which is thus a frame among others), and look at the determinant of the n vectors of a frame, hence a *FRAME* → *REAL* function. This function is basis-dependent, of course. But the equivalence relation defined by "$f \equiv f'$ if and only if frames f and f' have determinants of same sign" does not depend on the basis, and is thus intrinsic to the structure of V_n. There are two equivalence classes with respect to this relation.

Orienting V_n consists in designating one of them as the class of "positively oriented" frames. This amounts to defining a function, which assigns to each frame a label, like e.g., "direct" and "skew". There are two such functions, therefore two possible orientations. (Equivalently, one may define an oriented vector space as a pair {vector space, privileged basis}, provided it's well understood that this basis plays no other role than specifying the orientation.)

Subspaces of V_n also can be oriented, by the same procedure, and orientations on different subspaces are unrelated things. Affine subspaces are oriented by orienting the parallel vector subspace. For consistency, one agrees that the subspace {0} can be oriented, too, by giving it a sign, + 1 or − 1. Connected patches of affine subspaces, such as polygonal faces, or line segments (and also, after the previous sentence, points), can be oriented by orienting the supporting subspace. Lines and surfaces as a whole are oriented by conferring orientations to all their tangents or tangent planes in a consistent[27] way, if that can be done. (It cannot in the case of a Möbius band, for instance.)

There is another kind of orientation of subspaces (and hence, of lines, surfaces, etc.), called *outer* orientation. By definition, an outer orientation of a p-dimensional subspace W of V_n is an orientation of one of its complementary subspaces U, as previously defined (Remark A.3). As we saw in Chapter 5 in the case $n = 3$, this formalizes the concepts of "crossing direction" when $p = 2$, and of "way of turning around" a line when $p = 1$. *If* the ambient space is oriented, outer orientation of W determines an inner orientation: Given a frame in W, made of p vectors, one may add to them the $n - p$ vectors of a *positively* oriented frame in U, hence a

[27] I hope this makes intuitive sense. One cannot be more precise without introducing manifolds and charts, that is, starting a differential geometry course.

frame of $V_{n'}$ which falls into one of the two orientation classes, hence the orientation of the original frame.

A.2.3 Metric spaces

A *metric space* $\{X, d\}$ is a set X equipped with a *distance*, that is, a function $d : X \times X \to \mathbb{R}$ such that $d(x, y) = d(y, x) \geq 0 \ \forall\, x, y \in X$, with $d(x, y) > 0$ if $x \neq y$, and $d(x, z) + d(z, y) \geq d(x, y) \ \forall\, x, y, z$.

Metric-related notions we may have to use are *open ball* $B(x, r) = \{y \in \mathbb{R} : d(x, y) < r\}$, *open set* (a part A which if it contains x also contains an open ball centered at x), *closed* set (one the complement of which is open), *distance* $d(x, A)$ of a point x *to a part* A (which is $\inf\{d(x, y) : y \in A\}$), *adherence* or *closure* $\overline{A}$ of a part A (all points x of X such that $d(x, A) = 0$), *interior* of A (points such that $d(x, X - A) > 0$, set denoted $\mathrm{int}(A)$ when we need it), *boundary* ∂A of A (points for which $d(x, A) = 0$ and $d(x, X - A) = 0$). A sequence $\{x_n \in X : n \in \mathbb{N}\}$ *converges* if $\lim_{n \to \infty} d(x_{n'} x) = 0$ for some $x \in X$, called the *limit*, which one immediately sees must then be unique. By taking all the limits of all sequences whose elements belong to a part A, one obtains its closure.[28] A part A is *dense* in B if $\overline{A}$ contains B, which means (this is what counts in practice) that for any $b \in B$ and any $\varepsilon > 0$, there is some $a \in A$ such that $d(a, b) < \varepsilon$. This is strictly the same as saying that one can form a sequence $\{a_n \in A : n \in \mathbb{N}\}$ that converges towards b. A metric space X is *separable* if it contains a denumerable dense part.

Weaker than the notion of limit is that of *accumulation point*: Let us say (this is not part of the received terminology) that a family "clusters" at x if one can extract from it a sequence that converges to x (then called an accumulation point for this family). Convergent sequences cluster at their limit (and at no other point). Some sequences may not cluster at all. *Compact* parts of a metric space are closed parts in which any sequence must cluster at some point.

These notions are useful in approximation theory (Chapter 4). For instance, the union of all approximation spaces, for all imaginable meshes of a given domain, form a dense set in the set of all eligible fields, for the energy distance. Moreover, the family of approximate solutions, indexed over these meshes, clusters at the right solution. But knowing that is not enough. What one wants, which is more difficult, is to devise "refinement rules" which, starting from any mesh, generate a sequence of finer meshes

[28] Beware: This equivalence, like some others this list suggests, may not hold in topological spaces whose topology cannot be described by a distance.

with the property of convergence of the corresponding approximate solutions. (Incidentally, the functional space must be separable for this to be possible, since elements of the form $\sum_{i \in \mathcal{J}} \varphi_i \lambda^i$, where the coefficients φ_i take rational values only, and $\mathcal{J}$ is the union of the sequence of Galerkin bases, form a dense denumerable set. Most usual functional spaces are separable, as a corollary of the Weierstrass theorem[29] on polynomial approximation of continuous functions.)

A function f from {X, d} to {X', d'} is *continuous* if it maps converging sequences of X to converging sequences of X'. This is equivalent to saying that the pre-image of a closed set is closed. The function f is *uniformly* continuous if for each $\varepsilon > 0$, there exists $\delta(\varepsilon)$ such that, for any pair {x, y} of points taken in dom(f), $d(x, y) < \delta(\varepsilon)$ implies $d'(f(x), f(y)) < \varepsilon$. Obviously (this is a standard exercise in manipulating quantifiers), uniform continuity is logically stronger than simple continuity, but the two notions coincide when f is affine. An *isometry* from {X, d} to {X', d'} is a function f such that $d'(f(x), f(y)) = d(x, y)$ for all {x, y} in dom(f). This implies one-to-oneness, and uniform continuity of f and of its reciprocal, and therefore, *homeomorphism* (existence of a one-to-one map continuous in both directions), but is stronger.[30]

Remark A.4. You may be excused for guessing that continuous functions are functional relations with a closed graph, for this seems so natural. But it's wrong . . . For instance, the "weak gradient" of Chapter 5 has a closed graph in $L^2(D) \times \mathbb{L}^2(D)$, but is not continuous. (Relations between continuity of functions and closedness of their graphs are governed by the "Banach theorem", a deep result which belongs to the hard core of functional analysis, but is not used here. See [Br].) ◊

If a sequence clusters, its image by a continuous function clusters, too, so the continuous image of a compact part is compact. In particular, a real-valued continuous function whose domain is compact reaches its minimum and its maximum, since its image is closed.

Remark A.5. This obvious result is important in proving existence of equilibrium configurations in many physical situations. Suppose the set of states S can be described as a normed space (see below), the norm of a state being precisely its energy. (This is what we do in Chapters 2, 3, and 4.) States of bounded energy that satisfy specific constraints (of the kind $f(x) = 0$, where f is a continuous function) then form a closed bounded set.

[29] Given a nonempty compact subset K of $\mathbb{R}^n$, a continuous function $f : K \to \mathbb{R}$, and $\varepsilon > 0$, there exists a polynomial $p : \mathbb{R}^n \to \mathbb{R}$ such that $|f(x) - p(x)| < \varepsilon$ for all $x \in K$.

[30] So strong actually, that it implies much more. For instance, a theorem by Mazur and Ulam asserts that a surjective isometry between real Banach spaces is affine [MU].

Such a set is compact if S is of finite dimension, because then any bounded sequence must cluster somewhere (as shown by the well-known Bolzano–Weierstrass proof). But such a sequence need not cluster in infinite dimension, since there are an infinity of directions in which to go. This is why existence proofs for variational problems in infinite-dimensional functional spaces always require some structural element to supply, or replace, the missing compactness. Quite often (as will be the case with the projection theorem proved a little later, and in 3.2.1) *convexity,* associated with completeness (see A.4.1 below) is this element. ◊

The *support* of a real- or vector-valued function f on a metric space X is the closure of the set $\{x \in X : f(x) \neq 0\}$. Don't confuse support and domain.

A useful density result, often invoked is this book, can informally be expressed as follows: *Smooth* fields over E_3 form a dense set among fields, in the energy norm. To be more precise, let h be a vector field such that $\int_{E_3} |h(x)|^2 \, dx < \infty$, and denote by $||h||$ the square root of this energy-like integral. Take a *mollifier* ρ, that is, a real-valued C^∞ function on E_3, nonnegative, with bounded support, and such that $\int \rho = 1$. Define the sequence h_n by[31]

$$(*) \qquad h_n(x) = \int_{E_3} \rho(y)\, h(x - y/n)\, dy,$$

or in more compact notation, $h_n = \rho_n * h$, where $*$ denotes the *convolution product*, and $\rho_n = x \to n^3 \rho(nx)$. The outcome of such a product is as smooth as the smoothest factor ("regularizing property" of convolution), so h_n is $\mathbb{C}^\infty$. (This result, if not its proof, is intuitive: If ρ is smooth, one can differentiate under the summation sign in $(*)$, indefinitely.) Now, evaluate the quadratic norm of $h - h_n$: A tricky computation, which makes use of Fubini's theorem (see, e.g., [Pr]), and thus roots deeply in Lebesgue integration theory, will show that this norm tends to 0. Hence the density. The result is easily extended to fields over D by restriction.

We made repeated use of this when invoking the following argument, either in this form or in a closely related one: Suppose $f \in L^2(D)$ and $\int f \varphi = 0$ for all $\varphi \in C^\infty(D)$. By density, there is a sequence φ_n in $C^\infty(D)$ which converges towards f. Each term of the sequence $\int f \varphi_n$ is zero, and its limit is $\int |f|^2$ by continuity of the scalar product, hence $f = 0$ a.e. See for instance the proof of Prop. 2.1, p. 43.

[31] This way, $h_n(x)$ is a weighted average of values of h at points close to x. One often assumes a nice shape for the graph of ρ (centered at the origin, invariant by rotation, etc.), but this is not required, as far as theory is concerned.

A *path* in a metric space is a continuous mapping $c : [0, 1] \to X$. A *circuit*,[32] or *loop*, is a path that closes on itself ($c(0) = c(1)$). Let's call *patch* a continuous mapping $C : [0, 1] \times [0, 1] \to X$. A part of a metric space is *connected* if two of its points, x and y, can always be joined by a path c, that is, with $c(0) = x$ and $c(1) = y$. A connected open set is called a *domain*. We know the word in a different sense already, but this dual use should not be too confusing, for the "domain of definition" of a function or a field is often a domain in the present topological sense, and the context always makes clear what one means.

Two circuits c_0 and c_1 are *homotopic* if there is a patch C such that $c_0 = s \to C(s, 0)$ and $c_1 = s \to C(s, 1)$, with $C(0, t) = C(1, t)$ for all t in $[0, 1]$. This means one can be continuously deformed into the other, all intermediate steps being loops.[33] A metric space X is *simply connected* if any circuit is continuously reducible to a point, that is, homotopic to a circuit of the form $c(t) = x \ \forall\, t \in [0, 1]$, where x is a point of X.

A.2.4 Normed spaces, Euclidean norms

Being metric and being a vector or affine space are two different things, but if a set bears both structures, they had better be compatible. Suppose X is an affine space, with associated vector space V, and a distance d. The two structures are *compatible* if $d(x + v, y + v) = d(x, y)$, for all points x, y and all translation vectors $v \in V$. Then, once selected an origin 0, the real-valued function on V defined by $\|v\| = d(0, 0 + v)$ has the following properties, which characterize, by definition, a *norm:* $\|v\| > 0$ unless $v = 0$, $\|\lambda v\| = |\lambda|\,\|v\|$ for all v in V and real λ, and $\|v + w\| \le \|v\| + \|w\|$. If, conversely, a vector space has a norm $\| \ \|$, the distance this induces on the associated affine space, $d(x, y) = \|x - y\|$, is compatible with the affine structure. A *normed space* is, in principle, a vector space V equipped with a norm, but you will often realize, thanks to the context, that what is really implied is the associated affine metric space.

Norms often stem from a *scalar product,* that is, a real-valued mapping, denoted $(\ , \ )$, from $V \times V$ to $\mathbb{R}$, linear with respect to both arguments, symmetrical (i.e., $(v, w) = (w, v)$ for all $\{v, w\}$), and overall, *positive*

[32] For some, "circuit" implies more smoothness than mere continuity of c.

[33] More generally, two *maps* g_0 and g_1 from Y into X are *homotopic* if there is a continuous map f from $Y \times [0, 1]$ into X such that $f(s, 0) = g_0(s)$ and $f(s, 1) = g_1(s)$. One of the rare merits of some recent science-fiction movies has been to popularize this notion, since it happens to be the formalization of the concept of "morphing". The case of loops is when Y is a circle.

definite, that is

$$(v, v) > 0 \Leftrightarrow v \neq 0.$$

The norm of a vector v is then defined as $\|v\| = [(v, v)]^{1/2}$. Thereby, notions such as *orthogonality* (v and w are orthogonal if $(v, w) = 0$, which one may denote $v \perp w$) and *angle* (the angle of two nonzero vectors v and w is $\arccos((v, w)/\|v\| \, \|w\|)$), make sense. A vector space with scalar product is a "pre-Hilbertian space", a structure we shall study in its own right later.

The most familiar example is when $V = V_n$. Then we rather denote the norm $\|v\|$ by $|v|$, the scalar product (v, w) by $v \cdot w$, call them the *modulus* and the *dot product* respectively, and say that they confer a *Euclidean structure* (or a *metric*) on V_n and its affine associate A_n, via the *Euclidean distance* $d(x, y) = |x - y|$. *Euclidean geometry* is the study of the affine metric space $\{A_n, d\}$, called n-dimensional *Euclidean space*. (Why the singular in "space" will be discussed below.)

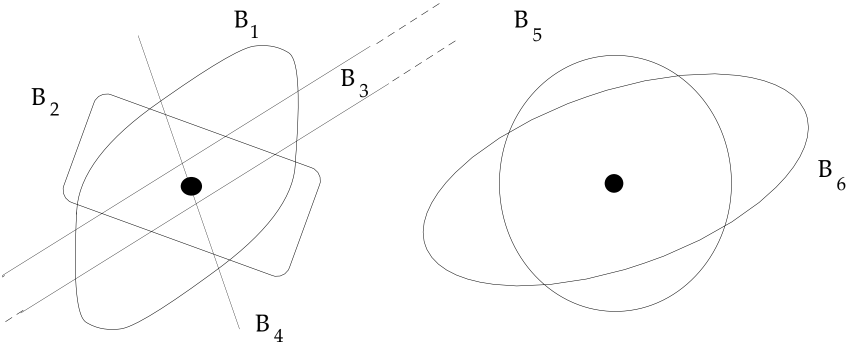


FIGURE A.13. Convex sets B_1, B_2, B_5, and B_6 are barrels, but B_3 (not bounded) and B_4 (not absorbing) are not. Observe the various degrees of symmetry of each barrel. Only those on the right generate *Euclidean* norms.

Other norms than Euclidean ones can be put on V_n, as suggested by Fig. A.13. All it takes is what is aptly called a *barrel*, that is a bounded, closed, absorbing, and balanced convex set B: *Balanced* means that $-v \in B$ if $v \in B$, *absorbing* that for every v, there exists $\lambda > 0$ such that λv belongs to B, *closed*, that the real interval $I(v) = \{\lambda : \lambda v \in B\}$ is closed for all v, and *bounded*, that $I(v)$ is bounded for all $v \neq 0$. One then sets $|v|_B = 1/\sup\{\lambda : \lambda \in I(v)\}$, and this function (the inverse of which is called the *gauge* associated with B) is easily seen to be a norm. (Check that $|v|_B = \inf\{\lambda : v \in \lambda B\}$.) The closed unit ball for this norm is then the barrel B itself. (Notice how the two different notions of closedness we had up to now are thus reunited.)

Barrels that generate Euclidean norms (that is, ellipsoids) are obviously "more symmetrical" than others (Fig. A.13). A bit of group theory confirms this intuition. Let's denote by GL_n, and call *linear group*, the group of all bijective linear maps on V_n. (It's isomorphic, via a choice of basis, with the group of $n \times n$ regular matrices, but should not be confused with it.) Let us set $G(B) = \{g \in GL_n : \ v \in B \Leftrightarrow gv \in B\}$, the subgroup of linear transforms that leave B globally invariant, that is to say, the little group of B relative to the action of GL_n on subsets of V_n. We see immediately, on the examples of Fig. A.13, where $n = 2$, what "more symmetrical" means: $G(B_1)$ has only two elements (the identity and the reflection $v \rightarrow -v$ with respect to the origin), $G(B_2)$ has four, whereas $G(B_5)$ and $G(B_6)$ have an infinity (both groups are isomorphic with SO_2, actually). It can be shown (but this is beyond our reach here) that Euclidean barrels are those with *maximal*[34] isotropy groups, and thus indeed, the most symmetrical barrels that can exist.

This symmetry is what is so special about Euclidean norms. A bit earlier, we remarked that physical homogeneity of the space around us was reflected in the choice of affine space as framework for most our modellings. One may add to this that *isotropy* of physical space is reflected in the use of Euclidean norms, hence their prominent role. Indeed, a Euclidean norm *privileges no direction*:[35] If v and w both belong to the surface of a Euclidean barrel B, there is a linear transform $g \in G(B)$ such that $w = gv$. In other words, the action of $G(B)$ on the surface of B, that is to say, its action on *directions*, is transitive. This is not true of other barrels.

Alternatively (as one easily sees, the two properties are equivalent), one may say that GL_n acts transitively on the set of all Euclidean barrels. In other words, given two scalar products " $\cdot$ " and " $\circ$ ", there is a linear

[34] More precisely: Define the *unimodular* group SL_n as the subgroup of linear maps that are represented, in some basis, by matrices of determinant 1. (This definition is in fact basis independent.) Then maximal proper subgroups of SL_n, all isomorphic with the group of orthogonal $n \times n$ matrices, are the isotropy groups of Euclidean barrels.

[35] *In spite* of what Fig. 13, right part, seems to suggest: One may feel like objecting "what about the principal directions of the ellipsoids B_5 or B_6 ?" But there is nothing special with these directions. They are spurious, due to the fact that we commit ourselves to a specific basis *just to do the drawing*. They will disappear if one selects the eigenvectors as basis (then B becomes a disk). Contrast this with the two "axes" of B_2, which no change of basis can erase. Denizens of a flat world with a metric governed by the barrel B_2 would be able to recognize the existence of privileged directions in their universe. Ask any New Yorker.

invertible map L such that $v \circ w = Lv \cdot Lw$. Which is why one can speak of *the* Euclidean geometry, *the* Euclidean norm, in the face of apparent evidence about their multiplicity. For after all, each choice of basis in V_n generates a particular dot product, to wit $v \cdot w = \sum_i \mathbf{v}^i \mathbf{w}^i$, where **v** and **w** are the component **vector**s, so there seems to be as many Euclidean geometries as possible bases. Why then *the* Euclidean space? Because all Euclidean structures on V_n are equivalent, up to linear transformation. We shall see this in more concrete terms in the next Section.

A.3 OUR FRAMEWORK FOR ELECTROMAGNETISM: E_3

In this book, we work in 3D affine space, and it is assumed all along that a specific choice of dot product *and* orientation has been made once and for all. Thus, what is called E_3 in the main text is always *oriented Euclidean three-dimensional space.*

Note that the all-important notion of *cross product* would not make sense without orientation: By definition, $u \times v$ is orthogonal to both u and v and its modulus is $|u|\,|v| \sin\theta$, where $\theta = \arccos(u \cdot v/|u|\,|v|)$, but all this specifies $u \times v$ only up to sign, hence the rule that u, v, and $u \times v$, in this order, should make a "positively oriented" frame (cf. p. 287). This assumes that one of the two classes of frames has been designated as the "positive", or "direct" one.

A.3.1 Grad, rot, and div

This subsection discusses the classical differential operators in relation with these structures.

We pointed out the essential uniqueness of Euclidean space, all Euclidean structures being equivalent via linear transformations. This is so ingrained in us that we forget about the multiplicity of Euclidean metrics, and it may be appropriate to tip the scales the other way for a while.

Consider two different Euclidean structures on A_3, as provided by two different dot products " $\cdot$ " and " $\circ$ ", and denote them E_3 and $\mathbb{E}_3$ respectively. (No orientation yet.) Let $\varphi : A_3 \to \mathbb{R}$, a smooth scalar field, be given. Its existence owes nothing to the Euclidean structure, obviously. But what of its gradient? If we define grad φ as the vector field such that, for all v, $(\text{grad}\,\varphi)(x) \cdot v = \lim_{\lambda \to 0} (\varphi(x + \lambda v) - \varphi(x))/\lambda$ —and who would object to that?[36]—then grad φ *does* depend on the Euclidean structure, and we have two different gradients: one, grad, for E_3, and another one, $\mathbb{grad}$ say, for $\mathbb{E}_3$. Coming back to A_3, where the notions of scalar field and vector field do make intrinsic sense, without any need for a metric, we have obtained two differential operators of type *SCALAR_FIELD* → *VECTOR_FIELD*, respectively grad and $\mathbb{grad}$, which are *different*. Of course—and this is where the equivalence of Euclidean structures lets itself be felt—they are closely related: As a consequence of $(\mathbb{grad}\,\varphi)(x) \circ v = (\text{grad}\ \varphi)(x) \cdot v$, one has $L^tL(\mathbb{grad}\,\varphi)(x) = (\text{grad}\ \varphi)(x)$, at all points, hence $L^tL\,\mathbb{grad} = \text{grad}$, which corresponds to a change of basis.

Such equivalence suggests that all these different gradients are mere avatars of a *single*, intrinsically defined, operator that would make sense on A_3. Indeed, this operator exists: It's the "exterior derivative" of differential geometry, denoted $\mathtt{d}$; but to develop the point would lead us into this "radical way" alluded to in 2.2.3, but not taken.

The situation with the curl operator is even worse, because not only the Euclidean structure, but the orientation of space plays a role in its definition.

Given u, we may—by using the Stokes theorem backwards—define rot u as the vector field such that its flux through a surface equals the circulation of u along this surface's boundary. Both words "through" and "along" refer to orientation, but the former connotes outer orientation of the surface, and the latter, inner orientation of the rim. Since both orientations can be defined independently, defining rot requires they be related in some arbitrary but definite way. When the ambient space is oriented, it becomes possible to establish such a relation (by the corkscrew rule), as we saw in 5.2.1 and p. 287. So it's only in *oriented* three-dimensional Euclidean space that rot makes sense. Another way to express this is to say that for a given dot product, there are *two* curl operators in three-space, one for each possible orientation, which deliver opposite fields when fed with one.

So if we insist on imitating the previous development on E_3, $\mathbb{E}_3$, grad and $\mathbb{grad}$, we must distinguish ${}^{+}E_3$ and ${}^{-}E_3$, say, to account for the two possible orientations, as well as ${}^{+}\mathbb{E}_3$ and ${}^{-}\mathbb{E}_3$, hence four operators ${}^{-}\text{rot}$, ${}^{+}\text{rot}$, ${}^{-}\mathbb{rot}$, ${}^{+}\mathbb{rot}$, of type *VECTOR_FIELD* → *VECTOR_FIELD*, defined in A_3. Again, if the new metric $\circ$ is given by $u \circ v = Lu \cdot Lv$, and if L preserves orientation (which one can always assume), one has[37] ${}^{+}\mathbb{rot} =$

[36] Well . . . One often sees (grad f)(x) defined as the vector of coordinates $\{\partial_1 f, \partial_2 f, \partial_3 f\}$ at point x. This is a different notion: The entity thus defined is a *covector*, that is to say, an element of the dual of V_3, not a vector. The $\partial_i f$'s are what is called "covariant components" of grad f. Only in the case of an orthonormal frame do they coincide with its ordinary ("contravariant") components.

[37] The new cross-product $\mathbb{\times}$ is then given by $L(u \mathbb{\times} v) = Lu \times Lv$. The reader is challenged to prove the formulas of Fig. A.14 by proper application of the Stokes theorem.

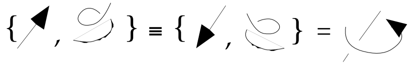
$\det(L^{-1})\ {}^{+}\mathrm{rot}\, L^{t}Lu$, and this plus the obvious relations ${}^{+}\mathrm{rot}\, u = -\,{}^{-}\mathrm{rot}\, u$ and ${}^{+}\mathbb{rot}\, u = -\,{}^{-}\mathbb{rot}\, u$ makes such changes of metric and orientation manageable (cf. Fig. A.14), but the lesson is clear: *Classical differential operators are definitely impractical*[38] when it comes to such changes. Better stay with the same metric and the same orientation all over. Problems where this is too cumbersome, such as computations involving moving (and possibly, deformable) conductors, call for the more elaborate framework provided by differential geometry, as discussed in 2.2.3.

$$\begin{tikzcd}
\bullet \arrow[r, "\mathrm{grad}"] & \bullet \arrow[r, "\mathrm{rot}"] & \bullet \arrow[r, "\mathrm{div}"] & \bullet \\
\bullet \arrow[u, "1"] \arrow[r, "\mathbb{grad}"] & \bullet \arrow[u, "L^{t}L"] \arrow[r, "\mathbb{rot}"] & \bullet \arrow[u, "\det(L)"] \arrow[r, "\mathbb{div}"] & \bullet \arrow[u, "\det(L)"]
\end{tikzcd}$$

FIGURE A.14. Relations between the differential operators associated with two different dot products. This is what is called a *commutative diagram*: Each arrow is marked with an operator, and by composing operators along a string of arrows that joins two dots, one obtains something which depends only on the extreme points, not on the path followed. Note that $\mathbb{div}\, v = \mathrm{div}\, v$.

A.3.2 Digression: the so-called "axial vectors"

As if this was not complicated enough, someone invented the following devilish device. Let's start from V_3 with a metric, but no orientation. Using our freedom to create new geometric objects thanks to the mechanism of equivalence relations and classes, let's introduce pairs {v, Or}, where v is a vector, and Or one of the two classes of frames, decree that {v, Or} and {– v, – Or}, where – Or is of course the other class, are equivalent, and call the equivalence class an *axial vector*. To soothe the vexed ordinary vector, call it a *polar* vector. (As suggested in inset, the right icon for an axial vector is not the arrow, but a segment with a sense of rotation around it. Note how, just as a polar vector orients its supporting line, an axial vector "outer orients" this line. Note also that axial and polar vectors can be associated in one-to-one correspondence, but in two different ways, one for each orientation of



[38] Some solace can be found in the invariance of the divergence with respect to changes of metric: div v = $\mathbb{div}$ v. If v is interpreted as the velocity field of a fluid mass, its divergence is the rate of change of the volume occupied by this mass, and though the volume depends on the metric, volume *ratios* do not.

ambient space.) Now define a new operator $\overrightarrow{\mathrm{rot}}$ as follows. Start from a field $\overset{\curvearrowright}{\mathrm{v}}$ of axial vectors. Select a representative {v, Or}, and define $\overrightarrow{\mathrm{rot}}\,\overset{\curvearrowright}{\mathrm{v}}$ as ${}^{\mathrm{Or}}\mathrm{rot}\ \mathrm{v}$, where ${}^{\mathrm{Or}}\mathrm{rot}$ is the operator associated with *this* choice of orientation. This is a consistent definition, because ${}^{-\mathrm{Or}}\mathrm{rot}(-\mathrm{v}) = {}^{\mathrm{Or}}\mathrm{rot}(\mathrm{v})$, and thus $\overrightarrow{\mathrm{rot}}\,\overset{\curvearrowright}{\mathrm{v}}$ is a well-defined *polar* (yes, polar!) vector field. Now, lo and behold, the new operator $\overrightarrow{\mathrm{rot}}$ does not depend on orientation.[39]

To make use of it, one must then confer the axial status to some of the vector fields in Maxwell equations. The electric field, akin to a force, has nothing to do with orientation and is thus polar. Then b *must* be axial, and also h, because of $\mathrm{b} = \mu\mathrm{h}$, and of course d and j are polar. Excessive emphasis on such notions, sometimes combined with obscure considerations of the "axial character" of some physical entities, on "the way vectors behave under mirror reflection", and so on, generates much undue confusion. The tiny advantage of not depending on orientation ($\overrightarrow{\mathrm{rot}}$ continues to depend on the metric, anyway), is thus dearly paid for.

The key to clarity is to stay aware of the distinction between physical entities and their mathematical representations. A vector field is a vector field is a vector field . . . But it often happens to be just an element, the main one but not the only one, in the description of a physical entity, to which other elements, standing in background, also contribute.

For instance, the electric field, as a physical object, can be represented by three mathematical objects, acting in conjunction: affine space, a dot product, and (the main item) a vector field denoted e. The magnetic field, still as a physical object, demands a little more: space, dot product, *an orientation,* and (the main item, again) a vector field b. Among these four elements, the first three can be fixed once and for all, thus forming a background, or "mathematical framework", here symbolized by E_3, which can be used for all electromagnetic entities. Hence the expression of a physical law such as, for instance, Faraday's, as a differential relation between vector fields, namely $\partial_t \mathrm{b} + \mathrm{rot}\ \mathrm{e} = 0$.

However, there is some leeway in the choice of items that will be kept in background. As the concept of axial vector suggests, one may decide *not* to include orientation among them, and have the actors on the stage (now axial vectors and polar vectors, depending) carry this information with them all the time. Hence such orientation-free but also, terribly contrived, formulations as $-\,\partial_t \vec{\mathrm{d}} + \overrightarrow{\mathrm{rot}}\,\overset{\curvearrowright}{\mathrm{h}} = \vec{\mathrm{j}}$, and symmetrically, $\partial_t \overset{\curvearrowright}{\mathrm{b}} + \overset{\curvearrowright}{\mathrm{rot}}\,\vec{\mathrm{e}} = 0$, where $\overset{\curvearrowright}{\mathrm{rot}}$ is the operator of Note 39.

[39] A similar operator, $\overset{\curvearrowright}{\mathrm{rot}}$, also orientation-independent, will act on a polar vector to give an axial one: Just define $\overset{\curvearrowright}{\mathrm{rot}}\,\vec{\mathrm{v}}$ as the class of the pair $\{{}^{\mathrm{Or}}\mathrm{rot}\ \mathrm{v}, \mathrm{Or}\}$.

But then, why not also bring *metric,* which is at least as versatile as orientation, to the foreground? This is possible by treating b and e as differential forms. Then Faraday's law takes the form $\partial_t b + \mathrm{d}e = 0$, where d is the exterior derivative, which is metric- and orientation-independent. Axial vectors thus appear as an awkward device, which leaves us with a job less than half-done, at the price of considerable conceptual complexity.

A.3.3 The Poincaré lemma

Curl-free fields are gradients, locally, and divergence-free fields are curls. The Poincaré lemma is the precise statement of this well-known and important property.

A domain D of E_3 is *star-shaped* if it contains a privileged point x_0 such that if $x \in D$, then $x_0 + \lambda(x - x_0)$ belongs to D for all $\lambda \in [0, 1]$. One may always select x_0 as origin, which we do in what follows.

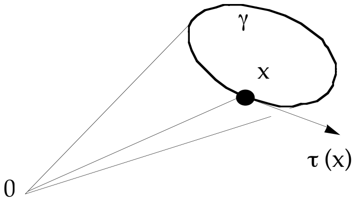
Poincaré's lemma. *Let* e, b, *and* q *be two vector fields and a function, smooth over a star-shaped domain* D, *such that* $\mathrm{rot}\, e = 0$ *and* $\mathrm{div}\, b = 0$ *in* D. *There exists a smooth function* ψ *and smooth fields* a *and* j *such that* $e = \mathrm{grad}\, \psi$, $b = \mathrm{rot}\, a$, *and* $q = \mathrm{div}\, j$, *in all* D.

There are explicit formulas for ψ, a, and j, as follows:

(p1) $\psi(x) = \int_0^1 x \cdot e(\lambda x)\, d\lambda,$

(p2) $a(x) = -\int_0^1 x \times b(\lambda x)\, \lambda\, d\lambda,$

(p3) $j(x) = \int_0^1 x\, q(\lambda x)\, \lambda^2\, d\lambda,$



where x is a fixed point of D and λ the integration variable, and the proof is a verification—not that straightforward. For (p2), for instance, take the circulation of a along a small loop γ (inset), compare the result with the flux of b across the surface of the cone centered at 0 generated by γ, and apply Stokes to the sole of the cone.

Note that an open ball is star-shaped, so the lemma is always valid *locally,* in the neighborhood of a point. What is at stake here is the *global* result ("in *all* D"). It holds in all dimensions, and studying the proof (as given in [BS], after pp. 94–95 of [Sp], or [Co], or [Sc], p. 140) reveals what is important in the hypothesis: not D being star-shaped in the strict sense, but the existence of a *deformation–retract,* that is, a family g_t of maps from D into itself, continuous with respect to t and x, which satisfies $g_1(x) = x$ and $g_0(x) = x_0$ for all $x \in D$. (In the language of Note

33, this is a homotopy between the identity map $x \to x$ and the constant map $x \to x_0$.) A metric space is *contractible* if it can be, so to speak, collapsed onto one of its points by such a deformation-retract (here, $g_t(x) = tx$). The Poincaré lemma is thus valid for contractible domains of E_3, actually, even if simple formulas like (p1)–(p3) may not be available.

All simply connected sets of E_2 are contractible. In E_3, this condition is implied by contractibility, but the latter is stronger. One can prove that bounded simply connected regular domains with a connected boundary are contractible. (This is the criterion we use in Chapter 5.) For simply connected regions with non-connected boundary, it is still true that curl-free fields are gradients, although solenoidal fields may not be curls.

Note that, contrary to what is often lightly asserted, domains where all irrotational fields are global gradients need *not* be simply connected. Figure 8.8, Chapter 8, offers a counter-example.

Formula (p2) is important in electromagnetism, where it is called "Poincaré gauge" [BS]. A "gauge", as we saw on p. 274, is a rule by which, given a solenoidal b, one can select a particular representative in the class of vector potentials a that satisfy $\operatorname{rot} a = b$ —if there is one, which Poincaré lemma shows to be the case in contractible regions. As pointed out in [Sk], the gauge implied by (p2) is the obvious condition $x \cdot a(x) = 0$, which does not coincide with either the Coulomb or the Lorenz gauge. In particular, note that $\operatorname{div} a \neq 0$ in (p2). Poincaré gauge might have useful applications in some modellings [Ma], and should be better known.

The central importance of Poincaré's lemma, however, lies elsewhere: the fact that, for a contractible domain of E_3, the sequence

$$C^\infty(\overline{D}) \xrightarrow{\text{grad}} \mathbb{C}^\infty(\overline{D}) \xrightarrow{\text{rot}} \mathbb{C}^\infty(\overline{D}) \xrightarrow{\text{div}} C^\infty(\overline{D})$$

is *exact*, in the sense of Chapter 5 (the codomain of each operator fills out the kernel of the next operator in the sequence). Moreover, when the sequence is *not* exact, i.e., when either of the quotients

$$\ker(\text{rot}\,;\ \mathbb{C}^\infty(\overline{D}))/\text{grad}(C^\infty(\overline{D})), \qquad \ker(\text{div}\,;\ \mathbb{C}^\infty(\overline{D}))/\text{rot}(\mathbb{C}^\infty(\overline{D}))$$

has nonzero dimension, some topological peculiarities of D (presence of "loops" and "holes", respectively, as explained in Chapter 5) can be inferred.

A.3.4 Symmetry in E_3

In most modellings, there is some geometrical symmetry of the domain of interest, that can be exploited to reduce the size of the computational domain, hence substantial economies. The idea is to perform the computation on a subdomain, called the *symmetry cell*, containing only one point per orbit under the action of the symmetry group. Thus, the union of images of the closure of the cell is identical with the closure of the original computational domain. But this supposes a proper setting of boundary conditions on "new boundaries" thus introduced (on symmetry planes, for instance), and for this, the formal notions about symmetry that follow may be helpful.

The *isometries* of a metric space X are the transformations (functions of type X → X, defined over all X) that preserve distances. (This implies bijectivity.) Isometries of E_3 are the rotations, the translations, the mirror symmetries, and their compositions. We'll say an isometry is *skew* or *direct*, according to whether it changes the orientation of a reference frame or not. (Alternatively, one could say *odd* or *even*, but we reserve these words for a different use.)

Let D be a regular bounded domain in E_3.

Definition A.1. *An isometry*[40] i *of* E_3 *is a* symmetry *of domain* D *if it leaves* D *globally unchanged:* i(D) = D.

Symmetries of D form of course a group (denoted G_D or simply G in what follows).[41] This group has two elements in the case that first comes to mind when symmetry is mentioned, which is bilateral symmetry: the identity and the mirror symmetry h with respect to a plane Σ (group denoted C_{1h}). But there may be much more: for instance, all the $2\pi/n$ rotations around some straight line a (called a "repetition axis of order n"), group denoted C_n. Other frequently encountered symmetry groups are D_n, C_{nh}, C_{nv}, obtained by combining the rotations in C_n with, respectively, the half-turn around an axis orthogonal to a, the reflection h with respect to a plane orthogonal to a, and the reflection v with respect to a plane containing a, and D_{nh}, which is obtained by composing the rotations of D_n with h. For concrete examples, think of a three-blade propeller (group D_3 or C_3, depending on whether the propeller's action is reversible or not), a triumph arch (C_{2v}), the Eiffel tower (C_{4v}), a brick (D_{2h}).

[40] Some interesting symmetries are not isometries. One may conceive of objects with "fractal" structure, invariant with respect to some non-distance-preserving transformations, dilatations for instance. The exploitation of symmetries of this kind is an open problem.

[41] The little group of D under the action of isometries on parts of E_3.

A symmetry of D is direct or skew according to whether the isometry it comes from is itself direct or skew. Elements of G_D which are direct symmetries form a subgroup of G_D.

Let i be an isometry of E_3. If v is a vector at x which has its tip at y (which is another way to say that y = x + v), it is natural to define the transform of v under i as the vector at ix which has its tip at iy, that is, iy – ix. We shall denote this vector by i_*v.

By restriction to D, one may similarly define the effect of a symmetry s of D on a vector at x, for x in D or its boundary. If now $v = x \to v(x)$ is a vector *field* over D, we'll denote by Sv the transform of v under the action of s, thus defined:

$$(1) \qquad (Sv)(sx) = s_*(v(x)),$$

that is to say, $Sv = x \to s_*(v(s^{-1}x))$. Thus if s is, for instance, the mirror reflection in a plane, and if v is represented, according to a popular (and quite unfortunate) graphic convention, by a bundle of arrows, Sv is imaged by the set of reflections of these arrows. Functions transform under a symmetry the same way vector fields do: If φ is a function defined over D, we may set $(S\varphi)(sx) = \varphi(x)$, on the model of (1), that is $S\varphi = x \to \varphi(s^{-1}x)$ (the "push-forward" of φ by s, cf. Note 7.10). All this suggests the following definition:

Definition A.2. *A symmetry* s *of* D *is a symmetry* of the vector field v *[resp.* of the function φ] *if and only if* Sv = v *[resp.* Sφ = φ].

Note how this provides a concrete example of a family of group actions, all different, of the same group, here G_D, on different geometrical objects. General notions as given earlier apply. In particular, the symmetries of a vector field or a function form a subgroup of G_D, denoted G_v or G_φ if a name is needed, called as we know the isotropy group (or little group) of v or φ. By the Stokes theorem, the little group of a function φ [resp. of a field h, a field j] can be embedded[42] in the little group of grad φ [resp. of rot h, of div j].

When we refer to the symmetries of a *problem*, it means more than the symmetries of D. Symmetries of the material properties also should be considered. This is, in all generality, a difficult subject, if one wishes to take into account the deformability of materials, and possible anisotropies. For homogeneous materials and non-changing geometries, however, it's simple. All we have to do is consider the symmetry groups

[42] It's not simply an isomorphism, because rot v, for instance, may be much more symmetrical than v. (Think of some undistinguished v for which rot v = 0.)

of the functions σ, μ, and ε, and take their intersection with G_D. The subgroup of G_D thus obtained is the symmetry group of the problem.

Many symmetries are *involutions*, in the sense that $s^2 = 1$ (the identity): symmetries with respect to a point, a straight line or a plane, are involutions. For these, the following notion applies:

Definition A.3. *A function* φ *is said to be* even *[resp.* odd*] with respect to the involutive symmetry* s *if* $S\varphi = \varphi$ *[resp.* $S\varphi = -\varphi$*]. A vector field* v *is* even *[resp.* odd*] if* $Sv = v$, *that is to say* $s_* v = v$ *at all points [resp.* $s_* v = -v$*].*

It's easy to see that if a function is even or odd, its gradient has the same property, and that the divergence of a vector field has the same parity as the field. In contrast, the curl of an even or odd field has *opposite* parity in the case of a *skew* symmetry (reflection with respect to a point or a plane) and the same parity in the case of a direct symmetry (half-turn around some axis). This reflects the "sensitivity to orientation" of rot, as earlier remarked.

These properties rule the setting of boundary conditions, in a quite simple way, at least as far as mirror symmetries are concerned. Suppose (which is the general case) the source of the field is a given current density j^g. If j^g is even[43] [resp. odd], j has the same property, and hence e (at least in conductors) is even [resp. odd], provided σ is even with respect to this mirror. By Faraday's law, b then has the symmetry of rot e, which means odd [resp. even]. And so forth, for all fields. Once the parity of all fields has thus been determined, boundary conditions follow from simple rules: For fields which are, like b, associated with surfaces (fields d and j), the boundary condition is $n \cdot b = 0$ in case of even fields, no condition at all in case of odd fields. For fields like h which are associated with lines (fields e and a), it's the opposite: The boundary condition is $n \times h = 0$ in case of odd fields, no condition at all in case of even fields. Since h and b (or d and e, or j and e) have same parity, boundary conditions on symmetry planes are complementary: $n \cdot b = 0$ on some, $n \times h = 0$ on others. We had a concrete example of this with the Bath-cube problem.

For more on this subject, see [B1, B2].

Let's now give a few other practical examples, also borrowed from the TEAM workshop trove.

[43] It's always possible to express j^g as the sum of an even and an odd component, and to do this repeatedly for all mirror symmetries, thus forming kind of "Fourier components" of the source. One then solves one reduced problem (on the symmetry cell) for each of these components, and adds the results.

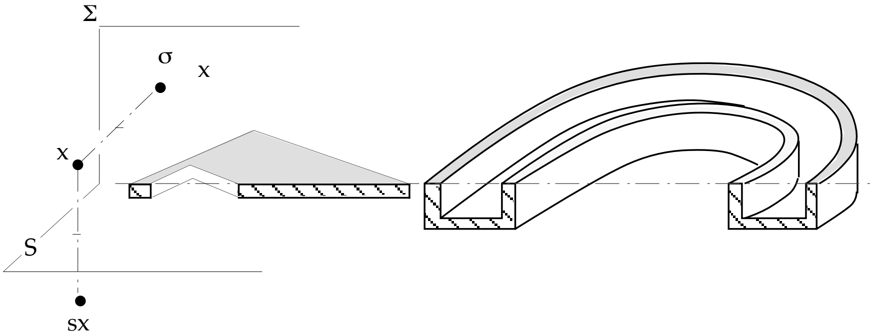


FIGURE A.15. Sketch of TEAM Pbs. 7, 14, with common symmetry D_2. (Only the "material symmetry cell", below S and behind Σ, is shown. This is a part of the passive conductor that generates all of it by letting the symmetry operations act.) The inductor, not represented here, may not share the symmetries of the passive conductor, but this does not impede the exploitation of symmetry [B1]. TEAM Problems 8, on the detection of a crack inside an iron piece, and 19, on a microwave cavity, have the same kind of symmetry.

Problems 7 (the misnamed "asymmetrical" plate with a hole), 8 (coil over a crack) and 14 (the "Euratom casing" [R&]) fall into a category described by the group C_{2v}, with four elements. It is generated by reflections s and σ with respect to two orthogonal planes S and Σ. Its elements are thus {1, s, σ, s σ} (Fig. A.15).

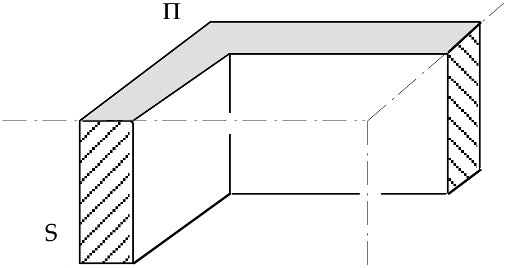


FIGURE A.16. Symmetry D_{2h}, common to Pbs. 3, 4, 1. (The example shown is the symmetry cell of the "Felix brick".)

When the group is generated by reflections s, σ, π with respect to *three* orthogonal planes, it is called D_{2h} (Fig. A.16). It has 8 elements, and is relevant to Pbs. 3 (the "Bath ladder") and 4 (the "Felix brick" [T&]). Problem 12 (the cantilevered flexible plate [C&]) should be included in this category, because all computations relative to it can be conducted in the so-called "reference configuration" of the conductive plate, with negligible error (because of small deformations), and the symmetry group of this reference configuration is D_{2h}.

Some problems are "much more symmetrical" than any of the above, having infinite symmetry groups. The frequent (and never formally explained . . .) references in this book to "2D modelling" have to do with geometrical symmetry: 2D modelling is relevant when the symmetry group of the problem contains all translations along some direction. Symmetry reduction by one spatial dimension also occurs in case of axisymmetry (group SO_2 or larger). For instance, TEAM Pbs. 1 and 2, the "Felix cylinders", and Pb. 9 on the far-field effect in a tube, have symmetry group O_2, composed of all rotations around a fixed axis, combined with reflections with respect to an axial plane.

Even more extended symmetry can happen. Problems 6 and 11 on eddy currents induced in a hollow sphere have symmetry group O_3: all rotations around a fixed point, combined with reflections with respect to the origin. Fourier series is the right (and well known) tool for such cases. It goes quite far, up to giving *exact* solutions, by formulas, in some cases (the hollow sphere, for instance).

A.4 GLIMPSES OF FUNCTIONAL ANALYSIS

A.4.1 Completion, principle of extension by continuity

Cauchy sequences in a metric space $\{X, d\}$ are sequences $\{x_n : n \in \mathbb{N}\}$ such that $d(x_n, x_m)$ tends to zero when both indices n and m tend to infinity. Convergent sequences are Cauchy. A space is *complete* if, conversely, all its Cauchy sequences converge. A normed vector space with this property is called a *Banach space.*

In applied mathematics, the only good spaces are complete spaces, as we experienced in Chapter 3. So let's give in full this construction of complete spaces that proved so important then:

Theorem A.1 (of completion). *Given a metric space* $\{X, d\}$, *there exists a space* $\{\hat{X}, \hat{d}\}$ *and an isometry* $i \in X \rightarrow \hat{X}$, *such that* $\hat{X}$ *be complete and* cod(i) *dense in* $\hat{X}$.

Proof. The key proof-ideas were given in 3.2.3, and we just fill in details. X° being the set of all Cauchy sequences $x^\circ = \{x_1, \ldots, x_n, \ldots\}$ of elements of X, set $x^\circ \sim y^\circ$ if $\lim_{n \rightarrow \infty} d(x_n, y_n) = 0$. This is an equivalence, because $x^\circ \sim y^\circ$ and $y^\circ \sim z^\circ$ imply that $d(x_n, z_n) \leq d(x_n, y_n) + d(y_n, z_n)$, tends to 0, by the triangular inequality. Now define $\hat{X}$ as the quotient $X^\circ / \sim$, and set

$\hat{d}(\hat{x}, \hat{y}) = \lim_{n \to \infty} d(x_n, y_n)$. Then $\hat{d}(\hat{x}, \hat{z}) \leq \hat{d}(\hat{x}, \hat{y}) + \hat{d}(\hat{y}, \hat{z})$, still by the triangular inequality, it's obvious that $\hat{d}(\hat{y}, \hat{x}) = \hat{d}(\hat{x}, \hat{y})$, and if $\hat{d}(\hat{x}, \hat{y}) = 0$, classes $\hat{x}$ and $\hat{y}$ coincide, since two representatives x° and y° will satisfy $\lim_{n \to \infty} d(x_n, y_n) = 0$ and thus be equivalent. Let $i(x)$ be the sequence $\{x, \dots, x, \dots\}$. Then $\hat{d}(i(x), i(y)) = d(x, y)$, so i is an isometry. The image $i(X)$ is dense in $\hat{X}$, because if $x^\circ = \{x_1, \dots, x_n, \dots\}$ is a representative of $\hat{x}$, then $\hat{d}(\hat{x}, i(x_n)) = \lim_{m \to \infty} d(x_m, x_n)$, which tends to 0 by definition of a Cauchy sequence. Finally, if $\{\hat{x}^n : n \in \mathbb{N}\}$ is a Cauchy sequence of $\hat{X}$, select a sequence $\{\varepsilon_n : n \in \mathbb{N}\}$ of reals which tend to 0, and for each n, choose $x_n \in X$ such that $\hat{d}(i(x_n), \hat{x}^n) \leq \varepsilon_n$, which the density of $i(X)$ makes possible. As $d(x_n, x_m) = \hat{d}(i(x_n), i(x_m)) \leq \hat{d}(i(x_n), \hat{x}^n) + \hat{d}(\hat{x}^n, \hat{x}^m) + \hat{d}(\hat{x}^m, i(x_m)) \leq \varepsilon_n + \hat{d}(\hat{x}^n, \hat{x}^m) + \varepsilon_m$, which tends to 0, the x_ns form a Cauchy sequence. Let $\hat{x}$ be its class. Then $\hat{d}((\hat{x}, \hat{x}^n) \leq \hat{d}(\hat{x}, i(x_n)) + \hat{d}(i(x_n), \hat{x}^n) \leq \lim_m \hat{d}(x_m, x_n) + \varepsilon_n$, which goes to 0 as n increases, showing that $\hat{x}$ is the limit of $\{\hat{x}^n\}$. ◊

Note that one can legitimately refer to *the* completion, because if one can find, by some other method, another dense injection j of X into some complete space $X\hat{}$, then elements of $\hat{X}$ and $X\hat{}$ are in isometric correspondence, so the completion is unique up to isometry. The proof is *constructive,* giving us one of these isometric complete spaces in explicit form. One can argue that $\hat{X}$ is not necessarily "the right one", however. Indeed, our intuitive notion of completion seems to require embedding X into a space made of objects of the *same* type as those of X. Hence the search, in most cases, for such a "concrete" complete space. For instance, if X is a space of functions defined on a domain of E_3, one will try[44] to identify its completion with a similar functional space. An important example will be given below, where $L^2(D)$, the completion of $C(D)$, is embedded in a space of functions defined on D, thanks to Lebesgue integration theory.

There is a companion result to the completion theorem:

Theorem A.2 (of extension by continuity). *Let* X *and* Y *be metric spaces, both complete,* U *a* dense *part of* X, *and* $f_U \in X \to Y$, *with* $\mathrm{dom}(f_U) = U$, *a* uniformly *continuous function. There is an extension* f *of* f_U *to all* X *that is continuous, and it's the only one.*

Proof. Take $x \in X$ and let $\{x_n \in U\}$ be a sequence that converges to x. Because of *uniform* continuity, the $f_U(x_n)$ form a Cauchy sequence, which converges, since Y is complete, towards a point that one can denote $f(x)$, because it does not depend on the chosen sequence. As $f(x) = f_U(x)$ if $x \in U$,

[44] And when this fails, never mind: A cunning extension of the very notion of function will often save the day.

one thus obtains an extension of f_U the domain of which is all X, and one easily checks that f is (uniformly) continuous. If g is another continuous extension of f_U, then $\lim_{n \to \infty} g(x_n) = g(x)$ by continuity, so $g(x) = f(x)$. ◊

Obviously, continuity of f_U might not be enough (Fig. A.17).

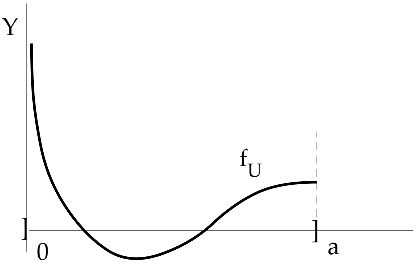


FIGURE A.17. Here, $X = [0, a]$, part of IR, $U =]0, a]$ and $Y =$ IR. In spite of its continuity, f_U has no continuous extension to X.

This result is almost often applied to the extension of *linear* (or affine) maps, between normed spaces, and *then* continuity is enough, because affine continuous maps are uniformly continuous. It works as follows: When a linear map (or as one prefers to say then, an "operator") $L : X \to Y$ of domain U is continuous, one can extend it into an operator from the completion X of U to the completion of Y, since U is dense in X; just apply the previous result to the composition $i_Y \circ L$, where i_Y is the canonical injection of Y into its completion.

A.4.2 Integration

I assume you know about integration, though not necessarily about *Lebesgue* integration *theory*. It's an ample and difficult theory, which cannot even be sketched here. And yet, some of its results are absolutely essential when it comes to weak formulations, complete spaces, existence proofs, etc. Fortunately, one can live in blissful ignorance of the theory, provided one is aware of what it does better than the older and (only apparently) easier Riemann theory.[45]

What Riemann's theory does, and does fairly well, is to give sense to the concept of *average value* of a *continuous* function over a set where concepts such as "length", "area", or "volume" make sense. (The generic

[45] As stressed in [Bo], the standard comment, "in one case you divide up the x-axis and in the other you divide up the y-axis", is totally misleading in its emphasis on a tiny technical difference.

term is "measure". For instance, on the real line, the measure of an interval [a, b] is $|b - a|$, in both theories.) After some work on so-called "Riemann sums", one obtains a sensible definition of the integral $I([a, b], f) = \int_{[a, b]} f$ of f over [a, b], also denoted $\int_{[a, b]} f(x)\, dx$, or $\int_a^b f(x)\, dx$, which embodies the concept of "area under the graph", when $f \geq 0$. (The average is then $I/|b - a|$.) Extensions to sets other than intervals, and to several variables, then follow; hence a map, the type of which is

$$PART\_OF\_A\_MEASURED\_SPACE \times FUNCTION \rightarrow REAL,$$

with the right properties: additivity with respect to the set, linearity with respect to the function. The integral of f over A is denoted $\int_A f$.

Remark A.6. This reduced notation, recommended in Note 1.8 and largely used in this book, reflects the functional character of integration: All that is left is the operator symbol $\int$ and the two arguments, A and f. There is no ambiguity when A is a part of a set X on which exists a standard measure (which is the case of E_3), and if A is all X, one may even not mention it. Developed notation such as $\int_X f(x)\, dx$ or $\int_X f(x)\, d\mu(x)$ may be useful when one must be explicit about the underlying measure (because several of them can appear simultaneously, for instance, or to sort out multiple integrals: several examples appear in Chapter 1), but in such cases, x is a bound variable, that must appear (at least) twice in an expression, as argument of the function *and* of the measure element. Expressions like $\int_X f(x)\, dV$, for instance, are not well-formed in this respect, and should not be used. ◊

There is however an essential flaw in this theory. When a sequence of functions f_n converges pointwise[46] towards some function f, one cannot assert that $\int_X f = \lim_{n \to \infty} \int_X f_n$, if only because the limit f may be outside the domain of the above mapping, and thus not be "integrable in the Riemann sense". Because of this shortcoming, one cannot safely permute integration and passage to the limit, like this: $\int_X \lim_{n \to \infty} f = \lim_{n \to \infty} \int_X f_n$. The Lebesgue theory corrects that by enlarging the domain of the map: There are more functions integrable "in the Lebesgue sense", on more exotic sets. This advantage, by itself, is marginal, for it's not so often that one *must* compute the average of an everywhere discontinuous function on a Cantor set, or similar. The point is elsewhere: In Lebesgue theory, one *can* permute limit and integration, under the condition of *dominated convergence*, that is, when there exists a function g, integrable itself, such that $|f_n(x)| \leq g(x)$ whatever x and n. This commutativity between two such fundamental operations is the great triumph of Lebesgue's theory,

[46] That is, $f_n(x)$ tends to f(x), as a sequence of real numbers, for a fixed x.

because it legitimizes a series of basic manipulations in calculus: differentiation under a summation sign (hence the possibility to permute differentiation and convolution alluded to p. 290), change of order of summation in multiple integrals (Fubini's theorem), and so forth.

How is this achieved? Very roughly, take the space of continuous functions $C(X)$ on some metric space X, and give it a norm, by setting $\|f\|_1 = \int_X |f(x)|\,dx$, where the integral is understood in the Riemann sense. This is not a *complete* space (as discussed in 3.2.1). So take its completion, call this enlarged space $L^1(X)$, and now that every Cauchy sequence does converge, define the integral of the limit f of $\{f_n\}$ (which has just been defined into existence by the process of completion) as the limit of the integrals $\int f_n$. It's so simple that one may wonder where the difficulty is that makes books so thick [Ha, Lo]: Again, it comes from the completion being an abstract space, not a priori a functional one (the above limit f is an abstract object, not yet a function), and the hard work consists in embedding this abstract completion $L^1(X)$ into some functional space.

One thus introduces (after a copious measure of measure theory) a concept of "measurable function", which is very encompassing,[47] and an equivalence relation, the "almost everywhere equal (a.e.)" relation alluded to at places in this book: $f \overset{\text{a.e.}}{=} g$ (or $f(x) = g(x)$ a.e.) if points x where $f(x) \neq g(x)$ form a "negligible" set, that is, one to which Lebesgue's measure theory attributes the measure 0. Once all this, which is an impressive piece of work, is said and done, one can identify the elements of $L^1(X)$ with *equivalence classes* of a.e.-equal measurable functions. Yet, one continues to call "functions" the elements of $L^1(X)$, and this abuse is natural enough: Two almost everywhere equal functions belong to the same class and have *the same integral*, so from the point of view of integration theory they are "the same" indeed. This is even more justified when one realizes that a *continuous* function is *alone* in its own class (because two a.e.-equal continuous functions must coincide).

Once in possession of $L^1(X)$, one can define $L^2(X)$ as the space of "functions" the square of which is in $L^1(X)$, or more precisely, as the completion of $C(X)$ with respect to the *quadratic* norm, $\|f\|_2 = (\int |f|^2)^{1/2}$ instead of the L^1-norm $\|f\|_1 = \int |f|$. This L^2-norm is associated with a scalar product, namely $(f, g) = \int f(x)\,g(x)\,dx \equiv \int fg$, so L^2 is pre-Hilbertian, and being also complete, is a Hilbert space, *the* essential concrete realization of this abstract notion.

[47] It's such a large class that no constructive examples of *non*-measurable functions exist; one must invoke the axiom of choice to get them.

Most of the time, it's convenient to think of elements of $L^2(X)$ as functions, though they are actually *classes* of functions. But there is one case in which awareness of the real nature of L^2 is important: when one tries to define the restriction of a function of $L^2(D)$, where D is our usual "computational domain", to its boundary S. The boundary being a negligible set, values of f on S can be changed at will without changing the *class* f belongs to, which means that "restriction to S" is a meaningless expression as regards elements of $L^2(D)$. And yet we need such a concept to deal with boundary-value problems! Hence the introduction of the relatively sophisticated notion of *trace*: The trace γf of a continuous function f is just its restriction to S. Now if f is a generic element of $L^2(D)$, there is, by construction, a Cauchy sequence of continuous functions f_n which tends towards f in quadratic norm. So we define γf, the trace of f, as the limit in $L^2(S)$ of the sequence of restrictions γf_n, *provided this sequence converges*, which may not be the case: Some "functions" of $L^2(D)$ have traces, some have not. The question is discussed in Chapter 7, where it is shown in detail how functions the gradient of which (in the weak sense) is square summable in D do have traces, even though they need not be continuous. All this makes only the beginning of the (difficult) theory of Sobolev spaces, but what precedes is enough baggage for our needs.

Apart from this all-important extension of scope, Lebesgue theory does not bring anything new when it comes to the more mundane aspects of integration as used in calculus, such as integration by parts, change of variables, and the like. Let's just stress two points of special importance, the definition of *circulations* and *fluxes*.

Let c denote a bounded curved line in E_3. On c, the Euclidean distance existing in E_3 induces a notion of length of curved segments, which turns c into a measured space, on which integration makes sense: If f is a function whose domain contains c, the integral $\int_c f$ is the average of f on c, multiplied by the length of c.

Now, let's equip c with a field of tangent vectors. For this, take a parameterization of c, that is to say, a smooth map, still denoted c, from $[0, 1]$ into E_3, having the curved line as codomain. (The deliberate confusion between the path $c : [0, 1] \rightarrow E_3$ and the curve proper, which is only the codomain of this path, has obvious advantages, provided one stays aware of the distinction.) Assume the derivative $\partial_t c(t)$, which is a vector of V_3, does not vanish for $t \in [0, 1]$. Set $\tau(t) = \partial_t c(t)/|\partial_t c(t)|$: This is the *unit tangent vector* at point $c(t)$. (Obviously, whatever the parameterization, there are only two possible fields τ, each corresponding to one of the two possible orientations of c. Cf. p. 287 and 5.2.1.)

Finally, let u be a smooth vector field, the domain of which contains c. By taking the dot product $\tau(x) \cdot u(x)$ for each point x of c, one obtains a smooth real-valued function of domain c, naturally denoted by $\tau \cdot u$. This function can be integrated on c, hence a number $\int_c \tau \cdot u$. This is, by definition, the *circulation of* the vector field u *along* c, *as oriented by* τ. (Of course it reverses sign with orientation.)

The same things exactly can be said about a smooth patch C, mapping $[0, 1] \times [0, 1]$ into E_3, and such that vectors $\partial_s C$ and $\partial_t C$ at point $C(s, t)$ don't vanish. One then forms a normal field $n(s, t) = N(s, t)/|N(s, t)|$, where $N(s, t) = \partial_s C \times \partial_t C$, again with only two possible outcomes, corresponding to orientations of C. Again, $n \cdot u$ is a scalar function on C, whose integral $\int_C n \cdot u$ is called the *flux* of u *through* C *as oriented by* n. By sewing patches together, and orienting them consistently, one can thus define fluxes relative to smooth orientable surfaces. This is the case, in particular, of the surface S of a computational domain D, and we have often had to deal with integrals like $\int_S n \cdot u$, especially when using the two basic integration by parts formulas, established in 2.3.1 and 2.3.2:

$$(2) \qquad \int_D \varphi \operatorname{div} b = -\int_D b \cdot \operatorname{grad} \varphi + \int_S n \cdot b \ \varphi,$$

$$(3) \qquad \int_D h \cdot \operatorname{rot} a = \int_D a \cdot \operatorname{rot} h - \int_S n \times h \cdot a.$$

These formulas concern smooth fields, but thanks to the good behavior of Lebesgue integrals with respect to passages to the limit, one can extend these formulas by continuity to $\varphi \in L^2_{\text{grad}}(D)$, $b \in \mathbb{L}^2_{\text{div}}(D)$, h and a in $\mathbb{L}^2_{\text{rot}}(D)$, as defined in Chapter 5, thus giving them enlarged validity. See Section 5.1 for this important development.

A.4.3 Hilbert spaces

A real[48] vector space X is *pre-Hilbertian* when equipped with a scalar product, as previously defined. The function $\|\ \| = x \to (x, x)^{1/2}$ is then a norm, which confers a metric on X. (The triangular inequality comes from

$$(4) \qquad |(x, y)| \le \|x\| \, \|y\|,$$

which is the *Cauchy–Schwarz* inequality.) Note that $(\ ,\ )$ is continuous with respect to both its arguments. A simple computation yields the following *parallelogram equality*:

[48] That is, built on $\mathbb{R}$ as scalar field. Complex spaces are not less important, but there is some gain in simplicity in treating them apart, as we do a little later.

$$(5) \qquad ||x-y||^2+||x+y||^2=2(||x||^2+||y||^2) \qquad \forall\, x, y\in X.$$

The existence of a scalar product gives sense to the notions of *orthogonality* in X (x and y are orthogonal if $(x, y) = 0$, which one may denote $x \perp y$) and *angle* (the angle of two nonzero vectors x and y is $\arccos((x, y)/||x||\,||y||)$), so all the concepts of Euclidean geometry make sense: The Pythagoras theorem holds, and (5) is nothing else than a generalization of the metric relation between median and sides in elementary geometry of the triangle (Fig. A.18). Such things cannot be said of any normed vector space, only if (5) is valid for the given norm $||\ ||$, for then one can prove that $\{x, y\} \to (||x+y||^2 - ||x-y||^2)/4$ is a scalar product. Pre-Hilbertian spaces, and their affine associates, are therefore those spaces in which notions and concepts of ordinary Euclidean geometry hold, without any restriction on the dimension: *their theory extends intuitive geometry to infinite dimension.*

A *Hilbert space* is a *complete* pre-Hilbertian space, and we saw many examples, almost all of them related with the spaces L^2 or $\mathbb{L}^2$.

The basic result about Hilbert spaces is this:

Theorem A.3 (of projection). *Let* C *be a* closed convex *part of a* Hilbert *space* X. *The function "distance to* C*", i.e.,* $d_C = x \to \inf\{||x-y|| : y \in C\}$, *reaches its lower bound at a unique point of* C, *called the* projection *of* x *on* C, *here denoted* $p_C(x)$.

Proof. Most of the proof appears in 3.2.1, the only difference being that there, C was not only convex but an affine subspace. In particular, the key concept of *minimizing sequence* was introduced there. So let's be terse: The lower bound $d = d_C(x) = \inf\{||x-y|| : y \in C\}$ can't be reached, if it is reached at all, at more than one point, for if $||x-y|| = ||x-z|| = d$ for $y \neq z$, then $u = (y+z)/2$ would belong to C by convexity, whereas $||x-u|| < d$ after (5), hence a contradiction. As for existence, let $y_n \in C$ be a minimizing sequence, i.e., $||x-y_n||$ converges towards $d_C(x)$. It's a Cauchy sequence, because

$$||y_n - y_m||^2 + 4\,d^2 \leq ||y_n - y_m||^2 + 4\,||x - (y_n + y_m)/2||^2$$
$$= 2(||x-y_n||^2 + ||x-y_m||^2),$$

thanks to (5) and to the convexity of C, and the right-hand side tends to $4d^2$. Since X is complete, there is a limit, which belongs to C, since C is closed. ◊

Remark A.7. The inequality that characterizes the projection, that is

$$(6) \qquad ||p_C(x) - x||^2 \leq ||y-x||^2 \quad \forall\, y \in C,$$

can also be written as (develop the scalar product)

$$(7) \qquad (x - p_C(x),\ y - p_C(x)) \le 0 \quad \forall\, y \in C.$$

This is called a "variational inequality", or *variational inequation,* if considered as the problem "given x, find $p_C(x)$". Observe how this is "read off" Fig. A.18, right, confirming the remark on Hilbertian geometry as the natural extension of Euclidean geometry to infinite dimensions. Equation (7) is called the *Euler equation* of the *variational problem* (6). ◊

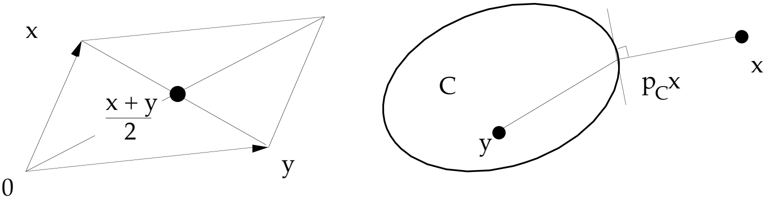


FIGURE A.18. The parallelogram equality (left) and inequality (7).

Remark A.8. The map p_C is a *contraction,* in the sense that

$$\|p_C(x) - p_C(y)\| \le \|x - y\| \quad \forall\, x, y \in X.$$

To see it, replace y by $p_C(y)$ in (7), permute x and y, add, and apply Cauchy–Schwarz. ◊

In the particular case when C is a closed subspace Y of X, (7) becomes an equality, or variational equation:

$$(x - p_Y x,\ y) = 0 \quad \forall\, y \in Y.$$

The vector subspace formed by all elements of X orthogonal to Y is called the *orthocomplement* of Y, or more simply, its *orthogonal,* denoted $Y^\perp$. It is closed, as easily checked (cf. Remark A.9). One can therefore apply Theorem A.3 to it, and the projection of x on $Y^\perp$ appears to be $x - p_Y x$. Thus, any x in X can be written as the sum of two orthogonal vectors, one in Y, one in its orthocomplement. Moreover, this decomposition is unique, for $y_1 + z_1 = y_2 + z_2$, with y_i in Y and z_i in $Y^\perp$, $i = 1, 2$, implies $y_1 - y_2 = z_2 - z_1$ at the same time as $y_1 - y_2 \perp z_2 - z_1$, and hence $y_1 = y_2$ and $z_1 = z_2$. One says that Y and $Y^\perp$ have X as *direct sum,* and this is denoted $X = Y \oplus Y^\perp$. Note that $Y^{\perp\perp} = Y$.

Remark A.9. If the subspace Y is *not* closed, one may still define its orthocomplement by $Y^\perp = \{z \in X : (y, z) = 0 \ \forall\, y \in Y\}$. It's closed, because if $z_n \in Y^\perp$ converges to z, then $(y, z) = \lim_{n \to \infty} (y, z_n) = 0$ for all y in Y, by continuity of the scalar product. By applying the projection theorem to $Y^{\perp}$, one sees that $Y^{\perp\perp}$ is not Y, but its *closure* in X. ◊

A second special case is when Y is the kernel of a linear continuous functional $f : X \to \mathbb{R}$. Then Y is closed indeed, and does not coincide with X if f is not trivial, so there exists in $Y^{\perp}$ some nonzero vector z. The equality $x = x - \theta z + \theta z$ then holds for all x and all real θ. But $x - \theta z$ belongs to Y if $\theta = f(x)/f(z)$, so $z \perp (x - \theta z)$ for this value θ, and hence $(x, z) = \theta \|z\|^2$, that is, finally,

$$f(x) = (x, z\, f(z)/\|z\|^2) \quad \forall\, x \in X.$$

So there exists a vector z_f that "represents f ", in the precise sense that its scalar product with x is f(x), and this vector is $z_f = z\, f(z)/\|z\|^2$. Moreover (apply (4)), $\|z_f\| = \sup\{|f(x)| / \|x\| : x \neq 0\}$, that is $\|z_f\| = \|f\|$. The correspondence between f and z_f thus achieved is therefore a linear isometry, and we may conclude:

Theorem A.4 (F. Riesz). *To each linear continuous real-valued function* f *on a real Hilbert space* X, *there corresponds a unique vector* z_f *such that* $f(x) = (x, z_f)\ \forall\, x \in X$, *and* $\|f\| = \|z_f\|$.

In this respect, a Hilbert space is "its own dual". But beware there can be other isomorphisms between a concrete Hilbert space and its dual than the Riesz one, which is both an asset (one can solve boundary-value problems that way) and an inexhaustible source of puzzlement. See for example the several isomorphisms between $H^{1/2}(S)$ and its dual in Section 7.4.

Third special case: when C is some *affine* closed subspace X^g, with X^0 as parallel vector subspace, and the point to be projected is the origin. Calling x the projection, we see that x solves the problem, *find* $x \in X^g$ *such that* $(x, x') = 0\ \forall\, x' \in X^0$. As the slight change in notation should help one to realize, this result is the paradigm of our existence proofs in Chapters 4 and 6: By adopting the energy-related scalar product, we were able to apply the projection theorem directly in this form. It's not always convenient, however, and the following generalization then comes handy:

Lax–Milgram's lemma. *Let* $a : X \times X \to \mathbb{R}$ *be a bilinear map, continuous with respect to both arguments, and such that*

$$(8) \qquad a(x, x) \geq \alpha\, \|x\|^2 \quad \forall\, x \in X,$$

where α *is a strictly positive real number* (coercivity *of* a). *Given a linear continuous functional* $f \in X \to \mathbb{R}$, *the problem* find $x \in X$ such that

$$(9) \qquad a(x, x') = f(x') \quad \forall\, x' \in X$$

has a unique solution x_f, *and the mapping* $f \to x_f$ *is* continuous.

Proof. Since $x' \to a(x, x')$ is continuous, there exists, by the Riesz theorem, some element of X, which can be denoted Ax (a single symbol, for the time being), such that $(Ax, x') = a(x, x')$ for all x'. This defines a linear continuous operator A from X into itself, injective by virtue of (8), and Eq. (9) can then be written as $Ax = z_f$, where z_f is the Riesz vector of f. This is equivalent to $x - \rho(Ax - z_f) = x$, where $\rho \neq 0$ is a parameter that can be chosen at leisure. Let $\{x_n\}$ be the sequence defined by $x_0 = 0$ and $x_{n+1} = (1 - \rho A)x_n + \rho z_f$. If it does converge, the limit is the solution x_f of $Ax = z_f$, and $\alpha \|x_f\| \leq \|z_f\| \equiv \|f\|$ after (9), hence the continuity of $f \to x_f$. The sequence will converge if $\|1 - \rho A\| < 1$, so let's compute:

$$\|x - \rho Ax\|^2 \leq \|x\|^2 - 2\rho\,(Ax, x) + \rho^2 \|Ax\|^2 \leq \|x\|^2 - 2\rho\alpha\|x\|^2 + \rho^2 \|Ax\|^2$$

after (8), so $\|1 - \rho A\| < 1$ if $0 < \rho < \|A\|^2/2\alpha$. (Note that no *symmetry* of a was assumed or used.) ◊

The standard application is then to the problem, *find* $x \in U^g$ *such that* $a(x, x') = 0 \ \forall\, x' \in U^0$, where a is a continuous bilinear map. By picking some x^g in U^g, this amounts to finding x in U^0 such that $a(x^0 + x^g, x') = 0 \ \forall\, x' \in U^0$. As seen by setting $f(x') = -\,a(x^g, x')$ and $X = U^0$, the lemma applies if the restriction of a to U^0 is coercive.

As mentioned in Note 48, the need arises to extend all these notions and results to complex spaces. This is most easily, if not most compactly, done by *complexification*. The *complexified* U^c of a vector space U is the set $U \times U$ with composition laws induced by the following prescription: An element $\text{U} = \{u_R, u_I\}$ of U^c being written in the form $\text{U} = u_R + iu_I$, one applies the usual rules of algebra, with $i^2 = -1$. Thus, $\text{U} + \text{U}' = u_R + iu_I + u'_R + iu'_I = u_R + u'_R + i(u_I + u'_I)$, and if $\Lambda = \lambda_R + i\,\lambda_I$, then

$$\Lambda \text{U} = (\lambda_R + i\,\lambda_I)(u_R + iu_I) = \lambda_R u_R - \lambda_I u_I + i(\lambda_R u_I - \lambda_I u_R).$$

The *Hermitian* scalar product (U, V) of two complex vectors $\text{U} = u_R + iu_I$ and $\text{V} = v_R + iv_I$ is by convention the one obtained by developing the product $(u_R + iu_I, v_R - iv_I)$ after the same rules, so $(\text{U}, \text{V}) = (u_R, v_R) + (u_I, v_I) + i(u_I, v_R) - i\,(u_R, v_I)$. The norm of U is given by $|\text{U}|^2 = (\text{U}, \text{U})$. (Be aware that a different convention is adopted in Chapters 8, where expressions such as $(\text{rot}\ \text{U})^2$ are understood as $\text{rot}\ \text{U} \cdot \text{rot}\ \text{U}$, not as $|\text{rot}\ \text{U}|^2$.)

Now, when X is complex, all things said up to now remain valid, if (x, y) is understood as the Hermitian scalar product, with obvious adjustments: f is *complex*-valued, and the Riesz vector is no longer linear, but *anti*-linear with respect to f (to multiply f by λ multiplies x_f by λ^*). The form a in the Lax–Milgram lemma becomes "sesqui"-linear (anti-linear with respect to the second argument), and the same computation as above yields the same result, provided

$$\mathrm{Re}[a(x, x)] \geq \alpha \, ||x||^2 \quad \forall \, x \in X,$$

with $\alpha > 0$, which is what "coercive" means in the complex case.

Remark A.10. The lemma remains valid if λa is coercive, in this sense, for some complex number λ. We make use of this in 8.1.3, where the problem is of the form *find* $x \in U^g$ *such that* $a(x, x') = 0 \quad \forall \, x' \in U^0$, with $(1 - i)a$ coercive over U^0. ◊

The theory does not stop there. Next steps would be about orthonormal bases and Fourier coefficients, whose treatment here would be out of proportion with the requirements of the main text. Let's just mention (because it is used once in Chapter 9) the notion of weak convergence: A sequence $\{x_n : n \in \mathbb{N}\}$ *weakly converges* toward x if

$$\lim_{n \to \infty} (x_n, y) = (x, y) \quad \forall \, y \in X.$$

This is usually denoted by $x_n \rightharpoonup x$. By continuity of the scalar product, convergence in the standard sense (then named "strong convergence" for contrast) implies weak convergence, but not the other way around: for instance, the sequence of functions $x \to \sin nx$, defined on $[-1, 1]$, converges to 0 weakly, but not strongly. However, weak convergence *plus* convergence of the *norm* is equivalent to strong convergence.

Compact operators are those that map weakly convergent sequences to strongly convergent ones. It's not possible to do justice to their theory here. Let's just informally mention that, just as Hilbert space is what most closely resembles Euclidean space among infinite-dimensional functional spaces, compact operators are the closest kin to matrices in infinite dimension, with in particular similar spectral properties (existence of eigenvalues and associated eigenvectors). An important result in this theory, *Fredholm's alternative*, is used in Chapter 9. Cf. (for instance) [Yo] on this.

A.4.4 Closed linear relations, adjoints

The notion of *adjoint* is essential to a full understanding of the relations between grad and div, the peculiarities of rot, and integration by parts formulas involving these operators.

We know (p. 284) what a linear relation $A : X \to Y$ is: one the graph $\mathcal{A}$ of which is a subspace of the vector space $X \times Y$. If the relation is functional, i.e., if the section $\mathcal{A}_x$ contains no more than one element, we have a linear operator. By linearity, this amounts to saying that the only pair in X × Y of the form {0, y} that may belong to $\mathcal{A}$ is {0, 0}.

Suppose now X and Y are Hilbert spaces, with respective scalar products $(\ ,\ )_X$ and $(\ ,\ )_Y$. Whether $\mathcal{A}$ is closed, with respect to the metric induced on X × Y by the scalar product $(\{x, y\}, \{x', y'\}) = (x, x')_X + (y, y')_Y$, is a legitimate question. If A is continuous, its graph is certainly closed, for if a sequence of pairs $\{x_n, Ax_n\}$ belonging to $\mathcal{A}$ converges to some pair {x, y}, then y = Ax. The converse is not true (Remark A.4), so we are led to introduce the notion of *closed* operator, as one the graph of which is closed.

Now if the graph of a linear relation $\{X, Y, \mathcal{A}\}$ is not closed, why not consider its *closure* $\{X, Y, \bar{\mathcal{A}}\}$? We get a new relation this way, which is an extension of the given one. But it may fail to be functional, because pairs of the form {0, y} with y ≠ 0 may happen to be adherent to $\mathcal{A}$. Hence the following definition: An operator is *closable* if the closure of this graph is functional. In Chapter 5, we work out in detail the case of div: $\mathbb{L}^2(D) \to L^2(D)$, with domain $\mathbb{C}^\infty(\bar{D})$, find it closable, and define the "weak" divergence as its closure. The new operator thus obtained has an enlarged domain (denoted $\mathbb{L}^2_{\text{div}}(D)$) and is, of course, closed, but not continuous on $\mathbb{L}^2(D)$.

There is a way to systematically obtain closed operators. Start from some operator A, and take the orthogonal $\mathcal{A}^\perp$ of its graph in X × Y. This is, as we know, a closed subspace of the Cartesian product. Now consider the relation $\{Y, X, \mathcal{A}^\perp\}$, with X as target space. *If* this happens to be a functional relation, we denote by $-\,A^*$ the corresponding operator, which thus will satisfy the identity

$$(11)\qquad (x, A^*\, y)_X = (y, Ax)_Y \quad \forall\ \{x, y\} \in \mathcal{A},$$

and call A*—an operator of type Y → X—the *adjoint*[49] of A.

So when is $\mathcal{A}^\perp$ functional? The following statement gives the answer:

Proposition A.1. *Let* $A = \{X, Y, \mathcal{A}\}$ *be a given linear relation. The relation* $\{Y, X, \mathcal{A}^\perp\}$ *is functional if and only if* dom(A) *is dense in* X.

Proof. If $\{x, 0\} \in \mathcal{A}^\perp$, then $(x, \xi)_X = (0, A\xi)_Y \equiv 0$ for all $\xi \in$ dom(A), after (11). So if dom(A) is dense, then x = 0, and $\mathcal{A}^\perp$ is functional. Conversely, if dom(A) is not dense, there is some x ≠ 0 in the

[49] Not to be confused with the *dual* of A, similarly defined, but going from the dual Y' of Y to the dual X' of X. The notion of adjoint is specifically Hilbertian.

orthocomplement of dom(A) with respect to X, and hence a nontrivial pair $\{x, 0\} \in \mathcal{A}^\perp$. ◊

Remark A.11. After (11), the domain of A^* is made of all y such that the linear partial function $x \to (y, Ax)_Y$ be continuous on dom(A), with respect to the metric of X. This can be used as an alternative definition of A^*: first define its domain this way, then define the image A^*y as the Riesz vector of the linear continuous mapping obtained by extending $x \to (y, Ax)_Y$ to the closure of dom(A), i.e., all X, by continuity. ◊

If dom(A) is not dense, we can always consider A as being of type $X^0 \to Y$, where X^0 is the closure of dom(A) (equipped with the same scalar product as X, by restriction), and still be able to define an adjoint, now of type $Y \to X^0$.

Note that $(\mathcal{A}^\perp)^\perp$ is the closure of $\mathcal{A}$. Therefore, if an operator A has an adjoint, and if dom(A*) is dense, the closure of A is A**, the adjoint of its adjoint. Therefore,

Proposition A.2. *Let* $A : X \to Y$ *be a linear operator with dense domain. If* dom(A*) *is dense in* Y, A *is closable.*

Its closure is then A**. This is how we proved that div was closable, in Chapter 5: The domain of its adjoint is dense because it includes all functions $\varphi \in C_0^\infty(D)$. Indeed, the map $b \to \int_D \varphi \,\mathrm{div}\, b \equiv -\int_D b \cdot \mathrm{grad}\, \varphi$ is $\mathbb{L}^2$-continuous for such a φ, due to the absence of a boundary term. As we see here, the weak divergence is simply the adjoint of the operator $\mathrm{grad} : C_0^\infty(D) \to \mathbb{C}_0^\infty(D)$, the closure of which in $L^2(D) \times \mathbb{L}^2(D)$, in turn, is a *strict* restriction (beware!) of the weak gradient.

The reader is invited to play with these notions, and to prove what follows: The boundary of D being partitioned $S = S^h \cup S^b$ as in the main chapters, start from grad and – div, acting on smooth fields, but restricted to functions which vanish on S^h and to fields which vanish on S^b, respectively. Show that their closures (that one may then denote grad_h and $-\mathrm{div}_b$) are mutual adjoints. Same thing with rot_h and rot_b.

REFERENCES

[AB] Y. Aharonov, D. Bohm: "Significance of Electromagnetic Potentials in the Quantum Theory", **Phys. Rev., 115** (1959), pp. 485–491.

[Bo] R.P. Boas: "Can we make mathematics intelligible?", **Amer. Math. Monthly, 88** (1981), pp. 727–731.

[B1] A. Bossavit: "The Exploitation of Geometrical Symmetry in 3-D Eddy-currents Computations", **IEEE Trans., MAG-21**, 6 (1985), pp. 2307–2309.

[B2] A. Bossavit: "Boundary value problems with symmetry, and their approximation by finite elements", **SIAM J. Appl. Math., 53,** 5 (1993), pp. 1352–1380.

[BS] W.E. Brittin, W.R. Smythe, W. Wyss: "Poincaré gauge in electrodynamics", **Am. J. Phys., 50**, 8 (1982), pp. 693–696.

[Bu] W.L. Burke: **Applied Differential Geometry**, Cambridge University Press (Cambridge, UK), 1985.

[Co] F.H.J. Cornish: "The Poincaré and related gauges in electromagnetic theory", **Am. J. Phys., 52,** 5 (1984), pp. 460–462.

[C&] Y. Crutzen, N.J. Diserens, C.R.I. Emson, D. Rodger: **Proc. European TEAM Workshop on Electromagnetic Field Analysis** (Oxford, England, 23–25 April 1990), Commission of the European Communities (Luxembourg), 1990.

[Ha] P.R. Halmos: **Measure Theory**, Van Nostrand (Princeton), 1950.

[Hl] P.R. Halmos: **Naive Set Theory**, Van Nostrand (Princeton), 1960.

[Hn] P. Henderson: **Functional Programming**, Prentice-Hall (Englewood Cliffs, NJ), 1980.

[Hr] R. Hersh: "Math Lingo vs. Plain English: Double Entendre", **Amer. Math. Monthly, 104**, 1 (1997), pp. 48–51.

[It] K. Ito (ed.): **Encyclopedic Dictionary of Mathematics** (2nd ed.), The MIT Press (Cambridge, MA), 1987.

[KB] A. Kaveh, S.M.R. Behfar: "Finite element ordering algorithms", **Comm. Numer. Meth. Engng., 11,** 12 (1995), pp. 995–1003.

[Kn] E.J. Konopinski: "What the electromagnetic vector potential describes", **Am. J. Phys., 46**, 5 (1978), pp. 499–502.

[Kr] J.-L. Krivine: "Fonctions, programmes et démonstration", **Gazette des Mathématiciens,** 60 (1994), pp. 63–73.

[Lo] M. Loève: **Probability Theory**, Van Nostrand (Princeton), 1955.

[Ma] I. Mayergoyz: **Nonlinear Diffusion of Electromagnetic Fields,** Academic Press (Boston), 1998.

[MU] S. Mazur, S. Ulam: "Sur les transformations isométriques d'espaces vectoriels normés", **C.R. Acad. Sci. Paris, 194** (1932), pp. 946–948.

[Me] B. Meyer: **Object-oriented Software Construction**, Prentice Hall (New York), 1988.

[Mr] B. Meyer : **Introduction to the Theory of Programming Languages**, Prentice-Hall (New York), 1990.

[NC] R.D. Nevels, K.J. Crowell: "A Coulomb gauge analysis of a wave scatterer", **IEE Proc.-H, 137**, 6 (1990), pp. 384–388.

[Pr] J.D. Pryce: **Basic Methods of Linear Functional Analysis**, Hutchinson & Co, Ltd. (London), 1973.

[R&] K.R. Richter, W.M. Rucker, O. Biro (Eds.): **4th IGTE Symposium & European TEAM 9** (Graz, Austria, 10–12 Oct. 1990), Technische Universität Graz (Graz), 1990.

[Sc] B. Schutz: **Geometrical Methods of Mathematical Physics**, Cambridge University Press (Cambridge, U.K.), 1980.

[Sk] B.-S.K. Skagerstam: "A note on the Poincaré gauge", **Am. J. Phys., 51,** 12 (1983), pp. 1148–1149.

[Sp] M. Spivak: **Calculus on Manifolds**, Benjamin, (New York), 1965.

[T&] L. Turner, H. Sabbagh et al. (Eds.): **Proceedings of the Toronto TEAM/ACES Workshop at Ontario Hydro** (25–26 Oct. 1990), Report ANL/FPP/TM-254, the Fusion Power Program at Argonne Nat. Lab., Argonne, Ill., 60439–4814.

[Yo] K. Yosida: **Functional Analysis**, Springer-Verlag (Berlin), 1965.

APPENDIX B

LDL^t Factorization and Constrained Linear Systems

Although the standard variational approach to magnetostatics led to a standard linear system of the form $Ax = b$, with A symmetric regular and positive definite, we had to realize that discrete models do not automatically come in this form, but rather constitute what we called "constrained linear systems". So there is a gap, however small, between equations as they emerge from the modelling and numerical methods as proposed by textbooks and software packages. Whether this gap is negligible or not, and how to bridge it in the latter case, are important issues. This appendix is an approach to this problem from the side of *direct* methods, based on Gaussian factorization, such as LDL^t. Some facts about the LDL^t method and its programming are recalled, and we find the adaptation to constrained linear systems feasible, if not totally straightforward.

B.1 NONNEGATIVE DEFINITE MATRICES

A standard result about the factorization of matrices is: A *positive definite* matrix A (see Def. B.1 below), not necessarily symmetric, *has an* LDM^t *decomposition,* i.e., one can express A as the product LDM^t, where D is diagonal, with strictly positive entries, and L and M are unit lower triangular (i.e., with all diagonal entries equal to 1). See for instance [GL], p. 86, for a proof. A corollary is Gauss's LU decomposition: $A = LU$, obtained by setting $U = \mathrm{DM}^t$.

There is a need, however, for an analogous result that would hold under the weaker assumption of *semi-positive* or *nonnegative* definiteness (Def. B.2 below): Several times, and notoriously in the case of the rot–rot equation, we found the system's matrix nonnegative definite, but not regular, because of the non-uniqueness of potentials representing the same field. Our matrices were also symmetric, so we shall make this assumption, although this is not strictly necessary. Then, $M = L$. As we shall see, the LDL^t factorization is an effective tool for the treatment of constrained linear systems of this category. The reason for this lies in a mathematical result, which we shall prove: *Non-negative definite symmetric matrices are the LDL^t-factorizable matrices with* $D \geq 0$ (i.e., all entries of D nonnegative).

As in the main text, we denote by V_n the real vector space of dimension n, but the boldface convention, pointless here, is shunned, and the scalar product is denoted either by (x, y) or by $x^t y$, where the superscript t stands for "transpose". Observe that xy^t is an $n \times \nu$ matrix, called the *dyadic product* of x by y.

For any $n \times n$ matrix A, we set $\ker(A) = \{x \in V_n : Ax = 0\}$ and $\operatorname{cod}(A) = \{Ax : x \in V_n\}$, i.e., the image of V_n under the action of A, also called the range of A. Let us recall that $\ker(A)$ and $\operatorname{cod}(A^t)$ are mutually orthogonal complementary subspaces of V_n, a fact which is expressed as

$$(1) \qquad V_n = \ker(A) \oplus \operatorname{cod}(A^t) \equiv \ker(A^t) \oplus \operatorname{cod}(A).$$

Definition B.1. *An* $n \times n$ *matrix* A *(with real entries) is said to be* positive definite *if*

$$(2) \qquad x^t A x > 0 \quad \forall x \in V_n, \ x \neq 0.$$

Definition B.2. *Matrix* A *is* nonnegative definite *if*

$$(3) \qquad x^t A x \geq 0 \quad \forall x \in V_n.$$

A positive definite matrix must be regular. Beware, a regular non-definite matrix may fail to satisfy $x^t A x \neq 0$: for instance, matrix $I = \{\{0, 1\}, \{-1, 0\}\}$ is regular, but also skew-symmetric, and hence $x^t I x = 0$ for all x. However,

Proposition B.1. *Among* symmetric *matrices, positive definite matrices are the* regular *nonnegative definite matrices.*

Proof. It's the same proof we did in Section 3.1. Assume (3), and suppose $x^t A x = 0$ for some $x \neq 0$. Then A cannot be regular, because for all $y \in V_n$ and all $\lambda \in \mathbb{R}$,

$$0 \leq (x + y)^t A (x + y) = 2\lambda \, y^t Ax + \lambda^2 y^t A y,$$

hence $y^t Ax = 0 \ \forall y$ (divide by λ, and let it go to 0), and hence $Ax = 0$. Thus, (3) and regularity, taken together, imply (2) in the case of symmetric matrices. ◊

From this point on, let us restrict ourselves to symmetric matrices. Let us recall the following:

Definition B.3. *An* $n \times n$ *matrix* A *is* Cholesky-factorizable *if there exists an upper triangular* $n \times n$ *matrix* S *such that* $A = S^tS$.

Such a matrix is symmetric and, obviously, nonnegative definite. Conversely,

Proposition B.2. *Nonnegative definite, symmetric matrices are Cholesky-factorizable.*

Proof. The proof is by recurrence on the order n. Let us write A, by rows of blocks, as $A = \{\{c, b^t\}, \{b, C\}\}$, where C is of order $n - 1$, and look for its factorization in the form

$$A = \begin{vmatrix} c & b^t \\ b & C \end{vmatrix} = \begin{vmatrix} a & 0 \\ d & T \end{vmatrix} \begin{vmatrix} a & d^t \\ 0 & T^t \end{vmatrix} = SS^t,$$

where T is lower triangular of order $n - 1$. If (3) holds, then, for any $(n - 1)$-vector y and any scalar z,

$$\text{(4)} \qquad y^t\, C\, y + 2\, y^t b\, z + c\, z^2 \geq 0.$$

If $c = 0$, this shows that $y^tb = 0$ for all $y \in V_{n-1}$, hence $b = 0$. Then $a = 0$ and $d = 0$, and since C is symmetric and nonnegative definite (take $x = \{0, y\}$ in (3)), we do have $C = TT^t$ by the recurrence hypothesis. Therefore, $S = \{\{0, 0\}, \{0, T\}\}$ (by rows of blocks). If $c > 0$, set $a = \sqrt{c}$, which forces $d = b/a$, and requires $C = c^{-1}bb^t + TT^t$, so all we have to do is show that $C - c^{-1}bb^t$ is nonnegative definite. This again results from (4), by setting $z = c^{-1}b^ty$. ◊

Remark B.1. A priori, ker(A) contains $\ker(S^t)$. But if $A\,x = 0$, then S^tx belongs to ker(S), hence $S^tx \perp \operatorname{cod}(S^t)$ after (1), which implies $S^tx \perp S^tx$, and therefore $S^tx = 0$. So $\ker(S^t) = \ker(A)$. ◊

Remark B.2. Diagonal terms of S are all nonnegative, with the foregoing choice of a. Note that whenever one of them vanishes, the whole column below it must vanish, too. ◊

Remark B.3. The proof works just as well if entries are complex (note that A is *not* Hermitian, then), provided $\mathrm{Re}\{x^t\, A\, x^*] \geq 0$ for all complex vectors x. *Bisymmetric* matrices (matrices such that both real and imaginary parts are symmetric) are thus LDL^t-factorizable under this hypothesis. This is relevant to the eddy-current problem of Chapter 8. ◊

Now what about the LDL^t factorization? Thanks to Remark B.2, we obtain it by the following sequence of assignments, where s_i, ℓ_i, and 1_i

denote the ith column of S, L, and the unit matrix respectively, p is real, and the ith component of some vector v is v^i:

```
forall $i \in [1, n]$ do
|     $p := s_i^{\,i}$; $d^i := p * p$;
|     $\ell_i :=$ if $p > 0$ then $s_i/p$ else $1_i$
```

This yields a diagonal matrix D, with nonnegative entries d^i, and a unit lower triangular matrix L. Clearly, $A = LDL^t$.

Remark B.4. A straightforward adaptation of the proof would lead us to the result that *nonnegative definite matrices are the* LDM^t-*factorizable matrices.* ◊

One may then solve $Ax = b$, provided b is in the range of A. First, compute $y = L^{-1}b$, then execute the code

```
forall $i \in [1, n]$ do
|     $z^i :=$ if $d^i \neq 0$ then $y^i/d^i$ else $0$,
```

and finally, solve $L^t x = z$. If b Œ cod(A), one *must* have $y^I = 0$ if $d^i = 0$. The choice $z^i = 0$ in such cases is arbitrary: It selects *one* of the solutions of $Ax = b$, and thus constitutes a gauging procedure. By referring to p. 182, one will see how this applies to the rot–rot equation.

Let us now face the question of how to turn these results into a practical algorithm. The problem is, of course, imperfect arithmetic: In spite of the theoretical proof that successive "pivots", i.e., the values of p, will all be nonnegative, there is no guarantee that small negative values or (perhaps worse, because the trouble is harder to spot and to cure) very small positive but non-zero values, will not appear.

B.2 A DIGRESSION ABOUT PROGRAMMING

The question belongs to the immense realm of program correctness in presence of floating-point computations [CC]. Some notions on program construction will help in the discussion.

Let's adhere to the discipline of object-oriented programming [Me]. We deal with abstract data types called *INTEGER, REAL, VECTOR, MATRIX,* etc. (a construct such as m : *MATRIX,* for instance, means that m is a program object of type *MATRIX)* and with operations such as

order : MATRIX → INTEGER,

column : MATRIX × INTEGER → VECTOR,

row : MATRIX × INTEGER → VECTOR,

$$\mathit{component} : \ \mathit{VECTOR} \times \mathit{INTEGER} \rightarrow \mathit{REAL},$$

$$\mathit{length} : \ \mathit{VECTOR} \rightarrow \mathit{INTEGER}\,,$$

and so forth. The idea is to program in terms of such operations exclusively. (Their practical *implementation*, of course, may require operations of lower level, those that are available in the target programming language.)

The formal definition of the universe of types and operations we need is quite a large job, and I don't attempt it. Let us just agree upon a few notational conventions: *component*(v, i) is abbreviated as v^i, *length*(v) as $|v|$, *column*(a, j) as a_j, *row*(a, i) as a^i, etc. This way, a_j^i refers to the entry on row i and column j of matrix a. We also have the ordinary multiplication $*$, which can be considered as acting either on a *SCALAR* and a *VECTOR*, or on a pair of *VECTORs*, this way:

$$* : \ \mathit{SCALAR} \times \mathit{VECTOR} \rightarrow \mathit{VECTOR},$$

$$* : \ \mathit{VECTOR} \times \mathit{VECTOR} \rightarrow \mathit{VECTOR},$$

and is defined as one may guess: $\lambda * v$ is the vector of components $\lambda * v^i$, that is, formally, $(\lambda * v)^i = \lambda\, v^i$, and for two vectors u and v, $(u * v)^i = u^i\, v^i$.

Just for practice (and also in order to introduce without too much fuss some syntactical conventions about programs), let us code a matrix-vector multiplication within this universe of types:

```
program mat–mul–vec (in a: MATRIX, x : VECTOR {order(a)
|       = length(x)}, out y : VECTOR )
|       local n: INTEGER, n := length(x) ;
|       for j := 1 to n do
|       |       $y := y + x^j * a_j$
```

Note the *assertion* between curly brackets (the program is not supposed to work if this input assertion is not satisfied). Mere *comments* also come within such brackets.

Once *mat–mul–vec* has thus been written, one may denote by $a * x$ the returned vector y. This overloading of $*$ is harmless and can be reiterated after the eventual construction of other similar programs like *mat–mul–mat*, etc. Along with other obvious operations, like *diag : VECTOR → MATRIX*, *transp : MATRIX → MATRIX*, etc., all these operations contribute to the step-by-step construction of an *algebra*, i.e., a consistent and organized universe in which to program [Ba].

Actually, the word "algebra" has connotations which suggest a little more. An algebra of types should be "complete", meaning that when the inverse of some operation can be defined, it is included in the algebra. (Stating this formally is difficult, and I shall not attempt it.) For instance, if *mat–mul–vec* is there to allow multiplication of a matrix a by a vector x, yielding vector y, there should also be something to get x from a and y, say,

$$\textit{solve}(\textbf{in}\ \mathrm{a}: \textit{MATRIX},\ \mathrm{y}: \textit{VECTOR}, \textbf{out}\ \mathrm{x}: \textit{VECTOR}\ \{\mathrm{y} = \mathrm{a} * \mathrm{x}\}),$$

for which the appropriate syntax[1] might be x := a\y. As one knows, this is not a primitive operation, and it requires stepping stones like triangular solvers and (for instance) the LDL^t factorization. Let us therefore return to this.

B.3 THE LDL^t FACTORIZATION

Let us introduce the *outer product* (or *dyadic product)* of vectors,

$$\times : \textit{VECTOR} \times \textit{VECTOR} \to \textit{MATRIX},$$

defined by

$$(\mathrm{u} \times \mathrm{v})^i_j = \mathrm{u}^i\, \mathrm{v}^j.$$

We are looking for a vector d and a lower triangular matrix ℓ such that $\ell^i_i = 1$ and $\ell * \textit{diag}(\mathrm{d}) * \textit{transp}(\ell) = \mathrm{a}$. This specification can be rewritten as

$$\Sigma_{j = 1, \ldots, n}\, \ell_j \times (\mathrm{d}^j * \ell_j) = \mathrm{a},$$

which immediately suggests an algorithm, on the model of the proof of Prop. B.2:

```
program LDLT(in a : MATRIX, out ℓ : MATRIX,
|       d : VECTOR)  {a is nonnegative definite}
|       local c : MATRIX ; c := a ;
|       for j := 1 to order(a)  do
|       |       $d^j := c^j_j$ ; $\ell_j := c_j$ ;
```

[1] This is the syntax of MATLAB [MW]. Obviously, a package such as MATLAB is the implementation of an algebra, in the above sense, and its writing has required at some stage the kind of abstract programming suggested here.

```
|     |     if $d^j > 0$ then {if $d^j = 0$ then $c_j = 0$}
|     |     |     $\ell_j := \ell_j / d^j$ ;
|     |     |     $c := c - \ell_j \times (d^j * \ell_j)$ {$c_j = 0$}
|     |     else $\ell_j^j := 1$ {$\ell_j = 1_j$} ;
|     |     {$c_j = 0$}
|     |{$c = 0$ ; $a = \ell * diag(d) * transp(\ell)$}
```

This is the way it works in perfect arithmetic: The crucial assertion **if** $d^j = 0$ **then** $c_j = 0$ is a consequence of the hypothesis on nonnegative definiteness. (Check this point before reading on, if necessary, by rferring to the proof of Prop. B.2.) Note how column j of ℓ happens to be equal to 1_j when the jth pivot is null.

Now, still with perfect arithmetic, we can do this exactly in the same way for *any* entry a, provided the assertions, now not automatically true, are *enforced* when necessary. Hence the following program, which does the same thing as the previous one when a is nonnegative definite, but still does something when this precondition is not satisfied:

```
program LDLT(in a : MATRIX, out $\ell$ : MATRIX, d : VECTOR)
|     local c : MATRIX ; c := a ;
|     for j := 1 to order(a) do
|     |     $d^j := c_j^j$ ; $\ell_j := c_j$ ;
|     |     if $d^j > 0$ then
|     |     |     $\ell_j := \ell_j / d^j$ ;
|     |     |     $c := c - \ell_j \times (d^j * \ell_j)$
|     |     else $\ell_j := 1_j$
```

But the question now is: Could this algorithm fail to be *stable*? Since we may have to divide by arbitrary small pivots d^j, should we fear uncontrolled growth of some terms, and eventual overflow?

The answer seems to be *no,* provided the standard precaution is taken of implementing the test (**if** $d^j > 0$) as **if** $1. + d^j > 1$. This way, the smallest possible pivot will be the machine–epsilon ε, i.e., the number such that $1 + \varepsilon$ is the machine number next to 1 in the (finite) system of numbers the machine offers as an approximation to the ideal *REAL* type. Since the algorithm is a variant of Cholesky, the classical error analysis by Wilkinson should be relevant, and give similar results (cf. [GL], p. 89). Giving a formal proof of this, however, looks like a tough challenge.

In a further attempt to extend the scope of this program, one may replace the clause **if** $d^j > 0$ by **if** $d^j \neq 0$, and **else** $\ell_j := 1$ by a loop exit. What we get then is a program that, *when it doesn't encounter a null pivot*, returns an LDL^t factorization with terms of both signs in D. Then A = LDL^t is regular. Such a program may be a useful tool,[2] but be aware that regularity of A is no guarantee that it will work to the end without falling on a zero pivot: A simple counter-example is given by A = {{0, 1}, {1, 0}} (by rows).

The reader is invited to complete the coding of *solve* (p. 324) by writing out the triangular solvers, and the division by D. Note how this object-oriented style automatically provides "vectorized" programs [Bo].

B.4 APPLICATION TO CONSTRAINED LINEAR SYSTEMS

If working with potentials leads to nonnegative definite system matrices, working with fields directly generates constrained linear systems, as we have seen. Actually, such systems are rather the rule than the exception in numerical modelling. Let us recall the paradigm: Given a symmetric, nonnegative definite matrix A of order N, an N-vector b, a rectangular matrix B, and a vector c of same height, find x such that (Ax, x) – 2(b, x) be minimized over the affine subspace {x : Bx = c}. By introducing a Lagrange multiplier λ, this problem is transformed into a linear system of the form

$$\begin{vmatrix} A & B^t \\ B & 0 \end{vmatrix} \begin{vmatrix} x \\ \lambda \end{vmatrix} = \begin{vmatrix} b \\ c \end{vmatrix} . \tag{5}$$

If ker(A) ∩ ker(B) = {0}, which we assume, x is unique. There is no loss of generality if we also assume that B is surjective (i.e., ker(B^t) = 0), in which case λ is unique too. The large block-matrix at the left-hand side of (5), M say, is thus regular,even when A is not.

However, standard off-the-shelf packages will not, in general, be able to factor M, in order to solve (5). Though regular, M is certainly not positive definite or even semi-positive definite, so Cholesky is out. The existence of an LDL^t-factorization, on the other hand, is not ruled out a priori, provided both signs are allowed for the entries of D. If A is regular, the modified version of *LDLT*, which accpets negative pivots,

[2] In particular, according to Sylvester's "law of inertia" [Kn], the number of positive entries of D is the number of positive eigenvalues of A. Running the program on A – σ1 thus allows one to count the eigenvalues larger than 0 . . . when the algorithm succeeds.

will work. But otherwise it can fail, as the counter-example $A = \{\{1, 0\}, \{0, 0\}\}$ and $B = \{0, 1\}$ will show: Though regular, the matrix $\{\{1, 0, 0\}, \{0, 0, 1\}, \{0, 1, 0\}\}$ has no LDL^t factorization.

So what is to be done? Remark that $A + B^tB$ is (strictly) positive definite, for $((A + B^tB)x, x) = 0$ implies $(Ax, x) = 0$ and $Bx = 0$, therefore $x \in \ker(A) \cap \ker(B)$. And since $Bx = c$ if x is solution, (5) and the following system are equivalent:

$$\begin{vmatrix} A+B^tB & B^t \\ B & 0 \end{vmatrix} \begin{vmatrix} x \\ \lambda \end{vmatrix} = \begin{vmatrix} b+B^tc \\ c \end{vmatrix}. \tag{6}$$

But now the new (augmented) matrix M is LDL^t-factorizable. Working by blocks to begin with, we get

$$\begin{vmatrix} A+B^tB & B^t \\ B & 0 \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ \beta & 1 \end{vmatrix} \begin{vmatrix} A+B^tB & 0 \\ 0 & -\gamma \end{vmatrix} \begin{vmatrix} 1 & \beta^t \\ 0 & 1 \end{vmatrix}, \tag{7}$$

with $\beta = B(A+B^tB)^{-1}$ and $\gamma = B(A+B^tB)^{-1}B^t$. Assuming the programming environment is an LDL^t package (even the standard one, that assumes $D > 0$, will do), complete with its factorizer and downward and upward triangular solvers, the essential task consists in factorizing $A + B^tB$ and $\gamma = B(A+B^tB)^{-1}B^t$, which are both positive definite, since we have assumed $\ker(B^t) = 0$. The factorization of γ is necessary in order to get λ, by solving

$$B(A + B^tB)^{-1}B^t\lambda = B(A + B^tB)^{-1}(b + B^tc) - c,$$

and that of $A + B^tB$ to obtain x, by solving

$$(A + B^tB)\,x = b + B^t(c - \lambda),$$

hence a solution in two steps. Since these factorizations allow passing from the block form (7) to the full-fledged LDL^t factorization of the augmented M, we may do it all in one stroke by applying the LDL^t package to system (5), provided the program lets negative pivots pass.

Remark B.5. The Lagrangian of (5) was $\mathcal{L}(x, \lambda) = (Ax, x) + 2(\lambda, Bx) - 2(b, x)$, and although strictly convex when x is restricted to ker(B), it was not strictly convex in x. The "augmented" Lagrangian of (6), $\mathcal{L}(x, \lambda) = (Ax, x) + |Bx|^2 + 2(\lambda, Bx) - 2(b, x)$, is. (Note that one may search for its saddle point $\{x, \lambda\}$ by some iterative method, such as the Arrow–Hurwicz–Uzawa algorithm [AH]. We don't discuss this alternative here, having direct methods in view.) One may think of a more general, possibly better form for it: $\mathcal{L}(x, \lambda) = (Ax, x) + \rho\,|Bx|^2 + 2(\lambda, Bx) - 2(b, x)$, where ρ is a positive constant, the optimal value of which depends of course on how B has been built. Note that $\rho = 0$ is allowed if A is regular, the easy case. ◊

Remark B.6. It has been proposed [Ve] that (1) be replaced by the obviously equivalent system

$$\begin{vmatrix} -1 & B & 1 \\ B^t & A & B^t \\ 1 & B & -1 \end{vmatrix} \begin{vmatrix} \mu \\ x \\ \lambda \end{vmatrix} = \begin{vmatrix} c \\ b \\ c \end{vmatrix}, \tag{8}$$

that is, *two* Lagrange multipliers instead of one. The matrix of (6) is indeed LDL^t-factorizable, under the above hypotheses:

$$\begin{vmatrix} -1 & B & 1 \\ B^t & A & B^t \\ 1 & B & -1 \end{vmatrix} = \begin{vmatrix} 1 & & \\ -B^t & 1 & \\ -1 & 2\beta & 1 \end{vmatrix} \begin{vmatrix} -1 & & \\ & A + B^t B & \\ & & -4\gamma \end{vmatrix} \begin{vmatrix} 1 & -B & -1 \\ & 1 & 2\beta^t \\ & & 1 \end{vmatrix},$$

with the same β and γ as above. So one can solve (1) this way, by running an LDL^t package on (8). But the numerical effort involved is no less than was required by the above method (a bit *more*, actually, since some arithmetic is wasted on numerically retrieving the first column block of L, that is $\{1, -B^t, -1\}$, which is already known). As the heuristic leading from (5) to (8) is quite obscure, in comparison with the easily motivated passage from (5) to (6), this "double-multiplier" approach is more of a curiosity than a real alternative. ◊

REFERENCES

[AH] K. Arrow, L. Hurwicz, H. Uzawa: **Studies in Nonlinear Programming**, Stanford U.P. (Stanford), 1958.

[BA] J. Backus: "Can Programming be Liberated from the Von Neumann Style? A Functional Style and its Algebra of Programs", **Comm. ACM, 21,** 8 (1978), pp. 613–641.

[CC] F. Chaitin-Chatelin, V. Fraysse: **Lectures on Finite Precision Computation,** SIAM (Philadelphia), 1996.

[GL] G.H. Golub, C.F. Van Loan: **Matrix Computations**, North Oxford Academic (Oxford) & Johns Hopkins U.P. (Baltimore), 1983.

[Kn] D.E. Knuth: "A permanent inequality", **Amer. Math. Monthly, 88** (1981), pp. 731–740.

[MW] **MATLAB™ for Macintosh Computers, User's Guide**, The MathWorks, Inc. (Natick, Ma, USA), 1991.

[Me] B. Meyer: **Object-oriented Software Construction**, Prentice Hall (New York), 1988.

[Ve] Int. report by M. Verpeaux, CEA-DEMT, Saclay. Cf. J. Pellet: **Dualisation des conditions aux limites**, Document ASTER R3.03.01 (EdF, Clamart), 27 11 91.

APPENDIX C

A Cheaper Way to Complementarity

For reference, let us state again the basic problem of Chapter 6: In domain D, the surface S of which is partitioned as $S^h \cup S^b$, *find among pairs* $\{h, b\}$ *which satisfy*

(1) $\operatorname{rot} h = 0$ in D, (3) $\operatorname{div} b = 0$ in D,

(2) $n \times h = 0$ on S^h (4) $n \cdot b = 0$ on S^b,

(6) $\int_c \tau \cdot h = I,$ (7) $\int_C n \cdot b = F,$

a minimizer for the error in constitutive law

$$E(b, h) = \int_D \mu^{-1} \, |b - \mu h|^2.$$

(Cf. Fig. 6.1 for the definition of the "link" c and the "cut" C.) As we saw, there is a minimizer whatever I and F, with h and b weakly solenoidal and irrotational, respectively, and a unique value of the ratio $R = I/F$ (the reluctance) for which the constitutive law

(5) $b = \mu h$ in D,

is satisfied, i.e., $E(b, h) = 0$.

Complementarity consists in simultaneously solving for $h = \operatorname{grad} \varphi$ by nodal elements for φ and for $b = \operatorname{rot} a$ by edge elements for a, hence (assuming one uses the same mesh m for both, which is not mandatory), a rot-conformal h_m satisfying (1–2)(6) and a div-conformal b_m satisfying (3–4)(7), "m-weakly" solenoidal and irrotational respectively, but not linked by (5). This gave us bilateral bounds for R and (by computing $E(b_m, h_m)$) upper bounds for both approximation errors $\int_D \mu \, |h - h_m|^2$ and $\int_D \mu^{-1} \, |b - b_m|^2$.

Alas, this nice approach has a serious drawback: As we saw in Section 6.3, the part of the computation that yields a is much more expensive[1] than the determination of φ.

Therefore, it would be interesting to be able to save on this effort, in the case of the a-method, by making good use of the information one has, once h_m has been determined. In quite fuzzy terms for the moment—but this will become more and more precise—can the solution in terms of φ somewhat be corrected to yield a truly solenoidal (not only m-solenoidal) approximation of b?

C.1 LOCAL CORRECTIONS

So let's suppose we have computed h_m, satisfying Eqs. (1), (2), and (6), and such that

$$(8) \qquad \int_D \mu\, h_m \cdot \operatorname{grad} \lambda^n = 0 \;\; \forall\, n \in \mathcal{N}_h,$$

where λ^n (preferred in this Appendix to w_n, for notational uniformity) is the barycentric function of node n, and $\mathcal{N}_h = \mathcal{N} - \mathcal{N}(S^h)$ the set of nodes not included in S^h. We want some $b \in W^2_m$, divergence-free, and—in order to make use of the knowledge of the solution we have already acquired—close to h_m.

What we have done in Chapter 6 seems to give an obvious solution: Look for a minimizer of the error in constitutive law,

$$(9) \qquad b = \operatorname{arginf}\{\int_D \mu^{-1} \,|b' - \mu h_m|^2 : \; b' \in \mathbb{B}^F_m\},$$

where $\mathbb{B}^F_m = \{b \in W^2_m(D) : \operatorname{div} b = 0, \; n \cdot b = 0 \text{ on } S^b, \; \int_C n \cdot b = F\}$. Vector fields in this space are linear combinations of face elements,

$$(10) \qquad b = \sum_{f \in \mathcal{F}_b} \mathbf{b}_f \, w_f,$$

where $\mathcal{F}_b$ abbreviates $\mathcal{F}(S - S^b)$, the set of faces not included in S^b. (This way, (10) implicitly takes the no-flux condition (4) into account.) But the remaining nonzero face-DoFs $\mathbf{b}_f$ are not independent for $b \in \mathbb{B}^F_m$. They are constrained by linear relations:

[1] Just for practice, let's do it again, this time with the ratio T/N equal to 6. Thanks to the Euler–Poincaré formula, one has $E \sim 7N$ and $F \sim 12\,N$. The average number of faces that contain a given edge is $3F/E$, so each edge has $9F/E$ "neighbors", if one defines as neighbors two edges that belong to a common tetrahedron. The number of off-diagonal entries of the edge-element matrix is thus $9F$, that is, $108\,N$, against $14\,N$ for the matrix created by the φ-method.

$$\sum_{f \in \mathcal{F}(T)} \mathbf{D}_{Tf}\, \mathbf{b}_f = 0, \qquad \sum_{f \in \mathcal{F}(C)} \mathbf{D}_{C\,f}\, \mathbf{b}_f = F,$$

where $\mathcal{F}(C)$ is the set of faces that pave the cut C, and $\mathbf{D}_{C\,f} = \pm 1$ according to relative orientation. As $\mathbb{B}^F_m = \text{rot } A^F_m$, (9) is equivalent to finding a minimizer

$$\text{(11)} \qquad a_m \in \text{arginf}\{\int_D \mu^{-1} \,|\text{rot a} - \mu h_m|^2 : \ a \in A^F_m\},$$

and there is no difference between doing that and directly solving for a by edge elements. The b = rot a thus obtained, which is the approximation b_m of Chapter 6, is indeed the closest to h_m in energy. But no use is made of the knowledge of h_m this way.

Remark C.1. Problem (11) is the same as problem (6.21): Since $\int_D \mu^{-1} \,|\text{rot a} - \mu h_m|^2 = \int_D \mu^{-1} \,|\text{rot a}|^2 - 2\int_D h_m \cdot \text{rot a} + \int_D \mu \,|h_m|^2$ and (by Lemma 6.1) $\int_D h_m \cdot \text{rot a} = 0$, the two functionals in (11) and (6.21) differ by a constant, and minimization is performed on the same subspace. ◊

We now introduce the *localization* heuristics. Let's have a partition of unity over D, i.e., a family of piecewise-smooth functions χ^i, indexed over some finite set $\mathcal{J}$, and satisfying $\sum_{i \in \mathcal{J}} \chi^i = 1$. Any b can be written as a sum $b = \sum_{i \in \mathcal{J}} \chi^i b \equiv \sum_{i \in \mathcal{J}} b^i$. For one that suits our needs (divergence free, and close to h_m), each b^i should satisfy $\text{div } b^i = \text{div}(\chi^i b) = b \cdot \text{grad}\, \chi^i$, and should be close to $\chi^i \mu h_m$. Not knowing b, we replace $b \cdot \text{grad}\, \chi^i$ by the next best thing, which is $\mu h_m \cdot \text{grad}\, \chi^i$, and try to achieve $\text{div } b^i = \mu h_m \cdot \text{grad}\, \chi^i$ as best we can, while looking for b^i in W^2. Since then $\text{div } b^i$ belongs to W^3, and is thus mesh-wise constant, the best we can do is to request

$$\text{(12)} \qquad \int_T \text{div } b^i = \int_T \mu\, h_m \cdot \text{grad}\, \chi^i \quad \forall\, T \in \mathcal{T},$$

for all indices i. Besides that, we also want b^i as close as possible, in energy, to $\chi^i \mu h_m$, hence

$$\text{(13)} \qquad b^i = \text{arginf}\{\int_D \mu^{-1} \,|b' - \mu \chi^i h_m|^2 : \ b' \in \mathbb{B}^i(h_m)\},$$

where $\mathbb{B}^i(h_m)$ is an ad-hoc and provisional notation for the set of b^is in $\mathbb{B}^F_m$ that satisfy the constraints (12). Intuitively (and we'll soon confirm this), computing b^i is a *local* procedure. (Notice that $\text{div } b = \text{div}(\sum_i b^i) = 0$, by summing (12) over i.) This is the principle.

For its formal application, now, let us call $\mathcal{T}^i$ the set of tetrahedra whose intersection with $\text{supp}(\chi^i)$ has a nonzero measure, D^i their set union, and $\mathcal{F}^i$ the collection of all faces of such tetrahedra *except* those contained in the boundary ∂D^i (but not in S^h, cf. Fig. C.1). In (13),

$\mu\,\chi^i h_m = 0$ outside D^i, so we may search b^i among the restricted set of fields that vanish outside D^i, which means (since by normal continuity of b^i, its normal component on ∂D^i must be null) those of the form $\sum_{f \in \mathcal{F}^i} \mathbf{b}_f w_f$. Let us therefore introduce the notation

$$W^2_m(D^i) = \{b \in W^2_m(D) : \ b = \sum_{f \in \mathcal{F}^i} \mathbf{b}_f w_f\},$$

and redefine $\mathbb{B}^i(h_m)$ as

$$\mathbb{B}^i(h_m) = \{b \in \mathbb{B}^F_m \cap W^2_m(D^i) : \int_T \operatorname{div} b^i = \int_T \mu h_m \cdot \operatorname{grad} \chi^i \ \ \forall\, T \in \mathcal{T}^{\,i}\}.$$

The b^is are given by

$$\text{(14)} \qquad b^i = \operatorname{arginf}\{\int_{D^i} \mu^{-1} \,|b' - \mu\chi^i h_m|^2 : \ b' \in \mathbb{B}^i(h_m)\},$$

which differs from (13) only by the integration domain being D^i instead of D.

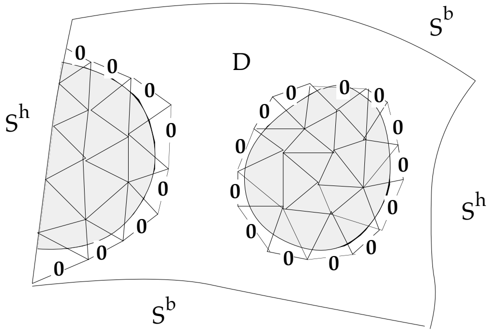


FIGURE C.1. Two examples showing the relation between $\operatorname{supp}(\chi^i)$ (shaded) and $\mathcal{T}^{\,i}$. Faces of $\mathcal{F}^i$ (appearing as edges in this 2D drawing) are those not marked with a 0.

Before going further, let us point to an easily overlooked difficulty: If D^i does not encounter S^h (we call $\mathcal{J}_h$ the subset of $\mathcal{J}$ for which this happens), then $\int_{D^i} \operatorname{div} b^i = 0$. So unless

$$\text{(15)} \qquad \int_{D^i} \mu h_m \cdot \operatorname{grad} \chi^i = 0 \ \ \forall\, i \in \mathcal{J}_h,$$

some of the affine sets $\mathbb{B}^i(h_m)$ may well be empty! Fortunately, there are easy ways to enforce condition (15). One is to use the barycentric functions as partition of unity, and then $\mathcal{J}_h$ coincides with $\mathcal{N}_h$, so (15) is equivalent to (8), which is indeed satisfied if h_m was computed by the φ-method. More generally, if all the χ^is are linear combinations of the λ^ns, which we shall assume from now on, (15) holds, and we are clear.

C.2 SOLVING PROBLEM (14)

The previous remark sheds some light on the algebraic structure of Problem (14), which determines b^i. The number of unknowns is the number of faces in $\mathcal{F}^i$, say F^i. There are T^i tetrahedra in $\mathcal{T}^i$, hence T^i constraints, but only $T^i - 1$ of those are independent, owing to (15), in the general case where $i \in \mathcal{J}_h$. This leaves $F^i - T^i + 1$ face DoFs with respect to which to minimize the energy error in (14). In the case of connected simply connected regions D^i, one has

$$N^i - E^i + F^i - T^i = -1,$$

where N^i and E^i are the numbers of *inner* nodes and edges in D^i (those not on ∂D^i), so the number of independent DoFs is $E^i - N^i$.

To go further in the identification of these DoFs, let $b^i(h_m)$ be some member of $\mathbb{B}^i(h_m)$, constructed by a procedure, the description of which we defer for an instant. Then $\text{div}\, b^i = \text{div}\, b^i(h_m)$ in D^i, hence $b^i = b^i(h_m) +$ rot a^i, with

$$(16) \qquad a^i = \sum_{e \in \mathcal{E}^i} \mathbf{a}^i_e w_e \,,$$

a linear combination of edge elements, indexed over the E^i inner edges of D^i. The a^is which are gradients yield rot $a^i = 0$, and the dimension of the subspace they span is N^i, so we have indeed $E^i - N^i$ independent degrees of freedom, as far as b^i is concerned. (One might use the N^i loose ones to "gauge" a^i. But this is not necessary, as stressed in Chapter 6.)

Let us now rewrite problem (14) in terms of a^i. We have

$$(17) \qquad a^i \in \text{arginf}\{\textstyle\int_{D^i} \mu^{-1} \,|\, \text{rot a}' + b^i(h_m) - \mu\chi^i h_m \,|^2 : \ a' \in A^i\},$$

where A^i is the set of fields of the form (16). When $\chi^i = \lambda^i$, the barycentric function of node number i, this simplifies a little, for

$$\int_{D^i} \mu^{-1} \,|\, \text{rot a}' + b^i(h_m) - \mu\lambda^i h_m \,|^2 = \int_{D^i} \mu^{-1} \,|\, \text{rot a}' + b^i(h_m) \,|^2$$
$$+ \int_{D^i} \mu \,|\, \lambda^i h_m \,|^2 - 2\int_{D^i} \lambda^i \text{ rot a}' \cdot h_m,$$

and the latter term is $\frac{1}{2} \int_{D^i} \text{rot a}' \,.\, h_m \equiv 0$, because both rot a' and h_m are piecewise constant, and the average value of λ^i is 1/4 over all tetrahedra. The term to be minimized is then[2] $\int_{D^i} \mu^{-1} \,|\, \text{rot a}' + b^i(h_m) \,|^2$.

[2] This is actually true in all cases when the χ^is are linear combinations of the λ^ns.

This only leaves the problem of determining $b^i(h_m)$. Refer back to (12), which says, equivalently, that

$$\int_{\partial T} n \cdot b^i = \int_{\partial T} \mu \chi^i\, n \cdot h_m \quad \forall\, T \in \mathcal{T}^{\,i}.$$

Select $T^i - 1$ faces in such a way that no more than three of them belong to the same tetrahedron, and attribute to them the DoF

$$\text{(18)} \qquad \mathbf{b}^i_f = \tfrac{1}{2} \textstyle\sum_{T \in \mathcal{T}^{\,i}} \int_f \mu \chi^i\, n \cdot {}^T h_m,$$

where ${}^T h_m$ is the value of h_m over T. (Only two tetrahedra give nonzero contributions to the sum, those sharing f, so $\mathbf{b}^i_f$ is the average flux through f.) Other DoFs will be determined from the linear relations (12), now in the right number.

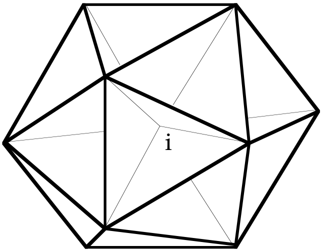


FIGURE C.2. A cluster of 20 tetrahedra around node i, with opaque inner faces and transparent surface faces. There are 12 inner edges, as many surface nodes, and 30 inner faces, in one-to-one correspondence with surface edges ($E^i = 12$, $F^i = 30$, $T^i = 20$).

As for the selection of faces, we shall describe this process only in the case where $\mathcal{J} \equiv \mathcal{N}$ and $\chi^i \equiv \lambda^i$. Then $N^i = 1$, and D^i is the cluster[3] of tetrahedra around node number i (cf. Fig. C.2). Its surface is a polyhedron with $E^{\,i}$ nodes (the number of inner edges), $F^{\,i}$ edges (for each inner face corresponds to a boundary edge) and $T^{\,i}$ faces (the number of tetrahedra in the cluster). By Euler–Poincaré, $E^{\,i} - F^{\,i} + T^{\,i} = 2$ (don't be confused by the notational shift), and since each boundary face has three edges, $F^{\,i} = 3T^{\,i}/2$. One has thus $T^{\,i} = 2E^i - 4$ (typically 20, on the average, if we assume 5 tetrahedra per node) and $F^{\,i} = 3E^i - 6$ (typically, 30). (This fits: $T^i - 1$ independent constraints, $E^i - 1$ "genuine" edge-DoFs, and $19 + 11 = 30$.) We must select $T^{\,i} - 1$ edges (out of $F^{\,i}$), leaving out $E^{\,i} - 1$. This is done by extracting a spanning tree from the graph, the nodes of which are the *faces* of ∂D^i (*not* the surface nodes!) with the surface edges as edges

[3] French readers, please use a French name, "grappe" or "agrégat", to translate cluster.

of the graph. I suppose a drawing could help: cf. Fig. C.3. Co-edges of the spanning tree point to the faces for which the computation (18) will be done (those which cut the surface along the thick lines of Fig. C.3).

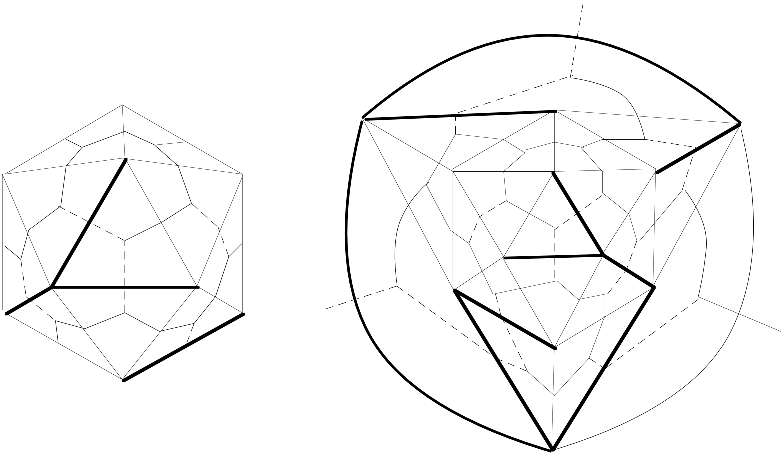


FIGURE C.3. Complementary spanning trees on the surface of an icosahedron, as seen in perspective view (left) and in stereographic projection from the center of the back face (right). The latter view shows only 19 faces, three of them with some distortion, as if one was peering inside the icosahedron from a point near the center of face 20 (the rear one on the left view), which thus corresponds to the outer region of the plane in the stereographic projection. The spanning tree of the face-to-edges graph of the text is in thin lines. Co-edges of this tree are in dotted lines; they intersect the thick-drawn edges of the polyhedron. Note that these edges themselves constitute a spanning tree for the node-to-edges tree on the surface. (This "complementarity" of the trees of both kinds is a general fact on closed simply connected surfaces; cf. Remark C.3.)

After this, Problem (17) amounts to solving a linear system of order E^i, with twice as many off-diagonal terms as there are edges on the surface of the cluster, that is, $2F^i = 2(3E^i - 6)$, typically 72. There are about N such problems (more precisely, N^h, the number of nodes in $\mathcal{N}^h$), so if we regroup them in a single large matrix, the latter will be block-diagonal (about N blocks), and contain about 60 N off-diagonal terms. This compares favorably with the 90 N we found in Chapter 6 for this typical mesh,[4] to say nothing of the intrinsic parallelism of the method.

[4] With T/N ~ 6, as in Note 1, typical figures are $E^i = 12$, $F^i = 36$, $T^i = 24$. One has then 72 N off-diagonal terms vs the 108 N computed in Note 1—the same ratio.

Remark C.2. All this cries out for some symmetrization: Suppose b_m has been obtained by the a-method. Could a local correction to $\mu^{-1}h_m$ be built, quite similarly, which would yield a curl-free companion to b_m? Indeed, this is straightforward, and the reader will easily do it by imitating all we have been doing up to now, systematically transposing h for b, rot for grad, etc. Economy is no more a factor there, since the global φ-method is likely to be cheaper than all alternatives, but the need for local corrections may arise in mesh-refinement procedures the same way. And in this respect, even though the a-method is not the preferred one in magnetostatics, due to its intrinsically high cost, we may hope a method elaborated in this context will transpose to the more general one of *eddy-current* problems, in which edge-elements and the curl-curl equation are natural ingredients, and the scalar potential method is not an option. A mesh-refinement procedure based on the present ideas would then take all its value. ◊

Remark C.3. Let's explain the intriguing "tree-complementarity" of Fig. C.3. (This will illustrate what was said in 5.3.3 about the relevance of graph-theoretical concepts.) Start from a spanning tree of the so-called "primal" graph, nodes to edges. Across each co-edge, there is a line joining the two nearby triangles. These lines form a subgraph of the "dual" graph (the faces-to-edges one). This "co-edge subgraph" visits all triangles, because no triangle can have all its edges in the spanning tree, since that would make a loop. It's connected, because the complement of the spanning tree is. It has no loops, because a loop would disconnect the primal spanning tree. (On other surfaces than spheres, the situation is completely different, as we observed in Section 5.3, Fig. 5.9.) ◊

C.3 CONCLUSION AND SPECULATIONS

We achieved our objective, which was to find some b, divergence-free, and close to μh, with moderate computational effort. In practice, one will thus solve for the scalar magnetic potential φ, hence h = grad φ, and compute b by the above procedure. Then everything one may wish to know about the error relative to this mesh is told by the estimator

$$(19) \qquad E(b, h) = \sum_{T \in \mathcal{T}} \int_T \mu^{-1} \, |b - \mu\, h|^2,$$

which can be used to get bilateral bounds on the reluctance, or to map the local error in constitutive law. The approximation error with *this* mesh is therefore well documented. If it is found too large, a mesh refinement, by appropriate subdivision of the "guilty" elements and their neighbors, cannot fail to improve the result.

But the estimator (19) *could,* conceivably, fail to register this fact when recomputed on the refined mesh. Let's call m' the new mesh, and denote by $m' < m$ the fact that m' is a refinement of m. That $h_{m'}$ is "better" than h_m is no proof that $E(b_{m'}, h_{m'}) < E(b_m, h_m)$, if b_m is computed by local correction, since the partition of unity has changed with the mesh. On the other hand, this equality would hold if b_m was computed by the "expensive", nonlocal edge-element procedure. We shall denote by $b^L{}_m$ and $b^G{}_m$ (for "local" and "global") these two approximations.

Hence the following question: Does the approximation $b^L{}_m$ converge, like $b^G{}_m$, towards the true b? The answer is likely to be *yes*, but the issue is not yet settled.

To make progress in this direction, let us establish two lemmas (useful by themselves):

Lemma C.1. *One has* $\sum_{i \in \mathcal{N}} \mathrm{vol}(D^i) = 4\,\mathrm{vol}(D)$, *where* vol *denotes the volume.*

Proof.
$$\sum_{i \in \mathcal{N}} \mathrm{vol}(D^i) = \sum_{i \in \mathcal{N}} \sum_{T \in \mathcal{T}(i)} \mathrm{vol}(T) = \sum_{T \in \mathcal{T}(i)} \sum_{i \in \mathcal{N}} \mathrm{vol}(T)$$
$$= 4 \sum_{T \in \mathcal{T}(i)} \mathrm{vol}(T) = 4\,\mathrm{vol}(D),$$

since each tetrahedron has four nodes. ◊

Lemma C.2. *Let* k *be the maximum number of edges adjacent to a node. Then, for a given vector field* u,

$$(20) \qquad \int_D |u|^2 \le (k+1) \sum_{i \in \mathcal{N}} \int_D |\lambda^i u|^2.$$

Proof. Let us set $u^i = \lambda^i u$. Then, for any definite node ordering,

$$\int_D |u|^2 = \int_D |\sum_{i \in \mathcal{N}} \lambda^i u|^2 = \sum_{i \in \mathcal{N}} \int_D |u^i|^2 + \sum_{i \neq j} \int_D u^i \cdot u^j$$
$$\le \sum_{i \in \mathcal{N}} \int_D |u^i|^2 + \sum_{i < j} \int_D (|u^i|^2 + |u^j|^2)$$
$$\le (k+1) \sum_{i \in \mathcal{N}} \int_D |u^i|^2,$$

hence (20). ◊

Now call $\gamma(m)$ the grain of the mesh. Recall the standard result of Chapter 4,

$$E(b^G{}_m, h_m) \le C_1\, \gamma(m)^2,$$

where the constant C_1 (like all similar ones to come) depends on D and the data, but not on the mesh. If one can prove that

$$(21) \qquad \int_{D^i} \mu^{-1} \, | b^i - \mu\chi^i h_m |^2 \le C_2 \ \mathrm{vol}(D^i)\, \gamma(m)^2,$$

it will follow from this and the lemmas that

$$E(b^L{}_m, h_m) \le C_3\, \gamma(m)^2,$$

a quite nice prospect, since the convergence speed would be the same, with less computation, than with the global complementarity method, only of course with a larger multiplicative factor in the error estimate.

So all depends on (21) being true, which seems likely, but I have no proof.

Another legitimate question is, could the local correction be obtained in a more straightforward way? After all, we know μh in D^i, so we have a fair estimate of the fluxes of b through faces of ∂D^i. Why not just look for the b in $W^2(D^i)$, with these fluxes as Dirichlet data, that is closest to μh in energy? What we get this way is a local b directly, not something like the earlier b^i, which was only the local contribution to $b = \sum_i b^i$. It's much better, in a way, since if we are interested in b around specified isolated nodes, *one* computation, instead of about 13 to 15, will be required for each of these.

But while this may be all right if all we want is some local divergence-free correction of μh, in order to make good-looking, flux-preserving displays, for instance, such a procedure is not able to yield a *globally* valid b: Even if we implement it on a family of D^is which pave D, and try patching the results together, the b thus obtained will not be div-conformal (and hence, not divergence-free over D), because the boundary fluxes taken as data on any given face differ for the two computations in the domains D^i and D^j adjacent to that face. (The variant in which the fluxes are used as Neumann data instead has the same drawback, and more degrees of freedom.) This procedure, therefore, is of no value as regards error bounds.

AUTHOR INDEX

SUBJECT INDEX